Nonparametric Regression Models and Bootstrap Inference

BY

ELLA REVZIN
B.A. (University of Chicago) 2003
M.S. (University of Illinois at Chicago) 2011

THESIS
Submitted as partial fulfillment of the requirements for the degree of Doctor of Philosophy in Mathematics in the Graduate College of the University of Illinois at Chicago, 2014

Chicago, Illinois

Defense Committee:

Jing Wang, Chair and Advisor
Dibyen Majumdar
Jie Yang
Min Yang
Gib Bassett, Finance
This dissertation is lovingly dedicated to my parents, Ida and Viktor Revzin.
ACKNOWLEDGMENT

Foremost, I would like to express my gratitude and deepest appreciation to my advisor, Dr. Jing Wang, for her continual encouragement and patience throughout the last several years. She has been a great advisor and is the reason I first became interested in studying Nonparametric methods. She was instrumental in getting me to travel to China, where we started the Bootstrap Spline work.

Further, I am deeply indebted to Drs. Dibyen Majumdar and Gib Bassett for their generous support and guidance on my research on Quantile Regression for Tumor Models, which appears as the third essay in this Thesis. They are both exemplary scholars and wonderful mentors.

I would also like to thank Dr. Lijian Yang and his wife Ulya Yang for their great encouragement and generous support during my research work in China in 2012 and beyond. Much of the work presented here on Spline regression models would not have been possible without the helpful guidance of Dr. Yang. He has been a great mentor and fount of immense knowledge on Nonparametric statistics and statistical research in general.

In addition, I would like to thank the other two members of my committee, Drs. Jie Yang and Min Yang for their encouragement and insightful comments. It has been an honor and invaluable learning experience to have both as teachers and as advisors on several statistical consulting projects.
ACKNOWLEDGMENT (Continued)

Acknowledgement must also be given to the NSF for including me in their 2012 cohort of East Asian and Pacific Summer Institutes Fellows and funding my research experience in China under NSF grant NSF DMS - 1107107.

Finally, much heartfelt thanks goes out to my family and friends for putting up with my hectic schedule and crazy work hours over the entire course of my graduate study, especially over the last several months.
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
</tr>
<tr>
<td>1.1</td>
<td>Polynomial Spline Regression</td>
</tr>
<tr>
<td>1.2</td>
<td>Confidence Bands for Univariate Regression Model</td>
</tr>
<tr>
<td>1.3</td>
<td>Confidence Bands for Generalized Linear Model Using Free-knot Splines</td>
</tr>
<tr>
<td>1.4</td>
<td>Conditional Quantile Regression Models of Melanoma Tumor Growth Curves</td>
</tr>
<tr>
<td>2</td>
<td>CONFIDENCE BANDS FOR UNIVARIATE REGRESSION MODEL</td>
</tr>
<tr>
<td>2.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>2.2</td>
<td>B-Splines</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Error Decomposition</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Knot Choice</td>
</tr>
<tr>
<td>2.3</td>
<td>The Wild Bootstrap for Regression Confidence Bands</td>
</tr>
<tr>
<td>2.3.0.1</td>
<td>Algorithm</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Consistency of the Bootstrap Estimators</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Knot choice and bias correction</td>
</tr>
<tr>
<td>2.4</td>
<td>Examples</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Simulation Study</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Canadian wage data</td>
</tr>
<tr>
<td>2.5</td>
<td>Conclusion</td>
</tr>
<tr>
<td>2.6</td>
<td>Proofs of Results</td>
</tr>
<tr>
<td>2.7</td>
<td>Conditional Variance of $t_n^*(x)$</td>
</tr>
<tr>
<td>2.8</td>
<td>Conditional Distribution of $t_n^*(x)$</td>
</tr>
<tr>
<td>3</td>
<td>FREE-KNOT SPLINES FOR GENERALIZED LINEAR MODELS</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>3.2</td>
<td>Polynomial Splines for Quasi-likelihood Estimation</td>
</tr>
<tr>
<td>3.3</td>
<td>Free-knot Spline for GLM</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Binary Response Model</td>
</tr>
<tr>
<td>3.4</td>
<td>Bootstrap for Binary Response Model</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Confidence Bands</td>
</tr>
<tr>
<td>3.5</td>
<td>A Simulation Study</td>
</tr>
<tr>
<td>3.6</td>
<td>Real Example: Horseshoe Crab Data</td>
</tr>
<tr>
<td>3.7</td>
<td>Conclusion</td>
</tr>
<tr>
<td>CHAPTER</td>
<td>CONDITIONAL QUANTILE REGRESSION MODELS OF MELANOMA TUMOR GROWTH CURVES</td>
</tr>
<tr>
<td>---------</td>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td>4</td>
<td>4.1 Introduction</td>
</tr>
<tr>
<td></td>
<td>4.2 Penalized Fixed Effects Quantile Regression for Longitudinal Data</td>
</tr>
<tr>
<td></td>
<td>4.2.1 Selection of tuning parameter $\lambda$</td>
</tr>
<tr>
<td></td>
<td>4.3 Modeling Quantile Tumor Growth Curves with Combined Studies Data</td>
</tr>
<tr>
<td></td>
<td>4.3.1 Data</td>
</tr>
<tr>
<td></td>
<td>4.3.2 Tumor Growth Model</td>
</tr>
<tr>
<td></td>
<td>4.3.3 Results</td>
</tr>
<tr>
<td></td>
<td>4.4 Conclusion</td>
</tr>
<tr>
<td></td>
<td>CITED LITERATURE</td>
</tr>
<tr>
<td></td>
<td>VITA</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>COVERAGE PROBABILITIES FROM 500 REPLICATIONS OF THE MODEL IN (16) WITH C=100</td>
</tr>
<tr>
<td>II</td>
<td>COVERAGE PROBABILITIES FROM 500 REPLICATIONS OF THE MODEL IN (16) WITH C=5</td>
</tr>
<tr>
<td>III</td>
<td>COVERAGE PROBABILITIES FROM 500 REPLICATIONS OF AR(1) ERROR MODEL AND C=100</td>
</tr>
<tr>
<td>IV</td>
<td>COVERAGE PROBABILITIES FROM 500 REPLICATIONS OF AR(1) ERROR MODEL AND C=5</td>
</tr>
<tr>
<td>V</td>
<td>COVERAGE RATES OF BOOTSTRAP CONFIDENCE BANDS, FREE VS. FIXED</td>
</tr>
<tr>
<td>VI</td>
<td>MICE COUNTS BY STUDY AND TREATMENT. MEL2, MEL29, AND MEL23 ARE THREE DIFFERENT MELANOMA CELL LINES WITH P53 EXPRESSION.</td>
</tr>
<tr>
<td>VII</td>
<td>ESTIMATED QUANTILE TREATMENT EFFECTS</td>
</tr>
<tr>
<td>VIII</td>
<td>ESTIMATED QUANTILE GROWTH RATES BY TREATMENT</td>
</tr>
<tr>
<td>IX</td>
<td>DOUBLING TIME BY TREATMENT AND QUANTILE</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
</tr>
<tr>
<td>3</td>
<td>38</td>
</tr>
<tr>
<td>4</td>
<td>39</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
</tr>
<tr>
<td>6</td>
<td>43</td>
</tr>
<tr>
<td>7</td>
<td>45</td>
</tr>
<tr>
<td>8</td>
<td>46</td>
</tr>
<tr>
<td>9</td>
<td>85</td>
</tr>
<tr>
<td>10</td>
<td>86</td>
</tr>
<tr>
<td>11</td>
<td>87</td>
</tr>
<tr>
<td>12</td>
<td>88</td>
</tr>
<tr>
<td>13</td>
<td>89</td>
</tr>
<tr>
<td>14</td>
<td>90</td>
</tr>
<tr>
<td>15</td>
<td>91</td>
</tr>
<tr>
<td>16</td>
<td>92</td>
</tr>
<tr>
<td>17</td>
<td>93</td>
</tr>
<tr>
<td>18</td>
<td>94</td>
</tr>
<tr>
<td>19</td>
<td>95</td>
</tr>
<tr>
<td>FIGURE</td>
<td>PAGE</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>20</td>
<td>Uniform Bands for Crabs Link Function,p=2</td>
</tr>
<tr>
<td>21</td>
<td>Linear Model and Uniform Spline Estimates of Satellites Probability .</td>
</tr>
<tr>
<td>22</td>
<td>Observed tumor volumes, Control and P28, 10mg daily groups, MEL2 and MEL 29 cell lines.</td>
</tr>
<tr>
<td>23</td>
<td>Observed tumor volumes, Control and P28, 10mg daily groups, MEL2 23 cell lines.</td>
</tr>
<tr>
<td>24</td>
<td>Observed tumor volumes, Control and P28, 10mg daily groups, All Studies.</td>
</tr>
<tr>
<td>25</td>
<td>Estimated Quantile Growth Curves for Control, P28, 10 mg/kg, and DTIC.</td>
</tr>
</tbody>
</table>
SUMMARY

