3-2006

Analysis of Patient Information: An Empirical Modeling Approach

Tony A. Murphy

ANALYSIS OF PATIENT INFORMATION:

AN EMPIRICAL MODELING APPROACH

THESIS

Tony A. Murphy, Captain, USAF

AFIT/GOR/ENS/06-14

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government.

AFIT/GOR/ENS/06-14

ANALYSIS OF PATIENT INFORMATION:
AN EMPIRICAL MODELING APPROACH

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Tony A. Murphy, B.S.

Captain, USAF

March 2006

AFIT/GOR/ENS/06-14

ANALYSIS OF PATIENT INFORMATION:

AN EMPIRICAL MODELING APPROACH

Tony A. Murphy, BS

Captain, USAF

Approved:

Dr. Sharif H. Melouk Thesis Advisor	Date
Dr. Marcus B. Perry Reader	Date

AFIT/GOR/ENS/06-14

Abstract

With rising costs and increasing complexities, many hospitals seek to better understand the intricate details of their operations. Increasingly, these organizations have a strong desire to accurately predict the resources required to effectively treat their patient load. This research investigates patient length-of-stay in a hospital neurological unit using an empirical modeling approach. Factors significantly affecting patient length of stay were identified and used to construct a regression model. The predictive model provides hospital decision makers with a compact tool to input what-if scenarios and predict future patient treatment lengths, thus, allowing the hospital to properly allocate resources.

Acknowledgments

To Dr Melouk and Dr Perry, thanks for all of your guidance, advice, and insight.

To my former supervisors, Master Sergeants Ertl, Bradbury, and Brown, thanks for giving me the opportunity to better myself by furthering my education. I would not be where I am today without you all.

To Mom and Dad, thanks for all of your love and support and for making me the person I am today.

Table of Contents

Page

List of Figures

Page

List of Tables

Page

ANALYSIS OF PATIENT INFORMATION:

AN EMPIRICAL MODELING APPROACH

I. Introduction

1.1 Overview

Currently, there is a growing need and interest to improve the quality and efficiency of health care, which emphasizes the necessity of good indicators to determine this. Furthermore, hospitals and other health care facilities are experiencing difficulties providing reliable care while maintaining sizable profits. As a result, many are looking at innovative ways to better utilize the often limited resources they have. Several have determined the best way to accomplish this is to better understand and ultimately exercise better control over their operations. Because it is often related to total costs, managing patient length-of-stay (LOS) has become the main focus in health care settings and many hospitals have turned to LOS prediction models to improve their resource utilization [Blais 2003, Omachonu 2004].

Hospitals have a strong desire to be able to accurately predict LOS and, more specifically, costs by using the often limited information from incoming patients. LOS is becoming an increasingly important and primary performance indicator for hospitals and other health care facilities and has often been used as an indicator of inpatient care efficiency [Blais 2003]. Due to its clear meaning as one of the main sources of hospital costs and because it can also be used as an indicator of quality patient care, it is often viewed as a measure of effective treatment and as a means for managing cost. Because of this, hospital administrators generally focus on LOS as an indicator of the quality of care and as an important variable in determining budgets based on access (i.e., admissions to the unit per year). As a result, many hospitals have dedicated countless efforts and resources to better understand what ultimately determines LOS [Omachonu 2004].

Ideally, these facilities would wish to rely on predetermined expected LOS values, which are calculated based on statistical patient data from the previous fiscal year [Centers for Medicare and Medicaid Services 2002]. However, this approach does not consider how these expected values may be impacted by certain patient characteristics or other factors. In most cases, a younger, healthier patient with little or no medical history is going to have a shorter LOS than a frailer, elderly one who has had prior related diagnoses. However, the current system makes no adjustment in these instances. Consequently, these expected LOS values tend to have excessive variability which obviously makes them highly unreliable.

An alternative is to model the current hospital operation, or system, to more accurately predict an incoming patient's LOS. These prediction models generally use historical patient data to determine the predominant factors and indicators that drive LOS and costs. While patient characteristics are the obvious predictors to use, clinical aspects such as physician experience and hospital policies have also been used.

Multiple regression analysis is typically used to build these models. After validation testing, these models can be analyzed using simulation software, which will derive statistical operating characteristics, such as mean LOS and patient load, of the entire hospital system as a whole. Ultimately, hospital administrators will have a model that not only characterizes their entire operation, but also has the capability to analyze various what-if scenarios affecting resource utilization.

1.2 Problem Objective

Using neurological patient data provided by a local hospital, a prediction model will be created to determine patient LOS and an indirect associated cost of treatment. Using this model, a hospital could match their performance compared to the government standard and consider dropping more costly treatments in exchange for more profitable ones. They could also allocate and schedule resources, such as beds and staff, accordingly. Other items the model may be used to determine:

- Are there certain patient profiles that are prone to have longer/shorter stays compared to the expected LOS?
- What, if any, insurance providers encourage a shorter stay compared to other providers?
- Are there certain profiles the hospital should target to reduce LOS and/or increase profit?

1.3 Method of Approach

1.3.1 Problem Background

Good Samaritan Hospital of Dayton, OH, is one of the leading health care facilities in southwestern Ohio. It is currently performing a major renovation to its facilities with its neurology department being expanded as a result of this. With this upgrade, the hospital will become a leading certified stroke care unit in the area and is expected to see a dramatic increase in not only stroke patients, but neurosurgical and neuromedical cases as well. Also, another neurosurgeon is expected to be hired in the near future, further increasing the influx of patients. Because of these changes, the hospital feels it needs a better understanding of its operations. They also desire to know the consequences of varying resource allocations, such as beds and staffing.

1.3.2 Overview of Neurology and Neurological Disorders

Neurology is the branch of medicine that deals with the central nervous system and its disorders. Neurological disorders are those affecting the central nervous system (brain, brainstem, cerebellum), the peripheral nervous system (peripheral nerves - cranial nerves), or the autonomic nervous system (parts of which are contained in both of the former). While stroke is the most common neurological ailment, other disorders include migraine headaches, epileptic seizures, and cerebral palsy [National Institutes of Health 2006].

Stroke is defined as rapidly developed clinical signs of focal (or global) disturbance of cerebral function, lasting more than 24 hours or leading to death. Though its cause has yet to be determined, it is usually due to a blockage of an artery in the brain, called a cerebral infraction. In minor instances, stroke results from a cerebral hemorrhage, or rupture of a blood vessel in the brain. More information on neurology and all neurological disorders can be found at the website of the National Institute of Neurological Disorders and Stroke: http://www.ninds.nih.gov [National Institutes of Health 2006].

1.3.3 Patient Classification (Diagnostic Related Groups)

Hospitals classify medical patients into clinically cohesive groups, called Diagnostic Related Groups (DRGs), which typically consume the same amount and type of hospital resources. The DRG system was implemented in 1983 when Congress mandated a national hospital prospective payment system (PPS) for all Medicare patients. The PPS utilizes DRGs to determine hospital reimbursement. The Centers for Medicare and Medicaid Services (CMS) administers the PPS and issues all rules and changes with regard to DRGs. Although the DRG system was originally intended exclusively for Medicare patients, many hospitals now use it as a means to charge all their patients. However, the actual payment of non-Medicare patients varies according to their individual insurance providers [Centers for Medicare and Medicaid Services 2002].

The DRG system is used to predict costs of treatment and length of stay. It is a tool that reflects the severity of the diagnosed ailment and difficulty of treatment and is expected to indicate an efficient payment mechanism for health care. A basic assumption is that a patient in the same DRG will require similar resources regardless of the type or location of the hospital. Relative weights (RW) are assigned to each DRG to indicate the expected resource consumption, based on historical data, required to treat a certain ailment. A high RW indicates the case requires a high level of hospital resources, and in most cases, a longer LOS. Base rate is the amount of money paid to the hospital per unit of RW. The formula for computing the total hospital payment for each DRG is then:

$$\text{DRG Relative Weight} \times \text{Hospital Base Rate} = \text{Payment to Hospital} \qquad (1\text{-}1)$$

For example, in 2005 the RW for a stroke was 1.2719 and the hospital base rate was $4500, which results in a payment of $5,723.55 (1.2719 * $4500). Each DRG also has an expected LOS value, which is partially used to determine the RW. In other words, the longer the LOS, the higher the charge will generally be. As stated earlier, the drawback of the RW and expected LOS values is that they give no indication of how they may be impacted by certain patient characteristics or other factors. If a hospital knew certain scenarios that dramatically affected LOS, they potentially could plan accordingly to take advantage of these situations. For example, a patient profile prone to have a much longer LOS than the expected value would most likely be avoided, if possible. Moreover, if certain diagnoses consistently result in a longer than expected LOS, the hospital may consider discontinuing these particular treatments [Centers for Medicare and Medicaid Services 2002].

1.3.4 Approach

The model will be implemented using a variety of multivariate linear regression analysis techniques. Since there is a considerable amount of data, software will be used extensively to accurately and efficiently handle the storage, sorting, and analysis of the data. All 2^k possible regressions will be studied, where k is the number of possible regressors, along with all the various interactions between them. Any categorical data will be quantified either arbitrarily or based on some type of ranking method.

Ultimately, all statistically significant predictor variables will be determined to develop a final regression model. With this representation, a simulation model describing the neurological unit will be implemented. Simulation software will be used to validate and verify the regression model before performing full-scale testing and sensitivity analysis with it.

1.4 Thesis Outline

This thesis consists of five chapters. Chapter 2 reviews related research in predicting and modeling LOS that is relevant to this study. Chapter 3 explains the methodology used in the development and testing of the regression and simulation models. Chapter 4 presents the results of the analysis. Chapter 5 presents insight and conclusions, based on this research, and makes recommendations for further study.

II. Literature Review

2.1 Overview

There has been broad and extensive research performed in determining what variables impact, and to what extent, patient LOS. Not only have various methods of regression analysis been used in these studies, there has also been a plethora of possible predictor variables studied. In fact, it was determined that no less than 22 different variables influenced LOS in various studies. While objective and quantitative indicators such as diagnosis and age were predominantly used, more subjective and qualitative types, such as patient severity, physician characteristics, and the patient's opinion of his or her overall health, were also commonly studied.

The objective of this review is not only to present relevant work related to this research, but also to study and understand the methods and techniques previously used for predicting LOS in order to possibly relate it to this analysis. While the primary focus of this review is on work pertaining specifically with LOS, those dealing with predicting patient costs are also examined. This was done primarily because the nature of most hospital billing policies tends to relate health care operating costs to LOS.

The remainder of this chapter examines relevant research in analyzing and predicting LOS relating to neurological cases (Section 2.2), followed by a review of other research relevant to predicting LOS (Section 2.3) and an overview of some of the data analysis methods and techniques used in these studies (Section 2.4). The chapter concludes with an overview of software applications used in this research (Section 2.5).

2.2 Related Work in Predicting LOS for Neurological Cases

Much of the work done in modeling LOS in neurological units has focused on stroke cases as they are the most common neurological ailment and typically require the most resources. Strokes are the third leading cause of death in the US among people aged 65 and older and are also a major cause of long-term disability and rehabilitation. Its cost to Medicare alone has been estimated to be as high as $18 billion per year [Monane 1996].

Many researchers [Bohannon et. al. 2002, Evers et. al. 2002, Hakim and Bakheit 1998 Herman et. al. 1984, Lee et al. 1997, Monane et. al. 1996, Wade and Langton 1985] have examined patient demographics, such as age, race, gender, and marital status, in an effort to determine if any are potential predictors of LOS in neurological cases. Several others [Brousseau et. al. 1996, Evers et. al. 2002, Hakim and Bakheit 1998, Herman et. al. 1984, Monane et. al. 1996, Parsons et. al. 2003, Wade and Langton 1985] have looked at other indicators, such as insurance status, history of hospitalization, physician experience and additional diagnoses and complications the patient may or may not have had.

A study performed at University Hospital Maastricht in the Netherlands dealt with 731 stroke patients over the period of 1996-1998. The hospital had recently implemented a DRG classification system similar to the US version and wanted to determine whether it provided an accurate prediction of the variance of costs in stroke patients. The results showed that DRGs accounted for 34% of the cost variance. Adding other variables, such as age, gender, and additional diagnoses the patient had, increased R^2 to over 61% [Evers et. al. 2002]

Another study [Monane et. al. 1996] looked at insurance data from 745 elderly stroke patients and divided them into three categories: Medicare, HMOs, and other (including Medicaid and private providers). It concluded that those belonging to HMOs tended to have a shorter LOS, although there was evidence that many of these patients may have been discharged to a rehabilitation unit sooner and more often than patients belonging to other insurance providers. Two other studies analyzed the relationship between insurance type and hospital utilization. One study [Lagoe and Lauko 1995] found no significant relationship, while the other [Lanska 1994] found that LOS is indeed related to insurance type, primarily those patients that belong to Medicare or a HMO.

The majority of studies have investigated the relationships between age and hospital costs, but the results have been at best, contradictory. One study [Brousseau et. al. 1996] reported that elderly patients require higher costs, concluding that the recovery of older stroke patients is longer, while other studies [Lee et. al. 1997, Wade and Langton 1985] determined that younger patients utilized hospital services more. Lee explained this by proposing that older patients receive a less aggressive approach to treatment and tend to expire at a higher rate and earlier in the treatment phase. He also presented evidence that some hospitals ‘pad’ their LOS statistics by discharging their terminally ill patients to Hospices and other similar facilities.

While four studies [Burns and Douglas 1991, Hakim and Bakheit 1998, Monane et. al. 1996, Wade and Langton 1985] examined gender relationships, only one found a significant difference between men and women. Burns determined that men of all ages generally have a longer LOS. Out of four studies that investigated marital status [Christina et. al. 1991, Herman et. al. 1984, Monane et. al. 1996, Wade and Langton

1985], two [Monane et. al. 1996, Wade and Langton 1985] concluded that being single led to a longer hospital stay. Several authors [Burns and Douglas 1991, Christina et. al. 1991, Hakim and Bakheit 1984, Wade and Langton 1985] examined physician characteristics and found that the more experienced doctors [Burns and Douglas 1991] and those with a more general background [Christina et. al. 1991] tended to keep their patients in the hospital longer. Hakim also concluded that LOS was generally shorter in smaller hospitals within metropolitan areas and also in hospitals with a high nurse to bed ratio.

