Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Alessandria Y. Shen Date

Predicting Low Back Pain Following Lumbar Interbody Fusion: A Comparison of Multivariate Model Selection Criteria

By

Alessandria Y. Shen
Master of Science in Public Health

Department of Biostatistics & Bioinformatics
Rollins School of Public Health
Emory University

Paul S. Weiss
Faculty Thesis Advisor

Predicting Low Back Pain Following Lumbar Interbody Fusion: A Comparison of Multivariate Model Selection Criteria

By

Alessandria Y. Shen
B.A. Mathematics
Emory University
2011

Paul S. Weiss, MS
Faculty Thesis Advisor

An abstract of
a thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements of the degree of
Master of Science in Public Health
in Biostatistics & Bioinformatics
2012

Abstract

Developing an adequate predictive regression model and indentifying appropriate functional and interaction relationships between its predictors is perhaps one of the most challenging tasks in regression analysis. Specifically, linear models applied to a high-dimensional dataset, such as those commonly found in clinical settings, are usually over-parameterized, unclear for interpretation, and have high variances for estimated parameters. Furthermore, traditionally used hypothesis testing frameworks for selection criteria have proven to be inconclusive and unreliable. The goal of this thesis was to overcome these challenges by integrating clinical experience with statistical theory and develop empirical predictive models using different variable selection criteria.

This thesis also contributes to research in health-related quality of life (HRQOL) outcomes for the assessment of lumbar interbody fusion (IBF) efficacy. Research regarding statistical methods in this field is preliminary and has been limited to univariate analysis. In this thesis, pre-operative predictors associated with 3-month post-operative low back pain (LBP) improvement were determined and multivariate linear regression models were selected for the data from the Georgia Spine Patient Outcomes Registry. The results were very sensitive to the selection criteria used, as well as noise variables included in the full dataset. Further work in the analysis of lumbar IBF HRQOL outcomes needs to be done to transform the knowledge and teaching base from theory and person experience to one of statistical evidence.

Predicting Low Back Pain Following Lumbar Interbody Fusion: A Comparison of Multivariate Model Selection Criteria

By

Alessandria Y. Shen
B.A. Mathematics
Emory University
2011

Paul S. Weiss, MS
Faculty Thesis Advisor

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements of the degree of
Master of Science in Public Health
in Biostatistics & Bioinformatics
2012

Table of Contents

1 Introduction 1

1.1 Low Back Pain 1

1.2 Lumbar Interbody Fusion 3

1.3 Evaluation of Efficacy 3

1.4 Study Motivation & Goals 8

2 Statistical Methods 11

2.1 Multivariate Linear Regression 11

2.2 Model Selection 12

2.3 Stepwise Selection 13

2.4 Mallows' C_p Criterion 15

2.5 Model Validation 18

3 Study Materials 20

3.1 Data Source 20

3.2 Predictor Variables 21

3.3 Outcome Variable 21

3.4 Sample Selection 23

4 Analysis & Results 24
4.1 Univariate Analysis 24
4.2 Stepwise Selection 26
4.3 Mallows' C_p Criterion 28
4.4 Model Validation 30

5 Discussion 31

References 33

A HRQOL Instruments 40
A.1 Numeric Rating Scale (NRS) 40
A.2 Oswestry Disability Index (ODI) 40
A.3 SF-36v2$^{\text{TM}}$ (SF-36) 41

B Proofs 42
B.1 Proof of Equation (2.7) 42
B.2 Lemma B.2 43
B.3 Proof of Equation (2.8) 44

C Supplemental Materials 45
C.1 Results of Univariate Analysis 46

List of Tables

Table 1.1: Example outcome measures relevant to the assessment of lumbar IBF efficacy for treating LBP. 5

Table 1.2: Inter-observer agreement of radiographic measures compared to that of HRQOL measures. 6

Table 1.3: Psychometric properties of HRQOL instruments and their corresponding tests statistics. 7

Table 1.4: Selected examples of commonly used HRQOL instruments in LBP patient populations. 9

Table 3.1: List and description of potential predictor variables, not including pre-operative HRQOL measures. 22

Table 4.1: Parameter estimates of models selected from stepwise selection procedures. 26

Table 4.2: Summary statistics for candidate models fit on training on training and validation sets. 30

Table C.1: Results of univariate analysis. 46

List of Figures

Figure 1.1: X-rays of common degenerative lumbar spine disorders. 2

Figure 1.2: Lumbar IBF in the treatment of degenerative scoliosis. 4

Figure 4.1: Scatter-plot matrix of pre-operative HRQOL variables. 25

Figure 4.2: Model 1 residual analysis of normality and heteroscedasticity... 27

Figure 4.3: Model 2 residual analysis of normality and heteroscedasticity... 27

Figure 4.4: Model 3 residual analysis of normality and heteroscedasticity... 27

Figure 4.5: Model 4 residual analysis of normality and heteroscedasticity... 28

Figure 4.6: Model 5 residual analysis of normality and heteroscedasticity... 28

Figure 4.7: C_p plot of best 50 models for each model size. 29

Figure 4.8: $\overline{C}_p$ plot of best 50 models for each model size. 29

Acknowledgements

First and foremost, to Dr. Kaveh Khajavi, thank you for giving me the amazing opportunity to work at Georgia Spine and for teaching me so much about medicine (and the real world). Without your strong dedication to clinical research and providing the best care for your patients, this project (among many, *many* others) would not have been possible.

To Professor Paul Weiss, thank you for your guidance and patience through the writing of this thesis.

To Felix, thank you for providing love and words of encouragement when I needed it the most.

Finally, to my parents, I am grateful beyond words for all that you have done for me. I would not be where I am today without your love and support.

Chapter 1

Introduction

1.1 Low Back Pain

Low back pain (LBP) is a common musculoskeletal disorder marked by pain in the lumbar region (L1-L5 vertebrae) and in some instances, can radiate to the buttocks, hips, and upper thigh region. LBP affects nearly 80% of all adults at some point in their lives[1] and is the most frequently reported type of pain in the United States[2], with an estimated 54 million American adults experiencing LBP of at least one day in duration every three months.[1]

LBP is classified according to the duration of symptoms from the time of onset to the time of resolution. Pain is classified as acute if the duration is less than 4 weeks, sub-acute if between 4 and 12 weeks, and chronic if greater than 12 weeks. Although the majority of LBP cases resolve within 8 to 12 weeks of onset, symptoms become chronic in up to 15% of patients, resulting in episodes of intense pain, significant physical limitations, and decreased quality of life (QOL).[2]

1.1.1 Cost Burden

The estimated economic burden of LBP in the U.S. ranges from $84.1 billion to $624.8 billion including both direct and indirect costs. Lost work productivity is the largest component of this economic burden, resulting in indirect costs ranging from $7.4 billion to $28 billion.[3,4] LBP also results in substantial direct medical costs associated with the

use of healthcare resources, including primary care physician or specialist visits, as LBP is the second most cited reason for physician visits in the US. Overall, cases of chronic LBP account for the majority of the cost burden associated with LBP.[3-5]

1.1.2 Pathophysiology

The etiology of LBP symptoms varies by classification, or duration of symptoms. Cases of acute or sub-acute LBP are typically due to isolated incidences of trauma or acute injury in otherwise healthy and active individuals. On the other hand, the prevalence of chronic LBP is highest for individuals aged between 45 and 64 years (23.7%)[1], and thus, these cases usually originate from one or more degenerative lumbar spine disorders.

Common degenerative lumbar spine disorders that result in chronic LBP are shown in Figure 1.1. Due to aging, inter-vertebral discs may become desiccated and lose height, resulting in a degradation of the stress-buffering affect of the disc. This can lead to abnormal motion, instability, and rapid wearing of the vertebral body endplates. Decreased interbody height results in change of the diameter of the bony foramina where the spinal nerve roots exit, causing compression and thereby leading to painful symptoms in the lower extremities.

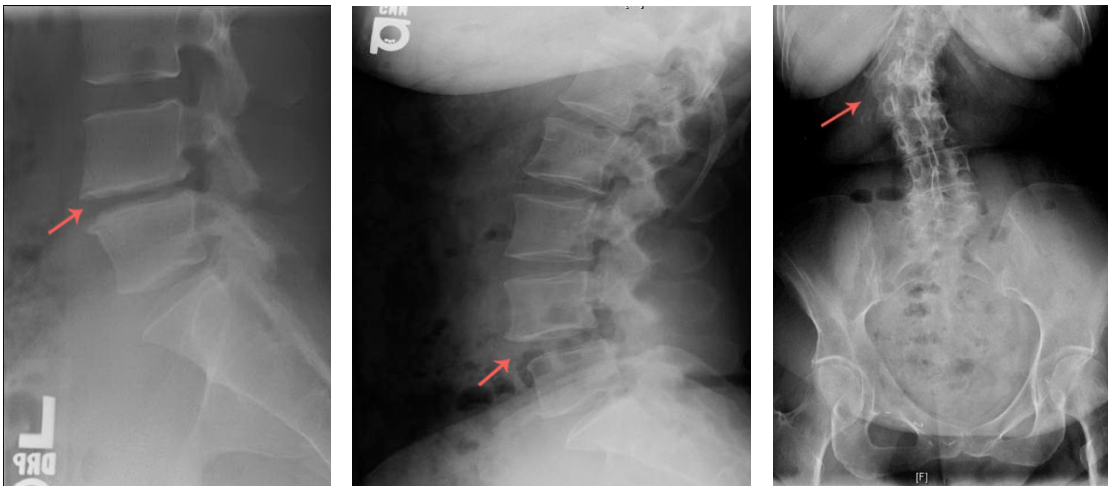

Figure 1.1: X-rays of common degenerative lumbar spine disorders. (Left) DDD resulting in severe L4-L5 disc space collapse; (middle) degenerative spondylolisthesis marked by anterior displacement of L4 vertebral body over L5; (right) degenerative scoliosis with apex of coronal deformity at L1-L2.

