Short-Term Occupancy Prediction at the Ottawa Hospital Using Time-Series Data for Admissions and Longitudinal Patient Data for Discharge

Lon Michel Luk Arbuckle

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of Master of Science in Mathematics [1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

[1] The M.Sc. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

The Ottawa Hospital cancels hundreds of elective surgeries every year due to a lack of beds, and has an average weekday occupancy rate above 100%. Our approach to addressing these issues, by way of informing administrators of resource needs, was to model the flow of patients coming and going from the hospital.

We used administrative data from the Ottawa Hospital to build a time-series model of emergency department admissions, and studied models that would predict next-day discharge of patients currently taking up hospital beds. In the latter, we considered population-averaged models for groups of patients based on their primary medical condition, as well as subject-specific models. We included the random effects from subject-specific variation to improve on predictive accuracy over the population-averaged approach. The result was a model that provided more realistic probabilities of discharge, and stable predictive accuracy over patient length of stay.

Acknowledgements

Carl van Walraven proposed the project, lead the charge for access to hospital data, and provided general support and a great positive attitude, for which I am very greatful. Alan Forster and the data warehouse support team gave us access to clean data, made variables workable, and helped us understand the information available. Without their help this project would not have been possible.

I wish to thank Rafal Kulik for providing support to the project as a whole and the writing of the thesis, as well as Mahmoud Zarepour for his critical eye and reviewing the thesis. Thank-you to Jason Neilsen and Pierre-Jérôme Bergeron for reviewing the thesis and providing valuable feedback.

Finally, I wish to express my deepest gratitude to my wife and children for their love and support. Without them life would have no meaning.

Contents

List of Figures

List of Tables

Introduction

"Surgical wait times increase in Canada", stated the headline of a leading article in the Ottawa Citizen on December 6, 2010. It went on to say that "two dozen elective surgeries had to be cancelled [on Monday] so the hospital could help patients with urgent needs." "This happens regularly, almost daily now," said Dr. Jeff Turnbull, the Ottawa Hospital's chief of staff. "Doctors and nurses are working very hard to keep the system going despite all these pressures. I can't see it getting better and we're very worried about that" [9]. According to The Ottawa Hospital 2010-2011 Annual Report, the hospital cancelled hundreds of elective surgeries every year for the last four due to a lack of beds (no data is provided before this time). In that same four-year span, the average weekday occupancy rate was over 100% in every year [34].

Currently an administrator at the Ottawa Hospital decides how many beds will be available for elective surgeries the next day. No analytic methods are used, highlighting an opportunity to increase the presence of automation and decision support tools. We believe that hospital administrators could better manage patient beds if they knew how many patients were to be admitted, and which patients currently in hospital were to leave the next-day. Managing hospital resources and beds could be significantly improved if the demand on those resources could be accurately predicted.

Our objective was to study the feasibility of using predictive models to predict next-day emergency department admissions and patient discharge. The two problems

are distinct, and required very different modelling approaches. The goal, however, would be to combine these two models with scheduled arrivals to predict hospital occupancy, in order to allocate resources more effectively. At the time of this study we did not have scheduled arrivals, based on transfers from other healthcare facilities or elective surgeries, and were therefore unable to combine results into a single prediction of the hospital's occupancy rate.

In Chapter 1 the data warehouse is described, and administrative data in general, to understand the context in which modelling and validation can take place. The data warehouse is, in a sense, our proxy to the hospital itself. It comprises administrative data that are collected for the day-to-day operations in the hospital. Ideally data would be collected specifically for predictive modelling, based on the knowledge of medical experts and previous modelling work. However, administrative data is easy to come by, as hospitals in Canada are required to collect information used for precise billing of medical procedures, and for auditing purposes. The downside is that the data is not always what one expects, or wants.

The data warehouse is pulled from many departments with different coding practices and goals. This in itself poses challenges, although the team in charge of managing it helped guide us as much as possible. For example, variables used in other studies were recommended, and in many cases the the job of getting useful information out of the data warehouse was greatly simplified by using previously written macros to derive variables of predictive value (such as laboratory scores, and indexes of disease severity). Nonetheless, detailed descriptions of the variables in the data warehouse, of which there are many, were not available and we were therefore not always sure what data would be in them.

It is worth noting that the data warehouse contains highly confidential information, and as such the hospital's Research Ethics Board (REB) must approve any and all access and use (according to the Health Information Protection Act, 2004). That is, permission to see the contents of variables must be requested and approved

before any summary statistics or the like can be obtained. For example, we originally thought we had found and requested useful variables for building a predictive model of emergency department admissions, only to learn that the variables contained free-form text (i.e., notes typed out from a chart, rather than a finite number of categorical values).

With a basic understanding of the data warehouse we next delve into the literature to find what others have done to model hospital admissions and discharge, with a particular interest in administrative sources of data. Data mining tools are naturally out of the question as access to the data is strictly guarded and only specific data elements can be requested (not open access to data sources). Concerns over requesting an excessive number of variables, should the study be denied due to risks of identity disclosure (see, for example, the online Canadian REB Wizard [14]), certainly had an influence on the types of models we considered. Also, a large number of variables can lead to problems in and of themselves, such as correlation between variables, and the need to consider interactions.

Given a lack of suitable variables for modelling next-day admissions, we restricted ourselves to daily counts, and time-series models. In terms of modelling next-day discharge, the simplest approach is arguably a logistic regression model. In our case, however, we wished to track patients over multiple days, in a longitudinal data framework, thus losing the basic assumption of independent observations. We therefore considered ways to modify the basic logistic model to account for the correlated observations in a longitudinal study. As such, we must also present metrics appropriate for all model types we considered. Finally we describe the creation of the datasets from the data warehouse, including details of the variables we ended up using.

It is worth noting, given the prominent role it played in this study, that we incorporated primary condition groups into our models of patient discharge. Primary conditions are groups of broad diagnostic categories, used for the sake of analysis. The groupings are based on the relative similarity of the diagnoses, and on observed

mortality rates (see Appendix A for a description of the diagnostic categories). We used primary conditions to group discharge data and results to determine how the data and results varied by group. We did this to take full advantage of the domain knowledge that indicated that primary conditions are an important grouping variable.

Before investigating models of hospital admissions or discharge, we needed to understand the variables created from the data warehouse and the dataset itself. In Chapter 2 we describe our data and covariate selection methods used in order to simplify our models in advance. We also describe how we reduced our data to ensure a relevant sample size was used, but without exhausting computing resources or overfitting our models. Given the confidential nature of the hospital data we worked with, and that it is not possible to share this data in any other way, we also provide many detailed plots of important variables and relationships. Differences between primary condition groups, in particular, are considered throughout.

We describe how we developed time-series models of emergency department admissions in Chapter 3. The primary source of unscheduled entry to the hospital is admissions from the emergency department. Combined with scheduled arrivals to the hospital, from transfers or elective surgeries, such models could be used to predict the number of patient admissions to the hospital on any given day. We begin with a brief overview of linear time-series model theory, then the seasonal extension, before presenting the results of our modelling efforts in Section 3.2.

We first attempted to predict next-day patient discharge using a population-averaged approach, described in Chapter 4. Our outcome variable was an indicator of whether a patient had or had not left the hospital within 24 hours. The population-averaged model was, therefore, a generalized linear model with binomial error distribution that essentially averaged over any random effects from subject-specific variation (using generalized estimating equations). We discuss the fitted model itself in Section 4.4, after a discussion of supporting theory. This approach did not meet our criteria of predictive accuracy, and the probabilities of discharge for individuals were

low and seemingly unrealistic. We were, however, able to improve on these results with subject-specific models.

In Chapter 5 we describe our efforts to model the random effects between patient encounters and improve predictive accuracy. Subject-specific models are well established for normal responses (conditional on predictors), but the extension to categorical data is relatively recent. Although we build on the theory of generalized linear models from the preceding chapter, estimating the random effects, or predictors, adds significant technicalities. We therefore present supporting theory before discussing the fitted models in Section 5.3. We also show how our subject-specific model, with random intercepts for primary conditions and for patient encounters, improves on predictive accuracy compared to the population-averaged models.

We did not include interactions between covariates in this study (i.e., we only considered main effects in our models). Often in smaller longitudinal studies, covariates that depend on time (e.g., a covariate for time) are included in interactions with covariates that do not depend on time (e.g., a covariate for gender, resulting in the interaction gender*time). However, the main-effects considered in this study were numerous, and resulted in a model that was stable and fast on the most basic of desktop computers. Although analytics at the hospital are deployed on servers, experimental models are not permitted on these. Only when we considered interactions with time did we find the computation time prohibitive and run into memory overflows with the more advanced subject-specific models, and even the population-averaged models were found to be unstable in this case. We were nonetheless able to demonstrate the feasibility of predicting next-day patient discharge using predictive models, which was our main goal.

Chapter 1

Developing Models for Hospital Occupancy Prediction

Pioneers in healthcare information technology are including automation and decision support tools into electronic health records[6]. Before one can tackle such an endeavour, it is important to understand the type of data hospitals collect, in order to understand the context in which models will need to exist, as well as limits and expectations. We introduce the "data warehouse" that collects the administrative data used in this study, before we discuss models that can be used to predict patient admissions and discharges. We then delve into modeling strategies specific to our problem, considering previous research to find opportunities for improvement, and how we chose to evaluate the models we eventually considered. Finally, we explore the specific data available to us for this study.

1.1 Rationale for the Study

Currently an administrator at the Ottawa Hospital decides, at 3 p.m., how many beds will be available for elective surgeries the next day. The decision is made based on

current hospital occupancy, time of year, basic trends noticed by the administrator, etc. The decision does not use any analytic methods, at least not to the extent of explicit mathematical or statistical models. We believe there is an opportunity to create a decision support tool to help guide this decision.

A predictive model of short-term hospital admissions and discharge (i.e., live discharge, transfer out of hospital, or in-hospital death) could be used by hospital administrators as an aid to determine the number of beds to allocate for elective surgeries. As seen in Figure 1.1—a simplified flow diagram of hospital occupancy—emergency admissions represent the unknown entry point to the hospital. Therefore, a predictive model of emergency admissions to the hospital would aid in determining the rate of admissions to the hospital, in an effort to maximize occupancy and allocate resources.

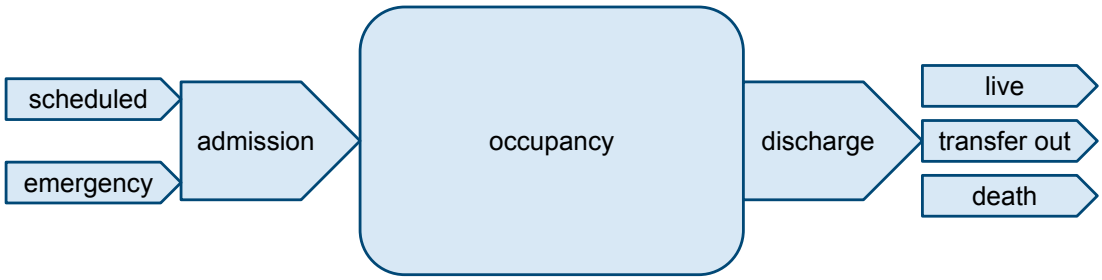

Figure 1.1: Admission versus discharge.

An example of fluctuations in daily hospital occupancy at the Ottawa Hospital is given by Forster et al. in [16]. In that study, average hospital occupancy from 1993–1999 was about 90%. Although a descriptive statistic of the variation in daily hospital occupancy was not reported (such as standard deviation), a time-series plot shows considerable day-to-day variation (with points at which occupancy fell below 70%, and rose above 100%).

It is interesting to note that the study in [5] found that increased automation and decision support tools resulted in "reductions in mortality, complications, and

costs" across 41 hospitals in Texas. This strongly suggests advantages to increasing automation in hospitals across Canada.

1.2 Data Collection and the Data Warehouse

Administrative data are collected at the hospital to ensure that financial and legal outcomes take place when there is a health-care transaction. An electronic database is used to record these transactions. Although clinical content of administrative data can be limited to patient diagnoses and codes of procedures, they are nonetheless useful in evaluating and improving the quality of health care.

At the Ottawa Hospital, the data warehouse stores the electronic transactions containing patient information from dozens of different databases in the hospital. The information is therefore integrated into a single source for both research and day-to-day operations. Administrative data are not, however, primarily collected for research—they are collected for the day-to-day operations in the hospital. This can limit the quantity and quality of clinical data available to the researcher. Also, given the size and complexity of hospital operations, it is not surprising to find "that significant inaccuracies in administrative data are common"[37].

Nonetheless, in the same way that data are needed to support evidence-based medicine, data are also needed to support evidence-based health services. "Only by measuring and comparing data—ranging from death rates to how much a patient's vision improves after cataract surgery—do we get the evidence we need to determine where there are quality problems and where there are effective approaches to care that more people should be using"[11].

1.2.1 Timeliness of Data

There are close to 900 variables in the data warehouse, providing a wealth of information for building predictive models. However, not all variables are available in time for predicting events the next day. In other words, the value of a variable in predicting events the next day depends in part on when that variable is entered into the hospital's administrative system. For example, patient abstracts cannot be used for predicting patient discharge, since they are created after the patient leaves the hospital and are therefore not available in time for prediction.

Of course that some variables not currently available in "real-time" could become available in the future. Increasing automation at the hospital, as described in [5], means that more and more data are becoming available in a timely fashion. As a result, variables not currently available for predictive models may be in the future. Therefore, we considered some variables from patient abstracts to see how they could be used, which will be described when we discuss relevant covariates.

Patient data are updated regularly, so information contained in the data warehouse may only reflect that last change. The census history records changes to patient data in the encounter table, but this adds to the confusion in terms of which variables to use and when. For the sake of this study, certain compromises needed to be made in order to simplify the data. Namely, it was decided to consider data on a daily basis, capturing information at midnight.

Timeliness of data is especially important when considering admissions from the emergency department, as it can occur very quickly. For example, a patient could be almost immediately admitted to hospital upon arrival, or be subject to extended wait times (typically less than 12 hours), depending on severity of illness. Given such a short stay in the emergency department, it is questionable whether variables would be available in time to predict admission at the subject-level. Therefore, it was decided that daily admissions from the emergency department would be modelled instead.

1.3 Modelling Strategies

1.3.1 Variable Selection and Access to Data

The issue of patient identification and confidentiality of health information means that individual analysts and researchers are limited in their knowledge of variables outside their immediate area (since they do not have access). Analysts and researchers on the data warehouse support team, however, offered valuable domain knowledge in the course of this study—important in the selection of variables. We also used the study by Escobar et al. in [15] for ideas, which will be described later.

Variable selection for modelling admissions from the emergency department is limited by the lack of reliable variables. Many variables are text entries (due to time constraints), and it was not our goal to develop keyword search criteria to extract information that could be used to create useful variables for modelling admissions (unlike the study [28]). We therefore limited ourselves to a daily count of admissions from the emergency department, without additional variables to model admissions.

To model patient discharge, the study in [28] used 30-60 covariates in logistic regression models of discharge, using an automated variable selection method, claiming they wanted the best predictions. For this many covariates, however, estimates will be unreliable, especially for small wards (as described by Gelman and Hill in [18], Chapter 1). And more variables does not imply better prediction in this type of model. For example, there is the concern of multicollinearity (i.e., strong linear dependencies between covariates, resulting in unstable coefficients for the variables that are collinear: standard errors may be large, and coefficient estimates may be small and therefore not practically significant, whereas those variables have a strong effect as a group, described by Allison in [3], Section 3.5).

Due to privacy concerns, the Ottawa Hospital does not allow access to variables in the data warehouse without full review by the data warehouse team, and then the

hospital's research ethics review board. Therefore, variables of interest—for creating suitable covariates in modelling discharge and admissions—needed to be selected based on domain knowledge, relevant literature, and some guesswork. Definitions of variables in the data warehouse are either not available, or with insufficient detail to understand how they might be used and what they contained (something the data warehouse team would eventually like to address). Also, no descriptive statistics could be provided for any of the variables in the data warehouse.

1.3.2 Model Types

Past research has shown mixed results trying to model hospital admissions—however, we found that this largely depended on their definition of admissions. In a recent study in [1], the researchers had good success with a time series regression model predicting emergency department occupancy rates (also shown in [46]), but suggested that admissions from the emergency department to hospital ward was not predictable (i.e., mostly random). Their definition of admissions, however, seem to be to a ward other than the emergency department, and not just to hospital in general. In other words, in their model of admissions, they seem to have excluded patients admitted to hospital but still in the emergency department. Under these circumstances, it is surprising they did not include hospital occupancy rates in their models, which was shown by Forster et al. in [16] to directly influence admission to a hospital ward. Also note that the study in [1] only considered admissions from the emergency department that resulted in an overnight stay.

We instead modelled admissions to hospital including patients that remained in the emergency department until discharged or transferred to a ward, and regardless of their length of stay. Given a lack of variables for modelling admissions, we began our study of admissions from the emergency department by considering linear time-series models and their extensions—that is, auto-regressive moving average (ARMA)

models and their integrated (linear) and seasonal (nonlinear) extensions, in Chapter 3.

In developing a logistic regression model of discharge, Littig and Isken in [28] make no mention of correcting for the longitudinal structure of the data (i.e., the repeated measures taken of patients over multiple days). If this has not been taken into account, standard errors will likely be underestimated, test statistics overestimated (because they are divided by underestimated standard errors), and coefficients will not be statistically efficient (i.e., there exists other methods that will produce estimates with smaller standard errors), as described by Allison in [3], Chapter 8. It is possible that they did take clustering into account, using perhaps generalized estimating equations, but the method used should be described as there are important details that need be considered. For example, the approach of generalized estimating equations makes full use of the data but does not adjust for bias from omitted covariates, whereas conditional logit models are the opposite (even though both correct standard errors and test statistics for clustered data).

Also, the authors in [28] highlight that they did not take covariance between patients into account, largely due to concerns regarding computation time, given the number of models they were running besides discharge and admissions. They recognize that this is problematic, as hospital staff suggested that hospital occupancy may affect patient discharge rates. They attempted to adjust for this factor using a "meta-model" (not to be described here).

A random coefficient, or multilevel, model lends itself well to structured data such as that in a hospital, and is very effective for prediction. The basic idea is to treat clustering as a random effect in a mixed model. For example, patient discharge will vary depending on hospital ward, meaning we need to consider group effects. A two-stage model, with group indicators and a regression of group effects on group-level covariates, can be used to predict outcomes for a new group, even if sample sizes are small for some groups (Gelman provides further details in [17]).

With a random coefficient model we can include indicators for clusters at multiple levels of a hierarchy, without having to worry as much about overfitting than with classical regression. It would make it possible to include all covariates in the model, since coefficients for each set of group indicators are fit to a probability distribution (i.e., a variance component is estimated for the set of group indicators, in order to predict random effects of clusters). Such a model also ensures a more accurate measure of predictive uncertainty, since we can deal with correlation between patient live discharges, transfers or deaths on a given day in a convenient way (explained by Gelman and Hill in [18], Chapter 1). It is also efficient and corrects for the effects of unobserved heterogeneity (whereby the exclusion of important covariates will lead to logit coefficients attenuated toward 0, due to increased variability of the random disturbance), described by Allison in [3], Chapter 8.

We therefore began our study of discharge by considering a similar approach to [28]—that is, using logistic regression—but included generalized estimating equations to correct for clustering in the longitudinal patient data—this will be our population-averaged model, described in Chapter 4. Building on that, we will then introduce a multilevel modelling approach for building a subject-specific model in Chapter 5.

1.3.3 Evaluating the Admissions Models

We consider measures of predictive accuracy for time series models in this section, following Hyndman et al. in [23].

Mean Absolute Percentage Error

One of the more common measures of predictive accuracy is the Mean Absolute Percentage Error (MAPE). If, at time t, y_t is our outcome (i.e., the number of admissions from the emergency department), and e_t is the error between prediction and outcome, then the percentage error is given by $p_t = 100e_t/y_t$, and MAPE is simply mean($|p_t|$).

This measure has the advantage of simplicity, and is independent of scale, but can prove problematic when the outcome y_t has zeros (which is not the case with our data). More important, however, is that basing the measure on percent error may not be the most desirable or relevant—for example, an error of 100% when the mean outcome is low, say 100, can seem less practically significant than an error of 100% when the mean outcome is large, say 10,000, depending on the context.

Mean Absolute Scaled Error

Another way to scale the measure of predictive accuracy is to use a relative error based on a naïve approach to prediction—i.e., a random walk. In this case the predicted value at time t is the outcome at the previous time step, y_{t-1}. Therefore, the relative error is e_t/e_t^{rw}, where $e_t^{rw} = y_t - y_{t-1}$ is the prediction error of the random walk. Note that e_t^{rw} can be be small, with a positive probability density at 0 resulting in infinite variance of the relative error.

Yet another option is to use a relative measure, instead of a relative error. For example, the mean absolute error, mean($|e_t|$), can be divided by the mean absolute error for the random walk, mean($|e_t^{rw}|$). An advantage of this method is that it is easily interpreted: a value less than one means that predictions are better than predictions from the random walk; a value greater than one means that predictions are worse than predictions from the random walk. However, relative measures cannot be used to measure predictive accuracy when only a single prediction point is being considered beyond in-sample data (i.e., beyond the data on which the model was built).

Rather, we considered a method proposed and recommended in [23], which scales the error using the mean absolute error of the random walk using the in-sample data (i.e., the data on which the model was built). Therefore, we scale the prediction error e_t by dividing $\text{mean}_{s\in\{2,\ldots,n\}}(|e_s^{rw}|) = \sum_{i=2}^{n} |y_i - y_{i-1}|/(n-1)$ (where the in-sample

data are made up of n observations). The scaled error is therefore less than one when predictions are better than in-sample predictions from the random walk, and greater than one when predictions are worse than in-sample predictions from the random walk.

The Mean Absolute Scaled Error (MASE) is then the mean of the absolute scaled errors. This measure is independent of scale, easy to interpret, and only degenerate when all in-sample data are equal (i.e., $e_t^{rw} = 0$ in-sample, resulting in a divide by 0). As it is broadly applicable, we included it in our evaluation of admissions models.

1.3.4 Evaluating the Discharge Models

In order to compare models of patient discharge, we need test statistics that work with the range of models we consider—from a simple logistic regression to a multilevel approach. And given that our focus was on developing predictive models, not explanatory models, we needed appropriate statistics for quantifying predictive ability.

Coefficient of Determination

We included a version of a generalized R^2 when considering simple logistic regression, which is equivalent to the classical R^2. In formulating the likelihood function, we assume that observations i are independent with binomial distribution, such that for the vector $\boldsymbol{y}$ of binary events it has the form

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_i \pi_i^{y_i}(1-\pi_i)^{1-y_i},$$

where $\boldsymbol{\beta}$ is the parameter vector for the model, π_i is the probability of success, and $1-\pi_i$ is the probability of failure (we ignore the aggregation of events used to simplify computations, leading to a strictly binomial rather then Bernoulli representation, as shown here). Note that the $\pi = \pi(\boldsymbol{\beta}; \boldsymbol{X_i})$ are in fact functions of the covariate values $\boldsymbol{X_i}$ and parameters $\boldsymbol{\beta}$(through the logit function, which we will see in Chapter 4).

If we let L be the likelihood estimate for the current model, with fixed parameter estimates, L_0 be the likelihood estimate of the model with intercept only (i.e., with no other covariates in the model), and n be the total number of observations, then the generalized R^2 is given by $R_g^2 = 1 - (L_0/L)^{2/n}$, with maximum attainable value of $R_{max}^2 = 1 - L_0^{2/n}$. Therefore, following Nagelkerke in [35], we used a scaled version given by

$$0 \leq R_N^2 = \frac{R_g^2}{R_{max}^2} = \frac{1 - (L_0/L)^{2/n}}{1 - L_0^{2/n}} \leq 1$$

(whereby larger values indicate a model that should have better predictive accuracy).

It is important to note that the generalized R^2 for logistic regression does not quantify the proportion of explained variation in the way classical R^2 does for ordinary linear regression. Harrell explains in Section 9.8 of [20] that in the case of a binary response R^2 should not be used to check fit, but to quantify predictive discrimination (i.e, "the model's ability to separate subject's outcomes", Section 5.2.2 of [20]). Also, given that it is based on maximum likelihood, it is not applicable to the model once generalized estimating equations are used (to account for clustering of the longitudinal patient data), or to the multilevel model we developed.

Receiver Operating Characteristic Curves

For an alternative measure of predictive discrimination, we turned to a popular method for evaluating binary classifiers, as noted by Agresti (Section 6.2.6 of [2]), known as Receiver Operating Characteristic (ROC) analysis. ROC curves plot the true positive rate, or sensitivity, against the false positive rate, or 1-specificity, as the threshold is varied (the threshold being the probability level used to distinguish between positive and negative outcomes). This is perhaps best understood by considering the plots in Figure 1.2.

Although we plotted ROC curves together to compare different models, we also used the area under the ROC curve as a summary statistic of predictive discrimi-

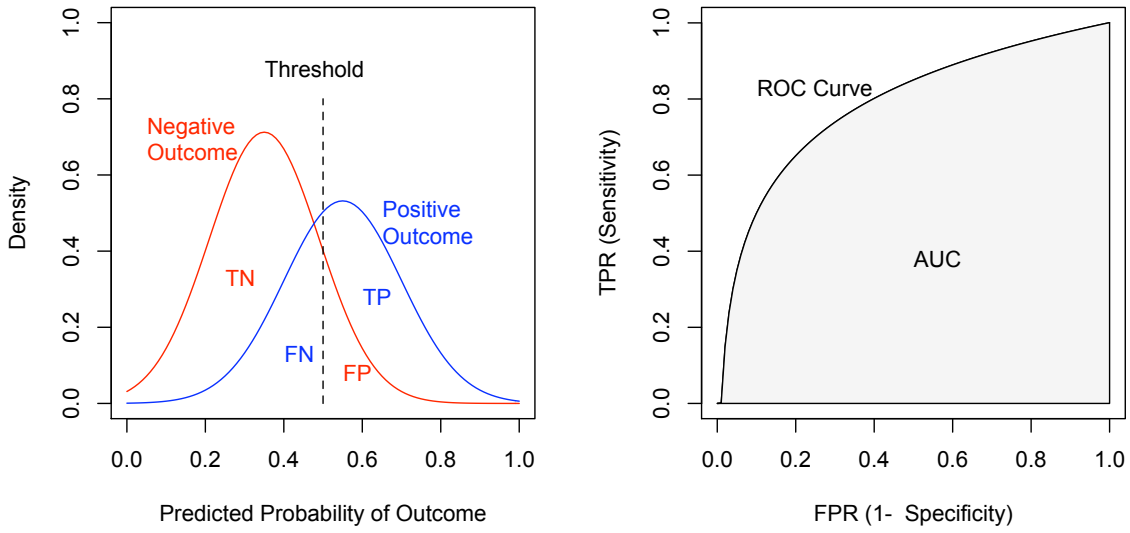

Figure 1.2: Receiver Operating Characteristic (ROC) curves are created by plotting the True Positive Rate (TPR) versus the False Positive Rate (FPR) of a binary classifier. Where true/false (T/F) positives/negatives (P/N) indicate the number of such outcomes, we have that TPR=TP/(TP+FN) and FPR=FP/(FP+TN). Each threshold value (distinguishing the classification of negative and positive outcomes) translates into a point on the ROC curve. The Area Under the ROC Curve (AUC) is often used as a measure of predictive power.

nation, which is identical to the concordance probability—often called the c-index, or c-statistic (included as standard output in many implementations of logistic regression, such as the `logistic` procedure in SAS). Considering all concordant pairs (negative outcomes paired with positive outcomes), the c-index is the probability that the pairs remain concordant (i.e., the lower predicted probability being assigned the negative outcome; the higher predicted probability being assigned the positive outcome).

Although the c-index can be found by approximating the area under the ROC curve, one other way is to use a derivation from the Mann-Whitney-Wilcoxon rank-sum test described in [30], [52]. Rank the predicted probabilities and sum the ranks paired to a positive outcome (call this R_1). Let n_0 be the number of negative outcomes, and n_1 be the number of positive outcomes. We then calculate $U_1 = R_1 - n_1(n_1 + 1)/2$, where $n_1(n_1 + 1)/2$ is the sum of $1, \ldots, n_1$ (the sum of n_1 ranks). The c-index, or area under the curve, is the then $AUC = U_1/(n_0 n_1)$. This is the derivation we used, via the R function `somers2` {Hmisc package}. (If we were deriving the U statistic, of the Mann-Whitney-Wilcoxon test, then we would repeat the above for the negative outcomes to find U_0, and add the two to get $U = U_0 + U_1 = n_0 * n1$.)

The c-index is "a unitless index of the strength of rank correlation between predicted probability of response and actual response"(Section 10.8 of [20]). A value of 0.5 is equivalent with random predictions, and models with a value above 0.8 are deemed to have "some utility", according to Harrell (Section 10.8 of [20]). A value of 1.0 is equivalent with perfect predictions, in the sense of having perfect separation between positive and negative outcomes. It is perhaps worth noting that the c-index has a direct relationship with Somer's D_{xy} rank correlation, by $D_{xy} = 2(c - 0.5)$.

More important, however, is noting that the c-index is a rank-order statistic, and will therefore not distinguish between a pair of subjects with predicted values of 0.01 and 0.99 versus 0.2 and 0.8 (where these are concordant pairs: the first subject in

the pair, with the lower probability, has a negative outcome; the second subject in the pair, with the higher probability, has a positive outcome). On the other hand, rank-order statistics are insensitive to the proportion of positive outcomes, and the ROC curve itself is invariant to monotone transformations of the outcome. We used the c-index to rank predictive performance of the competing models.

Brier Score

In contrast to a rank-order statistic, we also considered a measure of the error rate, known as the Brier (or quadratic) score, to assess predictive accuracy. It is the mean-squared error of predicting binary outcomes with probabilities, and combines calibration, "the statistical consistency between the predicted probability and the observations", with sharpness (or discrimination), "the concentration of the predictive distribution" [41]. If y_i is the binary outcome (with values of 0 or 1), and p_i is predicted probability, for observation i of a total of n observations, the Brier score is simply $\sum_{i=1}^{n} (y_i - p_i)^2 / n$. The Brier score is called a (strictly) proper score function since the best score is achieved when the predicted probabilities equal the binary observations. Clearly, a lower score indicates better model accuracy (from 0 for perfect predictions to 1 for awful predictions).

1.4 Creating a Dataset from the Data Warehouse

Creation of a dataset for the purposes of this study were approved by the Ottawa Hospital Research Ethics Board, with the support of the data warehouse support team. Figure 1.3 describes the data model used for the data warehouse, with the relationships between tables. The data warehouse integrates data from multiple sources at the Ottawa Hospital, storing it in a central location in such a way to facilitate research endeavours. Each table includes columns, or variables, which may be used either directly or indirectly (to create new variables) in analyses—resulting in almost

900 variables in the data warehouse.

1.4.1 The Data Warehouse

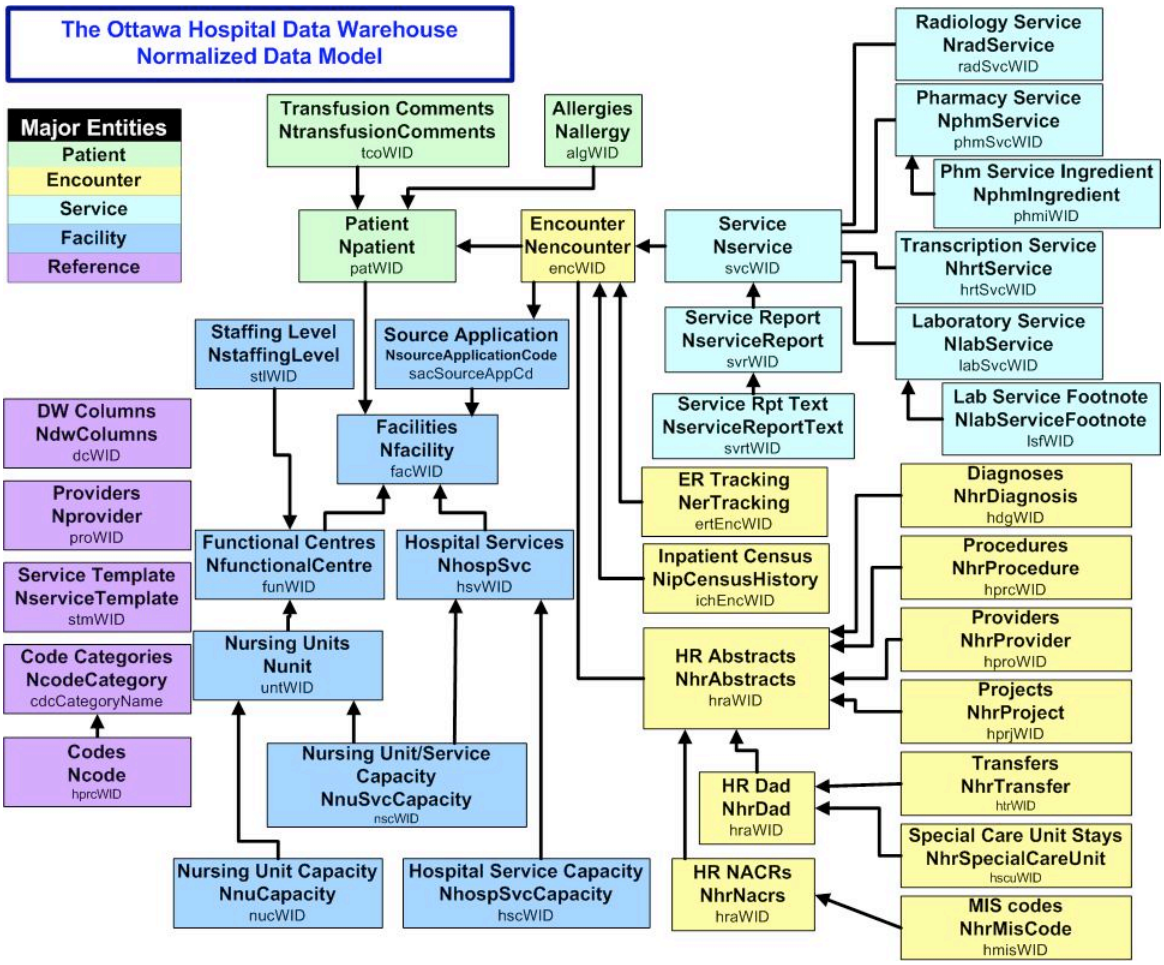

Figure 1.3: The data warehouse is a relational database that integrates several sources of information for the sake of analyses.

The most prominent source of data for this study was the encounter table (Nencounter), which describes patient level information regarding their encounter with the hospital. Each row of this table represents a single patient encounter. Data for this table mostly comes from registration systems. Variables in this table include

demographic details, the type of encounter, patient details, and disposition at time of discharge. Also of importance is the census history, (nIpCensusHistory), which is a record of changes made to the encounter data—crucial for creating the longitudinal data required in this study.

The diagnosis table (NhrDiagnosis) includes diagnosis type, grade, category, and class, wherein each row represents a diagnosis. It includes far too much detailed information to be used directly, but can be used to create comorbitity scores or for grouping patients by large diagnostic classes (both of which will be discussed shortly). It is important to note that diagnostic coding is abstracted data from patient files after they exit the hospital. This information could, however, be provided during patient admission, at least in principle. Also, for the sake of the variables needed for our model, only a subset of this information would be required at admission. Similarly, the Discharge Abstract Database (DAD) table (NhrDad) contains value-added output from abstracted data, such as case-mix groups, which were considered.

An important source of information for the daily evaluation of patient status is contained in service records (Nservice), which includes things such as radiology or pharmacy, but once again it contains too much detailed information to be used directly. Therefore we instead focused on lab results (NlabService), and used a score to evaluate patient well-being (described in the next section).

Admissions focussed on the Emergency Department (ED), the largest source of unscheduled admissions to the hospital, through the emergency tracking table (NerTracking). This table includes arrival, triage, and departure times, triage codes, presenting complaint, etc. Although these variables seem extensive, much of the detailed information regarding the patient, such as the admitting diagnosis, is in the form of text descriptions (not coded variables that can be used to create useful variables for modelling patient flow). This is understandable, given the timelines dealt with in the emergency department, whereby patients admitted to hospital are done so within 24 hours of presenting themselves to the emergency department.

1.4.2 Description of Population

No patient recruitment was required. This study used retrospective administrative data from the Ottawa Hospital Data Warehouse. For the discharge model, data included all discharge patients (i.e., live discharge, transfer out of hospital, or in-hospital death) over the age of 15, from 1 April 2006 to 31 March 2009, resulting in 122,883 patient encounters, and 987,163 observations (captured at midnight). Excluded were cadavers, stillbirths, and neonates, as well as records with incomplete abstracts and for which abstracts and encounters have discrepant admission and discharge dates. For the admissions model, data included a daily count of all patients over the age of 15 that were admitted to the hospital from the emergency department from 1 April 2004 to 31 March 2009, from the civic and general campuses only (the heart institute being excluded from consideration).

1.4.3 Variables Included

Having described the data warehouse and the most prominent sources of data for this study, we list the variables that were created for the discharge model and from what sources in Table 1.1, and their description in Table 1.2. Foreign keys—the unique references between tables in a relational database, used for linking entries—are not included (for example, encounter ID's, service ID's, etc.). This section does not discuss variables for the admissions model, as only a daily count of patients admitted to the hospital from the emergency department was used in this case.

Socio-economic factors were included where possible, as they have been shown in the literature to be good covariates of a range of health problems. For example, marital status, religion, and occupation are all covariates for wealth attainment. Postal codes are sometimes used to estimate socio-economic status but were excluded because they are of questionable value according to a study in [13]. They are also highly confidential and would require linking to Statistics Canada data to determine

Table 1.1: Source of variables from the data warehouse requested for modelling discharge.

Variable Name	Type	Table Name(s)	Column Name(s)
accident	indicator	Nencounter	encAccidentInd
accommodation	nominal class	Nencounter	encAccommodation-Given
admitPriority	nominal class	Nencounter	encAdmPriority
admitType	nominal class	Nencounter	encAdmType
age	continuous	Nencounter	encPatAgeAtAdm
ambulance	indicator	Nencounter	encAmbulanceInd
campus	nominal class	Nencounter	encCampusCd
charlson	ordinal class	NhrDiagnosis	hdgCd, hdgType
day	continuous	NipCensusHistory	ichEffectiveDtm
exit	indicator	NipCensusHistory, NhrDad	hraExitCd
gender	nominal class	Nencounter	encPatGenderCd
icu	indicator	NipCensusHistory	ichOldHospSvc, ich-NewHospSvc
laps†	continuous	Nservice, NlabService	labSpecimenDtm, labSingleResult, labStsCd
married	indicator	Nencounter	encMaritalSts
nursStaOcc	continuous	NipCensusHisotry	ichChangeType, ichEffectiveDtm, ichNewNursSta
occUnknown	indicator	Nencounter	encOccupationCd
ohip	indicator	Nencounter	encRfpCd
primaryCondition†	nominal class	NhrDad, Nservice	hraAdmCatCd, hProSvc
regularAdm	indicator	Nencounter	encAdmRoute
religion	indicator	Nencounter	encpatreligioncd
retired	indicator	Nencounter	encOccupationCd
surgery	indicator	NipCensusHistory	ichOldHospSvc, ich-NewHospSvc
transferIn	indicator	Nencounter	encTransferFrom-InstitutionCd
urgentAdm	indicator	NhrDad	hraAdmCatCd

† Also adjusted for age and admission type.

Table 1.2: Description of variables from the data warehouse requested for modelling discharge.

Variable Name	Description (% of patient encounters)
accident	Admission due to an accident? (3%)
accommodation	Accommodation at admission: p=private (28%) s=semi-private (42%) w=ward (30%)
admitType	Admission type: d=direct (5%) e=emergency (42%) l=elective (32%) s=same day admission (10%) u=urgent (11%)
admitPriority	Admission priority: r=routine elective (42%) s=semi-urgent (5%) u=urgent (11%) x=emergency (42%)
age	Age at admission ($\geq$ 15)
ambulance	Arrived by ambulance? (29%)
campus	Campus: c=civic (39%) g=general (47%) h=heart (14%)
charlson	Co-morbidity score at admission, measuring the effect of additional diseases to primary condition (higher being worse): 0 (54%), 1-2 (23%), 3-4 (10%), 5+ (13%)
day	day in hospital (from day 1 to last day)
exit	Inpatient exits (by discharge, transfer, or death) hospital within 24 hours?
gender	male (58%) or female (42%)
icu	Is patient currently in ICU service? (5%, representing 25% of deaths)
laps	Laboratory-based Acute Physiology Score (LAPS)
married	Is married? (64%)
nursStaOcc	Number of patients divided by number of available beds for the nursing station
occUnknown	Occupation unknown? (8%)
ohip	Covered by Ontario Health Insurance Plan (OHIP)? (92%)
primaryCondition	44 primary conditions (main reason for hospitalization)
regularAdm	Regular admission? (87%)
religion	Declared religion? (38%)
retired	Is retired? (13%)
surgery	Is patient currently in surgical service? (18%)
transferIn	Transferred from another institution? (27%)
urgentAdm	Urgent admission? (62%)

average incomes for the different postal codes.

Variables with labels "new" in the census history were needed in an attempt to have timely records for prediction on specific days—the same variables from the encounter table are probably the last updated version when the patient was discharged from the hospital. We also needed the variables from the encounter table, however, in case there were no updates to the variables in the census history, and to compare quality of data.

Note that the surgery indicator is not an indicator of whether the patient is currently undergoing surgery, but is in the care of a surgical service—general, cardiac, neuro, oral, orthopaedic, plastic, thoracic, vascular, and cardiothoracic. Similarly, the ICU indicator is an indicator of whether a patient is in the care of the Intensive Care Unit, and cannot be active at the same time as the surgery indicator (they are mutually exclusive). Also, the percent of patient encounters for these variables, given that they are time-varying, is whether an encounter resulted in a stay in either of those respective services (i.e., not the percent of patient days in ICU or a surgical service, based on the total number of observation by patient days).

Derived Variables

A data warehouse support team member processed the necessary data required to calculate the derived variables described below. Therefore, only the derived variables were provided, not the variables needed for their creation. This was done to simplify model development by providing relevant covariates (developed for previous research work), and ensure patient confidentiality.

Newly admitted patients to the hospital are given an initial diagnosis by a medical doctor. This diagnosis is then coded according to the International Classification of Diseases framework (with over 16,000 admission codes, administered by the World Health Organization), and can be grouped into 44 primary conditions representing

related diseases and observed mortality rates (see the Appendix A for a description of the primary conditions). Primary conditions provided us with an important grouping variable (see Escobar et al. in [15] for further details).