There have also been several scales developed to better, albeit subjectively, describe a patient's condition or functional level. The Rankin scale, developed by a group of neurologists, estimates a patient's level of functioning before the stroke from 0 'no symptoms' to 5 'severe handicap'. The Canadian Neurological Scale ranges from 1.5 'severely handicapped' to 10 'no symptoms'. [Evers et. al. 2002] found that men who score more severely on the Rankin scale induce significantly higher costs than do women with the same scale level.

Other studies also attempted to quantify patient severity levels. In his research, Bohannon [2002] utilized two scales: the National Institutes of Health Stroke Scale and Barthel ADL (Activity of Daily Living) Index score. The Barthel Index consists of two parts, pre-stroke and post-admission, and is used to gauge how well the patient independently performs 10 activities (e.g. dressing, walking). Bohannon used these two indexes, along with age and gender, to predict LOS, total patient charges, and discharge destination. Analysis showed that once the post-admission Barthel score was obtained, no other variable contributed to LOS prediction. Wade and Langton [1985] also utilized the Barthel Index but found very little relevance while using it.

2.3 Related Work in Predicting LOS in Non-Neurological Cases

Additional studies non-related to neurology were also analyzed. It was believed further insight into predicting LOS could be gained from these while at the same time avoiding ‘tunnel-vision’ from focusing too strictly on stroke cases. This section centers more on the techniques used to gain enhanced insight into patient characteristics and modeling of the treatment process itself.

A study [Omachonu et. al. 2004] of about 1500 Medicare patients at the University of Miami Medical Center investigated patient characteristics and clinical indicators for their top 5 DRGs (1, 127, 430, 462, 489) according to patient volume. They determined that approximately 60% of LOS variance is explained by patient characteristics and diagnosis. For DRG 1(Craniotomy, age > 17, except for trauma), patients admitted through the emergency room tended to have a longer LOS than non-emergency patients, while married patients have a shorter LOS than unmarried patients. For DRG 127 (Heart failure and shock), results indicate that being male, American Indian, Cuban, Hispanic, or Caucasian would have an increased LOS of 0.72, 0.48, 0.54, 0.97, and 0.57 days, respectively. For DRG 430 (Psychoses), older, white male patients, who at some point during their stay were transferred within the hospital, will generally stay longer.

Another study [Parsons et. al. 2002] that tested patients with respiratory problems developed a pre-admittance questionnaire consisting of 30 items. The answers were used to generate values for five patient functional scales: physical, role, emotional, social, and cognitive. These were then used to derive a global QL (Quality of Life) score. They also measured the patient's 6-minute walking distance (6MWD) and administered other physical tests that measured lung capacity and function. The results showed that for patients experiencing fewer complications, QL and 6MWD were the strongest predictors of LOS.

Still another study [Weingarten et. al. 1997] evaluated the relationship between socioeconomic status and hospital resource utilization as measured by LOS for elderly Medicare patients, age 65 and older, within Shelby County, Tennessee. Variations in length of stay were compared across income groupings for seven different Diagnosis Related Groups (DRGs) and relative effects are measured for socioeconomic status, age, race, gender, discharge status, and severity of illness. Despite the lack of provider specific and patient specific information, the analysis does suggest that, once patients access the medical care system, socioeconomic status has a limited effect on discharge decisions. The results also indicate that the effect of administratively necessary days on LOS needs further policy review.

In a large university hospital in Canada, one study [Keefler et. al. 2001] was done to examine the effects of psychosocial problems on LOS, controlling for patient demographics and medical condition. Mean LOS for DRGs were used as a response for severity of medical condition, and a subjective classification system called Person-in-Environment (PIE) was used to measure psychosocial problems. Data were collected on a sample of 160 patients: 78 in psychiatry and 82 in medical/surgical wards. In a regression analysis, the severity of the patient's psychosocial problem was a more significant predictor of LOS than the DRG variable. The identification of psychosocial problems and their severity add an important dimension to research into the effectiveness of social workers in reducing length of stay. Health providers found patients having significantly more problems related to their social role functioning tended to have a longer LOS than patients with problems in the hospital environment.

2.4 Data Analysis Methods and Techniques

Since most researchers tend to work with large quantities, many choose to use a top-to-bottom approach in analyzing their data. Once they determine exactly what indicators they were going to study and the techniques they are to use, they analyze the data using advanced software packages.

Some of the common software used was SAS (Statistical Analysis System) (Ver 6.12) [Monane et. al. 1996, Omachonu et. al. 2004], of SAS Institute Inc in Cary, NC, Statview II (Ver 1.03) [Inouye 2001], a statistical program developed by Abacus Concept in Berkeley, CA, and SPSS (Statistical Package for the Social Sciences) (Ver 10.0) [Bohannon et. al. 2002, Brousseau et. al. 1996, McKenna et. al. 2002] software, created by SPSS Inc. of Chicago, IL.

For simplicity, data was often grouped into bins, although the approaches vary considerably. For example, Omachonu et. al. [2004] divided their age data into 5-year increments while McKenna et. al. [2002] and Monane et. al. [1996] had only two age groups (<75 and >=75). Furthermore, while most chose to make LOS continuous, Monane divided his LOS response into 3 groups (1-5 days, 6-10 days, and >10 days).

2.5 Overview of Software Applications

2.5.1 MATLAB (Matrix Laboratory)

MATLAB v7.1 is a high-level programming language that uses matrix-based calculations and techniques to solve complex numerical problems. It utilizes high-quality graphics and also provides a convenient interface to built-in state-of-the-art subroutine libraries. It also has an interactive interface, reliable algorithmic foundation, and a fully extensible computing environment. For more information concerning MATLAB software, go to: http://www.mathworks.com/products/matlab/

2.5.2 ARENA

ARENA is a high-level graphical simulation language that uses hierarchical models to simulate complex real-world systems. Results are used to better understand the process(es) and to assist in making more informed and educated decisions relating to its operation. For more information concerning ARENA software, go to: http://www.arenasimulation.com/

III. Methodology

3.1 Overview

Using neurological patient data provided by Good Samaritan Hospital (GSH), an empirical model was developed using advanced linear regression analysis techniques. The data includes the following information for each patient: Relative Weight (RW), Geometric Mean Length-Of-Stay (GMLOS), and Arithmetic Mean Length-Of-Stay (AMLOS), age, gender, and Insurance Provider (IP). Details for each of these indicators is presented in the following section. The objective was to determine which, if any, of these patient indicators significantly impact LOS. This model was then compared to the output from a simulation model of the neurology unit along with general statistical information from the raw patient data, i.e. minimum/mean/maximum LOS, minimum/mean/maximum patient level. GSH could ultimately use this regression model to determine a reasonably accurate expected LOS for an incoming patient based on his or her personal data. Administrators will also be able to analyze various what-if scenarios relating to the operation of the neurology unit which will ultimately lead to better informed decisions concerning resources such as beds and staffing.

3.2 Background of Patient Information

The patient data was provided by GSH's finance office and consisted of 7319 in-residence patients treated for neurological symptoms from January 2002 to June 2005. The contents of the patient data are discussed in the following subsections.

3.2.1 Patient Age and Gender

The patient age is given in years and has a mean of 64.35, a standard deviation of 18.03, and a minimum and maximum of 14 and 105, respectively. There are 4174 (57.03%) females and 3145 (42.97%) males.

3.2.2 Patient Diagnosis and corresponding Relative Weight, GMLOS, and AMLOS

Shortly after being admitted, each patient is given a principal diagnosis code, based on his or her ailment(s). This code, along with other related codes, is assigned to one and only one DRG. As stated earlier in Chapter 1, each DRG has associated RW and LOS values. These LOS values consist of AMLOS, which is generally used as an expected LOS based on similar historical cases requiring the same type of treatment, and GMLOS, which is indirectly used to determine the actual payment. On average, AMLOS is 30% higher than GMLOS. Because RW, AMLOS, and GMLOS are updated at the beginning of every fiscal year, the model will have to be updated annually to reflect these changes. Note that there are 323 RWs, even though there are only 115 associated DRGs in this study. This is because RWs for most DRGs vary annually due largely to inflation and varying treatment costs. The changes are usually minimal and therefore should have little or no effect on the development and implementation of this model, at least for the near future. Eventually though, these increasing RW values may result in the model having to be adjusted to account for them.

3.2.3 Patient Insurance Provider (IP)

This data consists of 20 categories indicating the source of payment for all treatments. In the overwhelming majority of cases, this will be some health insurance provider, although in a few cases, the patient pays for treatment directly (‘Self’ category). Patient levels for each of the 20 categories vary from 19 (Other Governmental Insurance) to 4958 (Medicare). Although the DRG system was originally created exclusively for Medicare patients only, most hospitals, including GSH, use it for billing non-Medicare patients as well. However, while all IPs are billed the same, the *actual* payment varies for each particular IP and is subject to privacy restrictions. Therefore, this analysis can only determine which, if any, IPs affect LOS, but cannot explain the particular reasons for these irregularities. For instance, the fact that provider A pays a higher amount than provider B may induce a shorter or longer stay on patient A compared to patient B, assuming the two patients and their diagnoses are similar. Also, each individual IP may have several different coverage policies available, each with varying payment policies.

3.2.4 Patient Admit and Discharge times

Admit and discharge times are precise to the minute and assumed, in most cases, to be accurate and reflect the actual time the patient was present for treatment.

3.3 Exploratory Data Analysis

Since there is such a large amount of data, in some instances the model was not developed according to each observation of patient data. Instead, mean LOS based on certain patient characteristics was used. For example, if the regressor is RW, all the patients are grouped together based on his or her particular RW. The mean actual LOS values for each RW then become the “new” observations.

This technique was primarily used in early model development, particularly to find the relationships between the individual regressors and the response, actual LOS. However, as the model progressed, it was determined that the model could lose some of its information, particularly model fidelity, using this approach and eventually each of the patients was treated as a single observation.

Again, since there is an excessive amount of patient data, efficient and effective techniques were developed to handle it effectively. Microsoft Excel^{TM} software was used to store and sort the data, while MATLAB^{TM} software was used to calculate statistical information such as computing the means, variances, and develop all regression model information (analysis-of-variance (ANOVA) tables, model coefficients, and residual terms). For example, if GMLOS and patient age were the two regressors, Excel was used to sort patients, first by GMLOS, and secondly by age. An example of this is:

$$\begin{bmatrix} 1.6 & 46 \\ 2.1 & 75 \\ 1.8 & 56 \\ 1.6 & 27 \\ 2.1 & 44 \\ 1.6 & 67 \end{bmatrix} \qquad\qquad \begin{bmatrix} 1.6 & 27 \\ 1.6 & 46 \\ 1.6 & 67 \\ 1.8 & 56 \\ 2.1 & 44 \\ 2.1 & 75 \end{bmatrix}$$

where the first matrix is the original, unsorted data and the second is the resulting sorted data. Along with the corresponding actual LOS, this sorted data would then be placed in a text file representing an *n* x *3* matrix, where *n* is the number of observations (patients) and the three columns are represented by x_1, x_2, and y, respectively. MATLAB functions were then written to read in these text files and calculate statistical results for each combination of GMLOS and age. Both software packages were also used to plot the various associated graphs.

3.4 Simple Regression Analysis

Initially, each of the six regressors (RW, GMLOS, AMLOS, age, gender, and IP) was studied individually to determine what, if any, significance each of them had on LOS. It was established in Chapter 1 that RW and expected LOS (GMLOS, AMLOS) all have an intuitive relationship to LOS, i.e., generally the higher these values are, the longer LOS and higher total costs will be, and vice versa. Because of this, these regressors, which from this point on will be referred to as the DRG-regressors, were plotted directly against LOS. However, using similar reasoning, there is no corresponding relationship between the non-DRG regressors (age, gender, IP) and LOS. Therefore, for these particular regressors, the response used was average percent difference between expected LOS (GMLOS) and actual LOS to determine how particular categories of each impact LOS.

As seen in Figures 3.1, 3.2, and 3.3 below, there is indeed an increasing relationship in LOS vs. RW, LOS vs. GMLOS, and LOS vs. AMLOS, respectively.