1.2 Lumbar Interbody Fusion

While non-surgical treatments such as physical therapy and pain management (oral medication or steroid injections) exist for LBP, cases of acute or sub-acute LBP respond the most rapidly and favorably to these conservative treatments.[6] Surgical intervention is typically used to treat individuals with chronic LBP and severe degenerative lumbar spine disorder pathologies.

Lumbar interbody fusion (IBF) is a surgical procedure that involves a discectomy, or complete removal of the inter-vertebral disc, and placement of a bone-graft filled cage to fuse two vertebral bodies (Figure 1.2). The purpose of fusion is to immobilize the degenerated vertebrae and eliminate the abnormal motion that leads to pain. Although effective, lumbar IBF has the potential to be extremely invasive, often requiring up to two weeks post-operative hospital stay and three months rehabilitation.[6]

In patients who have additional leg pain or severely compressed nerve roots, a posterior decompression may be performed in addition to IBF. The decompression includes a direct mini-open approach to the posterior vertebra and removal of any bone or thickened ligaments that may be compressing on the nerve root. Finally, posterior instrumentation, such as percutaneous pedicle screw and rod fixation, may also be placed to further stabilize the spine.

1.3 Evaluation of Efficacy

The efficacy, usefulness, and value of a medical intervention are vital pieces of information for all parties involved in the healthcare system. Particularly in recent years as the current U.S. health economy has shifted towards managed care and other reforms intended to lower costs, an increased emphasis has been placed on assessing the benefits of an intervention against its costs and risks.[8-10] As a result, there is growing need in the healthcare field for standardized methods of evaluating treatment utility and effectiveness.

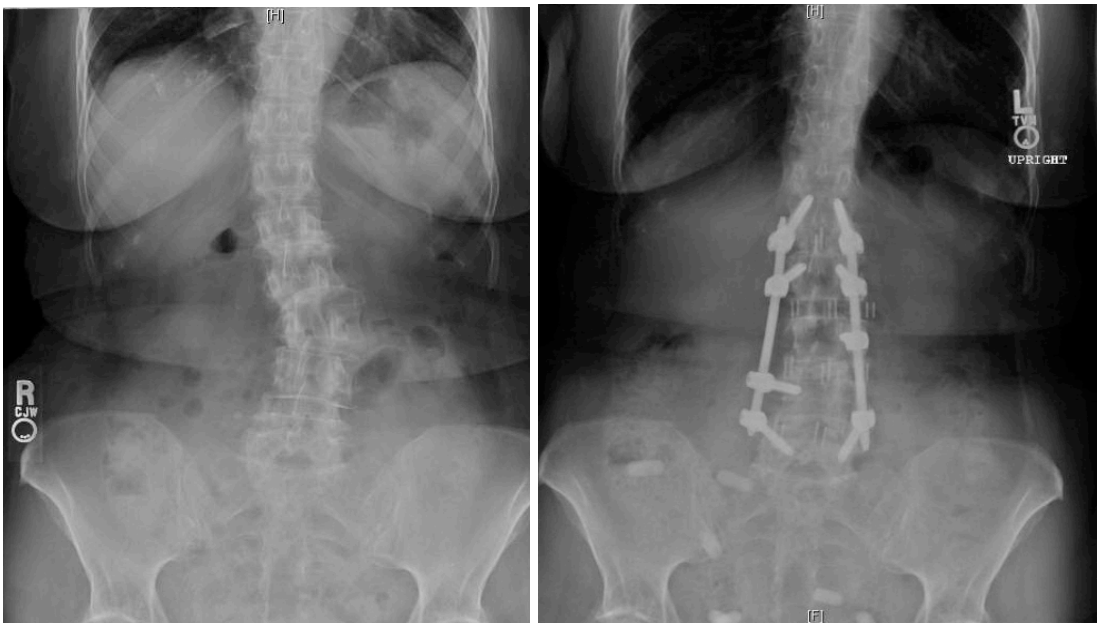

Figure 1.2: Lumbar IBF in the treatment of degenerative scoliosis. Pre-operative (left) and two post-operative lumbar IBF (right) with additional pedicle screw and rod fixation.

The first step in evaluating a medical intervention is to define a "successful outcome." For the treatment of diseases such as hypertension or hypercholesterolemia, a successful outcome can be defined using objective measures such as biomarkers. In other diseases that lack these objective measures, a successful outcome has traditionally been defined as the absence of pathology. However, many clinicians and researchers have come to the realization that this definition alone cannot comprehensively capture the goal of an intervention. They propose that the definition of outcomes must extend beyond the absence of pathology and also measure *health-related QOL* (HRQOL) factors, which include psychological and social well being.[8]

1.3.1 HRQOL Outcomes in Lumbar IBF

Perhaps the need for HRQOL outcomes is most evident in the treatment of chronic LBP using lumbar IBF, where the primary motivation for undergoing the procedure is often chronic LBP and excess disease burden on QOL. Although patient outcomes are multi-dimensional (as suggested in Table 1.1), lumbar IBF efficacy has predominantly been evaluated using physiologic, anatomic, and radiographic measures.

Category	Measure
Physiologic	Spine or extremity range of motion
	Muscle EMG activity
	Spinal fluid endorphin levels
	Muscle strength, endurance
Anatomic	Solid fusion mass
	Disc height
	Vertebral displacement
Physical Exam	Neurologic deficits
	Straight leg raising
Symptoms	Pain duration, severity, frequency
	Neurologic symptoms
Functional Status	Activities of daily living
	Psychological function
	Recreational activities
	Social function
	Health perceptions/general well-being
Role Function	Employment status
	Disability compensation
	Days of work absenteeism
	Days of limited activity

Table 1.1: Example outcome measures relevant to the assessment of lumbar IBF efficacy for treating LBP. Measures listed in the first three categories are considered "objective" measures, and those in the last three are considered "subjective" HRQOL outcomes.

If objective measures prove to be closely correlated with HRQOL outcomes, it may be justifiable to only collect the former. However, Deyo *et al.*[20] state that dissociations between the two are extremely common in the treatment of musculoskeletal conditions. In other words, while HRQOL outcomes are, on average, better for patients who achieve a solid fusion or experience substantial correction of their disease pathology, it is also a common observation that a patient may experience an improvement in LBP even in the case of pseudarthrosis. Conversely, a patient may report little or no change in HRQOL despite achieving an excellent physiologic outcome. Thus, in order to adequately assess the effectiveness and benefit of lumbar IBF, relevant HRQOL outcomes must be collected and analyzed as an individual entity instead of attempting to infer upon them from readily available information.

One major reason HRQOL measures have been rejected in the past may be due to its perceived "subjective" nature. Bias is, indeed, one of the major challenges of measuring HRQOL, as there is no way to reflect or account for the possibility of a patient exaggerating symptoms or disabilities. However, this is not to say that objective measures are immune to many of the same extraneous factors, as depression also confounds many physiologic measures, such as range of motion and strength. Deyo *et al.*[20-23] illustrate this bias using inter-observer variability in the interpretation of radiographic measures, which are the most widely used objective measures in lumbar IBF efficacy (Table 1.2). This data suggests that "subjective" HRQOL measures prove to be as reliable, if not more, as objective radiographic measures.

Category	Measure	Kappa
Lumbar Spine Interpretation by Musculoskeletal Radiologists	Any abnormality	0.51
	Facet joint sclerosis	0.33
	Any narrowed discs	0.49
Self-Administered Patient Questionnaire	Sickness impact profile	0.87
	Medical history	0.79

Table 1.2: Inter-observer agreement of radiographic measures compared to that of HRQOL measures.

1.3.2 HRQOL Measurement Instruments

Instruments used to measure and quantify HRQOL can be classified as either *generic* or *disease-specific*.[11] Generic instruments are related to health in general can be used regardless of the underlying pathology or intervention. One advantage of a generic instrument is that it allows comparisons to be made across a variety of disease populations.[12,13] In contrast, disease-specific instruments can only measure HRQOL attributable to a specific disease or pathology. These instruments are generally sensitive to subtle changes that may not otherwise be captured using generic instruments. For the majority of studies that include HRQOL outcomes, both classes of measures are used to optimize the amount of information that is acquired.

There are a vast number of both generic and disease-specific instruments that can be used to monitor a patient's HRQOL during an intervention. The adequacy of an instrument for a specified research goal is evaluated based on its psychometric criteria of validity, reliability, and sensitivity.[10-12] A description of these criteria and their corresponding statistical tests are shown in Table 1.3.

Property	Characteristic	Test Statistic
Reliability	Internal consistency	Cronbach's alpha
	Inter-rater reliability	Cohen's kappa
	Reproducibility	Intraclass correlation coefficient
Validity	Concurrent validity	Pearson correlation
	Predictive validity	Messick's model
Sensitivity	Able to detect change	Guyatt's statistic

Table 1.3: Psychometric properties of HRQOL instruments and their corresponding test statistics.

Selected examples of commonly used generic and LBP-specific HRQOL measurement instruments are presented in Table 1.4. Extensive research[11,13-18] has found that the following instruments are most pertinent for evaluating lumbar IBF for the treatment

of LBP: Numeric Rating Scales, Oswestry Disability Index, and SF-36. The construct and psychometric properties of these instruments are discussed in Appendix A.