As previously mentioned, we wished to use daily lab results to assess patient well-being. The Laboratory-based Acute Physiology Score (LAPS) is a continuous variable based on test results within 48 hours prior to admission or within 6 hours after admission, and therefore provides potentially timely information of patient health. Note that it is also adjusted for age and admission type, among other, laboratory and physiology-specific, factors. Details regarding its derivation can be found in [15].

The presence of one or more diseases, in addition to the primary disease and reason for hospitalization, is known as comorbidity. A comorbidity score attempts to measure the effect these additional diseases have on patient health. The most common comorbidity score is the Charlson index, which is proportional to the risk of one-year mortality [10]. We included the Charlson index to test how variables from patient abstracts could be used to improve the prediction of patient hospital discharge. Although diagnostic codes are not currently available for short-term prediction of patient events, increasing automation at the hospital suggests that they could at some future date.

Finally, nursing station occupancy was requested under the assumption that there is internal pressure at the hospital to discourage overcrowding, and optimize resource allocation. For example, the study in [16] demonstrated a correlation between patient length of stay (the total number of days a patient stayed in hospital) in the emergency department and hospital occupancy. The authors suggest that, in their sample, a hospital occupancy above 90% resulted in significant increases in emergency department length of stay. Similarly, we suspected that nursing station occupancy would affect patient length of stay for all departments. However, we used nursing station occupancy as opposed to department occupancy in order to focus the variable more closely on the patient.

1.4.4 Confidentiality

The utmost was done to ensure that data was protected against breaches in security and privacy. Access to all data was password protected within hospital property, and no data was stored at any time on portable computing devices. There were no printed copies of the data for this study. A data request form had been submitted to the data warehouse support team. Independent study numbers were not required, and the encounter-based population was uniquely identified by an encounter number (encWID). Information that was used was from variables in the data warehouse that were considered possible covariates of patient hospital discharge or admission. Patient identifying information was not included in the data set. Date of birth was not required—we used the patient's age at the time of the admission (to the emergency department or hospital). Diagnostic codes and procedure codes were not included in the data set.

Chapter 2

Dataset and Variables

Before describing our efforts to model hospital admissions and discharges, we need to understand the variables created from the data warehouse and the dataset itself. In the previous chapter we described their origin, and what was requested, but not any statistical descriptions as they were not available until after the request for data was made and approved.

2.1 Admissions Data

The admissions data were created by taking a daily count of patient admissions to the hospital from the emergency department. The result of this daily count was a time series for the civic campus, shown in Figure 2.1, and a time series for the general campus, shown in Figure 2.2.

The normality of the time-series data is important for the construction of linear models, which we will consider in Chapter 3. We needed to consider that the data did not have mean zero, and could exhibit a linear trend with time (which can, however, be easily handled with linear models). More important, however, is the seasonal variation of the data (although the day-to-day variation seems relatively constant

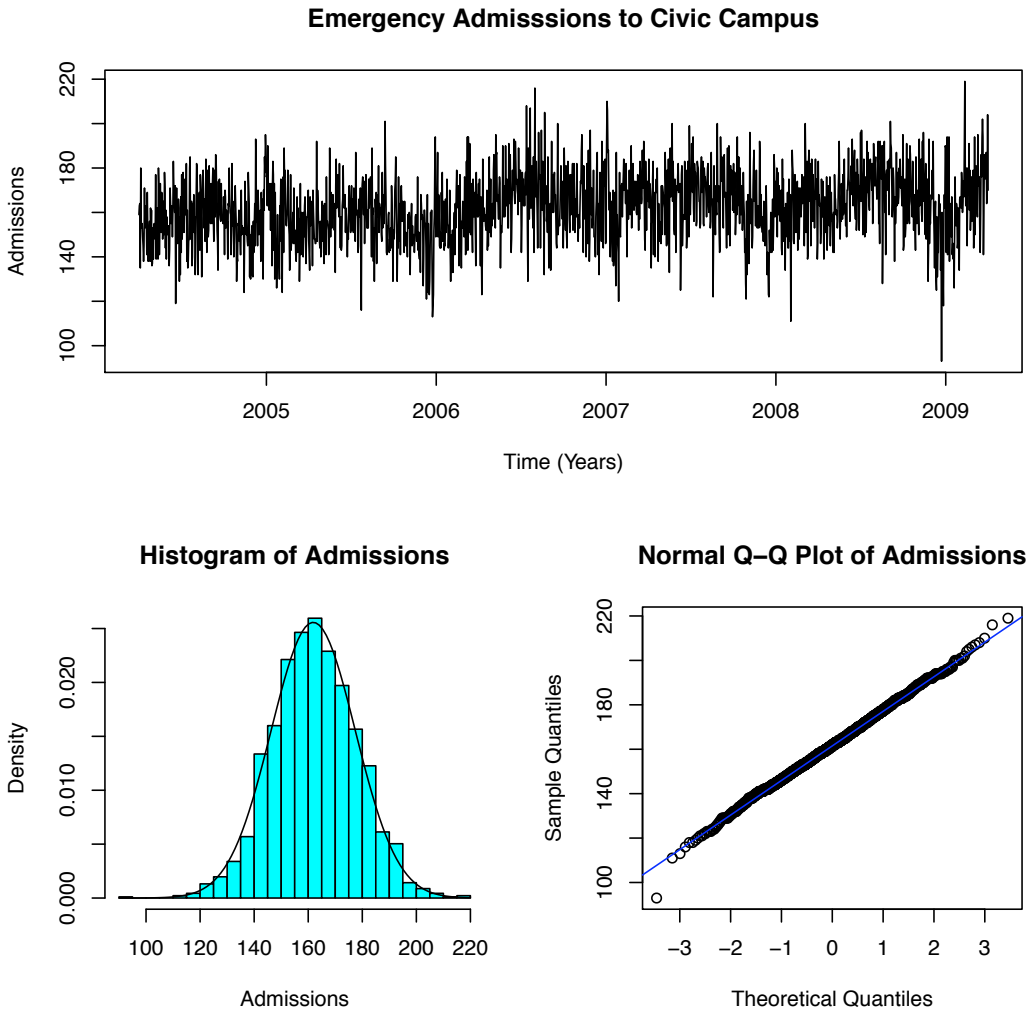

Figure 2.1: Histogram and q-q plot of emergency admissions to civic campus. Note that the time-series data appear to be normally distributed.

when following the seasonal trend).

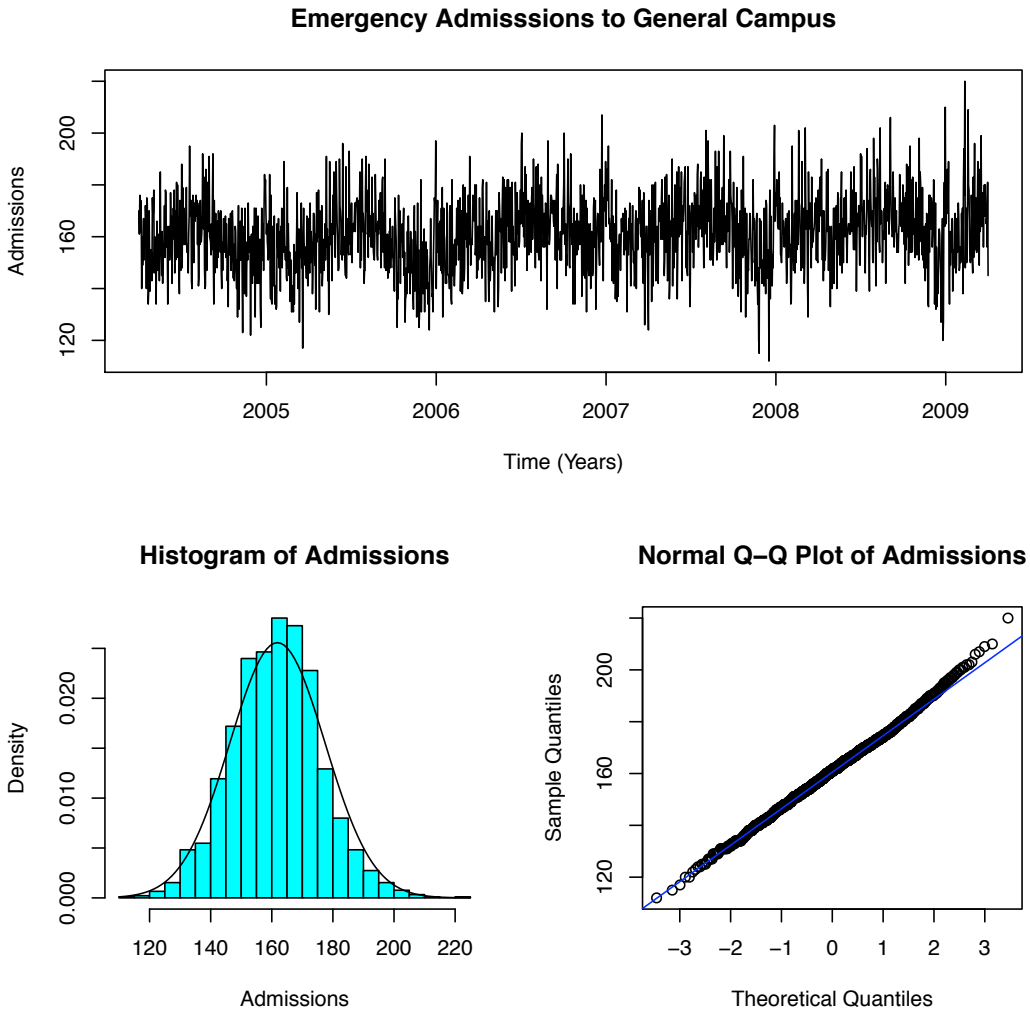

Figure 2.2: Histogram and q-q plot of emergency admissions to general campus. Note that the time-series data appear to be normally distributed.

2.2 Discharge Data

One of the first things we would like to do before modelling discharge is eliminate any covariates that will not have a practically significant effect on predictive performance. Reducing the potential number of covariates in the model will reduce the chance of collinearity. It will also reduce the number of interactions we may wish to consider. This may seem trivial with smaller datasets, but in this case we had data for an entire hospital, and the models we considered were computationally intensive. Therefore, reducing model complexity now will pay dividends later.

As can be seen in Table 1.2, some variables were only relevant for a small number of patient encounters: the accident indicator represented 3% of patient encounters; the ICU indicator represented 5% of patient encounters; the unknown-occupation indicator represented 8% of patient encounters; the retirement indicator represented 13% of patient encounters; and the surgery indicator represented 18% of patient encounters. We could guess that the unknown-occupancy indicator and the retirement indicator would not be important covariates for a discharge model, but the others could be important given their perceived severity. For example, 25% of in-hospital deaths occurred in the ICU, even though the ICU itself only represented 5% of patient encounters.

The problem with evaluating variables in this way is that we knew so little about how they were used in day-to-day operations at the hospital. For example, given the descriptions provided in Table 1.2, there seemed to be significant overlap between the variables admitType and admitPriority—both had entries for elective, emergency, and urgent admission. In speaking with data warehouse support team members, it was suggested that these two variables may even have been used differently in different departments, further complicating matters. However, note that admitType is used in deriving the Laboratory-based Acute Physiology Score (LAPS), which domain knowledge suggested would be an important time-dependent covariate in our models.

2.2.1 Covariate Selection

Given our limited knowledge of the variables extracted or created from the data warehouse, and our need to distil the number of variables to the most practically significant, we considered a covariate selection approach similar to that suggested in [42]. Recognizing the significant criticisms to using p-values in stepwise (forward or backward) selection routines, as described by Harrell in Section 4.3 of [20], we instead used an approached by Shtatland [42] that relies on information criteria. That is, we used a forward selection routine only to create an ordered sequence of covariates, based on p-values, and then used the Akaike Information Criterion (AIC) to find an optimum set of covariates. We treated this as a rough guide of the best covariates, and ignored other information criteria or methods of model selection discussed by Shtaltland in [42]. Again, our purpose was not to select a final set of covariates, just to create an ordered sequence of the most practically significant covariates for further consideration (using AIC to eliminate covariates).

However, we did not use this covariate selection approach on all of the data directly. Instead, following Escobar et al. in [15], we applied the stepwise forward selection routine for a logistic regression of discharge (using the outcome variable exit) on all covariates by primary condition (where the primary conditions group patients by related diseases and mortality rates, as described in the Derived Variables of Section 1.4.3, and in Appendix A). In other words, we had 44 logistic regressions, one for each primary condition, resulting in 44 sequences of covariates, ranked by AIC. We then considered the covariates that appeared repeatedly at the top of each sequence in order to distil our set of covariates to the most important ones, taking domain knowledge and relevant literature described earlier into consideration. The result is Table 2.1.

Luckily nothing in Table 2.1 came as a surprise, although it needed to be confirmed in a defensible way. Also, we needed to be sure that one group of covariates

Table 2.1: Covariates to be used for modelling discharge, and covariates dropped from consideration. We used stepwise forward selection for 44 logistic regression models of discharge, by primary condition, to rank sets of covariates based on AIC, in order to help in choosing the most practically significant covariates.

Covariates Kept (Type)	**Covariates Dropped (Type)**
accommodation (nominal class)	accident (indicator)
admitType (nominal class)	admitPriority (nominal class)
age (continuous)	married (indicator)
campus (nominal class)	occUnknown (indicator)
charlson (ordinal class)	ohip (indicator)
day (continuous)	regularAdm (indicator)
gender (nominal class)	religion (indicator)
icu (indicator)	retired (indicator)
laps (continuous)	transferIn (indicator)
surgery (indicator)	urgentAdm (indicator)
nursStaOcc (continuous)	

could reasonably represent each primary condition. We have already described how admitPriority (dropped) had significant overlap with admitType (kept), seen in Table 1.2, whereas the latter was used in deriving the Laboratory-based Acute Physiology Score. All other covariates dropped were indicators, most of which overlap with covariates kept, or with small sample sizes, or with little predictive plausibility.

We also tested the Variance Inflation Factor (VIF) of the resulting set of covariates, those kept in Table 2.1 for modelling discharge, in order to test for multicollinearity (i.e., strong linear dependencies between covariates). VIF measures how much the variance of an estimate has "inflated" due to collinearity. For covariate k it is $1/(1 - \mathrm{R}_k^2)$, where R_k^2 is the coefficient of determination for covariate X_k regressed on the remaining set of covariates. For an ordinary linear model of discharge with all covariates, including primary conditions, on the full dataset, the VIF did not exceed 10 (a typical threshold used to detect serious problems, as suggested in [24] and [3]).

On the other hand, for 44 ordinary linear models of discharge, by primary condi-

tion, we found the factor levels of a class variable would at times exceed a VIF of 10. However, only one class variable would be problematic for any one primary condition, and which covariate depended entirely on the primary condition. For example, for primary condition AMI, only the factor levels of admitType have VIF that exceed 10, and in this case they have the least statistically significant estimates.

Finally, having selected our covariates, we give an example in Table 2.2 of what the entries in the dataset might look like for an entirely fictional patient encounter.

Table 2.2: Example of entries in the dataset for one fictional patient encounter. Note that this example is in no way derived from the true dataset, and is entirely made up, else it would be a breach of patient confidentiality. Not all covariates are included, and we use full names in factor levels rather than abbreviated forms for clarity.

exit	day	laps	admitType	age	campus	charlson	gender	icu
0	1	140	direct	45	civic	3–4	male	1
0	2	120	direct	45	civic	3–4	male	1
0	3	120	direct	45	civic	3–4	male	1
0	4	100	direct	45	civic	3–4	male	1
0	5	90	direct	45	civic	3–4	male	0
0	6	80	direct	45	civic	3–4	male	0
0	7	60	direct	45	civic	3–4	male	0
1	8	60	direct	45	civic	3–4	male	0

2.2.2 Data Reduction

With 122,883 patient encounters, plus repeated measures (one observation for each day in hospital), resulting in 987,163 observations, we needed to reduce our dataset to a more reasonable size. When we ignored the large number of observations, wishing to make use of all the data (besides a 10% hold-back sample for validation), our algorithms ran out of memory on a desktop computer (both the population-averaged model of Chapter 4, and the subject-specific model of Chapter 5). We had hoped

to install software on a server at the hospital, with a 64-bit processor and larger amounts of memory, but the information technology manager had concerns regarding conflicting software.

We limited the data to the last 28 days of admission (95% of patients are discharged by that time). This was not to reduce the size of the dataset, per se, but to eliminate bias from long-term patients. However, by capturing their last 28 days in hospital, we also captured their discharge, and the state of variables leading to their discharge.

In order to determine an appropriate sample size for building our models, we started with criteria given by Harrell in Section 4.4 of [20] (which does not, however, consider correlated data, as in our case). Basically, to avoid overfitting and ensure reliability, Harrell recommends that at most 10 to 20 effective observations be included for each parameter in the model (depending on how narrowly distributed the covariates are). For an ordinary logistic regression model the number of effective observations is the minimum of the number of 0 or 1 outcome values.

For example, for any one primary condition, we needed to estimate at most 18 parameters for the model with the covariates defined in Table 2.1 (the main effects model). Note that we have more parameters than covariates because for class variables the number of parameters we needed to estimate was equal the number of factor levels minus one (the degrees of freedom). Except that not all factor levels were represented for each primary condition, which is why we say *at most* 18 parameters needed to be estimated in the main effects model. Therefore, we wanted at most 180 to 360 patient encounters for any one primary condition to construct a reliable model, according to [20].

However, we had longitudinal data, and therefore needed to determine a sample size that took repeated measures into account. Given that our ultimate goal was to build a multilevel model of discharge, the subject-specific model that will be introduced in Chapter 5, we considered the guidelines offered by Snijders and Bosker in

[45], Section 3.4. Following the above, we started by considering a model for only one primary condition group (a two-level model).

Suppose that we had an equal number of repeated measures for each patient encounter. The total sample size would be the number of repeated measures, n, times the number of patient encounters, N. The adjustment needed to account for the (two-level) sample design, called the design effect—the sample variance of the design divided by sample variance of a simple random sample—is given by $d_e = 1 + (n-1)\rho$, where ρ is the correlation between two units in a group, called the intraclass correlation (which we will revisit in Section 5.3.3, when we interpret the results of a multilevel model). The effective sample size is therefore $N_{eff} = N \cdot n/d_e$.

It is worth considering the intraclass correlation in a bit more detail (although we ignore some complications until later). When repeated measures for patient encounters are homogeneous, ρ approaches 1, and the design effect approaches n, which dilutes the effective sample size. Essentially, the patient encounter becomes the main sample unit, not the repeated measures in the group, and the effective sample size approaches N. On the other hand, when the repeated measures for patient encounters are heterogeneous, ρ approaches 0, and the design effect approaches 1, meaning the effective sample size approaches the total sample size, $N \cdot n$.

Next we considered a few scenarios, and made some assumptions, to determine appropriate sample sizes for building a model for only one primary condition group. The mean patient length of stay, defined as the total number of days a patient stayed in hospital during an encounter, was about 9. We therefore used this as our estimate for n. Given that our patient groups (i.e., encounters) were the repeated measures, we could assume a high intraclass correlation, and consider $\rho = 0.5, 0.75$, and 1.0. Finally, we considered 180 patient encounters, following our analysis above (in which we ignored the correlated observations), and calculated the effective sample sizes.

We also repeated the above exercise for the 44 primary condition groups, our new N, with group sizes determined by the number of patient encounters, our new n (the

latter could come from our effective sample size for patient encounters, considered above). This was our three-level sample design. In this case we could not assume a high intraclass correlation, and according to Twisk in [47], Section 8.4, a value of 0.20 is considered very high in practice (except for longitudinal data). Hence, we considered $\rho = 0.30, 0.20$, and 0.10. Also, we added 100 patient encounters to our scenarios, given the large total sample size for our three-level design. The results can be seen in Table 2.3, in which we use subscripts to distinguish between sampling levels.

Table 2.3: Sample size calculations for multilevel sample designs. Subscripts refer to design levels—1 for repeated measures, 2 for patient encounters, 3 for primary conditions.

By patient group					**By primary condition**				
N_2	n_1	$\rho_{1,2}$	$d_{e_{1,2}}$	$N_{2_{eff}}$	N_3	n_2	$\rho_{2,3}$	$d_{e_{2,3}}$	$N_{3_{eff}}$
100	9	1.00	9	100	44	100	0.30	30.7	143
		0.75	8	113			0.20	20.8	212
		0.50	5	180			0.10	10.9	404
180		1.00	9	180		180	0.30	54.7	145
		0.75	8	202			0.20	36.8	215
		0.50	5	324			0.10	18.9	419

Given the effective sample sizes $N_{3_{eff}}$ shown in Table 2.3 for the three-level design, and the recommended maximum number of patient encounters of 180 to 360 determined earlier, we decided to use a sample of 100 patient encounters by primary condition. That way we have a sufficient number of patient encounters for building models by primary condition (with effective sample sizes ranging from 100 to 180), and for building a three-level model encompassing all primary conditions (with effective sample sizes ranging from 143 to 404). The resulting dataset was made up of 4,400 patient encounters, and 33,259 total observations.

2.2.3 Histograms

In order to better understand our dataset, we considered histograms of some important patient variables. However, as we have described already, primary conditions represent an important grouping variable, following [15], and we therefore considered our histograms by primary condition (an initial classification of illness, as described in Appendix A). Where normal density curves are included, it is not to suggest that data are necessarily normally distributed, but as a reference against which to compare what one might expect, and deviations thereof.

We see from the histograms in Figures 2.3, 2.4, 2.5, and 2.6 that there were significant differences in mean and variance by primary condition, especially for the Laboratory-based Acute Physiology Score, which we considered to be one of our most important covariates. This confirms that primary condition was an important grouping variable, which we therefore used for modelling discharge.

Although patient age varied significantly at the hospital, as would be expected, differences in the distribution of patient age by primary condition were less pronounced than might be anticipated, as shown in Figure 2.4. Length of stay shown in Figure 2.5, defined as the total number of days a patient stayed in hospital (from admission to discharge), had more variability between primary conditions, but also note the skew to the right (we limit the plots to 30 days).

As can be seen in Figure 2.6, the most common occupancy rate was 0, which indicates that there were no fixed number of beds for the nursing station—in most, if not all, cases this means the patient was in the emergency department waiting for a bed in hospital. It may have been more appropriate to give this situation a very high occupancy rate, but this is the way it was coded by the data warehouse support team, so we left it as provided.

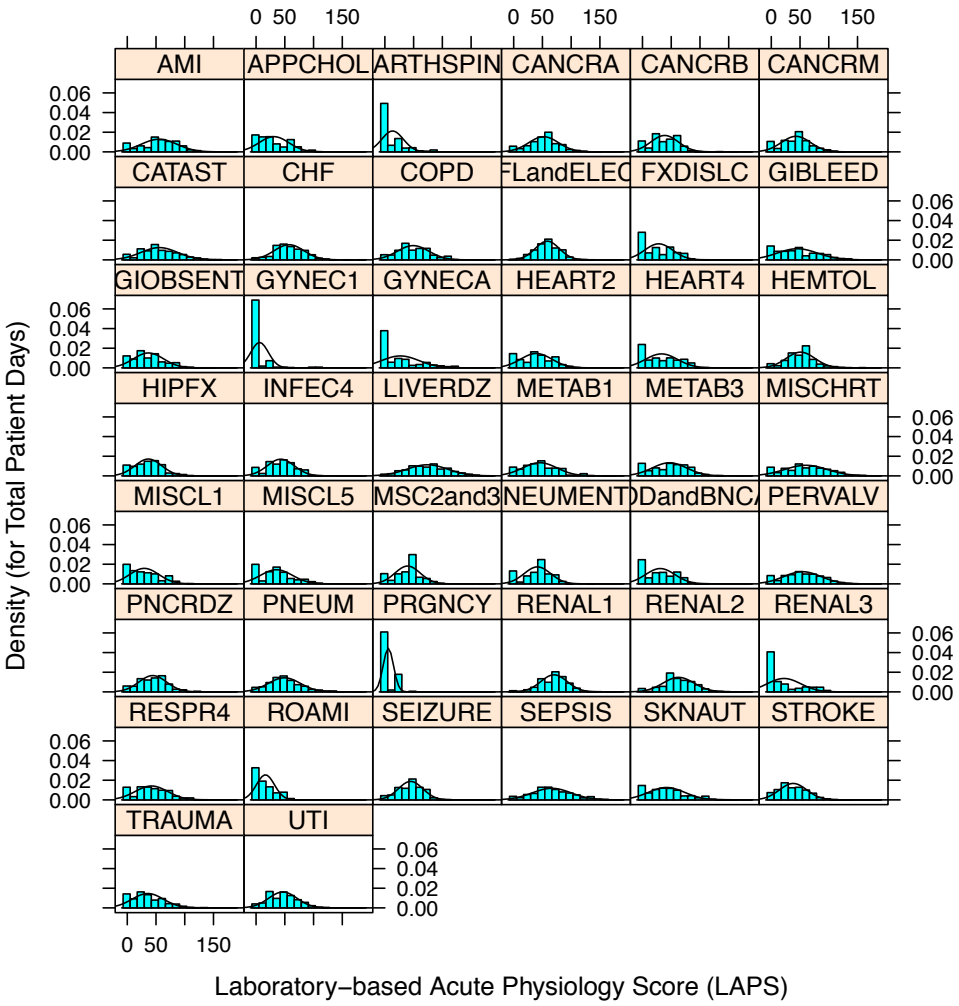

Figure 2.3: Histograms of Laboratory-based Acute Physiology Score, by primary condition. LAPS is adjusted for age and admit type, among other factors. As seen in these results, the distribution of LAPS varies considerably between primary condition groups.

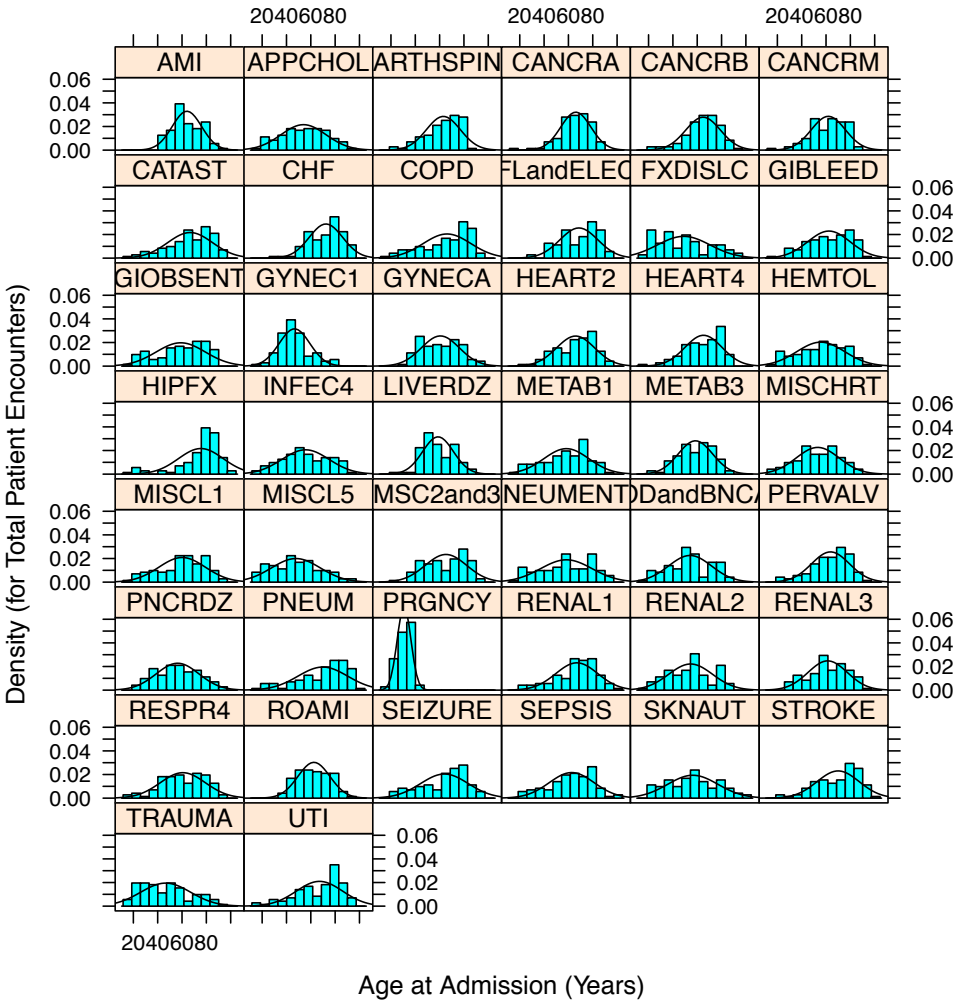

Figure 2.4: Histograms of patient age at admission, by primary condition. There does not seem to be a lot of variation in patient age between primary conditions ("seizure" being a clear exception).

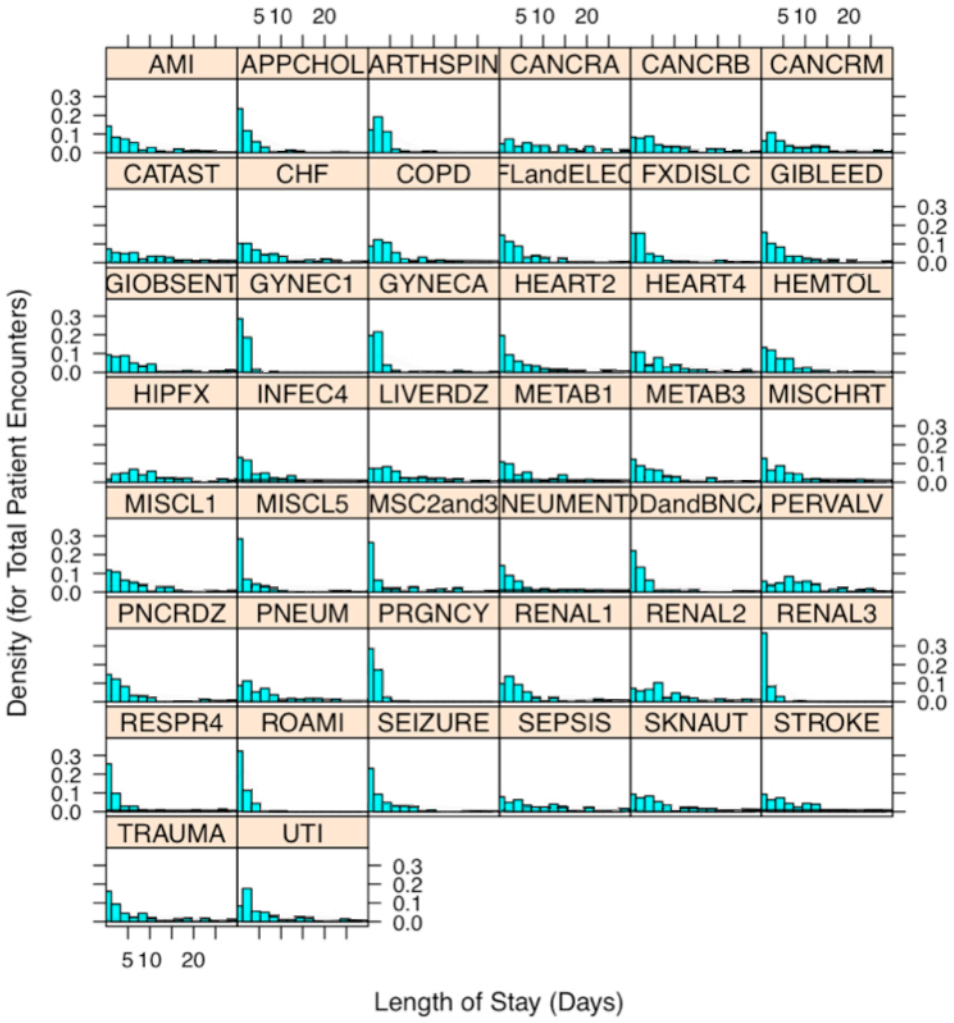

Figure 2.5: Histograms of patient length of stay, by primary condition. Length of stay is the total number of days a patient stayed in hospital. The distributions of length of stay are right skewed.

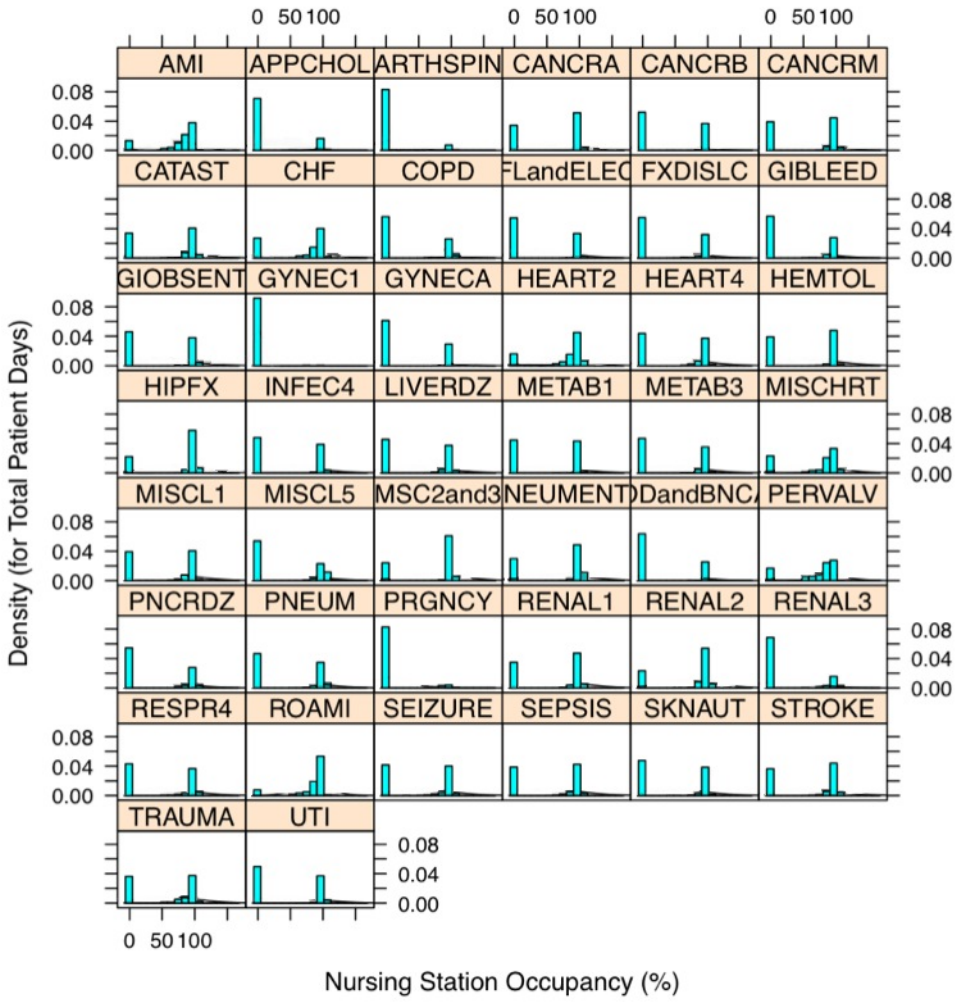

Figure 2.6: Histograms of nursing station occupancy, by primary condition. An occupancy rate of 0 indicates that there are no fixed number of beds (patients are therefore most likely in the emergency department).

2.2.4 Scatter Plots

There are certain effects that are typically accounted for in epidemiological data, such as age and gender. In particular, we considered the interactions described in [15], through the use of scatter plots, to determine if they needed to be considered in our models of discharge. We therefore expect to see some relationship between certain covariates, and the question is to what extent we can or cannot ignore them. Although a table summarizing the correlation coefficient between two variables would provide a measure of linear association, it is possible that a more complicated relationship will be seen from a visual inspection of scatter plots, or the detrimental impact of outliers on the correlation coefficient. Again, we consider these plots by primary condition (an initial classification of illness, as described in Appendix A).

The Charlson index is typically adjusted for by age, but the plots in Figure 2.7 do not seem to indicate the need to include them as an interaction in our model, at least for most groups. That is, the variation in age for each category of the Charlson index is more or less consistent, although there are some exceptions. For example, the group "respr4" (acute respiratory infections) shows less variability with higher Charlson index. Since a higher Charlson index implies higher risk of mortality, we can infer that the higher age groups are most affected by the sever cases of acute respiratory infections (justifying an interaction term between these covariates for this primary condition group).

Similarly, the plots in Figure 2.8 do not suggest the need to include an interaction effect for LAPS and age, which came as a surprise. However, when we looked at the specific derivation of LAPS used for this project, we found that it was already adjusted for by age (i.e., age was used to calculate LAPS). If a trend could be discerned from the scatter plots, then we would be more apt to consider interactions between the two. Although, as in the case of the Charlson index and age, there are exceptions, such as the "seizure" group, which could possibly benefit from a linear interaction.

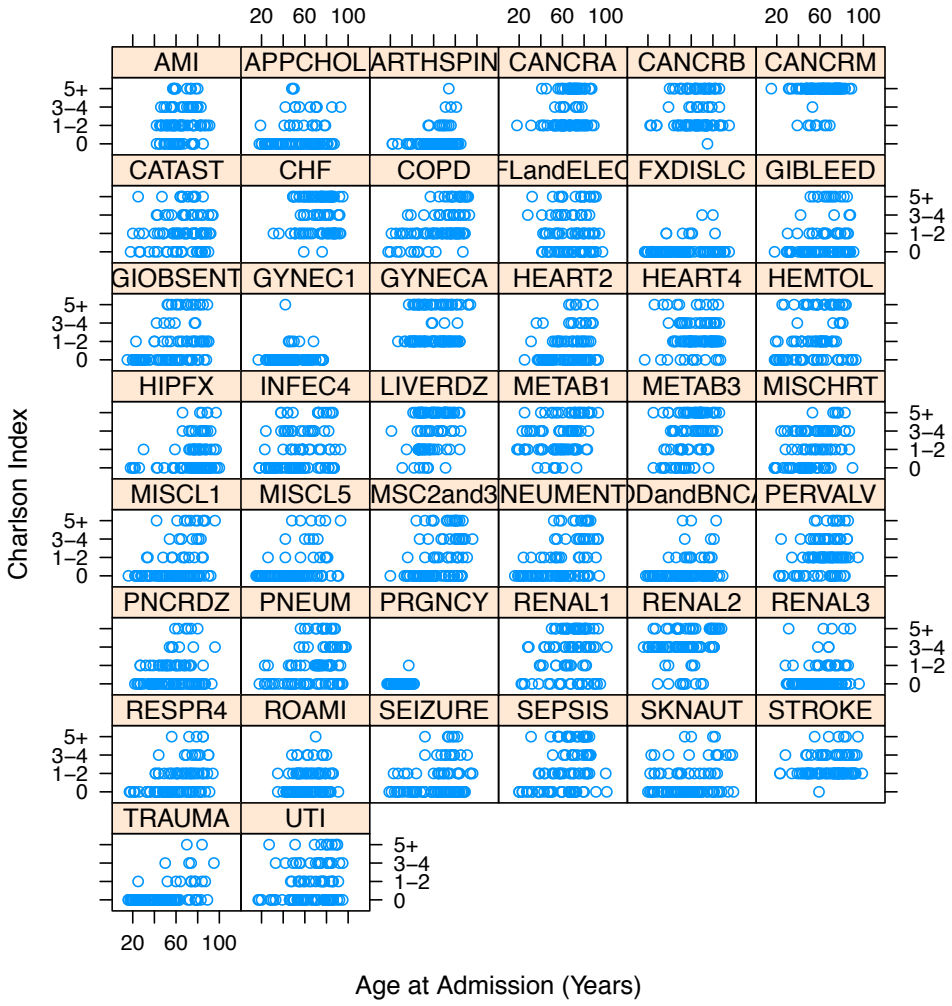

Figure 2.7: Scatter plots of Charlson index versus patient age at admission, by primary condition. Charlson is typically adjusted for by age, but these scatter plots do not justify an interaction term in our models.

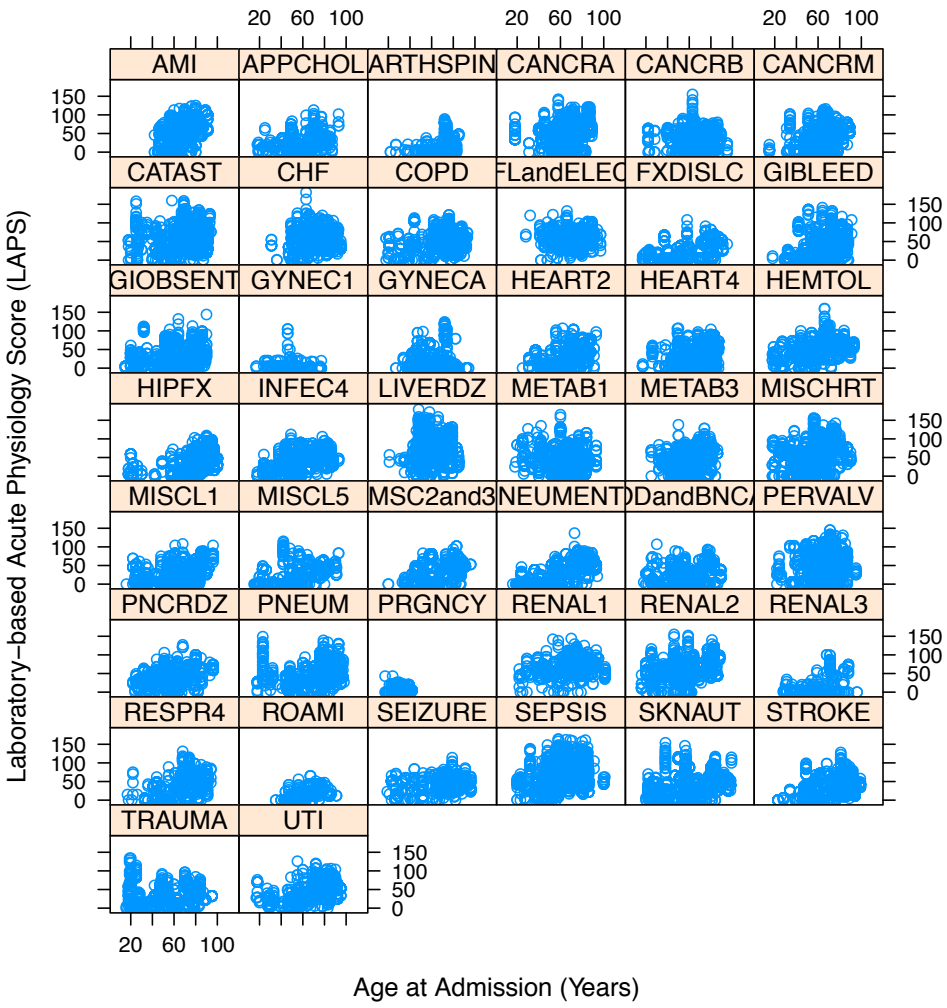

Figure 2.8: Scatter plots of Laboratory-based Acute Physiology Score versus patient age at admission, by primary condition. Age was already used in deriving LAPS.

