ABSTRACT

Title of dissertation: APPLICATIONS OF PARAMETRIC AND SEMI-PARAMETRIC MODELS FOR LONGITUDINAL DATA ANALYSIS

Hisham Talukder, Doctor of Philosophy, 2014

Dissertation directed by: Professor Héctor Corrada Bravo
Department of Computer Science

A wide range of scientific applications involve analyses of longitudinal data. Whether it is time or location, careful considerations need to be made when applying different statistical tools. One such challenge is to correctly estimate variance components in observed data. In this dissertation, I apply statistical tools to solve problems involving longitudinal data in the field of Biology, Healthcare and Networks.

In the second chapter, I apply SSANOVA models to find regions in the genome that have a specific biological trait. We introduce a direct approach of estimating genomic longitudinal data of two different biological groups. Using SSANOVA we then produce a novel method to estimate the difference between the two groups and find regions (location or time) where this difference is biologically significant.

In the third chapter, we analyze longitudinal network data using an overdispersed Poisson model. We build a network of musical writers yearly for a 42 year period. Using statistical models, we predict network level topology changes and

find covariates that explain these changes. Network level characteristics used for this chapter include average node degree, clustering coefficient and network density. We also build a visualization tool using R-Shiny.

The fourth chapter uses data partitioning to study the difference between insured patients and uninsured patients in health outcomes. There is a disparity in health outcomes depending on an individual's type of insurance. The level of risk for an injury is the longitudinal aspect of this dataset. We partition the data into four pre-defined risk categories and evaluate the disparity between insured and uninsured patients using logistic regression models.

APPLICATIONS OF PARAMETRIC AND SEMI-PARAMETRIC MODELS FOR LONGITUDINAL DATA ANALYSIS

by

Hisham Talukder

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory Committee:
Professor Héctor Corrada Bravo / Advisor
Professor Mihai Pop
Professor Benjamin Kedem
Professor Shawn Mankad
Professor Bruce Golden / Dean's Representative

To my mother, Nazmun Nahar Islam, my father, Dr. Mohammed Talukder M.D.,

and my brother Dr. Iehab Talukder M.D., Ph.D.

Acknowledgments

I appreciate all the people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost I'd like to thank my advisor, Professor Héctor Corrada Bravo, for giving me the opportunity to work with him on challenging problems. He always had a open door policy, which as a grad student, was more than I could ask for. I appreciate the lessons he has taught me in statistics, computational biology, doing research and most of all life. Without his help and guidance this would not be possible and for that I am forever grateful.

I would also like to thank my committee which consist of Professor Bruce Golden, Professor Benjamin Kedem, Professor Shawn Mankad and Professor Mihai Pop. I am thankful to have a committee filled with people who at any moment helped me understand the challenges at hand. I have collaborated with members in my committee in different projects and their guidance has helped me finish the projects throughout graduate school.

My colleagues and friends at CBCB deserve a special shout-out. They have all helped me inside and outside the lab throughout my years as a graduate student. I want to thank Kwame Okrah for being a great housemate, a great lab mate, and a great friend throughout my graduate career. I would also like to thank Joyce Hsiao and Joseph Paulson for making this lab a fun place to work during the weekdays and this city a fun place to go out in during the weekends. I would also like to thank the rest of my colleagues at CBCB. Finally I would like to thank my office mates;

Lee Mendelowitz and Senthil Muthiah, whom I have shared ideas and stories with for the past two years while at lunch or coffee breaks. I would also like to thank the friends I made during my stay in the Applied Math program especially the ones who entered the same year as me. A special thanks to Karamatou Yacoubou Djima and Marie Chau, who were like sisters I never wanted. Also thanks to Matthew Temba for all the late night drinks in DC.

None of this would ever be possible without the support from friends I have had since high school and undergraduate. In no particular order I want to thank Simon Rubinsky, Mikael Bialek, Jared Goldberg, Jafar Ahmed, Arkadiy Daviydov, Giorgio Medranda, John Herrera, and Shohan Pervaze. Without the support from all of you none of this would be possible.

I owe my deepest thanks to my family - my mother and father who have guided me throughout the ups and downs of graduate school. They were hard on me when I was being lazy and patted me on the back when I accomplished something. The sacrifices my parent have made has paved the way for all the success in my life right now and in the future. The life lessons they have taught me could never be repaid by any amount of gratitude. I would also like to thank my brother who has made me realize even with a PhD under my name he will always be the smart one in the family. His work ethic pushed me everyday to work harder and more efficiently.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Definition of Longitudinal data

A wide range of scientific problems involve longitudinal data analysis. It ranges from Biological problems [10, 12, 20, 32, 36, 38], to Healthcare related problems [51–53], to Network analyses [68, 74–76, 78, 79]. Longitudinal data are defined as an experiment where subject outcomes or treatments are collected at multiple time points, locations, or periods. Longitudinal analysis allows an appropriate solution for the following problems out of many:

1. How many days will it take to show physical effects after taking drug X?
2. What are specific regions in the genome where a certain trait is observed?
3. How is a variable changing over time and what are good predictors for these changes?

Because of the dependencies between repeated measurements, modeling of longitudinal data are a big challenge. Any model appropriate for longitudinal analysis must capture the complexities of variance structure of the subjects in order to give accurate predictions. Another challenge is collecting the data itself. Longitudinal

data needs multiple observations across time, which allows for multiple missing observations for lack of subject turnout in every time point. There are time varying covariates which makes the the direction of causality hard to define or estimate. Failure to do so will result in false estimation of parameters and reporting of significance [80].

There are many benefits to using longitudinal data [82]. There can be a measurement of individual change and prediction of individual variables across time. This allows for conclusions across time on an individual basis. A longitudinal study can simultaneously characterize multiple time scale factors as response and independent variables are changing over time. Also possible are comparisons of sub groups across time. These types of questions can not be addressed without longitudinal data being available.

1.2 Types of models used for longitudinal data analysis

Following the exposition in Liang [81], there are two main types of model used for longitudinal data analysis. They are subject specific models and population mean models. A subject specific model is used to model individual behavior as subjects in the same groups vary from person to person. Using the same notation as Hedeker [80], we let Y_{ij} be the response for subject i at time or location j. We assume data for subject i are observed from a stochastic process that differs through time or location from the group mean. In order to model these changes, a random variable is included in a model to account for individual variation. Observations

for this type of analysis consist of responses for an individual i across multiple time points. A linear model will follow:

$$Y_i = X\beta + Z_i b_i + e \tag{1.1}$$

where $X\beta$'s are group effects and b_i corresponds to individual random effects. We assume correlation within subject measurements but independence across subjects.

The second type of model used for longitudinal analysis involves group means model [80,81]. If we average over subjects for each group and then analyze the data we are in effect removing individual effects from the model. This can be the only way to analyze data in certain scenarios. For example; assume we have observations across time for two groups of patients using two different drugs. We know nothing about individual characteristics, instead only group characteristics. This will allow the use of:

$$Y_i = X\beta + e \tag{1.2}$$

where no individual observations are there for use. Any model, parametric or semi-parametric, used for longitudinal analysis will fall under one of the two above [80–82].

1.3 Overview of dissertation

For this dissertation we use a variety of models to analyze longitudinal data. In chapter 2, we use a semi-parametric model called Smoothing Splines Analysis of variance (SSANOVA) to model genomic data. We use SSANOVA to find regions of interest where differences between two groups are significant with regards to a

biological response. In chapter 3 we use a parametric model to characterize changes in network structures over time. We used an overdispersed Poisson model to predict changes of network characteristics in a music writer network. For this, we have individual and group observations and both are used in our analysis. For chapter 4, we use a logistic regression model to look at differences between insured and uninsured patients across different levels of injury risk or severity. Here we partition the longitudinal data (levels of risk) into four disjoint groups and use models to predict outcomes (mortality rates) across different ages. We use multiple patients with similar characteristics as repeated measurements across different levels of risks.

Chapter 2: Finding regions of interest in high throughput genomics data using smoothing splines

2.1 Overview

High-throughput methods, like microarrays and next-generation sequencing, are frequently used to obtain quantitative measurements at base-pair resolution in many important applications: e.g., enrichment scans (ChIP-seq [1] and ChIP-chip [2], DNAse-seq [3], etc.) and measurements of DNA methylation, a chemical DNA modification known to play a significant role in gene regulation [4], using either microarrays [5,6] or sequencing [7,8]. One very important use of these quantitative data in these applications is to find contiguous regions in the genome where measurements differ between two or more populations of interest. For example, methylation changes are widely understood to be an important part of tumorigenesis in solid tumors [9], and genomic regions where these differences occur have been widely reported [10, 11]. In this case, the inference of interest is to find regions in the genome where methylation changes in cancer occur.

A common approach to detect regions of interest of this type is to model differences between groups with respect to these quantitative measurements as smooth

functions along the genome and perform statistical inference on these models. In particular, widely used methods for region finding using DNA methylation data use local regression methods [5, 12–14] to estimate these smooth functions. An important aspect of these tools is their ability to incorporate sample characteristics as covariates in these models, e.g., sex and age in population studies, or technical factors like processing batches. Incorporating these sources of variability, both biological and technical [15] is essential in high-throughput studies. Therefore, these methods require that the models used can accommodate both smooth functions and sample-specfic characteristic.

The methods mentioned above use an indirect approach to estimate both the smooth functions underlying the measurement of interest and parameters that model these covariates: they first estimate point-wise models where a term that captures differences between groups is included and then fit a smooth function using a method like LOESS [16] to these point-wise estimates. This is an inefficient approach prone to removing important characteristics of the data. In this chapter, we introduce an alternative, direct, approach to this problem using semi-parametric regression tools.

Smoothing spline regression models [17] are commonly used to model longitudinal data and form the basis for methods used in a large number of applications [19–21]. Specifically, an extension of this methodology called Smoothing-Spline ANOVA [22] is capable of directly estimating a smooth function of interest while incorporating other covariates in the model.

We show in this chapter that a direct approach based on Smoothing-Spline ANOVA is better suited for region finding applications in high-throughput genomic

data. We show by simulation that our direct approach significantly improves the accuracy of detecting regions of interest. We apply our approach to methylation data of colon cancer and normal tissue from the Cancer Genome Atlas (TCGA) project [23]. Additionally, to demonstrate the generality of this methodology we also apply this to data from a longitudinal high-throughput metagenomic study characterizing how diet is associated to changes in gut microbial composition [24].

2.2 Methods

2.2.1 Problem Formulation

We assume data of the form:

$$Y_{itk} = f_i(t, x_k) + e_{tk} \tag{2.1}$$

where Y_{itk} is a measured response, $i = 0, 1$ represents group factor (diet, cell type, etc.), $t = 1, \ldots, T$ represents series factor (for example, time or location), $k = 1, \ldots, K$ represents replicate observations, x_k are covariates for sample k (including an indicator for group membership $I\{k \in i\}$) and e_{tk} are independent $N(0, \sigma^2)$ errors. We assume f_i to be a smooth function, defined in an interval $[a, b]$, that can be parametric, non-parametric or a mixture of both.

Our goal is to identify intervals where the absolute difference between two groups $\eta_d(t) = f_1(t, \cdot) - f_2(t, \cdot)$ is large, that is, regions, $R_{t1,t2}$, where:

$$R_{t1,t2} = \{t1, t2 \in x \text{ such that } \ |\eta_d(x)| \geq C\} \tag{2.2}$$

and C is a predefined constant threshold.

To identify these areas we use hypothesis testing using the area $A_{t1,t2} = \int_{R_{t1,t2}} \eta_d(t)dt$ under the estimated function of $\eta_d(t)$ as a statistic (Figure 2.1) with null and alternative hypotheses

$$\begin{aligned} H_0 : A_{t1,t2} &\leq K \\ H_1 : A_{t1,t2} &> K \end{aligned} \tag{2.3}$$

with K some fixed threshold.

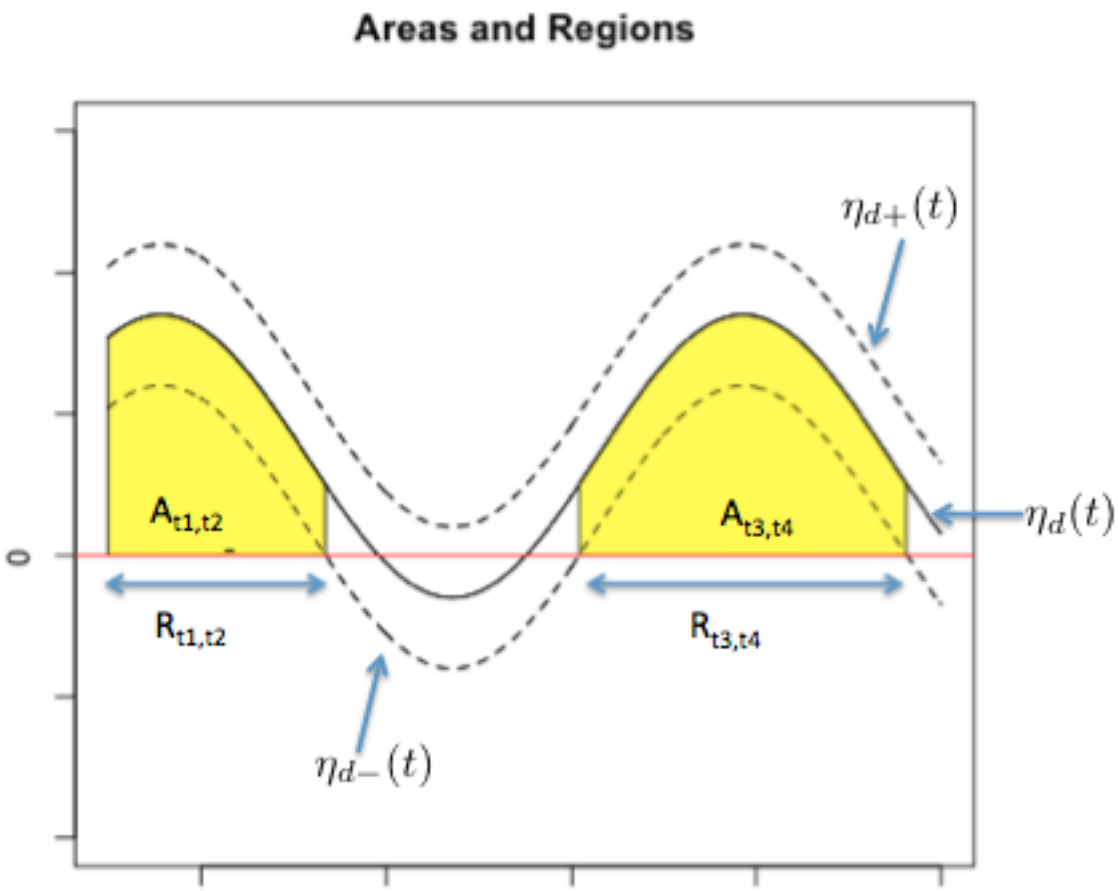

Figure 2.1: **Illustrative example of regions and areas** This example shows the difference function, η_d, with confidence intervals. We choose regions labeled $R_{t1,t2}$ and $R_{t3,t4}$ as possible locations where there are significant difference in response between two groups. The areas under the curve in these regions, $A_{t1,t2}$ and $A_{t3,t4}$, are calculated. These two areas are the test statistic being tested using permutation.

2.2.2 Region finding via smoothing functions

Recent, widely used, region-finding methods [5, 12–14] based on smoothing methods fit point-wise linear regression models at each t where there are multiple observations:

$$Y_{itk} = \beta^T x_k + \beta_t I\{k \in i\} + e_{tk}. \tag{2.4}$$

In this case β_t measures the difference between two groups at point t. Estimated $\hat{\beta}_t$ are then treated as realizations of the smooth difference function $\eta_d(t)$ of interest. Bumphunter use smoothing methods, e.g., LOESS to smooth $\hat{\beta}_t$'s and estimate the difference function of interest (η_d) using model $\beta_t = \eta_d(t) + \varepsilon_t$ which measures variability around the difference function of the marginal estimates $\hat{\beta}_t$.

This is an indirect approach to estimate both the smooth difference function $\eta_d(t)$ underlying the measurement of interest and parameters that model relevant technical or biological effects (β_t). This inefficient approach is prone to removing important characteristics of the data. Error estimates may be biased using this approach. For instance, these methods do not provide a clean way of interpreting the two variance estimates obtained above: one from the piece-wise linear model in equation 2.4, one from the smoothing method. This will effect down-stream inferences that rely on variability estimates, e.g., defining regions $R_{t1,t2}$ in equation 2.2. A direct approach that estimates all parameters without relying on point-wise estimates is needed.

Permutation-based methods are used to calculate a null distribution of the

area statistics $A_{t1,t2}$'s. To do this, the group-membership indicator variables (0-1 binary variable) are randomly permuted B times, e.g., $B = 1000$ and the method above is used to estimate the difference function η_d^b $(b = 1, \ldots, B)$ (in this case simulation the null hypothesis) and area statistics $A_{t1,t2}^b$ for each random permutation. Estimates $A_{t1,t2}^b$ are then used to construct an empirical estimate of $A_{t1,t2}$ under the null hypothesis. The observed area, $A_{t1,t2}^*$, is compared to the empirical null distribution to calculate a p-value. Figure 1 illustrates the relationship between $R_{t1,t2}$ and $A_{t1,t2}$. The key is to estimate regions $R_{t1,t2}$ where point-wise confidence intervals would be appropriate. However, due to the problem of variance estimation outlined above, this is not possible. We present a direct estimation methodology based on smoothing-spline methods that addressees these issues.

2.2.3 Smoothing Spline ANOVA models

Smoothing Spline analysis of variance (SSANOVA) [21] is a semiparametric method that models data generated from a smooth function $f(x)$ by assuming that f is a function in a Reproducible Kernel Hilbert Space (RKHS) of the form $\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_1$. The set of functions $(\phi_v(x))_{v=1}^m$ spans the finite dimensional subspace $\mathcal{H}_0$ and $\mathcal{H}_1$ is a RKHS induced by a given kernel function k. Therefore, f has a semiparametric form given by

$$f(x) = \sum_{j=1}^{m} d_j \phi_j(x) + g(x), \tag{2.5}$$

for some coefficients d_j, where functions ϕ_j have a parametric form and $g \in \mathcal{H}_1$

which is defined by:

$$g(x) = \sum_{\alpha} g_\alpha(x_\alpha) + \sum_{\alpha<\beta} g_{\alpha\beta}(x_\alpha, x_\beta) + \dots \tag{2.6}$$

where g_α and $g_{\alpha\beta}$ satisfy the standard ANOVA side conditions. g_α are main effects in the model and $g_{\alpha\beta}$ are the interactions in the model. An RKHS $\mathcal{H}_\alpha$ is associated with each term in the model along with the kernel function k_α. We can define a kernel function $k(\cdot,\cdot) = \sum_\alpha \theta_\alpha k_\alpha + \sum_{\alpha\beta} \theta_{\alpha\beta} k_{\alpha\beta} + \dots$, where the coefficients θ are hyper parameters that weigh the relative importance of each term in the model [25].

The SSANOVA estimate of f given data $(x_i, y_i), i = 1, ..., n$, is given by the solution of a penalized problem,

$$\min_{f\in H}(y_t - f(x))^2 + \lambda J(f(x)) \tag{1}$$

where the first term discourages the lack of fit of f and the second term penalizes the complexity of f with smoothing parameter λ controlling the trade-off between the two.

Following the representer theorem of [26] and the assumption of gaussian data, the minimizer of the problem in equation 1 has a finite representation of the form:

$$f(x) = \sum_{v=1}^{m} d_v \phi_v(x) + \sum_{j=1}^{n} c_j k(x_j, x) \tag{2.7}$$

for some coefficients c_i and d_v. Letting Y be the matrix of observations of size $N \times 1$, where N includes all observations, including repeated measurements

for different subjects; S is a $N \times m$ matrix where m represents the number of unpenalized terms in the model; and Q is a $N \times N$ matrix which accounts for all penalized terms in the model; estimation reduces to:

$$\min_{d,c}(Y - Sd - Qc)^T(Y - Sd - Qc) + n\lambda c^T Qc. \tag{2.8}$$

Here S is the matrix described above with the iv^{th} entry being $\phi_v(x_i)$ and Q is the penalized matrix with the ij^{th} entry being $k(x_i, x_j)$ [27].

We use Generalized Approximate Cross-Validation (GACV), an approximation to the leave-one-out estimate of the comparative Kullback-Leibler distance between $\hat{f}$ and the unknown true f to select regularization parameter λ and θ.

Under SSANOVA, f=$f_0 + f_1$, where f_0 has a diffuse prior in span$\{\phi_v, v = 1, ..., m\}$ and f_1 has a mean zero Gaussian process prior with covariance function $E[f_1(x)f_1(y)] = bR_j(x, y)$. The posterior variance for f is the respective element in the smoother matrix where the smoother matrix is:

$$A(\lambda) = I - n\lambda(M^{-1} - M^{-1}S(S^T M^{-1} S)^{-1} S^T M^{-1}) \tag{2.9}$$

where $M = Q + n\lambda I$. The 100(1-α)% confidence interval of $f_\lambda(x_i)$ based on the posterior distribution stated above is $f_\lambda(x_i) \pm z_{\frac{\alpha}{2}} \sigma \sqrt{a_{i,i}}$ where $a_{i,i}$ is the i, i^{th} element of $A(\lambda)$. Bayesian intervals are given to include some confidence to our estimates.

The interval estimates when used with GCV smoothing parameter λ demonstrate an across the function coverage property (ACP) for η fixed and smooth [22]. Over the sampled points, the average coverage proportion is defined as:

$$ACP(\alpha) = \frac{1}{n}\{i : |\hat{f}(x_i) - f(x_i)| \leq z_{\frac{\alpha}{2}}\hat{\sigma}\sqrt{a_{i,i}}\}, \tag{2.10}$$

where $a_{i,i}$ is the corresponding element of the smoother matrix $A(\lambda)$.