1.4 Study Motivation & Goals

Research in the efficacy of lumbar IBF has shifted towards determining factors that affect post-operative HRQOL outcomes. A wide range of variables have been found in the literature as predictive factors for outcomes following lumbar IBF.[18-26] While some factors have been generally accepted, other cited predictive factors are contradictory, and thus, strong conclusions are difficult to make. These discrepancies may be a consequence of the hypothesis testing frameworks most commonly used in analysis, as research regarding statistical methodology in this field is very limited.

Many researchers[24,30-32] have cited another research challenge as the lack of large, clinically and radiographically well-defined patient databases. To date, most existing large databases are claims-based and use ICD (Internal Classification of Diseases) codes to collect patient information. Numerous studies across various medical specialties have shown the misclassifications and inaccuracies that result from this method of data collection.[33,34] Furthermore, these databases are typically geared towards providing financial information to healthcare payers and hospitals, and thus, do not contain meaningful health status outcomes data.

This thesis seeks to fill the present gaps in knowledge by integrating clinical experience with statistical theory. The primary objective is to develop multivariate regression models using different selection criteria to predict the magnitude of a patient's LBP improvement following lumbar IBF. Secondary objectives are to statistically validate and compare the models' predictive abilities. Fulfilling these objectives and developing an appropriate predictive model could potentially help patients and doctors assess the benefits of using lumbar IBF to treat LBP.

Category	Instrument	Dimension
Pain Intensity	Numeric Rating Scales	Frequency
	Visual Analog Scales	Frequency
	Chronic Pain Grade	Perceived impact
	McGill Pain Questionnaire	Affective response
	Dallas Pain Score	Daily activities, work, leisure, social, etc.
Disease-Specific Functional Status	Roland-Morris Disability Scale	Various daily functions
	Oswestry Disability Index	Self-care, lifting, walking, standing, social, etc.
	Million Instrument	Various daily functions
Generic Functional Status	Short Form-36	Physical function, role function, pain, etc.
	Sickness Impact Profile	Ambulation, self-care, work, recreation, etc.
	Duke Health Profile	Social, psychological, physical
	Health Interview Survey	Days work absenteeism, days in bed, etc.
	Nottingham Health Profile	Physical mobility, pain, sleep, social, etc.

Table 1.4: Selected examples of commonly used HRQOL measurement instruments in LBP patient populations.

The remainder of this thesis is organized as follows:

- Chapter 2 presents the model selection process and its underlying theory. The proposed selection criteria, along with their derivations and utilization procedures, are also presented. Model validation techniques are also presented.
- Chapter 3 presents the dataset and other materials used in this thesis.
- Chapter 4 presents the results of model selection using each of the proposed criteria, as well as the validation of the identified models.
- Chapter 5 concludes the thesis with a discussion of the findings and suggestions for future research directions.

Chapter 2

Statistical Methods

2.1 Multivariate Linear Regression

The model assumed in this study is a multivariate linear regression model, given by

$$y = X\beta + \varepsilon \tag{2.1}$$

where

$$\begin{cases} y = n \times 1 \text{ response vetor} \\ X = n \times p \text{ design matrix} \\ \beta = p \times 1 \text{ coefficient vector} \\ \varepsilon = \text{error terms, where } \varepsilon \sim N(0, \sigma^2) \end{cases}$$

Classical assumptions for a linear regression model include:

1. **Linearity.** The mean of the response variable is a linear combination of the regression coefficients and the predictor variables.
2. **Homoscedasticity.** The variance of the error term is constant across all response variables.
3. **No multicollinearity.** All predictor variables are linearly independent, i.e. there is no correlation between any two predictors.

Satisfying these assumptions imply that the model's parameter estimates will be unbiased, consistent, and efficient. However, linear regression methods can still be used even when these assumptions are not true, as is the case in virtually all real-world datasets.[52]

2.2 Model Selection

The goal of model selection is to distinguish between authentic variables that are important in predicting a given outcome, and the noise variables that possess little to no predictive value.[53]

One key consideration in model selection is the size, or number of parameters, of the model.[52] In clinical settings, models with fewer predictors are easier to manage and use, as it requires less information to be collected. From a statistical standpoint, retaining excess predictors in a model tends to increase the variances of its estimated parameters. Similarly, the variances of the fitted values may also increase, causing a worsening of the model's predictive ability. On the other hand, eliminating or neglecting to include key predictors can also lead to serious consequences, most notably, biased estimates of regression coefficients and the error variance. This bias stems directly from the fact that error terms in an underfit model may reflect nonrandom effects of the omitted predictors. In these cases, a single predictor may have great explanatory power for observed variances in the response variable.[52] Ideally, a reasonable subset of variables is selected so that the resulting model is simple, parsimonious, and has strong predictive ability.

Another important point to consider is the criterion used to for model selection. Each criterion has its own advantages and disadvantages, and can directly affect the conclusions that are drawn. Because there is currently no convention in lumbar IBF HRQOL outcomes research regarding the selection criterion, this study uses the p-value-based stepwise selection procedures and the all-possible-regressions procedure, selecting variables based on Mallows' C_p.

2.3 Stepwise Selection

The stepwise regression approaches (forward stepwise, forward, backward) are based on standard hypothesis-testing frameworks, employing automatic search procedures to evaluate potential predictors based on *p*-values calculated from partial significance tests against a predetermined significance level, α. In datasets with a large number of potential predictors, this method is advantageous in that it provides a quick and simple way of assessing the significance of all predictors. While the theory upon which this approach is based is very widely understood and used, particularly in clinical spine research, this procedure also has several major drawbacks.[52]

First, because the predictor variables are evaluated in a linear fashion, stepwise selection procedures are only able to identify a single regression model. To put into perspective the magnitude of information lost in this procedure, consider that a dataset with 50 potential predictor variables will have 2^{50}, or 1.25×10^{15} possible first-order subset models.[54] Consequently, this approach often produces results that are misleading and counterintuitive with established clinical knowledge.

Another weakness of this procedure pertains to the determination of α. In basic hypothesis testing, it has become conventional to use $\alpha = 0.05$ or 0.10. In the context of model selection, however, there is currently no precise or methodical way of setting a value for α that balances significance with the power of the test against the alternative hypothesis.[54] This is an extremely important point to consider, as different levels of α may also lead to different conclusions.

2.3.1 Utilization

Forward Stepwise Selection

1. A simple linear regression (SLR) model is fit for each potential X variable:

$$Y = \beta_0 + \beta_1 X_1$$

2. For each SLR model, the F-statistic testing $H_0 : \beta_1 = 0$ is obtained:

$$F_k = \frac{SSR(X_k)}{MSE(X_k)}$$

3. The X_k with the largest F_k is added into the model, given that its corresponding p-value does not exceed α. If no X_k fit the criteria, then the model selection procedure is terminated, indicating that none of the potential variables are adequate for the regression model.
4. Suppose X_a is first entered into the model. The remaining potential X variables are each fit into the model already containing X_a:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

5. For each new model, the partial F testing $H_0 : \beta_2 = 0$ is obtained:

$$F_k = \frac{SSR(X_k \mid X_a)}{MSE(X_k, X_a)}$$

6. The X_k with the largest $F_{k \mid X_a}$, given that its corresponding p-value does not exceed α, is the second X variable added. Otherwise, the model selection procedure is terminated.
7. Suppose X_b is the second variable entered into the model. At this point, the procedure examines if any previously added X variables already in the model should be dropped by obtaining the following partial F statistic:

$$F_k = \frac{SSR(X_a \mid X_b)}{MSE(X_a, X_b)}$$

8. If the p-value corresponding to $F_{a \mid X_b}$ exceeds α, then X_a is dropped from the model. Otherwise, X_a is retained.
9. Steps 4-8 are then repeated for each of the remaining potential X variables until no further X variables can be added or deleted, at which point the procedure is terminated and the final model is identified.

Forward Selection

This procedure is a simplified version of forward stepwise selection, omitting the test whether previously entered variables should be dropped.

Backward Selection

1. A model containing all $p - 1$ potential X variables is fit and a partial F statistic is obtained for each variable. For example, the partial F statistic for X_1 is given by:

$$F_k = \frac{SSR(X_1 | X_2, \ldots, X_{p-1})}{MSE(X_1, X_s, \ldots, X_{p-1})}$$

2. The X_k with the smallest F_k is dropped from the model, given that its corresponding p-value exceeds α.
3. This step is repeated for all remaining potential X variables until no further variables can be dropped.

2.4 Mallows' C_p Criterion

In contrast to the stepwise selection procedures, where only a single model is identified, the *all-possible-regressions* procedure selects a subset of adequate models. All possible combinations of the predictors are evaluated against a full model that contains the complete set of potential predictors using the C_p statistic.[52] This selection method also directly addresses the concerns of model size and reduces the risk of overfitting by placing a "penalty" for increasing the number of X variables in a model, reflected in the model's C_p value.[55]

The major drawback of the C_p statistic is that since it is based on least squares estimation, it is very sensitive to outliers and other departures from the normality assumption of the error term. Additionally, it assumes the full model was developed such that its total MSE provides an unbiased estimator of σ^2. Fulfilling this assumption requires careful development of the k potential X variables, where important interactions are included and noise variables are excluded.[55] This can prove to be challenging, particularly in research fields where there is little background literature that can assist in this evaluation.[53] The determination of predictors to include in the full model is then essentially at the full discretion of the researcher.

2.4.1 Derivation

Suppose dataset D has a total of n observations and k potential X variables. The full model, Y, contains all k potential X variables and is given by

$$Y_i = \beta_0 + \sum_{i=1}^{k} \beta_j X_{ji} + \varepsilon_i \text{ for } i = 1, \ldots, n \tag{2.2}$$

where $\varepsilon \sim N(0, \sigma^2)$.

Let $\widehat{Y}_i$ be the i'th fitted value and $\widehat{\mu}_i$ be the true mean response for the i'th subject. The number of X variables in a given subset model, Y^*, is denoted by $p - 1$, such that $0 \leq p - 1 \leq k$.