This dissertation thesis presents three essays on several topics in nonparametric regression and the bootstrap. In Chapter 2, simultaneous confidence bands for univariate conditional mean regression are developed using a bootstrap of polynomial spline regression. Polynomial spline regression is a popular choice for nonparametric regression and can be used to construct tests for parametric model specifications. Simultaneous confidence bands centered around a non-parametric regression estimate are an especially simple way to assess the adequacy of model fit and the variance of estimators over the entire range of observed data. To construct bands, most approaches rely on either an asymptotic or bootstrap method. Although the asymptotic bands are valid approximations as the sample size tends to infinity, they can suffer from poor coverage for small and moderate sample sizes. Alternative bands are constructed from a Wild bootstrap of second order polynomial spline regression estimators. The bootstrap estimates and corresponding bands are shown to be valid for analysis of independent but not necessarily identically distributed data. Further, simulation results show that the bands have decent coverage for moderate sample sizes and are able to attain the nominal level for larger samples.

Chapter 3 discusses a computational study of bootstrap confidence bands based on a free-knot spline regression for the Generalized Linear Model. In free-knot spline regression, the knot locations of the usual polynomial splines enter into the model as additional parameters that need to be estimated from the data. This offers even greater flexibility and the potential to better account for rapid shifts in slope and other important structures in the target function.
SUMMARY (Continued)

However, in ‘freeing’ up the knots, the search for optimal solutions becomes very complicated. The most challenging part of free-knot computation is the existence of a ‘lethargy’ property in the objective function, which results in many local optima with coalescent knot solutions (ones in which two or more knots are the same). To prevent solutions with identical knots, a penalized Quasi-likelihood estimating equation is proposed that relies on both a Jupp transformation of knot locations and an added penalty on solutions with small minimal distances between knots. Focusing on logistic regression for binary outcome data, a parametric bootstrap is used to study the variability of the proposed estimator and to construct confidence bands for the unknown form of the logistic regression link function.

Finally, Chapter 4 illustrates the use of quantile regression to construct tumor growth curves for longitudinal data collected from small sample animal studies. The traditional approach to fitting tumor growth curves has been to estimate conditional mean regression functions which describe the average effect of covariates on growth. However, this method ignores the possibility that tumor growth dynamics are different for different quantiles of the possible distribution of growth patterns. Instead of fitting conditional mean curves, conditional quantile curves are estimated for data from several independent xenograft mouse studies of human Melanoma. In addition to estimating quantile curves, the work also incorporates a novel use of small sample data. In general, study designs with too few animals tend to have lower power to detect a statistically significant difference in tumor volume between treatment groups. To increase power, the different samples are pooled together. To construct the tumor reference growth curves, a longitudinal regression model for multiple quantiles is specified that incorporates individual
subject and study fixed effects. The curves are estimated using Penalized Fixed Effects Quantile Regression (PFEQR), which is a distribution free method that can be used to model multiple quantiles simultaneously (Koenker, 2004). This method extends the original quantile regression model to longitudinal data by incorporating fixed subject specific model parameters that are penalized to prevent overfitting and decrease model dimension. For inference, a stratified subject bootstrap is used wherein the bootstrap samples are constructed with resampling by subjects and not by time.
CHAPTER 1

INTRODUCTION

Regression analysis is one of the main statistical tools for the study of relationships among variables. A simple case is the univariate conditional mean regression model that specifies a relationship between the mean of a single dependent variable $Y$ and an independent variable $X$. Let \( \{Y_i, X_i, i = 1, \ldots, n\} \) be a given set of $n$ independent observations and \( \{\epsilon_i, i = 1, \ldots, n\} \) be random white noise with mean 0 and variance 1. A general form for conditional mean regression model can be written as

$$Y = m(X) + \sigma(X)\epsilon, \quad (1.1)$$

where the conditional mean of $Y$ is $E(Y|X) = m(X)$ and the conditional variance is $Var(Y|X) = \sigma^2(x)$.