Simulation results suggested in [22] that for n large,

$$E[ACP(\alpha)] \approx 1 - \alpha \tag{2.11}$$

This coverage property provides a principled way to incorporate interpretable variance estimates to detect regions of interest $R_{i,j}$ as defined in equation 2.2.

2.2.3.1 SSANOVA for region finding

We apply the SSANOVA model to region finding by modeling f as semiparametric function:

$$f_i(t, x_k) = \beta^T x_k + f_1(t) + f_2(I\{k \in i\}) + f_{12}(t, I\{k \in i\}), \tag{2}$$

where β are coefficients of a linear model of sample covariates (e.g., age, sex), f_1 is the main effect term for the series, f_2 is the main effect term for group i and f_{12} is the interaction term. By encoding group membership using a 0-1 binary variable, the ANOVA side conditions imply that we can directly estimate the difference function $\eta_d(t)$ as $\eta_d(t) = f_2(1) + f_{12}(t, 1)$. In contrast to the bumphunter [12] method above, we are able to directly estimate η_d. We use the bayesian confidence intervals above to tune our definition of $R_{t1,t2}$ from before as,

$$R_{t1,t2} = \{x \in [t1, t2] \text{ such that } \eta_{d+}(x) \leq C \text{ or } \eta_{d-}(x) \geq C\} \tag{2.12}$$

where η_{d+} and η_{d-} are the upper and lower 95% confidence intervals. We use this direct estimate of the difference function $\eta_d(t)$ to calculate area statistics $A_{t1,t2}$ used for testing as described above. For permutations, we treat negative η_d as negative area and positive η_d as positive area. For each cluster we adjust for multiple testing by using a Bonferroni correction (α/n). For example, if a given cluster has three candidate regions we would reject if the calculated p-value is less than .05/3.

2.2.4 Materials

2.2.4.1 Illumina HumanMethylation 450k beadarray data

IDAT files for 17 normal colon and 34 colon tumor samples were obtained from the TCGA project [23]. Pre-processing was performed using the `minfi` Bioconductor package [14]. Data were preprocessed and normalized using the standard Illumina method. Probes were grouped using the `bumphunter clusterMaker` function with a maximum gap parameter of 1000bp following the differential methylation region finder in `minfi`. Our SSANOVA region finder was run for each probe group.

2.2.4.2 Metagenomics

We use data from a metagenomic longitudinal study consisting of twelve germ-free adult male C57BL/6J mice. The twelve mice were all fed a low-fat, plant

polysaccharide-rich diet and gavaged with healthy adult human fecal material. Following fecal transplant the mice continued on the low-fat, plant polysacchaaride-rich diet for four weeks. After four weeks, a subset of 6 were switched to a high-fat and high-sugar diet. Weekly fecal samples for each mouse went through PCR amplification of the bacterial 16S rRNA gene V2 region. Further details of the experimental protocols and descriptions of the data can be found in [24]. Sequences can be downloaded from: `http://gordonlab.wustl.edu/TurnbaughSE_10_09/STM_2009.html` Count data were distributed as part of the the `metagenomeSeq` bioconductor package. Counts were normalized using cumulative sum scaling normalization [28].

2.3 Results

2.3.1 Simulation Study

For our simulation study we used the 17 control samples from the methylation study on the comprehensive molecular characterization of human colon and rectal cancer [23]. For 10 of the control samples we inserted bumps uniformly in different locations in the genome. The widths of the bumps were selected using a uniform distribution that matches real bumps and the magnitude of the bumps ranged from one to three. We chose 100 different clusters and inserted one bump by the above method. Another 100 different clusters were chosen and two random bumps were inserted this time. Figure 2.2 shows an illustration of a simulation with two bumps inserted.

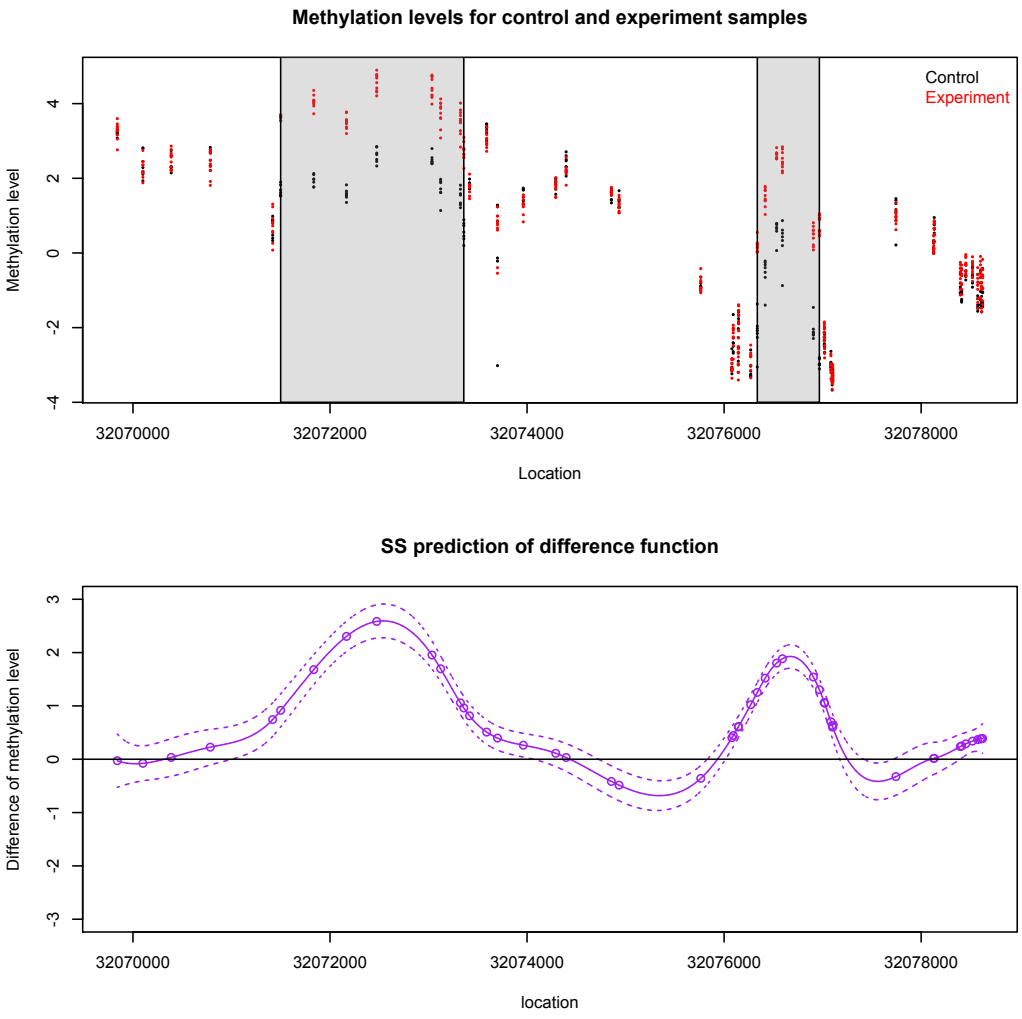

Figure 2.2: **An example of a pair of simulated bumps in a fixed cluster** Above is an illustration of a pair of simulated bumps in a fixed cluster. The two bumps are introduced in this simulation in the shaded region on the top half of the figure. On the bottom half of the figure in purple is the predicted difference function using SSANOVA with 95% confidence intervals.

2.3.1.1 Span parameter for Bumphunter

We used simulated data to identify an optimal span parameter to use for LOESS. The span parameter represents the percentage of data used in each polynomial fit. "Leave one out" cross validation was used to calculate an optimal span parameter. Figure 2.3 shows the results of the cross validation and from this point forward we used span = 0.3, which gave us the lowest error for all LOESS fitting.

2.3.1.2 Comparing Bumphunting to Splines

We used SSANOVA to estimate the difference function and repeat the same with Bumphunter (LOESS). We calculated the difference between the true curves and estimated curves using mean squared error (MSE). Table 2.1 shows the results of the simulations by comparing MSE of Bumphunter and SS in predicting the true difference function. When the magnitudes of the bumps are higher ($\geq$ 2 units) SS performs much better than Bumphunter. The MSE for SS when adding two units are .20 and .36 for single bump and double bumps simulations. For the same simulations the MSE for Bumphunter are .29 and .41 respectively. The improvement is more drastic when three units are added for the bumps. For single bump MSE for SS is .25 and MSE for Bumphunter is 1.42. For double bumps the MSE values are .42 and 1.76, respectively. That is an 82% and 75% improvement for higher magnitude of bumps from Bumphunter to SS.

In our simulations, the difference function, η_d, will be zero at most time points and away from zero at other time points. We compared the two methods to see which

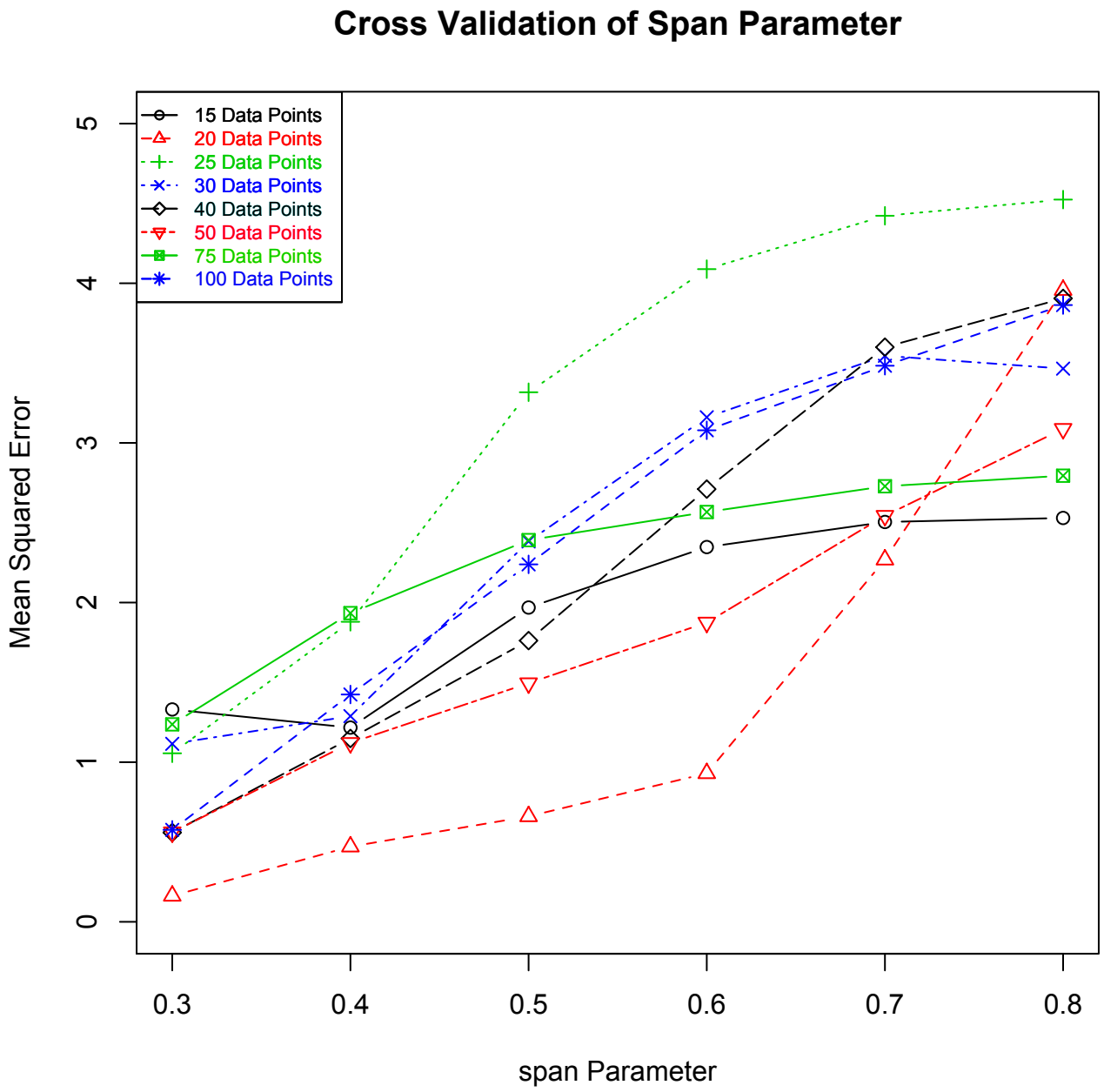

Figure 2.3: **Cross Validation selection of Span parameter** We applied "Leave one out" cross validation technique to calculate the optimal span to use for LOESS. We used LOESS to predict the true difference function, η_d from simulated data. We calculated the mean squared error loss for each span using multiple number of time points randomly chosen. Across different subset of time points span=0.3 is the optimal selection. For the rest of our comparisons we use span=0.3 for LOESS.

performed better in detecting these bumps. Half of our simulations had single bumps and the other half had double bumps. We show the detection performance using an ROC curve in Figure 2.4. AUC for splines was 0.9646 while for Bumphunter it was 0.9175 showing a 5% improvement for SSANOVA in detection. The same increase in performance is seen when a single bump is introduced to the simulation.

Region detection gives us candidate intervals to test using the permutation methods described above. In order to test these regions we compared the two methods at calculating the areas of these regions. Since in this simulation study, regions of interest are known in advance (i.e., where simulated η is not zero), we calculated the area within these intervals as estimated by SSANOVA and Bumphunter and compared it to the area from the simulated η. We performed one thousand replications with random intervals. We calculated the mean absolute error for each of the thousand simulations and compared the distributions of the two methods. Figure 2.5 shows that SSANOVA performs better than Bumphunter calculating areas under $\hat{\eta}$ at specific intervals.

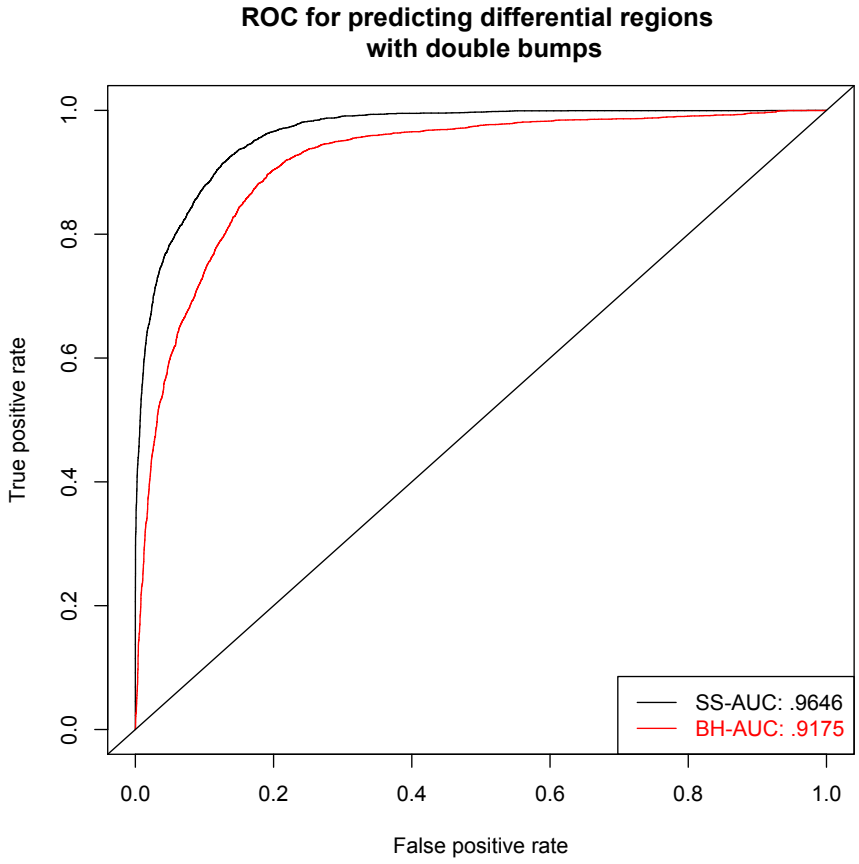

Figure 2.4: **ROC for predicting double bumps** We use the ROC curve to compare the two method's performance on detecting regions throughout the series where the difference between two groups is different from a fixed value (e.g., zero in this case). Spline (AUC: .96) performs much better than Bumphunting (AUC: 0.91).

Single bump	SS	BH	Double bumps	SS	BH
Add 1 unit	0.45	0.42	Add 1 unit	0.84	0.62
Add 2 units	0.20	0.29	Add 2 units	0.36	0.41
Add 3 units	0.25	1.42	Add 3 units	0.42	1.76

Table 2.1: **MSE for difference function with varying number of bumps.** We compare SS to Bumphunter by taking MSE (mean squared error loss) of both method in all types of simulation. In most simulations where the magnitude of the bump is high, we see a improvement using SS over BH. Sometimes the improvement is as high as 120%.

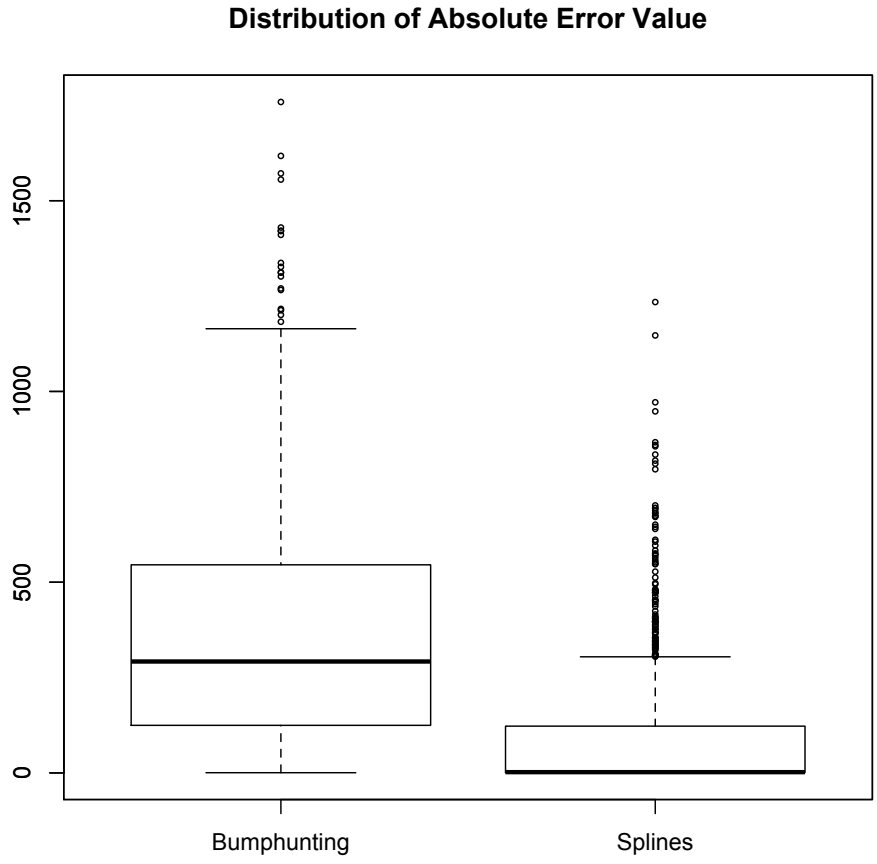

Figure 2.5: **Distribution of Absolute Value Error for both methods.** We used absolute value error value to compare the two methods on simulated data. This boxplot shows the distribution of these error values in calculating areas of region with true difference. For each of the 1000 simulations we picked random intervals that had definite difference between the two groups. The error distribution of splines have a lower mean compared to of the one in LOESS.

2.3.1.3 Comparing Bi-Seq to Splines

We also applied SS method to Bisulfite Sequencing data. We compared it to Bi-Seq [36] in detecting differential regions in Bisulfite sequencing data. Bi-Seq is a five step process that detects differentially methylated regions in targeted bisulfite sequencing data. It applies a modified indirect method in comparison to Bumphunter where smoothing is first performed independently for each individual sample and then testing is done using a beta regression model on the smoothed data.

We used the same RRBS data of bone marrow specimens published by [37]. We used the 12 control samples (four remission, four CD34+ and four promyelocyte) and simulated bumps by placing methylation differences with various lengths and magnitude. We only used chromosome 1 for our simulation. We picked 100 different clusters uniformly and inserted bumps, which we tried to predict using SS and Bi-Seq. We show the ROC curves in Figure 2.6. We report the AUC for SS at 0.782 and the AUC for Bi-Seq at 0.614, which shows a 27% improvement with SS.

2.3.1.4 Comparing DER-Finder to Splines

We also applied SS method to RNA-seq data. We used DER-finder [38] to compare our method at detecting differential regions. Like Bumphunter, DER-finder uses a linear regression model at each base of the genome to identify differential expressions. Then segmentation is done comprised of bases showing similar differential expression signal to find the differentially expressed regions. We used the data avail-

able from the R-Package to split the control samples in half. We introduced bumps uniformly with different magnitude and width and tried to predict these bumps with SS and DER-finder. Figure 2.7 shows the results of the simulations. The AUC for SS is 0.993, which is an improvement over DER-finder (AUC: 0.952).