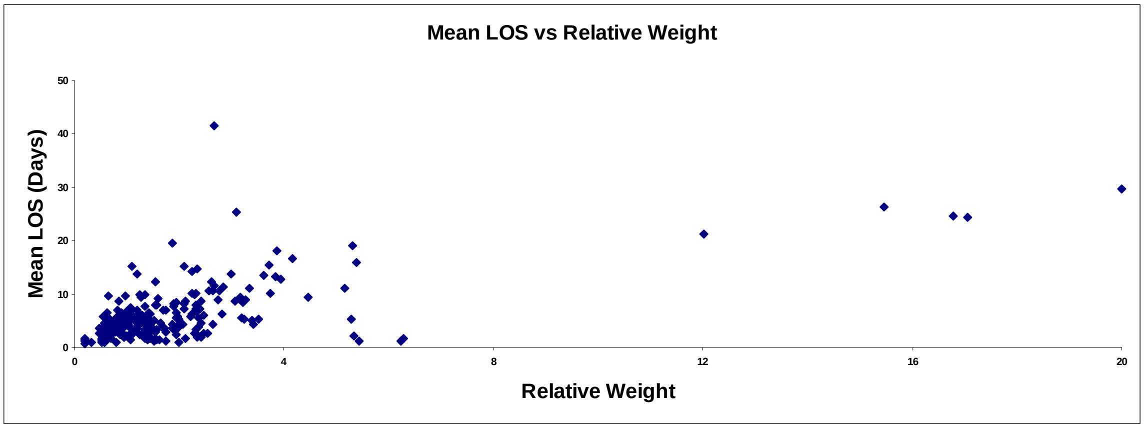

Figure 3.1 Mean LOS vs. Relative Weight

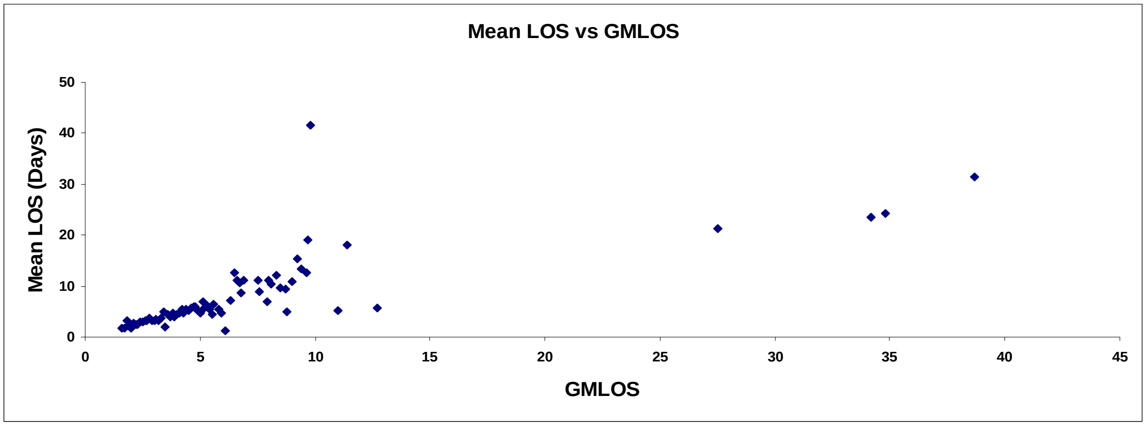

Figure 3.2 Mean LOS vs. GMLOS

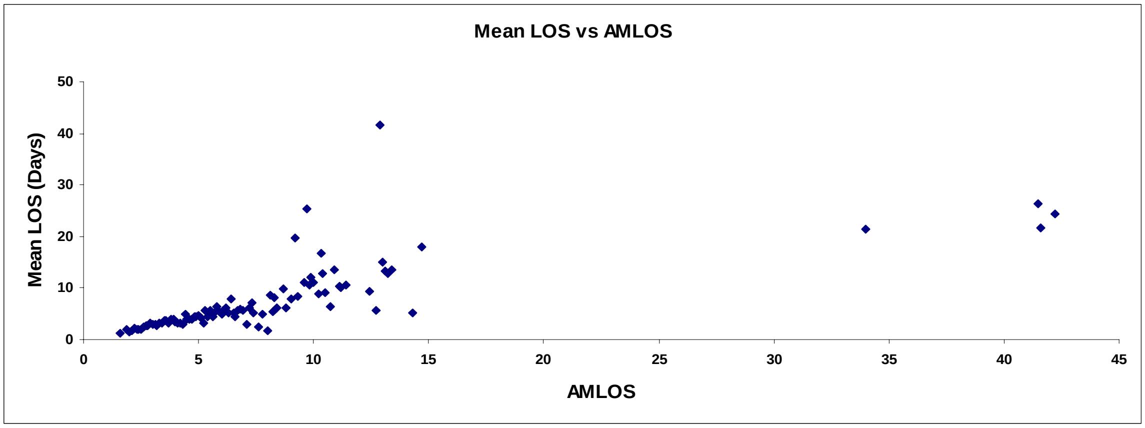

Figure 3.3 Mean LOS vs. AMLOS

Figure 3.4 below shows a logarithmic relationship between age and LOS.

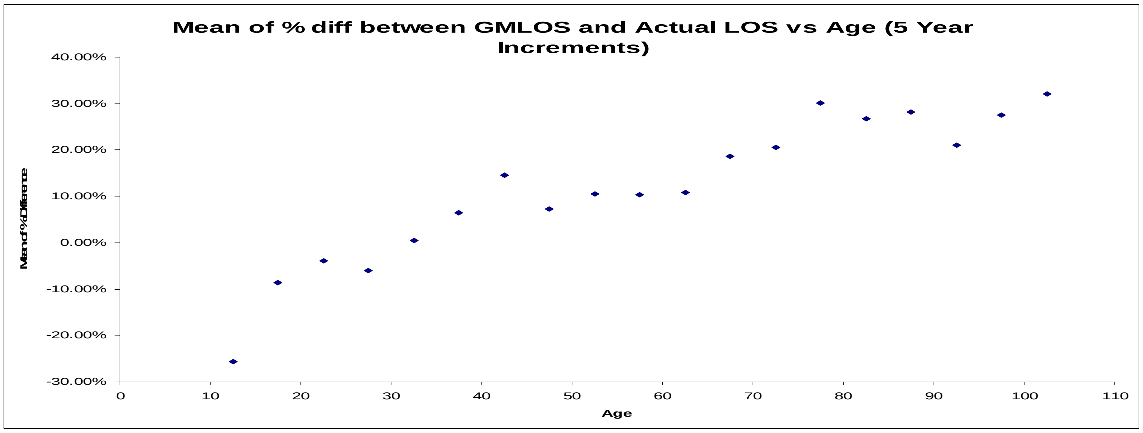

Figure 3.4 Mean % Difference between GMLOS and Actual LOS vs. Age

Note that the ages were grouped into 5-year increments (11-15, 16-20,…, 101-105) and are plotted at the midpoint of each group (13, 18,…, 103). It may be expected that a younger, healthier person will have a shorter LOS, but the "leveling off" at higher ages is not as easily explained. It could actually indicate elderly patients tend to expire more frequently. This could perhaps be due to their receiving less aggressive treatment due to frail health or a living will that has a "do not resuscitate" provision. Figure 3.5 below shows that various IPs can also impact LOS.

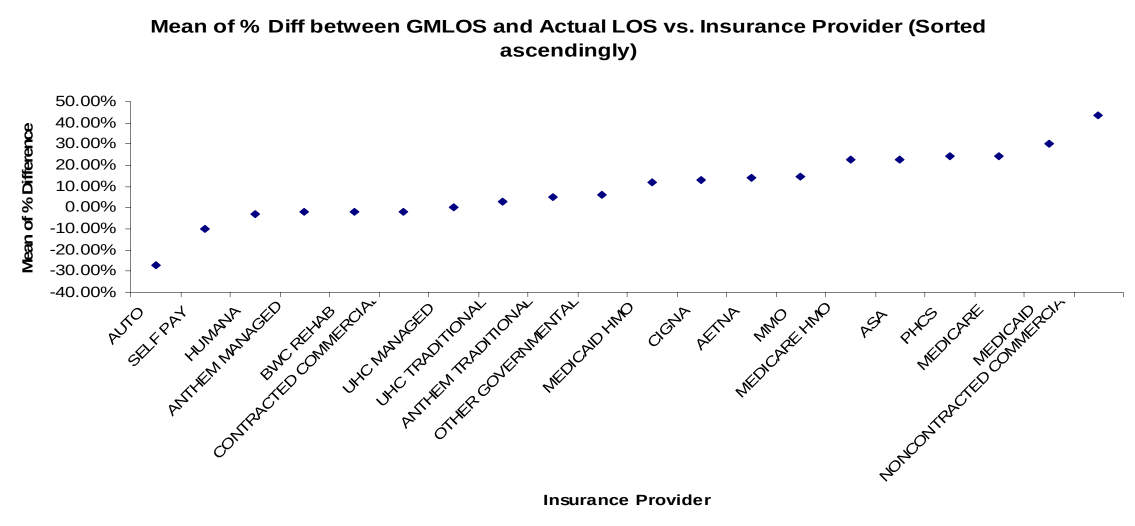

Figure 3.5 Mean % Difference between GMLOS and Actual LOS vs. IP

Again, not knowing the detail and policies of each provider, it is not possible to explain these trends, only that a relationship does indeed exist.

To examine the total variability of the data, a technique similar to the one used to calculate the means was used. Figures 3.6, 3.7, and 3.8 below show how LOS variability increases significantly as RW, GMLOS, and AMLOS increase, respectively.

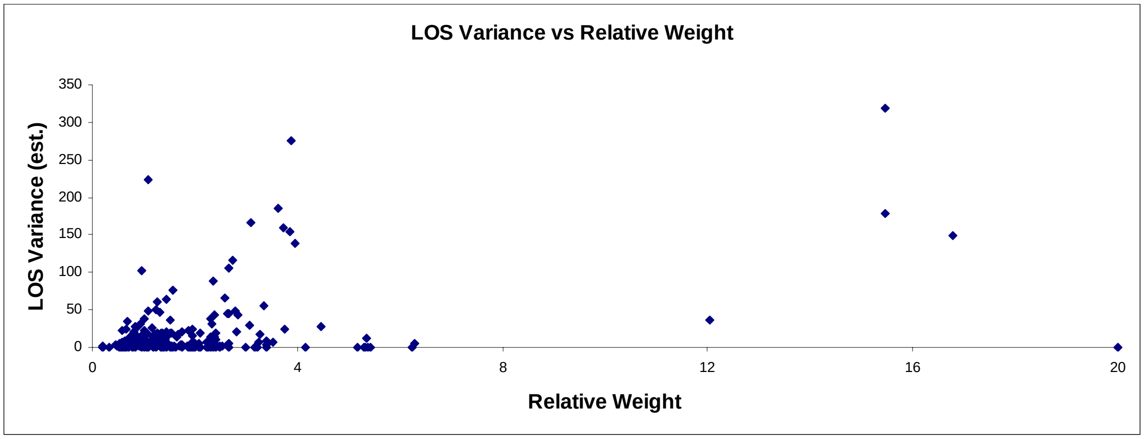

Figure 3.6 LOS Variance vs. Relative Weight

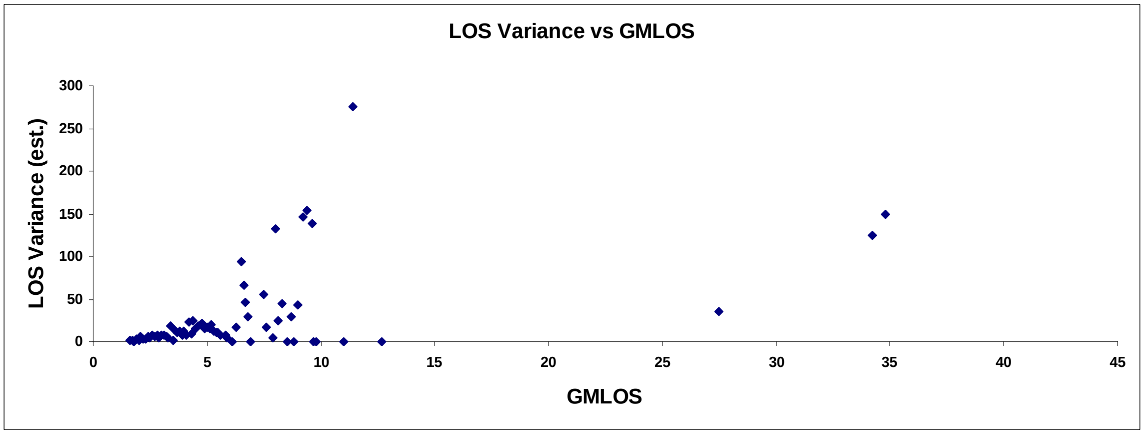

Figure 3.7 LOS Variance vs. GMLOS

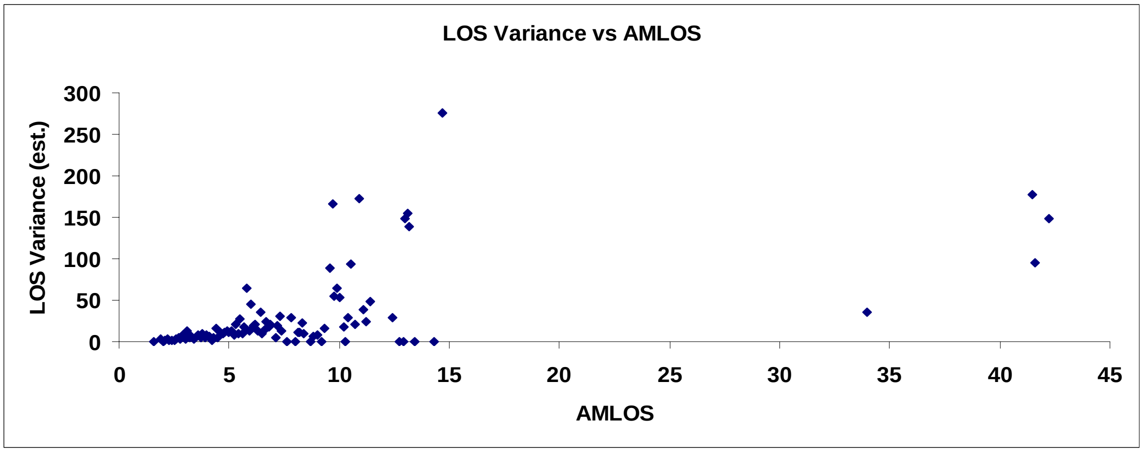

Figure 3.8 LOS Variance vs. AMLOS

Variability effects of the three non-DRG regressors are not nearly as significant. As shown in Figure 3.9 below, there is clearly no relationship between age and LOS variance.

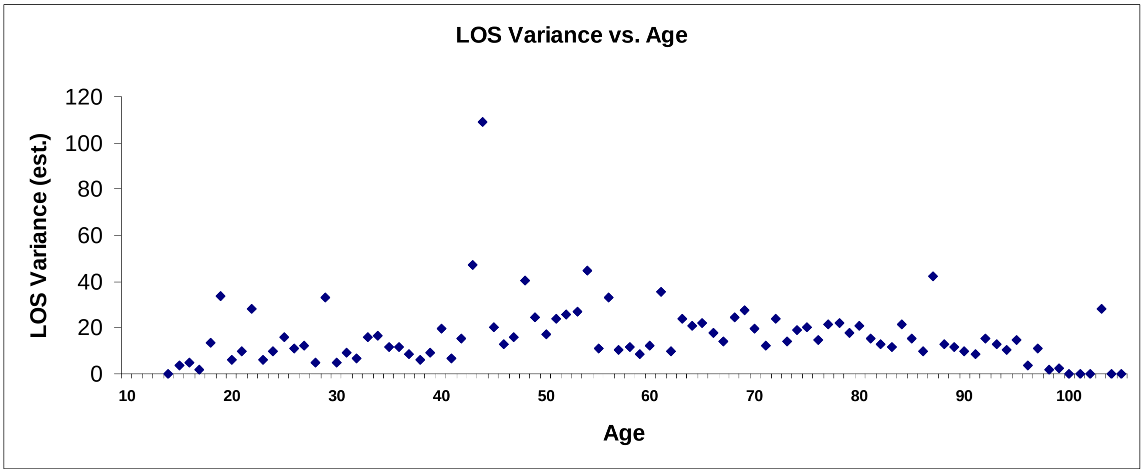

Figure 3.9 LOS Variance vs. Age

Using the previously defined ordering scheme for IP (Figure 3.5), Figure 3.10 reflects that there is also no relationship between IP and LOS variance.