The plots in Figure 2.9 show no significant correlation between LAPS and the Charlson index. Although both are attempts at measuring severity of illness, the former was determined daily, whereas the latter was determined at admission—see Table 2.3 for an example of how LAPS varies over time (as it is time-dependent), whereas the Charlson index remains fixed over time (it is time-independent).

We also considered plots of patient length of stay versus age in Figure 2.10, which suggests that age had an appreciable affect on length of stay. Knowledge of the health care domain suggested that age would be an important determinant of patient length of stay in hospital. This could suggest the need of an interaction term for age and day in hospital. It is typical in longitudinal studies (with a small number of covariates) to include interactions of fixed effects with time (in our case the covariate day).

The overall results of the scatter plots in this section do not suggest the need to include the interactions described in [15] (which did not include interactions with time). This implies a simpler model, and also a less computationally complex one. As was mentioned earlier, our algorithms ran out of memory on the full dataset, and also when we included several interactions (including time). However, some primary condition groups did suggest relationships between variables. In other words, interactions could improve on a main-effects model in some primary conditions, which could in turn improve the overall predictive accuracy, should the need arise.

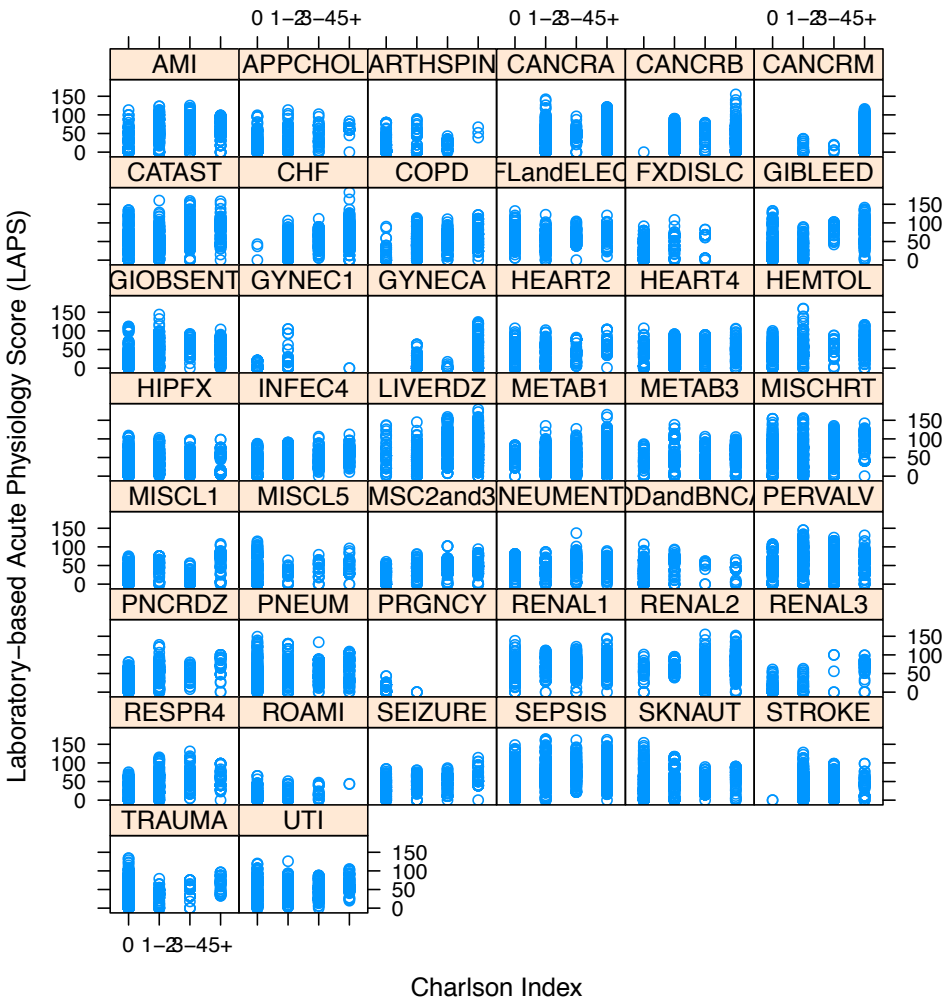

Figure 2.9: Scatter plots of Laboratory-based Acute Physiology Score versus Charlson index, by primary condition. The former is time-variant; the latter is time-invariant.

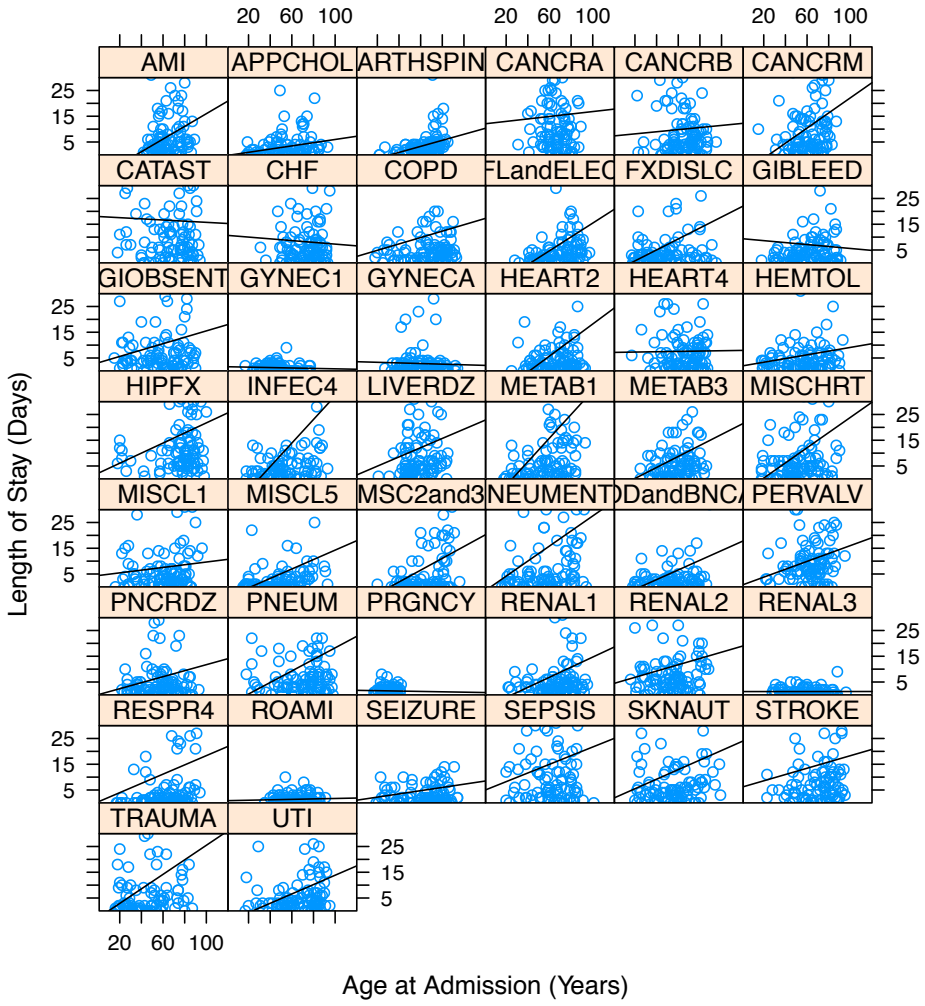

Figure 2.10: Scatter plots of patient length of stay versus age at admission, by primary condition. Length of stay is the total number of days a patient stayed in hospital.

Chapter 3

Time-Series Model for Admissions

The primary source of unscheduled entry to the hospital is the emergency department. As such, in this chapter we describe how we developed time-series models of admissions to the hospital from the emergency department (including patients that remain in the emergency department for treatment, waiting to be admitted to a ward). Combined with scheduled arrivals to the hospital, such models could be used to predict the number of patient admissions to the hospital on any given day. We begin with a brief overview of linear time-series model theory, and extensions, before presenting the results of our modelling efforts in Section 3.2.

3.1 Time-Series Models

Let $\{X_t\}$ be a sequence from $t = 1, \ldots, n$. Fundamental to inference in linear time-series models is the concept of (weak) stationarity, which imposes the conditions of time-invariance of the first and second moments—i.e., that the mean $\mu = \mathrm{E}(X_t)$ and autocovariance $\gamma(h) = \mathrm{Cov}(X_{t+h}, X_t)$ do not depend on time t, for lag h (and note that implicitly $\gamma(0) = \mathrm{Var}(X_t) < \infty$).

Given a stationary time series, the autocorrelation function (ACF) $\rho(h) = \gamma(h)/\gamma(0)$

estimates the linear dependence between elements X_t and X_{t+h} (i.e., the correlation depends only on the difference in time). In contrast, the partial autocorrelation function (PACF) estimates the correlation between the residuals of X_t and X_{t+h} after regressing on the intermediate elements $X_{t+1}, \ldots, X_{t+h-1}$. We will use plots of the ACF and PACF to identify our models, to be described in what follows.

3.1.1 Autoregressive Moving Average

We define an autoregressive (AR) model of order $p \geq 1$ as a linear combination of past X_t elements up to lag p by the equation

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \varepsilon_t.$$

The series $\{\varepsilon_t\}$ is called white noise, with $\mathrm{E}(\varepsilon_t) = 0$, $\mathrm{Var}(\varepsilon_t) = \sigma^2$, and $\mathrm{Cov}(\varepsilon_s, \varepsilon_t) = 0$ for all $s \neq t$. The PACF of an AR(p) process is zero for lags $h > p$.

We define a moving average (MA) model of order $q \geq 1$ as a linear combination of past white-noise elements up to lag q by the equation

$$X_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q}.$$

An MA process is clearly always stationary, given the properties of white noise. The ACF of an MA(q) process is zero for lags $|h| > q$.

The previous two models can be combined, and therefore generalized, into the autoregressive moving average (ARMA) model or order $p, q \geq 0$ defined by the equation

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q}.$$

A convenient way to express this model is using the backshift (or lag) operator, $B^k X_t = X_{t-k}$ (where k is an integer), with polynomials $\alpha(z) = 1 - \alpha_1 z - \cdots - \alpha_p z^p$ (of order p) and $\beta(z) = 1 + \beta_1 z + \cdots + \beta_q z^q$ (of order q), such that

$$\alpha(B) X_t = \beta(B) \varepsilon_t.$$

An ARMA(p, q) process is stationary if the complex roots of $\alpha(z)$ are strictly outside the unit circle (which also holds for a purely AR(p) process).

3.1.2 Autoregressive Integrated Moving Average

An increasing or decreasing time trend, or cyclical features, can often be removed from non-stationary time-series data via differencing, the result of which can then be modelled using a stationary ARMA process. The integration, or combining, of the differenced series (a stationary ARMA process) is the original series itself, and is therefore called an autoregressive *integrated* moving average (ARIMA).

Accordingly, we define the ARIMA model of order $p, d, q \geq 0$ as

$$\alpha(B)(1-B)^d Y_t = \beta(B)\varepsilon_t,$$

where

$$X_t = (1-B)^d Y_t \quad \text{is a stationary ARMA}(p, q) \text{ model}$$

(the series X_t is therefore the dth difference of the original series Y_t). Clearly this generalizes the ARMA model in a convenient and important way, given that many applications involving time-series data will include trends.

3.1.3 Seasonal Autoregressive Integrated Moving Average

Seasonal effects, however, can be accounted for more efficiently by taking differences at lags greater than one, and applying an ARIMA model to the resulting series. If we let s be the seasonal period, then this corresponds to replacing the backshift operator B by B^s in the definition of the ARIMA model. However, we combine seasonal and non-seasonal terms in defining a seasonal ARIMA (SARIMA) model.

In this case, we define the SARIMA model of order $(p, d, q)(P, D, Q)[s]$ as

$$\alpha^{AR}(B)\alpha^{SAR}(B^s)(1-B)^d(1-B^s)^D Y_t = \beta^{MA}(B)\beta^{SMA}(B)\varepsilon_t,$$

where

$$X_t = (1-B)^d(1-B^s)^D Y_t \quad \text{is a stationary ARMA}(p+sP, q+sQ) \text{ model}$$

(parameterized such that many coefficients are zero). Such a model is preferable to increasing the order of an ARIMA(p,d,q) model in an attempt to correct for seasonal effects (as increasing the order of the model could lead to instability in the estimation procedure, and will not correctly remove seasonal effects).

Therefore, to be clear, p, q define the order of the polynomials $\alpha^{AR}(z), \beta^{AR}(z)$ for the non-seasonal AR and MA components; P, Q define the order of the polynomials $\alpha^{SAR}(z), \beta^{SAR}(z)$ for the seasonal AR and MA components. Also, d is the order of the non-seasonal differencing; D is the order of the seasonal differencing (with period s).

3.2 Evaluating the Time-Series Model

In constructing our admissions models for the civic and general campuses of the hospital, we restricted ourselves to the time-series data from 1 April 2004 to 1 January 2008 (our in-sample data). The time-series data after this date, up to 31 March 2009, was reserved for forecasting (i.e., predicting future events, outside the in-sample data).

We used the Box-Jenkins method (originally published in [7], but reproduced in most books on the subject of time-series analysis since) to construct our time-series model of admissions from the emergency department for the civic campus. We started by plotting the (in-sample) data with ACF and PACF plots in Figure 3.1. Although we could see some form seasonality in the series data, this was much more clear in the ACF plot. Interestingly, the most prominent source of seasonality was a weekly effect (something that was not obvious from the time-series data extending over multiple years).

We therefore tried differencing the series with a seasonal period of 7, the results

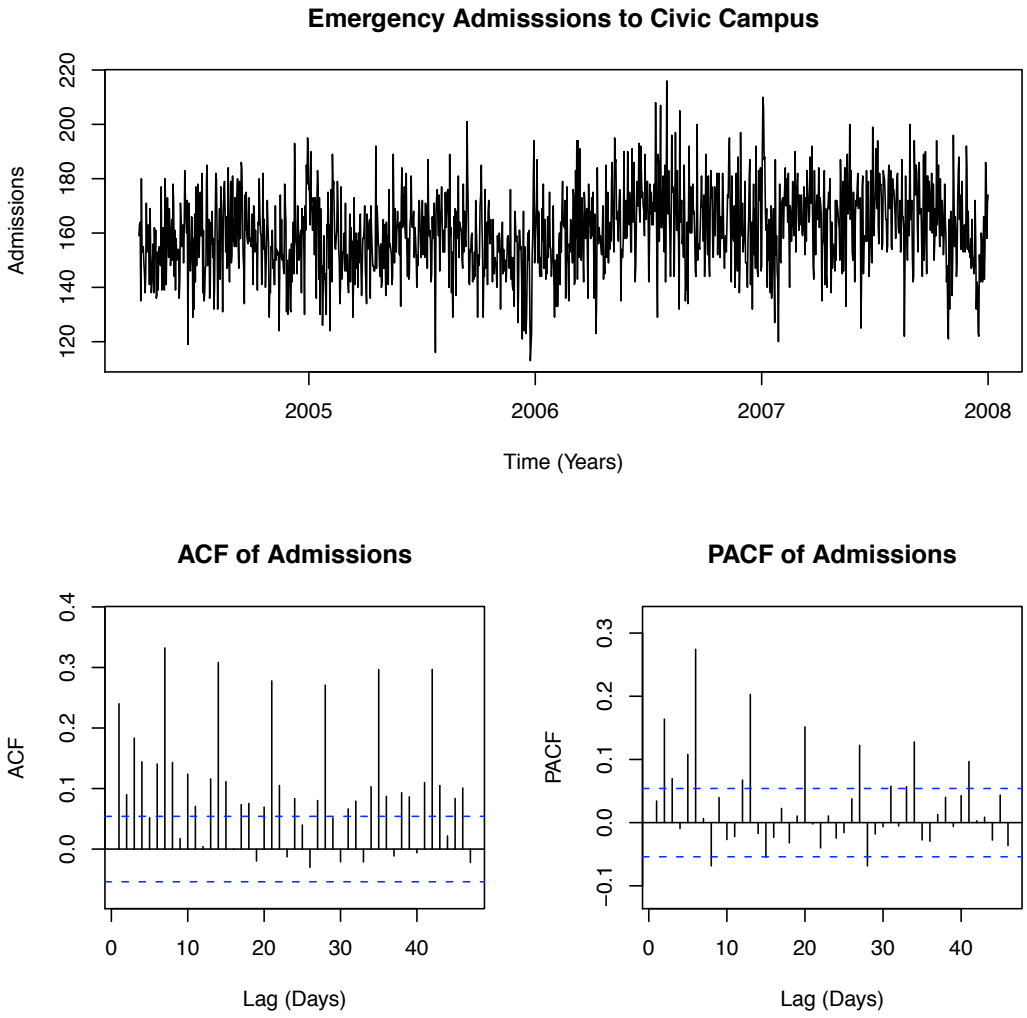

Figure 3.1: Emergency admissions to civic campus with ACF and PACF plots. A strong weekly effect is clearly present in the ACF and PACF plots, which could not be seen in the multiple years of data. The weekly effect is stable over time, suggesting the need for seasonal differencing).

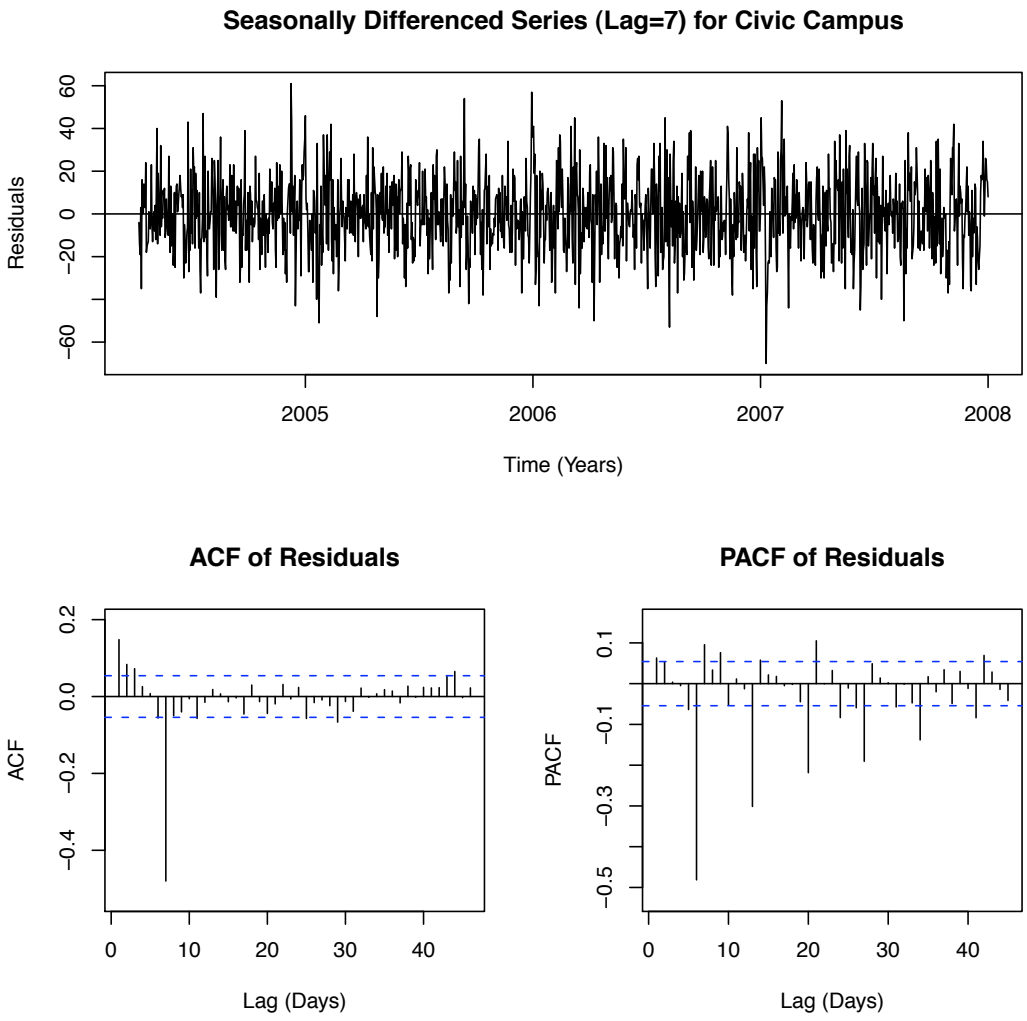

Figure 3.2: Seasonally differenced series for civic campus with ACF and PACF plots. Differencing at a period of 7 removed much of the trend from the data, although the ACF plot suggests over-differencing and the need of an AR term. The PACF plot does not cut-off after lag 7, but instead repeats seasonally, suggesting the need of an SMA term.

of which can be seen in Figure 3.2. From the ACF plot we concluded that the series was over-differenced, and needed an AR term; from the PACF plot, in which we could see a seasonal effect at weekly lags, we concluded the need of an SMA term. Hence, we turned our attention to constructing a SARIMA model, which offered us the most flexibility for modelling these aspects of the series.

In fitting the SARIMA model, we changed only one term at time to evaluate the ACF and PACF plots of the residuals at each step. We also considered, however, a plot of the p-values for the Ljung-Box statistic (a Portmanteau test) to evaluate stationarity, or fit—if we reject the null hypothesis that the autocorrelations up to lag s are 0, then we conclude the series is not stationary (see [29] for details). Rather than walk through the plots at each step of the fitting process, we give a brief summary in the next paragraph. Note that we included an intercept in the model to correct for the increasing trend, or drift, we believed might exist in the data.

We started by fitting a SARIMA model with order (1,0,0)(0,1,0)[7] (i.e., AR(1) plus weekly differencing), which again showed the need of an SMA term in the PACF plot (with negative, weekly PACF coefficients). We removed the weekly differencing to test the SMA term, resulting in a model of order (1,0,0)(0,0,1)[7], but the seasonal structure of the residuals confirmed the need of the seasonal differencing. The resulting model of order (1,0,0)(0,1,1)[7] had positive (decreasing) ACF coefficients, suggesting the need of an MA term. Although the ACF and PACF plots for the model of order (1,0,1)(0,1,1)[7] were reasonable, the Portmanteau test of fit failed, so we tried one more MA term (the result of which satisfied our fitting criteria).

The final SARIMA model of emergency admissions to the civic campus was of order (1,0,2)(0,1,1)[7]. The same process was repeated for the general campus, resulting in a SARIMA model of the same order. An analysis of residuals for each of these models is provided in Figures 3.3 and 3.4. We also include a summary of the coefficient values, with standard errors, in Table 3.1. Terms that were not statistically significant were not removed from the models, as the Portmanteau test of fit would

fail.

Table 3.1: Estimates from seasonal ARIMA$(1,0,2)(0,1,1)[7]$ models for civic and general campuses. An intercept term was included in the models to account for an increasing trend, or drift, which may have been present in the data. Statistically non-significant terms were not removed else the Portmanteau test of fit would fail.

Civic Campus

Parameter	**Estimate**	**S.E.**	**P-Value**
(Intercept)	0.0074	0.0034	0.0263
α_1^{AR}	0.9107	0.0310	<0.0001
β_1^{MA}	-0.7236	0.0415	<0.0001
β_2^{MA}	-0.0408	0.0322	0.2054
β_1^{SMA}	-0.9808	0.0071	<0.0001

(MAPE 6.17; MASE 0.657.)

General Campus

Parameter	**Estimate**	**S.E.**	**P-Value**
(Intercept)	0.0034	0.0040	0.3873
α_1^{AR}	0.9753	0.0112	<0.0001
β_1^{MA}	-0.8224	0.0292	<0.0001
β_2^{MA}	-0.0713	0.0283	0.0118
β_1^{SMA}	-0.9896	0.0083	<0.0001

(MAPE 6.00; MASE 0.647.)

We used the series beyond the in-sample data to evaluate the forecast accuracy of our SARIMA models. First we plot the forecasts (up to 31 March 2009), with 80% and 95% prediction intervals, in Figure 3.5. As can be seen, there were significant spikes in admissions (the dotted line) which cannot be accounted for in the model (unless additional time-series regressors were made available, such as outdoor temperature data). Note that the residuals showed no further seasonal behaviour, and the spikes are themselves not seasonal.

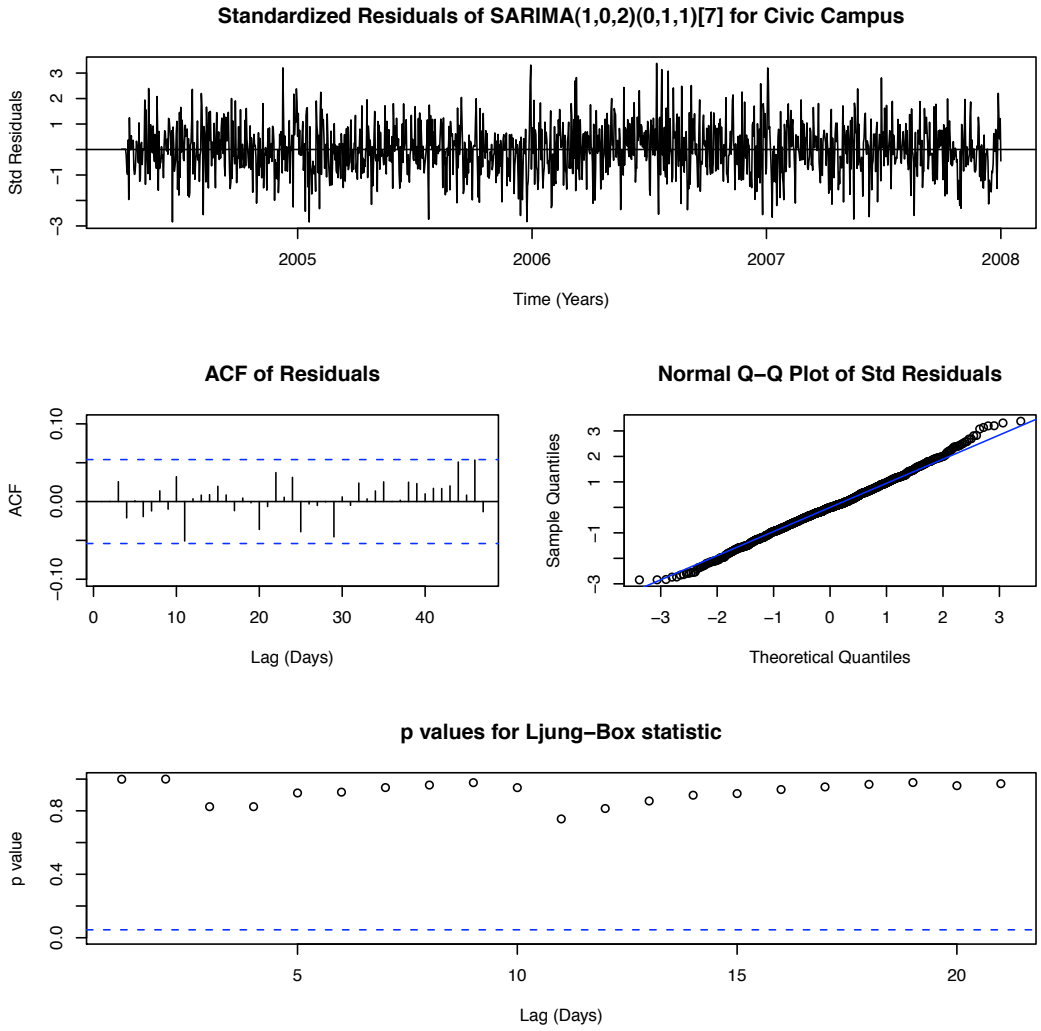

Figure 3.3: Residual analysis of SARIMA(1,0,2)(0,1,1)[7] model for civic campus. The ACF coefficients are not statistically significant, the q-q plot suggests normally distributed residuals, and the Portmanteau test suggests stationarity. Therefore we conclude that the residuals are Gaussian white-noise.

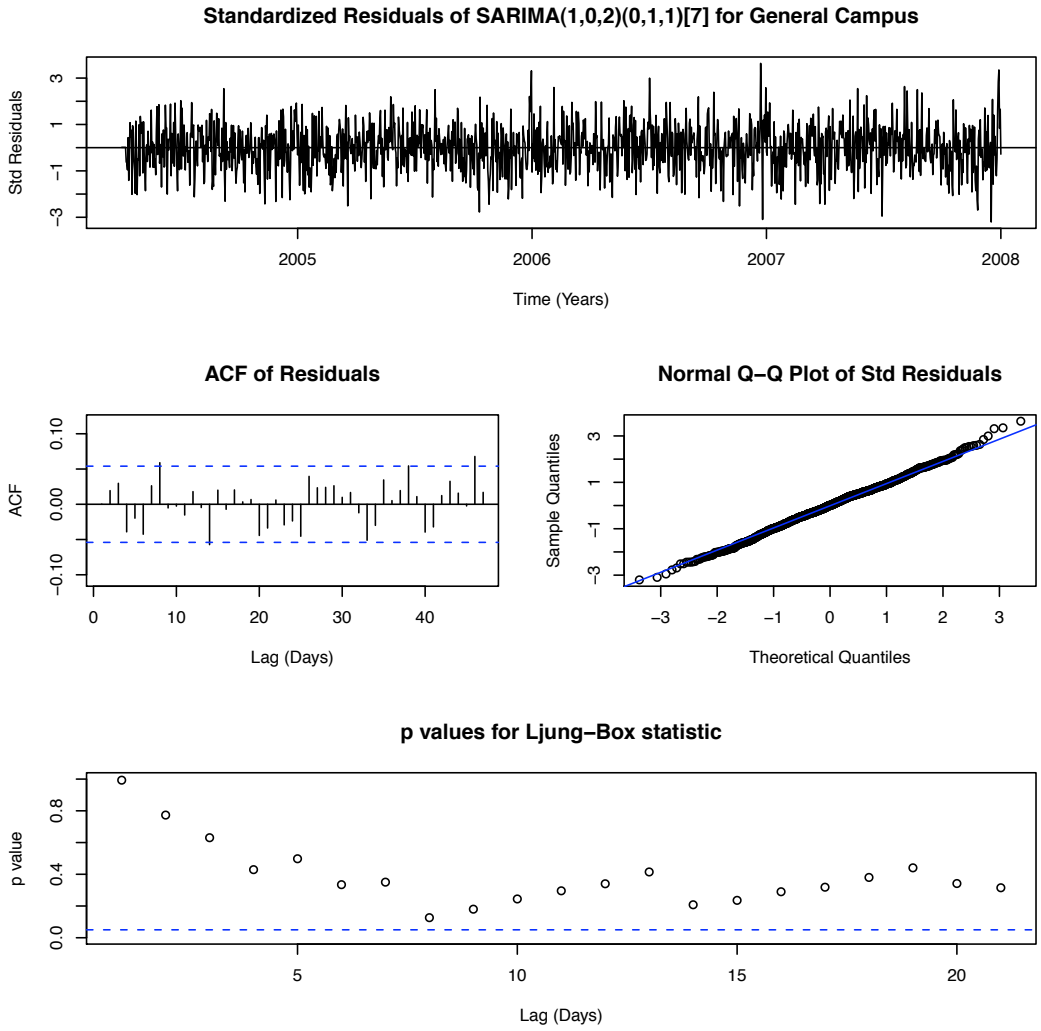

Figure 3.4: Residual analysis of SARIMA(1,0,2)(0,1,1)[7] model for general campus. Most of the ACF coefficients are not statistically significant (besides a minor few points), the q-q plot suggests normally distributed residuals, and the Portmanteau test of fit suggests stationarity. Therefore we conclude that the residuals are Gaussian white-noise.

We summarize forecast accuracy using the Mean Absolute Percent Error (MAPE) and the Mean Absolute Scaled Error (MASE), described in Section 1.3.3. We computed the forecasts from the first day of the out-of-sample data (the day after the last in-sample data on which the models were fit), and by increasing the forecast window by increments of 10 days. The results are plotted in Figure 3.6. The MASE was less than one throughout, which suggests that the models had smaller errors on average than the one-step random walk.

It is likely that an administrator wishing to predict emergency admissions would perform no better than the one-step random walk. Given the results in this section, we suggest that SARIMA models would improve on the accuracy of predicting emergency admissions to the hospital. Combined with the knowledge of scheduled admissions to the hospital, these models could be used to determine a more optimal allocation of beds and resources on a daily basis. The prediction intervals would also allow them to quantify uncertainty, which could be incorporated into hospital procedure regarding spare capacity in the face of shocks (spikes in emergency admissions).

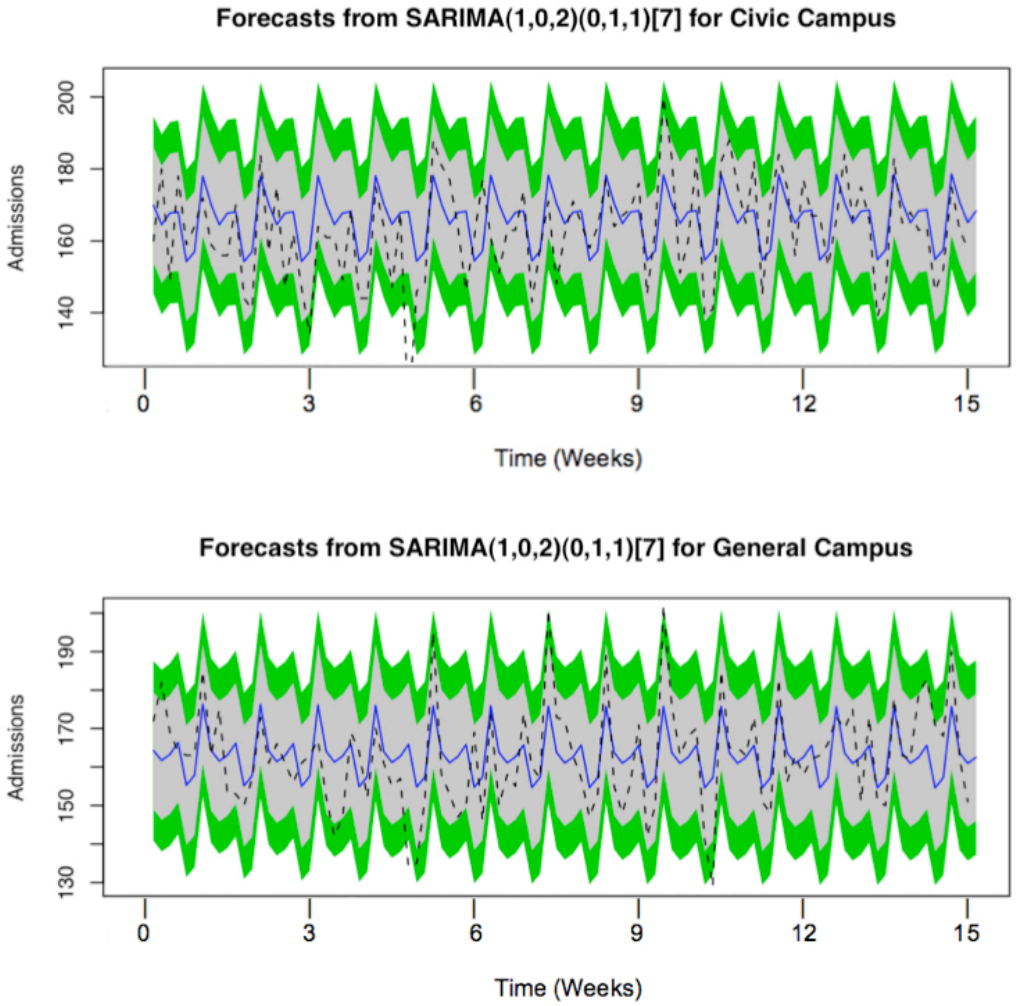

Figure 3.5: Forecasts for seasonal ARIMA models for civic and general campuses. The solid line is the forecasts, the dashed line the true number of admissions, and the inner and outer shaded areas are the 80% and 95% prediction intervals, respectively.

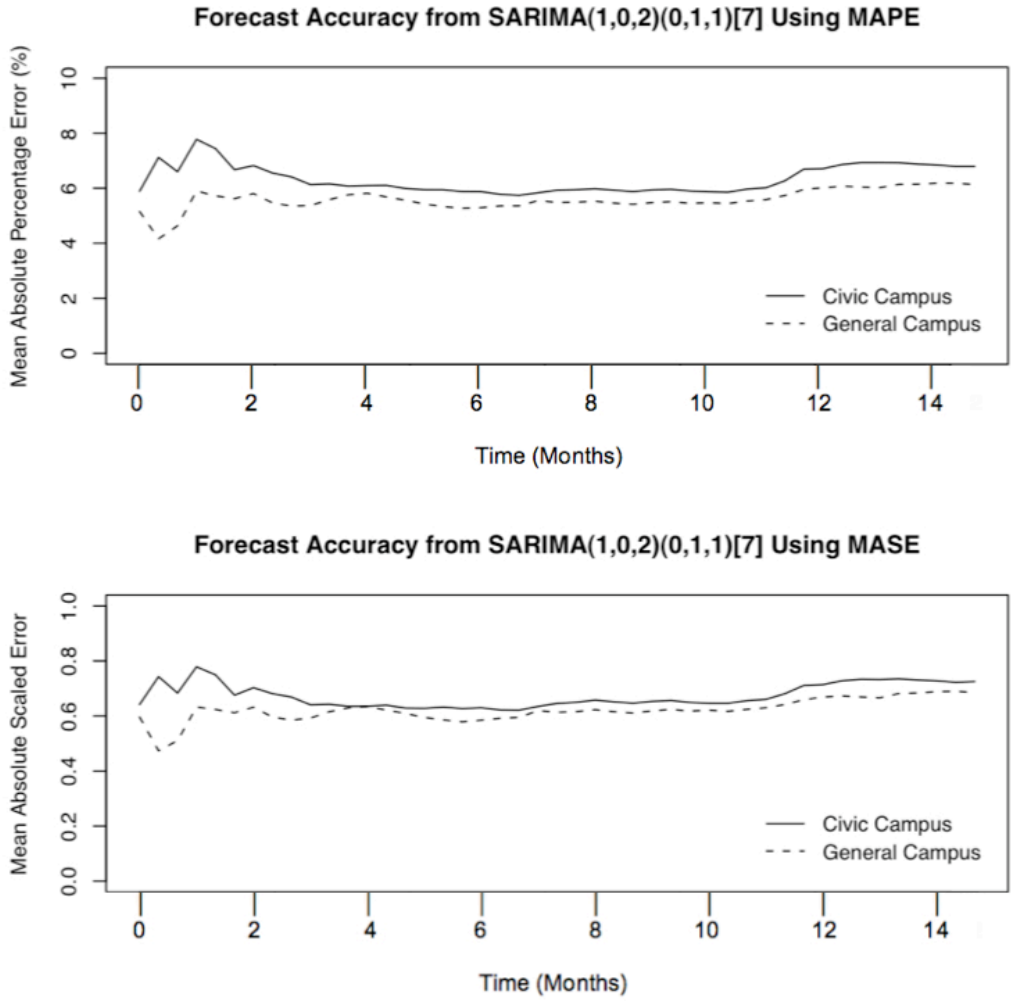

Figure 3.6: Forecasts accuracy for seasonal ARIMA models for civic and general campuses. The MAPE and MASE have the same shape, but the scaling in the latter provides a useful interpretation. Namely, a MASE less than one throughout suggests the models perform better than a one-step random walk.

Chapter 4

Population-Averaged Model for Discharge

Given that we wanted to predict patient hospital discharge, our outcome variable was an indicator of whether a patient had or had not left the hospital. The population-averaged model is essentially averaging over any random effects from subject-specific variation (which we will consider in the next chapter, to improve on predictive accuracy). Given that we had a binary outcome, and longitudinal data, there are certain technical elements that we need to consider. Therefore, before we discuss the fitted model itself, in Section 4.4, we will first look at supporting theory.

4.1 Linear Models

Let Y_i be our outcome variable, and $\boldsymbol{X}_i = [1, X_{i1}, \ldots, X_{ik}]'$ be our vector of k deterministic covariates following a place holder for our intercept term, for observations indexed from $i = 1, \ldots, n$. We will first assume that the outcome variable is continuous while we introduce notation and basic theory. Consider the multiple regression

model (which is linear in the parameters),

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i \\ &= \boldsymbol{X}_i'\boldsymbol{\beta} + \varepsilon_i, \end{aligned}$$

where $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]'$ is the vector of parameters for the covariates and an intercept term, and ε_i the error term (or random disturbance). When convenient we will use matrix notation and therefore introduce the notation $\boldsymbol{Y} = [Y_1, \ldots, Y_n]'$, and the n by $k+1$ design matrix,

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1' \\ \vdots \\ \boldsymbol{X}_n' \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{bmatrix}$$

To derive estimators for the parameters, we minimize the error sum of squares,

$$\mathrm{S}_\varepsilon = \sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i'\boldsymbol{\beta})^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}),$$

by taking the derivative with respect to β and setting it to 0. This leads to the least squares estimators $\widehat{\boldsymbol{\beta}}$ from the equations

$$\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y}$$

which can be solved by taking the inverse of $\boldsymbol{X}'\boldsymbol{X}$ if it is non-singular, or the generalized inverse if it is singular (see Section A.12 of [43] for details).

4.1.1 Gauss-Markov Conditions

We need to ensure the least squares estimators are good in some way. We therefore assume the following Gauss-Markov conditions (as given by Sen and Srivastava in Section 2.5 of [43]):

(i) $\mathrm{E}(\varepsilon_i) = 0$, for all i.

(ii) $\mathrm{E}(\varepsilon_i^2) = \sigma^2$, for all i (constant variance, called homoscedasticity).

(iii) $\mathrm{E}(\varepsilon_i \varepsilon_j) = 0$, for all $i \neq j$ (uncorrelated random errors).

The first Gauss-Markov condition allows us to write our model in terms of the expected mean for the outcome (which will be convenient for comparing to models we will see later), i.e.,

$$\begin{aligned} \mathrm{E}(Y_i) = \quad & \mu_i \\ & \mu_i \quad = \boldsymbol{X}_i' \boldsymbol{\beta}. \end{aligned} \tag{4.1.1}$$

The second and third Gauss-Markov conditions imply that $\mathrm{Var}(Y_i) = \sigma^2$, or $\mathrm{Var}(\boldsymbol{Y}) = \sigma^2 \mathrm{I}$ (where I is the n by n identity matrix, and $\boldsymbol{X}$ is deterministic).

More importantly, however, is that by assuming the Gauss-Markov conditions, it can be proven that the least squares estimators $\widehat{\boldsymbol{\beta}}$ are best linear unbiased estimators (BLUE) for the parameters (called the Gauss-Markov theorem, given in Section 2.9 of [43]). By this we mean that the least squares estimators are unbiased ($\mathrm{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$) with minimum variance among all unbiased linear estimators (i.e., if $\tilde{\boldsymbol{\beta}}$ is another linear estimator of the parameters, then $\mathrm{Var}(\widehat{\boldsymbol{\beta}}) \leq \mathrm{Var}(\tilde{\boldsymbol{\beta}})$—i.e., $\mathrm{Var}(\tilde{\boldsymbol{\beta}}) - \mathrm{Var}(\widehat{\boldsymbol{\beta}})$ is positive semi-definite).