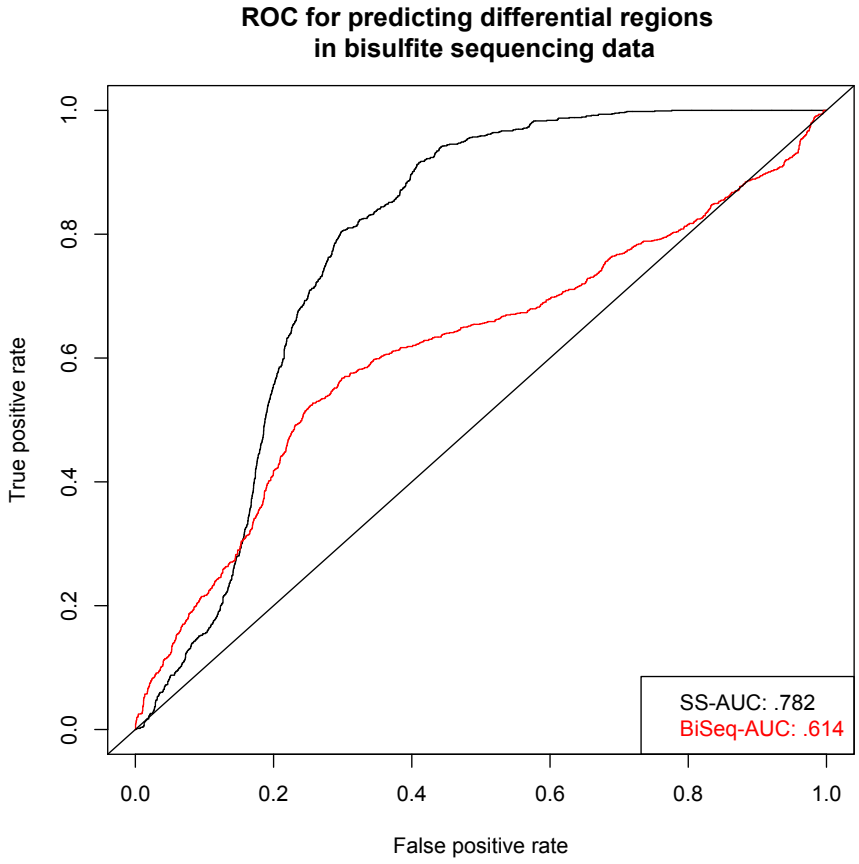

Figure 2.6: **ROC for predicting bumps in Bisulfite sequencing data** We compare the two method's performance on detecting regions throughout the series where the difference between two groups is different from a fixed value (e.g., zero in this case) by showing the individual ROC. Spline (AUC: 0.782) performs much better than Bi-Seq (AUC: 0.614).

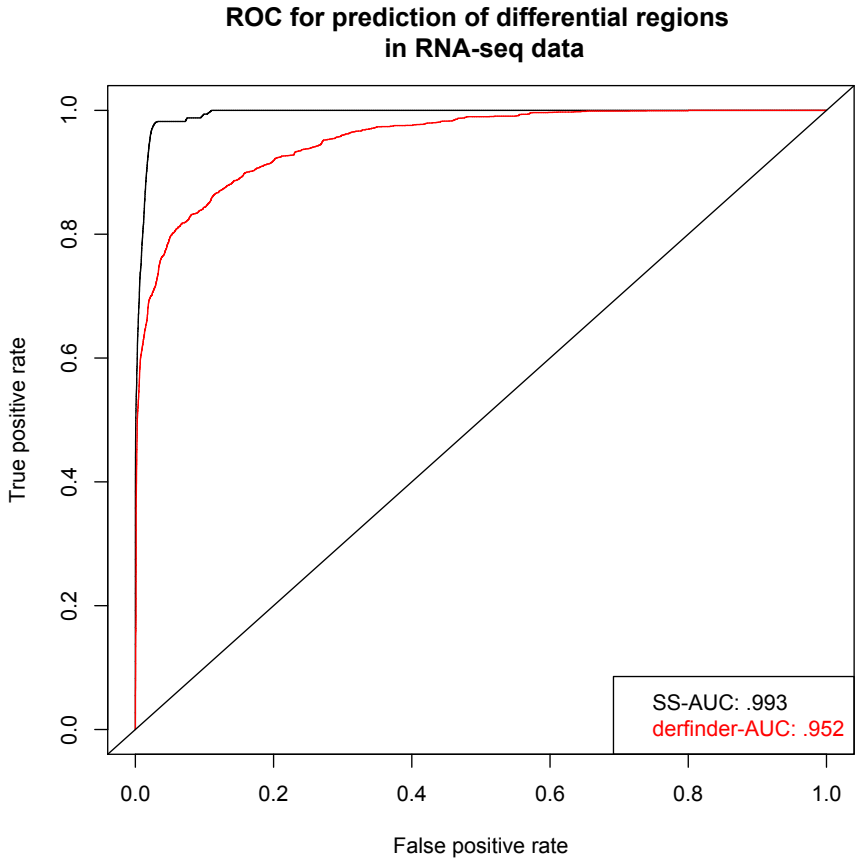

Figure 2.7: **ROC for predicting bumps in RNA-seq data** We compare the two method's performance on detecting regions throughout the series where the difference between two groups is different from a fixed value (e.g., zero in this case) by showing the individual ROC. Spline (AUC: 0.993) performs much better than derfinder (AUC: 0.952).

2.3.2 Applications

2.3.2.1 Colon Cancer Illumina HumanMethylation450k Beadarray Data

We applied the SSANOVA region-finding method to base-pair resolution DNA methylation data assayed on the Illumina HumanMethylation450k beadarray [6] from the Cancer Genome Atlas (TCGA) project [23]. The result is a methylation value for each of 485k locations in the genome. The resulting methylation values can be thought of as a data series with each nucleotide representing sampling point. The goal with methylation data are to identify regions along the genome that correspond to differentially methylated regions between groups of comparison.

We used SSANOVA to estimate the methylation level of each sample at each probe using the model in equation 2 . In this application, $f_1(t)$ represent main effect for probe location, $f_2(I\{k \in i\})$ represent main effect for cell type (tumor vs. normal) and $f_{12}(t, I\{k \in i\})$ represent interaction of location and cell type. Figure 2.9 shows the results of the methylation data. The top half of the plot shows $\hat{\eta}_d(t)$. We showed that using SSANOVA method we were able to find regions in the sequence where it was differentially methylated. The areas in these regions were tested using permutations to be significant or by chance. The bottom panel of Figure 2.9 of the plot shows the permutation results of each region.

2.3.2.2 Metagenomic Data

To illustrate the generality of this approach, we applied the SSANOVA region-finding methodology to a longitudinal metagenomic 16S marker-gene survey. Metagenomics is the study of genetic material recovered from an environmental sample and a field growing in its use of time-series analyses as microbial communities do not exist in equilibrium [39].

In metagenomic 16S marker-gene surveys, conserved regions of DNA from an environmental sample are amplified through a process known as polymerase chain reaction (PCR). The DNA is sequenced, usually with 454®, resulting in thousands of reads 200-400 nucleotide bases (depending on the technology) long genetic sequences representing various bacterial organisms. The reads are annotated to varying phylogenetic levels usually by using BLAST - a greedy search algorithm or a naive bayes classifier called RDP classifier against a database [40, 41]. The abundance of an organism is the number of sequenced reads annotated for a particular organism. Counts can be aggregated to determine the relative abundance of various levels including, genera, species or phyla. Metagenomic data are inherently biased data due to the variation in depth of coverage - the total number of sequences produced for each sample. Data normalization is an initial step in most differential abundance analyses aimed at making feature counts comparable across samples, but because of this variation and few time points per sample, smoothing splines are an ideal candidate to smooth observations. The goal in analyzing metagenomic time-series data are to identify organisms differentially abundant between groups of comparison

at biologically relevant intervals in time.

Using SSANOVA we tested the hypothesis that there was no difference in abundance for a particular class of bacteria due to a difference in diet. We considered each class of bacteria independent of each other. Twelve germ-free adult male C57BL/6J mice were fed a low-fat, plant polysaccharide-rich diet. Each mouse was gavaged with healthy adult human fecal material. Following the fecal transplant, mice remained on the low-fat, plant polysaccharide-rich diet for four weeks, following which a subset of 6 were switched to a high-fat and high-sugar diet for eight weeks. Fecal samples for each mouse went through PCR amplification of the bacterial 16S rRNA gene V2 region weekly. Details of experimental protocols and further details of the data can be found in [24]. Counts were normalized per the cumulative sum scaling method described in [28].

We used SSANOVA to estimate abundance of bacteria using the model in equation 2 . In this application, x_k represent individual mouse effect, $f_1(t)$ represent main effect for time, $f_2(I\{k \in i\})$ represent main effect for diet and $f_{12}(t, I\{k \in i\})$ represent interaction of diet and time. We estimated η_d which represents the difference between the two diet groups of mice with respect to their abundance of a specific type of bacteria. The observed area exceeds the 95% cutoff point, meaning at any given point in the interval, the two diets produce significantly different levels of this particular class of bacteria (Actinobacteria) Figure 2.10. The second part of the plot shows $\hat{\eta}_d(t)$ and the area colored in grey shows the region where there is a difference between the two groups of mice. The last part of the figure shows the results of the permutation test to show significant difference with α=.05. We

performed the same analysis for different classes of bacteria and found at least one differential region in the following classes of bacteria: Actinobacteria, Bacilli, Bacteroidetes, Deltaproteobacteria, Erysipelotrichi, and Gammaproteobacteria (Table 2.2).

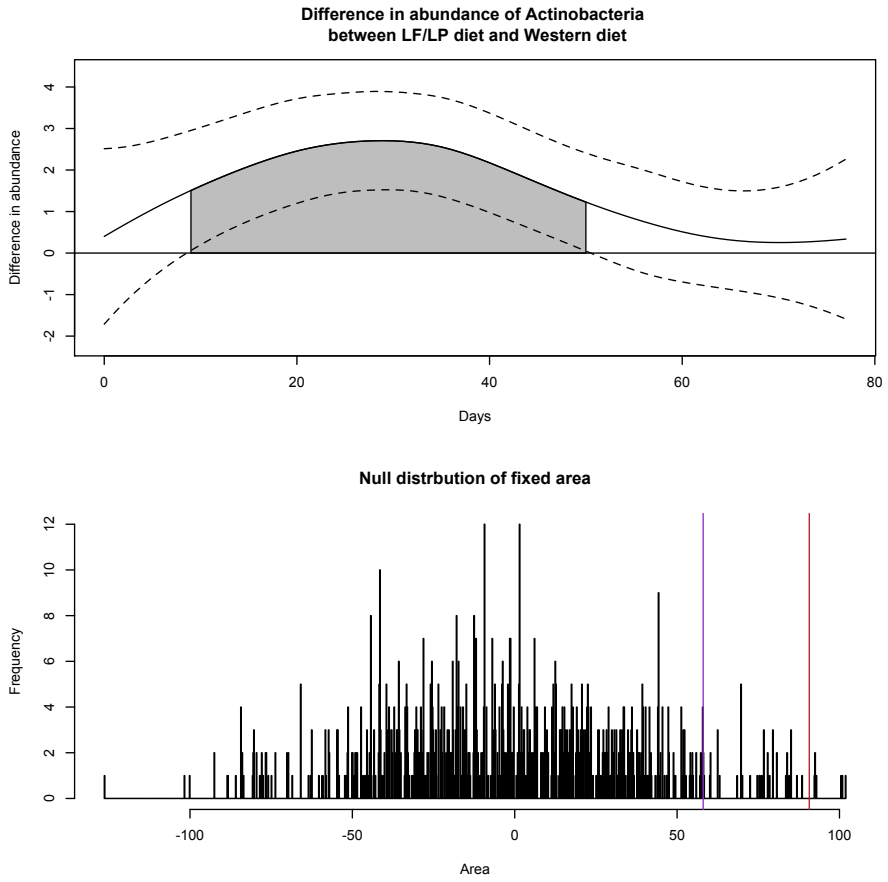

Figure 2.8: **Application of the SSANOVA pipeline on metagenomic data.** The top panel of the figure shows the estimated functions of control (black) and cancer (red) samples. The estimated difference function revealing three candidate regions is shown in the middle panel. Fixing these intervals and applying permutations to calculate a null distribution we observe in the bottom panel the distribution of these permutations and how they are used to detect significant differential methylation.

Candidate regions	Days start	Days end	Area	p-value	Adjusted p-value
Bacteroidetes interval: 1	18.00	20.00	6.43	0.00	0.00
Bacteroidetes interval: 2	22.00	72.00	-108.61	0.00	0.00
Bacteroidetes interval: 3	76.00	77.00	-1.45	0.00	0.00
Unknown interval: 1	63.00	77.00	5.30	0.03	0.07
Unknown interval: 2	0.00	22.00	-10.52	0.11	0.22
Bacilli interval: 1	21.00	77.00	472.29	0.00	0.00
Erysipelotrichi interval: 1	0.00	77.00	126.17	0.00	0.00
Betaproteobacteria interval: 1	24.00	34.00	-22.25	0.00	0.00
Epsilonproteobacteria interval: 1	9.00	27.00	-10.36	0.00	0.00
Gammaproteobacteria interval: 1	24.00	28.00	7.43	0.00	0.01
Gammaproteobacteria interval: 2	42.00	48.00	8.18	0.00	0.00
Gammaproteobacteria interval: 3	15.00	21.00	-26.49	0.01	0.04
Verrucomicrobiae interval: 1	15.00	50.00	29.55	0.03	0.03
Deltaproteobacteria interval: 1	14.00	77.00	122.71	0.00	0.00
Actinobacteria interval: 1	9.00	50.00	90.70	0.01	0.01
Cyanobacteria interval: 1	67.00	77.00	-1.96	0.03	0.03

Table 2.2: **Results of metagenomic data using the function `fitTimeSeries` as stated above.** The results for all class of bacteria along with their regions, calculated area under the curves, p-values and adjusted p-values.

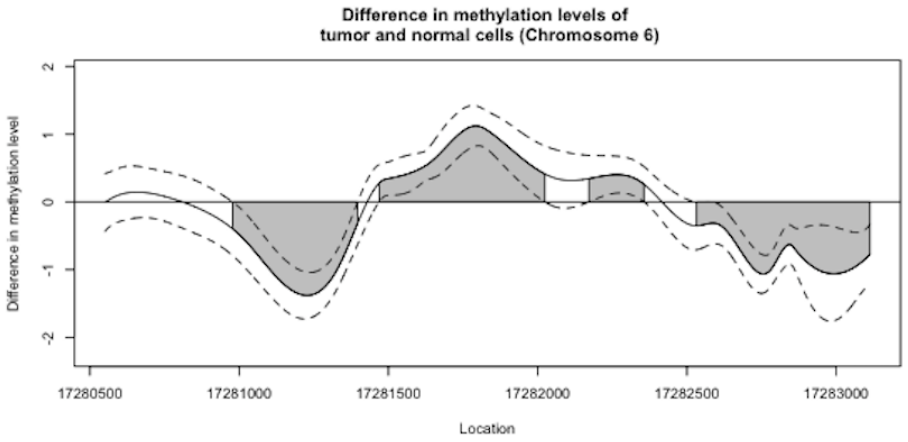

Figure 2.9: **Application of the SSANOVA pipeline on methylation data.** This figure shows the estimated difference function revealing four candidate regions. Fixing these intervals and applying permutations to calculate a null distribution we observe in the bottom panel the distribution of these permutations and how they are used to detect significant differential methylation. The exact areas and p-values for the four candidate regions can be found in Table 2.3.

Interval Start	Interval End	Area	P-value	Adjusted p-value
17281469	17282024	389.09	.001	.004
17282172	17282356	67.93	.023	.092
17280978	17281397	-412.46	.000	.000
17282532	17283113	-458.09	.000	.000

Table 2.3: **Results of methylation data in Figure 2.9** The calculated statistics and corresponding p values for the four candidate regions from Figure 2.9

2.4 Discussion

Splines have been used in the past to address bumphunting problems. [42] used splines to estimate the smoothing parameter λ, needed to find K number of bumps in a function. They used L-Splines (Linear differential operator splines) and other generic smoothing techniques to find the function with the exact number of bumps needed. They use bootstrap sampling to calculate a p-value for the hypothesis test:

$$\begin{aligned} H_0 &= \#\text{ of bumps in } f \leq k \\ H_1 &= \#\text{ of bumps in } f > k \end{aligned} \tag{2.13}$$

where k is fixed and in their application range from 0 to 4.

Friedman [43] used a different approach to solve bumphunting problems. Here instead of directly estimating the function of interest, f, they used a "box" approach to the problem. This is where a box is defined by a range of the whole data. It starts with the whole data being the 1st box. For each box an average is calculated which represents a weighted estimated value of the response variable in that box. Then the data are subsetted into smaller and smaller box till there are k number of boxes with each box having a weighted estimate of the response variable. That value is then compared to threshold and if its larger or smaller then some threshold it is defined as a bump in that range. Both a top down peeling and a bottom up pasting approach to the boxes was used for this bump hunting problem.

We have shown that smoothing spline models provide a direct approach for region-finding based on smoothing methods for genomics data. Indirect methods

require that data are present for both groups across the same time points. The direct SSANOVA does not have this restriction. Data can be distributed across sampling points for one group, and at the same time have data distributed across a different set of time points for the other group. This allows SSANOVA to be applied in a wider range of problems than the indirect approach.

2.5 Conclusion

We have presented a methodology for region-finding using high-throughput genomic data based on smoothing-spline regression methods. We have shown that this direct approach has specific advantages in estimation and the interpretation of these estimates over indirect approaches commonly used for this task. We have also shown the generality of these methods by applying our method to a cancer epigenetics study and a longitudinal metagenomics study. As region-finding applications continue to flourish with the advent of high-throughput assays, specifically next-generation sequencing, the general methodology presented here will address a rapidly increasing number of critical applications in genomics.

2.6 Software

All analyses were performed using R version 3.0.2 [29]. The software package `bumphunter` was downloaded through bioconductor `http://www.bioconductor.org/` and used with default settings in comparing the accuracy of SS region finding [30]. The package `BiSeq` package was used to compare with SS on bisulfite se-

quencing data [31]. The package `derfinder` was used for RNA-seq simulations [32]. The `metagenomeSeq` package was used to normalize the metagenomic data in the applications section [33]. The `gss` package was used to perform SSANOVA and in estimating the difference function [34]. The `pracma` package was used to calculate the area under the curve of η_d [35]. An implementation of the region / time series interval finder exists in the `metagenomeSeq` package version 1.7.18 and higher through the function `fitTimeSeries`.

2.6.1 `fitTimeSeries`

Implemented in the fitTimeSeries function is a method for calculating time intervals for which bacteria are differentially abundant. Fitting is performed using Smoothing Splines ANOVA (SSANOVA), as implemented in the gss package. Given observations at multiple time points for two groups the method calculates a function modeling the difference in abundance across all time. Using group membership permutations we estimate a null distribution of areas under the difference curve for the time intervals of interest and report significant intervals of time. An example is shows on Figure 2.10.

The R object produced from the above function gives the intervals in which there is a significant difference between the two groups of interest. It also provides the user with the areas calculated from each permutation, the fit and se (standard errors) of the predicted difference function, and a plotting function that can be used to plot the difference function. Figure 2.11 shows an example of the plotting

function and the plot it produces.

```
# The gnotobiotic mice come from a longitudinal study ideal for
# this type of analysis.
data(mouseData)

# We choose to perform our analysis at the class level and look
# for differentially abundant time intervals for 'Actinobacteria'.
# For time sake we perform only 10 permutations.
res = fitTimeSeries(obj = mouseData, lvl = "class", feature = "Actinobacteria",
    class = "status", id = "mouseID", time = "relativeTime", B = 10)

## Loading required package:  gss

# We observe a time period of differential abundance for
# 'Actinobacteria'
res$result

## NULL

str(res)

## List of 5
##  $ timeIntervals: num [1, 1:4] 9 50 90.7 0
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:4] "Interval start" "Interval end" "Area" "p.value"
##  $ data         :'data.frame': 139 obs. of  4 variables:
##   ..$ abundance: num [1:139] 0 3.82 3.13 7.4 0 ...
##   ..$ class    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##   ..$ time     : num [1:139] 21 22 28 0 35 6 42 49 56 63 ...
##   ..$ id       : Factor w/ 12 levels "PM1","PM10","PM11",..: 1 1 1 1 1 1 1 1
##  $ fit          :'data.frame': 78 obs. of  3 variables:
##   ..$ fit       : num [1:78] 0.401 0.537 0.67 0.8 0.928 ...
##   ..$ se        : num [1:78] 1.078 1.015 0.96 0.912 0.87 ...
##   ..$ timePoints: num [1:78] 0 1 2 3 4 5 6 7 8 9 ...
##  $ perm         : num [1:10, 1] -2.18 -36.1 -84.33 1.5 88.5 ...
##  $ call         : language fitSSTimeSeries(obj = obj, feature = feature, clas
```

Figure 2.10: **Example of using `fitTimeSeries` function in R** A screenshot that shows an example of using the `fitTimeSeries` function in R from the `metagenomeSeq` package.

```
plotTimeSeries(res)
```

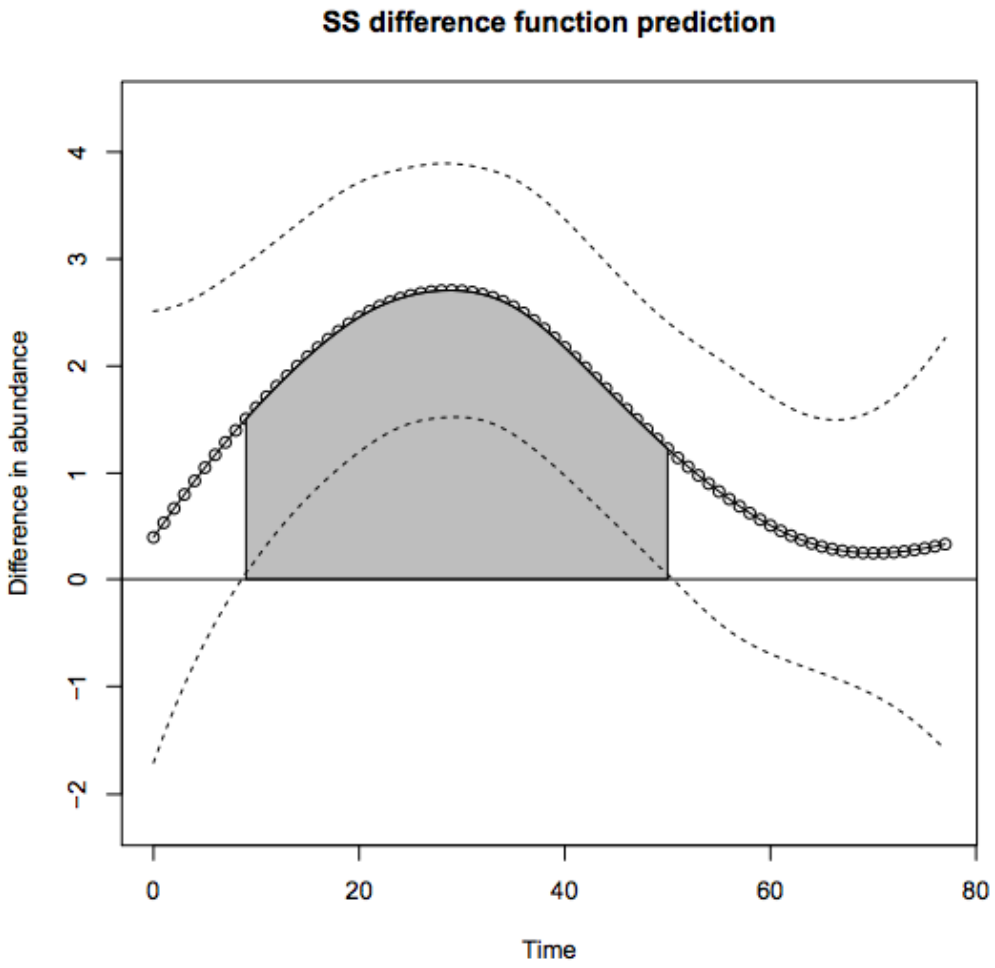

Figure 2.11: **Example of using `plotTimeSeries` function in R** A screenshot that shows an example of using the `plotTimeSeries` function in R from the `metagenomeSeq` package which plots the difference function from the object provided by using `fitTimeSeries`.