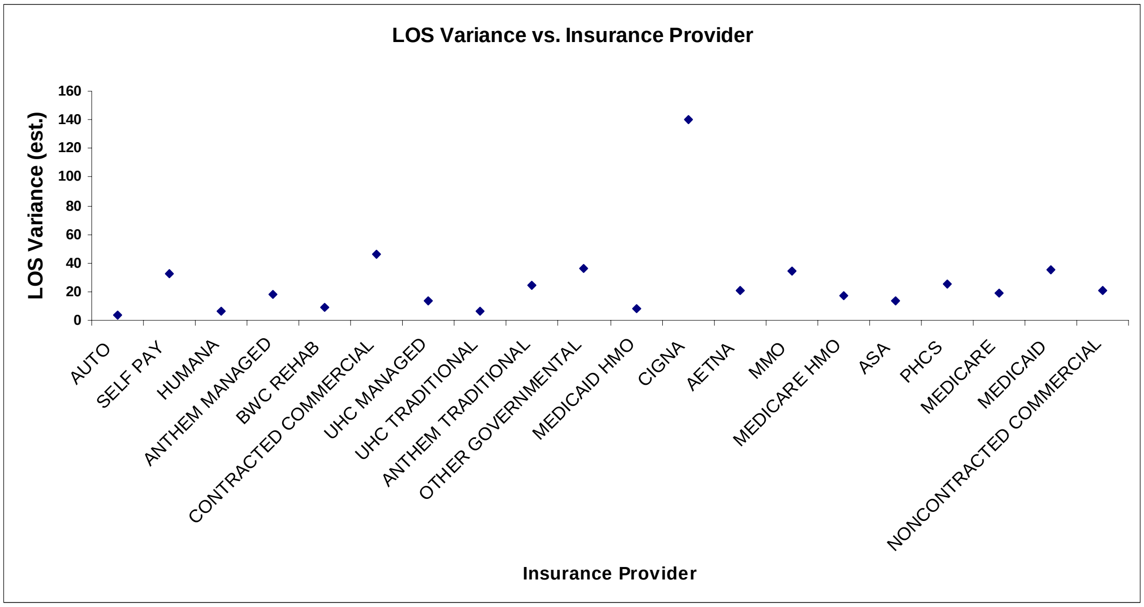

Figure 3.10 LOS Variance vs. Insurance Provider

Gender would prove to be more difficult to study. This is because even though males have a slightly higher variance (25.60) than females (17.79), having only two levels makes the results inconclusive. As a result, gender was combined with age by analyzing LOS variance at each age/gender combination. Figure 3.11 below shows no noticeable relationship between LOS variance and age and gender.

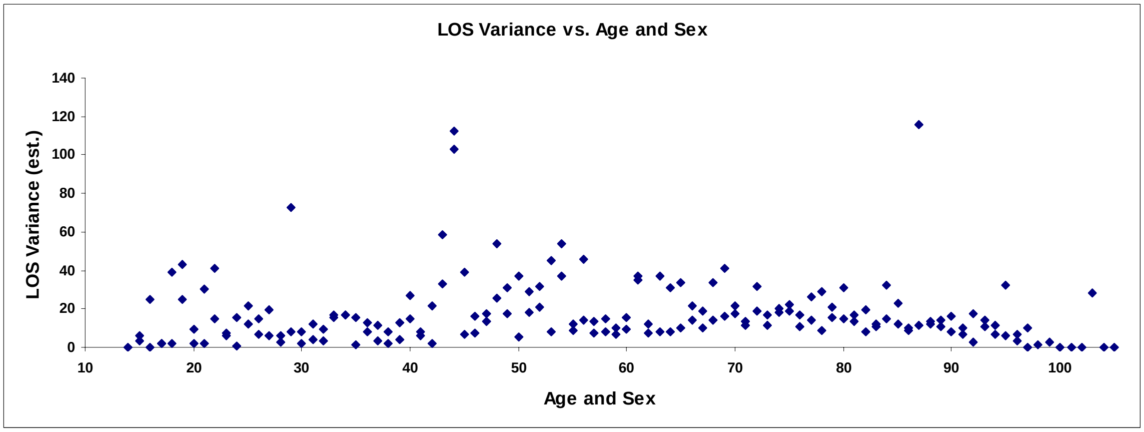

Figure 3.11 LOS Variance vs. Age and Gender

In general, the non-DRG regressors have constant variability while the DRG regressors do not. This is apparent even though there is a considerable degree of "unknown" variance throughout each range due to instances of only one observation at several values of RW, GMLOS, and AMLOS. The interactions between the DRG and non-DRG regressors may show other effects, but at this point it is not necessary to study and analyze them.

There are several approaches to account for this nonconstant variance. One of them used in general practice is weighted least squares (WLS). Three various methods using WLS are developed and implemented and will be discussed next few sections.

3.5 Least Squares Approaches

General assumptions usually made regarding the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ are that $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. This is commonly referred to as ordinary least squares (OLS) and assumes the error term has mean vector $\mathbf{0}$, the variance-covariance matrix $\sigma^2\mathbf{I}$ has uncorrelated and constant errors, and the least squares criterion is simply the squares of the error, or residual, terms:

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (3\text{-}1)$$

However, data typically possesses nonconstant error variance, making these earlier assumptions not only impractical, but also infeasible. The error terms still have an expected value of $\mathbf{0}$, but now $\text{Var}(\boldsymbol{\varepsilon}) = \mathbf{V}$, where $\mathbf{V}$ is the variance-covariance matrix. If $\mathbf{V}$ is strictly diagonal but with unequal diagonal elements, the observations $\mathbf{y}$ are uncorrelated with unequal variances; if some of the off-diagonal elements are nonzero, then the observations $\mathbf{y}$ are correlated. Since neurological diseases are generally not communicable in anyway, it is assumed all observations are uncorrelated, and therefore $\mathbf{V}$

is strictly diagonal. These diagonal elements represent the estimated weight of each corresponding observation. The resulting least-squares criterion now becomes:

$$S = \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 \tag{3-2}$$

The following sections describe three various methods of developing these weights. [Montgomery 2001].

3.5.1 Modeled Variance

In this approach, **V** is the variance-covariance matrix of **ε**, or Var(**ε**). However, as mentioned previously in this chapter, there is a considerable degree of unknown variance due to instances of only one observation at several values of RW, GMLOS, and AMLOS. Therefore, a linear model of the variance will be developed using the estimated variances. A potential drawback of this model is that it could generate negative values for the expected variance unless nonnegativity constraints are induced on the fitted values.

3.5.2 Estimating Weights via Observation Frequency

The premise behind this approach is to place more emphasis on observations that occur more frequently. For example, if a certain category of stroke occurs 50 times more often than another, more weight is placed on the former because the variance estimate is much more accurate and reliable compared to the latter. This method has several advantages. It is an intuitive approach, relatively simple to implement, and can easily be updated as more patient data becomes available. These weights are represented by **W**, an *n* x *n* diagonal matrix whose elements represent the weight estimate of each corresponding observation.

3.5.3 Isotonic Regression

Monotonic regression is a nonparametric regression method designed for cases where the expected value of a response variable changes isotonically (non-decreasing) or antitonically (non-increasing) in relation to one or more regressor variables. Isotonic regression satisfies the following: min $\sum_{i=1}^{N} C_i(x_i)$ subject to $x_1 \le x_2 \le \cdots \le x_n$, where C_i (x_i) is a convex function of x_i for each $i \in N = \{1,2, ..., n\}$. The process effectively "smoothes out" a response, resulting in a piecewise continuous step function which will reduce X to k ≤ *n* level sets [25].

Essentially, this technique will use information from non-zero adjacent variances to calculate a more accurate estimated variance at these points of unknown variances. Since the variability has been shown to be non-decreasing with RW, GMLOS and AMLOS, isotonic regression can be used to "fill in" the unknown variances for RW, GMLOS, and AMLOS with only one observation. Figure 3.12 below gives an example of isotonic regression by showing the previous LOS Variance vs. RW graph (Figure 3.6) with its corresponding isotonic regression plot.

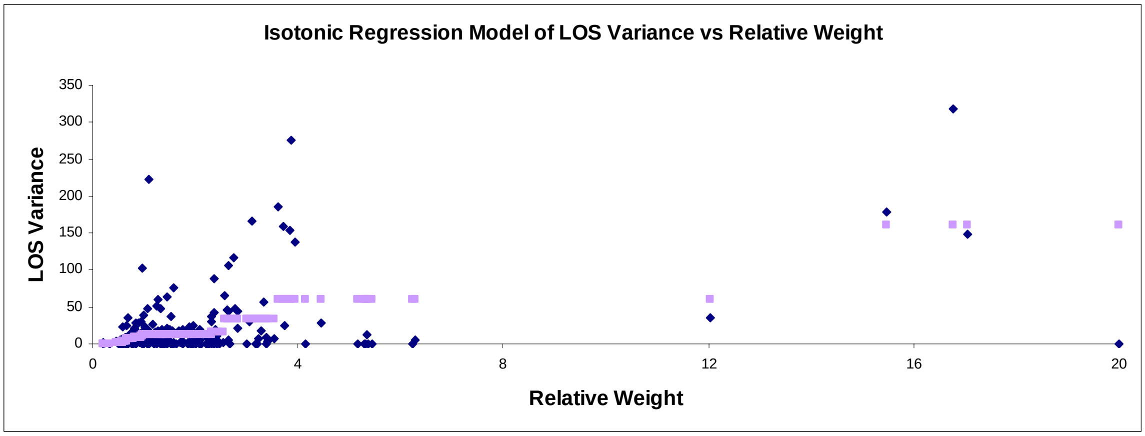

Figure 3.12 Isotonic Regression Model of LOS Variance vs. Relative Weight

The plot demonstrates how the original variance values are used to form a step function to effectively model points where the estimated variance is zero, or unknown.

3.5.4 Derivation of WLS model parameters and ANOVA terms

Nonconstant variance dramatically alters the derivation of the model. Before developing the updated model parameters and sum of square terms, a review of the ordinary least squares is given. Recall that to minimize the sum of the squared error terms,

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \varepsilon_i^2 \\ &= \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^T \mathbf{y} - 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X}\boldsymbol{\beta} \end{aligned} \qquad (3\text{-}3)$$

by satisfying:

$$\left.\frac{\partial S}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} = -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \qquad (3\text{-}4)$$

which, after solving the normal equations for **β**, yields:

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{3-5}$$

The formulas for SS$_R$ and SS$_{Res}$ are also modified to accommodate the $\mathbf{V}^{-1}$ matrix [19]. From basic regression analysis, the total sum of squares (SS$_T$) is a measure of the variability in **y** and is defined as:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{3-6}$$

SS$_T$ is actually the sum of two terms: regression sum of squares (SS$_R$) and residual sum of squares (SS$_{Res}$). The former is generally used to indicate the proportion (SS$_R$/SS$_T$) of variability that can be explained by the regression model, while the latter represents the unexplained variability. They are defined as:

$$SS_{\mathrm{R}} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \tag{3-7}$$

$$SS_{\mathrm{Res}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3-8}$$

However, in the weighted least squares case, the estimator and sum of square formulas are modified to accommodate the V (and W) matrix:

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \tag{3-9}$$

$$SS_R = (\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} - \mathbf{1}\bar{\mathbf{y}})^T\mathbf{V}^{-1}(\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} - \mathbf{1}\bar{\mathbf{y}}) \tag{3-10}$$

$$SS_{\mathrm{Res}} = \mathbf{y}^T\mathbf{V}^{-1}\mathbf{y} - \mathbf{y}^T\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \tag{3-11}$$

(Full derivations of these are presented in Appendix B)

3.6 Model Development

3.6.1 Regression Model Development

The main focus in the development of the model is a method for dealing with the nonconstant (and unknown) variances. It is imperative that these variances be accurately estimated and accounted for in the model. Weighted least squares regression is an effective and proven method to accomplish this due to its robustness and adaptability. Furthermore, there are several approaches to estimating the "weights".

Ultimately, four methods were studied:

1) Ordinary Least Squares Method (OLS)

   This is a general un-weighted regression method and was used as the baseline model for comparison to the subsequent weighted methods.

2) Modeled Variance Method (WLS 1)

   This is the first of the proposed weighted least squares methods. LOS variance is modeled with the inverse fitted values of the resulting model being used as the weight estimates.

3) Observed Frequency Method (WLS 2)

   The concept behind this method is that the more common occurring observations should be weighted more than those that occur less frequently. This is because there is a higher confidence in the variance estimate at these higher frequencies. Therefore, the weight estimates are the frequencies of each corresponding observation.

4) Isotonic Regression Method (WLS 3)

   Because empirical evidence suggests that LOS variance is a non-decreasing function of RW, GMLOS, and AMLOS, as evidenced in Figures 3.6, 3.7, and 3.8, isotonic regression can be used to derive the weight estimates, which are simply the inverse values of the resulting isotonic regression of the estimated LOS variance.

3.6.2 Simulation Model Development

Using the original patient data, a simulation model of the neurology unit was developed. This was done simply by using the admit and discharge times to develop inter-arrival and service (treatment) times. This model will be used for comparison with the eventual regression model(s).

3.7 Multiple Regression Analysis

Now that varying relationships between each of the regressors and LOS have been established, multiple regressors and their interactions can be added to the model. Before accomplishing this however, several issues must be resolved, such as data representation, coding of the qualitative IP regressor, calculation of the variance weights, and possible multicollinearity effects due to high correlation between some of the regressors.

3.7.1 Variable Scaling

A uniform scaling system for the regressors was implemented. This was done because it would be very difficult to establish and analyze interactions between, for example, RW, which ranges continuously from 0 to 20, and age, which ranges from 14 to 105 in integer increments. Scaling the data would make the magnitudes of the resulting estimates of the regression coefficients uniform, and therefore, easier to compare. All regressors were scaled using unit length scaling [Montgomery 2001]:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{S_{jj}^{1/2}},\ i = 1, 2,\ldots, n,\quad j = 1, 2,\ldots, k \tag{3-12}$$

$$\text{where } S_{jj} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \tag{3-13}$$

Note: x_{ij} is the value of regressor j at observation i, $\bar{x}_j$ is the mean of regressor j, $\bar{z}_j = 0$ and $\sqrt{\sum_{i=1}^{n} (z_{ij} - \bar{z}_j)^2} = 1$.