Models for the conditional mean function purport to answer the question of how the average response of $Y$ changes with $X$. Important characteristics of the conditional mean function that are of interest to researchers usually include monotonicity, the presence and location of different types of critical points, and the derivative function (Härdle, 1990). Further, good predictive properties are essential as well. In parametric regression, $m(\cdot)$ is assumed to be a specific function of a finite set of parameters that need to be estimated from the data. Often, parametric models are also augmented with a set of assumptions on the distribution of $X$ and the error.
vectors. In classic regression analysis, these types of assumptions are relatively stringent. The design vector $X$ is taken to be fixed and the unknown errors are assumed to be independent Normal random variables, with the variance usually set as $\sigma(x) \equiv \sigma$. The model parameters can then be estimated using Ordinary Least Squares (OLS), Maximum Likelihood (ML), or for nonlinear models via iterative OLS or ML. The ML procedure and its iterative versions can be used with or without the constant variance assumption.

However, these types of model assumptions are often inadequate to capture the true, albeit, unknown relationship between $Y$ and $X$ and of the underlying data generating process (DGP). Many types of methods have been proposed over the years that relax each of the various assumptions of the classic regression model. This dissertation thesis presents three independent studies that look at several applications that use some of these alternatives, including Polynomial Spline Regression, Generalized Linear Models, Quantile Regression, and the Bootstrap.

1.1 Polynomial Spline Regression

In Chapter 2, the bootstrap is applied to polynomial spline regression to construct confidence bands for a univariate conditional mean regression under the assumption of heterogeneous error variance. In general, there are many nonparametric methods available to fit smooth curves to observed data. The most common forms of smoothers include the kernel, local polynomial, nearest neighbor and splines (Härdle, 1990). The first three can be characterized as ‘local’ smoothers, with which one estimates many lower order polynomial regressions in small ‘local’ neighborhoods across the range of $X$ to produce a smoothed estimate. Although these methods work well in practice to fit curves, they do have several drawbacks. First, these fitting proce-
dures require that a individual regression must be estimated at each observation or prediction point, which comes with greater computation time. Second, they cannot be used to produce a regression estimate that is easily represented by a closed form formula. Predictions are always made on the basis of the locally weighted subsets of data.

Spline regression is an alternative idea that relies on basis expansion of the entire design matrix to fit a single global model. Further, spline regression is easily estimated using ordinary least squares regression. In particular, this method approximates the unknown mean function by using piecewise polynomials. The idea stems from the fact that any smooth function can be well approximated by a polynomial of some degree. The catch is that more complex structure requires polynomials of higher order, which increases model dimension and usually coefficient variances. Piecewise polynomials, on the other hand, provide both greater parsimony and good approximations. A spline of order $p > 0$ is a piecewise polynomial function that is composed of polynomial pieces of order $p$ (degree $p - 1$). It is continuous within and across pieces and can have up to $p - 1$ continuous derivatives at the points at which the pieces connect. More formally, let $\tau = \{t_j\}_{j=0}^N$ be a strictly increasing sequence of points on $[a, b] \in \mathbb{R}$, called ‘interior knots’, where

$$a = t_0 < t_1 < \cdots < t_N = b.$$  

Define $G_{p,\tau} = G_{p,\tau}[a, b]$ to be the space of piecewise polynomial functions of order $p$ defined on each $[t_j, t_{j+1})$. That is, if $g \in G_{p,\tau}$ then there is a sequence of $N + 1$ order $p$ polynomials
In polynomial spline regression, our aim is to find the 'best' approximation to \( m(x) \) in \( G_{p,\tau} \):

\[
\min_{g \in G_{p,\tau}} \sum_{i=1}^{n} (Y_i - g(X_i))^2
\]

### 1.2 Confidence Bands for Univariate Regression Model

In specifying a parametric model, the statistical analyst is free to choose from a wide variety of functional forms that can capture the relationship under study. However, one must have statistically valid methods to first decide which models to try and second to test which estimates are good approximations to the 'truth'. To this end, nonparametric approaches such as polynomial splines and the bootstrap, that place very minimal assumptions on \( m(x) \), can be used as a guide for model selection and data exploration. Polynomial spline regression provides an easy way for visualizing the shape of the unknown mean function and combined with the bootstrap is very useful in construction of inference procedures for parametric model specification and accuracy of model fit. In particular, simultaneous confidence bands centered around a nonparametric regression estimate are an especially simple way to assess the adequacy of a model fit and the variance of estimators over the entire range of observed data.

Distribution free methods such as spline regression lend themselves intuitively for use in inference and model selection. Return to the heterogenous variance model in (1.1). Suppose
the range of $X$ is a compact interval $[a, b]$. A pointwise confidence interval for $m(x)$ can be constructed using a nonparametric estimate $\hat{m}(x)$

$$P \left\{ \hat{m}(x) - l_{n, \alpha}(x) \leq m(x) \leq \hat{m}(x) + u_{n, \alpha}(x) \right\} = 1 - \alpha$$  \hfill (1.2)

where $l_{n, \alpha}(x), u_{n, \alpha}(x)$ are the lower and upper critical points from the sampling distribution of $\hat{m}(x)$. Similarly, to construct a uniform confidence band, we need to know the values $\{l^b(x), u^b(x)\}$ s.t.

$$P \left\{ \hat{m}(x) - l^b_{n, \alpha}(x) \leq m(x) \leq \hat{m}(x) + u^b_{n, \alpha}(x) \ \forall x \in [a, b] \right\} = 1 - \alpha. \hfill (1.3)$$

Much work exists on construction of confidence bands, with a majority of methods based on approximations to the distribution of the maximal deviation $\sup_{x \in [a, b]} |\hat{m}(x) - m(x)|$ and/or estimates of tail probabilities of Gaussian suprema. Asymptotic constructions rely on the limiting distribution of $\hat{m}(x)$ to formulate bands. Many nonparametric estimators, including the kernel and spline, are linear estimators and can be written as weighted sums of the observations. Under standard assumptions on the first two or three moments of the error distribution in model (1.1) and on the size of the weights, these types of estimators can be shown to converge pointwise to normal random variables as the sample size goes to infinity. Further, $\hat{m}(x)$ can also be characterized as a partial sum of a random process over the range of $X$. This will then have a Gaussian process limiting distribution. In turn, asymptotic methods use either the pointwise or process limit or both.
A popular asymptotic approach first formulated in (Bickel and Rosenblatt, 1973) relies on using a strong invariance principle for partial sums of random processes and related theory on extremes of random Gaussian processes ((Komlós et al., 1975), (Komlós et al., 1975), (Härdle, 1989), (Eubank and Speckman, 1993), (Claeskens and Keilegom, 2003), (Wang and Yang, 2009)). Another is to build bands from an estimate of the volume of a tube centered around a curve under the assumption of Normal errors ((Hotelling, 1939), (Naiman, 1990), (Johansen and Johnstone, 1990), (Sun and Loader, 1994)). Other methods include interpolation of simultaneous confidence intervals (Hall and Titterington, 1988); as well as Bayesian formulations ((Wang and Wahba, 1995), (Krivobokova et al., 2010)).