Chapter 3: Longitudinal network analysis shows the decline of pop music in the 21st century

3.1 Overview

On December 3, 2013, Spotify released for the first time the company's full business model, source of revenue and artist payouts [69]. Their system keeps 30% of the revenue and pay artists, labels and others the remaining 70%. In describing their business model they showed a plot, (Figure 3.1), summarizing data from the International Federation of Phonographic Industry (IFPI) to describe the decline in physical sales and increase in importance for revenue sharing systems and business models [72, 73].

Figure 3.1 shows a drastic change in the music industry that begins around the mid 90s. The revenues collected from physical music sales (Cassettes, Cd, etc.) start to level off around the mid 90s and decrease soon afterwards. Following the internet boom, downloads began to become a significant portion of the overall music sales. In January, CNN-Money [70] described the decline of revenues in music sales by using the sales of the number of cassettes, CDs, downloads, and others . The same declining trend in Figure 3.1 can be seen around the mid 90s for the number

of total units sold. The reasons for this trend vary from the general public losing interest in physical albums to the rise of technology and easier alternative access to music. What is more interesting is the response from the music industry to the decline in sales during the mid 90s.

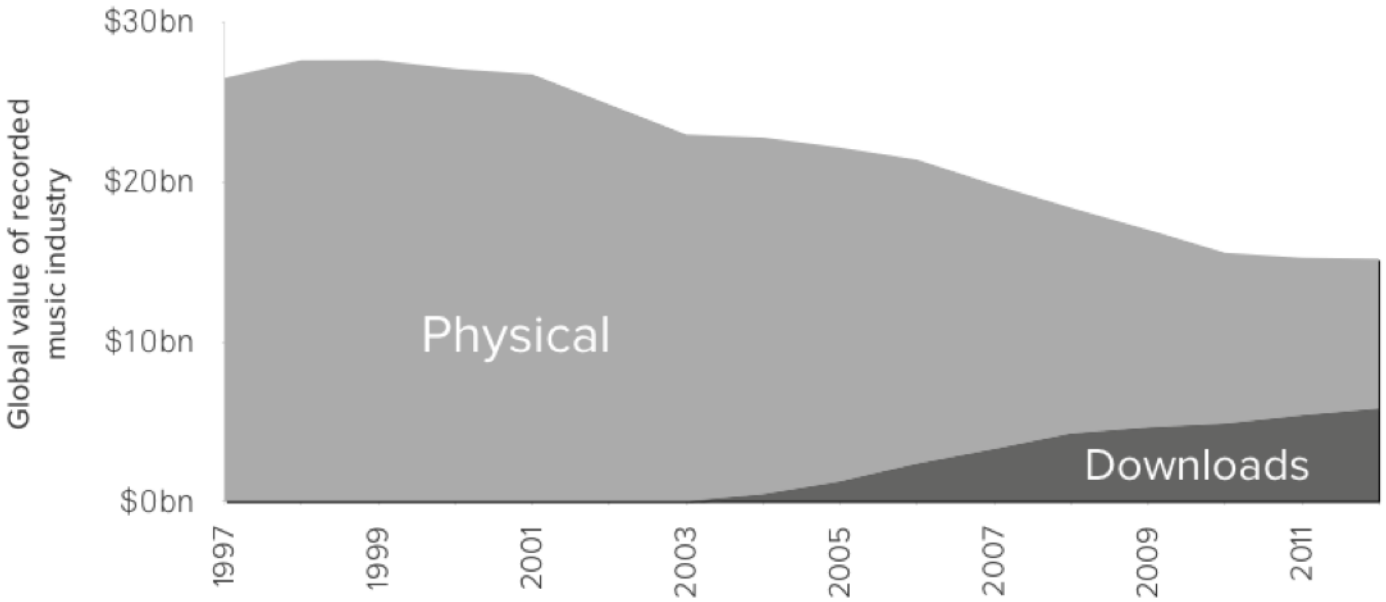

Figure 3.1: **Physical sales and download sales over time** This figure, posted by Spotify [69, 73], shows the trend of physical record sales versus download sales. The trend for physical record sale is decreasing starting in the mid to late 90's.

To increase album sales, music labels need to invest in advertising costs, including the cost of making and advertising singles. One could argue that singles play the role of trailers for the albums. To understand the music industry's response to the decline in physical sales we decided to investigate the number of writers involved in each hit single from 1970-2013. These are songs that have peaked at number one in the Billboard Hot 100 chart [71] released weekly. In a given year, at most 52 singles can make our list if each week there was a different number one song. We looked at the network formed by the writers of these hit singles. Nodes and edges define a network graph. In this case, nodes represent writers and edges represent collaboration for a song in that particular year.

Studies have used longitudinal network analysis to show the spread of information (happiness, obesity) in individuals over time [74]. Individual node degree and edge relationships over time have also been throughly written about in recent years [79]. Teng has used time series analysis to show that network structure alone can be highly revealing of the diversity of information being communicated [85]. Nodes need to be the same over time for this method to work, which is different from the music writer data. Sun [84] uses GraphScope to find communities forming through time in a larger network setting. This analysis is done without using any user specific parameters.

We take a similar approach to Hidalgo and Sun [75,84] to study the network as a whole and how it is changing over time. As noted by McCulloh [78], there are many types of dynamic changes occurring in a network over time. Markov chain models, multi agent simulations, and statistical models are used to study individual node

changes over time and characteristics of these nodes. There are four dynamic states that are changing in a network. A *stable* network is defined as a constant network that does not change over time except for random noise. An *evolution* occurs when interactions between node forces the dynamics of the network to change over time. A *shock* is an outside covariate affecting the social group in a network and finally, a *mutation* is incurred when a shock causes evolutionary changes in a network. For the music writers network we look at mutations caused by the music industry and lack of sales.

3.2 Methods

3.2.1 Data collection

The network studied in this chapter is writer networks of Number 1 hit songs in Billboard top 100 [71]. Billboard releases a top 100 chart weekly where they list the 100 best songs of that particular week. This chart takes the aggregate of record sales, radio airplay, downloads and others together to make the list of the 100 best songs. We looked at any song that went to the top spot of this chart. In a given year there could be at most 52 unique songs if each week there was a different number one song. We collected data from 1970-2013 of the Billboard top 100 chart and made a network out of these song writers.

3.2.2 Problem

Using the music writer data we construct a network of writers for 43 consecutive years. We want to model the changes in network structure using measures such as degree distribution and network density. In addition, we will analyze different covariate influences on these changes. We are not concerned with individual subject variables (node and edge presence or absence), instead we focus on how this music writer network is changing over time as a whole. These changes are called *mutations*, that are usually caused by *shock* events in a network [78].

3.2.3 Overdispersed Poisson model

Following methods in Zheng [67], we use a overdispersed Poisson model to estimate degree distribution through time using covariates. A Poisson model assumes that mean and variance are the same for the variable of interest. We assume:

$$Y_{ij}|\lambda_j = Poisson(\lambda_j) \tag{3.1}$$

where Y_{ij} is degree of node $i = 1, \ldots, I$ at time $j = 1, \ldots, J$. We assume each time frame has its own rate parameter, λ. Because of the difference between in mean and variance in our response variable we assume λ to also have a prior distribution:

$$\lambda_j \sim \text{gamma}(r, \frac{1-p}{p}) \tag{3.2}$$

and

$$f(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}exp(-\beta\lambda) \tag{3.3}$$

where α=r and β=$\frac{1-p}{p}$.

We can calculate the marginal distribution Y_{ij} as:

$$\begin{aligned} f(y) &= \int_0^{\infty} f(y|\lambda)f(\lambda)dy \\ f(y) &= \int_0^{\infty} \frac{\lambda^y}{y!}exp(-\lambda)\lambda^{r-1}\frac{exp(\frac{-\lambda(1-p)}{p})}{(\frac{p}{1-p})^r\Gamma(r)}d\lambda \\ f(y) &= \frac{(1-p)^r p^{-r}}{y!\Gamma(r)}p^{r+y}\Gamma(r+y) \\ f(y) &= \frac{\Gamma(r+y)}{y!\Gamma(r)}p^y(1-p)^r \end{aligned} \tag{3.4}$$

where $Y \sim \text{N.B}(r,p)$, also known as overdispersed Poisson.

3.3 Results

3.3.1 Number of hit songs and writers of hit songs

We show the trend of the music industry changing over the last forty years by looking at the number of songs that are number 1 along with the number of writers working on these songs (Figure 3.2). The number of songs making it to the Billboard top hit list follows a cyclic relationship over the timespan. It shows the general public music taste varies from year to year. Some years (around the mid 1970's) the general public responds to a large number of songs as their favorite and, in contrast, in other years (around 1980) the general public only listens to specific songs as the total number of hit single drastically decreases. The overall cyclic trend

for the total number of songs is seen in a smaller yearly interval while the overall forty year trend for the total number of songs decrease drastically in the mid 90's.

Also apparent in Figure 3.2 is the cyclic trend for the total number of writers working on hit singles on a yearly basis. This cyclic trend, like the trend for number of song writers, is more evident in smaller yearly intervals. The total number of writers over the 40 plus years stay relatively constant.

We show the number of writers working on hit singles drastically increase in the early to mid 90's. This increase can clearly be seen in Figure 3.3. There are 1-3 writers on average working on hit singles before the mid 1990's. This number drastically increases to 2.5 - 4 writers per hit single after that time period. The music industry changed something in the business model for that increase to occur. The number of writers working on hit single might have been increased due to low album sales [69].

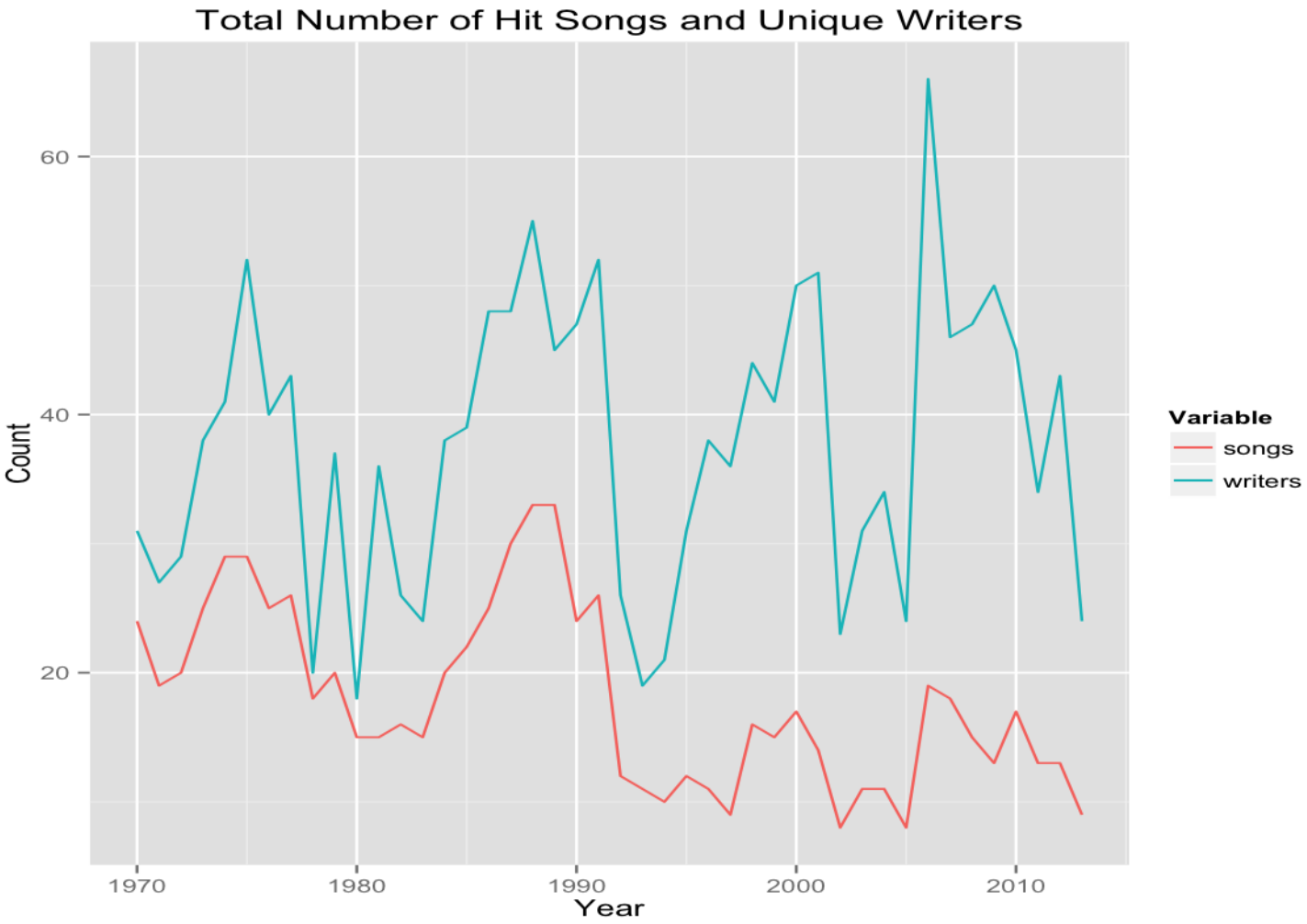

Figure 3.2: **Total number of hit songs and writers yearly** The total number of the hit songs yearly and writers working on these songs yearly follow a cyclic relationship. The overall forty year trend for number of songs is constant while the number of writers drastically decrease around the mid 90's.

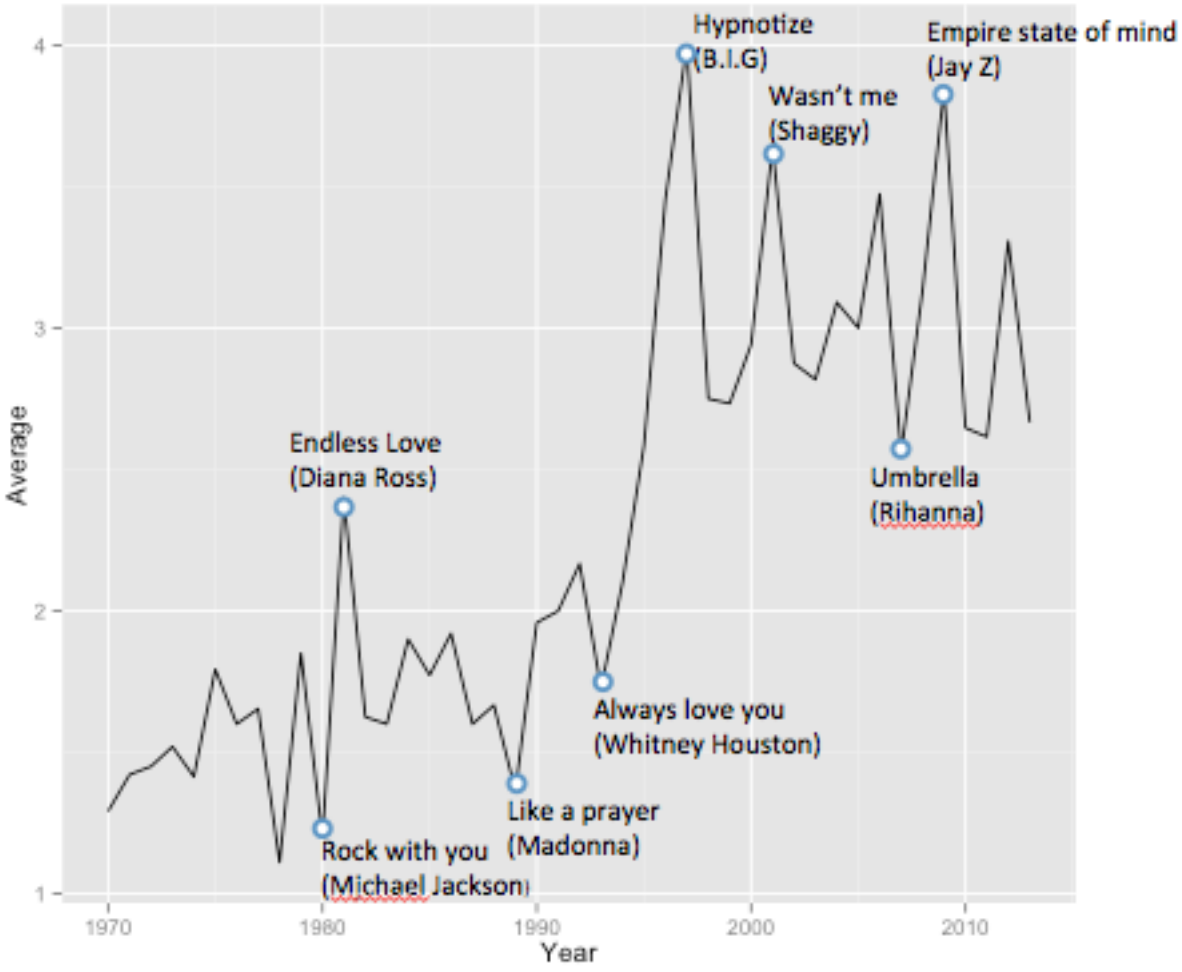

Figure 3.3: **Average number of writers per song** The number of writers working on hit songs started to increase drastically around the early to mid 90's. I highlighted some hit songs for each year such as Like a Prayer (1989), Hypnotize (1997), etc.

3.3.2 Network of musical song writers

We looked at the network of song writers on a year to year basis. We show the changes in the topology of music writer networks using characteristics of the yearly networks. Figure 3.4 shows the change in writer networks over time. Here we highlighted six network graphs over the years where it is clear the writer networks are becoming more and more dense over time. Not only are more writers being hired to work on individual singles, it is also the same writers being hired as can be seen from the network graph of 2010. Also clear are the disappearance of single nodes in the network. This indicates fewer songs is being written by only a single individual in 2010, whereas single songwriters were the norm before the 1990's.

Figure 3.5 shows the distribution of node degree for six selected years. This shows the evolution of network density over time. The number of nodes with degrees 1 and 0 is very rare after the mid 90's which corresponds to more and more collaborations in the music industry for hit songs. This change can also be seen by looking at other characteristics of the individual networks over time such as network density, clustering coefficient largest connected component.

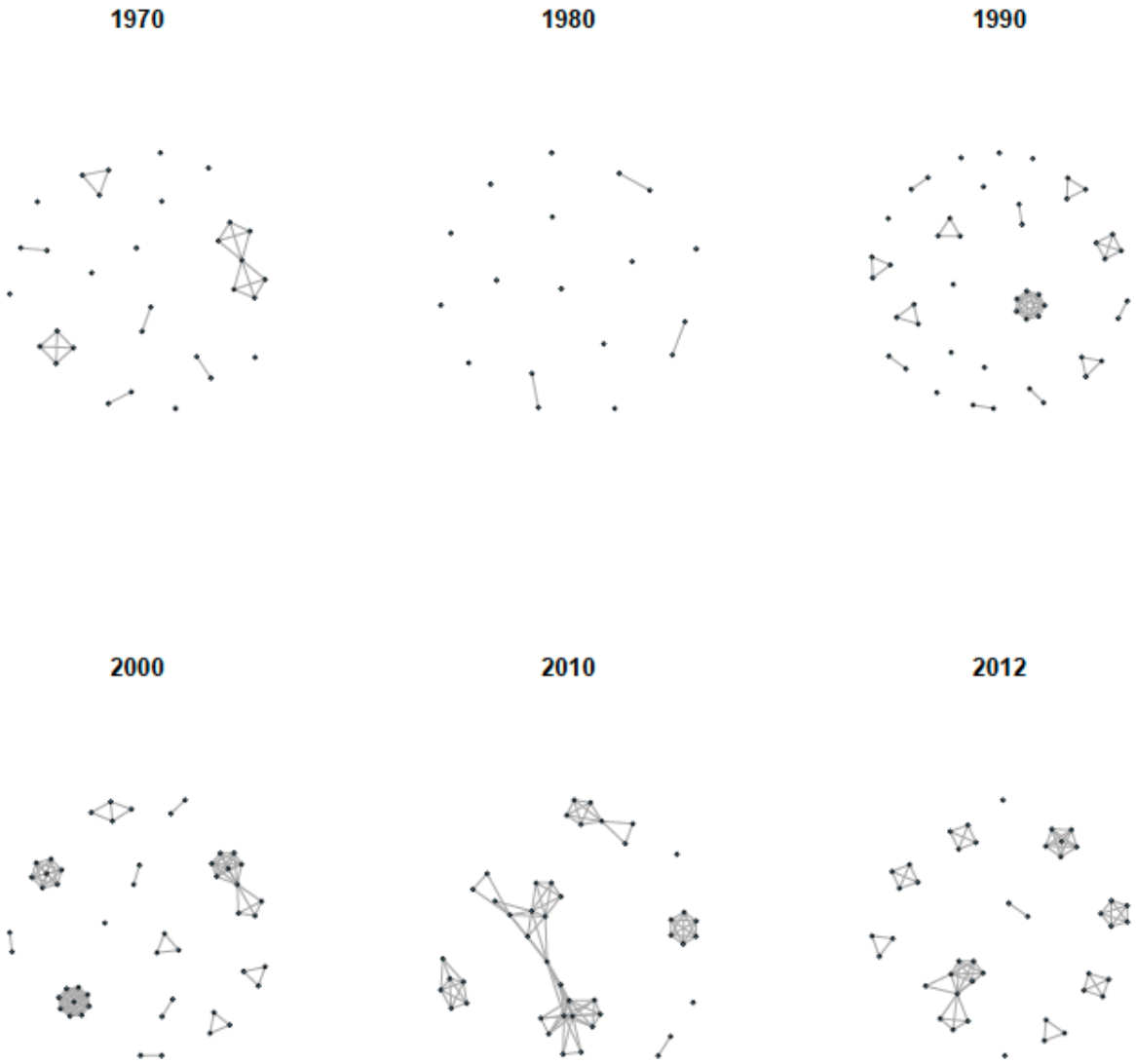

Figure 3.4: **Evolution of network overtime** Highlighted here are six yearly music writer networks. Nodes in the individual yearly networks represent writers and edges between writers represent a collaboration on a number 1 hit song. The drastic change is seen through time as the networks are getting more and more dense as time pass.