3.7.2 Coding of the Qualitative IP Regressor

Coding for the 20 levels of the IP regressor can either be arbitrary or logical. However, it was shown that LOS does vary with IP (Figure 3.5). As a result, it was believed a ranking scheme would be a slightly improved approach compared to a completely arbitrary one. Therefore, the IPs are ranked (using unit length scaling) ascendingly according to the average percent difference between expected LOS (GMLOS) and actual LOS, as found in Figure 3.5 (AUTO, SELF-PAY,…, MEDICAID, NON-CONTRACTED COMMERCIAL).

3.7.3 Data Binning

To calculate the weight estimates, the data was grouped into equally sized bins. Rather than use a complete arbitrary number of bins, $\sqrt{n}$ was used, which is often the value used in grouping data for statistical methods, such as histogram graphs. For this model, this value is ~85 ($\sqrt{7319}$). This turns out to be a useful benchmark because many of the predictors have approximately this many distinct values. For example, age has 92 values (14-105). Some predictors (IP, gender) do not have this many levels, so the number of bins for those are simply the total number of levels. RW, with 323 unique values, was grouped into 81 bins, with each bin having approximately four values of RW. Table 3.1 shows the number of bins for each predictor:

Table 3.1 Number of Bins for each Predictor

Predictor	Unique values	No of bins
RW	323	81
GMLOS	72	72
AMLOS	96	96
IP	20	20
Age	92	92
Gender	2	2

For example, male patients with GMLOS=6.7, IP=13, and Age=56, will all have the same weight estimate, which will be calculated using that particular group of patients. If the corresponding LOS variance estimate from this group is 25 days and the total number of patients is 10, then the corresponding weight estimates for WLS 1 and WLS 2 will be 0.04 (1/25) and 10/7319, respectively.

3.7.4 Multicollinearity Effects

It is highly suspected that the DRG regressors, particularly GMLOS and AMLOS, are highly linearly dependent and therefore, may be highly correlated with one another. Since $Var(\hat{\beta}) = \sigma^2 (X^T X)$ (for OLS), highly linear dependent columns of **X** will result in very large variances in the estimates of the model parameters.

Table 3.2 below shows that the DRG regressors are indeed highly correlated with one another (magnitude of coefficients > ~0.9):

Table 3.2 Correlation Matrix ($\mathbf{X}^T\mathbf{X}$)

	RW	GMLOS	AMLOS	Age	Gender	IP
RW	1	0.906	0.904	-0.072	0.073	-0.047
GMLOS		1	0.995	-0.018	0.035	0.008
AMLOS			1	-0.011	0.039	0.014
Age				1	-0.103	0.497
Gender					1	-0.080
IP	Sym.					1

However, these high values do not necessarily guarantee poorly estimated model parameters when using both of the corresponding regressors in the same model. Variance inflation factors (VIFs) are typically used to determine this. VIFs are simply the diagonal elements of $(\mathbf{X}^T\mathbf{X})^{-1}$, or $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$ and $(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}$ for WLS, accordingly. Montgomery [18] states that if any of these exceed 5, the model parameters are poorly estimated due to multicollinearity. As seen in Table 3.3, all three DRG regressors exceed this limit (RW = 5.788, GMLOS = 108.455, AMLOS = 106.354).

Table 3.3 Variance Inflation Factors $(\mathbf{X}^T\mathbf{X})^{-1}$

	RW	GMLOS	AMLOS	Age	Gender	IP
RW	5.788	-3.795	-1.448	0.214	-0.197	0.203
GMLOS		108.455	-104.528	0.459	0.658	0.223
AMLOS			106.354	-0.630	-0.516	-0.432
Age				1.347	0.080	-0.648
Gender					1.022	0.035
IP	Sym.					1.339

These high values indicate that any model with multiple DRG regressors will indeed have poorly estimated model parameters. Therefore, these models will be avoided altogether.

3.8 Model Ranking and Statistics of Performance

An approach has now been established to develop and study a regression model(s) based on this empirical patient data. The ideal model(s) should be fairly simple, easy to implement, and require low maintenance. While the main objectives are overall performance and simplicity, extensive comparisons of the OLS and WLS models will also be analyzed.

To rank each potential model, a commonly used statistic, called Mallow's $C_{p,}$ will be used. Mallow's C_p is a very good statistic to use when comparing models with different values of *p*. This statistic is a function of the residual sum of squares (SS_{Res}) for the full regression model and that for the reduced model, which will be a model containing a combination of, but not all, the regressors. The equation for C_p is:

$$C_p = \frac{SS_{\mathrm{Re}s}(p)}{\hat{\sigma}^2} - n + 2p \qquad (3\text{-}14)$$

where SS_{Res} (*p*)is the error sum of squares for the reduced model with *p* terms, $\hat{\sigma}^2$ is assumed to be an unbiased estimate of MSE (Mean Square Error = $SS_{Res}/(n\text{-}p)$) for the full model and $p = k + 1$, where *k* is the number of regressors in the model. Under the correct model, C_p is approximately equal to *p* and otherwise is typically greater than *p*, reflecting bias in the parameter estimates in the regression equation. (A value less than *p* indicates the model is over-fitted.) Thus, it is desirable to select a model in which the value of C_p is close to *p*.

IV. Results and Analysis

4.1 Overview

The results of the regression analyses and model comparisons follow. Section 4.2 outlines summary statistics for all possible regressions followed by further analysis of the best-fit models (Section 4.3). Final model analysis is performed in Section 4.4. The chapter concludes with a comparison of the original simulation and proposed regression models to the original data (Section 4.5) before sensitivity analysis of the regression coefficients is performed in Section 4.6.

4.2 Initial Regression Analysis

The initial objective was to determine the significant regressors by studying all possible regressions. However, the standard approach to performing this was modified somewhat. Because of the multicollinearity effects mentioned in Chapter 3, no more than one DRG regressor will be present in any of the studied models. Moreover, because it is assumed a DRG regressor must be present to accurately predict LOS, all models will contain one. Because there are 3 DRG regressors (RW, GMLOS, AMLOS), 4 methods (OLS, WLS 1, WLS 2, WLS 3), and up to K = 3 additional regressors, this consisted of 96 ($3 \cdot 4 \cdot 2^K$) total regression models with the eight possible baseline regression models shown below in Table 4.1. Model response is actual LOS.

Table 4.1 All Possible Regressions

DRG-regressor,
DRG-regressor, IP
DRG-regressor, Age
DRG-regressor, Gender
DRG-regressor, IP, Age
DRG-regressor, IP, Gender
DRG-regressor, Age, Gender
DRG-regressor, IP, Age, Gender

With so many models to be evaluated, a systematic approach to analyzing the various models was developed. Since Mallow's C_p statistic is based on the number of regressors in the model, or *k,* the idea was to analyze each group of models based on increasing values of *k*.

Table 4.2 below shows the top 2-regressor models based on Mallow's C_p statistic.

Table 4.2 Top 2-Regressor Models

Rank	Regressors	Method	Mallow's Cp
1	AMLOS, IP	WLS 2	7.7
2	AMLOS, IP	WLS 3	9.6
3	GMLOS, IP	OLS	9.7
4	AMLOS, IP	OLS	9.9
5	AMLOS, IP	WLS 2	13.5
6	GMLOS, IP	WLS 3	13.7
7	GMLOS, IP	WLS 1	16.4
8	RW IP	OLS	19.8
9	RW IP	WLS 1	25.7
10	RW IP	WLS 3	30.1

The best method overall (Mallow's $C_p = 7.7$) is the AMLOS, IP model using the WLS 2 method. Another AMLOS, IP model using the WLS 3 methods is the second ($C_p = 9.6$) best model. As indicated in the table, AMLOS is the most significant DRG regressor, appearing in 4 of the top 5 models. GMLOS and RW are the second and third best DRG regressors, respectively. IP is the most significant non-DRG regressor, while the effects of age and gender are effectively negligible. The WLS methods, particularly WLS 2, typically outperform OLS, with WLS 1 and WLS 3 performing comparatively to one another.

Results of the analysis using 3-regressor models are shown in Table 4.3 below.

Table 4.3 Top 3-regressor models

Rank	Regressors	Method	Mallow's Cp
1	AMLOS, IP, Age	WLS 2	3.7
2	AMLOS, IP, Age	WLS 3	4.0
3	AMLOS, IP, Age	OLS	4.1
4	GMLOS, IP, Age	OLS	4.9
5	GMLOS, IP, Age	WLS 3	6.6
6	GMLOS, IP, Age	WLS 2	6.7
7	RW, IP, Age	OLS	7.5
8	RW, IP, Age	WLS 2	8.1
9	RW, IP, Age	WLS 3	8.2
10	AMLOS, IP, Gender	WLS 2	8.7

AMLOS and IP, when combined with age, continue to perform as well as they do in the 2-regressor models, with the top 3 models consisting of all three. IP continues to be the most significant non-DRG regressor, appearing in the top 9 models, while gender appears only once. WLS 2 remains the best method, with WLS 1 and WLS 3 performing about the same.

Since MS_{res} from the full regression (4-regressor) model is being used for $\hat{\sigma}^2$, the Mallow’s statistic for it will always be p, or $k + 1 = 5$. However, the top four 3-regressor models are an improvement over this. This indicates that gender adds no significance to the model, and therefore the 3-regressor models are superior to the full regression ones.

This analysis shows that the best 2 and 3 regressor models are AMLOS, IP and AMLOS, IP, Age, respectively, using the observed frequency approach (WLS 2). Their corresponding ANOVA tables are found in Tables 4.4 and 4.5 below.

Table 4.4 ANOVA Table for AMLOS, IP Model

Source Variation	Sum of Squares	dof	Mean Square	F	p-value
Regression	186690	2	93344.8	6141.09	0
Residual	111203	7316	15.2		
Total	297893	7318			

Table 4.5 ANOVA Table for AMLOS, IP, Age Model

Source Variation	Sum of Squares	dof	Mean Square	F	p-value
Regression	182545	3	60848.5	4007.17	0
Residual	111078	7315	15.2		
Total	293623	7318			

Since the p-values from each of the corresponding F tests are very small, it is concluded that LOS is related to AMLOS, IP, and/or Age. Because it is not known exactly which of them is/are significant, further tests of model adequacy are required.

4.3 Tests on Individual Regression Coefficients

Adding a variable to any regression model increases the variance of the fitted value $\hat{y}$ so care must be used to include only those that of real value in explaining the response. Furthermore, adding an insignificant regressor may increase the residual mean square, possibly decreasing the model's utility.

The hypothesis for testing the significance of the individual regression coefficient β_j is:

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

If $H_0 : \beta_j = 0$ is not rejected, this indicates regressor x_j can be removed from the model. The corresponding test statistic is:

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \tag{4-1}$$

where C_{jj} is the diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$ corresponding to $\hat{\beta}_j$. Note that this will be different from the **X** matrix studied earlier in Chapter 3 as **X** now only contains the two or three regressors currently under study as opposed to the original six. The null hypothesis $H_0 : \beta_j = 0$ is rejected if $|t_0| > t_{\alpha/2,n-k-1}$. This is actually a partial test because coefficient $\hat{\beta}_j$ depends on all other regressor variables x_i ($i \neq j$) that are in the model. In more general terms, this is a test of the contribution of x_j given other variables are in the model. As seen in Table 4.6 below, all coefficients for both models are statistically significant (α = 0.05).

Table 4.6 Hypothesis Testing on the Individual Regression Coefficients

Model	Regressor	$\hat{\beta}_j$	$\hat{\sigma}^2$	C_{jj}	t_0	$t_{0.025,n-k-1}$	Result
2 regressor	AMLOS	203.55	15.2278	1.001	52.14	1.96	Reject H_0
	IP	41.63	15.2278	1.001	10.66	1.96	Reject H_0
3 regressor	AMLOS	203.82	15.2138	1.001	52.23	1.96	Reject H_0
	IP	30.73	15.2138	1.001	7.87	1.96	Reject H_0
	Age	10.39	15.2138	1.001	2.66	1.96	Reject H_0

To determine the contribution of regressor x_j, given that other regressors x_i ($i \neq j$) are included in the model, the extra-sum-of-squares method is generally used. This can also be used to investigate the contribution of a subset of regressors. For example, to determine if some subset of $r < k$ regressors contribute significantly, β is partitioned as:

$$\beta = \left[\frac{\beta_1}{\beta_2}\right]$$

where β is p x 1, β_1 is ($p - r$) x 1, and β_2 is r x 1. The hypothesis test is then:

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

The test statistic, which will not be derived, is:

$$F_0 = \frac{SS_R(\beta_2 \mid \beta_1)/r}{MS_{\mathrm{Re}\,s}} \tag{4-2}$$

where $SS_R(\beta_2 \mid \beta_1) = SS_R(\beta) - SS_R(\beta_1)$, and $SS_R(\beta)$ and $SS_R(\beta_1)$ are the regression sum of squares for β and β_1, respectively. If $F_0 > F_{\alpha,r,n-p}$, H_0 is rejected, and it is concluded at least one of the parameters of β_2 is not zero. In other words, at least one of these regressors contributes significantly to the model. Table 4.7 below shows the results of the extra-sum-of-squares method:

Table 4.7 Hypothesis Testing Using the Extra-Sum-of-Squares Method

Original Model	$SS_R(\beta_1)$	β_2	$SS_R(\beta)$	F_0	$F_{0.05,1,7315}$	Result
AMLOS	41566	IP	43297	113.67	3.84	Reject H_0
AMLOS, IP	43297	Age	43415	7.76	3.84	Reject H_0

As seen in the table, both IP and age add statistically significance to the model. However, based on the F-statistic, IP is much more significant than age.

4.4 Model Analysis

The two proposed models, AMLOS, IP, Age and AMLOS, IP, are defined by the following coefficients, respectively:

$$\beta = \begin{bmatrix} -1.68 \\ 0.951 \\ 0.081 \\ 0.006 \end{bmatrix} \qquad \beta = \begin{bmatrix} -1.45 \\ 0.956 \\ 0.092 \end{bmatrix}$$

The mathematical formulations for each respective model are then:

$$y = -1.68 + 0.951x_{AMLOS} + 0.081x_{IP} + 0.006x_{Age} \qquad (4\text{-}3)$$

$$y = -1.45 + 0.956x_{AMLOS} + 0.092x_{IP} \qquad (4\text{-}4)$$

Table 4.8 shows a few of these models' responses using observations from the original patient data.