The squared *total error* for $\widehat{Y}_i$, denoted by $(\widehat{Y}_i - \mu_i)^2$, is defined as

$$(\widehat{Y}_i - \mu_i)^2 = \left[\left(\mathrm{E}\{\widehat{Y}_i\} - \mu_i\right) + \left(\widehat{Y}_i - \mathrm{E}\{\widehat{Y}_i\}\right)\right]^2 \tag{2.3}$$

where

$$\begin{cases} \mathrm{E}\{\widehat{Y}_i\} - \mu_i = \text{model error for the } i\text{'th fitted value} \\ \widehat{Y}_i - \mathrm{E}\{\widehat{Y}_i\} = \text{random error for the } i\text{'th fitted value} \end{cases}$$

The mean squared error (MSE) for $\widehat{Y}_i$ is obtained by taking the expectation of equation (2.3), which reduces to

$$MSE(\widehat{Y}_i) = \left(\mathrm{E}\{\widehat{Y}_i\} - \mu_i\right)^2 + \sigma^2\{\widehat{Y}_i\} \tag{2.4}$$

where $\sigma^2\{\widehat{Y}_i\}$ is the variance of the i'th fitted value.

Summing equation (2.4) across all n gives the total MSE for the full model, or

$$MSE(X_1, \ldots, X_k) = \sum_{i=1}^{n} \left(\mathrm{E}\{\widehat{Y}_i\} - \mu_i\right)^2 + \sum_{i=1}^{n} \sigma^2\{\widehat{Y}_i\} \tag{2.5}$$

The criterion measure for a subset model Y^*, denoted by Γ_p, is obtained by dividing equation (2.5) by σ^2, the true error variance, or

$$\Gamma_p = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}\left(\mathrm{E}\{\hat{Y}_i\} - \mu_i\right)^2 + \sum_{i=1}^{n}\sigma^2\{\hat{Y}_i\}\right] \tag{2.6}$$

If it can be assumed that Y was chosen such that equation (2.5) is an unbiased estimator of σ^2, then C_p becomes an estimator of Γ_p (see Appendix B.1 for proof). Thus, the C_p value for subset model Y^* is given by:

$$C_p = \frac{SSR_p}{MSE(X_1,\ldots,X_k)} - (n - 2p) \tag{2.7}$$

where

$$\begin{cases} SSR_p = \text{residual sum of squares of } Y^* \\ p = \text{number of parameters in } Y^* \text{ (intercept } + p - 1) \end{cases}$$

2.4.2 Utilization

The "best" subset model is one where the total bias of the predicted values is minimized. When there is no bias in subset model Y^* so that $\mathrm{E}\{C_p\} \equiv \mu_i'$, then the expected value of $C_p \approx p$. The C_p value for full model with q parameters (i.e. $q - 1$ X variables) is, by definition, p. When the C_p value for each possible subset model is plotted against p, those models with little bias will fall near the line $C_p = p$. Models with substantial bias will fall above this line, and models below this line are interpreted as having no bias.

However, if the assumption that the MSE of the full model is an unbiased estimator of σ^2 is not met, then the values of C_p will be small. Additionally, the expectation of C_p will likely be negative when the number of potential predictors in the full model is significantly fewer than the number of observations. In response to these issues, Gilmour[56] proposed a modified statistic, denoted by $\overline{C}_p$, where

$$\overline{C}_p = C_p - \frac{2(k - p + 1)}{n - k - 3} \tag{2.8}$$

The expected values of $\overline{C}_p$ for models with little or no bias will be approximately p. This proof is given in Appendix B.3.

2.5 Model Validation

Model validation identifies the "best" predictive model by measuring a model's predictive performance in an independent dataset (known as the *validation sample*). One such measure of predictive ability is a model's *coefficient of determination*, denoted by R^2. The R^2 statistic measures the fraction of the total variability in the outcome variable that can be explained by the model and provides a measure of how well future outcomes are likely to be predicted by the model. Possible values of R^2 fall between 0 and 1, inclusive, with a value of 1 indicating that the model can explain all variability of the outcome variable. However, it is important to note that a high R^2 value alone is not sufficient to deem a model "adequate," as R^2 will inherently increase as more predictors are added to a model. The *adjusted* R^2 statistic addresses this problem by taking into consideration both the number of predictors in the models and the sample size.[57]

A model's predictive ability can also be measured using residuals. Recall that a model's MSE quantifies the difference between predicted and observed values, and is given by dividing its SSR by the total number of observations. In the context of model validation, if a model has good predictive power, then the MSE of fitting the model on a validation set should be approximately equal to its MSE obtained from the *training sample*, or the dataset upon which the model was selected.[52] The MSE obtained from the validation set is more commonly referred to as the *mean squared prediction error* (*MSPR*), and is given by

$$MSPR = \frac{SSR}{n^*} \tag{2.9}$$

where n^* is the number of observations in the validation sample.

Cross-validation is a commonly used validation method and entails partitioning a dataset into complementary subsets, performing the analysis on the training sample and validating the analysis on the validation sample.[58] To reduce variability, this process may be repeated a number of times, each time using different partitions. The *MSPR*'s are then averaged across all repetitions. One major disadvantage of cross-validation is the loss of precision that results from reducing the size of the training sample. This is evidenced by the fact that variances of the estimated regression coefficients developed from the training set will usually be larger than those obtained from the entire dataset, and thus this method usually requires the original dataset to be large enough so that this difference is minimized.[52,58,59]

To obtain unbiased least squares estimates of the regression coefficients, the model needs to be validated externally, wherein the validation sample is obtained by collecting new data. When a model is selection from given data, it is implied that the selected model fits the data well. However, for models used to predict future outcomes, this validation method is comparatively more useful as it assesses the model under broader circumstances than those related to the original data.[52]

Chapter 3

Study Materials

3.1 Data Source

Data for this study was obtained from the Georgia Spine Patient Outcomes Registry (Georgia Spine and Neurosurgery Center, LLC, Atlanta, GA). The GA Spine registry is comprised of all patients who underwent elective lumbar and/or cervical spine surgery at a single private neurosurgery practice starting from May 2006. All patients were de-identified and assigned a unique six-digit ID number upon entering into the registry.

The registry records variables pertaining to patient demographics, previous medical history, diagnoses, summary of surgical procedure, and any intra-operative and post-operative complications. This information was retrospectively adjudicated from previous medical records, physician assessments, operative notes, and anesthesiology reports.

HRQOL measures were obtained prospectively using standard paper forms, which are kept on record for a minimum of 5 years. These measures were collected pre-operative and prospectively at the following time points post-operative: 1 month, 3 months, 6 months, 12 months, 24 months, and 5 years. Scores for these HRQOL measures were calculated by an independent clinical data management company (PhDx Systems, Inc., Albuquerque, NM) using standard algorithms. The data is quality controlled and HIPAA (Health Insurance Portability and Accountability Act of 1996) compliant.

3.2 Predictor Variables

This database was created specifically for studies pertaining to lumbar IBF outcomes, and thus, predictor variables recorded were chosen according to clinical judgment and existing knowledge. With the exception of those pertaining to post-operative complications, all 32 available predictors from the database were used in the analysis. These predictors are listed and briefly described in Table 3.1. In addition, to those listed in the table, the following pre-operative HRQOL measures were also included in the pool of potential predictor variables: BPNRS, LPNRS, ODI, SF-36 PCS, and SF-36 MCS.

3.3 Outcome Variable

The predicted HRQOL outcome in this study is the difference between baseline BPNRS and 3-month post-operative BPNRS score. A positive value of the outcome variable indicates post-operative improvement, while a negative value indicates worsening of symptoms. This specific HRQOL measure was chosen for several reasons:

1. LBP is the primary indication for patients undergoing lumbar IBF and the primary goal of lumbar IBF is to resolve LBP.
2. The typical rehabilitation time for lumbar IBF patients is up to 3 months. Procedure-related pain and weakness may confound measures taken prior to this time point.
3. Patient reporting of leg pain may be confounded by inability to discriminate between degenerative lumbar spine disorder-specific pain and other unrelated pain such as knee pain, which is fairly typical in an older population.
4. ODI was not chosen because level of disability is often time-dependent, as patients tend to experience a certain amount of increase in disability over time as they age. Additionally, the primary goal of lumbar IBF is not to increase disability, but rather, to resolve pain.
5. SF-36 was not chosen because it is a generic HRQOL instrument and not specific to LBP or degenerative lumbar spine disorders.

Category	X	Variable	Description
Demographics	1	AGE	Age at time of surgery (years)
	2	BMI	Body mass index
	3	EDU	Highest education level
	4	GENDER	Gender
	5	SMOKE	Current smoking status
Co-Morbidities (Y/N)	6	ASTH	Asthma
	7	BLEED	Bleeding disorder
	8	COPD	COPD
	9	DEPRESS	Depression
	10	DIABETES	Diabetes mellitus
	11	GERD	Gastroesophageal reflux disease
	12	HDL	Hypercholesterolemia
	13	HEART	Cardiovascular disease
	14	HEP	Hepatitis
	15	HTN	Hypertension
	16	OA	Osteoarthritis
	17	OSP	Osteoporosis
	18	SEIZURE	Seizure
	19	STROKE	Stroke
	20	ULCER	Stomach ulcer
Surgical Summary	21	DIAGNOSIS	Primary indication for surgery
	22	EBL	Estimated blood loss (mL)
	23	FIX	Posterior fixation (Y/N)
	24	FUS	Fusion approach
	25	LOS	Length of post-op stay (days)
	26	NUMLEVS	# Levels fused
	27	OR_TIME	Length of operation (min)
	28	PRIOR	Prior lumbar surgery (Y/N)

Table 3.1: List and description of potential predictor variables, not including pre-operative HRQOL measures.