4.1.2 Normal Errors

Assume that the ε_i's are independent and identically distributed (iid) as $\mathrm{N}(0, \sigma^2)$, or equivalently $Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$ independently so that we may perform statistical inference. Consider, however, the probability density function of $\boldsymbol{Y}$, as a function of the parameters $\boldsymbol{\beta}$ and σ^2,

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2\sigma^2} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \right)$$

We call $L(\boldsymbol{\beta}, \sigma^2)$ the likelihood function and maximize it to find maximum likelihood estimators (by taking derivatives with respect to the parameters and setting equal to

0). Note, however, that in this case maximizing $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$ is equivalent to maximizing the likelihood function, which is the same problem as solving the least squares estimators.

Therefore, by assuming normality, we find that the maximum likelihood estimators are equal to the least squares estimators. Given that, under certain regularity conditions, maximum likelihood estimators are consistent (they converge in probability to the true value of the parameter, so that the estimators are unbiased in large samples) and asymptotically efficient (they converge in distribution to a normal distribution with minimal variance, the Cramér-Rao lower bound), they are the least squares estimators.

4.1.3 Binomial Errors

Now let us assume that our outcome variable Y_i is an indicator representing discharge, such that it is equal to 0 when the patient is in hospital, and 1 when the patient has been discharged from the hospital. This implies that the errors ε_i can only take on two values so that we are not dealing with normal errors. Consider,

$$\begin{aligned} \mathrm{E}(Y_i) &= 1 \cdot \Pr(Y_i = 1) + 0 \cdot \Pr(Y_i = 0) \;=\; p_i \qquad \text{by definition} \\ &= \boldsymbol{X}_i\boldsymbol{\beta} \qquad \text{by the first Gauss-Markov condition.} \end{aligned}$$

Therefore we have the linear probability model $p_i = \boldsymbol{X}_i'\boldsymbol{\beta}$. However, for a given set of covariates $\boldsymbol{X}_i$, we have that

$$\mathrm{Var}(Y_i) = p_i(1 - p_i) = (\boldsymbol{X}_i'\boldsymbol{\beta})(1 - \boldsymbol{X}_i'\boldsymbol{\beta}),$$

which implies that we do not have constant variance (called heteroscedasticity). This means that we lose the efficiency of the least squares estimators—our estimators no longer have minimum variance, and so other estimators exist with smaller standard errors. Also, standard errors may be too big or too small (since the variance estimators of the coefficients are no longer consistent), invalidating test statistics.

A more fundamental problem, however, is that the probability model is not realistic. Since we have not imposed constraints on our covariates, such as upper or lower bounds, $\boldsymbol{X}_i'\boldsymbol{\beta}$ may give values that are less than 0 or greater than 1 (which are not realistic probability values). What we need is to transform the probability into an unbounded set of real numbers, but this will require us to consider a non-linear regression model. However, this model fits into a larger class of models that will allow us to build on the linear regression theory presented thus far.

4.2 Generalized Linear Models

In the previous section we were able to model the mean of our outcome as a linear function in a set of covariates. Here we consider a similar idea, except we model a *function* of the mean that is linear in a set of covariates, i.e.,

$$\begin{aligned} \mathrm{E}(Y_i) = \quad & \mu_i \\ g(\mu_i) \quad &= \quad \boldsymbol{X}_i'\boldsymbol{\beta}. \end{aligned} \tag{4.2.1}$$

The function g *links* the expected value of the response $\mathrm{E}(Y_i)$ (which encapsulates the random component) to a linear combination of the covariates $X_{i1}, \ldots, X_{ip}$ (the systematic component).

The regression equation is therefore based on the inverse of the link function (so that the regression equation may not be linear any longer),

$$\mathrm{E}(Y_i) = \mu_i = g^{-1}(\boldsymbol{X}_i'\boldsymbol{\beta}).$$

This implies that our link function must be invertible. However, this latter form of the model is much less appealing, as it does not relate as directly to the linear models discussed earlier (unless the link function is the identity, in which they are the same, of course). As we will see shortly, the intrinsic linearity of the model for $g(\mu_i)$ will be used to develop an algorithm for finding estimators (wherein we will also need to

assume the link function is differentiable). It is important to note that we are not transforming the outcome variable Y_i, but its expected value μ_i. We will again assume the Y_i are independently distributed, but this time we consider them to be from the same exponential family (and under these assumptions, we call (4.2.1) a generalized linear model (GLM)).

4.2.1 Exponential Family

We consider the exponential family because it covers a large class of probability functions that arise naturally in the study of probability (for example, the binomial, Poisson, exponential, and gamma distributions). It has nice properties with known mean and variance, and we will use it to develop a unified approach to finding estimators of the parameters.

Following McCullagh and Nelder in [31], Chapter 2, we assume the probability density function for the observations is of the form

$$f(Y_i) = \exp\left(\frac{Y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(Y_i, \phi)\right),$$

which defines the exponential family with parameters θ_i, ϕ and known functions a_i,b, and c. The parameters θ and ϕ act like location and scale (or dispersion) parameters, respectively. Often the function $a_i(\phi)$ can be simplified as

$$a_i(\phi) = \phi/p_i,$$

where p_i is a known prior weight.

The mean and variance of Y_i in this exponential family are given by

$$\begin{aligned} \mathrm{E}(Y_i) &= b'(\theta_i) \\ \mathrm{Var}(Y_i) &= b''(\theta_i)a_i(\phi), \end{aligned}$$

where $b'(\theta_i)$ and $b''(\theta_i)$ are the first and second derivatives of the function $b(\theta_i)$ (this can be shown by taking the first and second derivatives of the log-likelihood $l(\theta)$, then

using the identities $\mathrm{E}\left(\frac{\partial l(\theta)}{\partial \theta}\right) = 0$ and $\mathrm{E}\left(\frac{\partial^2 l(\theta)}{\partial \theta^2}\right) + \mathrm{E}\left(\frac{\partial l(\theta)}{\partial \theta}\right)^2 = 0$). Therefore, unlike the linear regression model, we are not assuming constant variance, but the variance is a function of the covariates through the mean outcome.

4.2.2 Maximum Likelihood Estimation

The practical reason for using GLMs is that they can be fit to data using a common procedure of iterative weighted least squares, which leads to maximum likelihood estimates, paralleling our linear model in (4.1.1).

Consider a first-order approximation about μ_i of the link function applied to the data,

$$g(y_i) \simeq \underbrace{g(\mu_i) + (y_i - \mu_i)g'(\mu_i)}_{z_i}.$$

The right-hand side of this equation is a new dependent variable, z_i, with a linear representation. We can regress this dependent variable on the covariates, given that it is approximately equal to the link function applied to the data, and consider minimizing the sum of squared residuals. Given data y_i and a known link function g, we need to estimate the mean response μ_i in the above approximation. Of course, our interest lies in approximating the parameters $\boldsymbol{\beta}$, which we will now describe how to do.

Given an initial estimate of the parameters, $\widehat{\boldsymbol{\beta}}$, we calculate the estimated linear covariate $\widehat{\eta}_i = \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}$ and use it to fit the mean response $\widehat{\mu}_i = g^{-1}(\widehat{\eta}_i)$. Using these values, we calculate the adjusted dependent variable

$$\begin{aligned} z_i &= \widehat{\eta}_i + (y_i - \widehat{\mu}_i)\left.\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i}\right|_{\mu_i=\widehat{\mu}_i} \\ &= \widehat{\eta}_i + (y_i - \widehat{\mu}_i)g'(\widehat{\mu}_i) \end{aligned}$$

Next we calculate the weights, which are proportional to the variance of the adjusted dependent random variables Z_i, assuming η and μ are fixed. Adding weights

to the regression model ensures constant variance (when the variance of the residuals differ only by a known constant, the inverse of the weights). For example,

$$\begin{aligned}\mathrm{Var}(Z_i) &= \mathrm{Var}(Y_i)\left(\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i}\right)^2 \\ &= b''(\theta)a(\phi)\left(g'(\mu)\right)^2 \\ &= \mathrm{V}(\mu)\frac{\phi}{p_i}\left(g'(\mu)\right)^2,\end{aligned}$$

where we assume the function of dispersion parameter ϕ is $a(\phi) = \phi/p_i$, as previously described, and $\mathrm{V}(\mu) = b''(\theta)$ is called the variance function. Therefore, in order to achieve constant variance ϕ, we add the weights

$$\widehat{W_i} = \frac{p_i}{\mathrm{V}(\widehat{\mu})\left(g'(\widehat{\mu})\right)^2}.$$

(Note that in practice, however, the variance of the adjusted dependent variable is estimated based on the link function and the distribution of the outcome variable.)

We can now use weighted least squares to improve our estimate of the parameters, by regressing the adjusted dependent variable on the covariates with known weights. Therefore, we have

$$\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{Z},$$

where $\boldsymbol{Z} = [Z_1, \ldots, Z_p]'$ is the response vector (the adjusted dependent variables) and $\boldsymbol{W} = \mathrm{Diag}(W_1, \ldots, W_p)$ is the diagonal matrix of accompanying weights.

With a new approximation for the parameters $\boldsymbol{\beta}$, we can repeat the calculations for the adjusted dependent variables $\boldsymbol{Z}$ and weights $\boldsymbol{W}$, and again get an improved estimate of the parameters. This process is repeated until the difference between subsequent parameter estimates are within some small value. This algorithm, known as iterative reweighted least squares, is shown in Section 2.5 of [31] to lead to maximum likelihood estimates, which therefore means we again have estimators that are consistent and asymptotically efficient. Also note that natural starting values for the algorithm are to take $\widehat{\mu}_i = Y_i$ and $\eta_i = g(\widehat{\mu}_i)$, although it may be necessary to adjust

the starting values to ensure we do not take the log of 0, for example. Another option would be to use linear regression on the adjusted dependent variable.

4.2.3 Binary Outcome

As in Section 4.1.3, assume our outcome variable Y_i is an indicator of discharge. However, in this case, we transform the expected outcome $\mathrm{E}(Y_i) = \mu_i$ using the logit function (a canonical link function), $\mathrm{logit}(\mu_i) = \log(\mu_i/(1-\mu_i))$. Since μ_i is the probability of discharge, $\mathrm{logit}(\mu_i)$ is the log-odds of discharge, which we can use to model a linear function of a set of covariates. Following the methods of this section, we can therefore find maximum-likelihood estimates for a regression model with binary outcome, typically referred to as a logistic regression model.

4.3 Marginal Models

Our dataset for modelling patient discharge, described in Chapters 1 and 2, is made up of longitudinal collections of daily patient information. The generalized linear models described in the previous section, however, assume independent observations (i.e., the uncorrelated random errors in the third Gauss-Markov condition of Section 4.1.1). Estimates that ignore correlated errors can, in some cases, consistently estimate the parameters for a population-averaged model, if the actual correlation is weak. Standard errors will, however, still need adjustment.

In this section we recognize the clustering between patients and their repeated measures, but we model the outcomes *marginalized* over all other outcomes. We would like to know the marginal distributions of the repeated measures, in order to estimate the probability of discharge over time, up to time of discharge T (the number of repeated measures). These probabilities are referred to as the T first-order marginal distributions (as described by Agresti in Chapter 11 of [2]).

Several likelihood-based methods exist for fitting marginal models, but they are awkward and computationally intensive. With maximum likelihood, we would need to consider 2^T joint probabilities for each combination of the covariates, in order to arrive at the T marginal probabilities (the probability of discharge over time). Such an approach is useful when there is interest in the joint probability of events (see the discussion by Molenberghs and Verbeke in [33], Chapter 7, for details). It is not practical, however, if even possible, when there are many repeated measures and covariates, as is our case.

Given that we are only interested in predicting discharge, and have no need for the joint probability, we drop likelihood-based approaches from further consideration. Instead we will consider what is arguably the most popular estimating approach for longitudinal models.

4.3.1 Generalized Estimating Equations

Estimating equations were developed by Wedderburn in [50] to find regression parameters when the distribution is not specified. Rather, the method requires that the variance of the outcome be defined as a function of its expectation. As in Section 4.2, a link function is applied to the marginal expectation of the outcome, the result of which we assume is a linear function of the covariates (the generalized linear model in equation 4.2.1). An iterative method then needs to be used to arrive at the estimates, such as iterative reweighted least squares in Section 4.2.2, by replacing the variance $\text{Var}(Y_i) = v(\mu_i)$.

If the variance in the above approach is assumed to be from the exponential family, with the appropriate link function, as in Section 4.2.1, then it will lead to maximum-likelihood estimates (since nothing will have changed from the last section). Changing the variance of the outcome cannot, however, lead to maximum likelihood estimates, since none exist given that no distribution is specified. Instead, we have a

generalized form of likelihood, called quasi-likelihood, which results in quasi-likelihood estimates (with asymptotic covariance matrix of the same form, namely $(\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}$). The approach is useful, for example, when observations exceed the variability defined for a binomial distribution, known as overdispersion (see Section 4.7 of [2] for further details).

Liang and Zeger developed a form of multivariate quasi-likelihood in [26] to deal with longitudinal data, which they called generalized estimating equations (the general details of which are well explained by the same authors in [53]). The term estimating equations refers to the quasi-likelihood equations, and generalized for their extension to the case of repeated measures. The approach is the same as with quasi-likelihood, except that vectors and matrices are used in order to encapsulate the repeated measures within each subject. To account for the correlation of the repeated measures, a "working" correlation matrix is defined for each subject (of size equal to the number of repeated measures).

Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})'$ be the vector of n_i longitudinal outcomes for subject i. Let $\boldsymbol{R}_i(\boldsymbol{\alpha})$ be the $n_i \times n_i$ working correlation matrix of $\boldsymbol{Y}_i$. We assume the working correlation matrix is fully specified by the vector $\boldsymbol{\alpha}$ of unknown parameters. The number of longitudinal observations per subject, n_i, can vary by subject, but $\boldsymbol{\alpha}$ is fixed for all subjects. The working correlation matrix is given by

$$\boldsymbol{V}_i = \phi \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{A}_i^{1/2},$$

where $\boldsymbol{A}_i$ is an $n_i \times n_i$ diagonal matrix with entries $g(\mu_{ij})$ in the jth diagonal element, for $\mathrm{E}(Y_{ij}) = \mu_{ij}$. The working correlation matrix need not be the true correlation matrix of the $\boldsymbol{Y}_i$—it is a starting point only.

The generalized estimating equations (GEE), for n subjects, are given by

$$\sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{S}_i = 0,$$

where $\boldsymbol{S}_i = \boldsymbol{y}_i - \boldsymbol{\mu}_i$, and $\boldsymbol{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$, for $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})'$. If $n_i = 1$ for

all i, then these equations reduce to the quasi-likelihood equations. Further, if a distribution of Y_i is assumed, then the quasi-likelihood equations are the likelihood equations, through the relationship

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i x_{ij}}$$

An estimator of $\boldsymbol{\beta}$ is obtained as a solution to the GEE. Liang and Zegger showed in [26] that consistent estimates of $\boldsymbol{\beta}$ can be found using iterative reweighted least squares, as in Section 4.2.2. Also, in each iteration, standardized (or Pearson) residuals can be calculated using

$$r_{ij} = \frac{y_{ij} - \widehat{\mu}_{ij}}{\sqrt{[\widehat{\boldsymbol{V}}_i]_{jj}}},$$

which are then used to update the parameters $\boldsymbol{\alpha}$ and ϕ that characterize the working correlation matrix, as they are consistent estimates.

A sandwich estimator is used to find the $\boldsymbol{V}_i$ in the above, in which empirical evidence is squeezed between model-driven covariance. That is,

$$\left[\sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right]^{-1} \left[\sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \mathrm{Var}(\boldsymbol{Y}_i) \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right]^{-1} \left[\sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right]^{-1},$$

where the estimated $\boldsymbol{\mu}_i$ are taken from the fitted model, and $\mathrm{Var}(\boldsymbol{Y}_i)$ by the empirical covariance matrix of $\boldsymbol{y}_i$. Therefore, the dependence in the data (found in the residuals) is used to adjust for the correlation between repeated measures (details on the computations are provided in [26]).

It is important to note that estimates arrived at using generalized estimating equations can therefore be consistent even if the working correlation is not correctly specified. Even the repeated measures do not need to have the same correlation structure between subjects. For example, if the correlation between repeated measures is weak, defining the working correlation to be pairwise independent will nonetheless adjust the standard errors (as the estimates can achieve good efficiency).

4.4 Evaluating the Population-Averaged Model

Having considered all relevant theory, we can now formulate a population-averaged model of patient discharge. However, we will improve on this model in the next chapter, and therefore only consider the main effects—no interactions with time-varying covariates, as is typically done in longitudinal studies—although we will consider an alternative formulation, based on the work of Escobar et al. in [15], looping over primary conditions (an initial classification of illness, as described in detail in Appendix A).

4.4.1 Model Formulation

Our outcome variable is an indicator of discharge, called exit. We consider all main effects described in Table 2.1 (our main effects model), but we do not explicitly outline a parameter for the factor levels in our class variables. For example, admitType has five factor levels, and therefore has four parameters to estimate (the number of factor levels minus one reference level). Instead, we only show one parameter, for simplicity. We include a subscript for time t for time-varying covariates. Therefore, our model for patient encounter i, with repeated measures for time t, is given by

$$\begin{aligned}\text{logit}(\text{E}(\text{exit}_{ti}) = \beta_0 &+ \beta_1 \text{primaryCondition}_i + \beta_2 \text{day}_t + \beta_3 \text{laps}_{ti} + \beta_4 \text{age}_i \\ &+ \beta_5 \text{charlson}_i + \beta_6 \text{icu}_{ti} + \beta_7 \text{surgery}_{ti} + \beta_8 \text{nursStaOcc}_{ti} + \beta_9 \text{gender}_i \\ &+ \beta_{10} \text{campus}_i + \beta_{11} \text{accommodation}_i + \beta_{12} \text{admitType}_i. \end{aligned} \tag{4.4.1}$$

To improve on our predictive model of patient discharge, we also considered fitting one model for each primary condition, resulting in 44 models of discharge (as was done in [15]). In this case we can use the above model, except that we drop the covariate for primary condition. Also, factor levels will change under this model formulation, because some factor levels are missing for specific primary conditions. An obvious example is gender for the primary condition of pregnancy (in this case the

entire covariate must be dropped, since there is only one factor level, i.e., the gender female). This is another reason we did not explicitly give the parameters for factor levels in the above model.

For our working correlation matrix, as described in Section 4.3.1, we assume the repeated measures are highly correlated and dependent on time. Hence, we chose an auto-regressive correlation structure of order 1, $\text{corr}(Y_{ij}, Y_{i(j-h)}) = \rho^h$ for $h = 0, \ldots, T$, which assumes correlations decay exponentially with time. We also tried fitting the models with an exchangeable, or compound symmetry, correlation structure, $\text{corr}(Y_{ij}, Y_{ik}) = \rho$ for $j \neq k$, which assumes pairwise correlations are the same (and therefore the order of the repeated measures is of no importance), but the estimates converged to the same values. An unstructured correlation structure, $\text{corr}(Y_{ij}, Y_{ik}) = \rho_{jk}$, which assumes all pairwise correlations are different, led to convergence failures (a common problem) in both the R functions `gee` {gee package} and `geeglm` {geepack package}.

4.4.2 Analysis Steps

Since the joint distribution is not specified when using generalized estimating equations, there is no likelihood function, and therefore we could not use likelihood ratios for inference, testing fit, or comparing models. However, due to the asymptotic normality of the estimators, Wald statistics are used for inference (which are fine for large samples). Namely, for a null hypothesis $H_0 : \beta = \beta_0$, the Wald test statistic is

$$z = \frac{\widehat{\beta} - \beta_0}{\sqrt{\widehat{\text{Var}}(\widehat{\beta})}},$$

which has an approximate standard normal distribution under H_0.

Estimators and standard errors that ignore the correlated structure of longitudinal data are termed "naïve" in the literature, as opposed to "robust" methods that take it into account. We adopt this terminology in the tables and discussion that

follow.

One Model for All Primary Conditions

Table 4.1 summarizes the estimates and standard errors of our model of patient discharge, with a covariate for primary conditions. We do not include the covariate for primary conditions in the table of results, although it was included in the model, because it would result in an extra 43 lines in the table (44 primary conditions minus one reference level). The scale parameter is a multiplier on the variance-covariance, and the correlation parameter is the correlation between the first two repeated measures, but across all patient encounters (it does not indicate the actual correlation structure of the data). These last two are nuisance parameters, so we ignore them for the most part.

Although there does not seem to be a significant difference between the naïve and robust estimates (and corresponding standard errors), we need to keep in mind that we had 4,400 patient encounters in our sample dataset used in fitting this model. Such a large sample, for a population-averaged model, will "smooth out" most effects. The difference between naïve and robust measures will be more pronounced when we consider fitting one model per primary condition, with 100 patient encounters per primary condition.

Residuals plots allow us to visualize the model fit and discrimination (between outcomes), as well as diagnose problems. Although residual plots for binary outcomes are less informative than with linear models, they could nonetheless provide us with important information. We considered standardized (or Pearson) residuals—the difference between observed and fitted values, divided by their standard errors—because they are also available for the subject-specific models we consider in Chapter 5. For

Table 4.1: Estimates from logistic regression (naïve) and generalized estimating equations (robust) for main-effects population-averaged model for all primary conditions, described by equation (4.4.1). Note that we have not included the covariate for primaryCondition in the table—although it was used in fitting the model—because it would result in 43 additional rows (44 factor levels minus one reference level). Refer to Table 1.2 for a description of covariates. C-index of 0.685, and Brier score of 0.107.

Covariate	**Naive Estimate**	**Robust Estimate**	**Naive S.E.**	**Robust S.E.**	**P-Value**
(Intercept)	-0.65934	-0.66180	0.17893	0.19132	0.0005
day	0.03272	0.03265	0.00280	0.00264	<0.0001
laps	-0.01092	-0.01093	0.00077	0.00084	<0.0001
age	-0.00656	-0.00656	0.00104	0.00115	<0.0001
charlson1-2	-0.28650	-0.28622	0.04829	0.05215	<0.0001
charlson3-4	-0.42506	-0.42462	0.05821	0.05964	<0.0001
charlson5+	-0.37558	-0.37510	0.06154	0.06271	<0.0001
icu1	-0.21381	-0.21373	0.10579	0.11191	0.0561
surgery1	-0.01615	-0.01616	0.05375	0.06079	0.7902
nursStaOcc	-0.00580	-0.00580	0.00045	0.00053	<0.0001
genderM	0.01836	0.01835	0.03538	0.03742	0.6237
campusG	0.03696	0.03694	0.04337	0.04512	0.4130
campusH	0.12083	0.12055	0.10242	0.10794	0.2640
accommodationS	0.09612	0.09610	0.04416	0.04503	0.0328
accommodationW	0.11922	0.11896	0.04809	0.05282	0.0243
admitTypeE	-0.24245	-0.24232	0.07577	0.08248	0.0033
admitTypeL	0.08133	0.08125	0.08668	0.09788	0.4064
admitTypeS	0.04439	0.04421	0.10215	0.11543	0.7017
admitTypeU	0.10592	0.10594	0.09654	0.10532	0.3144

Parameter	**Estimate**	**S.E.**
Scale	1.02	0.0541
Correlation	-0.00303	0.00155

outcomes y_i, with fitted means $\widehat{\mu}_i$, the standardized residuals are

$$r_i^s = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\text{Var}}(Y_i)}}$$

In the case of a GLM, it is also worth considering deviance residuals. Let L be the likelihood of the fitted model, and L_s be the likelihood of the saturated model (i.e., the model with a separate parameter for each observation). The deviance is then $D = -2\log(L/L_s) = \sum d_i$, and the deviance residuals are

$$r_i^d = \sqrt{d_i}\ \text{sign}(y_i - \widehat{\mu}_i).$$

It is worth noting that deviance residuals were between -2 and 3, suggesting that the model fit was acceptable (a typical criteria is that the deviance residuals fall between -3 and 3).

In Figure 4.1 we present two plots of the standardized residuals. The first is a plot of the residuals versus the predicted probability of discharge. If discharge is plotted on the logit scale, the curvature seen would mostly disappear. The residual plot is not unreasonable for a logistic regression with unsimulated data. The histogram shows that the residuals are skewed. This could indicate overdispersion, that the variance is greater than theory would suggest for a binomial distribution, or missing covariates (such as interactions).

We cannot assume, however, that the standardized residuals will be normally distributed, even when the model is correct. As described in [39], standardized residuals converge asymptotically to normality if the dispersion parameter (ϕ in Sections 4.2.1 and 4.2.2) converges to zero for a fixed sample size (referred to as small-dispersion asymptotics). Otherwise, we cannot be sure that the residuals will be closely normal, as shown in [12]. To complicate matters further, our observations were not independent, as we had repeated measures for patient encounters, which is not considered in the asymptotic theory that has been developed for the standardized residuals of GLMs.

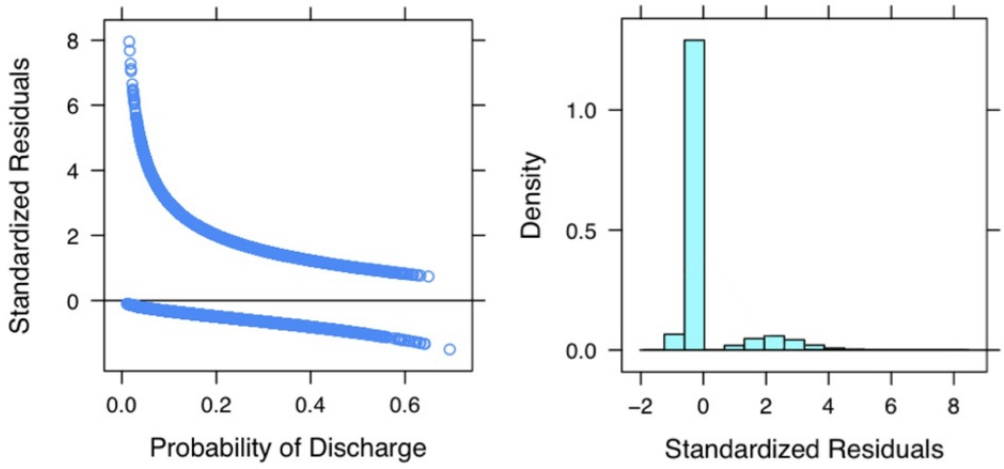

Figure 4.1: Scatter plot and histogram of residuals for main-effects population-averaged model including all primary conditions, described by equation (4.4.1). The standardized residuals are not normal, but we used these plots as a baseline upon which to judge our subsequent models.

Regardless, the deviance residuals suggested an acceptable fit. The adequacy of the fit of the model is also subjective, and dependent on its purpose. Our goal was not to develop an explanatory model, but a predictive model, and so we put more emphasis on predictive discrimination than fit. We used the residuals shown here as a baseline upon which to consider subsequent models, in this and the next chapter (to see how the fit changed, good or bad).

We also include a plot of the standardized residuals by primary condition in Figure 4.2. Our model included a covariate for primary condition, but we improved on our results by considering one model per primary condition, resulting in 44 models of discharge. Plotting the residuals by primary condition therefore allows us to consider the model fit for each primary condition now and then show the fitted model for each primary condition. We can see from these plots that the model did fit certain primary conditions better than others (since the residuals are less skewed and closer to 0). These plots also serve as the basis upon which we will compare improvements to the model, in this and the next chapter.

Finally we consider the predictive accuracy of our model using the c-index and Brier score in Table 4.2. Recall that we are using the c-index to rank predictive performance. It measures the area under the ROC curve (i.e., the true positive rate versus false positive rate), explained in Section 1.3.4. A value of 0.8 is considered a minimum if a model is to be deemed useful for prediction. Since the c-index is a rank order statistic, we also use the Brier score—the mean-squared error of prediction—as a second measure upon which to evaluate the predictive performance of the model. In this case, however, there is no threshold upon which to judge the model. Rather, we use it to compare models, where a lower score is better (again described in Section 1.3.4).

In order to improve on the predictive accuracy of the discharge model, we adjusted the coefficient estimates by fitting the model on different samples of our dataset. Namely, we fit the model using the last 7 days before discharge, and the last 14 days

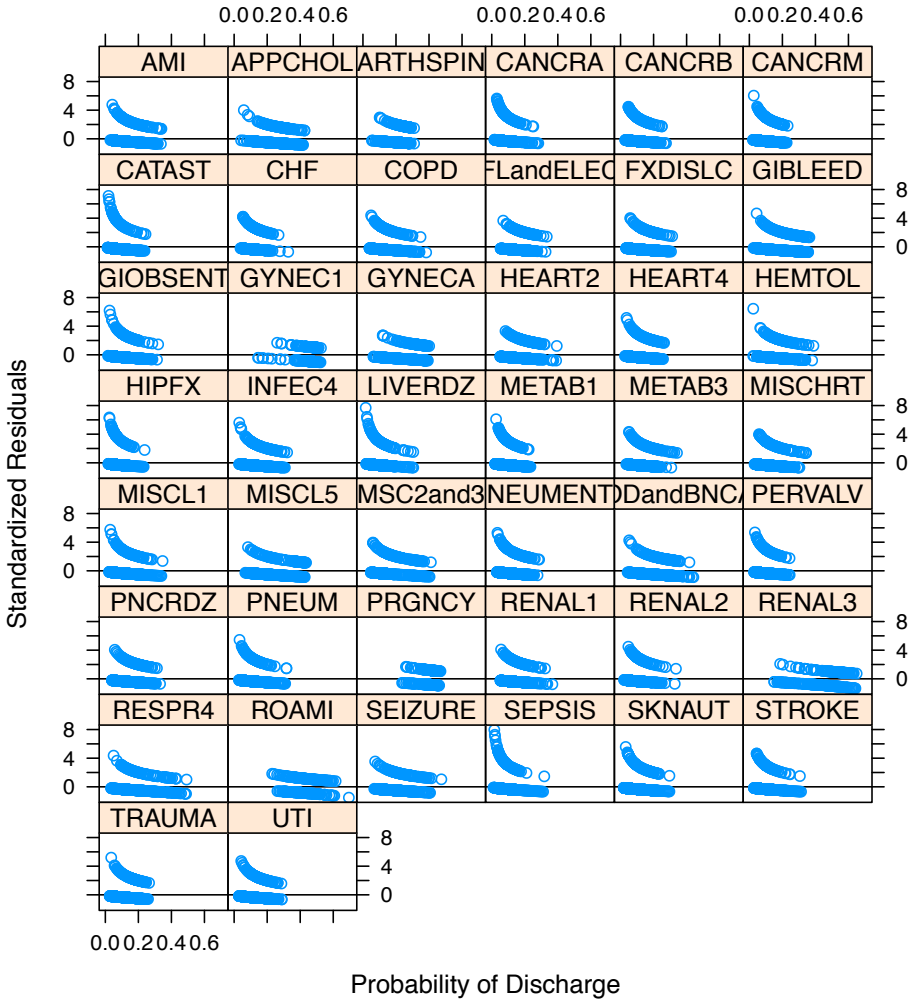

Figure 4.2: Scatter plots of residuals versus probability of discharge, by primary condition, for main-effects population-averaged model including all primary conditions, described by equation (4.4.1). In order to compare with other models, we consider the residuals as individual groups, by primary condition, even though our model is for all primary conditions.

Table 4.2: C-index (and Brier score) for main-effects population-averaged model including all primary conditions, described by equation (4.4.1). Training refers to the data on which the model was built, and validation refers to the data on which the model was applied.

	Validation		
Training	Last 7 Days	Last 14 Days	Last 28 Days
Last 7 Days	0.774 (0.143)	0.627 (0.258)	0.508 (0.407)
Last 14 Days		0.713 (0.127)	0.615 (0.143)
Last 28 Days			0.685 (0.107)

Adj.-R^2: 0.250 for last 7 days, 0.130 for last 14 days, and 0.101 for last 28 days.

before discharge, to eliminate the effect of long lengths of stay. We kept the last day for each patient encounter in order to keep the positive outcome (otherwise we would have skewed our sample significantly—the mean patient length of stay was 9 days). It was determined that the model needed to perform well on 28 days, and so we applied the models with adjusted coefficients, regardless of the number of days on which they were fit, on the larger samples as well. As we can see in Table 4.2, although the measures of predictive accuracy improved on the smaller datasets (with fewer days) compared to the model fit on 28 days, they did not improve once applied to the 28-day sample, and none achieved the requisite c-index of 0.8.

One Model for Each Primary Condition—44 Models Total

We now consider fitting models by primary condition in order to improve on predictive accuracy. Our model for patient encounter i, with repeated measures for time t, is given by

$$\begin{aligned}\text{logit}(\text{E}(\text{exit}_{ti}) = \beta_0 &+ \beta_1 \text{day}_t + \beta_2 \text{laps}_{ti} + \beta_3 \text{age}_i + \beta_4 \text{charlson}_i \\ &+ \beta_5 \text{icu}_{ti} + \beta_6 \text{surgery}_{ti} + \beta_7 \text{nursStaOcc}_{ti} + \beta_8 \text{gender}_i + \beta_9 \text{campus}_i \\ &+ \beta_{10} \text{accommodation}_i + \beta_{11} \text{admitType}_i. \end{aligned} \tag{4.4.2}$$

In this case we cannot give a single table of estimates and standard errors, and

there is very little value in considering 44 tables, by primary condition. It is of value to consider one primary condition, however, to see how one model of discharge for a single primary condition, with a sample of 100 patient encounters, compares to our model of discharge for all primary conditions, considered previously. For example, take the primary condition UTI (urinary tract infections) in Table 4.3.

Ignoring the time-dependency of longitudinal data will typically overestimate the standard errors of time-dependent covariates, and underestimate the standard errors of time-independent covariates. We see this effect for the time-dependent covariates for day and LAPS, and some of the time-independent covariates (but not all) in Table 4.3. More dramatic are the differences in naïve and robust standard errors, especially for highly predictive covariates with small sample sizes. For this particular primary condition, the estimates and standard errors for the surgery indicator and factor level L (elective) of the admit type demonstrate the importance of correcting for correlation in longitudinal data. The large confidence intervals could be due to rare events when assuming complete independence (e.g., elective surgeries, combining both covariates, will often result in a twenty-four hour stay, making them highly predictive; but the rare cases where elective surgeries result in longer hospital stays will significantly increase the confidence intervals when observations are assumed to be independent even for the same patient). Or they could be due to instabilities in the estimation routine used to fit the model.

It is worth noting that the difference in sample sizes between the model of discharge for one primary condition (100 patient encounters) and the model for all primary conditions (4,400 patient encounters) implies that we cannot compare statistical significance for covariates. With such a large sample size, it is no wonder that most covariates were highly statistically significant for the model with all primary conditions. On the other hand, we cannot expect to find the same levels of statistical significance for the model that was fit for only one primary condition.

Also, we could not eliminate covariates from all 44 models, by primary condition,

Table 4.3: Estimates from logistic regression (naïve) and generalized estimating equations (robust) for main-effects population-averaged model of primary condition UTI only, described by equation (4.4.2). Note the significant difference between naïve and robust estimates—factor levels surgery1 and admitTypeL, in particular. No estimates exist for campusH and admitTypeS because these factor levels do no exist for the UTI sample. Refer to Table 1.2 for a description of covariates. C-index of 0.675, and Brier score of 0.107.

Covariate	**Naive Estimate**	**Robust Estimate**	**Naive S.E.**	**Robust S.E.**	**P-Value**
(Intercept)	0.29330	0.27363	0.78870	0.70521	0.6980
day	0.04985	0.04928	0.02136	0.01903	0.0096
laps	-0.01214	-0.01214	0.00535	0.00511	0.0175
age	-0.00895	-0.00893	0.00644	0.00744	0.2301
charlson1-2	-0.27926	-0.27607	0.32740	0.28544	0.3334
charlson3-4	-0.11524	-0.11388	0.34620	0.30381	0.7077
charlson5+	-0.12581	-0.13005	0.42781	0.42894	0.7617
icu1	-0.31697	-0.30131	1.08303	1.05467	0.7751
surgery1	12.39452	42.02631	132.2	1.05654	<0.0001
nursStaOcc	-0.00896	-0.00898	0.00336	0.00367	0.0145
genderM	0.49594	0.49426	0.27264	0.25017	0.0481
campusG	-0.42757	-0.42325	0.26761	0.28425	0.1364
campusH	.	.	.	.	.
accommodationS	-0.53237	-0.52766	0.33764	0.35739	0.1398
accommodationW	-0.31053	-0.30628	0.38053	0.41951	0.4653
admitTypeE	-0.60532	-0.60340	0.57890	0.45963	0.1892
admitTypeL	11.31243	41.23495	132.2	1.07272	<0.0001
admitTypeS	.	.	.	.	.
admitTypeU	1.77627	1.79254	1.55000	0.52770	0.0006

Parameter	**Estimate**	**S.E.**
Scale	1.01	0.395
Correlation	-0.0165	0.00808

by considering the estimates, standard errors, or statistical significance of the fitted model for one primary condition, since the results are very different for each. Not only do sample sizes change for covariates between primary conditions, but so do the estimates themselves. An extreme example is those covariates, or factor levels, for which there was no data, such as factor level H in campus, or factor level S in admit type, in Table 4.3. Evaluating covariates by comparing 44 tables of estimates and standard errors by primary conditions would be far too cumbersome, and make implementation more difficult. Not to mention that missing factor levels in our sample may not be missing in another.

Next we consider plots of standardized residuals in Figures 4.3 and 4.4. The former groups residuals across the 44 models of discharge, by primary condition, into a single set of residuals; the latter plots residuals by primary conditions. Figure 4.3 allows us to compare it to the previous model for all primary conditions. In this case we see that the residuals near a predictive probability of discharge of 0.5 were closer to 0, which suggests better predictive discrimination (also seen in Figure 4.4). However, the maximum value of the residuals, near predictive probability of discharge of 0, increased. The histogram suggests that our residuals were slightly more normally distributed than before, but still skewed.

Although the residuals may suggest the possibility that our models had better predictive discrimination, Table 4.4 seems to confirm it. The improvement in the c-index is moderate, but these models still did not achieve a c-index greater than 0.8 on the 28 day sample, and the Brier score worsened (having increased). We were able to achieve a c-index just above 0.8 (and mean adjusted-R^2 just above 0.5) by including all interactions with day and LAPS (our primary time-varying covariates), as is commonly done in longitudinal models. This led, however, to convergence problems (the details of which are described by Allison in [4]). We ignored interactions for our population-averaged models, preferring to improve performance by way of a subject-specific model, as will be described in Section 5.3.

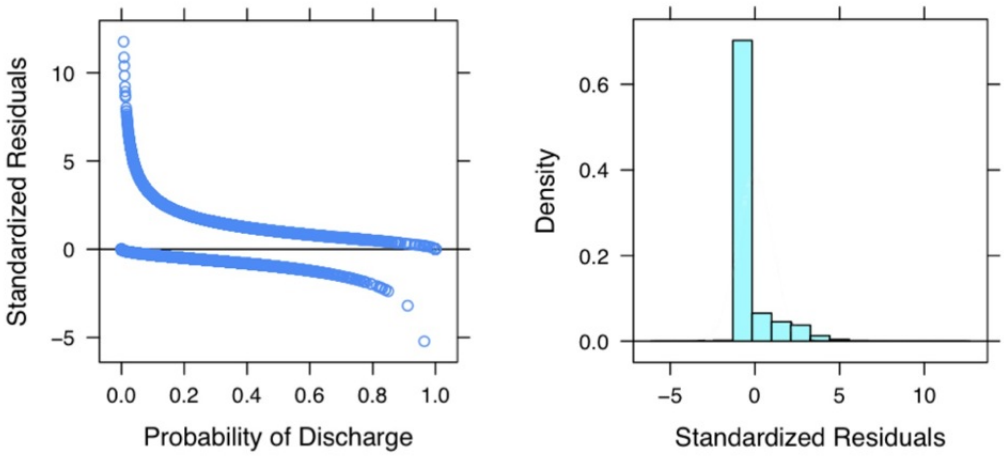

Figure 4.3: Scatter plot and histogram of residuals for 44 main-effects population-averaged models, one model for each primary condition, described by equation (4.4.2). In order to compare with other models, we consider the residuals as a single group, rather than by primary condition. The standardized residuals are not normal.

Table 4.4: Mean c-index (and Brier score) for 44 main-effects population-averaged models, one model for each primary condition, described by equation (4.4.2). Training refers to the data on which the models were built, and validation refers to the data on which the models were applied. Since there were 44 models, we calculated the mean c-index and Brier score. Although an improvement to the previous population-averaged model, the subject-specific model of the next chapter will achieve even better predictive performance.

	Validation		
Training	Last 7 Days	Last 14 Days	Last 28 Days
Last 7 Days	0.805 (0.135)	0.677 (0.266)	0.570 (0.393)
Last 14 Days		0.745 (0.126)	0.654 (0.168)
Last 28 Days			0.717 (0.113)

Mean adj.-R^2: 0.321 for last 7 days, 0.188 for last 14 days, and 0.147 for last 28 days.