Table 4.8 Examples of Model Responses Using Original Patient Data

AMLOS	Insurance Provider	IP Level	Age	Expected Value (M1)	Expected Value (M2)	Observed Value
4.6	MEDICARE	19	83	4.732	4.696	4.263
4.6	SELF PAY	2	58	3.205	3.132	3.630
13.0	ANTHEM TRADITIONAL	9	64	11.796	11.806	12.915

These results show the models will generally give a more accurate expected LOS than strictly using AMLOS. The table also shows the models produce very similar outputs. This confirms the results of the extra-sum-of-squares method that showed age adds very little significance to the model.

4.5 Model Comparison

Each of the two proposed regression models was compared to the original simulation model provided to GSH. The only significant difference between the two models is that the GSH model is based on the most recent 18 months of patient data (January 2004 to June 2005), while the regression models consist of all 42 months of available data. However, there was no significant change in hospital policy during this time that would have affected the flow and type of incoming patients. Therefore, the patient distributions, and thus the resulting models, are assumed to be effectively similar.

All three models were simulated using ARENA^{TM} software. The statistics of primary interest are total and mean patient level and LOS, and to a lesser extent, max number of patients and min/max LOS. The models were simulated in replicates of 42 months, which is the total time range of the original patient data and is also long enough to give an accurate steady-state representation of the model. Initially, 10 replicates were run to obtain sample variances which were then used to determine the eventual number of required replicates using the confidence interval half-width:

$$H = t_{\alpha/2,R-1}\frac{S}{\sqrt{R}} \tag{4-5}$$

Using a predetermined error criterion (ε), the objective is then to determine the minimum R which satisfies:

$$R \geq \left(\frac{z_{\alpha/2}S_0}{\varepsilon}\right)^2 \tag{4-6}$$

For $\alpha = 0.05$, $\varepsilon_1 = 0.15$ patients for average patient level, and $\varepsilon_2 = 0.01$ days for average LOS, 59 replications is the required minimum that satisfies all error criterion. Table 4.9 compares statistics from the simulation and regression models to the original data.

Table 4.9 Comparison of Original Data to Simulation and Regression Models

	Total Patients	% Diff	Avg Patient Level	% Diff	Max Patient Level	% Diff	Avg LOS (Days)	% Diff	Min LOS (Days)	Max LOS (Days)
Original Data	7319	----	24.98	----	42.5	----	4.443	----	0.002	94.549
Simulation Model	7211.1	1.50%	23.838	4.79%	42.7	-0.47%	4.425	0.41%	0.095	72.103
AMLOS,IP Regression Model	7267.9	0.70%	25.465	-1.90%	46.6	-8.80%	4.543	-2.20%	0.421	49.289
AMLOS, IP, Age Regression Model	7292.6	0.36%	25.099	-0.47%	45.5	-6.59%	4.517	-1.49%	0.562	48.925

All three models are comparable and represent the original data well, with percent differences for most statistics on the order of 1-3%.

4.6 Sensitivity Analysis

To test the sensitivity of the model coefficients, confidence intervals were calculated for each of them. This was accomplished by taking a random sample (without replacement) of half the total patients (3660 out of 7319) and in turn deriving a new model with this "new" pool of data. Since it is assumed these coefficients are normally distributed, the following formula was used to calculate the CIs:

$$\bar{Y} \pm t_{\alpha/2,R-1}\sqrt{\frac{S^2}{R}} \tag{4-7}$$

where S^2 is the sample variance and R is the number of replicates. Using R = 100, and $\alpha = 0.05$, the CIs are as shown below in Table 4.10:

Table 4.10 Confidence Intervals of the Model Coefficients

	AMLOS, Age, IP			AMLOS,IP		
	Y bar	+/-	Percent Diff	Y bar	+/-	Percent Diff
β_0	-1.652	0.109	6.60%	-1.434	0.076	5.30%
β_1	0.954	0.038	3.98%	0.957	0.043	4.49%
β_2	0.086	0.011	12.79%	0.094	0.009	9.57%
β_3	0.007	0.001	14.29%	-------	-------	-------

The percentage difference values represent half of the confidence interval and indicate how much the model coefficient is expected to vary in each direction. As expected, the variability of the coefficient estimates is generally less for the 2-regressor model compared to the 3-regressor values. To study the potential effects of this variability, both lower and upper bound estimates were simulated to see if any of the major simulation statistics (average LOS and patient level) from Table 4.9 vary significantly. While the overall patient levels did not change (the patient arrival distribution remains the same), Table 4.11 below shows how average LOS varied widely from 3.976-5.187 days compared to the narrow 4.517-4.543 day range (Table 4.9) of the original models.

Table 4.11 Comparison of Confidence Interval Bounds on Original Regression Models

	AMLOS, Age, IP Model			AMLOS,IP Model		
	Model Coefficients	Average LOS (Days)	Max Patients	Model Coefficients	Average LOS (Days)	Max Patients
Lower Bound (of CI)	β_0 = -1.761 β_1 = 0.916 β_2 = 0.075 β_3 = 0.006	3.976	45	β_0 = -1.510 β_1 = 0.914 β_2 = 0.085 -----	4.012	43
Original Regression Model	β_0 = -1.680 β_1 = 0.951 β_2 = 0.081 β_3 = 0.006	4.534	62	β_0 = -1.450 β_1 = 0.956 β_2 = 0.092 -----	4.556	52
Upper Bound (of CI)	β_0 = -1.543 β_1 = 0.992 β_2 = 0.097 β_3 = 0.008	5.167	> 150	β_0 = -1.358 β_1 = 1.000 β_2 = 0.103 -----	5.187	118

The level of patients in the upper bound once reached the maximum resource level in ARENA. This causes the software to terminate prematurely and thus prevented the determination of a true maximum at the upper bound. Regardless, this value is excessively larger than the top mark (62) of the original regression models. Clearly these results illustrate two things. Not only are the model coefficients very sensitive to change, but also the original model coefficients developed earlier appear to be estimated accurately, as shown by the data from the simulation runs in Table 4.9.

V. Discussion

5.1 Conclusions

The preceding research was performed to determine what, if any, relationship exists between general patient indicators and LOS in a neurology unit at a local hospital. A wide variety of analysis techniques were used to study, develop, and test regression models that will predict actual LOS more accurately than the hospital's current system.

The process involved analyzing 6 various patient predictors, three related to the patient's diagnosis (RW, AMLOS, and GMLOS) which are designated as DRG regressors. The remaining three non-DRG regressors correspond directly to the patient (age, gender, and IP). The data consists of 7319 patients treated in-residence at Good Samaritan Hospital in Dayton, OH, from January 2002 to June 2005.

Regression analysis involved comparing two different approaches: ordinary (OLS) and weighted least squares (WLS). OLS was employed for its ease of implementation and as a baseline for comparing the three subsequent models. WLS was used as a result of extensive analysis showing increasing variability in LOS as the DRG regressors increased. Three various WLS methods were applied.

Developing these models involved studying all possible models using all available regressors and using Mallow's C_p statistic to rank them. Normally, minimizing Mallow's C_p statistic over all possible regression models will determine the best subset model. However, it was established early in the research process of the desire to develop a relatively simple, maintainable model. Serious consideration was therefore given to a minimal-regressor model with a comparatively low Mallow's C_p value.

This analysis involved not only determining the top performing models, but more importantly, feasible ones. Analysis determined that all three DRG regressors are highly linearly dependent and, consequently, highly collinear. This typically translates into the model coefficients being poorly estimated, which was eventually demonstrated using Variance Inflation Factors. It was ultimately concluded the final models must be limited to having a single DRG-regressor.

With this restriction in place and also presuming this particular DRG regressor would be the model's most significant, choosing the most significant regressor is critical. Throughout the entire analysis, AMLOS consistently outperformed the two remaining DRG regressors (RW and GMLOS). This is consistent with the policies of most health-care facilities as AMLOS is typically used to predict LOS, while RW and GMLOS (indirectly) are used for calculating treatment costs.

For the non-DRG regressors, IP was the most significant, appearing in virtually all of the top candidate models. In one of the final models, LOS varies over 1.75 days when using two particular IPs. However, since there is very little information known about each IP's policies due to privacy issues, it can only be concluded that there is indeed a relationship, but any possible account or explanation as to why there is cannot be offered at this time. Any attempt to do so at this point would be pure speculation.

The two remaining non-DRG regressors, age and gender, were less significant, although age was eventually used in one of the models. A possible reason for this is that many DRGs are only applicable to certain age groups, which means age has already been accounted for in many diagnoses. The insignificance of gender, on the other hand, may easily be explained by the nervous system, unlike other systems of the human body, being effectively the same for each gender, and therefore will tend to have similar diagnoses, treatments, and recovery times.

The WLS methods were very effective, with many of the associated WLS models often outperforming OLS. The drawback to WLS is that these models generally require more maintenance and upgrades than OLS. However, WLS provides the user with a robust and powerful, yet simple approach to developing a regression model with nonconstant error variance that OLS cannot offer. Furthermore, as more patient data becomes available, the WLS models will continue to be improved considerably.

Each of the WLS approaches adds a particular element to regression analysis. WLS 1 (modeled variance) can be used in cases where little information about the total variance is known. If enough points are known or can be determined by some method, an effective variance model can be developed to account for these unknown variances. WLS 2 (probability model), on the other hand, is a very effective, albeit less theoretical, approach to use when there is little or insufficient information known about the variance by simply “weighting” each observation based on its frequency of occurring. WLS 3 (isotonic) is a more innovative approach to modeling error variance, where a step function of the variance is developed by minimizing a convex function of the original data. Its one drawback is that it must be thoroughly shown that the error variance is either nonincreasing (antitonic) or nondecreasing (isotonic).

Because of its simplicity and ease of maintainability, the recommended model to implement is WLS 2. As more patient data becomes available, the various weights can easily be adjusted to continuously update and improve the model. Because WLS 1 and

WLS 3 consist of variance models, updating them is much more involved as the entire model has to be redeveloped to derive the new variance weights.

The final regression models are relatively simple, highly effective and have been extensively analyzed and tested. Statistical data from their simulation runs was compared to both that of the original patient data and a simulation model developed from the raw patient data. Results are quite favorable, with most statistics, such as average LOS and patient level, within 1-3% of one another. Furthermore, these models are easily maintainable and updated as more patient and IP data becomes available.

The neurology department at Good Samaritan Hospital can now use these models to more effectively predict LOS for an incoming patient. Inputting the patient's AMLOS, age, and IP into the regression model generates an expected LOS that is much more accurate than their current guidelines for predicting LOS. Furthermore, the respective simulation models offer a powerful tool for analyzing various what-if scenarios:

- How does the system react to the anticipated increase in neurological patients?
- What is the optimal number of beds that ensures minimal occurrence of patient overflow?
- How is average LOS and average patient level impacted by small adjustments to the model?

This research also offers GSH a basis to possibly investigate the disparities in LOS using two different IPs. Certain IPs may have policies that possibly induce a shorter or longer LOS. If so, it may be possible to alter or revise these procedures.

5.2 Future Research

Because this patient data is used primarily for billing purposes, the information it offers is rather limited. Ideally, more descriptive information of the patient would be desired, such as his or her past medical history and lifestyle (history of tobacco use, recent obesity problems, etc). Furthermore, an admitted patient may have several additional symptoms that may or may not impact LOS. However, policy states the patient can only be charged for one, and only one, diagnosis.

Therefore, the finance office only tracks the one diagnosis code corresponding to the billing process. Nevertheless, this will usually be the one reflecting the most substantial and costliest treatment(s) the patient receives, and therefore will generally impact LOS more significantly than the others. However, if the patient has several additional diagnosis codes, it could significantly impact LOS.

Also, since there has been a significant relationship established between patient IP and LOS, more information about each of the individuals IPs is desired. As mentioned in Chapter 3, although each of the IPs is charged the *actual* amount it pays varies. If certain details about these various payment methods were known, it may better explain the impact IP has on LOS.

Appendix A

List of Acronyms

ADL	Activity of Daily Living
AMLOS	Arithmetic Mean LOS
ANOVA	Analysis-of-Variance
CI	Confidence Interval
CMS	Centers for Medicare and Medicaid Services
DRG	Diagnostic Related Group
GMLOS	Geometric Mean LOS
GSH	Good Samaritan Hospital (Dayton, OH)
HMO	Health Maintenance Organization
IP	Insurance Provider
LOS	Length-Of-Stay
MATLAB	Matrix Laboratory
MWD	Minimum Walking Distance
PIE	Person-In-Environment
PPS	Prospective Payment System
PRESS	Prediction Error Sum of Squares
QL	Quality of Life
RW	Relative Weight
SAS	Statistical Analysis System
SPSS	Statistical Package for the Social Sciences
SS_R	Regression Sum of Squares
SS_{Res}	Residual Sum of Squares
SS_T	Total Sum of Squares
VIF	Variance Inflation Factor
WLS	Weighted Least Squares

Appendix B

The following are the derivations for the Regression Sum of Squares (SS$_{\mathbf{R}}$) and the Residual Sum of Squares (SS$_{\mathbf{Res}}$) formulas, respectively:

$$
\begin{aligned}
SS_R &= (\hat{\mathbf{y}} - \mathbf{1}\bar{\mathbf{y}})^T \mathbf{V}^{-1} (\hat{\mathbf{y}} - \mathbf{1}\bar{\mathbf{y}}) \\
&= (\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{1}\bar{\mathbf{y}})^T \mathbf{V}^{-1} (\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{1}\bar{\mathbf{y}}) \\
&= (\mathbf{X}(\mathbf{X}^T \mathbf{V}^{-1}\mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1}\mathbf{y} - \mathbf{1}\bar{\mathbf{y}})^T \mathbf{V}^{-1} (\mathbf{X}(\mathbf{X}^T \mathbf{V}^{-1}\mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1}\mathbf{y} - \mathbf{1}\bar{\mathbf{y}})
\end{aligned}
$$

$$
\begin{aligned}
SS_{\mathrm{Res}} &= \mathbf{e}^T \mathbf{V}^{-1} \mathbf{e} \\
&= (\mathbf{y} - \hat{\mathbf{y}}^T)\mathbf{V}^{-1}(\mathbf{y} - \hat{\mathbf{y}}^T) \\
&= \mathbf{y}^T \mathbf{V}^{-1}\mathbf{y} - \mathbf{y}^T \mathbf{V}^{-1}\hat{\mathbf{y}} - \hat{\mathbf{y}}^T \mathbf{V}^{-1}\mathbf{y} + \hat{\mathbf{y}}^T \mathbf{V}^{-1}\hat{\mathbf{y}} \\
&= \mathbf{y}^T \mathbf{V}^{-1}\mathbf{y} - \mathbf{y}^T \mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{V}^{-1}\mathbf{y} + \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \mathbf{y}^T \mathbf{V}^{-1}\mathbf{y} - 2\hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{V}^{-1}\mathbf{y} + \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \mathbf{y}^T \mathbf{V}^{-1}\mathbf{y} - 2[(\mathbf{X}^T \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T \mathbf{V}^{-1}\mathbf{y}]^T \mathbf{X}^T \mathbf{V}^{-1}\mathbf{y} + [(\mathbf{X}^T \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T \mathbf{V}^{-1}\mathbf{y}]^T \mathbf{X}^T \mathbf{V}^{-1}\mathbf{X}[(\mathbf{X}^T \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T \mathbf{V}^{-1}\mathbf{y}] \\
&= \mathbf{y}^T \mathbf{V}^{-1}\mathbf{y} - 2\mathbf{y}^T \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T \mathbf{V}^{-1}\mathbf{y} + \mathbf{y}^T \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T \mathbf{V}^{-1}\mathbf{y} \\
&= \mathbf{y}^T \mathbf{V}^{-1}\mathbf{y} - 2\mathbf{y}^T \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T \mathbf{V}^{-1}\mathbf{y} + \mathbf{y}^T \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T \mathbf{V}^{-1}\mathbf{y} \\
&= \mathbf{y}^T \mathbf{V}^{-1}\mathbf{y} - \mathbf{y}^T \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T \mathbf{V}^{-1}\mathbf{y}
\end{aligned}
$$