3.4 Sample Selection

All patients who underwent lumbar IBF performed by a single neurosurgeon were identified and evaluated for inclusion based on the following eligibility requirements:

- Baseline and 3-month post-operative BPNRS scores were available.
- Primary indication for surgery was LBP (BPNRS > 2).
- No reported intra- or post-operative complications.

From these eligible patients, a training sample and a validation sample were assembled. The training sample included 177 patients who underwent surgery between January 2007 and October 2011. The validation sample included 32 patients who underwent surgery between November 2011 and January 2012.

Chapter 4

Analysis & Results

4.1 Univariate Analysis

Simple linear regression was used to evaluate each potential predictor variable's independent association with the outcome variable and the results are presented in Appendix C.1. Only three predictors, HDL (p=0.032), HEP (p=0.040), and LPNRS (p<0.001), were found to be significantly associated with the response variable at the 0.05 significance level. Another three predictors, BMI (p=0.075), EBL (p=0.058), and PCS (p=0.056) are marginally significant in their association with the outcome variable.

Collinearity among the HRQOL measures was also assessed using a scatter-plot matrix, shown in Figure 4.1. The figure indicates possible correlations between ODI and each component score of the SF-36, such that a decrease in LBP-related disability is correlated with increases in physical and mental QOL.

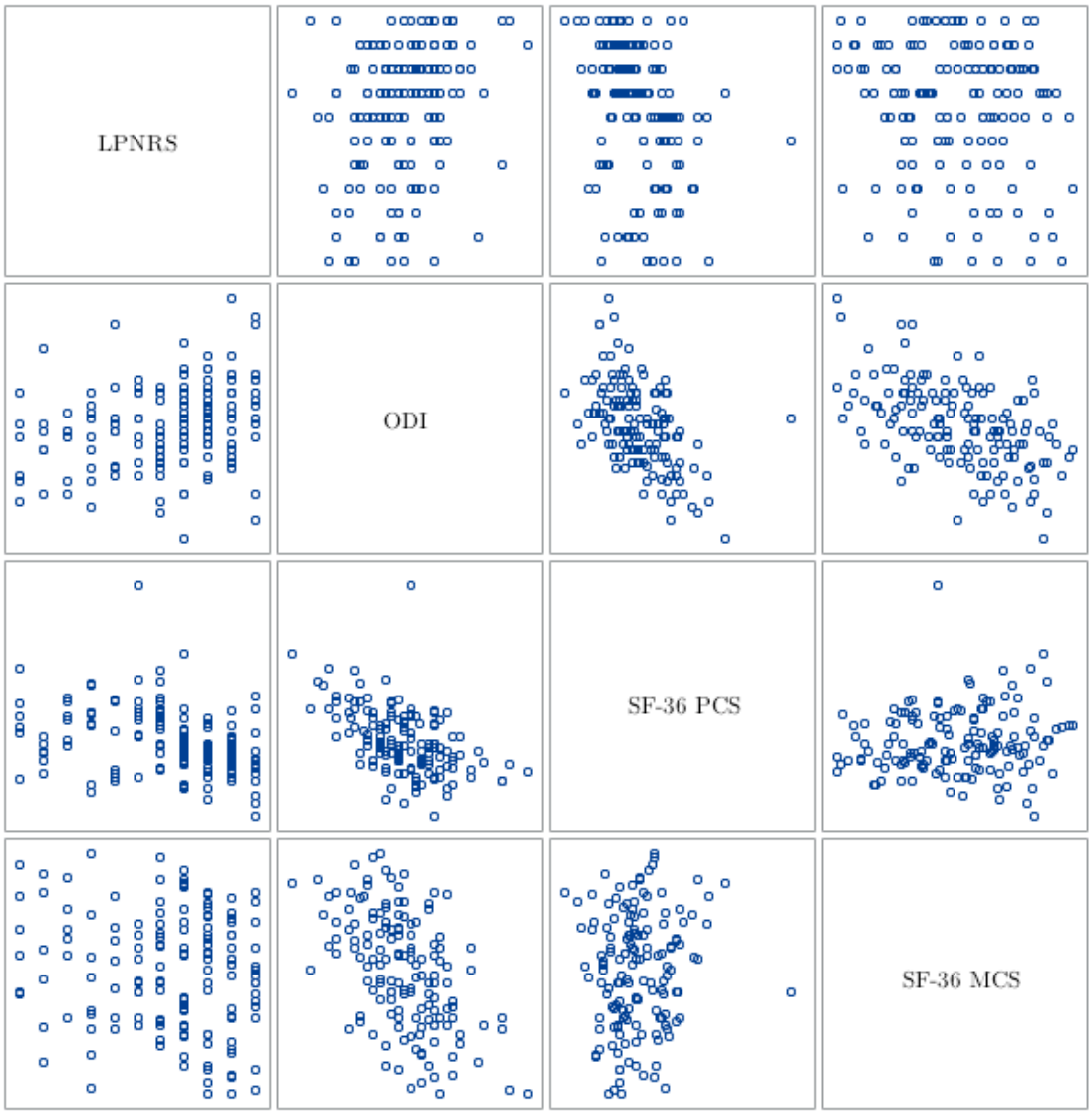

Figure 4.1: Scatter-plot matrix of pre-operative HRQOL variables.

4.2 Stepwise Selection Procedures

Models were fit using each of the stepwise selection procedures (forward stepwise, forward, backwards) at significance levels 0.05, 0.10, 0.15, 0.20 and resulted in five unique models (Table 4.1).

Model	Int	LPNRS	HDL	OA	PCS	AGE	BMI	PRIOR
1	2.651	0.244	---	---	---	---	---	---
2	2.421	0.241	0.763	---	---	---	---	---
3	3.257	0.171	---	-1.188	-0.050	0.030	---	---
4	4.918	0.168	0.636	-1.172	-0.052	0.024	-0.051	---
5	5.075	0.176	0.586	-1.105	-0.054	0.028	-0.058	-0.639

Table 4.1: Parameter estimates of models selected from stepwise selection procedures.

4.2.1 Regression Diagnostics

Residuals for each model were checked for normality and heteroscedasticity, the plots for which are presented in Figures 4.2 through 4.6. The normal quantile plots model show there are several residual outliers for each model, particularly when a greater difference in LBP is predicted. In addition, the distribution of the residuals seems to be most normal for models 1 and 3. Residual variance is reasonably heteroscedastic in models 2 to 5, but the residual variances of model 1, which contains only one X variable, seems to increase as the predicted value increases, suggesting a possible lack-of-fit for this model.

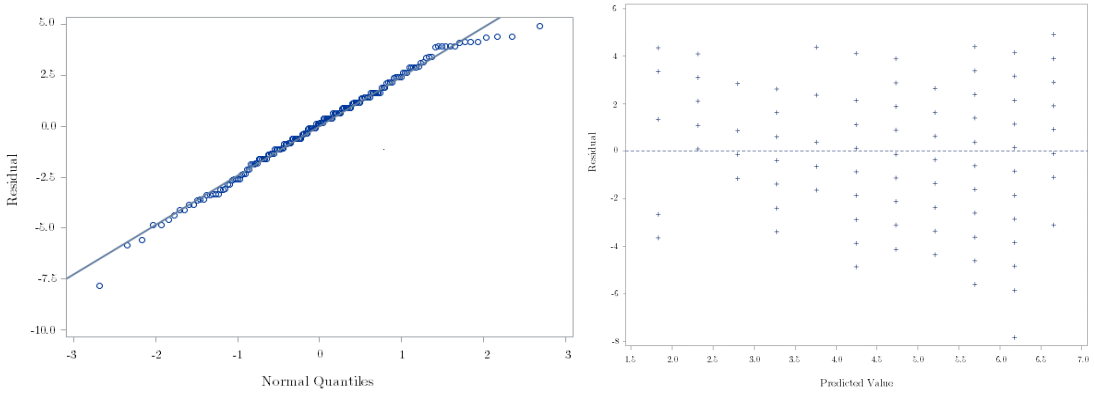

Figure 4.2: Model 1 residual analysis of normality and heteroscedasticity.

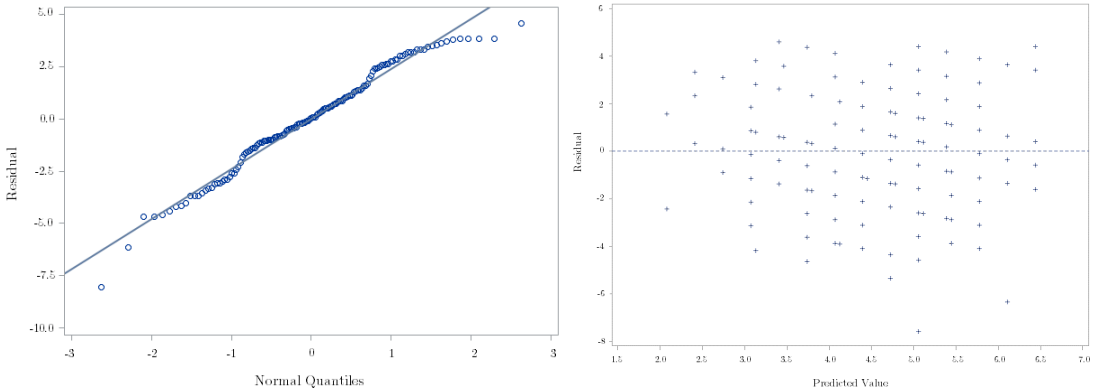

Figure 4.3: Model 2 residual analysis of normality and heteroscedasticity.