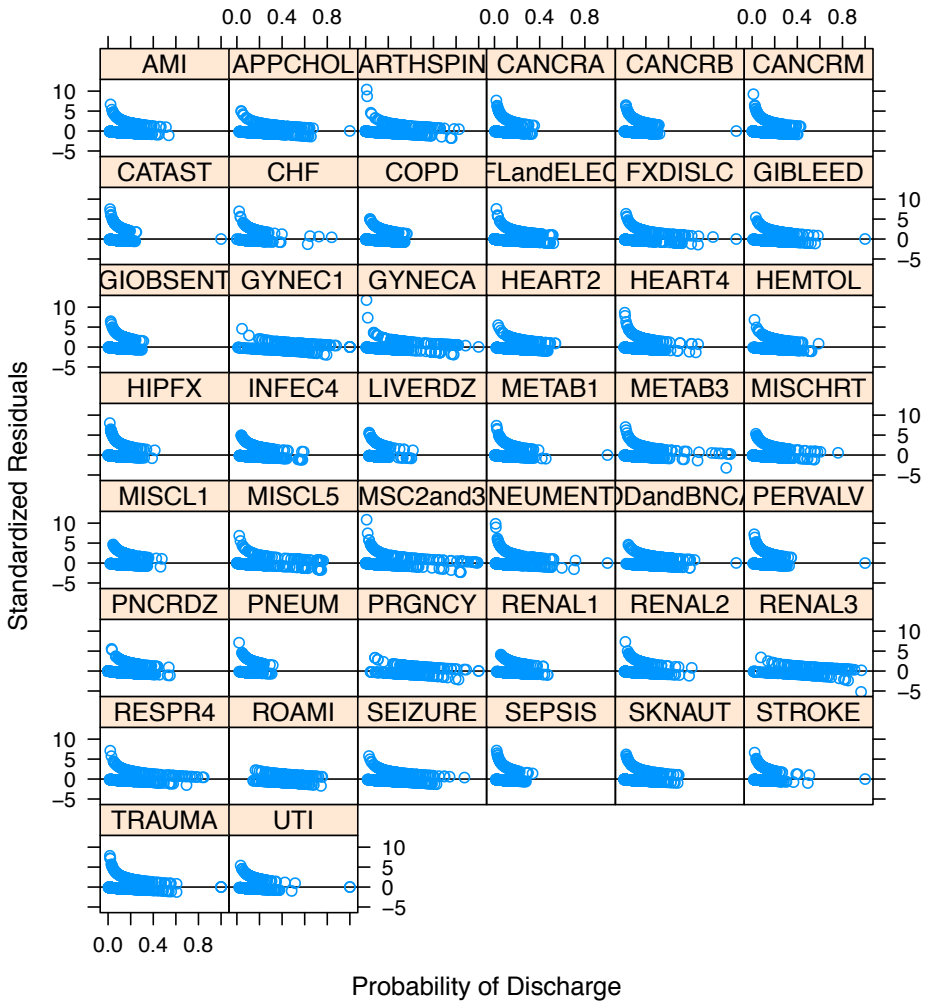

Figure 4.4: Scatter plots of residuals versus probability of discharge, by primary condition, for 44 main-effects population-averaged models, one model for each primary condition, described by equation (4.4.2). Each scatter plot is of one distinct model, since one model was fit to each primary condition.

Given that the approach of fitting 44 models moderately improved the c-index, but also moderately worsened the Brier score, we were uncertain as to which approach was better. We were, however, able to significantly improve on these results with using subject-specific models, as described in the next chapter.

4.4.3 Interpreting the Model

It is important to recognize that the parameters in a marginal model provide us with the average trend in the population. That we have fit one model for each primary condition allowed us to get better estimates of discharge for the primary conditions, as we have seen—but patients in the hospital are not grouped by primary condition. Even a nursing station can have patients from multiple hospital services, let alone primary conditions. This makes the population-averaged model difficult to employ in practice (whereby an administrator may wish to get an estimate of the number of patients that will be discharged for a single nursing station, or a hospital service, or just the floor of a specific building). The model we consider in the next chapter will improve on this (at which point we consider interpreting the model in greater detail).

Chapter 5

Subject-Specific Model for Discharge

In the previous chapter, we considered a population-averaged model of patient discharge, which essentially averages over any random effects from subject-specific variation. In this chapter, however, we wish to model the random effects to improve on predictive accuracy. As noted by Agresti, in Chapter 12 of [2], subject-specific models are well established for normal responses (conditional on predictors), but the extension to categorical data is relatively recent. Although we are building on the theory of generalized linear models presented in the previous chapter, estimating the random effects, or predictors, adds significant technicalities. Therefore we present supporting theory before discussing the fitted model in Section 5.3.

5.1 Linear Mixed Models

In this section we will discuss linear mixed models in general, but with a particular focus on multilevel structures, typically referred to as random coefficient or random effects models in statistics. First we consider a linear mixed model with only one

random effect term. To motivate this section, we provide a graphical example of a random-intercept model with a single fixed effect, as show in Figure 5.1.

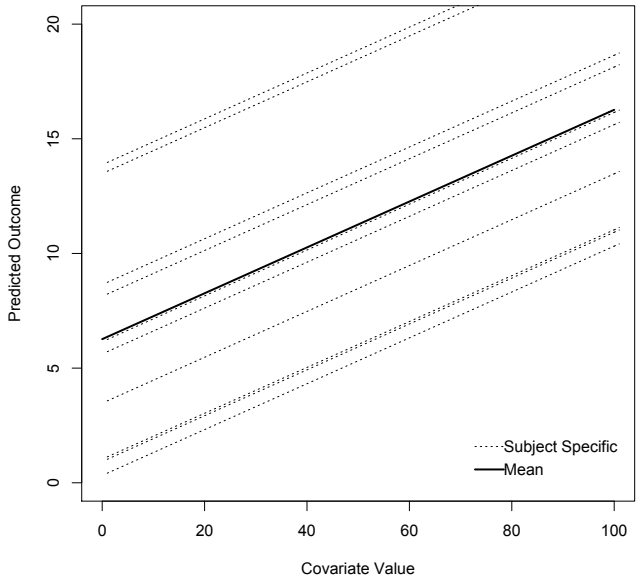

Figure 5.1: Example of linear regression with random intercepts. The mean response is what we would expect if clustering were ignored (or marginalized). However, using a random-intercept model we can find subject-specific responses (based on subject-specific clustering).

Say, for example, that we have a continuous outcome variable representing severity of illness, and an explanatory variable representing age. We could formulate a linear model to study the affect of age on severity of illness. However, assume domain knowledge suggests that severity of illness will vary by primary condition (an initial classification of illness, as described in Appendix A). We may therefore wish to study how age affects severity of illness by primary condition. That is to say, we may wish to group patients by primary condition (our clustering variable) and study the effect of age on severity of illness for each group of patients.

5.1.1 Introducing Random Effects

Let Y_{ij} be our continuous outcome variable, and $\boldsymbol{X}_{ij}$ be our vector of covariates, similar to our formulation for linear models. In this case, however, note that we have two subscripts on our outcome and covariates. The first subscript, i, is used to identify a pre-defined cluster, and the second subscript, j, is used to identify elements nested within a pre-defined cluster. In our example above, i would be primary conditions, and j would be patients. Therefore, X_{ij} would be the age of patient j in the primary condition group i, with a severity of illness captured by Y_{ij}.

In the terminology of multilevel modelling, the index i represents the second level of our data (the clustering), and index j represents the first level of our data (the unit of analysis). This allows us to define covariates specific to each level of our data. Also, we add the random effect u_i to the model, which will relate the influence of clusters (e.g., primary conditions) on elements in the clusters (e.g., patient age) not already captured by the covariates. An important element of estimating coefficients in a linear mixed model is thus the variance of the random effects (i.e., the variability between clusters).

Our linear mixed model, with random intercept, can be written as

$$Y_{ij} = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + u_i + \varepsilon_{ij},$$

Again, we can compare the linear model in 4.1.1 to the form of the linear mixed model below (split among two lines to relate it to the generalization with a link function that will follow),

$$\begin{aligned} \mathrm{E}(Y_{ij}|u_i) = \quad & \mu_{ij} \\ & \mu_{ij} \quad = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + u_i. \end{aligned} \tag{5.1.1}$$

In our example, we could think of the fixed intercept β_0 as the baseline severity of illness for all groups (for age $X_{ij} = 0$), and the random intercept u_i as the specific

effect on severity of illness for primary condition i. Severity of illness would therefore have a specific linear profile for each primary condition, such as that shown in Figure 5.1.

Given that we have introduced a new random element to the model, it is important to consider distributional assumptions. We again assume the errors ε_{ij} are iid $\mathrm{N}(0, \sigma^2)$, but also that the random effects u_i are iid $\mathrm{N}(0, \sigma_u^2)$, although with a different variance. Furthermore, we assume the errors and random effects are independent of one another. This implies that the outcomes Y_{ij} are also normally distributed.

We can extend the linear mixed model to have any number of random effects (and in vector notation for subject i)

$$\begin{aligned} \mathrm{E}(\boldsymbol{Y}_i|\boldsymbol{u_i}) = \;& \boldsymbol{\mu}_i \\ \boldsymbol{\mu}_i \;& = \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{Z}_i'\boldsymbol{u_i}, \end{aligned} \tag{5.1.2}$$

and assume the random effects $\boldsymbol{u}_i$ are iid $\mathrm{N}(\boldsymbol{0}, \boldsymbol{G})$, and the errors $\boldsymbol{\varepsilon}_i$ are iid $\mathrm{N}(\boldsymbol{0}, \boldsymbol{R}_i)$, where $\boldsymbol{G}, \boldsymbol{R}_i$ are the respective covariance matrices. Therefore, the conditional distribution of $\boldsymbol{Y}_i|\boldsymbol{u}_i$ is $\mathrm{N}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i, \boldsymbol{R}_i)$, where $\boldsymbol{X}_i$ is the design matrix for our covariates, and $\boldsymbol{Z}_i$ is the design matrix for our random effects. Outside of a Bayesian framework, estimation involves the marginal distribution of the outcome vector $\boldsymbol{Y}_i$ which is $\mathrm{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R}_i)$ (since random effects and errors are independent and normal).

5.1.2 Residual Maximum Likelihood Estimation

Often a common variance term is factored from the covariance matrices $\boldsymbol{G}$ and $\boldsymbol{R}_i$, so that it can be estimated while assuming the covariance matrices are known. In more specific multilevel models, authors will at times assign a set of parameters in $\boldsymbol{\phi}$ for the covariance matrices, specific to the structure of their model, instead of a single common variance. The main point, however, is that we need to estimate the

common variance or common variance parameters. We therefore make explicit the $\boldsymbol{\phi}$ in the variance of the marginal model as $V_i(\boldsymbol{\phi}) = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R}_i$.

We can use maximum likelihood of the marginal model to find estimates for the fixed effects $\boldsymbol{\beta}$ and the variance components in $\boldsymbol{\phi}$. Assuming independence, the marginal likelihood is given by

$$L(\boldsymbol{\beta}, \boldsymbol{\phi}) = \prod_{i=1}^{n} \left[(2\pi)^{-n_i/2} |V_i(\boldsymbol{\phi})|^{-1/2} \exp\left(-\frac{1}{2} (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})' V_i^{-1}(\boldsymbol{\phi}) (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \right) \right],$$

where i is the subject, n is the number of subjects, and n_i is the number of observations for subject i. Joint maximization will then yield estimates for the $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$, which requires the use of numerical optimization.

However, estimates for the variance components $\boldsymbol{\phi}$, using maximum likelihood, are often biased downwards (underestimating the variance) due to a lack of degrees of freedom from the estimation of the fixed effects $\boldsymbol{\beta}$, especially for highly unbalanced data (i.e., data for which the sample size n_i varies considerably between subjects), as noted by Harville in [21], and Patterson and Thompson in [36]. This leads to smaller confidence intervals for the variance components than is justified. Instead we use restricted or residual maximum likelihood (REML) to estimate the variance components $\boldsymbol{\phi}$, which is essentially maximum likelihood on ordinary least squares residuals.

The idea of restricted maximum likelihood is to transform the data using a matrix that is orthogonal to the design matrix of the fixed effects, i.e., $\boldsymbol{A}_i'\boldsymbol{X}_i = \boldsymbol{0}$. Patterson and Thompson in [38] and [36] motivated the concept of REML by considering the matrix $\boldsymbol{A}_i = \boldsymbol{I} - \boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'$, which transforms the outcomes in $\boldsymbol{Y}_i$ to ordinary least squares residuals (since $\boldsymbol{A}_i = \boldsymbol{I} - \boldsymbol{H}_i$, where $\boldsymbol{H}_i$ is the hat matrix). In doing as such, the transformed outcomes $\boldsymbol{A}_i'\boldsymbol{Y}_i$ have marginal distribution $N(\boldsymbol{0}, \boldsymbol{A}_i'V_i(\boldsymbol{\phi})\boldsymbol{A}_i)$.

The degrees of freedom used in the maximum likelihood estimation of the fixed effects $\boldsymbol{\beta}$ are therefore repurposed in REML to the estimation of the variance components $\boldsymbol{\phi}$. Furthermore, the estimated variance components from REML are not

biased, even for highly unbalanced data. Using the estimates of the variance components $\boldsymbol{\phi}$, the fixed effects $\boldsymbol{\beta}$ can be estimated more readily using marginal maximum likelihood or, equivalently, weighted least squares.

5.1.3 Best Linear Unbiased Predictors

Estimators of the random effects are typically referred to as predictors, since we are dealing with unobservables sampled from a population. The idea of estimating random effects, however, was not always accepted by the classical school of statistics, and may still be controversial in some stricter circles. Given that it is of practical use, it has therefore progressed regardless of some ideological views. The Bayesian school of statistics, however, has no apparent problem with the idea (given that probability distributions are used to describe variation and uncertainty equally, as noted by Robinson in [40]). We will discuss this again shortly, but for now we ignore possible objections to estimating random effects.

One way to calculate estimators for the random effects is to apply weighted least squares (to the untransformed model), but this would require calculating the inverse of the covariance matrix for the outcomes, where $\text{Var}(\boldsymbol{Y}_i) = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R}_i$, which is computationally difficult and costly. Instead, we can maximize the joint density of the outcomes $\boldsymbol{Y}_i$ and random effects $\boldsymbol{u}_i$, by minimizing

$$\begin{pmatrix} \boldsymbol{u}_i \\ \boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{u}_i \end{pmatrix}' \begin{bmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_i \end{bmatrix}^{-1} \begin{pmatrix} \boldsymbol{u}_i \\ \boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{u}_i \end{pmatrix}$$

with respect to the fixed effects $\boldsymbol{\beta}$ and random effects $\boldsymbol{u}_i$.

Differentiating with respect to the fixed effects $\boldsymbol{\beta}$ and random effects $\boldsymbol{u}_i$ and setting them equal to zero results in the following set of equations:

$$\begin{bmatrix} \boldsymbol{X}_i'\boldsymbol{R}_i^{-1}\boldsymbol{X}_i & \boldsymbol{X}_i'\boldsymbol{R}_i^{-1}\boldsymbol{Z}_i \\ \boldsymbol{Z}_i'\boldsymbol{R}_i^{-1}\boldsymbol{X}_i & \boldsymbol{Z}_i'\boldsymbol{R}_i^{-1}\boldsymbol{Z}_i + \boldsymbol{G}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{u}}_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_i'\boldsymbol{R}_i^{-1}\boldsymbol{Y}_i \\ \boldsymbol{Z}_i'\boldsymbol{R}_i^{-1}\boldsymbol{Y}_i \end{bmatrix}.$$

These equations are known as the mixed model equations, or the equations for the best linear unbiased predictors (BLUP). It is important to note, however, that unbiased here means that the expected value of the predictors is equal to the expected value of the random effects (i.e., $\mathrm{E}(\widehat{\boldsymbol{u}}) = \mathrm{E}(\boldsymbol{u})$). Note that the left-hand matrix is also the covariance matrix of estimation and prediction errors, assuming the design matrix $\boldsymbol{X}_i$ is full column rank (as there are more observations than there are parameters to estimate), as shown by Henderson in [22].

The BLUP equations given above were first presented in this way by Henderson in 1950 [22], who was doing work to estimate genetic differences in cows. He cautioned that this should not be called maximum likelihood given that unobservables are being considered, and instead called his derivation joint maximum likelihood estimation. His derivation was later generalized and not limited to his idea of joint likelihood, proving it to be quite robust. Also, Harville in [21] showed that BLUP estimators can be used to derive maximum likelihood or REML estimators for the variance components $\boldsymbol{\phi}$ (whereas all three methods arrive at the same estimates for the fixed effects $\boldsymbol{\beta}$).

The term BLUP is attributed to Goldberger in 1962 [19] who considered the problem of having a new set of observations and a new unobservable prediction error, correlated with previous errors. In his development the same set of solution equations are given, which provides a derivation from the classical school of statistics. In fact, there are a few different methods to derive the BLUP equations, including from the Bayesian school. As a result, this method of finding estimators for random effects is given multiple names, including BLUP likelihood, BLUP procedure, quasi-likelihood, and the method of most likely unobservables, among others. An excellent discussion on the topic of BLUP and its history is given by Robinson in [40].

5.2 Generalized Linear Mixed Models

As in the case of linear models, we wish to broaden our models to a larger class of distributions. Interestingly, McCullagh and Nelder provided a motivating example for generalized linear mixed models in Section 14.5 of [31], in 1989, before such models were given considerable attention.

Similar to our introduction to generalized linear models, we introduce a link function and consider the conditional model

$$\begin{aligned} \mathrm{E}(\boldsymbol{Y}_i|\boldsymbol{u_i}) = &\quad \boldsymbol{\mu}_i \\ g(\boldsymbol{\mu}_i) &= \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{Z}_i'\boldsymbol{u}_i, \end{aligned} \tag{5.2.1}$$

where the conditional distribution of $\boldsymbol{Y}_i|\boldsymbol{u}_i$ is not normal, and not necessarily from the exponential family either. We still assume, however, that the random effects $\boldsymbol{u}_i$ are iid $N(\boldsymbol{0}, \boldsymbol{G})$ (where $\boldsymbol{G}$ is fully described by a set of variance components $\boldsymbol{\phi}$).

The problem is that it is generally difficult to determine the marginal distribution of the outcomes. Consider the likelihood

$$L(\boldsymbol{\beta}, \boldsymbol{\phi}) = \prod_{i=1}^{n} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\beta}, \boldsymbol{\phi}) f(\boldsymbol{u}_i) \mathrm{d}\boldsymbol{u}_i,$$

where i is the subject, n is the number of subjects, n_i is the number of observations for subject i, $f_{ij}(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\beta}, \boldsymbol{\phi})$ is the conditional density of $Y_{ij}|\boldsymbol{u}_i$, and $f(\boldsymbol{u}_i)$ is the normal density of the random effects $\boldsymbol{u}_i$. The integral is the marginal density of $\boldsymbol{Y}_i$, which can rarely be solved analytically. Furthermore, there are n such integrals to solve in an effort to maximize the likelihood. This can be done, for example, using numerical integration (as done in the SAS procedure `nlmixed` [27] or R function `glmer` {lme4 package}).

5.2.1 Penalized Quasi-Likelihood

An alternative approach, and the one we used in this study, is to consider a first-order approximation about $\boldsymbol{\mu}_{ij}$ of the link function to the data, and estimate the mean response by applying the link function to the data and random effects (i.e., $g^{-1}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u})$) for current values of estimators and predictors. This allows us to consider a weighted least squares regression for the linear mixed model, and equivalently BLUP (where the covariance matrix $\boldsymbol{R}_i$ for the errors is now replaced by a weight matrix that depends on the current estimates). The idea is therefore similar to that used for generalized linear models, as seen in Section 4.2.2.

Consider the outcomes

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i = g^{-1}(\boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{Z}_i'\boldsymbol{u}_i) + \boldsymbol{\varepsilon}_i,$$

where the error terms $\boldsymbol{\varepsilon}_i$ have the relevant distribution (e.g., from an exponential family) with conditional covariance matrix given by $\mathrm{Var}(\boldsymbol{Y}_i|\boldsymbol{u}_i) = \mathrm{Diag}(v(\boldsymbol{\mu}_i)) = \boldsymbol{V}_i$ (i.e., each error term ε_{ij} has conditional variance $\mathrm{Var}(Y_{ij}|\boldsymbol{u}_i) = v(\mu_{ij})$, and they are conditionally independent). Now we consider a first-order approximation about the current estimates of the fixed effects $\boldsymbol{\beta}$ and random effects $\boldsymbol{u}_i$,

$$\boldsymbol{Y}_i \approx \widehat{\boldsymbol{\mu}}_i + \widehat{\boldsymbol{V}}_i\boldsymbol{X}_i(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + \widehat{\boldsymbol{V}}_i\boldsymbol{Z}_i(\boldsymbol{u}_i - \widehat{\boldsymbol{u}}_i) + \boldsymbol{\varepsilon}_i.$$

Rearranging we get

$$\underbrace{\widehat{\boldsymbol{V}}_i^{-1}(\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_i) + \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\widehat{\boldsymbol{u}}_i}_{\boldsymbol{Z}_i} \approx \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i + \underbrace{\widehat{\boldsymbol{V}}_i^{-1}\boldsymbol{\varepsilon}_i}_{\boldsymbol{\varepsilon}_i^z},$$

where $\mathrm{E}(\boldsymbol{\varepsilon}_i^z) = 0$, and $\boldsymbol{Z}_i$ is our new dependent variable. The form of this equation is that of a linear mixed model, for pseudo-data $\boldsymbol{Z}_i$.

An algorithm for solving then becomes obvious, as we turn to the estimation process from Section 5.1 for linear mixed models, and iterate in a similar fashion to the weighted least squares procedure seen in Section 4.2.2 for generalized linear

models (the approach taken by the SAS procedure `glimmix` [27] or R function `glmmPQL` {MASS package} [48]). Breslow and Clayton provide additional details in [8].

5.3 Evaluating the Subject-Specific Model

We formulate our model in terms of a hierarchy of simpler models, corresponding to the levels of clustered longitudinal data (i.e., repeated measures data), to make it easier to describe. In our case, we group patient encounters by their primary condition (an initial classification of illness, as described in detail in Appendix A), and measure each patient's status daily—resulting in longitudinal data for each patient encounter, clustered by primary condition.

The structure of the data can be best seen in Figure 5.2, for three patient encounters with three repeated measures each (although our model and data are not limited to balanced data—i.e., equal cluster sizes, with measures taken at the same time, which can often be used, at least in simpler cases, to derive closed-form solutions of estimators).

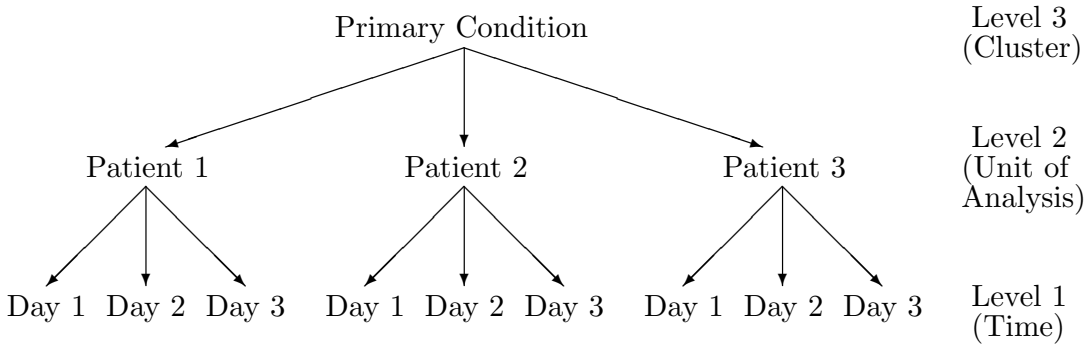

Figure 5.2: Multilevel structure of the clustered longitudinal patient data.

The structure shown in the figure explains the name *multilevel model*, although it is often referred to as a *random-coefficient model* in statistics, given its mathematical description. Formulating our model this way allowed us to include random effects

associated with the clusters, primary conditions, and the units of analysis, patient encounters, while taking into account correlation between the residuals of the longitudinal measures for the same unit of analysis, resulting in a subject-specific model (i.e., a patient-specific model). It would allow the hospital to estimate the probability of discharge for a single patient, or any group of patients—e.g., by nursing station, service, or ward.

5.3.1 Model Formulation

What follows is our formulation of the model, using all suitable covariates in Table 2.1 and interactions (following [51], Section 7.3). As with the population-averaged model of Section 4.4.1, we do not explicitly outline a parameter for the factor levels in our class variables. For example, admitType has five factor levels, and therefore has four parameters to estimate (the number of factor levels minus one reference level). Instead, we only show one parameter, for simplicity.

We could have included all time-varying covariates at level 1, but chose to consider only day and LAPS in order to reduce the number of interactions that needed to be considered (also, ICU and surgery indicators would normally only change once, if at all, during a patient's stay, and nursing station occupancy did not typically follow a trend that would require consideration at level 1 of our model). Note that our outcome exit_{tij} is binary, and that n_{ij} is the number of repeated measures for a patient encounter i in primary condition j. Also note that the subscript t is used on time-varying covariates only.

Level 1 model—time t

$$\begin{aligned} \text{E}(\text{exit}_{tij}|\boldsymbol{u}_{ij}) = \quad & \mu_{tij} \\ \text{logit}(\mu_{tij}) \quad & = b_{0i|j} + b_{1i|j}\text{day}_t + b_{2i|j}\text{laps}_{tij} \end{aligned}$$

where $\boldsymbol{u}_{ij} = (u_{0ij}, u_{0j}, u_{1j}, u_{2j})'$ are the random effects.

Level 2 model—patient encounter i

(a) Intercept and main effects, plus a random intercept for patient encounter (in primary condition):

$$\begin{aligned} b_{0i|j} &= b_{0j} + \beta_3 \text{age}_{ij} + \beta_4 \text{charlson}_{ij} + \beta_5 \text{icu}_{tij} + \beta_6 \text{surgery}_{tij} + \beta_7 \text{nursStaOcc}_{tij} \\ &\quad + \beta_8 \text{gender}_{ij} + \beta_9 \text{campus}_{ij} + \beta_{10} \text{accommodation}_{ij} + \beta_{11} \text{admitType}_{ij} + u_{0ij} \end{aligned}$$

$$\text{where } u_{0ij} \sim \text{N}(0, \sigma^2_{ij})$$

(b) Slope and interaction effects for day:

$$\begin{aligned} b_{1i|j} &= b_{1j} + \beta_{12} \text{age}_{ij} + \beta_{13} \text{charlson}_{ij} + \beta_{14} \text{icu}_{tij} + \beta_{15} \text{surgery}_{tij} + \beta_{16} \text{nursStaOcc}_{tij} \\ &\quad + \beta_{17} \text{gender}_{ij} + \beta_{18} \text{campus}_{ij} + \beta_{19} \text{accommodation}_{ij} + \beta_{20} \text{admitType}_{ij} \end{aligned}$$

(c) Slope and interaction effects for laps:

$$\begin{aligned} b_{2i|j} &= b_{2j} + \beta_{21} \text{age}_{ij} + \beta_{22} \text{charlson}_{ij} + \beta_{23} \text{icu}_{tij} + \beta_{24} \text{surgery}_{tij} + \beta_{25} \text{nursStaOcc}_{tij} \\ &\quad + \beta_{26} \text{gender}_{ij} + \beta_{27} \text{campus}_{ij} + \beta_{28} \text{accommodation}_{ij} + \beta_{29} \text{admitType}_{ij} \end{aligned}$$

Level 3 model—primary condition j

$$\begin{aligned} b_{0j} &= \beta_0 + u_{0j} \quad \text{(fixed plus random intercept)} \\ b_{1j} &= \beta_1 + u_{1j} \quad \text{(fixed plus random slope for day)} \\ b_{2j} &= \beta_2 + u_{2j} \quad \text{(fixed plus random slope for laps)} \end{aligned}$$

$$\text{where } (\varepsilon_{1ij}, \ldots, \varepsilon_{n_{ij}ij})' \sim \text{N}(\mathbf{0}, \boldsymbol{R}_{ij}), \text{ and } (u_{0j}, u_{1j}, u_{2j})' \sim \text{N}(\mathbf{0}, \boldsymbol{D})$$

The above model is given in terms of a multilevel formulation, but we can just as well incorporate the levels into a single equation (the disadvantage being that we do not readily see the nesting and their relation to the hierarchy in the data, as well as the size of the model in general).

Collapsed model—combining levels 1, 2, and 3

$$\text{logit}(\mu_{tij}) = \beta_0 + u_{0j} + u_{0ij} + \beta_3 \text{age}_{ij} + \beta_4 \text{charlson}_{ij} + \beta_5 \text{icu}_{tij}$$

$$
\begin{aligned}
&+\beta_6 \text{surgery}_{tij} + \beta_7 \text{nursStaOcc}_{tij} + \beta_8 \text{gender}_{ij} \\
&+\beta_9 \text{campus}_{ij} + \beta_{10} \text{accommodation}_{ij} + \beta_{11} \text{admitType}_{ij} \\
&+\beta_1 \text{day}_t + u_{1j} \text{day}_t + \beta_{12} \text{age}_{ij} \text{day}_t + \beta_{13} \text{charlson}_{ij} \text{day}_t + \beta_{14} \text{icu}_{tij} \text{day}_t \\
&+\beta_{15} \text{surgery}_{tij} \text{day}_t + \beta_{16} \text{nursStaOcc}_{tij} \text{day}_t + \beta_{17} \text{gender}_{ij} \text{day}_t \\
&+\beta_{18} \text{campus}_{ij} \text{day}_t + \beta_{19} \text{accommodation}_{ij} \text{day}_t + \beta_{20} \text{admitType}_{ij} \text{day}_t \\
&+\beta_2 \text{laps}_t + u_{2j} \text{laps}_t + \beta_{21} \text{age}_{ij} \text{laps}_t + \beta_{22} \text{charlson}_{ij} \text{laps}_{tij} + \beta_{23} \text{icu}_{tij} \text{laps}_{tij} \\
&+\beta_{24} \text{surgery}_{tij} \text{laps}_{tij} + \beta_{25} \text{nursStaOcc}_{tij} \text{laps}_{tij} + \beta_{26} \text{gender}_{ij} \text{laps}_{tij} \\
&+\beta_{27} \text{campus}_{ij} \text{laps}_{tij} + \beta_{28} \text{accommodation}_{ij} \text{laps}_{tij} + \beta_{29} \text{admitType}_{ij} \text{laps}_{tij}
\end{aligned}
$$

$$
\text{where } (\varepsilon_{1ij}, \ldots, \varepsilon_{n_{ij}ij})' \sim \text{N}(\mathbf{0}, \boldsymbol{R}_{ij}), \text{ and } (u_{0ij}, u_{0j}, u_{1j}, u_{2j})' \sim \text{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_{ij}^2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{D} \end{bmatrix}\right)
$$

The above formulation is therefore of the full model with main effects, interactions, random intercepts, and random slopes. Formulating our model in this way is sometimes called "top-down", as we start with the largest model and then work to simplify it (as described by Verbeke and Molenberghs in [49], Section 9.2). This, however, can prove difficult, if not impossible, given problems of convergence and computational load for complex multilevel models, such as we had.

Instead of the above formulation, we chose to use a bottoms-up approach. We began with all main effects, then added random intercepts to account for correlation between groups (which we clearly have in a longitudinal study). Assuming the random intercepts were needed, we then determined whether or not random slopes would improve the model, and subsequently considered dropping covariates with low p-values (following a similar approach to that described by Twisk in Section 5.3 of [47] for predictive models). We will see this in the next section. Our goal, ultimately, was to approach the law of parsimony—to build the simplest model necessary for predicting discharge.

Of course, since the model is a form of GLMM, estimation of fixed and random effects follows the methods presented in Sections 5.1 and 5.2. However, it is difficult to estimate random variance for logistic multilevel models, which therefore explains all the different algorithms that exist. As noted by Snijders and Bosker in Section 14.2.5 of [45], the choice of algorithm to use for estimation should be based on stability. In our case, only `glmmPQL` {MASS package}, in R, would converge for the three-level model.

5.3.2 Analysis Steps

Since we used penalized quasi-likelihood to estimate our model, the method of `glmmPQL` {MASS package}, there was no likelihood function, and therefore we could not use likelihood ratios for inference, testing t, or comparing models. However, due to the asymptotic normality of the estimators, Wald statistics were used for inference (which are fine for large samples).

Following Verbeke and Molenberghs in [49], Section 9.3, we plotted the residuals of 44 ordinary linear models for main effects, by primary condition, to determine if random slopes were required. Residual plots with a linear trend (besides a flat line) suggest the need of random slopes for the time-varying covariates, which we only see in Figure 5.3 for primary conditions GYNEC1 and PRGNCY. Given that this only represents 2 of 44 primary conditions, and that random slopes for time (our day covariate) can cause convergence problems, we chose not to consider random slopes any further (greatly simplifying our model).

In Section 4.4.2 we found that 44 main-effects population-averaged models, one model for each primary condition, achieved a moderately better c-index than a single model encompassing all primary conditions (although the Brier score worsened). We therefore decided to try estimating 44 main-effects subject-specific models, one model for each primary condition. Therefore, our model for patient encounter i, with

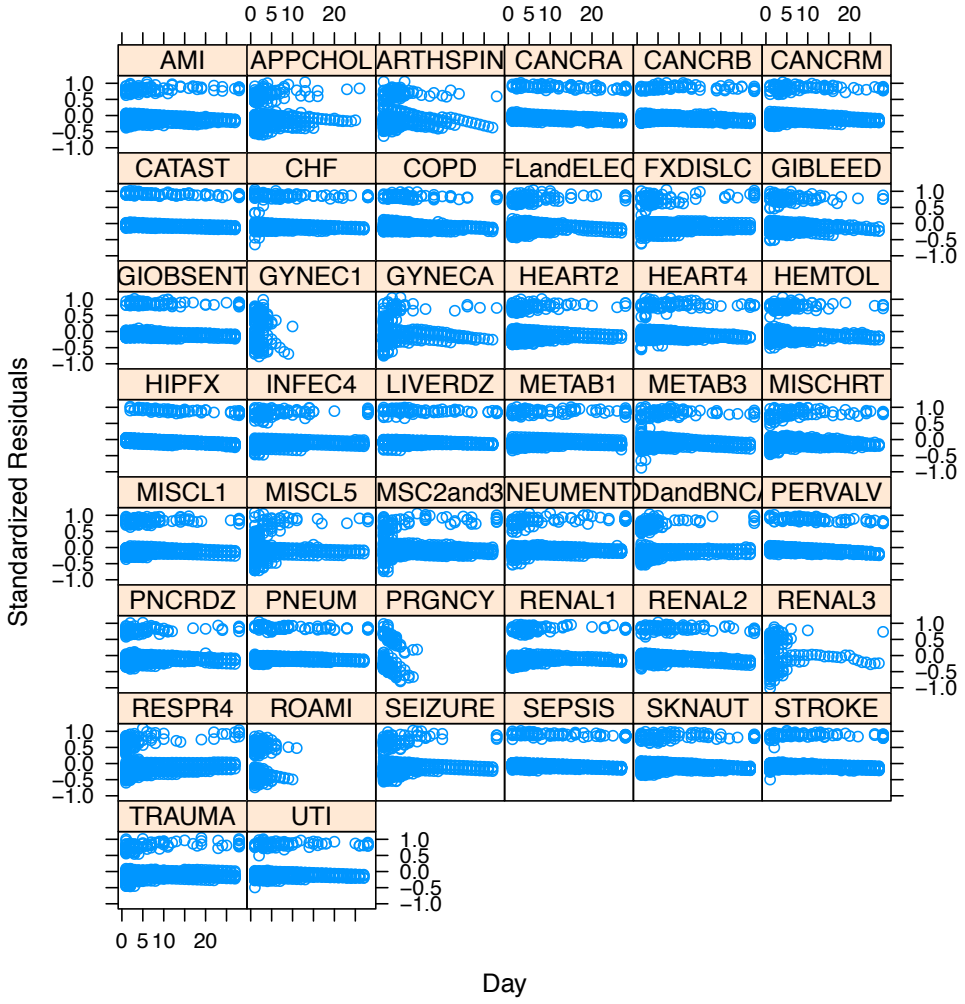

Figure 5.3: Scatter plot of residuals versus day, by primary condition, for 44 main-effects linear models, one for each primary condition. Linear trends would suggest the need for random slopes on time-varying covariates.

repeated measures for time t, and random intercept u_{0i}, is given by

$$\begin{aligned}\text{logit}(\mu_{ti}) &= \beta_0 + u_{0i} + \beta_1 \text{day}_t + \beta_2 \text{laps}_t + \beta_3 \text{age}_i + \beta_4 \text{charlson}_i \\ &+\beta_5 \text{icu}_{ti} + \beta_6 \text{surgery}_{ti} + \beta_7 \text{nursStaOcc}_{ti} + \beta_8 \text{gender}_i + \beta_9 \text{campus}_i \\ &+\beta_{10} \text{accommodation}_i + \beta_{11} \text{admitType}_i \end{aligned} \tag{5.3.1}$$

Two-level (patient encounters and repeated measures), main effects, random intercept models, by primary condition, achieved a mean c-index of 0.803 and a mean Brier score of 0.085. Compared to the population-averaged models which achieved a mean c-index of 0.717 and a mean Brier score of 0.113, our multilevel approach was a definite improvement.

For sake of comparison to the population-averaged models, and the subsequent three-level model we will consider next, we include residual plots of our two-level models in Figures 5.4 and 5.5. Figure 5.4 collects residuals into a single group for comparing with other models. The inclusion of random intercepts created a series of residuals, rather than one line of them (such as the residuals for the population-averaged models). The residuals were also more normally distributed than before. In Figure 5.5 we see that the two-level models achieved excellent predictive discrimination for a few primary conditions, such as AMI, whereby the predicted probability of discharge was almost exclusively 0 and 1.

Next we considered our three-level model, which adds only one estimate to the model—the variance of the random effects for primary conditions. Therefore, our model for patient encounter i, in primary condition j, with repeated measures for time t, and random intercepts for patient encounter u_{0ij} and primary condition u_{0j}, is given by

$$\begin{aligned}\text{logit}(\mu_{tij}) &= \beta_0 + u_{0ij} + u_{0j} + \beta_1 \text{day}_t + \beta_2 \text{laps}_t + \beta_3 \text{age}_{ij} + \beta_4 \text{charlson}_{ij} \\ &+\beta_5 \text{icu}_{tij} + \beta_6 \text{surgery}_{tij} + \beta_7 \text{nursStaOcc}_{tij} + \beta_8 \text{gender}_{ij} + \beta_9 \text{campus}_{ij}\end{aligned}$$

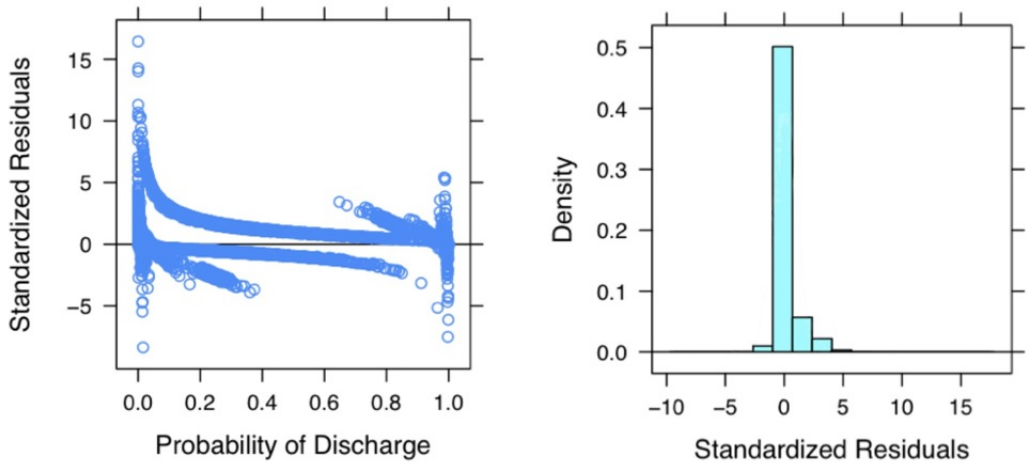

Figure 5.4: Scatter plot and histogram of residuals for 44 two-level main-effects subject-specific models, one model for each primary condition, described by equation (5.3.1). In order to compare with other models, we consider the residuals as a single group, rather than by primary condition. The multilevel model includes main effects, and random intercepts for patient encounters. The standardized residuals are not normal.

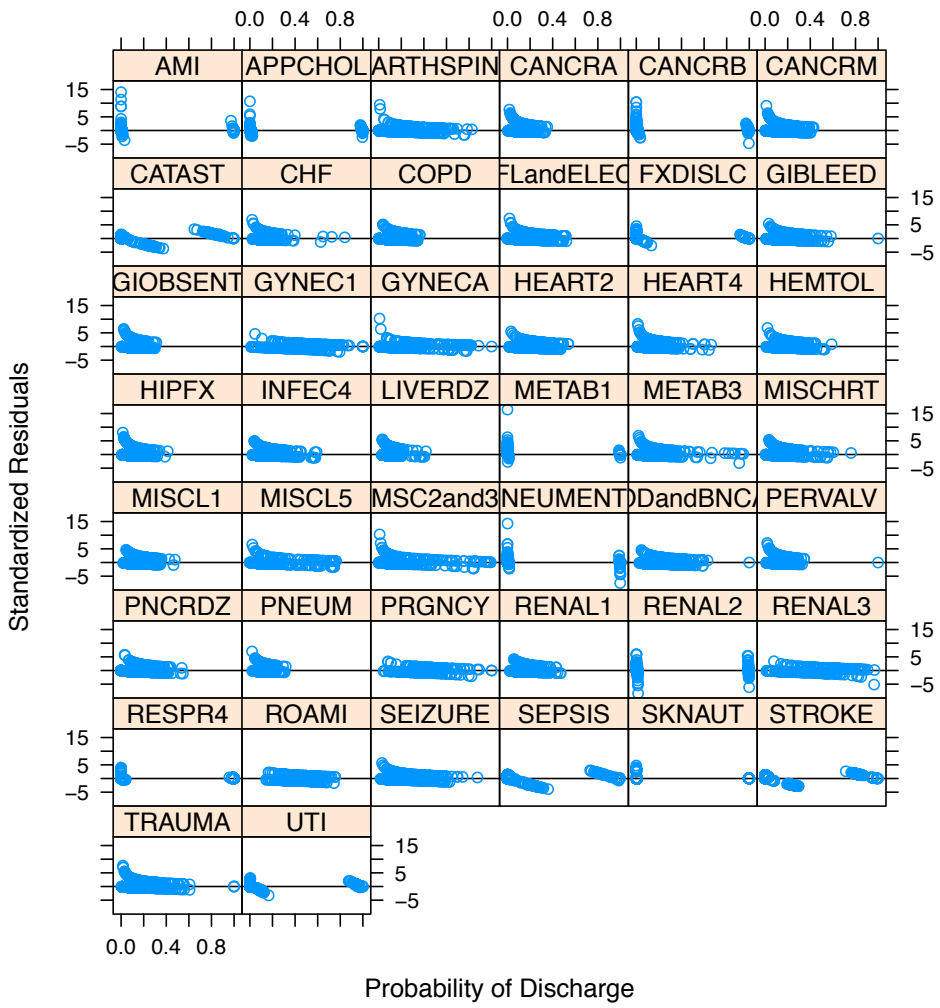

Figure 5.5: Scatter plots of residuals versus probability of discharge, by primary condition, for 44 two-level main-effects subject-specific models, one model for each primary condition, described by equation (5.3.1). The multilevel model includes main effects, and random intercepts for patient encounters. Each scatter plot is of one distinct model, since one model was fit to each primary condition.

$$+\beta_{10}\text{accommodation}_{ij} + \beta_{11}\text{admitType}_{ij} \tag{5.3.2}$$

This model achieved a c-index of 1.0, the highest possible, and Brier score of 0.00009. Table 5.1 summarizes the estimates, standard errors, and odds ratios for our three-level, main effects, random intercept model of patient discharge. Including random intercepts for primary conditions meant estimating a single parameter—the variance of the random effects—as opposed to estimating parameters for the primary conditions themselves (which would estimate each intercept separately). The same is true of the random intercepts for patient encounters.