Appendix C

The data used in this research consists of the following Insurance Providers:

Insurance Provider	# of Patients
AETNA	64
ANTHEM MANAGED	545
ANTHEM TRADITIONAL	51
ASA	208
AUTO	63
BWC REHAB	67
CIGNA	38
CONTRACTED COMMERCIAL	182
HUMANA	36
MEDICAID	560
MEDICAID HMO	63
MEDICARE	3731
MEDICARE HMO	412
MMO	112
NONCONTRACTED COMMERCIAL	39
OTHER GOVERNMENTAL	13
PHCS	66
SELF PAY	393
UHC MANAGED	521
UHC TRADITIONAL	155
Total Patients	7319

Appendix D

The following is the MATLAB code used in this research:

```
% This function runs all possible regression models
function [x] = RunAllModels()

        % Read in data matrix
        [datain] = textread('AMLOS_ulsx_rawy.txt');
        DataMatrix = datain;

        % Set number of bins
        r1 = 85;
        r2 = 97;
        r3 = 95;
        r4 = 92;
        r5 = 2;
        r6 = 20;

        % Run all 32 Regression Models
        % AMLOS
        Stats(1,:)   = RunModel(1, DataMatrix, 1, r1);
        % AMLOS, RW
        Stats(2,:)   = RunModel(2, DataMatrix, 1, r1, 2, r2);
        % AMLOS, GMLOS
        Stats(3,:)   = RunModel(2, DataMatrix, 1, r1, 3, r3);
        % AMLOS, Age
        Stats(4,:)   = RunModel(2, DataMatrix, 1, r1, 4, r4);
        % AMLOS, Gender
        Stats(5,:)   = RunModel(2, DataMatrix, 1, r1, 5, r5);
        % AMLOS, IP
        Stats(6,:)   = RunModel(2, DataMatrix, 1, r1, 6, r6);
        % AMLOS, RW, GMLOS
        Stats(7,:)   = RunModel(3, DataMatrix, 1, r1, 2, r2, 3, r3);
        % AMLOS, RW, Age
        Stats(8,:)   = RunModel(3, DataMatrix, 1, r1, 2, r2, 4, r4);
        % AMLOS, RW, Gender
        Stats(9,:)   = RunModel(3, DataMatrix, 1, r1, 2, r2, 5, r5);
        % AMLOS, RW, IP
        Stats(10,:) = RunModel(3, DataMatrix, 1, r1, 2, r2, 6, r6);
        % AMLOS, GMLOS, Age
        Stats(11,:) = RunModel(3, DataMatrix, 1, r1, 3, r3, 4, r4);
        % AMLOS, GMLOS, Gender
        Stats(12,:) = RunModel(3, DataMatrix, 1, r1, 3, r3, 5, r5);
        % AMLOS, GMLOS, IP
        Stats(13,:) = RunModel(3, DataMatrix, 1, r1, 3, r3, 6, r6);
        % AMLOS, Age, Gender
        Stats(14,:) = RunModel(3, DataMatrix, 1, r1, 4, r4, 5, r5);
        % AMLOS, Age, IP
        Stats(15,:) = RunModel(3, DataMatrix, 1, r1, 4, r4, 6, r6);
        % AMLOS, Gender, IP
        Stats(16,:) = RunModel(3, DataMatrix, 1, r1, 5, r5, 6, r6);
```

```
% AMLOS, RW, GMLOS, Age
Stats(17,:) = RunModel(4, DataMatrix, 1, r1, 2, r2, 3, r3, 4, r4);
% AMLOS, RW, GMLOS, Gender
Stats(18,:) = RunModel(4, DataMatrix, 1, r1, 2, r2, 3, r3, 5, r5);

% AMLOS, RW, GMLOS, IP
Stats(19,:) = RunModel(4, DataMatrix, 1, r1, 2, r2, 3, r3, 6, r6);
% AMLOS, RW, Age, Gender
Stats(20,:) = RunModel(4, DataMatrix, 1, r1, 2, r2, 4, r4, 5, r5);
% AMLOS, RW, Age, IP
Stats(21,:) = RunModel(4, DataMatrix, 1, r1, 2, r2, 4, r4, 6, r6);
% AMLOS, RW, Gender, IP
Stats(22,:) = RunModel(4, DataMatrix, 1, r1, 2, r2, 5, r5, 6, r6);
% AMLOS, GMLOS, Age, Gender
Stats(23,:) = RunModel(4, DataMatrix, 1, r1, 3, r3, 4, r4, 5, r5);
% AMLOS, GMLOS, Age, IP
Stats(24,:) = RunModel(4, DataMatrix, 1, r1, 3, r3, 4, r4, 6, r6);
% AMLOS, GMLOS, Gender, IP
Stats(25,:) = RunModel(4, DataMatrix, 1, r1, 3, r3, 5, r5, 6, r6);
% AMLOS, Age, Gender, IP
Stats(26,:) = RunModel(4, DataMatrix, 1, r1, 4, r4, 5, r5, 6, r6);
% AMLOS, RW, GMLOS, Age, Gender
Stats(27,:) = RunModel(5, DataMatrix, 1, r1, 2, r2, 3, r3, 4, r4, 5, r5);
% AMLOS, RW, GMLOS, Age, IP
Stats(28,:) = RunModel(5, DataMatrix, 1, r1, 2, r2, 3, r3, 4, r4, 6, r6);
% AMLOS, RW, GMLOS, Gender, IP
Stats(29,:) = RunModel(5, DataMatrix, 1, r1, 2, r2, 3, r3, 5, r5, 6, r6);
% AMLOS, RW, Age, Gender, IP
Stats(30,:) = RunModel(5, DataMatrix, 1, r1, 2, r2, 4, r4, 5, r5, 6, r6);
% AMLOS, GMLOS, Age, Gender, IP
Stats(31,:) = RunModel(5, DataMatrix, 1, r1, 3, r3, 4, r4, 5, r5, 6, r6);
% AMLOS, RW, GMLOS, Age, Gender, IP
Stats(32,:) = RunModel(6, DataMatrix, 1, r1, 2, r2, 3, r3, 4, r4, 5, r5, 6, r6);

% Set n
n = length(randMat);

% Total number of regression models
totalruns = 32;

% Generate matrix of 32 p-values corresponding to each regression model
p = [1 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 6];

% Set values for all 32 models
for(i = 1:totalruns)
        for(j = 1:4)
                % R^2 value
                R2(i,j) = Stats(i,1+6*(j-1));

                % Adjusted R^2 value
                adjR2(i,j) = Stats(i,2+6*(j-1));

                % Mean Square Error
                MS(i,j) = Stats(i,4+6*(j-1));
```

```
                    % Sum of Squares Total
                    SStot(i,j) = Stats(i,5+6*(j-1));

                    % PRESS Statistic
                    PRESS(i,j) = 1 - ((Stats(i,6+6*(j-1)))/(SStot(i,j)));

                    % Mallow's Cp statistic
                    Mallows_Cp(i,j) = ((Stats(i,3+6*(j-1)))/Stats(totalruns,4+6*(j-1))) - n +
                                                                                (2*(p(i)+1));
          end
     end
return


% This function generates a random list of patients
function z = GenDataMatrix(In, size_of_matrix, width)

     % Set a temp list from 1 to n
     List = zeros(size_of_matrix,width);

     % Set j to size of list
     j = length(In);

     % For 1 to size of list
     for(i = 1:size_of_matrix)

          % Find a random element and move it to the list
          index = ceil((j)*rand(1));
          List(i,:) = In(index,:);

          % Fill in the newly opened slot
          for(n = index:(length(In)-1))
               In(n,:) = In(n+1,:);
          end;

          % Update j
          j = j - 1;
     end

     % Return list
     z = List;
return;


% This function generates the Model (Beta) coefficients, ANOVA table, and residual plots
function z = RunModel(number_of_reg, InMatrix, x1, t1, x2, t2, x3, t3, x4, t4, x5, t5 )

     % Set y column
     y = 7;

     % Generate X and Y Matrices
     X(:,1) = ones(length(InMatrix),1);
     X(:,2) = InMatrix(:,x1);
```

```
Y       = InMatrix(:,y);
if(number_of_reg >= 2)
            X(:,3) = InMatrix(:,x2);
end
if(number_of_reg >= 3)
            X(:,4) = InMatrix(:,x3);
end


if(number_of_reg >= 4)
            X(:,5) = InMatrix(:,x4);
end
if(number_of_reg >= 5)
            X(:,6) = InMatrix(:,x5);
end
if(number_of_reg >= 6)
            X(:,7) = InMatrix(:,x6);
end

% Calculate Variances and Weights and set V and W Vectors
VarsAndWeights = GenerateVarsAndWeights(InMatrix, x1, t1, y);
V  = VarsAndWeights(:,1);
W = VarsAndWeights(:,2);

% Derive Variance Model for WLS 2
G = GLMfit(X(:,2),V);
Var = X(:,1:2)*G;

% Set n
n = length(VarsAndWeights);

% Generate VV, WW, and VarVar (nxn) matrices from the V, W, And Var vectors
VV = zeros(n,n);
WW = zeros(n,n);
VarVar = zeros(n,n);
for(i = 1:n)
            % Elements cannot be exactly zero
            if(V(i,1) == 0)
                        VV(i,i) = 0.00000001;
            else
                        VV(i,i) = V(i,1);
            end
            if(W(i,1) == 0)
                        WW(i,i) = 0.00000001;
            else
                        WW(i,i) = 1/W(i,1);
            end
            if(Var(i,1) == 0)
                        VarVar(i,i) = 0.00000001;
            else
                        VarVar(i,i) = Var(i,1);
            end
end
```

```
% Model 1: OLS Model ------------------------------------------

% Calculate Beta coefficients
OLS_Model = (inv(X'*X))*X'*Y

% Calculate ANOVA Table
ANOVA(1,1:6) = ANOVAstats(OLS_Model,X,eye(n),Y,n,number_of_reg,1);

% Calculate standardized and studentized residuals
sta1 = Stand_res(OLS_Model,X,Y,ANOVA(1,4),n);
stu1 = Stud_res(OLS_Model,X,eye(n),Y,ANOVA(1,4),n,number_of_reg);

% Calculate y_hats
for(i = 1:n)
        y_hat(i,1) = X(i,:)*OLS_Model;
end

% Plot residuals
figure (1)
subplot(1,2,1)
scatter(y_hat(:,1), sta1),title('M1: OLS Model Residuals'),
                                        xlabel('y_hat'), ylabel('Standardized residual');
subplot(1,2,2)
scatter(y_hat(:,1), stu1),title('M1: OLS Model Residuals'),
                                        xlabel('y_hat'),ylabel('Studentized residual');

% Model 2: Modeled Variance ----------------------------------

% Calculate Beta coefficients
Var_Model = inv(X'*inv(VarVar)*X)*X'*inv(VarVar)*Y

% Calculate ANOVA Table
ANOVA(1,7:12) = ANOVAstats(Var_Model,X,VarVar,Y,n,number_of_reg,2);

% Calculate standardized and studentized residuals
sta2 = Stand_res(Var_Model,X,Y,ANOVA(1,10),n);
stu2 = Stud_res(Var_Model,X,VarVar,Y,ANOVA(1,10),n,number_of_reg);

% Calculate y_hats
for(i = 1:n)
        y_hat(i,1) = X(i,:)*Var_Model;
end

% Plot residuals
figure (2)
subplot(1,2,1)
scatter(y_hat(:,1), sta2),title('M2: Variance Model Residuals'),
                                        xlabel('y_hat'),ylabel('Standardized residual');
subplot(1,2,2)
scatter(y_hat(:,1), stu2),title('M2: Variance Model Residuals'),
                                        xlabel('y_hat'),ylabel('Studentized residual');

% Model 3: Probability Model ----------------------------------
% Calculate Beta coefficients
Prob_Model = inv(X'*inv(WW)*X)*X'*inv(WW)*Y

% Calculate ANOVA Table
ANOVA(1,13:18) = ANOVAstats(Prob_Model,X,WW,Y,n,number_of_reg,2);

% Calculate standardized and studentized residuals
sta3 = Stand_res(Prob_Model,X,Y,ANOVA(1,16),n);
stu3 = Stud_res(Prob_Model,X,WW,Y,ANOVA(1,16),n,number_of_reg);

% Calculate y_hats
for(i = 1:n)
        y_hat(i,1) = X(i,:)*Prob_Model;
end

% Plot residuals
figure (3)
subplot(1,2,1)
scatter(y_hat(:,1), sta3),title('M3: Probability Model Residuals'),
                                        xlabel('y_hat'),ylabel('Standardized residual');
subplot(1,2,2)
scatter(y_hat(:,1), stu3),title('M3: Probability Model Residuals'),
                                        xlabel('y_hat'),ylabel('Studentized residual');

% Model 4: Isotonic Variance -----------------------------------

% Calculate Isotonic vector
Isot = pavaI(V(:,1));

% Generate Isotonic matrix
Isot_matrix = zeros(n,n);
for(i = 1:n)
        % Elements cannot be exactly zero
        if(Isot(i) == 0)
                Isot_matrix(i,i) = 0.00000001;
        else
                Isot_matrix(i,i) = Isot(i);
        end
end

% Calculate Beta coefficients
Isot_Model = inv(X'*inv(Isot_matrix)*X)*X'*inv(Isot_matrix)*Y

% Calculate ANOVA Table
ANOVA(1,19:24) = ANOVAstats(Isot_Model,X,Isot_matrix,Y,n,number_of_reg,2);

% Calculate standardized and studentized residuals
sta4 = Stand_res(Isot_Model,X,Y,ANOVA(1,22),n);
stu4 =  Stud_res(Isot_Model,X,Isot_matrix,Y,ANOVA(1,22),n,number_of_reg);

% Calculate y_hats
for(i = 1:n)
        y_hat(i,1) = X(i,:)*Isot_Model;
        end

        % Plot residuals
        Figure (4)
        subplot(1,2,1)
        scatter(y_hat(:,1), sta4),title('M4: Isotonic Model Residuals'),
                                            xlabel('y_hat'),ylabel('Standardized residual');
        subplot(1,2,2)
        scatter(y_hat(:,1), stu4),title('M4: Isotonic Model Residuals'),
                                            xlabel('y_hat'),ylabel('Studentized residual');

        % Return ANOVA table
        z = ANOVA;

return
% This function generates the ANOVA table
function z = ANOVAstats(Model,X,V,Y,n,k,mode)

        % Calculate y_hats
        for(i = 1:length(Y))
                Y_hat(i,1) = X(i,:)*Model;
        end

        % Calculate vector of y_bar’s
        sum = 0;
        for(i = 1:length(Y))
                sum = sum + Y(i,1);
        end
        average = sum/length(Y);
        for(i = 1:length(Y))
                Y_bar(i,1) = average;
        end

        % Calculate SSr and SS_res
        SSr     = (Y_hat-Y_bar)T*inv(V)*(Y_hat-Y_bar)
        SS_res  = Y'*inv(V)*Y - Y'*inv(V)*X*inv(X'*inv(V)*X)*X'*inv(V)*Y;

        % Create ANOVA Table
        ANOVA = zeros(3,5);
        ANOVA(1,1) = SSr;
        ANOVA(2,1) = SS_res;
        ANOVA(3,1) = SSr + SS_res;
        ANOVA(1,2) = k+1;
        ANOVA(2,2) = n - (k+1);
        ANOVA(3,2) = n;
        ANOVA(1,3) = ANOVA(1,1)/ANOVA(1,2);             % Mean Square Regression
        ANOVA(2,3) = ANOVA(2,1)/ANOVA(2,2);             % Mean Square Residual
        ANOVA(1,4) = ANOVA(1,3)/ANOVA(2,3);             % F statistic

        % Set output statistics
        Out(1,1) = SSr/(SSr + SS_res)                   % R^2
        Out(1,2) = 1-((n-1)/(n-k-1))*(1-Out(1,1));      % adj R-squared
        Out(1,3) = ANOVA(2,1);                          % Sum of Squares residual
        Out(1,4) = ANOVA(2,3);                          % Mean Square
        Out(1,5) = ANOVA(3,1);                                  % Sum of Squares total
        Out(1,6) = 1 - (PRESS(Model,X,V,Y,n,k,mode))/Out(1,5);  % PRESS statistic

        % Return statistics
        z = Out;
return


% This function computes the standardized residuals
function z = Stand_res(Model,X,Y,MS_res,n)
        for(i = 1:n)
                residuals(i) = (Y(i,1) - (X(i,:)*Model))/sqrt(MS_res);
        end
        z = residuals;
return


% This function computes the studentized residuals
function z = Stud_res(Model,X,V,Y,MS_res,n,k)

        % Calculate H and covariance matrices
        H = X*inv(X'*inv(V)*X)*X'*inv(V);
        cov = (eye(n)-H)*V*(eye(n)-H)T;

        for(i = 1:n)
                residuals(i) = (Y(i,1) - (X(i,:)*Model))/(sqrt(MS_res*cov(i,i)));
        end
        z = residuals;
return


% This function computes the PRESS statistic
function z = PRESS(Model,X,V,Y,n,k,mode)

        % Calculate H matrix
        H = X*inv(X'*inv(V)*X)*X'*inv(V);

        % OLS
        if(mode == 1)
                newH = H;
        else % WLS 2
                newH = (eye(n)-H)*V*(eye(n)-H)T;
        end

        % Calculate PRESS statistic
        PRESS_stat = 0;
        for(i = 1:n)
                PRESS_stat = PRESS_stat + ((Y(i,1) - (X(i,:)*Model))/(1 - newH(i,i)))^2;
        end
        z = PRESS_stat;
return


% This function calculates the variance and probability weight matrices
```

```
function z = GenerateVarsAndWeights( In, x1, t1, y )

        % Generate Used vector
        Used = zeros(length(In),1);

        % Set lower bound
        low = -1;

        % For each bin
        for(j = 1:t1)

                % Set upper bound and reset counter
                upp = -1 + (2*j/t1);
                cntr = 0;

                % Traverse entire x1 vector
                for(k = 1:length(In))

                        % If element falls within bounds and has not already been used
                        if((Used(k,1) == 0) && (In(k,x1) >= low) && (In(k,x1) <= upp))

                                % Increment counter
                                cntr = cntr + 1;

                                % Load y and patient values
                                yVector(cntr,1) = In(k,y);
                                patientVector(cntr,1) = k;

                                % Set element as being used
                                Used(k,1) = 1;
                        end
                end

                % For all applicable elements in this bin
                for(k = 1:cntr)
                        % Set Variance and weight values
                        Out(patientVector(k,1),1) = var(yVector(1:cntr,1));
                        Out(patientVector(k,1),2) = cntr;
                end

                % Reset y and patient vectors
                for(k = 1:cntr)
                        yVector(k,1) = 0;
                        patientVector(k,1) = 0;
                end
        end

        % Return matrix
        z = Out;
return
```