Although the asymptotic bands are valid approximations as the sample size tends to infinity, they can suffer from poor coverage for small and moderate sample sizes. According to a study in (Hall, 1991), the convergence rates of suprema of stationary Gaussian process is logarithmically slow. In particular, the maximal deviation for a Kernel regression estimator was found to converge at a rate of \( \log(n)^{-1} \) for iid data. And it was further found that a bootstrap approximation can achieve a much faster convergence rate of \( (nh)^{-1/2} \log(n)^2 \), where \( h \) is the bandwidth parameter of the Kernel function used to fit the regression (see (Fan and Gijbels, 1996) for reference on Kernel regression).

Similarly, the suprema of a polynomial spline estimator would also suffer from slow convergence to its asymptotic limit. For example, (Wang and Yang, 2009) and (Song and Yang, 2009) studied asymptotic bands based on first and second order polynomial splines. In simulations, these bands were found to have much different coverage than the asymptotic nominal level for
samples sizes ranging from 100 to 500 observations. Another drawback of asymptotic bands is that they often require plug-in estimators for the variance of the regression estimator as well as for the distribution of X. This in turn, has the potential to further decrease accuracy for smaller samples.

In Chapter 2, the central research question is how to use the bootstrap to construct confidence bands that may offer a better approximation for smaller samples. Since being first introduced in (Efron, 1979), the bootstrap has become a popular method to estimate the finite sample distribution and other properties of various estimators. The method consists in generating multiple samples with replacement from the original data and using these to calculate a Monte Carlo distribution for the statistic under study. Under appropriate sampling schemes, the resulting bootstrap distribution can provide a valid and often better approximation than the asymptotic alternative (Bickel and Freedman, 1981),(Hall, 1992). Here, a bootstrap is valid if the bootstrap distribution approaches the true distribution asymptotically. And of course, the bootstrap yields a better approximation if the two distributions get closer faster than the rate at which the sampling distribution gets closer to its asymptotic limit.

A systematic study is presented of the asymptotic characteristics and finite sample performance of bands based on a bootstrap of second order polynomial spline regression estimates. To account for the data heterogeneity, the Wild bootstrap resampling scheme is used. The Wild bootstrap (WB) is a simple alternative that accounts for the heterogenous variance structure. Originally proposed by (Wu, 1986), it was shown to be consistent for the simple linear regression model in (Liu, 1988) under the heterogenous variance assumption. This method was
successfully used in (Härdle and Mammen, 1993) to construct hypothesis tests of parametric models using a bootstrap of Kernel regression estimators. It is shown that the Wild bootstrap yields consistent estimates of the variance and maximal deviation of the linear spline estimator. In particular, Theorem 2.3.1(p.32) states that the bootstrap variance estimator converges to the true variance function in the order of \( O_p(n^{-6/5} \sqrt{\log n}) \) uniformly over the range of \( X \). Further, Theorem 2.3.2(p.33) shows that conditional on the observed data, the distribution of the bootstrap estimate for the maximal deviation converges to the true distribution, albeit at a very much slower rate of \( O(n^{-1/5}) \). Using this, confidence bands are constructed that are symmetric but are not uniform, which is very useful in studying the variability of the spline estimator at different points of the range as well as testing for regression fit. For knot selection, generalized cross validation is used. Further, to correct for finite sample bias, the original linear spline estimator is undersmoothed to determine the centering point and regression residuals.

1.3 **Confidence Bands for Generalized Linear Model Using Free-knot Splines**

For continuous responses, estimating the regression function becomes equivalent to smoothing a curve over the entire range of \( X \), which can be multidimensional. For binary, count, and other types of discontinuous responses, the nonparametric form can be incorporated into a Generalized Linear Model (GLM) framework. Under this approach, the response variable is mapped to the real line using an invertible link function that is then estimated nonparametrically. Chapter 3 extends some of the work in Chapter 2 to estimating confidence bands for the linear predictor of the transformed conditional mean in a Generalized Linear Model (GLM). The GLM is a popular generalization of ordinary linear regression to data from non-Normal distributions. GLM was
first introduced in (Nelder and Wedderburn, 1972) to estimate conditional mean models for data whose likelihood function has the exponential family form. Formally, an independent sample \( \{(X_i, Y_i)\}_{i=1,...,n} \) is from a scaled exponential family if the conditional density of each \( Y_i \) given \( X_i = x \) can be written as

\[
f_Y(y) = \exp\left\{ \frac{y \cdot \theta(x) - b(\theta(x))}{\phi} + c(y, \phi) \right\}
\]

where the support is independent of \( \theta \) and \( \phi \) is known. This family is quite large and includes many of the distributions used in practice, including the exponential, normal, Poisson, and multinomial. For any \( Y \) that has a distribution with the exponential family form, the conditional mean is \( \mu(x) = \mathbb{E}(Y|X = x) = b'(\theta(x)) \), and the conditional variance \( \text{Var}(Y|X = x) = \phi b''(\theta(x)) \).