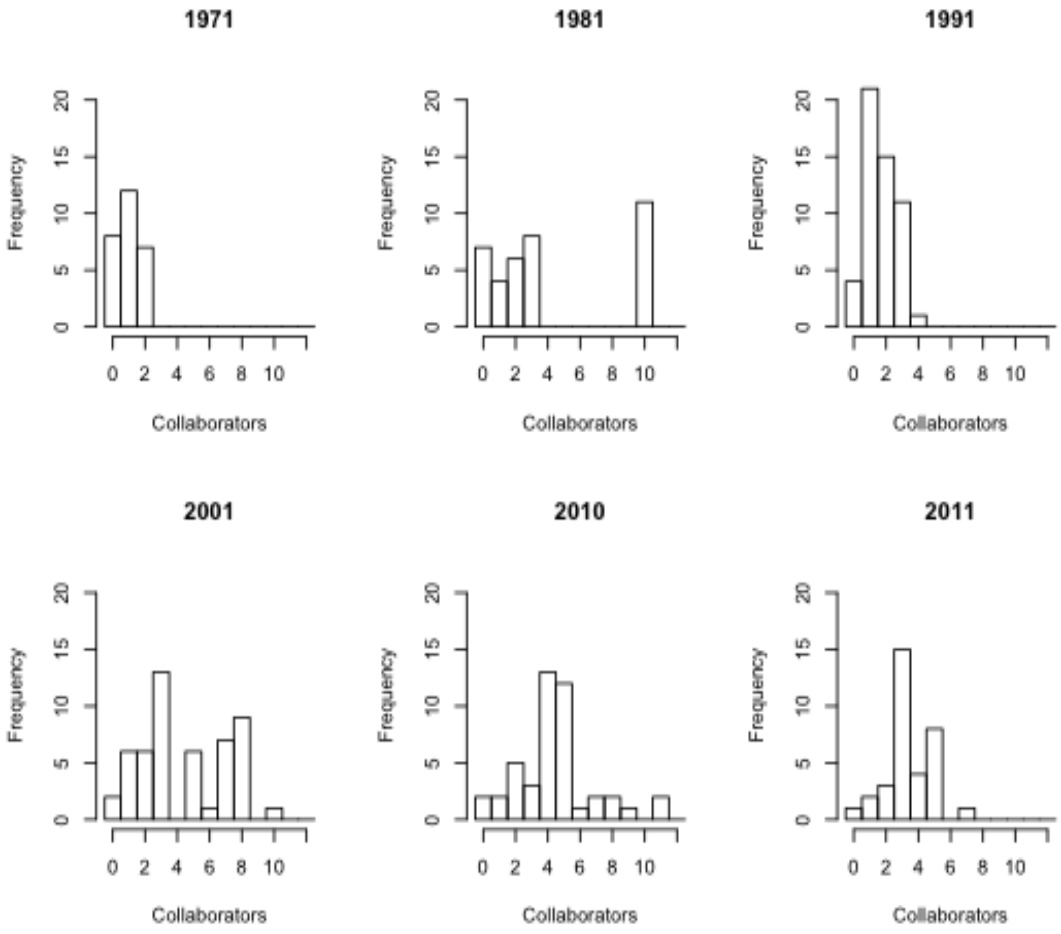

Figure 3.5: **Evolution of node degree distribution overtime** Distribution of node degree is shown here for six of the years. Before the mid 90's most of the writers (nodes) had 1-2 collaborators (edges) yearly in comparison to after the mid 90's where most of the nodes have more than 2 edges. The distribution of node degree is overdispersed after the mid 90's.

3.3.3 Prediction of node degree with covariates

We predict node degree for a given year using some chosen covariates. We chose multiple variables that we think can help explain these changes in network density over time. Figure 3.6 shows the predicted values of average node using the selected independent variables. Even though, these variables can predict the mean of node degree we wanted to use a overdispersed Poisson model to also predict the variances of node degree for a particular year.

We used overdispersed Poisson Model [67] to estimate the mean and variance of node degree for a particular year. Since movie sales provided the best results using a regular Poisson model we chose this independent variable to predict node degree using the overdispersed model. Figure 3.8 shows the difference between estimated variance and observed variance using the two models (regular and overdispersed Poisson). It clearly shows the overdispersed Poisson model captures the mean and variance relationship of node degree much better than a regular Poisson model.

We used the overdispersed Poisson distribution to predict node distribution on a year to year basis. We predicted the node distribution for all the years using the Poisson models and the overdispersed Poisson model. We observe in the years where overdispersion occurs, the overdispersed Poisson distribution predicts the node distribution much better. Figure 3.7 shows an example of such a year where overdispersion occurs in observed node degree. We clearly show the negative binomial (overdispersed Poisson) is a more suitable distribution assumption for node degree where overdispersion is occurring. For the years where overdispersion is not

occurring a regular Poisson distribution will suffice.

3.3.4 Overdispersed Poisson distribution and Poisson distribution

Figure 3.9 shows the Q-Q plots for different distribution fits using Poisson and three levels of Negative Binomial distributions. For all four distribution a $\mu = 3$ was used. α was changed for each of the three NB distribution fits. For the NB fits the mean and variance is defined as $E(y) = \mu$ and $Var(y) = \mu + \frac{\mu^2}{\alpha}$. The negative binomial fit with $\alpha = 4.39$ estimated from movie sales as the independent variable has the best fit.

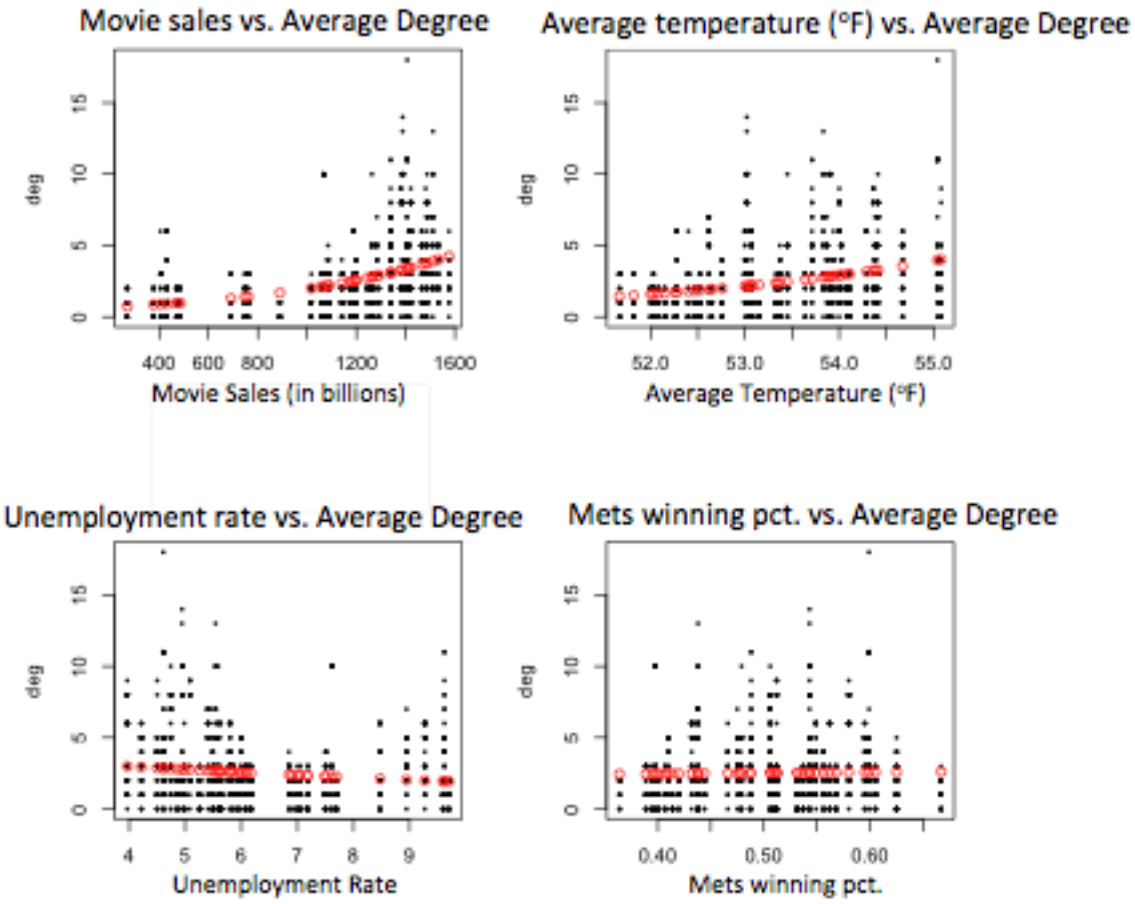

Figure 3.6: **Estimated Poisson regression models** In red are predicted values of average node using the four selected independent variables. Poisson regression was used to evaluate these predicted lines.

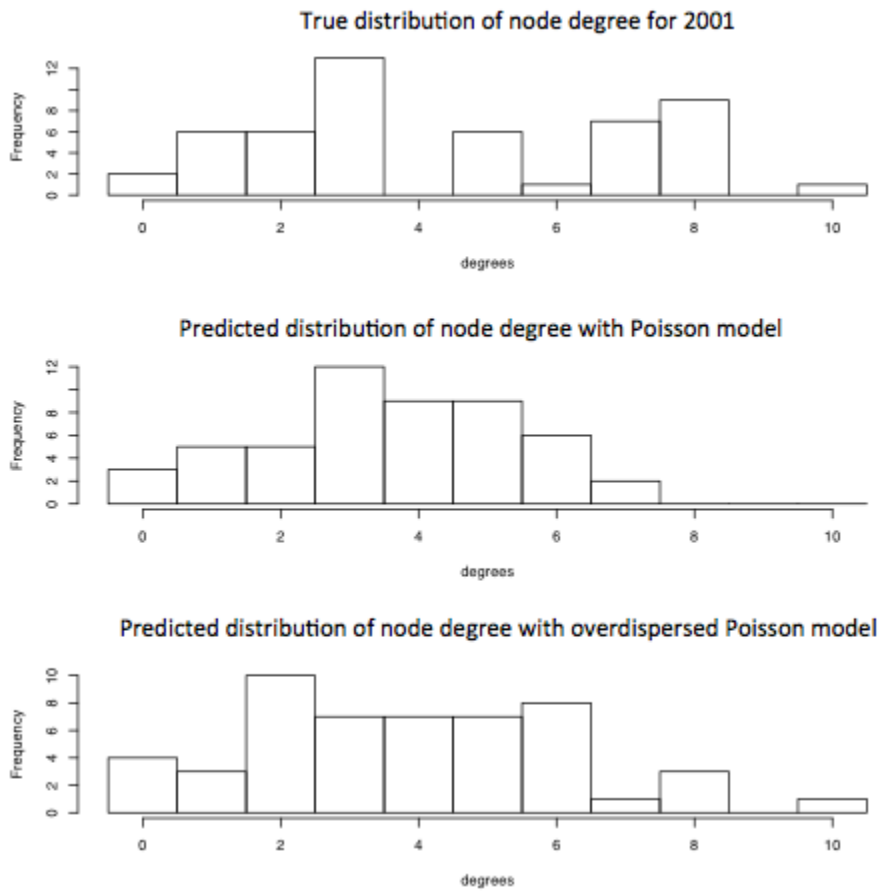

Figure 3.7: **Comparison of Poisson and overdispersed Poisson in estimating node degree distribution** For the year 2001, the observed node distribution is the top panel of this figure. The middle panel is the estimated node distribution assuming a Poisson distribution and the last panel is the estimated node distribution assuming a negative binomial (overdispersed Poisson) distribution.

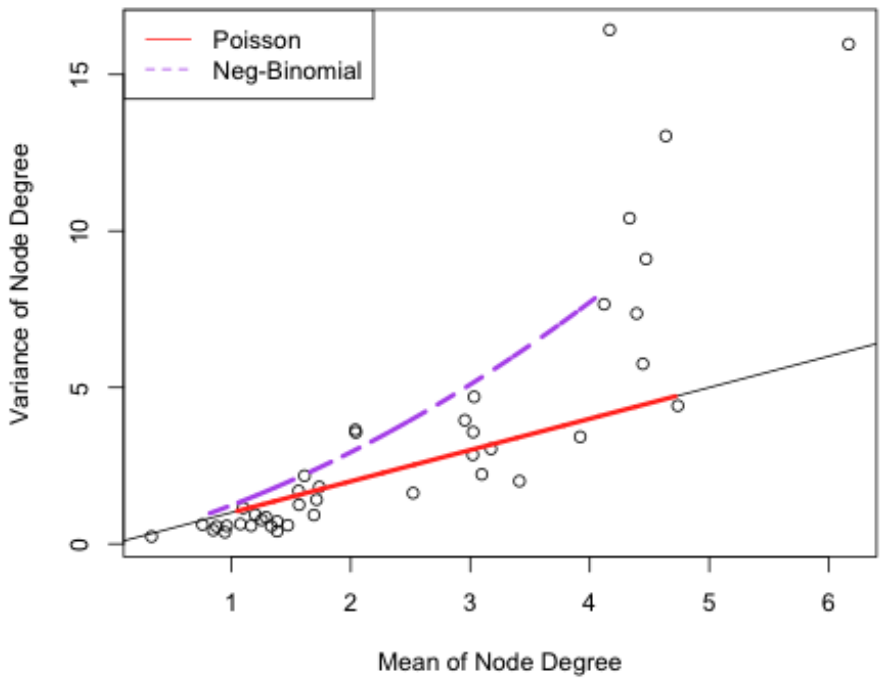

Figure 3.8: **Comparison of Poisson and overdispersed Poisson in estimating node degrees.** We plot the average degree and variance of degrees for each year and try to estimate it using Poisson and overdispersed Poisson model. The overdispersed Poisson model (in purple) fits the relationship much better.

	AIC (Poisson)	AIC (NB)	overdispersed estimate
Movie Sales	6642.4716	6409.2316	0.2277
Unemployment %	7280.0963	6776.8428	0.3847
Mets Win %	7321.2478	6795.2497	0.3950

Table 3.1: **AIC between two models and overdispersion parameter estimates** overdispersed Poisson model (NB) performs much better than a regular Poisson regression as shown by the lower AIC across all NB models. Also shown is the estimation of the overdispersion parameter which scales the variance of the response.

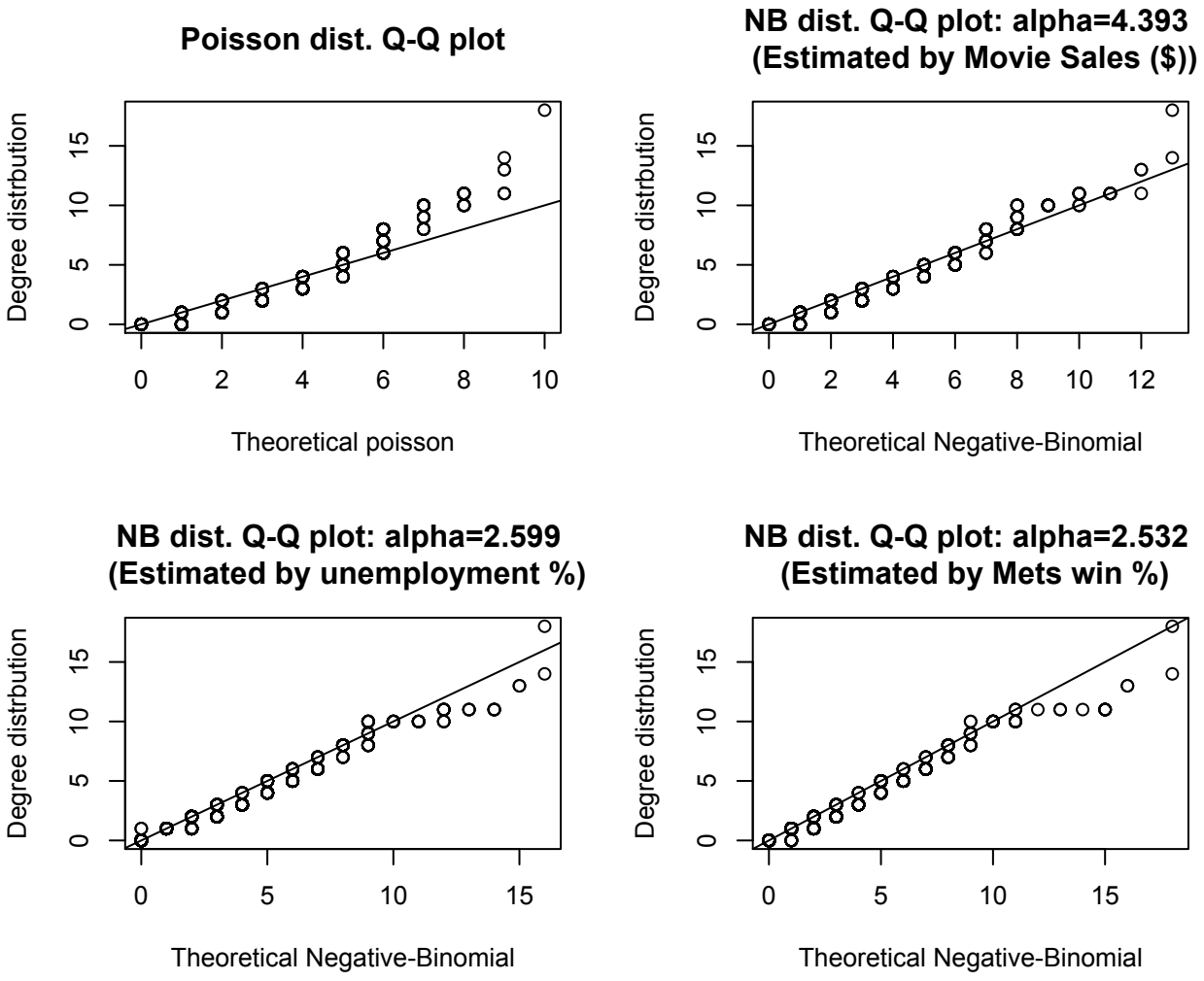

Figure 3.9: **Comparison of Poisson and negative binomial distributions with node degrees.** We show the four Q-Q plots we compare the node degree distribution with. The negative binomial distribution with the three selected independent variable performs better. The best fit appears to be with Movie Sales ($) as the independent variable.

3.4 Discussion

The observations in network dynamics can be partially explained by the revenue loss starting in the mid 90's by the music industry. There was a clear decision made in the mid 90s to invest more money (implying more writers) into hit singles. For the first time in more than twenty years, revenues were not increasing, so something needed to be changed. Once the revenues started to level off again with the increase of downloads and single sales, the music labels decided to continue the trend and hire more writers per song. As reported in the the CNN-Money article, it can be seen that downloads and single sales sum up to a significant portion of music sales starting in the mid 2000s. This is why even though revenues started to level off once again, this time, singles became even more important so the number of writers working on these songs stayed steady.

Also important to note are the writers being hired by the music labels. The same writers appear in the top singles list over and over again. During the last five years Max Martin and Lucasz Gottwald appear in the top singles every year more than once. The music labels are hiring the whole team for top singles. These singles are bringing in more money into the music labels. The recent Grammy nominations for 2014 nominated both Lucasz Gottwald and Max Martin for song of the year for the contributions to Roar by Katy Perry.

3.5 Conclusion

We used an overdispersed Poisson model to estimate the node degree distribution of musical writers over time. We show the best predictors of node degree of musical writers are movie sales in dollars. We also use graphical network analysis to show there was a *mutation*, cause by declining sales in music industry, in the musical writer network.

3.6 Software

All analyses were performed using R version 3.0.2 [29]. The software package `igraph` was used for all network plots and figures. The function `glm` was used to perform Poisson regression and overdispersed Poisson regression on the data. The package `XML` was used to scrape the data from Wikipedia.

3.6.1 R-Shiny website

We used the R package `Shiny` to make an interactive website that summarizes the musical network. The website can be found in `http://epiviz-dev.cbcb.umd.edu/shiny/musicwriters/` with the code found in `https://github.com/htalukder/musicwriters`. Figure 3.10 shows a screenshot of the website. Panel A has a time bar where the user can select the year from 1970-2013. Having selected an year, panel B and D will simultaneously change automatically. Panel B displays the network of music writers of that particular year selected in panel A.