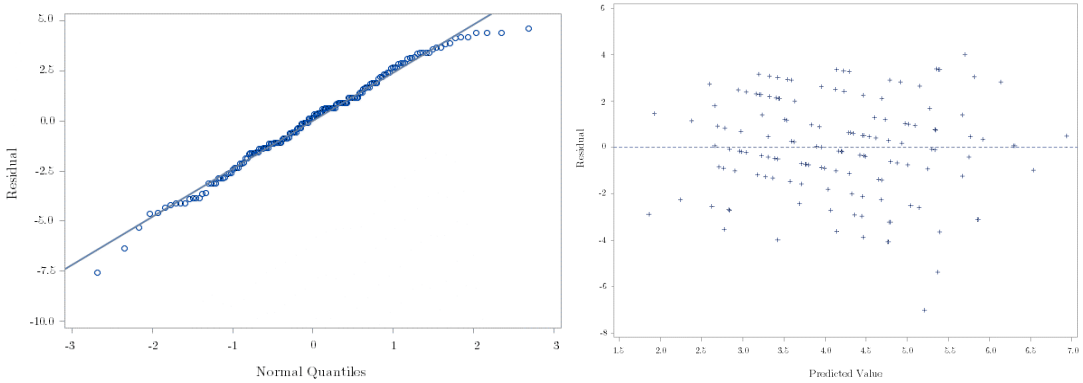

Figure 4.4: Model 3 residual analysis of normality and heteroscedasticity.

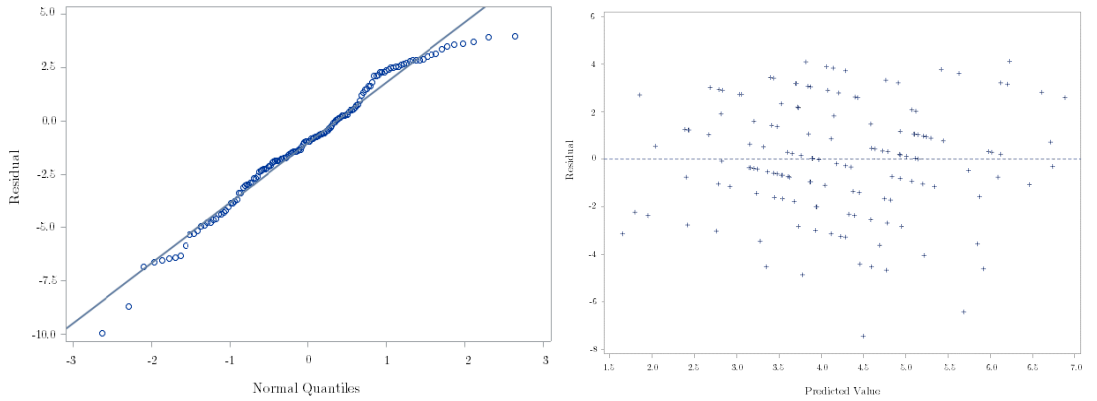

Figure 4.5: Model 4 residual analysis of normality and heteroscedasticity.

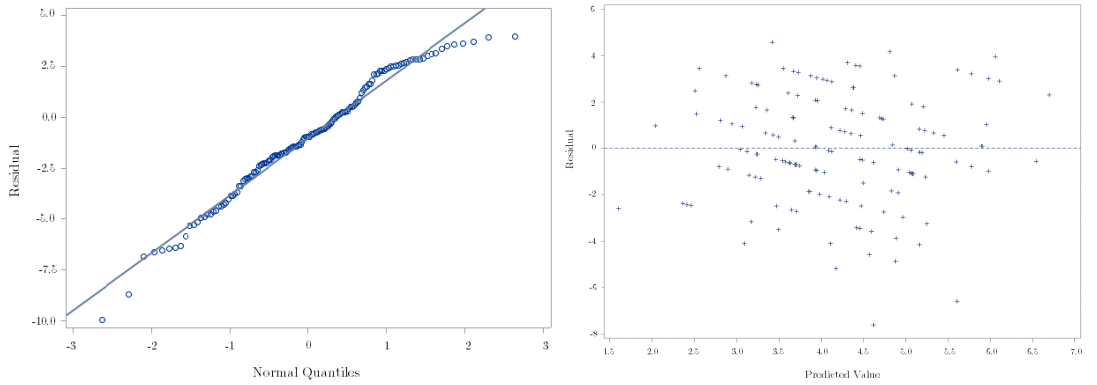

Figure 4.6: Model 5 residual analysis of normality and heteroscedasticity.

4.3 Mallows' C_p Criterion

For this selection procedure, dummy variables were created for all non-dichotomous categorical predictor variables using reference cell coding methods, giving 50 total variables in the full model. All possible models were generated from the dataset and the corresponding C_p statistic was calculated for each model. Figure 4.7 presents the C_p plot of the 50 best models for each size. The plot shows that with the exception of the full model, all subset models fell beneath the reference line $C_p = p$. This suggests that the full model contained a significant amount of potential predictor variables that do not contribute to the model, and consequently, the MSE of the model was a biased estimator of σ^2. Thus, Gilmour's modified $\overline{C}_p$ statistic was plot against the number of

model parameters, and again, all but the full model fell below the line $\overline{C}_p = p$. Therefore, the model with the lowest C_p and $\overline{C}_p$ was assigned as model 6 for validation, and includes the variables BMI, DEPRESS, FIX, HDL, HEART, OA, PCS, STROKE, and ULCER.

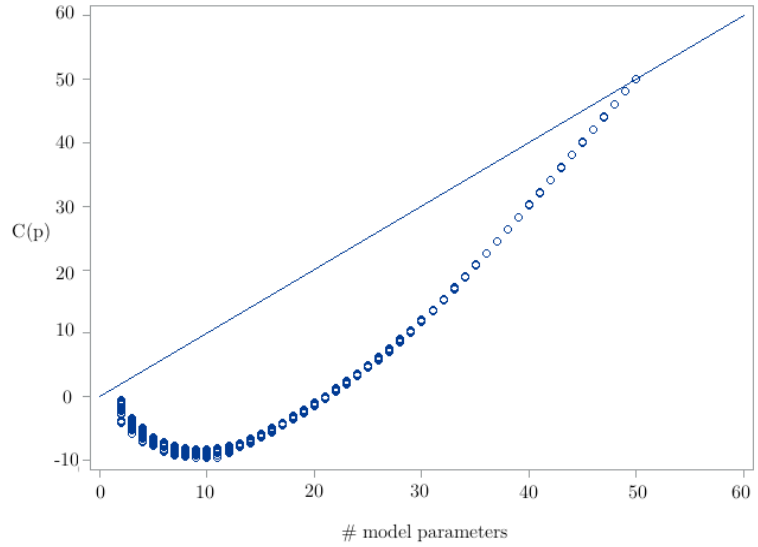

Figure 4.7: C_p plot of best 50 models for each model size.

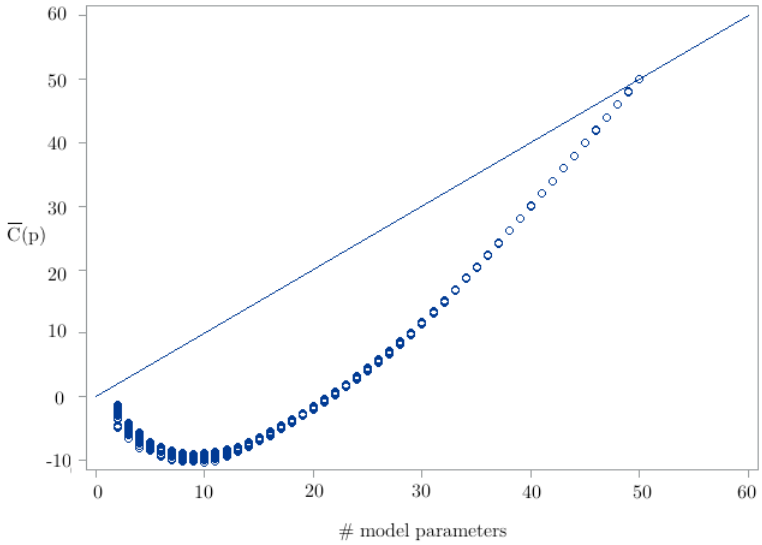

Figure 4.8: $\overline{C}_p$ plot of best 50 models for each model size.

4.4 Model Validation

The summary statistics for the candidate models fit on the training and validation sets are given in Table 4.2.

Model	R2		Adj R2		MSE/MSPR	
	Training	Valid	Training	Valid	Training	Valid
1	0.068	0.163	0.063	0.108	5.915	28.978
2	0.089	0.015	0.078	-0.058	5.819	8.126
3	0.109	0.163	0.084	0.010	5.866	6.329
4	0.132	0.172	0.095	-0.077	5.800	6.222
5	0.144	0.174	0.101	-0.071	5.757	5.926
6	0.163	0.240	0.108	-0.097	5.711	5.741

Table 4.2: Summary statistics for candidate models fit on training and validation sets.

Chapter 5

Discussion

Lumbar IBF has become a common treatment for patients with chronic LBP or severe disease pathologies who have been unresponsive to less invasive methods. This thesis utilized model selection methods to determine important patient factors predictive of post-operative LBP following lumbar IBF.

Pre-operative LPNRS and SF-36 PCS scores were found to be independently associated with LBP improvement. Hepatitis and high cholesterol were also found to be significant univariate analysis; however, these factors have never been cited in the relevant literature. Furthermore, the parameter estimates of high cholesterol suggest that patients with this co-morbidity experience greater LBP, which is clinically counterintuitive.