In GLM, the main interest is to estimate a regression model for the conditional mean. This is done using a transformation of \( \mu(\cdot) \) via an invertible ‘link’ function \( g(\cdot) \):

\[
\eta(x) = g(\mu(x)),
\]

where \( g(x) \) is a known and monotonic function but \( \eta(x) \) must be estimated from the data. Usually, \( \eta(x) \) is assumed to have some (linear) parametric form, e.g. \( \eta(x) = \beta^T X \). A general case is to relax the parametric assumptions and model the data structure nonparametrically i.e. \( \eta(\cdot) \in C^{p-1}[a,b] \) for some \( p > 0 \).
The conditional mean function $\mu(x)$ as specified by the model in (Equation 1.4) and (Equation 1.5) can be estimated using maximum likelihood. However, in many practical situations, it is not readily known whether the data follows an exponential family distribution. For example, when the dispersion parameter, $\phi$, is not known, the distribution in (Equation 1.4) is no longer in the exponential family (McCullagh, 1983). When a full likelihood is not available, but the relationship between the mean and variance can still be specified, the conditional mean regression can be estimated using the Quasi-likelihood method ((Wedderburn, 1974), (McCullagh, 1983)). Assume that the conditional variance function of the response is a known function of $\mu$ and can be written as $\text{Var}(Y) = \phi V(\mu(x))$. The quasi-likelihood (QL) function is defined as

\[
Q(\mu, y) = \int_{y}^{\mu} \frac{y - t}{\phi V(t)} dt.
\]

For a given link function $g$ that is monotone, the unknown function $\eta(\cdot)$ can be estimated by maximizing the empirical QL function.

\[
\hat{\eta}(x) = \arg \max_{\eta \in C^{-1}[a,b]} \sum_{i=1}^{n} Q\left(g^{-1}\left(\eta(X_i)\right), Y_i\right).
\]

In turn, the conditional mean estimate is $\hat{\mu}(x) = g^{-1}(\hat{\eta}(x))$. For data from an exponential family, the maximum QL estimates coincide with the maximum log likelihood. More generally, under some weak technical assumptions, the maximum QL estimates in the parametric model for the conditional mean are asymptotically unbiased and have minimum asymptotic variance (McCullagh, 1983).
Similar to the reasons for using a nonparametric model in ordinary regression, the transformed mean in GLM can also be estimated using polynomial regression splines without specifying a specific parametric form. The QL estimating equations for a $p$th order spline estimate for the unknown $\eta(\cdot)$ is given by:

$$\max_{\eta \in G_{p, \tau}} \sum_{i=1}^{n} Q\left(g^{-1}\left(\eta(X_i)\right)\right), Y_i.$$  

The work presented in Chapter 3 departs from the set-up in Chapter 2 by ‘free-ing’ up the knots in the spline regression model and including knot location as additional parameters. As other nonparametric methods, the accuracy and optimality of spline regression estimation relies on careful choice of its tuning parameters. The number of knots and their placement plays a critical role in minimizing estimator bias and variance in a balanced manner. Spline models that incorporate knot locations as additional parameters are called free-knot splines (Jupp, 1978). Properly placed knots can account locally for rapid shifts in slope and other significant changes in the target function. However, the added flexibility comes with a high computational cost and instability of the requisite nonlinear optimization methods.

Much of the published literature in free-knot splines has focused on simple linear or multivariate regression and resulting computational challenges. Further, no work exists on constructing nonparametric confidence bands using free-knot splines. The main contribution of the research presented here is the use of free-knot splines for QL estimation, with particular attention to constructing confidence bands for binary response data. A penalized polynomial
spline QL estimation method is proposed for choosing optimal knot locations and estimating model parameters for the general GLM. A step-wise model selection algorithm is developed to choose an optimal number of knots. The proposed method is applied to estimating a Logistic regression model. Similar to the method in Chapter 2, confidence bands are constructed by bootstrapping the QL estimates for $\eta(x)$. However, a specific parametric bootstrap is specified under the assumed binary response model.

### 1.4 Conditional Quantile Regression Models of Melanoma Tumor Growth Curves

Chapter 4 presents a methodological study of applying quantile regression to construct tumor growth curves from small sample animal studies. Quantile regression (Koenker, 2005) directly models the $\tau^{th}$ population quantile as a function of covariates. Quantile regression offers a more robust method as opposed to the conditional mean approach, which is more influenced by outliers and missing data. Within this framework, very little is assumed about the underlying error distribution; and in fact the estimated quantiles can be used to answer questions in this regard. This is especially important for the specific application to small sample longitudinal studies which is often the case in pre-clinical experiments of cancer treatments. The traditional approach to fitting tumor growth curves has been to estimate conditional mean regression functions which describe the average effect of covariates on growth. However, this method ignores the possibility that tumor growth dynamics are different for different quantiles of the possible distribution of growth patterns.
Instead of fitting conditional mean curves, conditional quantile curves are estimated for data from several independent xenograft mouse studies of human Melanoma. The primary aim is to provide a more complete understanding of how treatment affects the growth of tumors over time for different parts of the population. In particular, the average response of a cancer tumor to treatment gives no information about how well that drug will work against small, slow growing tumors and more importantly whether it will help patients with larger, more aggressive forms of the cancer. Therefore, a set of conditional reference growth curves corresponding to different quantiles of the population under study are fit for different tumor treatments.

In addition to estimating quantile curves, the work also incorporates a novel use of small sample data. Typical individual preclinical cancer drug study designs have very small sample sizes and can have lower power to detect a statistically significant difference in tumor volume between treatment groups. To increase power, the different samples are pooled together. To construct the tumor reference growth curves, a longitudinal regression model for multiple quantiles is specified that incorporates individual subject and study fixed effects. The curves are estimated using Penalized Fixed Effects Quantile Regression (PFEQR), which is a distribution free method that can be used to model multiple quantiles simultaneously (Koenker, 2004). This method extends the original quantile regression model to longitudinal data by incorporating fixed subject specific model parameters that are penalized to prevent overfitting and decrease model dimension. However, relatively little work exists on appropriate inference for PFEQR estimates. Therefore, for inference, a stratified subject bootstrap is used wherein the bootstrap samples are constructed by resampling the subjects and not time.
This research is motivated by a series of small sample studies that investigated the effect of a naturally derived biological peptide, P28, on tumor volumes in mice grafted with human Melanoma cells. As opposed to conventional chemotherapies, P28 is potentially much less toxic, and therefore it is especially of interest to test for its effect on larger tumors. Under the proposed approach, there appears to be a statistically significant quantile treatment effect on tumor volume trajectories and baseline values. In particular, the experimental treatment and a corresponding conventional chemotherapy had different effects on tumor growth by quantile. The conventional treatment, DTIC, tended to inhibit growth for smaller quantiles, while the experimental treatment P28 produced slower rates of growth in the upper quantiles, especially in the 95th quantile.
CHAPTER 2

CONFIDENCE BANDS FOR UNIVARIATE REGRESSION MODEL

2.1 Introduction

Nonparametric regression is a powerful tool for exploratory data analysis and model checking. In nonparametric regression, instead of specifying a specific functional form and estimating a finite set of parameters, $m(\cdot)$ is only assumed to be smooth and is estimated directly. Estimating the regression function becomes equivalent to curve smoothing over the entire range of $X$, which can be multidimensional. Popular methods are flexible and data driven; and given current computational power offer quick and efficient ways to capture nonlinearities, inflection points, and other important features of the unknown mean function. Under proper tuning, they can yield asymptotically unbiased and consistent estimates for $m(x)$, which is impossible to do with a mispecified parametric model. In addition, they are a popular way to impute missing values in the data and are usefull in identifying outliers. Finally, and perhaps even more importantly for the practioner, it is an easy way to visualize the shape of $m(x)$ over the entire range of $X$.