If the user hovers over the network nodes, the name of the writers will be visible. Panel C has four network dynamic characteristic the user can select from. They are clustering coefficient, average degree, network density and largely strongest connected component. Here, interactively, the user can hover over the graph and the point will display the amount for that particular year. Panel D also changes with the year selection from panel A. A Spotify playlist of the songs that were number 1 hits for the chosen year will show on panel D. The user can listen to each of the songs as they are learning the writer relationships for a particular year.

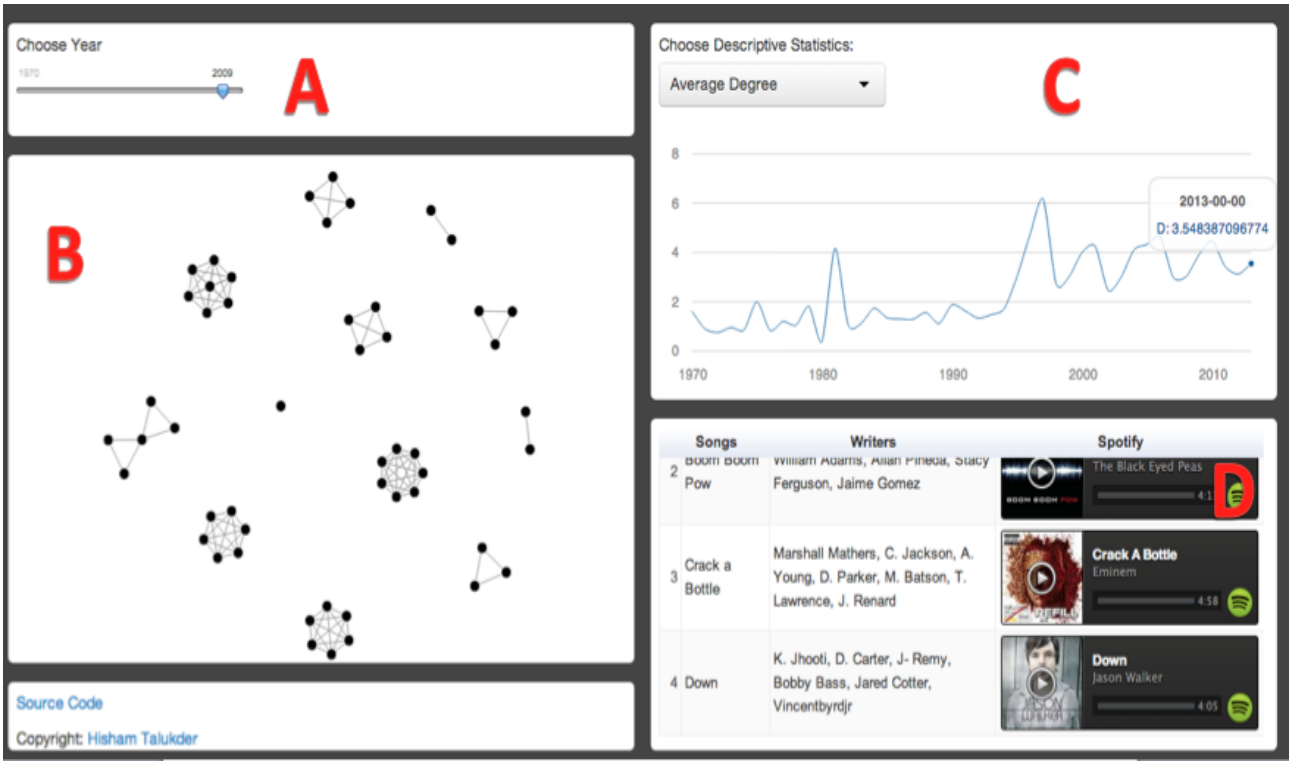

Figure 3.10: **Screenshot of Shiny website.**

Chapter 4: Does health insurance matter? Establishing insurance states as a risk factor for mortality rate.

4.1 Overview

Trauma is the most common cause of death in persons between the ages of 1 and 44 in the United States, and the fifth most common cause of death overall (CDC). According to the National Hospital Ambulatory Medical Care 2010 Survey, approximately 37.9 million Americans are treated for traumatic injuries annually. Insurance status, a surrogate for socioeconomic level, has been shown to have pervasive effects on outcomes in trauma, particularly mortality. Race, income, and insurance status impact mortality in trauma and play a role in rehabilitation placement after brain surgery [48, 54]. Insurance status and race predict mortality in pediatric patients with traumatic injuries [56]. Arriving at off-hours worsens outcomes in trauma, and access to care matters, as studies have shown that level I trauma centers have better survival outcomes than their level II counterparts [52, 77].

In this chapter, we examined disparities in mortality rates between self-pay and insured adult (18-64) patients using data from the National Trauma Data Bank (NTDB), a repository of patient data compiled from trauma centers across the

United States. We examined variables that might act as confounders age, sex, race, mechanism, severity of injury, region of the country, hospital size and tested whether self-pay patients and insured patients differ in mortality rates after controlling for relevant trauma center and patient characteristics.

4.2 Methods

4.2.1 Data

The NTDB is a repository of trauma related data voluntarily submitted by participating trauma center across the United States. This particular version (V7.2) contains data on individual patient cases in over 900 trauma centers. It includes all the data submitted to NTDB with admission year 2002-2006. We obtained the following variables from the database: sex, race, gender, age, insurance status, in-hospital mortality, and patient disposition after treatment. Patient disposition after treatment was divided into those who were discharged to home and those who were transferred to a rehabilitation hospital. Following methods outlined in extant literature [55], insurance status was divided into a binary variable: insured or self-pay patient. The following groups are considered to have insurance: Automobile, Blue Cross, CHAMPUS, Government Military, Liability insurance, Medicare, Medicaid, MCO, Crippled Childrens, No fault insurance, other commercial indemnity plan, Private charity, and Workers compensation.

Different characteristics of trauma centers were also obtained from the database. For each trauma center, the total number of beds was used as a proxy for the size

of the trauma center. NTDB includes the location of the trauma centers into four regions across the United States: Northeast, Midwest, South, and West. The American College of Surgeons Committee on Trauma designates trauma centers as one of four different levels. For this analysis, we focused on level I and II centers, which serve as referral centers, receive the greatest number of patients, and have the most resources. Level II trauma centers are regional facilities that have 24-hour emergency medicine and trauma services that may initiate definitive care, but they have limited on-site availability of surgical subspecialties (e.g., otorhinolaryngology and oral maxillofacial surgery are not present on-campus at all times). Level I trauma centers are tertiary-care referral centers with 24-hour staffing of all surgical and medical specialties. Because they must also host clinical training programs and conduct research, level 1 trauma centers are generally staffed by leaders in the field who have early access to new treatments.

A logistic regression analysis was done using the mortality outcome of patients as the response variable. A significance level of 5% was used for all hypothesis testing throughout the data analysis. We use the logistic regression model to specifically compare outcomes between insured patients and self-pay patients, while controlling for age, race, gender, size of facility, region of facility, facility level, mechanism of injury (blunt or penetrating), and time of admittance.

We limited our investigation to adults less than 65 years of age, since that is when all US citizens qualify for insurance through Medicare. The Injury Severity Score (ISS) is a non-linear and anatomy-based scale that quantifies the seriousness of a patients injuries. For this analysis, we followed several examples in the literature

[8, 9] and categorized the scores into a clinically-relevant scale. ISS$<$ 9 was defined as being mild, 9-16 was moderate, 16-25 was considered severe, and patients with an ISS $>$25 were considered to have a critical injury. Although patients that are admitted to trauma centers receive an ISS, there is variation in injuries receiving the same ISS scores. Most of the patients being admitted to facilities either had a blunt trauma or a penetrating trauma. Penetrating trauma is defined as an injury in which an object pierces through the skin and tissue. Blunt trauma is physical trauma caused to a body part without any piercing of the body.

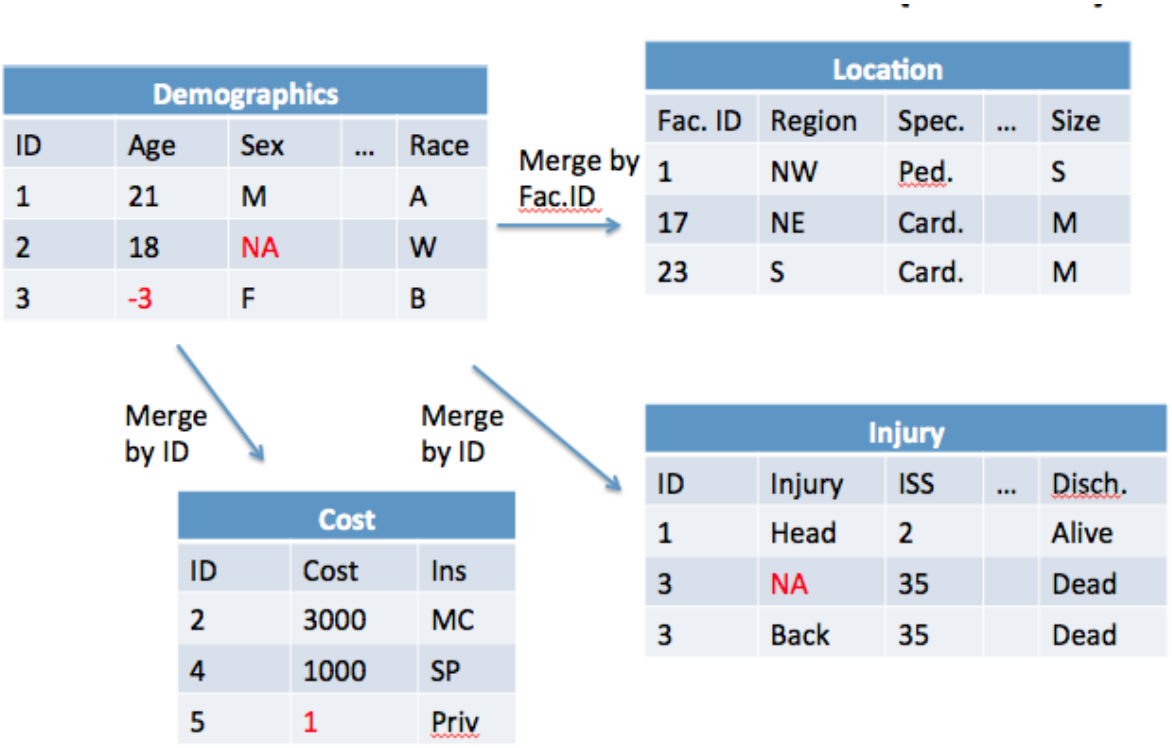

Figure 4.1: **Illustration of data cleanup and merging** The separate data sets from NTDB were cleaned first then merged using the appropriate variables. All the analyses were done using the result datasets.

Different characteristics of trauma centers were obtained from the database. For each trauma center, the total number of beds was used as a proxy for the size of the trauma center. The location of the trauma center was divided into four different regions across the United States; Northeast, Midwest, South, and West. Trauma centers were designated as two different levels. Level II trauma centers are regional facilities that have 24-hour emergency medicine and trauma services that may initiate definitive care, but they are limited in the on-site availability of surgical subspecialties (i.e., otorhinolaryngology and oral maxillofacial surgery are not present on-campus at all times). Level I trauma centers are tertiary-care referral centers with 24-hour staffing of all surgical and medical specialties. Because they must also host clinical training programs and conduct research, leaders in the field who have early access to new treatments generally staff level I trauma centers.

4.2.2 Review of Logistic Regression

A logistic regression [18] analysis was done using the mortality outcome of patients as the response variable. We wanted to specifically look at the trauma outcomes between insured patients and self-pay patients. We controlled for the following variables: age, race, gender, size of facility, region of facility, facility level, type of injury (blunt or penetrating), and time of admittance. Because of interactions between ISS scores and other variables we looked at mortality outcomes in different ISS groups.

Payment source	Number of cases
Automobile Insurance	4201
Blue Cross/Blue Shield	6827
CHAMPUS	786
Government/Military Insurance	2094
Liability Insurance/Under Litigation	687
Managed Care Organization	16228
MCH and Crippled Children's	27
Medicaid	10568
Medicare	3708
No Fault Insurance	44
Other	22839
Other Commercial Indemnity Plan	7866
Private Charity	63
self-pay	37501
Worker's Compensation	6683

Table 4.1: **Number of cases belonging to each payment source** : The payment source types are taken directly from the NTDB dataset.

The logistic regression model we use is:

$$\begin{aligned} ln(\frac{\pi}{1-\pi}) = \beta_0 X_0 + \beta_{\text{age}} X_{\text{age}} + \beta_{\text{race}} X_{\text{race}} + \beta_{\text{gender}} X_{\text{gender}} + \\ \beta_{\text{bedsize}} X_{\text{bedsize}} + \beta_{\text{region}} X_{\text{region}} + \\ \beta_{\text{injury-type}} X_{\text{injury-type}} + \beta_{\text{time-of-admit}} X_{\text{time-of-admit}} + \\ \beta_{\text{payment-source}} X_{\text{payment-source}} \end{aligned} \tag{4.1}$$

where the last line is the variable of interest (self-pay vs. insured), π is the probability of survival.

4.2.3 Estimation of parameters

The likelihood function for the above model (4.1) is:

$$L(\beta|y) = \prod_{n=1}^{N} \frac{n_i}{y_i!(n_i - y_i)!} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} \tag{4.2}$$

where y_i corresponds to individual groups of patients and n_i is the total number of patients in these groups. After rearranging terms above we have:

$$L(\beta|y) = \prod_{n=1}^{N} (\frac{\pi_i}{1-\pi_i})^{y_i} (1-\pi_i)^{n_i} \tag{4.3}$$

After simplifying the above equation and taking log of both sides we get:

$$l(\beta|y) = \sum_{i=1}^{N} y_i (\sum_{k-1}^{K} X_{ik}\beta_k) - n_i log(1 + exp(\sum_{k=1}^{K} X_{ik}\beta_k)) \tag{4.4}$$

By taking derivatives of both sides of 4.4 and solving for each β separately does not yield a closed form solution. A variation of the Newton-Raphson method of scoring can be used to get estimates for β [18, 60].

4.3 Results

4.3.1 Raw results

We used 120,123 cases in our final analysis, of which 82,622 were insured and 37,501 were self-pay patients. Table 4.1 shows a breakdown of how many patients had each type of insurance. Figure 4.3 shows the distribution of patients by age and insurance type. The proportion of self-pay patients is the highest for 19 year old patients, yet remains essentially constant until approximately 40 years of age. After 40 years, the proportion decreases until age 65.

The overall mortality rate for all patients was 3.69%. The mortality rates varied through different injury types and insurance types. Across all levels of injury severity and forms of payment, the mortality rate was the highest in self-pay patients (4.3). As ISS increases, mortality rate for both groups dramatically increases. However, this increase is disproportionately greater in self-pay patients. For instance, patients with severe (ISS 16-25) and critical injuries (ISS>25) and who were self-pay, had twice the mortality (12% and 42%, respectively) of insured patients with similar injuries (6% and 22%). These differences are statistically significant, with p-values < 0.05.

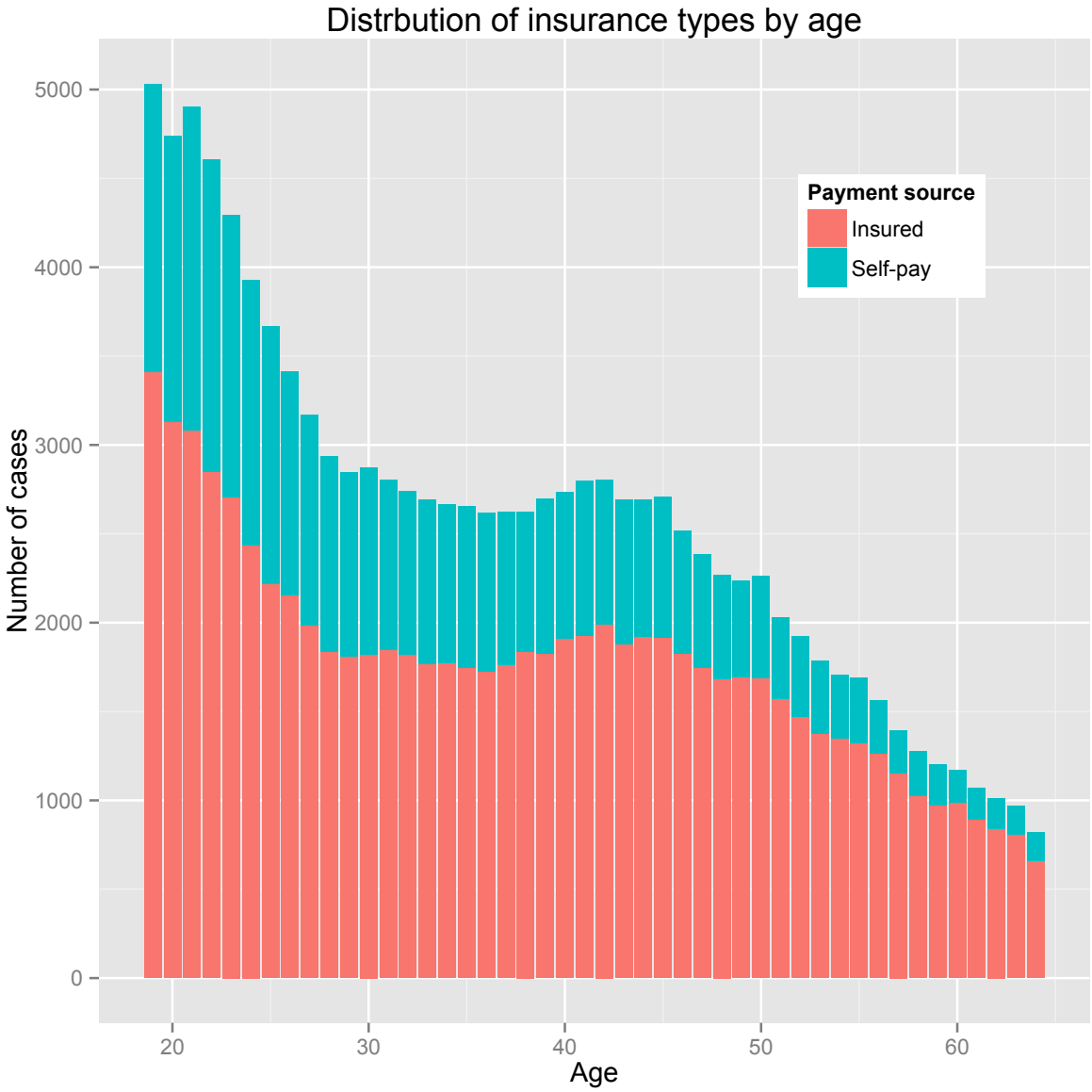

Figure 4.2: **Distribution of cases by insurance form and age.** This figure shows the distribution of patients by their age and state of insurance. The proportion of self-pay patients is the highest at age 19 and decreases with older age.

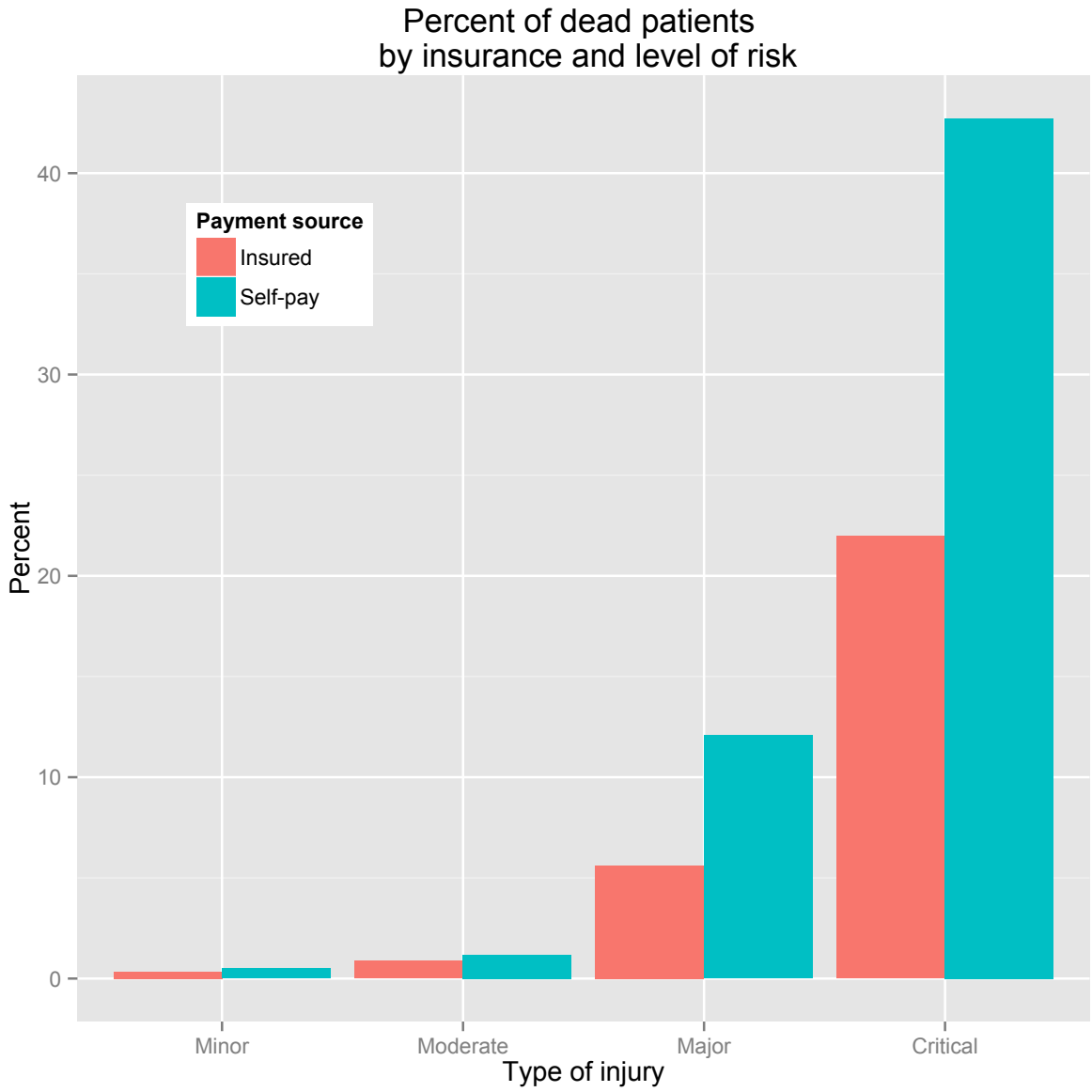

Figure 4.3: **Percentage of dead patients by insurance type and injury risk.** This figure shows the mortality rates by insurance type and level of risk. Across all levels of risk the mortality rate of self-pay patient is higher than insured patients. For the higher risk patients (Major, critical) there is a 100% increase in mortality rate from insured patients to self-pay.