The estimated standard deviation of the random intercept terms do not include a test of statistical significance in Table 5.1, because there is no legitimate test for variance parameters. However, Twisk suggests a rule of thumb in Section 4.2 of [47]—if the variance of the random effects is more than double its standard error, the variance should be considered important (i.e., assume that the random effects are represented by a sufficient number of independent random variables with finite mean and variance so that, by the central limit theorem, we can assume they follow a normal distribution). As can be seen in the table, the estimated standard deviations for primary conditions and for patient encounters were both far greater than double their respective standard errors. We can therefore conclude that both the random intercepts were important to our model, and should therefore be kept in the model. This also implied that we needed a three-level model of discharge rather than a two-level model.

We consider the residuals of our three-level model in Figures 5.6 and 5.7. The residuals decreased in magnitude, and, more importantly, demonstrated excellent predictive discrimination, with predicted probabilities of discharge at nearly 0 and 1 exclusively. Given that the model is able to set the random-intercept for each individual subject, so that it captures the transition from no discharge (0) to discharge (1), this is not entirely surprising. But it does raise concerns about over-fitting, which

Table 5.1: Estimates from penalized quasi-likelihood for three-level main-effects subject-specific model including all primary conditions, described by equation (5.3.2). The multilevel model includes main effects, and random intercepts for primary conditions and patient encounters. Refer to Table 1.2 for a description of covariates. C-index of 1.0, and Brier score of 0.00009.

Covariate	**Estimate**	**S.E.**	**P-Value**	**Odds Ratio (95% CI)**	
				Lower	**Upper**
(Intercept)	-22.58060	5.879031	0.0001	4.367e-13	5.580e-8
day	7.82945	0.001152	<0.0001	2.521e3	2.516e3
LAPS	-0.00145	0.000110	<0.0001	9.984e-1	9.987e-1
age	-0.47879	0.050903	<0.0001	5.888e-1	6.519e-1
charlson1-2	-14.13844	2.394972	<0.0001	6.601e-8	7.941e-6
charlson3-4	-28.56480	2.919981	<0.0001	2.120e-14	7.287e-12
charlson5+	-32.16305	2.988766	<0.0001	5.417e-16	2.137e-13
icu1	2.21571	0.019821	<0.0001	8.988	9.351
surgery1	-8.50347	0.019814	<0.0001	1.988e-4	2.068e-4
nursStaOcc	0.00778	0.000069	<0.0001	1.008	1.008
genderM	0.41481	1.766523	0.8144	2.588e-1	8.858
campusG	1.73975	2.185778	0.4261	6.401e-1	5.068e1
campusH	1.86866	4.515751	0.6790	7.086e-2	5.925e2
accommodationS	6.54282	2.197003	0.0029	7.716e1	6.246e3
accommodationW	7.78697	2.430670	0.0014	2.119e2	2.738e4
admitTypeE	-1.40996	3.625932	0.6974	6.500e-3	9.170
admitTypeL	15.14029	4.249582	0.0004	5.368e4	2.636e8
admitTypeS	21.44307	4.665760	<0.0001	1.933e7	2.182e11
admitTypeU	5.89657	4.884240	0.2274	2.752e0	4.809e4

Random Intercept	**Estimated S.D.**	**S.E.**	**Odds Ratio (95% CI)**	
			Lower	**Upper**
primaryCondition	16.4304	4.3482	1.767e5	1.057e9
patient	55.4003	1.1514	3.631e23	3.632e24

is why we consider validation on external data in Section 5.3.4, when we compare models against one another.

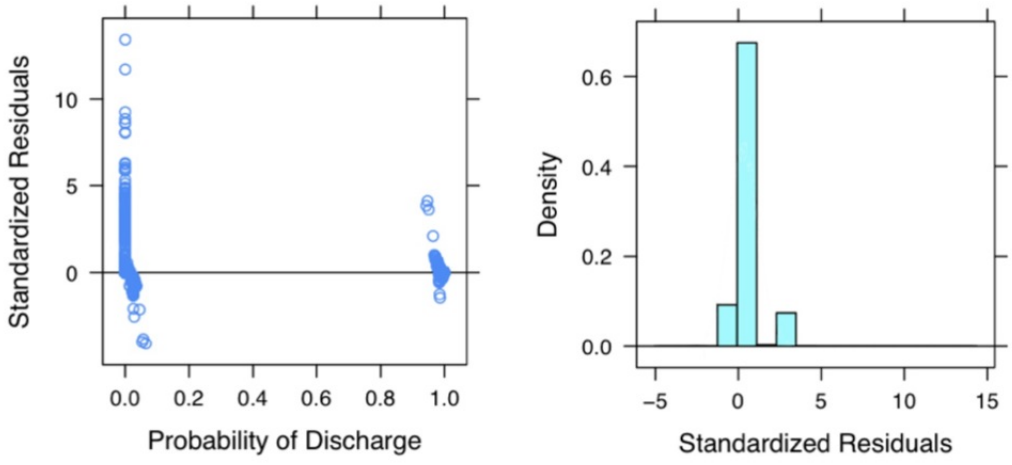

Figure 5.6: Scatter plot and histogram of residuals for three-level main-effects subject-specific model, described by equation (5.3.2). The multilevel model includes main effects, and random intercepts for primary conditions and patient encounters. The standardized residuals are not normal.

Another useful diagnostic tool, unique to mixed models, is plots testing normality of the random effects. Recall that we assume normality of the random effects in our model formulation of Section 5.3.1, and this assumption is used in estimating our predictors (the parameters of the random effects). This is a fundamental assumption, going back to the BLUP equations described in Sections 5.1 and 5.2, at the beginning of this chapter. We consider the normality of our random intercepts in Figures 5.8, and use q-q plots to assess and view possible deviations from normality.

The random intercepts for primary conditions of the three-level model were as close to being normally distributed as we could possibly expect, as shown in Figure 5.8. The random intercepts for patient encounters, however, showed some skeweness

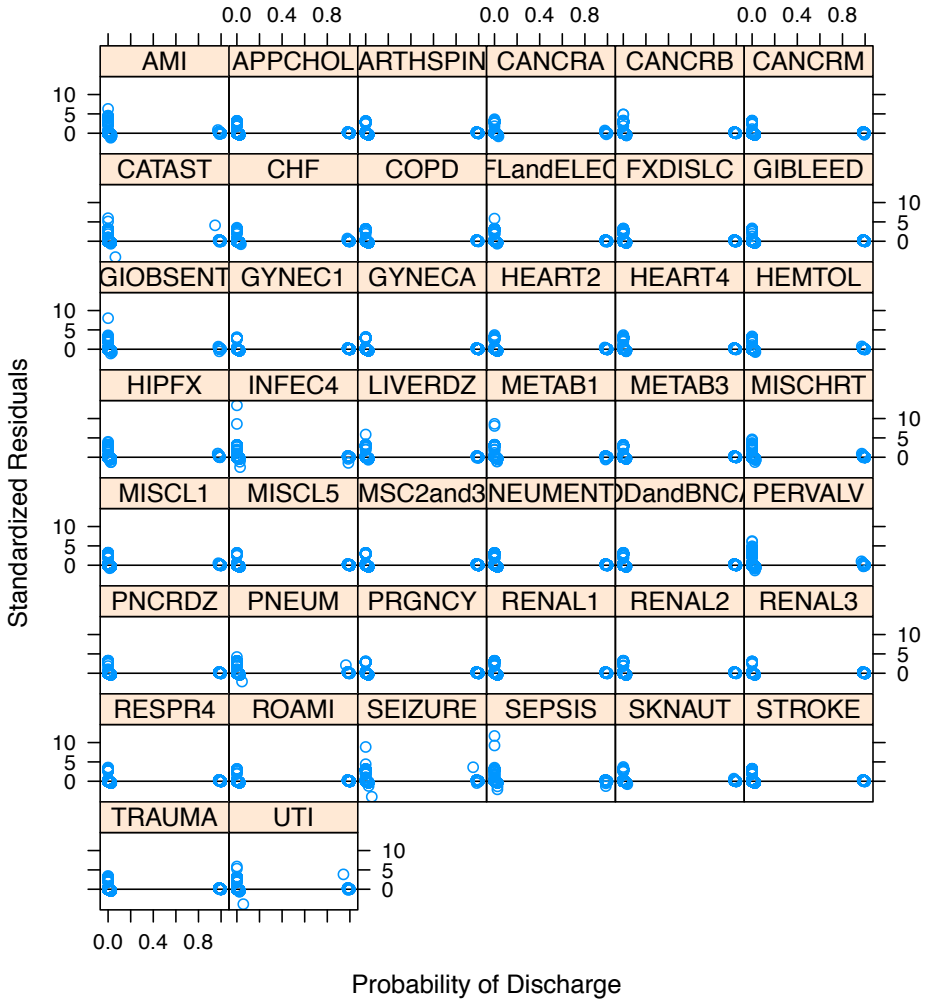

Figure 5.7: Scatter plots of residuals versus probability of discharge for three-level subject-specific model, by primary condition, described by equation (5.3.2). We divide the residuals by primary condition, even though we have a single model of discharge in this case, in order to compare with our other models, but also because primary conditions are a level in the formulation of our multilevel model. The multilevel model includes main effects, and random intercepts for primary conditions and patient encounters.

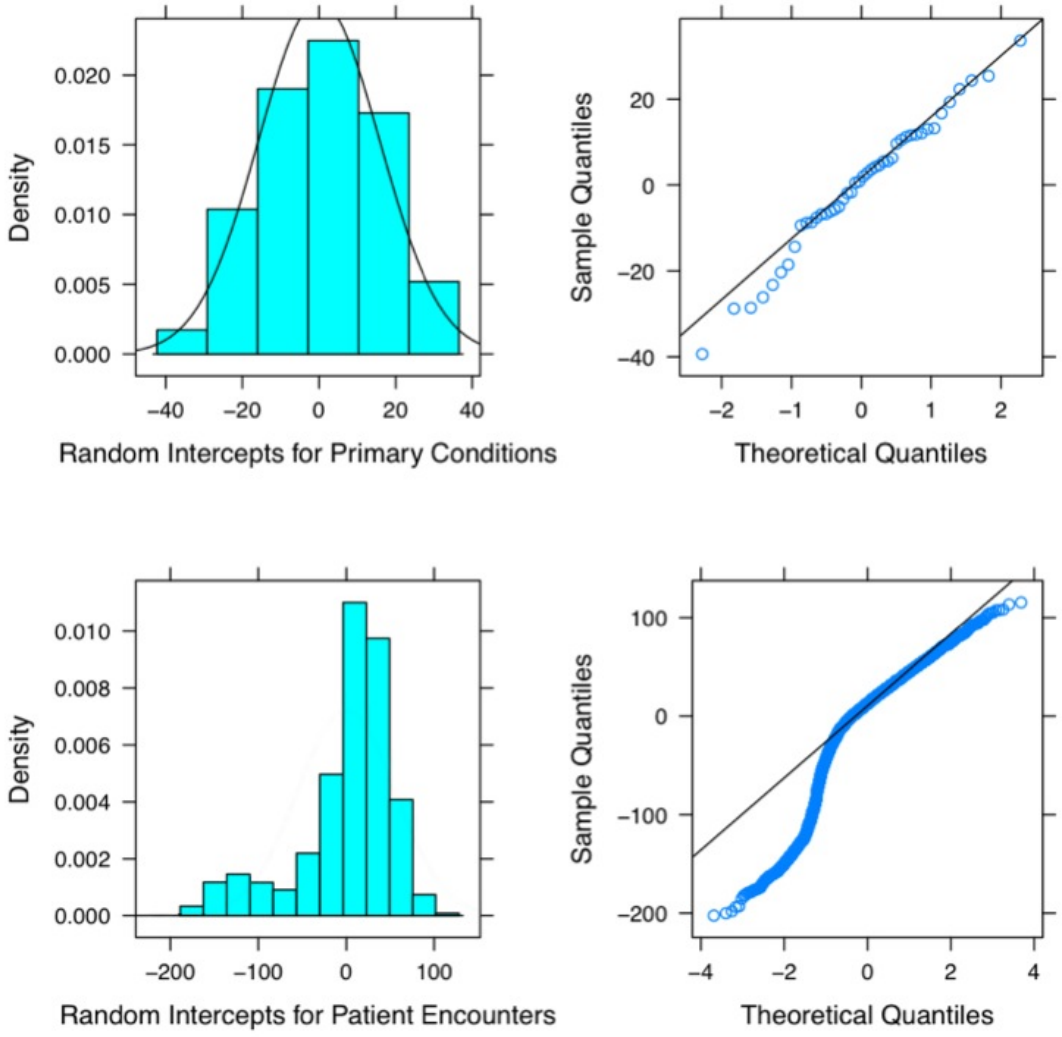

Figure 5.8: Histogram and q-q plot of random effects for three-level main-effects subject-specific model including all primary conditions, described by equation (5.3.2). Ideally random effects should be normally distributed, given our model formulation. The multilevel model includes main effects, and random intercepts for primary conditions and patient encounters.

(in fact they look to be bi-modal). This could be the same problem noted for the residuals of the population-averaged model. That is, the skeweness may be caused by overdispersion or missing covariates (such as interactions), such that the model is predicting discharge when there is none, which could affect the estimate of predictors. Or it could imply that there are two distinct entities in the model, perhaps differences based on one or more primary condition groups. But given a c-index of 1.0 we do not feel it necessary to consider it further. If we were concerned, however, hierarchical generalized linear models do not assume the distribution of random effects are necessarily normal, as described in by Lee and Nelder in [25] (which is outside the scope of this thesis).

Given the lack of statistical significance for the factor levels of the campus and gender covariates, we tried removing them from the model. The c-index and Brier score did not change, and estimates changed by only insignificant amounts. Also, the function `glmmPQL` {MASS package} returned an estimate of the correlation matrix for the covariates, which we include for completeness in Table 5.2. None of the correlations between covariates are very high, which would suggest that the main effects do not necessarily overlap in their contribution to the model. Covariate pairs that did correlate would need to be investigated further, as they could be combined into a single covariate to reduce the effects of multicolinearity (which could lead to estimation problems). However, correlation between covariate pairs for the full-model may mask important relationships when one considers them at the level of the primary condition groups.

Table 5.2: Correlation in percent between covariates in three-level subject-specific model, described by equation (5.3.2). We index the covariates (0 for intercept) to simplify cross-referencing of elements. Correlation between covariate pairs is not high, which would suggest they make distinct contributions to the model.

Covariate		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
intercept	0																		
day	1	0																	
laps	2	0	-6																
age	3	-48	0	0															
charlson1-2	4	-8	0	0	-17														
charlson3-4	5	-5	0	0	-20	41													
charlson5+	6	-5	0	0	-24	45	40												
icu1	7	-0	1	-4	0	0	0	0											
surgery1	8	-0	-8	4	0	0	0	0	18										
nursStaOcc	9	-0	-9	0	0	0	0	0	14	-9									
genderM	10	-15	0	0	5	-3	-4	-5	0	0	0								
campusG	11	-43	0	0	12	-3	-5	-8	0	0	0	0							
campusH	12	-16	0	0	-0	-2	2	-1	0	0	0	-3	27						
accommodationS	13	-25	0	0	-2	0	4	2	0	0	0	0	11	14					
accommodationW	14	-28	0	0	0	3	4	3	0	0	0	-4	25	23	61				
admitTypeE	15	-56	0	0	-3	2	9	1	0	0	0	0	19	2	-1	-3			
admitTypeL	16	-56	0	0	2	1	1	7	0	0	0	0	30	-6	0	-2	76		
admitTypeS	17	-38	0	0	-1	4	6	13	0	0	0	2	-5	-6	-7	-6	62	58	
admitTypeU	18	-45	0	0	0	0	0	4	0	0	0	-1	19	-22	0	-2	67	65	47

5.3.3 Interpreting the Model

We expect estimates for a subject-specific model with binary outcome to be greater than for a population-averaged model. We certainly see this when we compare Tables 5.1 and 4.1. The estimates for the three-level, subject-specific model are in fact quite high. However, the interpretation is very different from before. See the graphical example of a random-intercept logistic regression model with a single fixed effect (time) shown in Figure 5.9.

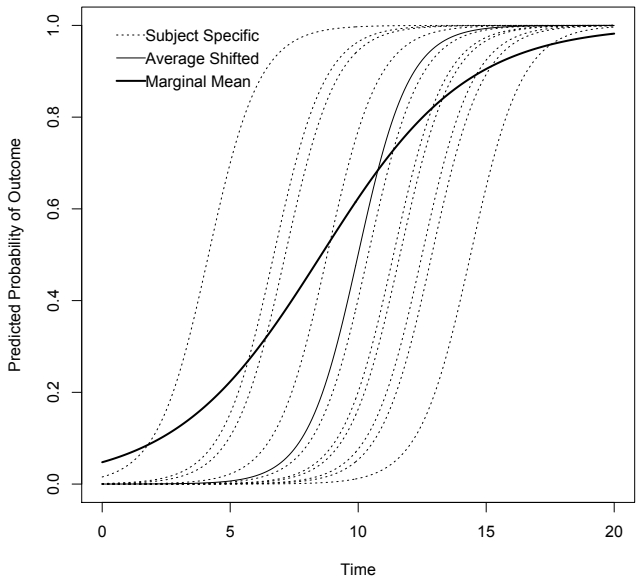

Figure 5.9: Example of logistic regression with random intercepts. The random-intercept responses are subject-specific (based on subject-specific clustering), and result in shifted probability curves because of the inverse link function (inverse logit). The conditional mean $\mathrm{E}(Y_{ti}|u_i = 0)$ (with mean random intercept) is the average-shifted curve, whereas the marginal mean $\mathrm{E}(Y_{ti})$ averages over the random effects resulting in a curve with an attenuated slope.

We follow an example given by Molenberghs and Verbeke in Section 16.3 of [33]. Consider a simple random-intercept model with a single fixed effect. The linear

predictor is $\text{logit}(P(Y_{ti})) = \beta_0 + u_i + \beta_1 t$, where β_0 is the fixed intercept, u_i is the random intercept for subject i, β_1 is the fixed slope, and t is the covariate for time. The conditional means are given by

$$\text{E}(Y_{ti}|u_i) = \frac{\exp(\beta_0 + u_i + \beta_1 t)}{1 + \exp(\beta_0 + u_i + \beta_1 t)}.$$

We average the conditional means over the random effects to get the marginal average

$$\begin{aligned} \text{E}(Y_{ti}) &= \text{E}\left(\text{E}(Y_{ti}|u_i)\right) \\ &= \text{E}\left(\frac{\exp(\beta_0 + u_i + \beta_1 t)}{1 + \exp(\beta_0 + u_i + \beta_1 t)}\right) \\ &\neq \frac{\exp(\beta_0 + \beta_1 t)}{1 + \exp(\beta_0 + \beta_1 t)} = \text{E}(Y_{ti}|u_i = 0). \end{aligned}$$

The marginal time trend will be much less steep for a random-intercept model than for each subject-specific time trend. The greater the variability between subjects, or variance of random effects, the more pronounced will be the differences between the estimates of the population-averaged model and the subject-specific model. In general, there is no simple relation between the parameters in either model type, except for special cases.

Intraclass Correlation

In general, the degree of likeness between elements in the same group can be expressed by the intraclass correlation (as originally described in Section 2.2.2). It is typically described as the proportion of total variance attributable to the group, as summarized in the equation

$$\rho = \sigma_b^2/(\sigma_b^2 + \sigma_w^2),$$

where ρ is the intraclass correlation, σ_b^2 is the between-group variance, and σ_w^2 is the within-group variance.

This interpretation is more applicable to continuous variables than dichotomous ones. For a continuous variable, we can easily determine the variance at the subject

level (e.g., a subject's blood pressure), as well as for a group (e.g., blood pressure for subjects with the same primary condition). However, for a dichotomous variable, the subject-level variance needs to be determined using prevalence for the group, confounding the subject and group-level variances.

Although there is some confusion over the correct interpretation of the intraclass correlation for a dichotomous variable (a concern raised by Twisk in Section 4.2.1 of [47]), it is still a useful measure of the relative weight of variances at each level of a multilevel model. To approximate the intraclass correlation, we can assume there is a continuous latent variable (i.e., an unobservable variable whose existence we infer) underlying the binary outcome. In our model, the latent variable represents the patient's "readiness" of discharge. Therefore, as noted by Snijders and Bosker in Section 14.3.3 of [45], we can use the variance of the latent variable to calculate the intraclass correlation, which is the variance of a logistic distribution ($\sigma^2 = \pi^2/3$).

For our three-level logistic model of discharge, we have three intraclass correlation coefficients, which we can approximate as follows (where the subscript 2 represents the patient-encounter level, and 3 the primary-condition level—the variance for level 1, repeated measures, has been replaced by the variance of its latent variable, $\pi^2/3$, as described in the preceding paragraph):

$$\begin{aligned} \rho_3 &= \sigma_3^2/(\sigma_2^2 + \sigma_3^2 + \pi^2/3) \\ \rho_{2,3} &= (\sigma_2^2 + \sigma_3^2)/(\sigma_2^2 + \sigma_3^2 + \pi^2/3) \\ \rho_2 &= \sigma_3^2/(\sigma_2^2 + \sigma_3^2) \end{aligned}$$

From the estimates in Table 5.1, we find that $\rho_3 = 0.08$ (8% of total variance was at the level of the primary conditions), $\rho_{2,3} = 0.99$ (99% of total variance was at the level of patient encounters and primary conditions), and $\rho_2 = 0.08$ (8% of total variance was at the level of patient encounters). In other words, we can (loosely) interpret these results to mean that the likeness of repeated measures in the same primary conditions was 8%, the likeness of repeated measures in the same patient

encounters and the same primary conditions was 99%, and the likeness of patient encounters in the same primary conditions was 8% (since the latter is much less than 50%, we can say that the primary-condition level contributed much less to variability than the patient-encounter level).

5.3.4 Comparing Models

We finish by comparing our subject-specific models of this chapter to the population-averaged models of the previous one. Consider the ROC plot, the true positive rate versus false positive rate, in Figure 5.10 (created using the ROCR package in R, from [44]). Although we have cited the c-index when evaluating our models, it is nonetheless important to inspect the ROC plot to ensure there is nothing out of the ordinary. It is also a useful plot for comparing models, as in our case. From this plot we see that we have progressively improved on our model of patient discharge.

Accuracy for binary classifiers is the number of true positives and true negatives, divided by the number of true and false positives and negatives. We include a plot of accuracy versus cutoff in Figure 5.11. Such a plot can be used to determine an appropriate cutoff between positive and negative outcomes in a binary classifier. The high accuracy of the three-level model speaks to its high level of predictive discrimination.

A true test of predictive accuracy, however, needs to consider out-of-sample data. In others words, we needed to apply the models to data on which they had not been fit. Although the population-averaged models we considered can easily be applied to a new sample of discharge data, the subject-specific models require that new estimates be calculated so that the random-intercepts can be predicted for the specific patient encounters under consideration. This complicates matters significantly, and there is unfortunately no standard cross-validation approach for such models. Researchers therefore create cross-validation methods they deem appropriate for their particular

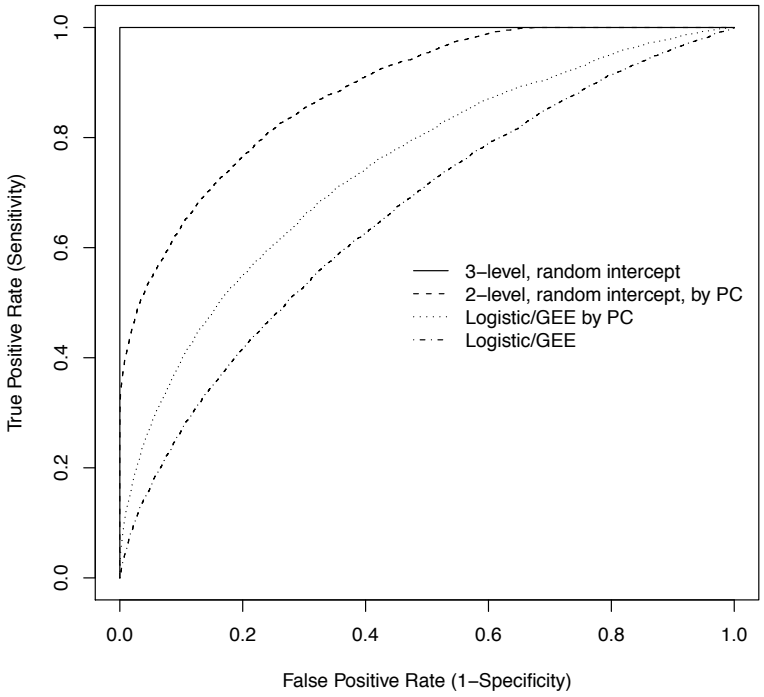

Figure 5.10: ROC plot comparing different models of patient discharge. Area under curve, or c-index: 0.685 for Logistic/GEE; 0.717 for Logistic/GEE by PC; 0.803 for 2-level, random intercept, by PC; 1.0 for 3-level, random intercept.

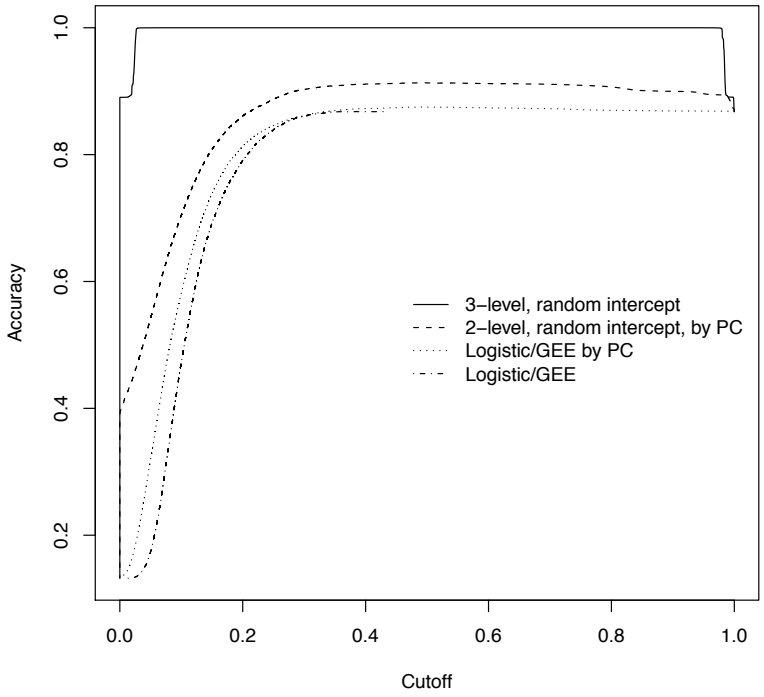

Figure 5.11: Accuracy plot comparing different models of patient discharge. Accuracy is true positive outcomes and true negatives outcomes divided by the total of all positive and negative outcomes.

application.

We tested the models by considering next-day forecasts on a new dataset, one day at a time. We incorporated the previous day into the data on which the models were fit so that the random intercepts at the patient level could be predicted. For each day we calculated the c-index and Brier score. We defined day 0 as the day on which the new patients were admitted to hospital, and therefore their data was not available for fitting the models. We then used the models to predict discharge on day 1. On day 1, however, we included patient data from day 0, the previous day, into the original dataset, and re-fit the models in order to predict discharge on day 2. And so on for 7 days, at which point many patients had been discharged and our measures of predictive accuracy would begin to suffer from a lack of data.

We constructed an out-of-sample dataset by randomly selecting 100 new patient encounters per primary condition. We then predicted discharge, and refit the models, by day, as described in the preceding paragraph. In doing so, however, we found that the models constructed for one primary condition only, and therefore resulting in sets of 44 models, were unable to perform adequately. Because they had been fit on a random sample of 100 patient encounters, they would often have missing factor levels. However, when we attempted to predict discharge for new patient encounters, these factor levels would at times be present. For example, if the "elective" factor level of the admitType covariate was missing when fitting a model for a particular primary condition, but then found to be present for new patient encounters, the prediction would need to exclude this covariate. The results were so poor, often worse than chance (a c-index less than 0.5), that we decided to drop this approach from further consideration.

We therefore compared forecast accuracy for the population-averaged model (logistic with GEE), against the three-level subject-specific model (random intercept for patient and primary condition) in Figures 5.12 and 5.13. However note that we also included forecast accuracy for the three-level subject-specific model with random

intercept for primary condition only (i.e., ignoring the random intercept for patient). Given that the prediction of discharge on day 1 (next day) is based on data that does not include day 0, there was no random intercept for patient to include. For the following days we did predict the random intercept for each patient, but wanted to see its effect on predictive accuracy.

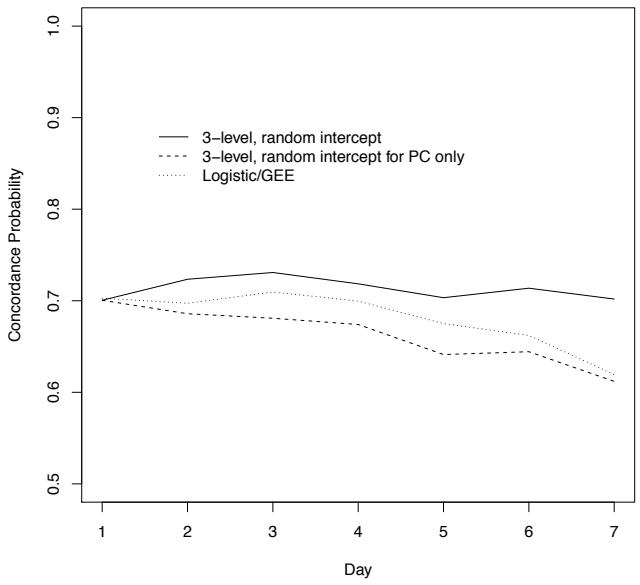

Figure 5.12: Predictive accuracy of forecasts using c-index by day. We forecast on out-of-sample data, but include patient data from the previous day so that the subject-specific model can predict the random intercept.

It is clear from Figure 5.12 that the subject-specific model was able to maintain an acceptable c-index by adjusting the patient-level random intercept accordingly (although not the 0.8 or more needed to achieve "some utility" as a predictive model, as stated by Harrell in Section 10.8 of [20]). The c-index for the population-averaged model, however, got progressively worse the longer a patient was in hospital. The Brier score seemed to arrive at the opposite conclusion. Except that the predicted probabilities from the population-averaged model never exceeded 0.6, the majority of

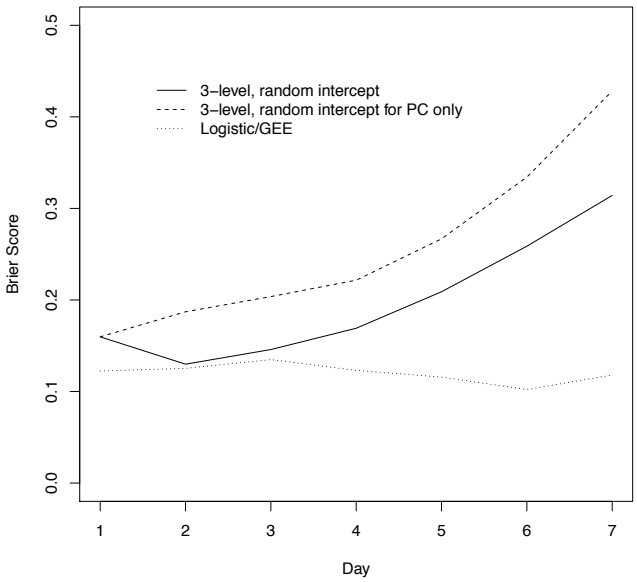

Figure 5.13: Predictive accuracy of forecasts using Brier score by day. The predicted probabilities of the population-averaged model remains relatively low for each day; the predicted probabilities of the subject-specific model increase with each day. This difference accounts for the difference in Brier scores.

which kept well below even 0.4, and remained relatively low throughout the days.

The predicted probabilities of the subject-specific model, on the other hand, ranged from 0 to 1, and the majority gradually increased over time. Intuitively this is what we expected—the longer a patient is in hospital, the greater the probability that they will soon be discharged. This is also not surprising given that the coefficient estimates were much higher for the subject-specific model than the population-averaged model. Accuracy plots by day also showed that the subject-specific model would performed better with a higher probability threshold used to distinguish between positive and negative outcomes.

Given the predictive accuracy by day of Figure 5.12, and how the probability of discharge increased by day, we feel the three-level, main effects, random intercept model performed best. Although the model used sophisticated estimation routines, building on the theory of generalized linear models, it required fewer estimates than the equivalent generalized linear model (i.e., estimating the variance of a random effect for a cluster variable instead of factor levels). If the complexity of a model is judged by the number of parameters to estimate, then multilevel models are an improvement towards the law of parsimony.

We recommend that interactions between covariates be studied further in an attempt to improve the subject-specific model's predictive performance (and achieve the desired c-index of 0.8 or above). Normally in smaller longitudinal studies time-dependent covariates are included in interactions with time-independent covariates. Other interactions between covariates may also improve model performance for individual primary condition groups, raising the overall level of performance.

In considering interactions, the residuals plots of the subject-specific model are not informative, due to the almost perfect predictive discrimination of the model for all primary condition groups. The residuals of the population-averaged models, however, did show that certain primary condition groups fit the data significantly better than others. The primary condition groups that do not fit the data as well

would likely be good candidates for further study. One could also look at scatter plots of important interactions by day (up to the first seven days, for example) to determine if relationships exist through time.

Including such interactions was not possible in our case as the much older 32-bit desktop computers on which we tested these models would run out of memory. Nonetheless, the main-effects subject-specific model was able to achieve about 70% concordance, and could be a valuable tool in the day-to-day operations of the hospital.

Chapter 6

Conclusions

Hospital administrators need to know how many patients will be coming and going on a daily basis so that they can schedule surgeries and assign resources. Admissions to the emergency department represent the main driver to unscheduled hospital admissions. Patients are often being cared for in one ward, but seeking treatment in another, due to high occupancy levels. Others are forced to wait in the emergency department until a bed is available in a ward. Knowing in advance the patient load expect in the emergency department would help assign resources accordingly. And knowing in advance which patients, in which wards, are more likely to be discharged would help plan their exit and free up beds for incoming patients.

Given the lack of appropriate covariates (continuous or categorized) available from the emergency department, and the relatively short amount of time before an emergency patient is assigned to a bed, we chose to model admissions using a time-series model. Specifically, we fit Seasonal Auto Regressive Integrated Moving Average (SARIMA) models for the civic and general campuses of the hospital. Our results showed about a 30% improvement on the one-step random walk. If we assume an administrator has no analytical tools at their disposal, our results showed about a 30% improvement on predicting admissions compared to using the previous day as the

predicted outcome (known as the one-step random walk). We therefore suggest that SARIMA models would improve on the accuracy of predicting emergency admissions to the hospital. Also, the prediction intervals would also allow administrators to quantify uncertainty, which could be incorporated into hospital procedure regarding spare capacity in the face of shocks (spikes in emergency admissions). To study the problem and potential solutions further, we could investigate the literature on integer-based time-series models, and consider modeling hourly demands in the emergency department.

Our next challenge was to predict patient discharge, but here we had many more useful covariates at our disposal. Our primary dataset for building models of discharge encompassed longitudinal patient data. We used a sample of 100 patient encounters per primary condition, wherein observations were taken each day, for each patient, for their last 28 days before discharge. Therefore each longitudinal series of patient data included their positive outcome, indicating discharge. A sample of the last 28 days of patient data removed the effect of outliers (i.e., patient encounters with long lengths of stay), whereas 95% of patients are discharged within 28 days.

We used the c-index, a unitless measure proportional to rank correlation, and the Brier score, the mean squared error between predicted probability and the binary outcome, to assess predictive accuracy of the discharge models. We considered population-averaged models, using logistic regression with a correction for the dependency between patient observations using generalized estimating equations (wherein the random effects between patient encounters is treated as a nuisance). We also considered subject-specific models, using logistic regression and predicting the random effects between patient encounters, and the random effects between primary conditions.

We had slightly better overall results with the population-averaged approach when we fit individual models by primary condition group. However, because each model was fit on smaller samples than the whole (100 patient encounters per primary

condition), many factors would include only small sample sizes, or be missing entirely from covariates. For example, the ICU indicator was missing from some primary conditions, leading to missed opportunities to better predict patient discharge. Doubling the sample size to 200 patient encounters per primary condition was still not enough to rectify this particular problem.

We achieved the best predictive accuracy with a three-level subject-specific model of main effects—a patient-specific random-intercept, and primary-condition-specific random-intercept, logistic regression model. The resulting model comprised a random intercept for patient clustered in a random intercept for primary condition, as well as the following statistically significant covariates: day, LAPS, age, charlson, icu, surgery, nursStaOcc, accommodation, and admitType. We found the in-sample predictive accuracy to be very high, perhaps even misleading. The random-intercepts, in this case, are fit so well that they capture the transition from no-discharge to discharge. Since no standard cross-validation methods exist for such models, we created one that suited our application. Namely, we compared predictive accuracy for next-day discharge over multiple consecutive days, each time re-fitting the model on the days prior.

Based on an out-of-sample dataset of 100 new patient encounters per primary condition, the subject-specific model was able to maintain an acceptable c-index of about 0.7 for each day, compared to a population-averaged model that steadily decreased to about 0.6 by the 7th day. Furthermore, we found that the probability of discharge progressively increased each day using the subject-specific model (from 0 to 1), as one would expect, whereas the population-averaged model did not (almost staying constant, and never increasing above 0.6). Specifying a threshold on the probability of discharge, at the optimum for each model, in fact showed that the subject-specific model would produce better predictions compared to the population-averaged model.

We believe the three-level subject-specific model could be improved upon by

including interactions in the model, thus raising the c-index above 0.8 in the cross-validation study. We had no problems running our main-effects models on an older desktop computer, but including interactions with time-dependent covariates (such as day) led to memory overflows and convergence failures (for the population-averaged models as well). Although the hospital deploys analytics using servers, these production devices were not available for our research project. However, interactions with time-independent covariates (such as age) could possibly be investigated further without access to a server if a suitable computer were made available (with a sufficient amount of memory and computing power).

Although scatter plots did not suggest the need for interactions between some covariates for the overall model, there were some primary condition groups that could have benefited from them (and thus raise the overall level of predictive accuracy). One could consider, for example, a cubic splines for age, which is commonly done in healthcare models, and interactions with other covariates. And scatter plots through time could be used in an attempt to find time-dependent relatiohships. We could also consider hierarchical generalized linear models, which do not assume normally distributed random effects, in an effort to improve the discharge model. Another area of interest would be methods to predict discharge for new patients in a cluster (i.e., predicting the random intercept for a new patient encounter, as opposed to relying solely on the random intercept for the primary condition the new patient belongs to).

In order to merge the discharge model more directly to a model of admissions, we believe relevant covariates would need to be collected when patients present themselves to the emergency department. This alone requires much further investigation, in terms of what type of relevant data could be collected, in a timely manner. In particular, a model is needed that would not only predict admissions, but admissions to specific primary condition groups. Relevant data could include the presenting complaint, symptoms as soon as they are identified, test results, and initial diagnoses. An admissions model that could place a patient into a primary condition group, perhaps

even predict many of the covariate values used in the discharge model, would be a powerful tool for administrators in planning resource usage.

We believe that the models presented, predicting overall admissions and subject-specific discharge, could be included in a decision support tool to improve resource allocation at the hospital. There is a push in healthcare to reduce surgical wait-times, and introduce more information technology and analytics solutions. Only with the introduction of such tools will more relevant data be gathered, and will researchers learn more about the specific needs of hospitals. Ours is but a first step toward these goals.

Appendix A—Primary Conditions

When a patient is admitted to a hospital, a medical doctor must assign an initial diagnosis. Such a diagnosis is then fit within the framework of the International Classification of Diseases (ICD) codes, which include over 16,000 admission codes. Primary conditions are groups of 44 broad diagnostic categories, used for the sake of analysis. The groupings are based on the relative similarity of the diagnoses from the perspective of the disease, and on observed mortality rates (see Escobar et al. in [15] for details).

Table A.1: Description of primary conditions.

Primary Condition	**Description**
AMI	Myocardial infarction
APPCHOL	Appendicitis, hernias, cholecystitis, and cholangitis
ARTHSPIN	Arthropathies and spine disorders (but no infections or autoimmune conditions), except pathologic fracture
CANCRA	Malignant neoplasms of respiratory tract and intrathoracic organs; leukemias, non-Hodgkins lymphomas, and other histiocytic malignancies
CANCRB	All other malignant neoplasms not in 202-208 or gynecologic ones (including Hodgkins disease); radiation therapy and chemotherapy encounters where CA not specified
CANCRM	Ovarian cancer and metastatic cancer

continued on next page

Table A.1: Description of primary conditions—*continued*.

Primary Condition	Description
CATAST	Catastrophic conditions: dissecting aneurysms, cardiac arrest, respiratory arrest, all shock except septic; intracranial and subdural hemorrhages
CHF	Congestive heart failure and some related illnesses
COPD	Chronic obstructive pulmonary disease and some less common respiratory conditions
FLandELEC	Typical fluid and electrolyte disorders and dehydration
FXDISLC	All other fractures and dislocations, including pathologic fractures
GIBLEED	Gastro-intestinal hemorrhage; miscellaneous disorders of stomach and duodenum; diverticulitis; abdominal symptoms, nausea with vomiting; blood in stool
GIOBSENT	Inflammatory bowel disease and malabsorption; gastrointestinal obstruction; enteritides
GYNEC1	Non-malignant, non-infectious gynecologic diseases, including benign neoplasms, and miscellaneous (must be female patient)
GYNECA	Gynecologic malignancies other than ovarian cancer; female breast cancer (must be female patient)
HEART2	Diseases of pulmonary circulation and cardiac dysrhythmias, and miscellaneous
HEART4	Atherosclerosis (including that affecting precerebral arteries) and other forms of peripheral vascular disease
HEMTOL	Hematologic problems other than malignancies, and miscellaneous
HIPFX	Hip fracture
INFEC4	All other infections with the exception of hepatitis; unspecified fever, multiple others, including joint infections and muscle infections
LIVERDZ	Liver disorders, including hepatitis
METAB1	Diabetic ketoacidosis, with and without coma; hypoglycemic coma; unspecified coma and alteration of consciousness
METAB3	All other endocrine, metabolic and miscellaneous immune disorders (but not including SLE or RA)

continued on next page

Table A.1: Description of primary conditions—*continued*.