Simultaneous confidence bands centered around a nonparametric regression estimate are an especially simple way to assess the adequacy of model fit and the variance of estimators over the entire range of observed data. A large literature already exists on construction of nonparametric confidence bands, with much of the work concerned with kernel type estimators.
Examples include (Claeskens and Keilegom, 2003), (Eubank and Speckman, 1993), (Neumann and Polzehl, 1998), and (Xia, 1998) for local polynomial smoothing; and (Härdle, 1989) for the Nadaraya-Watson regression estimator. However, kernel type methods require fitting a model at each prediction point and do not offer a closed-form formula to easily calculate future predictions. Polynomial spline regression is a less computation heavy nonparametric method that offers similar flexibility and approximation accuracy.

To construct bands, most approaches rely on either an asymptotic or bootstrap method. In asymptotic constructions, the bands are calculated using quantiles from the theoretical asymptotic distribution of mean function estimators over the entire range of data. Although the asymptotic bands are valid approximations as the sample size tends to infinity, they can suffer from poor coverage for small and moderate sample sizes. See (Wang and Yang, 2009) and (Song and Yang, 2009) for some examples of coverage rates in simulation studies. Furthermore, asymptotic bands often require plug-in estimators for the variance of the mean function estimator, which in turn effects accuracy and ease of implementation.

The bootstrap provides a simple alternative to asymptotic approximations and is a very easy method to implement in cases where the asymptotic distribution is not available. Currently, the technique comes in many flavors, each appropriate to a different data generating process, but all varieties of the method rely on some sort of random perturbations of the original data or regression residuals to approximate the finite sample distribution of an estimator. As noted above, spline regression is less computationally intensive than local polynomial and kernel regression, requiring only a single least squares estimate. However, to the authors knowledge,
no results are available on the validity of the bootstrap method to construct confidence bands from polynomial spline regression estimators.

In this work, we take up systematic study of confidence bands based on a bootstrap of second order polynomial spline regression estimates. For our data model, we assume heteroscedastic errors. To account for this heterogeneity we use the Wild bootstrap to approximate the distribution of the maximum deviation, which is defined as $\sup_{x \in B} |\hat{m}(x) - m(x)|$, where $\hat{m}(x)$ is the nonparametric estimate of the mean function $m(x)$ and $B$ is a compact set. ((Liu, 1988), (Härdle and Mammen, 1993), (Neumann and Polzehl, 1998)).

We show that the Wild bootstrap yields consistent estimates of the variance and maximal deviation of the linear spline estimator. Using this, we construct bands that are symmetric but are not uniform, which is very useful in studying the variability of the spline estimator at different points of the range as well as testing for regression fit. To choose the number of knots, we use generalized cross validation. To correct for finite sample bias, we use undersmoothing to determine the centering point and regression residuals.

The rest of this chapter is organized as follows. In Section 2.2 we introduce polynomial spline regression. Section 2.3 is devoted to our bootstrap algorithm for constructing confidence bands. It includes consistency results for the bootstrap estimators required to construct the bands. In Section 2.4, we report simulation results and empirical results using Canadian labor wage data. Section 2.5 provides a brief conclusion of the overall project, while proofs of the asymptotic results are detailed in Section 2.6.
2.2 B-Splines

Throughout this paper, we assume that the observations \( \{(X_i, Y_i)\}_{i=1}^n \) and unobserved errors \( \{\varepsilon_i\}_{i=1}^n \) are i.i.d. copies of \((Y, X, \varepsilon)\) satisfying the model

\[
Y = m(X) + \sigma(X)\varepsilon, \tag{2.1}
\]

where \( E(\varepsilon|X) = 0, E(\varepsilon^2|X) = 1 \). In what follows, let \( \epsilon_i = \sigma(X_i)\varepsilon_i \). To define polynomial spline regression, divide \([a, b]\), which is the range of \(X\), into \((N + 1)\) subintervals \(J_j = [t_j, t_{j+1}), j = 0, ..., N - 1, J_N = [t_N, b]\). The \( \{t_j\}_{j=1}^N \) are called interior knots and are determined by

\[
t_j = a + jh, j = 0, ..., N + 1,
\]

where \( h = (b - a)/(N + 1) \). For any \( x \in [a, b] \) define its location index as \( j(x) = j_n(x) = \min \{[\frac{x-a}{h}, N]\} \). Let \( G^{(p-2)}[a, b] \) be the space of functions that are polynomials of degree \((p-1)\) on each \(J_j\) and have \(p-2\) continuous derivatives.

Under this set-up, \( G_{p, \tau} \) is a linear space of dimension \(p + (N + 1)\). It can be shown that any function in this space can be uniquely represented by a linear combination of Basis spline functions of degree \(p - 1\) (De Boor, 2001 Rev ed). Basis spline functions or B-splines as they
are more commonly called can be defined recursively starting from the constant spline basis for $G_{0,\tau}$. The constant basis is simply a set of indicator functions

$$B_{1,\tau}(x) = \{ I_{(a,t_1)}(x), \ I_{[t_1,t_2)}(x), \ I_{(t_2,t_3]}(x), \cdots, \ I_{[t_{N-1},b)}(x) \}^T$$

This set of functions is very easy to work with in practice. For any $x$, the above vector has exactly 1 nonzero component. Under a reasonable choice of knots, the associated model matrix will be full rank.

Now define the following vectors $X = \{ X_1, \cdots, X_n \}^T$ and $B_{p,\tau}(X) = \{ B_{p,\tau}(X_1), \cdots, B_{p,\tau}(X_n) \}^T$.