Bibliography

1 Blais, Mark A., John Matthews, Robin Lipkis-Orlando, Erin Lechner, Michelle Jacobo, Robert Lincoln, Christina Gulliver, John B. Herman, Alyson F. Goodman. “Predicting length of an acute care medical psychiatric inpatient service,” *Administration and Policy in Mental Health*, 31(1):15-29 (September 2003).

2 Bohannon, RW, N. Lee, R. Maljanian. “Postadmission Function Best Predicts Acute Hospital Outcomes After Stroke’” *American Journal of Physical Medicine and Rehabilitation*, 81:726-730 (2002).

3 Brousseau, Lucie, Pierre Philippe, Louise Potvin, Yves-Louis Boulanger. “Post-stroke inpatient rehabilitation. I. Predicting length of stay,” *American Journal of Physical Medicine and Rehabilitation*, 75:422-430 (1996).

4 Burns, LR, RW. Douglas. “The effects of patent, hospital, and physician characteristics on length of stay and mortality,” *Medical Care*, 29:251-271 (1991).

5 Centers for Medicare and Medicaid Services. *Federal Register.* 67(148): 49981-50289 (August 2002).

6 Christina, A, A. Allevi, E. Taioli, N. Anzalone, A. Nicolosi, E. Polli. “Analysis of diagnostic procedure costs for cerebrovascular disease admission to a highly specialized hospital,” *Italian Journal of Neurological Science*, 12:397-405 (1991).

7 Evers, Silvia, Gemma Voss, Fred Nieman, Andre Ament, Tom Groot, Jan Lodder, Anita Boreas, Gerhard Blaauw. “Predicting the cost of hospital stay for stroke patients: the use of diagnostic related groups,” *Health Policy*, 61:21-42 (2002).

8 Hakim, EA, AMO Bakheit. “A study of factors which influence the length of hospital stay of stroke patients,” *Clinical Rehabilitation*, 12:151-156 (1998).

9 Herman, JM, L. Culpepper, P. Franks. “Patterns of utilization, disposition, and length of stay among stroke patients in a community hospital setting,” *Journal of the American Geriatrics Society*, 32:421-426 (1984).

10 Inouye, M. “Predicting outcomes of patients in Japan after first acute stroke using a simple model,” *American Journal of Physical Medicine and Rehabilitation,* 80:645-649 (2001).

11 Keefler, Joan, Sydney Duder, Constance Lechman. “Predicting Length of Stay in an Acute Care Hospital: The Role of Psychosocial Problems,” *Social Work in Health Care,* 33(2):1-16 (2001).

12 Kleijnen, Jack P.C. “An overview of the design and analysis of simulation experiments for sensitivity analysis,” *European Journal of Operational Research*, 164:287-300 (2005).

13 Lagoe, RJ, SL Lauko. “Stroke hospitalization under prospective payments: analysis of diagnosis related group 14. *Archives of Physics and Medical Rehabilitation*. 66:773-776 (1985).

14 Lanska, DJ. “Length of hospital stay for cerebrovascular disease in the United States: Professional Activity Study, 1963-1991,” *Journal of Neurological Science.* 127:214-220 (1994).

15 Lee, AJ, JH Huber, WB Stason. “Factors contributing to practice variation in post-stroke rehabilitation,” *Health Service Research,* 32:197-221 (1997).

16 McKenna, K, L. Tooth, J. Strong, K. Ottenbacher, J. Connell, M. Cleary. “Predicting discharge outcomes for stroke patients in Australia,” *American Journal of Physical Medicine and Rehabilitation*. 81:47-56 (2002).

17 Monane, Mark, Daniel S. Kanter, Robert J. Glynn, Jerry Avorn. “Variability in length of hospitalization for stroke: The role of managed care in an elderly population,” *Archives of Neurology,* 53:875-880 (1996).

18 Montgomery, Douglas C, Elizabeth A. Peck, G. Geoffrey Vining, *Introduction to Linear Regression Analysis.* John Wiley & Sons, Inc. 2001.

19 National Institutes of Health. *National Institute of Neurological Disorders and Stroke.* 5 January 2006. http://www.ninds.nih.gov/

20 Omachonu, V. K., S. Suthummanon, M. Akcin, S. Asfour. “Predicting length of stay for Medicare patients at a teaching hospital,” *Health Services Management Research*, 17:1-12 (2004).

23 Osberg, JS, SM Haley, GE McGinniss, G. Dejong. “Characteristics of cost outliers who did not benefit from stroke rehabilitation,” *American Journal of Physical Medicine and Rehabilitation*, 69:117-125 (1990).

24 Parsons, Janet A., Michael R. Johnston, Arthur A. Slutsky. “Predicting length of stay out of hospital following lung resection using preoperative health status measures,” *Quality of Life Research*. 12:645-654 (2003).

25 Perry, Marcus B., J.J. Pignatiello, Jr., J.R. Simpson. “Change Point Estimation for Monotonically Changing Poisson Rates in SPC.” International Journal of Production Research (2006). To appear.

26 Wade, Derick T., Richard Langton Hewer. “Hospital admission for acute stroke: who, for how long, and to what effect,” *Journal of Epidemiology and Community Health*, 39:347-352 (1985).

27 Weingarten, JP, JC Clay, DA Heckert. “Impact of socioeconomic status on health care utilization: factors influencing length of stay,” *Journal of Health and Human Services Administration*. 19(4):384-409 (1997).

Vita

Captain Tony A. Murphy graduated from Deubrook High School in White, SD. He enlisted in the Air Force a month later and, after receiving F-16 avionics systems training at Lowry AFB, CO, and Shaw AFB, SC, was assigned to the 57th Aircraft Generation Squadron at Nellis AFB, NV. Later he was accepted to the USAF Aerial Demonstration Squadron, "The Thunderbirds", where he was an aircraft avionics specialist and later a crew chief on one of the aircraft. In October 1993, he was assigned to the 36th Fighter Squadron at Osan AB, Republic of Korea, before later transferring to the 39th Flight Test Squadron at Eglin AFB, FL, in November 1994, where he participated in the testing and development of several premiere aircraft weapon systems.

While at Eglin, he earned an Associate of Arts degree in Pre-Engineering from Okaloosa-Walton Community College in Niceville, FL, in May 1998. He was then accepted into the Airman Education and Commissioning Program and began undergraduate studies at the University of Florida in Gainesville, FL, where he later graduated Magna Cum Laude with dual Bachelor of Science degrees in Electrical Engineering and Computer Engineering in December 2000. He was commissioned through Officer Training School at Maxwell AFB, AL, in April 2001.

He was then assigned to the Air Force Research Laboratory at Kirtland AFB, NM, as a member and later team lead of a high-powered microwave weapon development team. In August 2004, he entered the Graduate School of Engineering and Management, Air Force Institute of Technology, Wright-Patterson AFB, OH. Upon graduation, he will be assigned to Air Combat Command Headquarters at Langley AFB, VA.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 074-0188*

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of the collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to an penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* 23-03-2006	**2. REPORT TYPE** **Master's Thesis**	**3. DATES COVERED** *(From – To)* Sep 2005 - Mar 2006

4. TITLE AND SUBTITLE ANALYSIS OF PATIENT INFORMATION: AN EMPIRICAL MODELING APPROACH	**5a. CONTRACT NUMBER** **5b. GRANT NUMBER** **5c. PROGRAM ELEMENT NUMBER**
6. AUTHOR(S) Murphy, Tony, A., Captain, USAF	**5d. PROJECT NUMBER** **5e. TASK NUMBER** **5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S) Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 Hobson Street, Building 642 WPAFB OH 45433-7765	**8. PERFORMING ORGANIZATION REPORT NUMBER** AFIT/GOR/ENS/06-14
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSOR/MONITOR'S ACRONYM(S)** **11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

12. DISTRIBUTION/AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

With rising costs and increasing complexities, many hospitals seek to better understand the intricate details of their operations. Increasingly, these organizations have a strong desire to accurately predict the resources required to effectively treat their patient load. This research investigates patient length-of-stay in a hospital neurological unit using an empirical modeling approach. Factors significantly affecting patient length of stay were identified and used to construct a regression model. The predictive model provides hospital decision makers with a compact tool to input what-if scenarios and predict future patient treatment lengths, thus, allowing the hospital to properly allocate resources.

15. SUBJECT TERMS
Health, Regression Analysis, Mathematic Prediction, Isotonic Regression

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Sharif H. Melouk (ENS)
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-3636, ext 4525; e-mail: Sharif.Melouk@afit.edu
U	U	U	UU	78	