Primary Condition	Description
MISCHRT	Miscellaneous cardiac conditions and congenital heart disease
MISCL1	Miscellaneous conditions not classified previously
MISCL5	Miscellaneous non-cardiac congenital anomalies; miscellaneous symptoms other than fever; miscellaneous tooth and tongue disorders
MSC2and3	Remaining V codes; all E codes
NEUMENT	All other neurologic problems and mental disorders (other than drug overdoses); senility
ODandBNCA	Non-gynecologic benign; drug overdoses, drug abuse, adverse drug reactions, and poisonings neoplasms
PERVALV	Pericarditis and valvular heart disease
PNCRDZ	Pancreatic disorders
PNEUM	All forms of pneumonia; empyema; pleurisy; lung abscess; also pulmonary TB; pulmonary congestion and hypostasis
PRGNCY	Pregnancy and related conditions (must be female patient)
RENAL1	Acute renal failure, nephritic syndrome, and related conditions
RENAL2	Chronic renal failure, end stage renal disease, and kidney transplants
RENAL3	All other renal diseases other than infections, and miscellaneous
RESPR4	Acute respiratory infections and miscellaneous respiratory diseases
ROAMI	Chest pain, myocardial infarction not specified
SEIZURE	Seizure disorders, and miscellaneous
SEPSIS	Sepsis, meningitis, septic shock, and major infections
SKNAUT	Systemic lupus erythematosus, rheumatoid arthritis, skin disorders, related autoimmune diseases, and sialoadenitis
STROKE	Stroke and post-stroke complications
TRAUMA	Traumatic injuries not included elsewhere, including head injuries without intracranial or subdural bleeds
UTI	Urinary tract infections, not including pregnancy related ones

Appendix B—Source Code

B.1 Dataset and Variables

B.1.1 Creating Covariates in SAS

```
libname datain "T:\DW Research\DW Projects\larbuckle\datain";

* link laps to census (both daily obs);
proc sort data=datain.censusTransformed out=censusSorted
      (rename=(ichhospSvc=hospSvc));
   by encwid date;
   run;
proc sort data=datain.dailyLapsForCohort out=lapsSorted;
   by encwid date; *labelled hraencwid;
   run;
data censusLaps;
   merge censusSorted lapsSorted;
   by encwid date;
   drop ichattendingprowid admDtm dschDtm;
      *labelled hraadmdtm hradschdtm;
   run;
*no missing entries after merge;

* link occupancy (by nursing station) to census and laps;
proc sort data=censusLaps out=censusLapsSorted
      (rename=(ichnurssta=nurssta encCampusCd=campus));
   by date encCampusCd ichnurssta;
   run;
proc sort data=datain.nursStaOccupancy out=occupancySorted
      (rename=(hospOccupancy=nursStaOcc));
```

```
   by date campus nurssta;
   run;
data datain.censusLapsOccupancy;
   merge censusLapsSorted (in=here) occupancySorted;
   by date campus nurssta;
   if here=1 then output;
   run;
proc sort data=datain.censusLapsOccupancy;
   by encwid date;
   run;

* create covars from inpatient data (before creating daily obs);
data inpatient_tmp;
   set datain.inpatientCohort25feb2010 (rename=(charlIndex=chrlsn)
      drop=encHospSvcCd where=((primaryCondition ne '*ERROR*')
      and (primaryCondition ne '')));
   run;
data inpatientCovars;
   set inpatient_tmp (drop=charl: hraadmdtm hradschdtm encstartdtm
      encenddtm encAttendingProWid encexpectedlos);

   *group primary condition code;

   if hraAdmCatCd = 'U' then urgentAdm = 1;   *62% of encs;
   else urgentAdm = -1;

   if encMaritalSts in ('M','C') then married = 1;   *64% of encs;
   else married = -1;

   if encpatreligioncd in ('REF','U','NIL') then religion = -1;
   else religion = 1;   *38% of encs;

   if encRfpCd = 'H' then ohip = 1;   *92% of encs;
   else ohip = -1;

   if encOccupationCd = '5' then retired = 1;   *13% of encs;
   else retired = -1;

   if encOccupationCd = '98' then occUnknown = 1;   *8% of encs;
   else occUnknown = -1;
```

```
   if encAdmRoute = 'RG' then regularAdm = 1;   *87% of encs;
   else regularAdm = -1;

   if encTransferFromInstitutionCd = '' then transferIn = -1;
   else transferIn = 1;   *27% of encs;

   if encAmbulanceInd = 'Y' then ambulance = 1;  *29% of encs;
   else ambulance = -1;

   if encAccidentInd = 'Y' then accident = 1;  *< 3% of encs;
   else accident = -1;
   run;

proc sort data=inpatientCovars out=covarsSorted
      (drop=hraentrycd hraentrycd_ hraAdmCatCd hraAdmCatCd_
         encCampusCd encMaritalSts encpatreligioncd encRfpCd
         encOccupationCd encAdmRoute encTransferFromInstitutionCd
         encAmbulanceInd encAccidentInd encNursSta
       rename=(encAdmPriority=admitPriority encAdmType=admitType
          encPatAgeAtAdm=age encPatGenderCd=gender));
   by encwid;
   attrib _all_ label='';  *remove labels;
   run;
data dailyInpatientStatus_tmp;
   merge datain.censusLapsOccupancy (in=inhere) covarsSorted;
   by encwid;
   if inhere then output;
   attrib _all_ label='';  *remove labels;
   run;
*no missing obs;

proc sort data=dailyInpatientStatus_tmp;
   by encwid date;
   run;
data datain.dailyInpatientStatus;
   set dailyInpatientStatus_tmp
      (where=((age ge 15) and (hraExitCd ne '12')
         and (nursSta ne '')));
      *hraExitCd 12 dne, don't need patient status on exit;
   by encwid;
   if last.encwid then do; *sorted by encwid date;
```

```
      exit24hrs = 1;
      select;
         when (hraExitCd in ('01','02','03'))
            exit24hrsHow = 'transfered';
         when (hraExitCd in ('04','05','06'))
            exit24hrsHow = 'discharged';
         when (hraExitCd in ('07'))
            exit24hrsHow = 'deceased';
         end;
      end;
   else do;
      exit24hrs = 0;
      exit24hrsHow = 'na';
   end;

   if hospSvc in ('ACS','CSG','GSA','GSB','GSC','GSD','GSU','NRS',
         'PLA')
      then surgery = 1;
   else surgery = -1;

   if hospSvc eq 'ICU' then icu = 1;  *25% of deaths;
   else icu = -1;

   if surgery eq 1 then riskgroup = 'surgery';
   else do;
      if icu eq 1 then riskgroup = 'icu';
      else riskgroup = 'none';
      end;

   drop hraExitCd hraExitCd_;
   rename daysAfterAdm=los chrlsn=charlson
      encAccommodationGiven=accommodation;
run;
*no missing obs;

proc freq data=datain.dailyinpatientstatus
      (where=(exit24hrsHow eq 'transfered'
         or exit24hrsHow eq 'deceased'
         or exit24hrsHow eq 'discharged'));
   tables ichhospsvc*exit24hrsHow;
   run;
```

```
proc freq data=datain.dailyInpatientStatus;
   tables exit24hrs exit24hrsHow;
   run;

*create flat file of encwids from dailyInpatientStatus;
proc freq data=datain.dailyInpatientStatus noprint;
   tables encwid / out=datain.inpatientEncwids;
   run;

* subset for last 28 days of encounter, to remove effect of outliers
  (patients with long LOS), but capture status nearing discharge
;
proc freq data=datain.dailyInpatientStatus noprint order=freq;
   tables encwid / out=encwiddays;
   run;
proc freq data=encwiddays;
   tables count;
   run;
*74% of encounters have los of one week or less;
*87% of encounters have los of two weeks or less;
*92% of encounters have los of three weeks or less;
*95% of encounters have los of four weeks or less;

*add days until exit;
proc sort data=datain.dailyInpatientStatus;
   by encwid descending los;
   run;
data datain.dailyInpatientStatusDte;
   set datain.dailyInpatientStatus;
   by encwid;

   if first.encwid then daysUntilExit=1;
   output;
   daysUntilExit = daysUntilExit + 1;
   retain daysUntilExit;
   run;

*create dataset with 28-day look-back;
data dailyInpatientStatus28days;
   set datain.dailyInpatientStatusDte (where=(daysUntilExit le 28));
```

```
   run;

*check that it worked;
proc freq data=dailyInpatientStatus28days noprint order=freq;
   tables encwid / out=encwid28days;
   run;
proc freq data=encwid28days;
   tables count;
   run;

*reorder into standard form;
proc sort data=dailyInpatientStatus28days;
   by encwid los;
   run;

*add count of day;
data datain.dailyInpatientStatus28days;
   set dailyinpatientstatus28days;
   by encwid;

   if first.encwid then day=1;
   output;
   day = day + 1;
   retain day;
   run;
```

B.1.2 Covariate Selection in SAS

```
libname datain "T:\DW Research\DW Projects\larbuckle\datain";

proc sort data=datain.dailyInpatientStatus28days;
   by primaryCondition encwid day;
   run;

ods output ModelBuildingSummary=datain.logitfit_allcovars_seq_withby;
ods output FitStatistics=datain.logitfit_allcovars_seq_withby;
ods rtf file =
   "T:\DW Research\DW Projects\larbuckle\dataout\
      stepwise_logistic_withby.rtf";

proc logistic data=datain.dailyInpatientStatus28days;
```

```
   by primaryCondition;
   class campus charlson gender admitType admitPriority accommodation;

   model exit24hrs(event='1') = los campus charlson admitType age
      gender admitPriority urgentAdm ambulance accident transferIn
      retired accommodation married ohip occUnknown religion
      regularAdm laps icu surgery nursStaOcc
      / selection=stepwise slentry=0.99 slstay=0.99;

   *following shtatland to create selection sequence, use AIC/SC for
      best models, selection score for the best model (used loosely);
   *harrold recommends slentry/slstay=0.5 for prediction
      (if you must use p-values);
   *economists only drop covars whose parameters are in the wrong
      direction (contrary to expert knowledge) or small;

   run;
ods rtf close;

ods rtf file =
   "T:\DW Research\DW Projects\larbuckle\dataout\
      logitfit_allcovars_seq_withby_minaic.rtf";

proc means data=datain.logitfit_allcovars_seq_withby
      (where=(Criterion eq 'AIC'));
   by primaryCondition;
   var InterceptAndCovariates;
   output out=aicoutint minid(InterceptAndCovariates(step))=minstep;
   run;
ods rtf close;

ods rtf file =
   "T:\DW Research\DW Projects\larbuckle\dataout\
      stepwise_logistic_withby_aic_min_ints.rtf";

proc print data=datain.logitfit_allcovars_seq_withby noobs;
   by primaryCondition;
   var Criterion InterceptAndCovariates;
   run;
ods rtf close;
```

B.1.3 Data Reduction in SAS

```
libname datain "T:\DW Research\DW Projects\larbuckle\datain";
libname dataout "T:\DW Research\DW Projects\larbuckle\datain";

* create dataset for fitting model:
  - 100 encs by primary condition
  - subset for last 28 days of encounter, to remove effect of outliers
    (patients with long LOS), but capture status nearing discharge
;

*create dataset of last day for encwid;
proc sort data=datain.dailyinpatientstatus;
   by encwid descending los;
   run;
data datain.dailyinpatientlastday;
   set datain.dailyinpatientstatus;
   by encwid;
   if first.encwid then output;
   run;

*create flat file of encwids with primaryCondition;
data dataout.encwidsbypc;
   set datain.dailyinpatientlastday (keep=encwid primaryCondition);
   run;
proc sort data=dataout.encwidsbypc;
   by primaryCondition encwid;
   run;
*check results;
proc freq data=dataout.encwidsbypc;
   tables primaryCondition;
   run;

*sample of 100 encounters per primaryCondition;
proc surveyselect data=dataout.encwidsbypc method=srs sampsize=100
      out=dataout.encwidSample100bypc;
   strata primaryCondition;
   id _all_;
   run;
proc sort data=datain.dailyInpatientStatus;
   by encwid;
   run;
```

```
proc sort data=dataout.encwidSample100bypc;
   by encwid;
   run;
data dataout.dailyInpatientStatus100bypc;
   merge datain.dailyInpatientStatus dataout.encwidSample100bypc
      (in=inhere);
   by encwid;
   if inhere then output;
   run;
*check results;
proc freq data=dataout.encwidSample100bypc;
   tables primaryCondition;
   run;

* sample of 28 days, 100 encs bypc;
proc sort data=dataout.encwidSample100bypc;
   by encwid;
   run;
data dataout.dis100bypc28days;
   merge datain.dailyInpatientStatus28days dataout.encwidSample100bypc
      (in=inhere);
   by encwid;
   if inhere then output;
   run;

*export for R, 28 days 100 encs by primary condition;
proc sort data=dataout.dis100bypc28days;
   by primaryCondition encwid day;
   run;
data dis100bypc28days;
   set dataout.dis100bypc28days (
      keep=exit24hrs encwid los campus charlson admitType age gender
         admitPriority urgentAdm ambulance accident transferIn retired
         accommodation married ohip occUnknown religion regularAdm
         laps icu surgery nursStaOcc primaryCondition day
         );
   format date MMDDYY10.; *r does not recognize date9. format;
   run;

proc export data=dis100bypc28days outfile=
   "T:\DW Research\DW Projects\larbuckle\datain\dis100bypc28days.csv"
```

```
   DBMS=CSV REPLACE;
   run;

* last day only (for analyzing dataset);
proc sort data=dataout.encwidSample100bypc;
   by encwid;
   run;
data dataout.dis100bypclastday;
   merge datain.dailyinpatientlastday dataout.encwidSample100bypc
      (in=inhere);
   by encwid;
   if inhere then output;
   run;

*export for R;
proc sort data=dataout.dis100bypclastday;
   by primaryCondition encwid;
   run;
data dis100bypclastday;
   set dataout.dis100bypclastday (
      keep=exit24hrs encwid los campus charlson admitType age gender
         admitPriority urgentAdm ambulance accident transferIn retired
         accommodation married ohip occUnknown religion regularAdm
         laps icu surgery nursStaOcc primaryCondition
         );
   format date MMDDYY10.; *r does not recognize date9. format;
   run;

proc export data=dis100bypclastday outfile=
   "T:\DW Research\DW Projects\larbuckle\datain\dis100bypclastday.csv"
   DBMS=CSV REPLACE;
   run;
```

B.1.4 Variance Inflation Factor in R

```
# load packages
library(lattice)
library(Design)

# set working directory, for printouts
setwd("T:/DW Research/DW Projects/larbuckle/dataout/")
```

```
# load data
datain <- read.csv("T:/DW Research/DW Projects/larbuckle/datain/
   dis100bypc28days.csv", header = TRUE, sep = ",")

# indicators coded with 1/-1 not recognized as factors
datain$accident   <- factor(datain$accident)
datain$ambulance  <- factor(datain$ambulance)
datain$married    <- factor(datain$married)
datain$religion   <- factor(datain$religion)
datain$ohip       <- factor(datain$ohip)
datain$transferIn <- factor(datain$transferIn)
datain$regularAdm <- factor(datain$regularAdm)
datain$retired    <- factor(datain$retired)
datain$urgentAdm  <- factor(datain$urgentAdm)
datain$occUnknown <- factor(datain$occUnknown)
datain$icu        <- factor(datain$icu)
datain$surgery    <- factor(datain$surgery)

# linear model
model.lin <- ols(exit24hrs~primaryCondition+day+LAPS+age+charlson+icu
   +surgery+nursStaOcc+gender+campus+accommodation+admitType,
   data=datain)
# check variance inflation factor
summary(model.lin)
vif(model.lin)
# result: no vif above 4 (very good)

model.ols <- lm(exit24hrs~primaryCondition+day+LAPS+age+charlson+icu
   +surgery+nursStaOcc+gender+campus+accommodation+admitType,
   data=datain)
summary(model.ols)

# run ols by pc, gather residuals for plot, and check vif

# create vector of primary conditions
pc <- unique(datain$primaryCondition)
# create atomic vector for residuals
rs <- NULL

# loop over primary conditions
```

```
for (i in 1:length(pc)) {
   # subset data by primary condition
   datasub <- subset(datain, primaryCondition==pc[i])

   # remove unused factors after subsetting using factor()
   datasub$charlson      <- factor(datasub$charlson)
   datasub$surgery       <- factor(datasub$surgery)
   datasub$gender        <- factor(datasub$gender)
   datasub$campus        <- factor(datasub$campus)
   datasub$accommodation <- factor(datasub$accommodation)
   datasub$admitType     <- factor(datasub$admitType)

   # load distribution summaries for covariates (used by ols)
   dd<-datadist(datasub)
   options(datadist="dd")

   # fit using ordinary linear regression
   # remove factors with less than 2 levels
   if (any(i==c(1,2,3,11,15,16,27,30,36))) {
      # remove icu
      model.ols <- ols(exit24hrs~day+LAPS+age+charlson+surgery
         +nursStaOcc+gender+campus+accommodation+admitType,
         data=datasub, tol=1e-10)
      model.lm  <-  lm(exit24hrs~day+LAPS+age+charlson+surgery
         +nursStaOcc+gender+campus+accommodation+admitType,
         data=datasub)
      if (any(vif(model.ols)>10)) print(i)
   } else if (i==9) {
      # remove surgery
      model.ols <- ols(exit24hrs~day+LAPS+age+charlson+icu
         +nursStaOcc+gender+campus+accommodation+admitType,
         data=datasub, tol=1e-10)
      model.lm  <-  lm(exit24hrs~day+LAPS+age+charlson+icu
         +nursStaOcc+gender+campus+accommodation+admitType,
         data=datasub)
      if (any(vif(model.ols)>10)) print(i)
   } else if (any(i==c(14,38))) {
      # remove icu & surgery
      model.ols <- ols(exit24hrs~day+LAPS+age+charlson+nursStaOcc
         +gender+campus+accommodation+admitType,
         data=datasub, tol=1e-10)
```

```
      model.lm  <-  lm(exit24hrs~day+LAPS+age+charlson+nursStaOcc
         +gender+campus+accommodation+admitType,
         data=datasub)
      if (any(vif(model.ols)>10)) print(i)
   } else if (i==33) {
      # remove icu & surgery & gender
      model.ols <- ols(exit24hrs~day+LAPS+age+charlson+nursStaOcc
         +campus+accommodation+admitType,
         data=datasub, tol=1e-10)
      model.lm  <-  lm(exit24hrs~day+LAPS+age+charlson+nursStaOcc
         +campus+accommodation+admitType,
         data=datasub)
      if (any(vif(model.ols)>10)) print(i)
   } else {
      # nothing removed
      model.ols <- ols(exit24hrs~day+LAPS+age+charlson+icu+surgery
         +nursStaOcc+gender+campus+accommodation+admitType,
         data=datasub, tol=1e-10)
      model.lm  <-  lm(exit24hrs~day+LAPS+age+charlson+icu+surgery
         +nursStaOcc+gender+campus+accommodation+admitType,
         data=datasub)
      if (any(vif(model.ols)>10)) print(i)
   }

   # collect residuals
   rs[length(rs)+1:length(residuals(model.lm))] <- residuals(model.lm,
      type="pearson")
}

# check for linear trend with time, indicating need for random slope
pdf("ols-residuals-vs-day-by-pc.pdf")
xyplot(rs~day|primaryCondition, data=datain, as.table=TRUE,
   layout=c(6,8), aspect=c(0.6),
   xlab="Day", ylab="Standardized Residuals"
)
dev.off()
```

B.1.5 Histograms and Scatter Plots in R

```
# load packages
library(lattice)
```

```
library(Design)

# set working directory, for printouts
setwd("T:/DW Research/DW Projects/larbuckle/dataout/")

# load data
dis<-read.csv("T:/DW Research/DW Projects/larbuckle/datain/
   dis100bypc28days.csv", header = TRUE, sep = ",")
disld<-read.csv("T:/DW Research/DW Projects/larbuckle/datain/
   dis100bypclastday.csv", header = TRUE, sep = ",")

# indicators coded with 1/-1 not recognized as factors
dis$accident     <- factor(dis$accident)
dis$ambulance    <- factor(dis$ambulance)
dis$married      <- factor(dis$married)
dis$religion     <- factor(dis$religion)
dis$ohip         <- factor(dis$ohip)
dis$transferIn   <- factor(dis$transferIn)
dis$regularAdm   <- factor(dis$regularAdm)
dis$retired      <- factor(dis$retired)
dis$urgentAdm    <- factor(dis$urgentAdm)
dis$occUnknown   <- factor(dis$occUnknown)
dis$icu          <- factor(dis$icu)
dis$surgery      <- factor(dis$surgery)
dis$exit24hrs    <- factor(dis$exit24hrs)

disld$accident     <- factor(disld$accident)
disld$ambulance    <- factor(disld$ambulance)
disld$married      <- factor(disld$married)
disld$religion     <- factor(disld$religion)
disld$ohip         <- factor(disld$ohip)
disld$transferIn   <- factor(disld$transferIn)
disld$regularAdm   <- factor(disld$regularAdm)
disld$retired      <- factor(disld$retired)
disld$urgentAdm    <- factor(disld$urgentAdm)
disld$occUnknown   <- factor(disld$occUnknown)
disld$icu          <- factor(disld$icu)
disld$surgery      <- factor(disld$surgery)
disld$exit24hrs    <- factor(disld$exit24hrs)

# print summaries of data
```

```
summary(dis)
summary(disld)

pdf("nursStaOcc-histogram-by-pc.pdf")
histogram(~nursStaOcc|primaryCondition, data=dis, as.table=TRUE,
   layout=c(6,8), aspect=c(0.6),
   xlab="Nursing Station Occupancy (%)",
   ylab="Density (for Total Patient Days)",
   type="density",
   panel = function(x,...){
      panel.histogram(x,...)
      panel.mathdensity(dmath = dnorm,
      args = list(mean=mean(x),sd=sd(x)),
      col = "black")
   },
)
dev.off()

pdf("age-histogram-by-pc.pdf")
histogram(~age|primaryCondition, data=disld, as.table=TRUE,
   layout=c(6,8), aspect=c(0.6),
   xlab="Age at Admission (Years)",
   ylab="Density (for Total Patient Encounters)",
   type="density",
   panel = function(x,...){
      panel.histogram(x,...)
      panel.mathdensity(dmath = dnorm,
      args = list(mean=mean(x),sd=sd(x)),
      col = "black")
   },
)
dev.off()

pdf("laps-histogram-by-pc.pdf")
histogram(~LAPS|primaryCondition, data=dis, as.table=TRUE,
   layout=c(6,8), aspect=c(0.6),
   xlab="Laboratory-based Acute Physiology Score (LAPS)",
   ylab="Density (for Total Patient Days)",
   type="density",
   panel = function(x,...){
      panel.histogram(x,...)
```

```
      panel.mathdensity(dmath = dnorm,
      args = list(mean=mean(x),sd=sd(x)),
      col = "black")
   },
)
dev.off()

pdf("los-histogram-by-pc.pdf")
histogram(~los|primaryCondition, data=disld, as.table=TRUE,
   layout=c(6,8), aspect=c(0.6), nint=200,xlim=c(0:30),
   xlab="Length of Stay (Days)",
   ylab="Density (for Total Patient Encounters)",
   type="density",
   panel = function(x,...){
      panel.histogram(x,...)
      panel.mathdensity(dmath = dnorm,
      args = list(mean=mean(x),sd=sd(x)),
      col = "black")
   },
)
dev.off()

pdf("los-vs-age-by-pc.pdf")
xyplot(los~age|primaryCondition, data=disld, as.table=TRUE,
   layout=c(6,8), aspect=c(0.6),
   type=c("p","r"), col.line="black", xlim=c(1,120), ylim=c(0,30),
   xlab="Age at Admission (Years)", ylab="Length of Stay (Days)"
)
dev.off()

pdf("laps-vs-age-by-pc.pdf")
xyplot(LAPS~age|primaryCondition, data=dis, as.table=TRUE,
   layout=c(6,8), aspect=c(0.6),
   xlim=c(1,120), xlab="Age at Admission (Years)",
   ylab="Laboratory-based Acute Physiology Score (LAPS)"
)
dev.off()

pdf("charlson-vs-age-by-pc.pdf")
xyplot(charlson~age|primaryCondition, data=disld, as.table=TRUE,
   layout=c(6,8), aspect=c(0.6),
```

```
   xlim=c(1,120), xlab="Age at Admission (Years)",
   ylab="Charlson Index"
)
dev.off()

pdf("laps-vs-charlson-by-pc.pdf")
xyplot(LAPS~charlson|primaryCondition, data=dis, as.table=TRUE,
   layout=c(6,8), aspect=c(0.6),
   xlab="Charlson Index",
   ylab="Laboratory-based Acute Physiology Score (LAPS)"
)
dev.off()
```

B.1.6 Time-Series Data in R

```
# load packages
library(forecast)
library(MASS) # truehist

# set working directory, for printouts
setwd("T:/DW Research/DW Projects/larbuckle/dataout/")

# campus names
campus <- c('civic','general')
# set input data file
datafile1 <- "T:/DW Research/DW Projects/larbuckle/datain/admissions"
# set working directory, for output
setwd("T:/DW Research/DW Projects/larbuckle/dataout/")

# load data for civic
datafile <- paste(datafile1, campus[1], ".csv", sep="")
admisciv <- read.csv(datafile, header = TRUE, sep = ",")
# create ts object, starting at 01 April 2004 (day 92 of year)
admisciv <- ts(admisciv[[1]], start=c(2004,92), frequency=365)

# load data for general
datafile <- paste(datafile1, campus[2], ".csv", sep="")
admisgen <- read.csv(datafile, header = TRUE, sep = ",")
# create ts object, starting at 01 April 2004 (day 92 of year)
admisgen <- ts(admisgen[[1]], start=c(2004,92), frequency=365)
```

```
# ts, hist, qq for civic
pdf("ts-hist-qq-plots-civic.pdf")
layout(matrix(c(1,2, 1,3), nc=2))
maintitle <- "Emergency Admisssions to Civic Campus"

plot.ts(admisciv,  main=maintitle, ylab = "Admissions",
   xlab="Time (Years)")
truehist(admisciv, main="Histogram of Admissions", xlab="Admissions",
   ylab="Density")
curve(dnorm(x, mean=mean(admisciv),sd=sd(admisciv)), add=TRUE)
qqnorm(admisciv, main="Normal Q-Q Plot of Admissions")
qqline(admisciv, col=4)

layout(matrix(c(1,1, 1,1), nc=2))
dev.off()

shapiro.test(admisciv)
# p-value = 0.18 > 0.01 => can't reject normality

# ts, hist, qq for general
pdf("ts-hist-qq-plots-civic.pdf")
layout(matrix(c(1,2, 1,3), nc=2))
maintitle <- "Emergency Admisssions to General Campus"

plot.ts(admisgen,  main=maintitle, ylab = "Admissions",
   xlab="Time (Years)")
truehist(admisgen, main="Histogram of Admissions", xlab="Admissions",
   ylab="Density")
curve(dnorm(x, mean=mean(admisciv),sd=sd(admisciv)), add=TRUE)
qqnorm(admisgen, main="Normal Q-Q Plot of Admissions")
qqline(admisgen, col=4)

layout(matrix(c(1,1, 1,1), nc=2))
dev.off()

shapiro.test(admisgen)
# p-value = 0.025 > 0.01 => can't reject normality
```

B.2 Population-Averaged Models in R

B.2.1 One Model for All Primary Conditions

```
# load packages
library(doBy)
library(geepack)
library(lattice)
library(Hmisc)
library(ROCR)

# load data 100bypc28days
datain <- read.csv("T:/DW Research/DW Projects/larbuckle/datain/
   dis100bypc28days.csv", header = TRUE, sep = ",")

# indicators coded with 1/-1 not recognized as factors
datain$icu<-factor(datain$icu)
datain$surgery<-factor(datain$surgery)

# create vector of primary conditions
pc <- unique(datain$primaryCondition)

# set working directory, for output
setwd("T:/DW Research/DW Projects/larbuckle/dataout/")

# open a file for writing out results
write("logistic regression with gee, for 100 patients up to their
   last 28 days, main effects, correlation ar1",
   file="T:/DW Research/DW Projects/larbuckle/dataout/tmp/
   gee28days-maineffects-ar1.txt")
write(c("c-ind", "brier"), sep="\t", append=TRUE, ncolumns=2,
   file="T:/DW Research/DW Projects/larbuckle/dataout/tmp/
   gee28days-maineffects-ar1.txt")

# create atomic vectors
rs   <- NULL # residuals
pp   <- NULL # predicted probabilities (of discharge)
cind <- NULL # c-index (c-stat)
br   <- NULL # brier score

# remove unused factors, if any
```

```
datain$charlson      <- factor(datain$charlson)
datain$surgery       <- factor(datain$surgery)
datain$gender        <- factor(datain$gender)
datain$campus        <- factor(datain$campus)
datain$accommodation <- factor(datain$accommodation)
datain$admitType     <- factor(datain$admitType)

# fit using logistic regression
model.gee <- geeglm(exit24hrs~primaryCondition+day+LAPS+age+charlson
   +icu+surgery+nursStaOcc+gender+campus+accommodation+admitType,
   data=datain, family=binomial, id=encWID, corstr="ar1")

# calculate c-stat
# use plogis to go from log-odds to probability
cind <- somers2(plogis(predict(model.gee)), datain$exit24hrs)[1]

# brier score
br <- mean((plogis(predict(model.gee))-datain$exit24hrs)^2)

# write to file
write(c(format(cind,digits=3), format(br,digits=3)), sep="\t",
   append=TRUE, ncolumns=2,
   file="T:/DW Research/DW Projects/larbuckle/dataout/tmp/
   gee28days-maineffects-ar1.txt")

# collect residuals (using pearson to compare to other models)
rs[1:length(residuals(model.gee))] <- residuals(model.gee,type="pearson")
# collect predicted probabilities of discharge
pp[1:length(predict(model.gee))] <- plogis(predict(model.gee))

# residual plot
xy1 <- xyplot(rs~pp, xlab="Probability of Discharge",
   ylab="Standardized Residuals",
   panel = function(...){
      panel.abline(0,0)
      panel.xyplot(...)
   },
)

# histogram of residuals
hi1 <- histogram(~rs, xlab="Standardized Residuals", type="density",
```

```
   panel = function(x,...){
      panel.histogram(x,...)
      panel.mathdensity(dmath = dnorm,
      args = list(mean=mean(x),sd=sd(x)),
      col = "black")
   },
)

#qqplot of residuals
qq1 <- qqmath(~rs, xlab="Theoretical Quantiles",
   ylab="Sample Quantiles",
   panel = function(x,...){
      panel.qqmath(x,...)
      panel.qqmathline(x,...)
   },
)

# panel of histos/qqpltos for random factors
pdf("gee28days-maineffects-ar1-resids-histo-qq.pdf")
plot(xy1, split=c(1,1,2,2))
plot(hi1, split=c(1,2,2,2), newpage=FALSE)
plot(qq1, split=c(2,2,2,2), newpage=FALSE)
dev.off()

# residual plot by pc
xy2 <- xyplot(rs~pp|datain$primaryCondition, as.table=TRUE,
   layout=c(6,8), aspect=c(0.6),
   xlab="Probability of Discharge", ylab="Standardized Residuals",
   panel = function(...){
      panel.abline(0,0)
      panel.xyplot(...)
   },
)
pdf("gee28days-maineffects-ar1-residualplot-bypc.pdf")
plot(xy2)
dev.off()

# roc plot
pred <- prediction(pp, datain$exit24hrs)
roc.gee <- performance(pred, measure = "tpr", x.measure = "fpr")
```

```
# accuracy plot
acc.gee <- performance(pred, measure = "acc")

# print model summary
summary(model.gee)
```

B.2.2 One Model for Each Primary Condition—44 Models

```
# load packages
library(doBy)
library(geepack)
library(lattice)
library(Hmisc)
library(ROCR)

# load data 100bypc28days
datain <- read.csv("T:/DW Research/DW Projects/larbuckle/datain/
   dis100bypc28days.csv", header = TRUE, sep = ",")

# indicators coded with 1/-1 not recognized as factors
datain$icu<-factor(datain$icu)
datain$surgery<-factor(datain$surgery)

# create vector of primary conditions
pc <- unique(datain$primaryCondition)

# need to remove factors with less than 2 levels
# can remove from data by subsetting with levels() and table()
# but must remove from model function
for (i in 1:length(pc)) {
   datasub <- subset(datain, primaryCondition==pc[i])
   if (length(table(factor(datasub$charlson)))<2) {
      print(c(i, 'missing charlson'))
   }
   if (length(table(factor(datasub$icu)))<2) {
      print(c(i, 'missing icu'))
   }
   if (length(table(factor(datasub$surgery)))<2) {
      print(c(i, 'missing surgery'))
   }
   if (length(table(factor(datasub$gender)))<2) {
```

```
      print(c(i, 'missing gender'))
   }
   if (length(table(factor(datasub$campus)))<2) {
      print(c(i, 'missing campus'))
   }
   if (length(table(factor(datasub$accommodation)))<2) {
      print(c(i, 'missing accommodation'))
   }
   if (length(table(factor(datasub$admitType)))<2) {
      print(c(i, 'missing admitType'))
   }
}
# results:
   # missing icu (only) in groups 1,2,3,11,15,16,27,30,36
   # missing surgery (only) in group 9
   # missing icu & surgery (only) in groups 14,38
   # missing icu & surgery & gender in group 33


# open a file for writing out results
write("gee by primary condition, for 100 patients up to their last 28
   days, main effects, ar1 correlation",
   file="T:/DW Research/DW Projects/larbuckle/dataout/
   geebypc28days-maineffects-ar1.txt")
write(c("pc", "c-ind", "brier"), sep="\t", append=TRUE, ncolumns=3,
   file="T:/DW Research/DW Projects/larbuckle/dataout/
   geebypc28days-maineffects-ar1.txt")

# create atomic vectors
rs   <- NULL # residuals
pp   <- NULL # predicted probabilities (of discharge)
cind <- NULL # c-index (c-stat)
br   <- NULL # brier score

# loop over primary conditions
for (i in 1:length(pc)) {
   # subset data by primary condition
   datasub <- subset(datain, primaryCondition==pc[i])

   # remove unused factors after subsetting using factor()
   datasub$charlson     <- factor(datasub$charlson)
```

```
datasub$surgery       <- factor(datasub$surgery)
datasub$gender        <- factor(datasub$gender)
datasub$campus        <- factor(datasub$campus)
datasub$accommodation <- factor(datasub$accommodation)
datasub$admitType     <- factor(datasub$admitType)

# fit using gee, removing factors with less than 2 levels
if (any(i==c(1,2,3,11,15,16,27,30,36))) {
   # remove icu
   model.geebypc <- geeglm(exit24hrs~day+LAPS+age+charlson+surgery
      +nursStaOcc+gender+campus+accommodation+admitType,
      data=datasub, family="binomial", id=encWID, corstr="ar1")
} else if (i==9) {
   # remove surgery
   model.geebypc <- geeglm(exit24hrs~day+LAPS+age+charlson+icu
      +nursStaOcc+gender+campus+accommodation+admitType,
      data=datasub, family="binomial", id=encWID, corstr="ar1")
} else if (any(i==c(14,38))) {
   # remove icu & surgery
   model.geebypc <- geeglm(exit24hrs~day+LAPS+age+charlson
      +nursStaOcc+gender+campus+accommodation+admitType,
      data=datasub, family="binomial", id=encWID, corstr="ar1")
} else if (i==33) {
   # remove icu & surgery & gender
   model.geebypc <- geeglm(exit24hrs~day+LAPS+age+charlson
      +nursStaOcc+campus+accommodation+admitType,
      data=datasub, family="binomial", id=encWID, corstr="ar1")
} else {
   # nothing removed
   model.geebypc <- geeglm(exit24hrs~day+LAPS+age+charlson+icu
      +surgery+nursStaOcc+gender+campus+accommodation+admitType,
   data=datasub, family="binomial", id=encWID, corstr="ar1")
}

# calculate c-stat
# use plogis to go from log-odds to probability
cind[i] <- somers2(plogis(predict(model.geebypc)),
   datasub$exit24hrs)[1]

# write to file
write(c(as.character(pc[i]), format(cind[i],digits=3), format(br[i],
```

```
      digits=3)), sep="\t", append=TRUE, ncolumns=3,
      file="T:/DW Research/DW Projects/larbuckle/dataout/
      geebypc28days-maineffects-ar1.txt")

   # calculate mean stats
   if (i==length(pc)) {
      write(c("mean", format(mean(cind,na.rm=TRUE),digits=3),
         format(mean(br,na.rm=TRUE),digits=3)), sep="\t", append=TRUE,
         ncolumns=3,
         file="T:/DW Research/DW Projects/larbuckle/dataout/
         geebypc28days-maineffects-ar1.txt")
   }

   # collect residuals
   rs[length(rs)+1:length(residuals(model.geebypc))] <-
      residuals(model.geebypc,type="pearson")
   # collect predicted probabilities of discharge
   pp[length(pp)+1:length(predict(model.geebypc))] <-
      plogis(predict(model.geebypc))
}

# set working directory, for printouts
setwd("T:/DW Research/DW Projects/larbuckle/dataout/")

# residual plot
xy1 <- xyplot(rs~pp, xlab="Probability of Discharge",
   ylab="Standardized Residuals",
   panel = function(...){
      panel.abline(0,0)
      panel.xyplot(...)
   },
)

# histogram of residuals
hi1 <- histogram(~rs, xlab="Standardized Residuals", type="density",
   panel = function(x,...){
      panel.histogram(x,...)
      panel.mathdensity(dmath = dnorm,
      args = list(mean=mean(x),sd=sd(x)),
      col = "black")
   },
```

```
)

#qqplot of residuals
qq1 <- qqmath(~rs, xlab="Theoretical Quantiles",
   ylab="Sample Quantiles",
   panel = function(x,...){
      panel.qqmath(x,...)
      panel.qqmathline(x,...)
   },
)

# panel of histos/qqpltos for random factors
pdf("geebypc28days-maineffects-resids-histo-qq.pdf")
plot(xy1, split=c(1,1,2,2))
plot(hi1, split=c(1,2,2,2), newpage=FALSE)
plot(qq1, split=c(2,2,2,2), newpage=FALSE)
dev.off()

# residual plot by pc
xy2 <- xyplot(rs~pp|datain$primaryCondition, as.table=TRUE,
   layout=c(6,8), aspect=c(0.6),
   xlab="Probability of Discharge", ylab="Standardized Residuals",
   panel = function(...){
      panel.abline(0,0)
      panel.xyplot(...)
   },
)
pdf("geebypc28days-maineffects-residualplot-bypc.pdf")
plot(xy2)
dev.off()

# roc plot
pred <- prediction(pp, datain$exit24hrs)
roc.geebypc <- performance(pred, measure = "tpr", x.measure = "fpr")

# accuracy plot
acc.geebypc <- performance(pred, measure = "acc")

# print model summary
pc[i]
summary(model.geebypc)
```