A constant spline estimate for $m(x)$ can be defined as:

$$\hat{m}(x) = \{ \lambda_0, \lambda_1, \cdots, \lambda_{N+1} \} B_{1,\tau}(x),$$

where $\{ \lambda_0, \lambda_1, \cdots, \lambda_{N+1} \}$ is the solution to

$$\min_{\lambda \in \mathbb{R}^{N+1}} \sum_{i=1}^{n} \{ Y_i - \lambda^TB_{1,\tau}(X_i) \}^2 \quad (2.2)$$

Moving onto the B-Spline basis for $G_{1,\tau}$, let $K(u) = (1 - |u|)_+, h_j = t_{j+1} - t_j$. Figure 1 illustrates an example of these based on a uniform sequence of knots over the interval $[0,1]$.

Then set of 2nd order or linear B-Spline basis functions are:

$$B_{j,2}(x) = K \left( \frac{x - t_{j+1}}{h_j} \right), \ j = -1, 0, ..., N.$$
These can actually be generated from the constant by

\[ B_{j,2}(x) = \frac{(x - t_j)}{t_{j+1} - t_j} B_{j,1}(x) + \frac{(t_{j+1} - x)}{t_{j+2} - t_{j+1}} B_{j+1,1}(x), j = -1, \ldots, N \]

In general, the \( p \)th order B-Spline basis can be generated from the \( p - 1 \)th order basis as

\[ B_{i,p}(x) = \frac{(x - t_i)}{t_{i+p} - t_i} B_{j,p-1} + \frac{(t_{i+p+1} - x)}{t_{i+p+1} - t_{i+1}} B_{j+1,p-1}(x), i = 1 - p, \ldots, N \]
For a general $p$, boundary knots are added $t_{1-p} = \cdots = t_0 = a.$ and $t_{N+1}, \ldots, t_{N+p-1} = b$.

The resulting basis possesses some nice properties that make it easy to work with analytically. First, for any observation, \( \sum_{j=1-p}^{N} B_{j,p,\tau} = 1 \) and $0 \leq B_{j,p,\tau} \leq 1$. In addition, $B_{p,\tau}$ has at least one nonzero entry and at most $p$, so the empirical inner product matrix, $\left( \frac{1}{n} B_{p,n}^T B_{p,n} \right)$ has a 'nice' banded structure with bounded entries.

Figure 2. B-Splines, $p = 1, 2, 3, 4, 5, 6$

To define the spline regression estimator $\hat{m}_p(x)$ of $m(x)$, first fix a set of knot locations $\tau$ and then define the following vectors and vector valued functions $Y = (Y_1, \ldots, Y_n)^T$, $X = (X_1, \ldots, X_n)^T$, and $B_p(\cdot) = \{B_{1-p,p}(\cdot), \ldots, B_{N,p}(\cdot)\}^T \in \mathbb{R}^{N+p}$. In addition, let the the spline
regression design matrix be $B_{p,n} = \{B_p(X_1), \ldots, B_p(X_n)\}^{T}_{n \times (N+p)}$. Further denote the empirical inner product matrix of the spline of $B_{p,n}$ as $V_n = (\frac{1}{n} B^T_{p,n} B_{p,n})$ and its inverse as $S_n = V^{-1}_n$.

With this notation in hand, the $p$ order spline estimate for $m(x)$ can be written as

$$\hat{m}_p(x) = \frac{1}{n} B_p(x)^T S_n B_p(X)^T Y.$$

### 2.2.1 Error Decomposition

In this section, we decompose the regression error $\hat{m}_2(x) - m(x)$ into a bias and variance part. We then use this decomposition in Section 3 to study how well the bootstrap approximates the distribution of the original estimates. Throughout the rest of this paper, we assume a uniform knot sequence $\tau$ s.t. $h_j \equiv h \forall j = -1, \ldots, N$ and $h = (b - a)/N$. In what follows, we also use the following technical assumptions

(A1) The regression function $m(x) \in C^{(p)}[a, b]$.

(A2) The density function $f(x)$ of $X$ is continuous and positive on $[a, b]$.

(A3) The number of interior knots $N \sim n^{1/(2p+1)} \iff h \sim n^{-1/(2p+1)}$

(A4) The joint distribution $F(x, \epsilon)$ of random variables $(X, \epsilon)$ satisfies the following:

(a) The error is conditional white noise: $E(\epsilon|X = x) = 0, E(\epsilon^2|X = x) = \sigma^2(x)$. Further assume $E(\epsilon^4|X = x) = \mu_4(x)$ is a positive function on $[a,b]$ with bounded variation.

The variance function $\sigma^2(x) \in C^{(p)}[a, b]$ and is bounded below on $[a,b]$.

(b) There exists a positive value $\delta > 1/p$ and finite positive $M_\delta$ such that $E|\epsilon|^{2+\delta} < M_\delta$ and $\sup_{x \in [a,b]} E(|\epsilon|^{2+\delta}|X = x) < M_\delta$. 
Assumptions (A1)-(A4) are adapted from (Wang and Yang, 2009) and (Song and Yang, 2009) and the results that follow rely on spline regression asymptotics that were studied in both of these papers.

Denote as \( \| \cdot \| \) the supremum norm of a function on \([a,b]\). Define the theoretical and empirical inner products by

\[
\langle \phi, \varphi \rangle = \int_a^b \phi(x)\varphi(x)f(x)dx = E\{\phi(x)\varphi(x)\}, \quad \langle \phi, \varphi \rangle_n = \frac{1}{n} \sum_{i=1}^n \phi(X_i)\varphi(X_i).
\]

Fix \( p=2 \) and denote the theoretical inner product matrix of the 2\(^{nd}\) order B-spline basis as

\[
V = (\langle B_{j,2}, B_{j,2} \rangle)_{j,j=-1}^N, \quad \text{and its inverse as } S = (s_{jj})_{j,j=-1}^N = V^{-1}.
\]

For polynomial spline regression the estimation error \( \hat{m}_p(x) - m(x) \) can be decomposed into a bias and a noise term.