Figure 4.4 shows the distribution of patients admitted to trauma centers by insurance status and time of day. For all self-pay patients, 57% were admitted to trauma facilities between 6pm and 6 am. By comparison, only 47% of insured patients were admitted during the same time period (p-value<.005).

Figure 4.5 shows the proportion of patients with penetrating trauma for each insurance group. Self-pay patients have a higher proportion of penetrating trauma incidents than insured patients across all levels of severity. Thus, we have a number of interesting differences between self-pay and insured patients. Next, we examine the effect of these differences on survival outcome more formally within the logistic regression framework.

4.3.2 Estimated results from logistic regression

The survival probability was the lowest in self-pay patients across all levels of severity and both facility levels (Table 4.2), which explains why the effect of insurance status is significant across all regressions, except level I patients with minor or moderate (ISS: 0-15) injuries (Figure 4.6). Moreover, the disparity in survival rate between self-pay patients and insured patients becomes wider as the severity of injury increases. For instance, for patients with major injuries (ISS 16-25), self-pay patients in type I facility have a survival probability of 0.5527, while insured patients in the same facility type have a survival probability of 0.5811. This amounts to a 4.88% drop in survival probability (p-value<.05). For level II trauma centers, the effect is even more pronounced: the survival probabilities are 0.5914 and 0.7106 for

self-pay and insured patients, respectively, or a 16.77% decrease in the survival of self-pay patients (p-value<.05). The difference in estimated survival probability is greatest for those patients with critical injuries (ISS>25). In level I trauma centers, the survival probability of self-pay and insured patients is 0.1958 and 0.2689, or a 27.18% decrease in survival for being self-pay (p-value<.05). The decrease for type II facilities from insured to self-pay patients is 26.74% (p-value<.05).

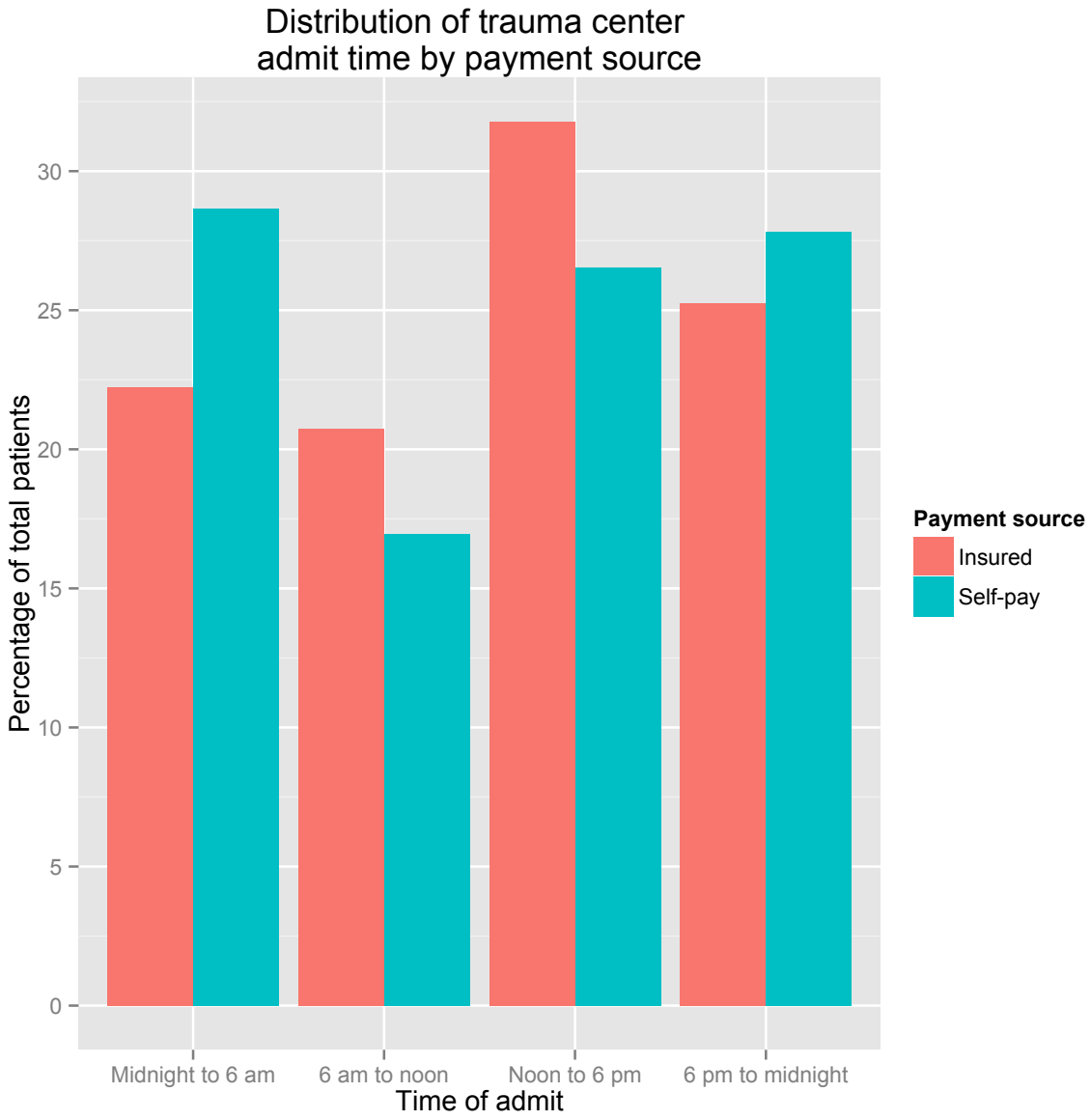

Figure 4.4: **Time on admit by payment source.** This figure shows the hourly breakdown of patients within the two groups of payment source. There are approximately 56% of patients without insurance getting admitted into trauma centers from 6 pm to 6 am. The same time slot account for 47% of insured patients.

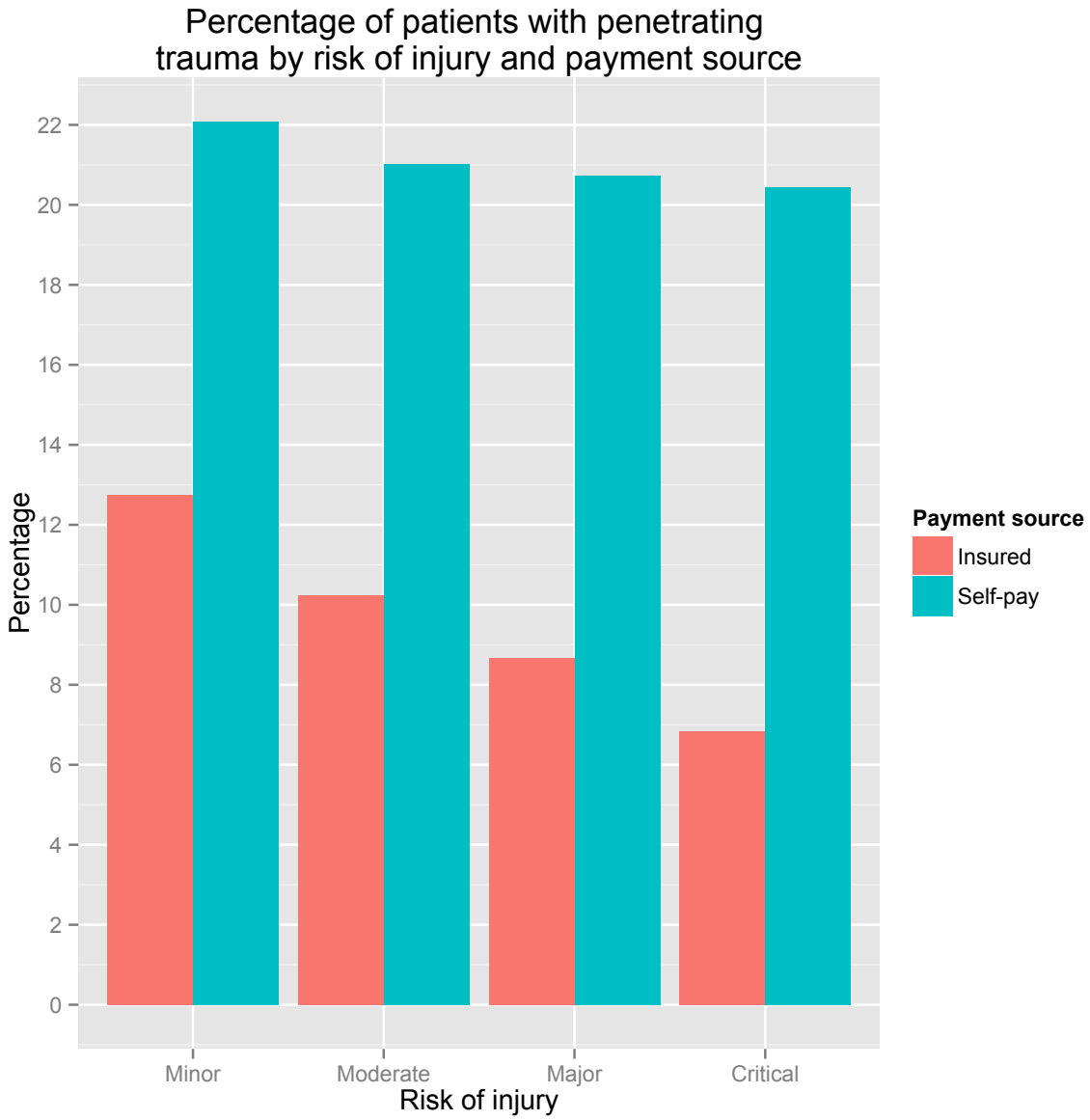

Figure 4.5: **Proportion of patients with penetrating trauma by payment source.** This figure shows the proportion of patients with penetrating trauma depending on payment source and risk of injury. The percentage of patients with a penetrating trauma is much higher in self-pay patients than insured patients across all age levels.

	Facility Level I		Facility Level II	
	Insured	self-pay	Insured	self-pay
Minor (ISS: 0-8)	0.9950	0.9942	0.9706	0.9530
Moderate (ISS:9-15)	0.9890	0.9849	0.9797	0.9660
Major (ISS: 16-25)	0.5811	0.5527	0.7106	0.5914
Critical (ISS: 26-75)	0.2689	0.1958	0.5458	0.3998

Table 4.2: **Estimated survival probabilities from logistic regression** The table shows the estimated survival probabilities from logistic regression averaged over other factors for facility level and injury risks.

The disparity in estimated survival probability between injury types is also shown in Table 4.3, grouped by mechanism of injury. Penetrating trauma has a lower chance of survival in comparison to blunt trauma, which helps explain why the regression analysis shows the risk factor is higher in penetrating trauma than for blunt trauma, even if the ISS and other control variables are the same (p-value$<$.05).

Table 4.4 shows the percentage change from insured to self-pay patients in survival probability within different racial groups. The greatest racial disparity between self-pay patients and insured patients occurred in non-white male patients, a drop in survival probability of 26.91%. The largest gender disparity overall occurred in males (p-value$<$.05 for level II facility and Major injuries).

Tables 4.5 and 4.6 shows the estimated coefficients and p-value for all logistic regressions performed on major and critically injured patients (Level I and II, ISS of 26-75) are consistent with the findings above. The effect of self-pay is statistically significant (p-value$<$.05) and among the largest in magnitude coefficients. As expected, race, sex, age, ISS and some regions (Northeast and South) also have significant coefficients (p-value$<$.05). Across all regression coefficients, the self-pay variable had a significant impact on the logistic regression.

	Penetrating trauma	Blunt Trauma
Minor (ISS: 0-8)	0.9953	0.9931
Moderate (ISS:9-15)	0.9456	0.9729
Major (ISS: 16-25)	0.5634	0.8355
Critical (ISS: 26-75)	0.2285	0.4185

Table 4.3: **Estimated survival probabilities from logistic regression by injury type.** The table shows the estimated survival probability from logistic regression by facility level and type of injury.

We validated the estimated logistic regressions by predicting survival probability on a held-out test sample consisting of only major and critical patients, since lower risk patients would result in artificially high scores due to low mortality rates. Figure 4.7 shows the accuracy of the models with different sets of independent variables for predicting survival outcomes using an ROC curve. Model 1 in the figure represents only individual demographics in the model, Model 2 additionally includes facility level characteristics, and Model 3 is the full model that includes injury type, time of admittance, insurance status, and all variables in Model 2. For both levels, we see that the models predictive performance on the held-out test set improves as more variables are included. For facility levels I and II, the full logistic regressions have area under the curve (AUC) scores of 0.79 and 0.80 respectively, indicating accurate predictive models.

Another choice to validating the model is to calculate the pseudo R-squared values. Table 4.7 shows the pseudo R-squared values for all regression types. It varies from .05-.21.

	White	Non white
Male	-24.63	-26.91
Female	-23.82	-25.61

Table 4.4: **Estimated survival probability disparity between payment source by race and gender** This table shows the percentage drop, in regards to survival probability estimated from logistic regression, from insured to self-pay by gender (male, female) and race (white, non white). The biggest drop in survival probability happens for non white males at approximately 27%. The data are restricted to facility level I and ISS of 26-75.

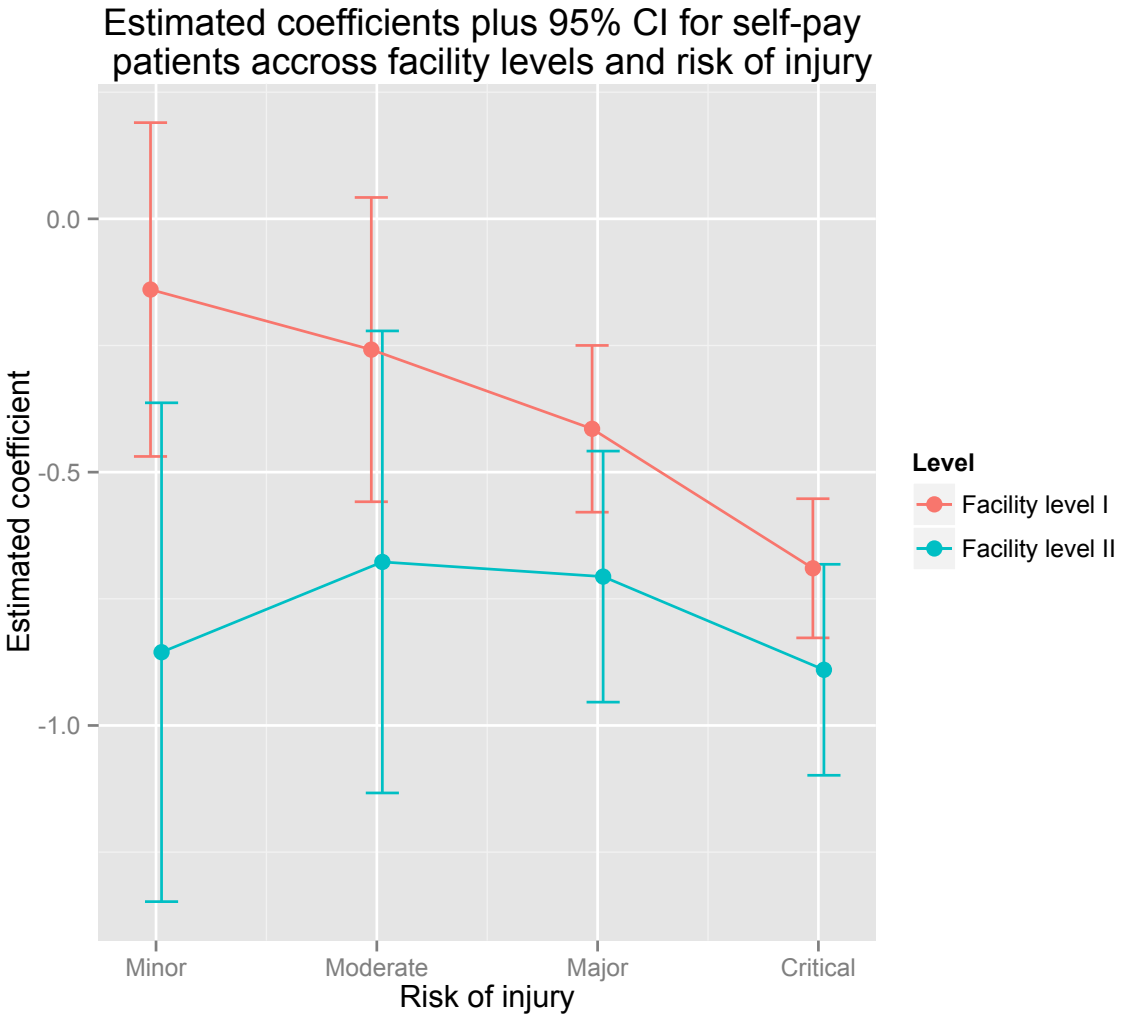

Figure 4.6: **Estimated coefficients plus 95% CI for self-pay patients across facility levels and risk of injury** This figure shows the estimated coefficients for self-pay patients from all the regression models. For minor and moderate injuries in facility level I the self-pay coefficient is not significant. For all other regression models this coefficient was significant. Level I facility have a higher coefficient for self-pay patient but the difference is not significant at α=.05.

Facility Level I		ISS: 15-25			ISS: >25	
Coefficient	Estimate	Z-value	P-value	Estimate	Z-value	P-value
Intercept	4.02	11.39	0.00	1.67	5.00	0.00
AGE	-0.01	-3.74	0.00	-0.01	-3.89	0.00
White	0.04	0.52	0.60	0.22	3.34	0.00
Male	-0.11	-1.14	0.26	-0.10	-1.39	0.17
Bedsize (>600)	-0.51	-1.72	0.09	-0.10	-0.32	0.75
Bedsize (201-400)	-0.41	-1.29	0.20	-0.13	-0.40	0.69
Bedsize (401-600)	-0.02	-0.07	0.94	0.36	1.17	0.24
Region NE	0.69	3.73	0.00	0.59	3.94	0.00
Region S	0.06	0.54	0.59	-0.02	-0.18	0.86
Region W	0.22	1.87	0.06	-0.28	-2.85	0.00
Injury (Penetrating)	-1.95	-22.68	0.00	-1.13	-12.44	0.00
Hour (6-12)	-0.25	-2.12	0.03	-0.20	-2.10	0.04
Hour (12-18)	-0.27	-2.51	0.01	-0.15	-1.73	0.08
Hour (18-24)	-0.18	-1.74	0.08	-0.09	-1.02	0.31
Self-pay	-0.43	-5.32	0.00	-0.70	-10.38	0.00

Table 4.5: **Coefficients for Level I regression models** Includes estimates of coefficients, estimated z-values and p-values.

Facility Level II		ISS: 15-25			ISS: >25	
Coefficient	Estimate	Z-value	P-value	Estimate	Z-value	P-value
Intercept	4.48	12.21	0.00	2.11	8.13	0.00
AGE	-0.02	-3.65	0.00	-0.02	-4.77	0.00
White	-0.09	-0.75	0.45	0.01	0.14	0.89
Male	-0.27	-1.88	0.06	0.08	0.74	0.46
Bedsize (>600)	-1.24	-5.29	0.00	-0.63	-3.79	0.00
Bedsize (201-400)	-0.64	-2.46	0.01	-0.11	-0.63	0.53
Bedsize (401-600)	-0.30	-1.16	0.25	0.13	0.74	0.46
Region NE	-0.98	-4.12	0.00	-0.88	-3.47	0.00
Region S	0.57	2.84	0.01	0.36	2.43	0.02
Region W	0.53	2.70	0.01	-0.03	-0.17	0.86
Injury (Penetrating)	-2.09	-15.78	0.00	-1.24	-8.11	0.00
Hour (6-12)	-0.03	-0.17	0.86	0.29	2.09	0.04
Hour (12-18)	0.09	0.56	0.58	-0.11	-0.91	0.36
Hour (18-24)	0.18	1.15	0.25	-0.18	-1.45	0.15
Self-pay	-0.71	-5.84	0.00	-1.03	-10.36	0.00

Table 4.6: **Coefficients for Level II regression models** Includes estimates of coefficients, estimated z-values and p-values.