BMI and patient age were also deemed significantly associated with LBP improvement, which have been shown to influence a surgeon's decision as to whether or not a patient can be expected to receive any benefits after lumbar IBF. Prior surgery has been shown in many studies to be predictive of post-operative HRQOL improvement, but was included in one model. The same is also true for the factor of depression.

Mallows' C_p was not the most adequate selection criterion for the purposes of this thesis, as the dataset contained too many noise variables. For future studies, the full dataset either needs to be developed such that only important variables are included, or a robust modification of the C_p statistic needs to be derived. Information criterion (BIC) are alternatives that have been shown in sociological and behavioral science such

as the Akaike information criterion (AIC) and the Bayesian information criterion research to produce intuitively reasonable results when p-values did not. For such a complex dataset, the information criterions are advantageous in that they base model selection on more theoretical considerations and objective variable assessment.

Future studies can also use various criterion with the *bootstrapping* method of model selection, where repeated samples of cases are randomly selected with replacement, and the criterion is evaluated for subsets to select the best model. Studies have found that this method is fairly accurate in identifying the true model, particularly in larger datasets.

From a clinical standpoint, a point that should be considered is the complexity of pain, as it has the potential to mean discomfort, burning, pins and needs, or even numbness. There is also an inherent difficulty of measuring the intensity of pain. Specifically, pain occurs within a context, and the intensity can be influenced by the meaning of the pain to the patient and its expected duration. Furthermore, there is no way of calibrating pain intensity measures between patients, as one patient may interpret an BPNRS score of "5" very differently from the next patient.

Finally, the selection procedures in this thesis did not consider functions or interactions of variables. Future research may use this thesis as the starting point for more sophisticated models that consider both functions and interactions of the selected variables.

Further work in the analysis of LBP HRQOL outcomes needs to be done to transform the knowledge and teaching base from theory and person experience to one of statistical evidence.

References

[1] Deyo RA, Mirza SK, Martin BI. Back pain prevalence and visit rates: estimates from U.S. national surveys, 2002. *Spine (Phila Pa 1976)* 2006; 23:2724-2727.

[2] Bogduk M. Management of chronic low back pain. *Medical J of Australia* 2003; 180:79-83.

[3] Dagenais S, Caro J, Haldman S. A systematic review of low back pain cost of illness studies in the United States and Internationally. *Spine J* 2008; 8:8-20.

[4] Gore M, Sadosky A, Stacey BR, *et al.* The burden of chronic low back pain: clinical comorbidities, treatment patterns, and healthcare costs in usual care settings. *Spine (Phila Pa 1976)* 1997; 22:11-19.

[5] Hart LG, Deyo RA, Cherkin DC. Physician office visits for low back pain: frequency, clinical evaluation, and treatment patterns from a U.S. national survey. *Spine (Phila Pa 1976)* 1995; 20:11-13.

[6] Youssef JA, McAfee PC, Patty CA, *et al.* Minimally invasive surgery: lateral approach interbody fusion: results and review. *Spine (Phila Pa 1976)* 2010; 35:302-311.

[8] Boden SD. Outcome assessment after spinal fusion: why and how? *Orthop Clin North Am* 1998; 29:717-729.

[9] Soegaard R, Christensen FB, Christiansen T, *et al.* Costs and effects in lumbar spinal fusion. A follow-up study in 136 consecutive patients with chronic low back pain. *Eur Spine J* 2007; 16:657-668.

[10] Blount KJ, Krompinger WJ, Maljanian R, *et al.* Moving toward a standard for spinal fusion outcomes assessment. *J Spinal Disord Tech* 2002; 15:16-23.

[11] Suarez-Almazor ME, Kendall C, Johnson JA, *et al.* Use of health status measures in patients with low back pain in clinical settings. Comparison of specific, generic and preference-based instruments. *Rheumatology (Oxford)* 2000; 39:783-790.

[12] Nemeth G. Health related quality of life outcome instruments. *Eur Spine J* 2006; 15:S44-S51.

[13] Fuhrer MJ. Subjectifying quality of life as a medical rehabilitation outcome. *Disabil Rehabil* 2000; 22:481-489.

[14] Hagg O, Fritzell P, Oden A, *et al.* Simplifying outcome measurement: evaluation of instruments for measuring outcome after fusion surgery for chronic low back pain. *Spine (Phila Pa 1976)* 2002; 27:1213-1222.

[15] Deyo RA, Battie M, Beurskens AJ, *et al.* Outcome measures for low back pain research. A proposal for standardized use. *Spine (Phila Pa 1976)* 1998; 23:2003-2013.

[17] Christensen FB. Lumbar spinal fusion. Outcome in relation to surgical methods, choice of implant, and postoperative rehabilitation. *Acta Orthop Scand Suppl* 2004; 75:2-43.

[18] Johnson L. Outcomes analysis in spinal research. How clinical research differs from outcome analysis. *Orthop Clin North Am* 1994; 25:205-213.

[19] Reeve BB, Hays RD, Bjorner JB, *et al.* Psychometric evaluation and calibration of health-related quality of life item banks: plans for the patient-

reported outcomes measurements information system. *Med Care* 2007; 45:S22-S31.

[20] Deyo RA, Andersson G, Bombardier C, *et al.* Outcome measures for studying patients with low back pain. *Spine (Phila Pa 1976)* 1994; 19:2032S-2036S.

[21] Deyo RA, McNiesh LM, Cone RO. III: Observer variability in interpretation of lumbar spine radiographs. *Arthritis Rheum* 1985; 28:1066-1070.

[22] Deyo RA, Inui TS, Leininger JD, *et al.* Measuring functional outcomes in chronic disease: A comparison of traditional scales and a self-administered health status questionnaire in patients with rheumatoid arthritis. *Med Care* 1983; 21:180-192.

[23] Pecoraro RE, Inui TS, Chen MS, *et al.* Validity and reliability of a self-administered health history questionnaire. *Public Health Reports* 1979; 94:231-238.

[24] Lettice JJ, Kula TA, Derby R, *et al.* Does the number of levels affect lumbar fusion outcome? *Spine (Phila Pa 1976)* 2005; 30:675-681.

[25] Anderson PA, Schwaegler PE, Cizek D, *et al.* Work status as a predictor of surgical outcome of discogenic low back pain. *Spine (Phila Pa 1976)* 2006; 31:2510-2515.

[26] Peolsson A, Vavruch L, Oberg B. Predictive factors for arm pain, neck pain, neck specific disability, and health after anterior cervical decompression and fusion. *Acta Neurochir* 2006; 148:167-73.

[27] Peolsson A, Hedlund R, Vavruch L, et al. Predictive factors for the outcome of anterior cervical decompression and fusion. *Eur Spine J* 2003; 12:274-280.

[28] Gehrchen PM, Dahl B, Katonis P, *et al.* No difference in clinical outcome after posterolateral lumbar fusion between patients with isthmic spondylolisthesis and those with degenerative disc disease using pedicle screw instrumentation: a comparative study of 112 patients with 4 years of follow-up. *Eur Spine J* 2002; 11:423-427.

[29] DeBeard MS, Masters KS, Colledge AL, *et al.* Outcomes of posterolateral lumbar fusion in Utah patients receiving workers' compensation: a retrospective cohort study. *Spine (Phila Pa 1976)* 2001; 26:738-746.

[30] Jenkins LT, Jones AL, Harms JJ. Prognostic factors in lumbar spinal fusion. *Contemp Orthop* 1994; 29:173-180.

[31] Andersen T, Christensen FB, Laursen M, *et al.* Smoking as a predictor of negative outcome in lumbar spinal fusion. *Spine (Phila Pa 1976)* 2001; 26:2623-2628.

[32] Hagg O, Fritzell P, Ekselius L, *et al.* Predictors of outcome in fusion surgery for chronic low back pain: a report from the Swedish Lumbar Spine Study. *Eur Spine J* 2003;12:22-33.

[33] Kern EFO, Maney M, Miller DR, *et al.* Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney diseases in diabetes. *Health Serv Res* 2006; 41:564-580.

[34] Cooke CR, Joo MJ, Anderson SM, *et al.* The validity of using ICD-9 codes and pharmacy records to identify patients with chronic obstructive pulmonary disease. *BMC Health Serv Res* 2011; 11:37.

[35] Zanoli G, Stromqvist V, Jonsson B, *et al.* Pain in low-back pain. Problems in measuring outcomes in musculoskeletal disorders. *Acta Orthop Scand Suppl* 2002; 73:54-57.

[36] Lara-Munoz C, De Leon SP, Feinstein AR, *et al.* Comparison of three rating scales for measuring subjective phenomena in clinical research. I. Use of experimentally controlled auditory stimuli. *Arch Med Res* 2004; 35:43-48.

[37] Vianin M. Psychometric properties and clinical usefulness of the Oswestry Disability Index. *J Chiropr Med* 2008; 7:161-163.

[38] Krishnan J, Chipchase L. Orthopedic surgery outcomes assessment model. *J Qual Clin Pract* 1997; 17:109-116.

[39] Naughton MJ, Anderson RT. Outcomes research in orthopedics: health-related quality of life and the SF-36. *Arthroscopy* 1998; 14:127-129.

[40] Ekman P, Moller H, Hedlund R. Predictive factors for the outcome of fusion in adult isthmic spondylolisthesis. *Spine (Phila Pa 1976)* 2009; 34:1204-1210.

[41] Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. *Control Clin Trials* 1991; 12:142S-158S.

[42] Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol* 1995; 50:741-749.

[43] Albert TA, Purtill J, Mesa J, *et al.* Health outcome assessment before and after adult deformity surgery: A prospective study. *Spine (Phila Pa 1976)* 1995; 20:2002-2004.

[44] Bombardier C. Outcome assessments in the evaluation of treatment of spinal disorders: summary and general recommendations. *Spine (Phila Pa 1976)* 2000; 25:3100-3103.