B.3 Subject-Specific Models in R

B.3.1 One Model for Each Primary Condition—44 Models

```
# load packages
library(MASS)
library(nlme)
library(lattice)
library(Hmisc)
library(ROCR)

# set correlation structure
cors <- "ind" # independent
#cors <- "ar1" # ar(1)
#cors <- "exc" # exchangeable (compound symmetric)
#cors <- "uns" # unstructured (symmetric)--convergence problems

# load data 100bypc28days
datain<-read.csv("T:/DW Research/DW Projects/larbuckle/datain/
   dis100bypc28days.csv", header = TRUE, sep = ",")

# indicators coded with 1/-1 not recognized as factors
datain$icu     <- factor(datain$icu)
datain$surgery <- factor(datain$surgery)

# create vector of primary conditions
pc <- unique(datain$primaryCondition)

# need to remove factors with less than 2 levels
# can remove from data by subsetting with levels() and table()
# but must remove from model function
for (i in 1:length(pc)) {
   datasub <- subset(datain, primaryCondition==pc[i])
   if (length(table(factor(datasub$charlson)))<2) {
      print(c(i, 'missing charlson'))
   }
   if (length(table(factor(datasub$icu)))<2) {
      print(c(i, 'missing icu'))
   }
   if (length(table(factor(datasub$surgery)))<2) {
      print(c(i, 'missing surgery'))
```

```
   }
   if (length(table(factor(datasub$gender)))<2) {
      print(c(i, 'missing gender'))
   }
   if (length(table(factor(datasub$campus)))<2) {
      print(c(i, 'missing campus'))
   }
   if (length(table(factor(datasub$accommodation)))<2) {
      print(c(i, 'missing accommodation'))
   }
   if (length(table(factor(datasub$admitType)))<2) {
      print(c(i, 'missing admitType'))
   }
}
# results:
   # missing icu (only) in groups 1,2,3,8,11,15,16,19,27,30,36
   # missing surgery (only) in group 9
   # missing icu & surgery (only) in groups 14,38
   # missing icu & surgery & gender in group 33

# set working directory, for output
setwd("T:/DW Research/DW Projects/larbuckle/dataout/")

# open a file for writing out results
fn <- "glmmPQLbypc28days-maineffects-"
fn <- paste(fn, cors, ".txt", sep="")
write(c("glmmPQL by primary condition, for 100 patients up to their
   last 28 days, main effects, random intercept, correlation",cors),
   ncolumns=2, file=fn)
write(c("pc", "c-ind", "brier"), sep="\t", append=TRUE, ncolumns=3,
   file=fn)

# create atomic vectors
rs   <- NULL # residuals
pp   <- NULL # predicted probabilities (of discharge)
acf  <- NULL # auto-correlation
cind <- NULL # c-index (c-stat)
br   <- NULL # brier score

# set correlation structure (corClasses) for glmmPQL (random intercept)
if (cors=="ind") {
```

```
   corr <- NULL
} else if (cors=="ar1") {
   corr <- corAR1(form=~1|encWID)
} else if (cors=="exc") {
   corr <- corCompSymm(form=~1|encWID)
} else if (cors=="uns") {
   corr <- corSymm(form=~1|encWID)
}

# start of main loop (if algo fails, we skip it and update start)
start <- 1

# loop over primary conditions
for (i in start:length(pc)) {
   # subset data by primary condition
   datasub <- subset(datain, primaryCondition==pc[i])

   # remove unused factors after subsetting using factor()
   datasub$charlson     <- factor(datasub$charlson)
   datasub$surgery       <- factor(datasub$surgery)
   datasub$gender        <- factor(datasub$gender)
   datasub$campus        <- factor(datasub$campus)
   datasub$accommodation <- factor(datasub$accommodation)
   datasub$admitType     <- factor(datasub$admitType)

   # fit using glmmPQL, removing factors with less than 2 levels
   # (data assumed to be orderd by encwid and day)
   if (any(i==c(1,2,3,8,11,15,16,19,27,30,36))) {
      # remove icu
      model.2level <- glmmPQL(exit24hrs~day+LAPS+age+charlson+surgery
         +nursStaOcc+gender+campus+accommodation+admitType,
         random=~1|encWID, data=datasub, family="binomial", cor=corr)
   } else if (i==9) {
      # remove surgery
      model.2level <- glmmPQL(exit24hrs~day+LAPS+age+charlson+icu
         +nursStaOcc+gender+campus+accommodation+admitType,
         random=~1|encWID, data=datasub, family="binomial", cor=corr)
   } else if (any(i==c(14,38))) {
      # remove icu & surgery
      model.2level <- glmmPQL(exit24hrs~day+LAPS+age+charlson
         +nursStaOcc+gender+campus+accommodation+admitType,
```

```
      random=~1|encWID, data=datasub, family="binomial", cor=corr)
} else if (i==33) {
   # remove icu & surgery & gender
   model.2level <- glmmPQL(exit24hrs~day+LAPS+age+charlson
      +nursStaOcc+campus+accommodation+admitType,
      random=~1|encWID, data=datasub, family="binomial", cor=corr)
} else {
   # nothing removed
   model.2level <- glmmPQL(exit24hrs~day+LAPS+age+charlson+icu
      +surgery+nursStaOcc+gender+campus+accommodation+admitType,
      random=~1|encWID, data=datasub, family="binomial", cor=corr)
}

# calculate c-stat
# use plogis to go from log-odds to probability
cind[i] <- somers2(plogis(predict(model.2level)),
   datasub$exit24hrs)[1]

# brier score
br[i] <- mean((plogis(predict(model.2level))-datasub$exit24hrs)^2)

# write to file
write(c(as.character(pc[i]), format(cind[i],digits=3), format(br[i],
   digits=3)), sep="\t", append=TRUE, ncolumns=3, file=fn)

if (i==length(pc)) {
   # mean of all metrics, ignoring NAs (ie, algo failures)
   write(c("mean", format(mean(cind,na.rm=TRUE),digits=3),
      format(mean(br,na.rm=TRUE),digits=3)), sep="\t", append=TRUE,
      ncolumns=3, file=fn)
   # mean of all metrics, penalizing for NAs (ie, algo failures)
   write(c("sm/44", format(sum(cind,na.rm=TRUE)/length(pc),
      digits=3), format(sum(br,na.rm=TRUE)/length(pc),digits=3)),
      sep="\t", append=TRUE, ncolumns=3, file=fn)
}

# collect residuals
rs[length(rs)+1:length(residuals(model.2level))] <-
   residuals(model.2level,type="pearson")
# collect predicted probabilities of discharge
pp[length(pp)+1:length(predict(model.2level))] <-
```

```
      plogis(predict(model.2level))

   # if glmmPQL fails: copy/paste text inside if-statement,
   # then repeat main for loop
   # (write to file, write NULL to rs and pp (for residual plot),
   # then update start of loop)
   if (FALSE) {
      write(c(as.character(pc[i]), format(cind[i],digits=3),
         format(br[i],digits=3)), sep="\t", append=TRUE, ncolumns=3,
         file=fn)
      rsNA <- rep(NA,length(datasub$exit24hrs))
      rs[length(rs)+1:length(rsNA)] <- rsNA
      ppNA <- rep(NA,length(datasub$exit24hrs))
      pp[length(pp)+1:length(ppNA)] <- ppNA
      start <- i+1
   }
}

# residual plot
xy1 <- xyplot(rs~pp, xlab="Probability of Discharge",
   ylab="Standardized Residuals",
   panel = function(...){
      panel.abline(0,0)
      panel.xyplot(...)
   },
)

# histogram of residuals
hi1 <- histogram(~rs, xlab="Standardized Residuals", type="density",
   panel = function(x,...){
      panel.histogram(x,...)
      panel.mathdensity(dmath = dnorm,
      args = list(mean=mean(x),sd=sd(x)),
      col = "black")
   },
)

#qqplot of residuals
qq1 <- qqmath(~rs, xlab="Theoretical Quantiles",
   ylab="Sample Quantiles",
   panel = function(x,...){
```

```
      panel.qqmath(x,...)
      panel.qqmathline(x,...)
   },
)

# panel of histos/qqpltos for residuals
pn <- "glmmPQLbypc28days-maineffects-"
pn <- paste(pn, cors, "-resids-histo-qq.pdf", sep="")
pdf(pn)
plot(xy1, split=c(1,1,2,2))
plot(hi1, split=c(1,2,2,2), newpage=FALSE)
plot(qq1, split=c(2,2,2,2), newpage=FALSE)
dev.off()

# residual plot by pc
xy2 <- xyplot(rs~pp|datain$primaryCondition, as.table=TRUE,
   layout=c(6,8), aspect=c(0.6),
   xlab="Probability of Discharge", ylab="Standardized Residuals",
   panel = function(...){
      panel.abline(0,0)
      panel.xyplot(...)
   },
)
pn <- "glmmPQLbypc28days-maineffects-"
pn <- paste(pn, cors, "-residualplot-bypc.pdf", sep="")
pdf(pn)
plot(xy2)
dev.off()

# roc plot
pred <- prediction(pp, datain$exit24hrs)
roc.2level <- performance(pred, measure = "tpr", x.measure = "fpr")

# accuracy plot
acc.2level <- performance(pred, measure = "acc")

# print model summary
pc[i]
summary(model.2level)
```

B.3.2 One Model for All Primary Conditions

```
# load packages
library(MASS)
library(nlme)
library(lattice)
library(Hmisc)
library(ROCR)

# set correlation structure
cors <- "ind" # independent
#cors <- "ar1" # ar(1)
#cors <- "exc" # exchangeable (compound symmetric)
#cors <- "uns" # unstructured (symmetric)--convergence problems

# load data 100bypc28days
datain <- read.csv("T:/DW Research/DW Projects/larbuckle/datain/
   dis100bypc28days.csv", header = TRUE, sep = ",")

# indicators coded with 1/-1 not recognized as factors
datain$icu     <- factor(datain$icu)
datain$surgery <- factor(datain$surgery)

# set working directory, for output
setwd("T:/DW Research/DW Projects/larbuckle/dataout/")

# open a file for writing out results
fn <- "glmmPQL3lev28days-maineffects-"
fn <- paste(fn, cors, ".txt", sep="")
write(c("glmmPQL using three levels, for 100 patients up to their last
   28 days, main effects, random intercept, correlation",cors),
   ncolumns=2, file=fn)
write(c("c-ind", "brier"), sep="\t", append=TRUE, ncolumns=2, file=fn)

# create atomic vectors
rs   <- NULL # residuals
pp   <- NULL # predicted probabilities (of discharge)
cind <- NULL # c-index (c-stat)
br   <- NULL # brier score

# set correlation structure (corClasses) for glmmPQL (random intercept)
if (cors=="ind") {
```

```
   corr <- NULL
} else if (cors=="ar1") {
   corr <- corAR1(form=~1|encWID)
} else if (cors=="exc") {
   corr <- corCompSymm(form=~1|encWID)
} else if (cors=="uns") {
   corr <- corSymm(form=~1|encWID)
}

# remove unused factors, if any
datain$charlson     <- factor(datain$charlson)
datain$surgery       <- factor(datain$surgery)
datain$gender        <- factor(datain$gender)
datain$campus        <- factor(datain$campus)
datain$accommodation <- factor(datain$accommodation)
datain$admitType     <- factor(datain$admitType)

# fit using glmmPQL, removing factors with less than 2 levels
model.3level<-glmmPQL(exit24hrs~day+LAPS+age+charlson+icu+surgery
   +nursStaOcc+gender+campus+accommodation+admitType,
   random=list(primaryCondition=~1, encWID=~1), family=binomial,
   data=datain)

# calculate c-stat
# use plogis to go from log-odds to probability
cind <- somers2(plogis(predict(model.3level)), datain$exit24hrs)[1]

# brier score
br <- mean((plogis(predict(model.3level))-datain$exit24hrs)^2)

# write to file
write(c(format(cind,digits=3), format(br,digits=3)), sep="\t",
   append=TRUE, ncolumns=2, file=fn)

# collect residuals
rs[1:length(residuals(model.3level))] <- residuals(model.3level,
   type="pearson")
# collect predicted probabilities of discharge
pp[1:length(predict(model.3level))] <- plogis(predict(model.3level))

# residual plot
```

```
xy1 <- xyplot(rs~pp, xlab="Probability of Discharge",
   ylab="Standardized Residuals",
   panel = function(...){
      panel.abline(0,0)
      panel.xyplot(...)
   },
)

# histogram of residuals
hi1 <- histogram(~rs, xlab="Standardized Residuals", type="density",
   panel = function(x,...){
      panel.histogram(x,...)
      panel.mathdensity(dmath = dnorm,
      args = list(mean=mean(x),sd=sd(x)),
      col = "black")
   },
)

#qqplot of residuals
qq1 <- qqmath(~rs, xlab="Theoretical Quantiles",
   ylab="Sample Quantiles",
   panel = function(x,...){
      panel.qqmath(x,...)
      panel.qqmathline(x,...)
   },
)

# panel of histos/qqpltos for residuals
pn <- "glmmPQL3lev28days-maineffects-"
pn <- paste(pn, cors, "-resids-histo-qq.pdf", sep="")
pdf(pn)
plot(xy1, split=c(1,1,2,2))
plot(hi1, split=c(1,2,2,2), newpage=FALSE)
plot(qq1, split=c(2,2,2,2), newpage=FALSE)
dev.off()

# residual plot by pc
xy2 <- xyplot(rs~pp|datain$primaryCondition, as.table=TRUE,
   layout=c(6,8), aspect=c(0.6),
   xlab="Probability of Discharge", ylab="Standardized Residuals",
   panel = function(...){
```

```
      panel.abline(0,0)
      panel.xyplot(...)
   },
)
pn <- "glmmPQL3lev28days-maineffects-"
pn <- paste(pn, cors, "-residualplot-bypc.pdf", sep="")
pdf(pn)
plot(xy2)
dev.off()

# roc plot--need to combine with other models (gee, 2-level)
pred <- prediction(pp, datain$exit24hrs)
roc.3level <- performance(pred, measure = "tpr", x.measure = "fpr")
pn <- "glmmPQL3lev28days-maineffects-"
pn <- paste(pn, cors, "-rocplot-allmodels.pdf", sep="")
pdf(pn)

plot(roc.3level, lty=1, xlab="False Positive Rate (1-Specificity)",
   ylab="True Positive Rate (Sensitivity)")
plot(roc.2level, lty=2, add=TRUE)
plot(roc.geebypc, lty=3, add=TRUE)
plot(roc.gee, lty=4, add=TRUE)
legend(0.45,0.6, bty="n", c("3-level, random intercept",
   "2-level, random intercept, by PC", "Logistic/GEE by PC",
   "Logistic/GEE"), lty=c(1,2,3,4)
)
dev.off()

# accuracy plot--need to combine with other models (gee, 2-level )
acc.3level <- performance(pred, measure = "acc")
pn <- "glmmPQL3lev28days-maineffects-"
pn <- paste(pn, cors, "-accplot-allmodels.pdf", sep="")
pdf(pn)

plot(acc.3level, lty=1)
plot(acc.2level, lty=2, add=TRUE)
plot(acc.geebypc, lty=3, add=TRUE)
plot(acc.gee, lty=4, add=TRUE)
legend(0.45,0.6, bty="n", c("3-level, random intercept",
   "2-level, random intercept, by PC", "Logistic/GEE by PC",
   "Logistic/GEE"), lty=c(1,2,3,4)
```

```
)
dev.off()

# extract random intercepts
re <- random.effects(model.3level)
repc <- re$primaryCondition[[1]] # for primaryConditions
reenc <- re$encWID[[1]] # for encWIDs

# plot histogram of random effects for pc
hi2 <- histogram(~repc, type="density",
   xlab="Random Intercepts for Primary Conditions",
   panel = function(x,...){
      panel.histogram(x,...)
      panel.mathdensity(dmath = dnorm,
      args = list(mean=mean(x),sd=sd(x)),
      col = "black")
   },
)

# qqplot of random effects for pc
qq2 <- qqmath(~repc, xlab="Theoretical Quantiles",
   ylab="Sample Quantiles",
   panel = function(x,...){
      panel.qqmath(x,...)
      panel.qqmathline(x,...)
   },
)

# plot histogram of random effects for encwid
hi3 <- histogram(~reenc, type="density",
   xlab="Random Intercepts for Patient Encounters",
   panel = function(x,...){
      panel.histogram(x,...)
      panel.mathdensity(dmath = dnorm,
      args = list(mean=mean(x),sd=sd(x)),
      col = "black")
   },
)

# qqplot of random effects for encwid
qq3 <- qqmath(~reenc, xlab="Theoretical Quantiles",
```

```
   ylab="Sample Quantiles",
   panel = function(x,...){
      panel.qqmath(x,...)
      panel.qqmathline(x,...)
   },
)

# panel of histos/qqplots for random factors
pn <- "glmmPQL3lev28days-maineffects-"
pn <- paste(pn, cors, "-histo-qq-random.pdf", sep="")
pdf(pn)
plot(hi2, split=c(1,1,2,2))
plot(qq2, split=c(2,1,2,2), newpage=FALSE)
plot(hi3, split=c(1,2,2,2), newpage=FALSE)
plot(qq3, split=c(2,2,2,2), newpage=FALSE)
dev.off()

# print model summary
summary(model.3level)
intervals(model.3level)
```

B.3.3 Creating Out-of-Sample Dataset in SAS

```
* create a dataset for validation:
  - small since it will be added to the dataset used for fitting
  - new encounters, outside dataset used for fitting
;

* encounters not in sample of 100 encs by pc;
proc sort data=dataout.encwidsbypc;
   by encwid;
   run;
proc sort data=dataout.encwidSample100bypc;
   by encwid;
   run;
data dataout.enciwdsbypcNOT100bypc;
   merge dataout.encwidsbypc dataout.encwidSample100bypc (in=inhere);
   by encwid;
   if ^inhere then output;
   run;
```

```
* create flat file of encwids with primaryCondition
  (not in sample of 100 encs by pc);
data dataout.enciwdsbypcNOT100bypc;
   set datain.dailyinpatientlastday (keep=encwid primaryCondition);
   run;
proc sort data=dataout.enciwdsbypcNOT100bypc;
   by primaryCondition encwid;
   run;
* check results;
proc freq data=dataout.enciwdsbypcNOT100bypc;
   tables primaryCondition;
   run;

* validation sample of 100 encounters per primaryCondition;
proc surveyselect data=dataout.enciwdsbypcNOT100bypc method=srs
   sampsize=100 out=dataout.encwidSample100bypc_val;
   strata primaryCondition;
   id _all_;
   run;
proc sort data=datain.dailyInpatientStatus;
   by encwid;
   run;
proc sort data=dataout.encwidSample100bypc_val;
   by encwid;
   run;
data dataout.dailyInpatientStatus100bypc_val;
   merge datain.dailyInpatientStatus dataout.encwidSample100bypc_val
      (in=inhere);
   by encwid;
   if inhere then output;
   run;
*check results;
proc freq data=dataout.encwidSample100bypc_val;
   tables primaryCondition;
   run;

*export for R, validation sample of 100 encs by pc;
proc sort data=dataout.dailyInpatientStatus100bypc_val;
   by primaryCondition encwid los;
   run;
data dis100bypc_val;
```

```
   set dataout.dailyInpatientStatus100bypc_val (
      keep=exit24hrs encwid date los campus charlson admitType age
      gender accommodation laps icu surgery nursStaOcc
      primaryCondition
   );
format date MMDDYY10.; * r does not recognize date9. format;
   run;

proc export data=dis100bypc_val outfile=
      "T:\DW Research\DW Projects\larbuckle\datain\dis100bypc_val.csv"
      DBMS=CSV REPLACE;
   run;
```

B.3.4 Forecasting and Predictive Accuracy in R

```
# load packages
library(MASS)
library(nlme)
library(doBy)
library(geepack)
library(lattice)
library(Hmisc)

# set correlation structure
cors <- "ind" # independent
#cors <- "ar1" # ar(1)
#cors <- "exc" # exchangeable (compound symmetric)
#cors <- "uns" # unstructured (symmetric)--convergence problems

# load data 100 encs by pc, last 28days, for fitting
datain  <- read.csv("T:/DW Research/DW Projects/larbuckle/datain/
   dis100bypc28days.csv", header = TRUE, sep = ",")
datain$date <- as.Date(datain$date,format="%m/%d/%Y")
# convert icu and surgery into indicators
datain$icu <- factor(datain$icu)
datain$surgery <- factor(datain$surgery)

# load data 100 new encs by pc for validation
dataval <- read.csv("T:/DW Research/DW Projects/larbuckle/datain/
   dis100bypc_val.csv", header = TRUE, sep = ",")
dataval$date <- as.Date(dataval$date,format="%m/%d/%Y")
```

```
dataval$day  <- dataval$los+1 # create day covar (los starts at 0)
# convert icu and surgery into indicators
dataval$icu <- factor(dataval$icu)
dataval$surgery <- factor(dataval$surgery)

# set correlation structure (corClasses) for glmmPQL (random intercept)
if (cors=="ind") {
corr <- NULL
} else if (cors=="ar1") {
corr <- corAR1(form=~1|encWID)
} else if (cors=="exc") {
corr <- corCompSymm(form=~1|encWID)
} else if (cors=="uns") {
corr <- corSymm(form=~1|encWID)
}

# atomic vectors to collect daily c-stat and brier scores
#   for subject-specific model
cval_ss  <- NULL
brval_ss <- NULL
cval_ss2  <- NULL # random intercept at pc level only
brval_ss2 <- NULL
#   for population-averaged model
cval_pa  <- NULL
brval_pa <- NULL

# loop over each day
for (i in 1:7) {

  # create dataset for fitting: datain + dataval before today
  # (will add today, with outcome removed, later)
  datafitpre <- subset(dataval, day<i)
  datafit <- merge(datain, datafitpre, all=TRUE)

  # extract today from dataval, for prediction
  datapred <- subset(dataval, day==i)
  # extract outcomes for today, to compare with predictions
  dataout  <- datapred[,"exit24hrs"]
  # remove outcome for today (for prediction of next day discharge)
  datapred[,"exit24hrs"] <- NA
  # merge with datain for fitting
```

```
  datafit  <- merge(datafit, datapred, all=TRUE)

  # remove unused factors, if any
  datafit$icu           <- factor(datafit$icu)
  datafit$surgery       <- factor(datafit$surgery)
  datafit$charlson     <- factor(datafit$charlson)
  datafit$surgery       <- factor(datafit$surgery)
  datafit$accommodation <- factor(datafit$accommodation)
  datafit$admitType     <- factor(datafit$admitType)

  # fit using glmmPQL, removing factors with less than 2 levels
  # (model will only fit using data without NAs as outcome)
  model.ss <- glmmPQL(exit24hrs~day+LAPS+age+charlson+icu+surgery
    +nursStaOcc+accommodation+admitType,
    random=list(primaryCondition=~1, encWID=~1), family=binomial,
    data=datafit)

  # fit using logistic (for comparison)
  model.pa <- geeglm(exit24hrs~primaryCondition+day+LAPS+age+charlson
     +icu+surgery+nursStaOcc+accommodation+admitType, family=binomial,
   data=datafit, family=binomial, id=encWID, corstr="ar1")

  # calculate c-stat of validation sample
  # use plogis to go from log-odds to probability
  cval_ss[length(cval_ss)+1]   <- somers2(pp_ss, dataout)[1]
  cval_ss2[length(cval_ss2)+1] <- somers2(pp_ss2, dataout)[1]
  cval_pa[length(cval_pa)+1]   <- somers2(pp_pa, dataout)[1]

  # brier score of validation sample
  brval_ss[length(brval_ss)+1]   <- mean((pp_ss-dataout)^2)
  brval_ss2[length(brval_ss2)+1] <- mean((pp_ss2-dataout)^2)
  brval_pa[length(brval_pa)+1]   <- mean((pp_pa-dataout)^2)
}

# set working directory, for output
setwd("T:/DW Research/DW Projects/larbuckle/dataout/")

# plot c-stat for validation by day (zoom)
day <- 1:length(cval_ss)
plot(day, cval_ss, ylim=c(0.5,1), type="l", lty=1,
   xlab="Day", ylab="Concordance Probability")
```

```
lines(day, cval_ss2, lty=2)
lines(day, cval_pa, lty=3)
legend(1.5,0.9, bty="n",
c("3-level, random intercept",
        "3-level, random intercept for PC only",
        "Logistic/GEE"),
lty=c(1,2,3)
)

# plot brier score for validation by day
day <- 1:length(cval_ss)
plot(day, brval_ss, ylim=c(0,0.5), type="l", lty=1,
   xlab="Day", ylab="Brier Score")
lines(day, brval_ss2, lty=2)
lines(day, brval_pa, lty=3)
legend(1.5,0.45, bty="n",
c("3-level, random intercept",
        "3-level, random intercept for PC only",
        "Logistic/GEE"),
lty=c(1,2,3)
)
```

B.4 Time-Series Models in R

```
# modified form of tsdisplay
tsdisplaynew = function(xdata, main="") {
   op <- par(no.readonly = TRUE)
   layout(matrix(c(1,2, 1,3), nc=2))
   num <- sum(!is.na(xdata))

   #plot.ts(xdata,  main = main, ylab = "Residuals",
   #   xlab="Time (Years)")
   plot.ts(xdata,  main = main, ylab = "Admissions",
      xlab="Time (Years)")
   abline(h=0)

   alag <- 10+sqrt(num)
   ACF = acf(xdata, alag, plot=FALSE, na.action=na.pass)$acf[-1]
   LAG = 1:alag#/frequency(xdata)
   L=2/sqrt(num)
```

```
   #plot(LAG, ACF, type="h", ylim=c(min(ACF)-.05,min(1,max(ACF+.05))),
   #   main = "ACF of Residuals", xlab="Lag (Days)")
   plot(LAG, ACF, type="h", ylim=c(min(ACF)-.05,min(1,max(ACF+.05))),
      main = "ACF of Admissions", xlab="Lag (Days)")
   abline(h=c(0,-L,L), lty=c(1,2,2), col=c(1,4,4))

   #qqnorm(stdres, main="Normal Q-Q Plot of Std Residuals");
   #   qqline(stdres, col=4)

   PACF = acf(xdata, alag, plot=FALSE, na.action = na.pass,
      type="partial")$acf[-1]
   LAG = 1:(alag-1)#/frequency(xdata)
   L=2/sqrt(num)
   #plot(LAG, PACF, type="h", ylim=c(min(PACF)-.05,
   #   min(1,max(PACF+.05))), main = "PACF of Residuals",
   #   xlab="Lag (Days)")
   plot(LAG, PACF, type="h", ylim=c(min(PACF)-.05,
      min(1,max(PACF+.05))), main = "PACF of Admissions",
      xlab="Lag (Days)")
   abline(h=c(0,-L,L), lty=c(1,2,2), col=c(1,4,4))

   on.exit(par(op))
}

# modified form of tsdiag
tsdiag.sarima = function(fitit, xdata,main="Standardized Residuals") {
   p<-fitit$arma[1]
   q<-fitit$arma[2]
   P<-fitit$arma[3]
   Q<-fitit$arma[4]
   S<-fitit$arma[5]
   d<-fitit$arma[6]
   D<-fitit$arma[7]
   op <- par(no.readonly = TRUE)
   layout(matrix(c(1,2,4, 1,3,4), nc=2))

   rs <- fitit$residuals
   stdres <- rs/sqrt(fitit$sigma2)
   num <- sum(!is.na(rs))
   plot.ts(stdres,  main = main, ylab = "Std Residuals",
      xlab = "Time (Years)")
```

```
   abline(h=0)

   alag <- 10+sqrt(num)
   ACF = acf(rs, alag, plot=FALSE, na.action = na.pass)$acf[-1]
   LAG = 1:alag#/frequency(xdata)
   L=2/sqrt(num)
   plot(LAG, ACF, type="h", ylim=c(min(ACF)-.05,min(1,max(ACF+.05))),
      main = "ACF of Residuals", xlab="Lag (Days)")
   #plot(LAG, ACF, type="h", ylim=c(min(ACF)-.05,min(1,max(ACF+.05))),
   #   main = "ACF Plot", xlab="Lag")
   abline(h=c(0,-L,L), lty=c(1,2,2), col=c(1,4,4))

   qqnorm(stdres, main="Normal Q-Q Plot of Std Residuals");
   qqline(stdres, col=4)
   #qqnorm(stdres, main="Normal Q-Q Plot"); qqline(stdres, col=4)

   # calculation as used in tsdiag
   nlag <- ifelse(S<4, 20, 3*S)
   ppq <- p+q+P+Q
   pval <- numeric(nlag)
   for (i in 1:nlag) {u <- Box.test(rs, i, type = "Ljung-Box")
      $statistic
      pval[i] <- pchisq(u, i, lower=FALSE)
   }

   plot( 1:nlag, pval[1:nlag], xlab = "Lag (Days)", ylab = "p value",
      ylim = c(0,1), main = "p values for Ljung-Box statistic")
   abline(h = 0.05, lty = 2, col = "blue")

   on.exit(par(op))
}

# load packages
library(forecast)
# load tsdisplaynew.R tsdiagsarima.R (improved versions)

# set working directory, for printouts
setwd("T:/DW Research/DW Projects/larbuckle/dataout/")

# campus names
campus <- c('civic','general')
```

```
# set input data file
datafile1 <- "T:/DW Research/DW Projects/larbuckle/datain/admissions"
# set working directory, for output
setwd("T:/DW Research/DW Projects/larbuckle/dataout/")

# load data for civic
datafile <- paste(datafile1, campus[1], ".csv", sep="")
admisciv <- read.csv(datafile, header = TRUE, sep = ",")
# create ts object, starting at 01 April 2004 (day 92 of year)
admisciv <- ts(admisciv[[1]], start=c(2004,92), frequency=365)

# load data for general
datafile <- paste(datafile1, campus[2], ".csv", sep="")
admisgen <- read.csv(datafile, header = TRUE, sep = ",")
# create ts object, starting at 01 April 2004 (day 92 of year)
admisgen <- ts(admisgen[[1]], start=c(2004,92), frequency=365)

# sample up to 2008, for forecasting with sarima
admciv <- window(admisciv, end=c(2008,1))
admgen <- window(admisgen, end=c(2008,1))

# ts, acf, pacf for civic and general
pdf{"ts-acf-pacf-plots-civic.pdf")
tsdisplaynew(admciv, main="Emergency Admisssions to Civic Campus")
dev.off()

pdf{"ts-acf-pacf-plots-general.pdf")
tsdisplaynew(admgen, main="Emergency Admisssions to General Campus")
dev.off()

# seasonaly differenced series (tried various lags, 7 by far the best)
pdf("ts-diff7-acf-pacf-plots-civic.pdf")
tsdisplaynew(diff(admciv,7), main="Seasonally Differenced Series
   (Lag=7) for Civic Campus")
# over differenced; going to need ar term;
dev.off()

pdf("ts-diff7-acf-pacf-plots-general.pdf")
tsdisplaynew(diff(admgen,7), main="Seasonally Differenced Series
   (Lag=7) for General Campus")
dev.off()
```

```
# over differenced; going to need ar term;

ar(diff(admciv,7))
# order 29 => seasonal effect still present
ar(diff(admgen,7))

# calculate correlation without drift
x <- admciv[1:length(admciv)-1]
y <- admciv[2:length(admciv)]
lsfit(x,y,intercept=FALSE)$coeff
# calculate drift (slope)
lsfit(x,y)$coeff # 0.2401688 for civic, 0.1791178 for general

### build civic sarima model (box-jenkins method)
# xreg used to account for drift
tsdiag.sarima(arima(admciv, order=c(1,0,0),
   seasonal=list(order=c(0,1,0), period=7),
   xreg=1:length(admciv)), admciv)
# portmanteau test fails
# negative acf for seasonal period => add sma term
# try without seasonal differencing

tsdiag.sarima(arima(admciv, order=c(1,0,0),
   seasonal=list(order=c(0,0,1), period=7),
   xreg=1:length(admciv)), admciv)
# portmanteau test fails
# seasonal structure in residuals => add seasonal differening

tsdiag.sarima(arima(admciv, order=c(1,0,0),
   seasonal=list(order=c(0,1,1), period=7),
   xreg=1:length(admciv)), admciv)
# portmanteau test fails at lag 2
# postive acf terms (decreasing) => add ma term

tsdiag.sarima(arima(admciv, order=c(1,0,1),
   seasonal=list(order=c(0,1,1), period=7),
   xreg=1:length(admciv)), admciv)
# portmanteau test fails
# try another ma term

tsdiag.sarima(arima(admciv, order=c(1,0,2),
```

```
   seasonal=list(order=c(0,1,1), period=7),
   xreg=1:length(admciv)), admciv)
# fits well, according to portmanteau

### final civic sarima model
mysarimadciv <- arima(admciv, order=c(1,0,2),
   seasonal=list(order=c(0,1,1),period=7), xreg=1:length(admciv))

### build general sarima model (box-jenkins method)
# try the same model
tsdiag.sarima(arima(admgen, order=c(1,0,2),
   seasonal=list(order=c(0,1,1), period=7),
   xreg=1:length(admgen)), admgen)
# fits well, according to portmanteau

### final sarima model for general
mysarimadgen <- arima(admgen,order=c(1,0,2),
   seasonal=list(order=c(0,1,1), period=7), xreg=1:length(admgen))

# plot diagnostics
pdf{"ts-sarima102011-drift-civic-residuals.pdf")
tsdiag.sarima(mysarimadciv, admciv,
   main="Standardized Residuals of SARIMA(1,0,2)(0,1,1)[7] for
   Civic Campus")
dev.off()

pdf{"ts-sarima102011-drift-general-residuals.pdf")
tsdiag.sarima(mysarimadgen, admgen,
   main="Standardized Residuals of SARIMA(1,0,2)(0,1,1)[7] for
   General Campus")
dev.off()

shapiro.test(mysarimadciv$res)
# p-value = 0.02426 > 0.01 => can't reject normality
shapiro.test(mysarimadgen$res)
# p-value = 0.1587  > 0.01 => can't reject normality

# print summaries
summary(mysarimadciv)
summary(mysarimadgen)
```

```
# forecasts
pdf("ts-sarima102011-drift-civic-general-forecasts-100days.pdf")
layout(matrix(c(1,2, 1,2),nc=2))

plot(forecast(mysarimadciv, include=0, shadecols=c(3,8),
   xreg=(length(admciv)+1):(length(admciv)+101)),
   main="Forecasts from SARIMA(1,0,2)(0,1,1)[7] for Civic Campus",
   xlab="Time (Years)", ylab="Admissions")
lines(window(admisciv, start=c(2008,2), end=c(2008,101)), lty=2)

plot(forecast(mysarimadgen, include=0, shadecols=c(3,8),
   xreg=(length(admgen)+1):(length(admgen)+101)),
   main="Forecasts from SARIMA(1,0,2)(0,1,1)[7] for General Campus",
   xlab="Time (Years)",ylab="Admissions")
lines(window(admisgen,start=c(2008,2),end=c(2008,101)),lty=2)

par(mfrow=c(1,1))
dev.off()

# out-of-sample accuracy for civic
start <-length(admciv)+1
step <-10
stop <- start+step

# create atomic vectors
mapec <- NULL
masec <- NULL

# number of iterations
its<-floor((length(admisciv)-start)/step)
for (i in 1:its) {
   acc  <- accuracy(forecast(mysarimadciv, xreg=(start:stop)),
      admisciv[start:stop])
   mapec[i] <- acc[['MAPE']]
   masec[i] <- acc[['MASE']]

   stop  <- stop+step
}
mapec[its] # 6.79
masec[its] # 0.725
```

```
# out-of-sample accuracy for general
start <-length(admgen)+1
step <-10
stop <- start+step

# create atomic vectors
mapeg <- NULL
maseg <- NULL

# number of iterations
its<-floor((length(admisgen)-start)/step)
for (i in 1:its) {
   acc  <- accuracy(forecast(mysarimadgen, xreg=(start:stop)),
      admisgen[start:stop])
   mapeg[i] <- acc[['MAPE']]
   maseg[i] <- acc[['MASE']]

   stop  <- stop+step
}
mapeg[its] # 6.13
maseg[its] # 0.685

# plot out-of-sample accuracy
pdf("ts-sarima102011-drift-civic-general-forecast-accuracy.pdf")
time <- seq(from=2008, by=10/356, length.out=its)
layout(matrix(c(1,2, 1,2),nc=2))

plot(time, mapec, type='l', ylim=c(0,10), main="Forecast Accuracy
   from SARIMA(1,0,2)(0,1,1)[7] Using MAPE",
   xlab="Time (Years)", ylab="Mean Absolute Percentage Error (%)")
lines(time, mapeg,lty=2)
legend(2008.9,3,"Civic Campus",bty='n',lty=1)
legend(2008.9,2,"General Campus",bty='n',lty=2)

plot(time, masec, type='l', ylim=c(0,1), main="Forecast Accuracy
   from SARIMA(1,0,2)(0,1,1)[7] Using MASE",
   xlab="Time (Years)", ylab="Mean Absolute Scaled Error")
lines(time, maseg,lty=2)
legend(2008.9,.3,"Civic Campus",bty='n',lty=1)
legend(2008.9,.2,"General Campus",bty='n',lty=2)
```

```
layout(matrix(c(1,1, 1,1),nc=2))
dev.off()

# MAPE: If all data are positive and much greater than zero,
# then MAPE may still be preferred for reasons of simplicity

# MASE: A scaled error is less than one if it arises from a better
# forecast than the average one-step naive (random walk) forecast
# computed in-sample.
```

Bibliography

[1] G. Abraham, G.B. Byrnes, and C.A. Bain , Short-Term Forecasting of Emergency Inpatient Flow, *IEEE Transactions on Information Technology in Biomedicine*, **3**, No. 3, pp. 380–388, 2009.

[2] A. Agresti, **Categorical Data Analysis**, 2nd Ed., Wiley Series in Probability and Statistics, Wiley-Interscience, Hoboken, New Jersey, USA, 2002.

[3] P.D. Allison, **Logistic Regression Using the SAS System: Theory and Application**, SAS Institute, Cary, NC, USA, 1999.

[4] P.D. Allison, Convergence Failures in Logistic Regression, *Proceedings of the 2008 SAS Global Forum*, Paper 360-2008, 2008.

[5] R. Amarasingham, L. Plantinga, M. Diener-West, D.J. Gaskin, and N.R. Powe, Clinical Information Technologies and Inpatient Outcomes, *Archives of Internal Medicine*, **169**, No. 2, pp. 108–114, 2009.

[6] H.J. Anderson, EHR Pioneers Try to Stay Out Front, *Health Data Management*, **15**, No. 5, pp. 26+, 2007.

[7] G. Box, and G. Jenkins, **Time Series Analysis: Forecasting and Control**, Holden-Day, San Francisco, USA, 1970.

[8] N.E. Breslow, and D.G. Clayton, Approximate Inference in Generalized Linear Mixed Models, *Journal of the American Statistical Association*, **88**, No. 421, 1993.

[9] C. Chai, and N. Haesler, Surgical Wait Times Increase in Canada, *Ottawa Citizen*, December 6, 2010.

[10] M.E. Charlson, P. Pompei, K.L. Ales, and C.R. MacKenzie, A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation, *Journal of Chronic Diseases*, **40**, No. 5, pp. 373-383, 1987.

[11] Canadian Health Services Research Foundation, *The Digital Warehouse: An Improvement Story*, 2008.

[12] G.M. Cordeiro, On Pearson's Residuals in Generalized Linear Models, *Statistics & Probability Letters*, **66**, No. 3, pp. 213–219, 2004.

[13] R. Deonandan, K. Campbell, T. Ostbye, I. Tummon, and J. Robertson, A Comparison of Methods for Measuring Socio-economic Status by Occupation or Postal Area, *Chronic Diseases in Canada*, **21**, No. 3, pp. 114–118, 2000.

[14] K. El Emam, and L. Gaudette, Canadian REB Wizard, Privacy Analytics Inc., 2010, visited July 2011, ¡http://www.ehealthinformation.ca/rebwizard/ca/¿.

[15] G.J. Escobar, J.D. Greene, P. Scheirer, M.N. Gardner, D. Draper, and P. Kipnis, Risk-Adjusting Hospital Inpatient Mortality Using Automated Inpatient, Outpatient, and Laboratory Databases, *Medical Care*, **46**, No. 3, pp. 232–239, 2008.

[16] A. Forster, I. Stiell, G. Wells A. Lee, and C. van Walraven, The Effect of Hospital Occupancy on Emergency Department Length of Stay and Patient Disposition, *Academic Emergency Medicine*, **10**, No. 2, pp. 127–133, 2003.

[17] A. Gelman, Multilevel (Hierarchical) Modeling: What It Can and Cannot Do, *Technometrics*, **48**, No. 3, pp. 241–251, 2006.

[18] A. Gelman, and J. Hill, **Data Analysis Using Regression and Multilevel/Hierarchical Models**, Analytical Methods for Social Research, Cambridge University Press, New York, USA, 2007.

[19] A.S. Goldberger, Best Linear Unbiased Prediction in the Generalized Linear Regression Model, *Journal of the American Statistical Association*, **57**, No. 298, pp. 369–375, 1962.

[20] F.E. Harrell, **Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis**, Springer Series in Statistics, Springer Verlag, New York, USA, 2001.

[21] W. Harville, Maximum Likelihood Approaches to Variance Components Estimation and to Related Problems, *Journal of the American Statistical Association*, **72**, No. 358, pp. 320–338, 1977.

[22] C.R. Henderson, Best Linear Unbiased Estimation and Prediction Under a Selection Model, *Biometrics*, **31**, pp. 423–447, 1975.

[23] R.J. Hyndman, and A.B. Koehler, Another Look at Measures of Forecast Accuracy, *International Journal of Forecasting*, **22**, No. 4, pp.679–688, 2006.

[24] J. Neter, M. Kutner, C. Nachtsheim, and W Wasserman , **Applied Linear Statistical Models**, 4th Ed., Irwin, Chicago, USA, 1996.

[25] Y. Lee, and J.A. Nelder, Likelihood Inference for Models with Unobservables: Another View, *Statistical Science*, **24**, pp. 255–269, 2009.

[26] K.Y. Liang, and S.L. Zeger, Longitudinal Data Analysis Using Generalized Linear Models, *Biometrika*, **73**, No. 1, pp. 13–22, 1986.

[27] R.C. Littell, G.M. Milliken, W.W. Stroup, R.D. Wolfinger, and O. Schabenberger, **SAS For Mixed Models**, 2nd ed., SAS Institute, Cary, NC, USA, 2006.

[28] S.J. Littig, and M.W. Isken, Short term hospital occupancy prediction, *Health Care Management Science*, **10**, pp. 47–66, 2007.

[29] G.M. Ljung, and G.E.P. Box, On a measure of lack of fit in time series models, *Biometrika*, **65**, No. 2, pp. 297–303, 1978.

[30] H.B. Mann, and D.R. Whitney, On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other, *Annals of Mathematical Statistics*, **18**, No. 1, pp. 50–60, 1947.

[31] P. McCullagh, and J.A. Nelder, **Generalized Linear Models**, Monographs on Statistics and Applied Probability 37, 2nd ed., Chapman & Hall, London, UK, 1989.

[32] C.A. McGilchrist, Estimation in Generalized Mixed Models. *Journal of the Royal Statistical Society: Series B*, **56**, pp. 61–89, 1994.

[33] G. Molenberghs, and G. Verbeke, **Models for Discrete Longitudinal Data**, Springer Series in Statistics, Springer, New York, USA, 2005.

[34] Quality Indicators, 2010-2011 Highlights, The Ottawa Hospital 2010-2011 Annual Report, visited July 2011, ¡http://www.worldclasscare.ca/en/highlights/quality-indicators.hml¿.

[35] N.J.D. Nagelkerke, A Note on a General Definition of the Coefficient of Determination, *Biometrika*, **78**, 3, pp. 691–692, 1991.

[36] H.D. Patterson, and R. Thompson, Maximum Likelihood Estimation of Components of Variance, *Proceedings Eighth International Biometric Conference*, Biometric Society, pp. 197–207, 1975.

[37] J. Peabody, J. Luck, S. Jain, D. Bertenthal, and P. Glassman, Assessing the Accuracy of Administrative Data in Health Information Systems, *Medical Care*, **42**, No. 11, pp. 1066–1072, 2004.

[38] H.D. Patterson, and R. Thompson, Recovery of Inter-Block Information when Block Sizes are Unequal, *Biometrika*, **58**, No. 3, pp. 545–554, 1971.

[39] A.P. Pierce, and D.W. Schafer, Residuals in Generalized Linear Models, *Journal of the American Statistical Association*, **81**, No. 396, pp. 977–986, 1986.

[40] G.K. Robinson, That BLUP Is a Good Thing: The Estimation of Random Effects, *Statistical Science*, **6**, pp. 15–51, 1991.

[41] K. Rufibach, Use of Brier Score to Assess Binary Predictions, *Journal of Clinical Epidemiology*, **63**, 8, pp. 938–939, 2010.

[42] E.S. Shtatland, K. Kleinman, and E.M. Cain, A New Strategy of Model Building in Proc Logistic With Automatic Variable Selection, *Proceedings of the 29th Annual SAS Users Group International Conference*, Paper 191-29, 2004.

[43] A. Sen, and M. Srivastava, **Regression Analysis: Theory, Methods, and Applications**, Springer Texts in Statistics, Springer Verlag, New York, USA, 1990.

[44] T. Sing, O. Sander, N. Beerenwikel, and T. Lengauer, ROCR: Visualizing Classifier Performance in R, *Bioinformatics*, **21**, No. 20, pp. 3940–3941, 2005.

[45] T.A.B. Snijders, and R.J. Bosker, **Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling**, Sage Publications, London, UK, 1999.

[46] D. Tandberg, C. Qualls, Time Series Forecasts of Emergency Department Patient Volume, Length of Stay, and Acuity, *Annals of Emergency Medicine*, **23**, No. 2, pp. 299–306, 1994.

[47] J.W.R. Twisk, **Applied Multilevel Analysis**, Practical Guides to Biostatistics and Epidemiology, Cambridge University Press, Cambridge, UK, 2006.

[48] W.N. Venables, and B.D. Ripley, **Modern Applied Statistics with S**, Statistics and Computing, 4th Ed., Springer, New York, USA, 2002.

[49] G. Verbeke, and G. Molenberghs, **Linear Mixed Models for Longitudinal Data**, Springer Series in Statistics, Springer Verlag, New York, USA, 2000.

[50] R.W.M. Wedderburn, Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method, *Biometrika*, **61**, No. 3, pp. 439–447, 1974.

[51] B.T. West, K.B. Welch, and A.T. Galecki, **Linear Mixed Models: A Practical Guide Using Statistical Software**, Taylor & Francis Group, Boca Raton, Florida, USA, 2007.

[52] F. Wilcoxon, Individual Comparisons by Ranking Methods, *Biometrics Bulletin*, **1**, No. 6, pp. 80–83, 1945.

[53] S.L. Zeger, and K.Y. Liang, Longitudinal Data Analysis for Discrete and Continuous Outcomes, *Biometrics*, **42**, No. 1, pp. 121-130, 1986.