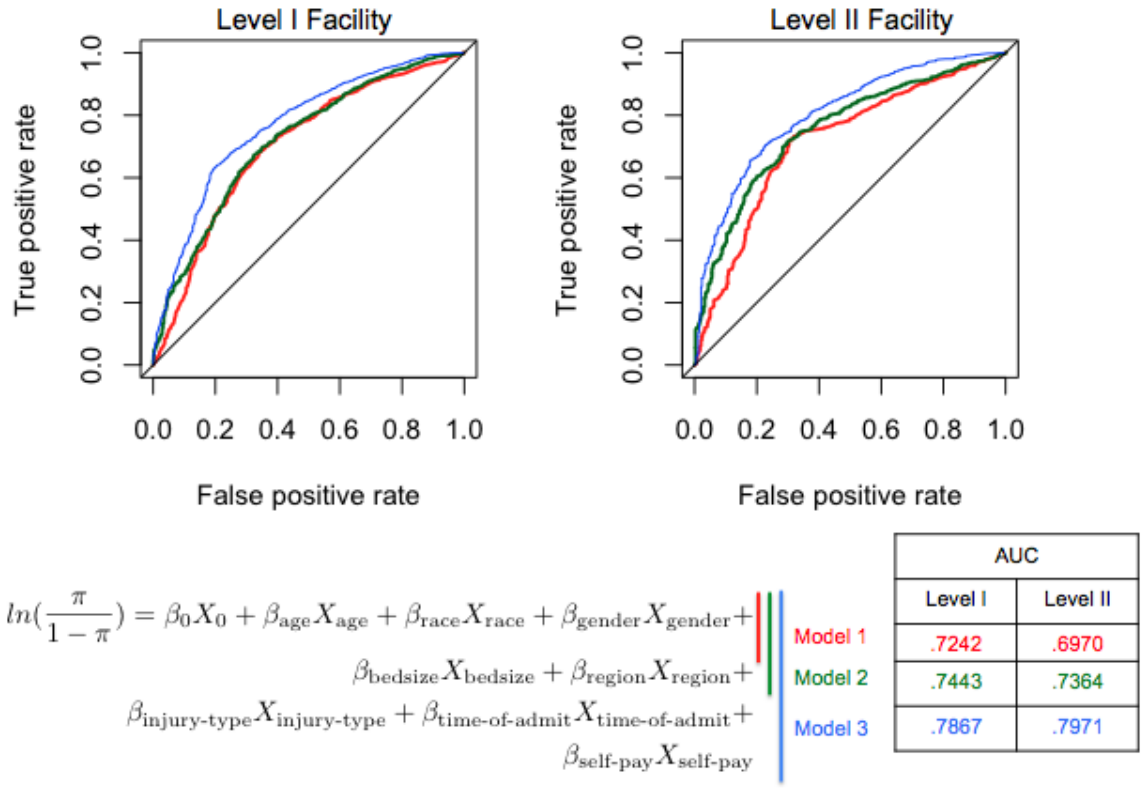

Figure 4.7: **ROC for facility level I and II regression models** The ROC is used to validate our model in predicting mortality outcomes for patients. The corresponding AUC for the curves above are .78 (Left, level I) and .79 (Right, level II) for the full model.

	Facility Level I	Facility Level II
Minor	0.0456	0.0855
Moderate	0.0414	0.1081
Major	0.1939	0.2136
Critical	0.1505	0.1439

Table 4.7: **Pseudo R-squared values for each regression model** This table shows the pseudo R-squared values for each regression model done.

4.4 Discussion

This study identifies a disparity in survival outcomes between insured and self-pay patients in trauma centers across the United States. This disparity increases with injury severity, a pattern that is consistently found across all races and trauma center levels. We also identified differences in arrival time and injury types (penetrating and blunt traumas) between insured and self-pay patients that may help explain the disparity in survival outcome. Previous work [77] has shown that time of arrival at trauma centers is directly correlated with survival probability of patients, since the number of resources available during late nights are much less than at peak hours of the day. We find that self-pay patients are more likely to arrive during late nights, thus lowering their probability of survival.

Similarly, the percentage of self-pay patients with penetrating trauma is significantly greater, which we find lowers their survival chances even if their ISS score is held fixed. This is a potentially important finding, since many ISS scores do not distinguish between penetrating and blunt trauma. We also found that type II trauma facilities show a bigger difference between self-pay patients survival probability and insured patients survival probability. Overall, cases in type II facilities had lower survival probability than cases in type I facilities, which may be expected since type I facilities typically have more resources and staffing.

We caution that there are also limitations to our findings due to the data. The NTDB does not contain unique patient identifiers. As such, it is likely impossible to track the movement of individual patients through the trauma center system.

However, given that the database contains large numbers of records drawn from every region of the US, it can be considered a representative sample of trauma care in the country. The NTDB data set is also not exhaustive, in that there are likely many missing socioeconomic factors that play a significant role in survival outcome and hence, confound our analysis [54]. Nonetheless, we controlled for a number of variables, such as time of arrival, injury types, race, and insurance status, which are proxies for socioeconomic status.

4.5 Conclusion

Using the NTDB V 7.2, we conclude self-pay patients have a lower probability of survival than insured patients across all facility levels. Two key, statistically significant factors that differ between the two groups of patients are identified, namely arrival time and injury type (penetrating or blunt traumas). The difference in survival outcome between self-pay and insured patients is not only statistically significant, but also large enough in magnitude to be practically meaningful, highlighting insurance status as an important topic of discussion in public policy and healthcare.

4.6 Software

All analyses were done using R 3.0.2 using the `glm` (generalized linear model) function and predicted values from our model are reported with significance. We also use a receiver operator characteristic (ROC) curve and the area under the curve (AUC) statistic [57] to show the accuracy of our model in predicting survival

outcome. We divided the data into a testing set (30% of data) and a training set (70% of data) to produce the ROC curves. The NTDB data contained a large number of records with missing data. Of the 1,926,245 unique incidents reported in the data, 120,123 complete records with no missing data in the fields of interest were kept for our analyses.

Chapter 5: Conclusion

For my PhD research, I have developed and applied multiple parametric and semi-parametric models to analyze longitudinal data. The tools I have developed can help solve problems from a variety of different fields.

For biological data, I developed a method using SSANOVA that can find regions of interest in biological data. Regions of interest here is defined by locations where the difference between two groups in some measurement is significant. For this project I also developed an R function, `fitTimeSeries`, which is currently part of `metagenomeSeq` package. This chapter was submitted as a paper recently and is currently under review [44]. I also used `fitTimeSeries` for another project analyzing longitudinal data of lung microbiota in monkeys over 15 months of SHIV infection [47].

For the second project, I developed a statistical method to analyze degree distribution in longitudinal network data. Using writers of hit singles, I build networks which was then analyzed using an overdispersed Poisson model. I developed a visualization tool using `shiny` package in R that illustrates the network data as well as other statistical graphs involved with the data. This chapter as a paper is currently under preparation and will be submitted soon [46].

For the final project, I used a partitioning technique to analyze longitudinal data in the field of healthcare. After partitioning the data, a logistic regression was used to analyze longitudinal data. I showed the disparity that exist in healthcare between insured and uninsured patients. Using different statistical methods I showed the accuracy of the regression model in predicting mortality. This chapter will submitted be as a paper [45].

Bibliography

[1] Johnson, D., Mortazavi, A., Myers, R., Wold, B.: Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science (June), 1497–1502 (2007)

[2] Buck, M.J., Lieb, J.D.: Chip-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics **83**(3), 349–360 (2004)

[3] Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., Crawford, G.E.: High-resolution mapping and characterization of open chromatin across the genome. Cell **132**(2), 311–22 (2008). doi:10.1016/j.cell.2007.12.014

[4] Laird, P.W.: Principles and challenges of genome-wide dna methylation analysis. Nature Reviews Genetics **11**(3), 191–203 (2010)

[5] Irizarry, R.A., Ladd-Acosta, C., Carvalho, B., Wu, H., Brandenburg, S.A., Jeddeloh, J.A., Wen, B., Feinberg, A.P.: Comprehensive high-throughput arrays for relative methylation (CHARM). Genome Research **18**(5), 780–790 (2008)

[6] Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., Fan, J.-B., Shen, R.: High density DNA methylation array with single CpG site resolution. Genomics **98**(4), 288–295 (2011)

[7] Lister, R., Pelizzola, M., Dowen, R., Hawkins, R., Hon, G., Tonti-Filippini, J., Nery, J., Lee, L., Ye, Z., Ngo, Q., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A., Thomson, J., Ren, B., Ecker, J.: Human DNA methylomes at base resolution show widespread epigenomic differences. Nature (2009)

[8] Meissner, A., Gnirke, A., Bell, G.W., Ramsahoye, B., Lander, E.S., Jaenisch, R.: Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis. Nucleic acids research **33**(18), 5868–5877 (2005)

[9] Feinberg, A.P., Tycko, B.: The history of cancer epigenetics. Nature Reviews Cancer **4**(2), 143–53 (2004). doi:10.1038/nrc1279

[10] Irizarry, R.A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., Ji, H., Potash, J.B., Sabunciyan, S., Feinberg, A.P.: The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nature Genetics **41**(2), 178–186 (2009)

[11] Hansen, K.D., Timp, W., Bravo, H.C., Sabunciyan, S., Langmead, B., McDonald, O.G., Wen, B., Wu, H., Liu, Y., Diep, D., Briem, E., Zhang, K., Irizarry, R.A., Feinberg, A.P.: Increased methylation variation in epigenetic domains across cancer types. Nature Genetics **43**(8), 768–775 (2011)

[12] Jaffe, A.E., Murakami, P., Lee, H., Leek, J.T., Fallin, M.D., Feinberg, A.P., Irizarry, R.A.: Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. International journal of epidemiology **41**(1), 200–209 (2012)

[13] Hansen, K.D., Langmead, B., Irizarry, R.A.: BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biology **13**(10), 83 (2012)

[14] Aryee, M.J., Jaffee, A.E., Corrada Bravo, H.J., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., Irizarry, R.A.: Minfi: A Flexible and Comprehensive Bioconductor Package for the Analysis of Infinium DNA Methylation Microarrays. (2014). doi:10.1093/bioinformatics/btu049

[15] Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., Irizarry, R.A.: Tackling the widespread and critical impact of batch effects in high-throughput data. Nature reviews Genetics **11**(10), 733–739 (2010)

[16] Cleveland, W.S.: Robust locally weighted regression and smoothing scatterplots. Journal of the American statistical association **74**(368), 829–836 (1979)

[17] Wahba, G.: Spline Models in Statistics. CBMS-NSF Regional Conference Series. London England Chapman and Hall 1983. Philadelphia, PA (1990)

[18] McCullagh, P., Nelder, J.: Generalized Linear Models. Chapman and Hall 1983., London, England (1989)

[19] Bravo, H.C.: Graph-based Data Analysis: Tree-structured Covariance Estimation, Prediction by Regularized Kernel Estimation and Aggregate Database Query Processing for Probabilistic Inference. ProQuest, (2008)

[20] Harezlak, J., Naumova, E., Laird, N.M.: Longcrisp: A test for bumphunting in longitudinal data. Statistics in Medicine, 1383–1397 (2007)

[21] Wahba, G., Wang, Y., Gu, C., Klein, R., Klein, B.: Smoothing spline anova for exponential families, with application to the wisconsin epidemiological study of diabetic retinopathy: the 1994 neyman memorial lecture. The Annals of Statistics **23**(6), 1865–1895 (1995)

[22] Gu, C.: Smoothing Spline Anova Model. Springer Series in Statistics. Springer, (2002)

[23] Cancer Genome Atlas Network: Comprehensive molecular characterization of human colon and rectal cancer. Nature **487**(7407), 330–337 (2012)

[24] Turnbaugh, P.J., Ridaura, V.K., Faith, J.J., Rey, F.E., Knight, R., Gordon, J.I.: The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. Science translational medicine **1**(6), 6–14 (2009)

[25] Bravo, H.C., Lee, K., Klein, B., Klein, R., Iyengar, S., Wahba, G.: Examining the relative influence of familial, genetic, and environmental covariate information in flexible risk models. PNAS, 8128–8133 (2009)

[26] Kimeldorf, G.S., Wahba, G.: A correspondence between bayesian estimation on stochastic processes and smoothing by splines. The Annals of Mathematical Statistics **41**(2), 495–502 (1970)

[27] Wang, Y.: Smoothing Splines: Method and Applications. CRC Press, (2011)

[28] Paulson, J.N., Stine, O.C., Bravo, H.C., Pop, M.: Differential abundance analysis for microbial marker-gene surveys. Nature methods (2013)

[29] R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2013). R Foundation for Statistical Computing. http://www.R-project.org/

[30] Irizarry, R.A., Aryee, M., Bravo, H.C., Hansen, K.D., Jaffee, H.A.: Bumphunter: Bump Hunter. (2013). R package version 0.99.34

[31] Hebestreit, K., Klein, H.-U.: BiSeq: Processing and Analyzing Bisulfite Sequencing Data. (2013). R package version 1.0.3

[32] Frazee, A., Collado-Torres, L., Leek, J.: Derfinder: Differential Expression Analysis of RNA-seq Data at Base-pair Resolution. (2013). R package version 1.0.2. `https://github.com/alyssafrazee/derfinder`

[33] Paulson, J.N., Talukder, H., Pop, M., Bravo, H.C.: metagenomeSeq: Statistical Analysis for Sparse High-throughput Sequncing. (2013). R package version 1.7-18. `http://cbcb.umd.edu/software/metagenomeSeq`

[34] Gu, C.: Gss: General Smoothing Splines. (2013). R package version 2.1-0. `http://CRAN.R-project.org/package=gss`

[35] Borchers, H.W.: Pracma: Practical Numerical Math Functions. (2014). R package version 1.6.1. `http://CRAN.R-project.org/package=pracma`

[36] Hebestreit, K., Dugas, M., Klein, H.-U.: Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. Bioinformatics **29**(13), 1647–1653 (2013)

[37] Schoofs, T., Rohde, C., Hebestreit, K., Klein, H.-U., Göllner, S., Schulze, I., Lerdrup, M., Dietrich, N., Agrawal-Singh, S., Witten, A., *et al.*: Dna methylation changes are a late event in acute promyelocytic leukemia and coincide with loss of transcription factor binding. Blood **121**(1), 178–187 (2013)

[38] Frazee, A.C., Sabunciyan, S., Hansen, K.D., Irizarry, R.A., Leek, J.T.: Differential expression analysis of RNA-seq data at single-base resolution. Biostatistics, 053 (2014)

[39] Handelsman, J., Tiedje, J., Alvarez-Cohen, L., Ashburner, M., Cann, I., Delong, E., Doolittle, W., Fraser-Liggett, C., Godzik, A., Gordon, J., *et al.*: The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet, (2007)

[40] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of molecular biology **215**(3), 403–410 (1990)

[41] Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R.: Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. Applied and Environmental Microbiology **73**(16), 5261–5267 (2007)

[42] Harezlak, J., Heckman, N.: Crisp: A tool for bump hunting. Journal of Computational and Graphical Statistics, 713–729 (2001)

[43] Friedman, J., Fisher, N.: Bump hunting in high-dimensional data. Statistics and Computing **9**, 123–143 (1998)

[44] Talukder, H., Paulson, J.N., Corrada Bravo, H.: Finding regions of interest in high throughput genomic data using smoothing splines. Submitted.

[45] Talukder, H., Corrada Bravo, H., Dezman, Z., Golden, B., Mankad, S.: Does health insurance matter? Establishing insurance states as a risk factor for mortality rate. In preparation.

[46] Talukder, H., Corrada Bravo, H.: Longitudinal network analysis shows the decline of pop music in the 21st century. In preparation.

[47] Elodie, G., Paulson, J.N., Talukder, H., Pop, M., et al., : Longitudinal analysis of the lung microbiota of cynomolgous monkeys over 15 months of SHIV infection. In preparation.

[48] Rodrigo F Alban, Cherisse Berry, Eric Ley, James Mirocha, Daniel R Margulies, Areti Tillou, and Ali Salim. Does health care insurance affect outcomes after traumatic brain injury? analysis of the national trauma databank. *The American Surgeon*, 76(10):1108–1111, (2010)

[49] Daniel Andersen, Gabriel Ryb, Patricia Dischinger, Joseph Kufera, and Kathleen Read. Self-reported health indicators in the year following a motor vehicle crash: a comparison of younger versus older subjects. In *Annals of Advances in Automotive Medicine/Annual Scientific Conference*, volume 54, page 359. Association for the Advancement of Automotive Medicine, (2010)

[50] Gerard F Anderson. From soak the richto soak the poor: recent trends in hospital pricing. *Health Affairs*, 26(3):780–789, (2007)

[51] Patricia C Dischinger, Kimberly A Mitchell, Joseph A Kufera, Carl A Soderstrom, and Albert B Lowenfels. A longitudinal study of former trauma center patients: the association between toxicology status and subsequent injury mortality. *Journal of Trauma-Injury, Infection, and Critical Care*, 51(5):877–886, (2001)

[52] Ross J Fleischman, Annette L Adams, Jerris R Hedges, O John Ma, Richard J Mullins, and Craig D Newgard. The optimum follow-up period for assessing mortality outcomes in injured older adults. *Journal of the American Geriatrics Society*, 58(10):1843–1849, (2010)

[53] Hassan Haghparast-Bidgoli, Soheil Saadat, Lennart Bogg, Mohammad Hossein Yarmohammadian, and Marie Hasselberg. Factors affecting hospital length of stay and hospital charges associated with road traffic-related injuries in iran. *BMC health services research*, 13(1):281, (2013)

[54] Adil H Haider, David C Chang, David T Efron, Elliott R Haut, Marie Crandall, and Edward E Cornwell. Race and insurance status as risk factors for trauma mortality. *Archives of Surgery*, 143(10):945–949, (2008)

[55] Adil H Haider, Zain G Hashmi, Syed Nabeel Zafar, Renan Castillo, Elliott R Haut, Eric B Schneider, Edward E Cornwell III, Ellen J Mackenzie, and David T Efron. Developing best practices to study trauma outcomes in large databases: An evidence-based approach to determine the best mortality risk adjustment model. *Journal of Trauma and Acute Care Surgery*, 76(4):1061–1069, (2014)

[56] Wael Hakmeh, Jarrod Barker, Susan M Szpunar, James M Fox, and Charlene B Irvin. Effect of race and insurance on outcome of pediatric trauma. *Academic emergency medicine*, 17(8):809–812, (2010)

[57] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. *The elements of statistical learning*, volume 2. Springer, (2009)

[58] Daithi S Heffernan, Roberto M Vera, Sean F Monaghan, Rajan K Thakkar, Matthew S Kozloff, Michael D Connolly, Shea C Gregg, Jason T Machan, David T Harrington, Charles A Adams Jr, et al. Impact of socioethnic factors on outcomes following traumatic brain injury. *Journal of Trauma and Acute Care Surgery*, 70(3):527–534, (2011)

[59] Gustavo Recinos, Joseph J DuBose, Pedro GR Teixeira, Galinos Barmparas, Kenji Inaba, David Plurad, DJ Green, Demetrios Demetriades, and Howard Belzberg. ACS trauma centre designation and outcomes of post-traumatic ards: NTDB analysis and implications for trauma quality improvement. *Injury*, 40(8):856–859, (2009)

[60] SJ Wright and J Nocedal. *Numerical optimization*, volume 2. Springer New York, (1999)

[61] Lewis, Kevin, Marco Gonzalez, and Jason Kaufman. Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences* 109.1, : 68-72, (2012)

[62] Batagelj, Vladimir, and Andrej Mrvar. Pajekanalysis and visualization of large networks. Springer Berlin Heidelberg, (2004)

[63] Christakis, Nicholas A., and James H. Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine* 357(4), 370-379, (2007)

[64] Kaiser, Marcus. Mean clustering coefficients: the role of isolated nodes and leafs on clustering measures for small-world networks. *New Journal of Physics* 10(8) (2008)

[65] Muchnik, Lev, Sinan Aral, and Sean J. Taylor. Social influence bias: A randomized experiment. *Science* 341.6146, 647-651, (2013)

[66] Saramki, Jari, et al. Persistence of social signatures in human communication. *Proceedings of the National Academy of Sciences* 111(3), 942-947, (2014)

[67] Zheng, Tian, Matthew J. Salganik, and Andrew Gelman. How many people do you know in prison? Using overdispersion in count data to estimate social structure in networks. *Journal of the American Statistical Association* 101(474), 409-423, (2006)

[68] Filippova, Darya, et al. Dynamic exploration of recording sessions between jazz musicians over time. *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom).* IEEE, (2012)

[69] Spotify. How is Spotify contributing to the music business. `http://www.spotifyartists.com/spotify-explained/`. (2013)

[70] CNN. I-tunes music decline `http://money.cnn.com/interactive/technology/itunes-music-decline/`.

[71] Billboard. `http://www.billboard.com/charts/hot-100`.

[72] International Federation of the Phonographic Industry (IFPI). `http://www.ifpi.org/global-artist-chart.php`.

[73] Spotify chart.
http://www.spotifyartists.com/site/wp-content/uploads/2013/09/First-Chart.png.

[74] James H Fowler and Nicholas A Christakis. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *Bmj*, 337, (2008)

[75] Cesar A Hidalgo and C Rodriguez-Sickert. The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, 387(12):3017–3024, (2008)

[76] Miranda J Lubbers, José Luis Molina, Jürgen Lerner, Ulrik Brandes, Javier Ávila, and Christopher McCarty. Longitudinal analysis of personal networks. the case of argentinean migrants in spain. *Social Networks*, 32(1):91–104, (2010)

[77] David Anderson, et al. Life is all about timing: An examination of differences in treatment quality for trauma patients based on hospital arrival time. *Production and operations management*, Forthcoming.

[78] Ian McCulloh and Kathleen Carley. Longitudinal dynamic network analysis. using the over time viewer feature in ora. Technical report, DTIC Document, (2009)

[79] Junzhou Zhao, John CS Lui, Don Towsley, Xiaohong Guan, and Yadong Zhou. Empirical analysis of the evolution of follower network: A case study on douban. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 924–929. IEEE, (2011)

[80] Hedeker, D.: Longitudinal data analysis. Hon Wiley and Sons (2006)

[81] Liang, K.: Longitudinal data analysis using generalized linear models. Biometrika Trust (1986)

[82] Peter Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott Zeger. *Analysis of longitudinal data*. Oxford University Press, (2002)

[83] Wei-Yin Loh, Wei Zheng, et al. Regression trees for longitudinal and multiresponse data. *The Annals of Applied Statistics*, 7(1):495–522, (2013)

[84] Jimeng Sun, Christos Faloutsos, Spiros Papadimitriou, and Philip S Yu. Graphscope: parameter-free mining of large time-evolving graphs. In *Pro-*

ceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 687–696. ACM, (2007)

[85] Chun-Yuen Teng, Liuling Gong, Avishay Livne Eecs, Celso Brunetti, and Lada Adamic. Coevolution of network structure and content. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 288–297. ACM, (2012)

[86] Duy Q Vu, David Hunter, Padhraic Smyth, and Arthur U Asuncion. Continuous-time regression models for longitudinal networks. In *Advances in Neural Information Processing Systems*, pages 2492–2500, (2011)