[45] LaCaille RA, DeBerard MS, Masters KS, *et al.* Presurgical biopsychosocial factors predict multidimensional patient outcomes of interbody cage lumbar fusion. *Spine J* 2005; 5:71-78.

[46] Pellise F, Vidal X, Hernandez A, *et al.* Reliability of retrospective clinical data to evaluate the effectiveness of lumbar fusion in chronic low back pain. *Spine (Phila Pa 1976)* 2005; 30:365-368.

[47] Nyiendo J, Haas M, Goodwin P. Patient characteristics, practice activities, and one-month outcomes for chronic, recurrent low-back pain treated by chiropractors and family medicine physicians: a practice-based feasibility study. *J Manipulative Physiol Ther* 2000; 23:239-245.

[48] Herbert J, Koppenhaver S, Fritz J, *et al.* Clinical prediction for success of interventions for managing low back pain. *Clin Sports Med* 2008; 27:463-79.

[49] Carreon LY, Glassman SD, Djurasovic M, *et al.* Are preoperative health-related quality of life scores predictive of clinical outcomes after lumbar fusion? *Spine (Phila Pa 1976)* 2009; 34: 725-730.

[50] Brunelli C, Zecca E, Martini C, *et al.* Comparison of numerical and verbal rating scales to measure pain exacerbations in patients with chronic cancer pain. *Health Qual Life Outcomes* 2010; 8:42.

[51] Fairbank JCT, Pynsent PB. The Oswestry Disability Index. *Spine (Phila Pa 1976)* 2000; 25:2940-2953.

[52] Kutner MH, Neter J, Nachtsheim CJ, *et al. Applied Linear Statistical Models.* 5th ed. New York, NY: McGraw-Hill/Irwin; 2005:95-110, 124-126, 217-236, 327-353, 385-388.

[53] Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research.* 4th ed. Malden, MA: Blackwell Science; 2002:261-264.

[54] Raftery AE. Bayesian model selection in social research. *Sociol Methodol* 1995; 25:111-196.

[55] Mallows CL. Some comments on C_p. *Technometrics* 1993; 15:661-675.

[56] Gilmour SG. The interpretation of Mallows' C_p statistic. *J Roy Stat Soc D-STA* 1996; 45:49-56.

[57] Nagelkerke N. A note on a general definition of the coefficient of determination. *Biometrika* 1991; 78:691-692.

[58] Picard R, Cook D. Cross-validation of regression models. *J Am Stat Assoc* 1984; 79:575-583.

[59] Mourad M, Bertrand-Krajewski JL, Chebbo G. Calibration and validation of multiple regression models for storm water quality prediction: data partitioning, effect of data sets size and characteristics. *Water Sci Technol* 2005; 52:45-52.

[60] Thompson WR. Variable selection of correlated predictors in logistic regression: investigating the diet-heart hypothesis [dissertation]. Tallahassee, FL: Florida State University; 2009.

[61] Chen XH. Comparisons of statistical modeling for constructing gene regulatory networks [master's thesis]. Vancouver, BC: University of British Columbia; 2008.

Appendix A

HRQOL Instruments

A.1 Numeric Rating Scale (NRS)

The NRS is an ordinal scale ranging from 0 to 10 that subjectively measures pain intensity. A score of 0 indicates the patient is experiencing "no pain," whereas a score of 10 indicates the patient is experiencing "worst possible pain." The Back Pain NRS (BPNRS) and Leg Pain NRS (LPNRS) were used to obtain the severities of LBP pain and leg pain, respectively, for each patient.

The NRS has not yet been thoroughly investigated, but its validity and reliability have been confirmed. There is no published information pertaining to the distribution or error of data obtained using the NRS.[30]

A.2 Oswestry Disability Index (ODI)

The ODI is a self-administered questionnaire measuring disability specific to back problems on a 10-item scale with 6 possible responses each. The 10 items include pain intensity: personal care, lifting, walking, sitting, standing, sleeping, work, social life, and traveling. Each item scores from 0 to 5, with higher scores indicating more severe disability, and an overall disability percentage is obtained. Patients scoring between 0-20% are minimally disabled, 21-40% are moderately disabled, 41-60% are severely disabled, 61-80% are crippled, and 81-100% are bed-bound or exaggerating their symptoms.[51]

The ODI has been extensively tested and has shown good psychometric properties. Specifically, the ODI shows good construct validity and internal consistency. Its test-retest reliability has also been shown to be high, but decreases the longer the wait between measurements.[49,51]

A.3 SF-36v2™ (SF-36)

The SF-36 is a general QOL survey comprising of 36 items that are organized into 8 subscales: physical functioning, physical role limitation, bodily pain, social functioning, general mental health, emotional role limitations, vitality, and general health perceptions. The SF-36 also includes two questions intended to estimate self-perceived overall change in health status over the past year. With the exception of the overall change in health status questions, patients are asked to respond with reference to the past 4 weeks. Scores for each of the subscales are summed and linearly transformed to derive a physical component score (PCS) and a mental component score (MCS), each on a scale of 0 to 100. A lower score indicates higher QOL.[49]

Although the psychometric properties of the SF-36 has been extensively tested across a range of patient populations, there has been limited research regarding its merit for degenerative spine disorder patients, and even less for lumbar IBF patients.[32,42,45,46] However, one study by Albert *et al.*[43] tested the viability of the SF-36 to measure health status changes in degenerative scoliosis patients one year after undergoing lumbar IBF. The study demonstrated that the SF-36 was sensitive enough to detect improvement in the study population.

Appendix B

Proofs

B.1 Proof of Equation (2.7)

Proof.

Given: $\begin{cases} \sum_{i=1}^{n} \sigma^2\{\widehat{Y}_i\} = p\sigma^2 \\ E\{SSR_p\} = \sum\left(E\{\widehat{Y}_i\} - \mu_i\right)^2 + (n-p)\sigma^2 \end{cases}$

Γ_p can then be expressed as

$$\begin{aligned} \Gamma_p &= \frac{1}{\sigma^2}\left[E\{SSR_p\} - (n-p)\sigma^2 + p\sigma^2\right] \\ &= \frac{E\{SSR_p\}}{\sigma^2} - (n-2p) \end{aligned}$$

Using estimators SSR_p and $MSE(X_1, X_2, \ldots, X_{q-1})$ for $E\{SSR_p\}$ and σ^2, respectively, gives the C_p criterion.

□

B.2 Lemma B.2

Given: $C_p = \dfrac{SSR_p}{SSR_{k+1}/(n-k-1)} - (n-2p)$

To obtain the distribution of C_p in the case where all important predictor variables are included in the model, assume without loss of generality, that $\beta_p = \ldots = \beta_k = 0$.

C_p can then be expressed as

$$\begin{aligned} C_p &= (n-k-1)\frac{SSE_{k+1}SS(\beta_p,\ldots,\beta_k\left|\beta_0,\ldots,\beta_{p-1}\right.)}{SSE_{k+1}} - (n-2p) \\ &= (n-k-1)\left\{1+\frac{SS(\beta_p,\ldots,\beta_k\left|\beta_0,\ldots,\beta_{p-1}\right.)}{SSE_{k+1}}\right\} - (n-2p) \\ &= (k-p+1)\frac{SS(\beta_p,\ldots,\beta_k\left|\beta_0,\ldots,\beta_{p-1}\right.)/(k-p+1)}{SSE_{k+1}/(n-k-1)} - (k+1-2p) \\ &= (k-p+1)\frac{U/(k-p+1)}{V/(n-k-1)} - (k+1-2p) \end{aligned}$$

where

$$\begin{cases} U \sim \chi^2_{k-p+1} \\ V \sim \chi^2_{n-l-1} \\ U \text{ and } V \text{ are independent} \end{cases}$$

Hence,

$$C_p = (k-p+1)F + 2p - k - 1 \qquad \text{(Lemma B.2)}$$

where $F \sim F_{k-p+1,n-k-1}$

□

B.3 Proof of Equation (2.8)

Given: $E\{F\} = \dfrac{n-k-1}{n-k-3}$

By Lemma B.2, $E\{C_p\}$ can then be expressed as

$$\begin{aligned} E\{C_p\} &= (k-p+1)\frac{n-k-1}{n-k-3} + 2p - k - 1 \\ &= p + \frac{2(k-p+1)}{n-k-3} \end{aligned}$$

Thus, for $E\{\overline{C}_p\} = p$ to be true, $\overline{C}_p$ is defined as

$$\overline{C}_p = C_p - \frac{2(k-p+1)}{n-k-3}$$

□

Appendix C

Supplemental Materials

C.1 Results of Univariate Analysis

Category	X	Variable	P-Value
Demographics	1	AGE	0.252
	2	BMI	0.075
	3	EDU	0.677
	4	GENDER	0.393
	5	SMOKE	0.515
Co-Morbidities (Y/N)	6	ASTH	0.770
	7	BLEED	0.458
	8	COPD	0.253
	9	DEPRESS	0.338
	10	DIABETES	0.841
	11	GERD	0.742
	12	HDL	0.032
	13	HEART	0.831
	14	HEP	0.040
	15	HTN	0.228
	16	OA	0.196
	17	OSP	0.210
	18	SEIZURE	0.474
	19	STROKE	0.220
	20	ULCER	0.207
Surgical Summary	21	DIAGNOSIS	0.418
	22	EBL	0.058
	23	FIX	0.481
	24	FUS	0.149
	25	LOS	0.309
	26	NUMLEVS	0.779
	27	OR_TIME	0.152
	28	PRIOR	0.267
HRQOL Measures	29	LPNRS	<0.001
	30	ODI	0.594
	31	PCS	0.056
	32	MCS	0.839