Write \( Y \) as the sum of a signal vector, \( M \), and a noise vector \( E \), where

\[
Y = M + E
\]

\[
M = \{m(X_1), \ldots, (X_n)\}^T, \quad E = \{\epsilon_1, \ldots, \epsilon_n\}^T.
\]

Projecting \( Y \) onto the linear space \( G_n^{(p-2)} \) spanned by \( B_{p,n} \), one gets

\[
\hat{m}_p = \{\hat{m}_p(X_1), \ldots, \hat{m}_p(X_n)\}^T = \text{Proj}_{G_n^{(p-2)}} Y = \text{Proj}_{G_n^{(p-2)}} M + \text{Proj}_{G_n^{(p-2)}} E
\]
Correspondingly, in the space $G^{(p-2)}$ of spline functions, one has $\tilde{m}_p(x) = \tilde{m}_p(x) + \tilde{\epsilon}_p(x)$, where

$$\tilde{m}_p(x) = \frac{1}{n} B_2(x)^T \left( \frac{1}{n} B_{p,n}^T B_{p,n} \right)^{-1} B_{2,n}^T M$$

$$\tilde{\epsilon}_p(x) = \frac{1}{n} B_2(x)^T \left( \frac{1}{n} B_{p,n}^T B_{p,n} \right)^{-1} B_{2,n}^T E \quad (2.3)$$

By Lemma 3.1 of (Wang and Yang, 2009), the empirical inner products of the sample B-spline basis for $p=1,2$ can be approximated by their corresponding theoretical inner products uniformly in $G^{(p-2)}$. In particular, for $p = 2$ the empirical inner product matrix can be decomposed into two parts

$$\left( \langle B_{j,2}, B_{j,2} \rangle_n \right)_{j, j=1}^N = V + \hat{B}, \quad (2.4)$$

where

$$|\hat{B}| = O_p \left( \sqrt{n^{-1} h^{-1} \log(n)} \right). \quad (2.5)$$
By definition, $S_n = \left(V + \tilde{B}\right)^{-1}$ and let $\tilde{\omega}_{i,n}(x) = B_2(x)^T S_n B_2(x_i), \; i = 1, \ldots, n$. Since $Y$ can be written as the sum of a mean vector $M$ and a noise vector $E$,

$$\hat{m}_2(x) = \frac{1}{n} B_2(x)^T S_n B_{2,n}^T Y = \frac{1}{n} B_2(x)^T S_n B_{2,n}^T \{M + E\}$$

$$= \tilde{m}_2(x) + \tilde{e}_2(x)$$

$$= \tilde{m}_2(x) + \frac{1}{n} \sum_{i=1}^{n} \tilde{\omega}_{i,n}(x) \epsilon_i$$

Define the matrix $\Sigma$ as

$$\Sigma = (\sigma_{jl})_{j,l=-1}^{N} = \left( \int \sigma^2(v) B_{j,2}(v) B_{l,2}(v) f(v) \partial v \right)_{j,l=-1}^{N} \quad (2.6)$$

By Lemma B.4 of (Wang and Yang, 2009) the pointwise variance of $\hat{m}_2(x)$ is approximated by the quantity

$$\sigma^2_{2,n}(x) = \sum_{j,j,l,l=-1}^{N} \frac{1}{n} B_{j,2}(x) B_{l,2}(x) s_{jj} s_{ll} \sigma_{jl}$$

$$= \frac{1}{n} B_2(x)^T \Sigma S^T B_2(x). \quad (2.7)$$

For the bootstrap procedure outlined in the sequel, we shall compare our bootstrap estimates of the estimator variance to this asymptotic approximation.
2.2.2 Knot Choice

There is no one single directive on knot choice. Popular solutions for choosing the distance between knots include uniform spacing and using data percentiles. Optimal knot number, $N$, can be based on asymptotic optimality and/or model selection. Knot number and location is an important factor in the accuracy of spline approximation. Even for fixed (nonrandom) data, the order of approximation for an arbitrary smooth function with $p-1$ continuous derivatives is of $O(|h|^p)$, where $h_i = t_{i+1} - t_i$ and $|h| = \max_i \{h_i\}$ (De Boor, 2001 Rev ed). Similarly, for random samples, the bias of approximation is $O_p(|h|^p)$ (Huang, 2003). Therefore for uniform knots (equal spacing), the asymptotic bias goes to 0 when the knot spacing is of order $n^{-1/(2p+1)}$.

However, uniform knot location can miss important aspects of the target function, especially inflection and extrema. Ideally, if the true regression function was known, good knot locations would reflect curvature. In practice, these are either eye-balled or more involved automatic methods can be used. The latter, known as free-knots, is an interesting and evolving area of research that is treated in Chapter 2 of this dissertation.

Finally, knot points do not need to be distinct. The number of knots at a break point sets restrictions on smoothness of the fit at the break. For $k$ multiple knots, the estimated function will have $p-k$ continuous derivatives at the break. Therefore, a single knot, guarantees that the estimate has $(p-1)$ continuous derivatives, 2 knots gives $(p-2)$ continuous derivatives and so on. Accounting for such discontinuities, however, affects efficiency of estimation and interpolation.
2.3 The Wild Bootstrap for Regression Confidence Bands

In this work we use the wild bootstrap to approximate the finite distribution of spline regression estimator $\hat{m}_2(x)$. We construct asymptotically exact bands for the conditional mean function by bootstrapping the second order polynomial spline estimator. Denote $\Delta_n = \sup_{x \in [a,b]} |\sigma^{-1}_{2,n}(x) (\hat{m}_2(x) - m(x))|$, where $\sigma^2_{2,n}(x)$ is the pointwise variance function (asymptotic) of the estimator defined in (8). An exact simultaneous confidence band for $m(x)$ is determined by

$$P\{\Delta_n \leq \theta_\alpha\} = P\{m(x) \in \hat{m}_2(x) \pm \theta_\alpha \sigma_{n,2}(x), \forall x \in [a,b]\} = 1 - \alpha \quad (2.8)$$

In order to construct this band, we need to know both the distribution of $\Delta_n$ and an estimate of $\sigma^2_{n,2}(x)$. Since the exact finite sample distribution of $\Delta_n$ is intractable, we approximate it using bootstrapping. Further, we also use a bootstrap estimate for the pointwise variance of $\hat{m}_2(x)$. As our problem concerns data with a heterogenous variance structure, the width of the confidence band varies with $x$ through the pointwise variance function.

A good bootstrap works by mimicking the underlying data generating process. For regression estimators, it is important to capture the underlying distribution of the unobserved random errors $\{\epsilon_i\}$. In the case of an IID regression model with a homogeneous variance assumption, bootstrapping is usually done by sampling from the empirical distribution of the centered regression residuals: $\hat{\epsilon} = Y - \hat{m}(X)$. However, the resulting bootstrap estimates are inconsistent.
when the variances are heterogenous (Freedman, 1981), (Liu, 1988), (Härdle and Mammen, 1993).

The Wild bootstrap (WB) is a simple alternative that accounts for the heterogenous variance structure. Originally proposed by (Wu, 1986), it was shown to be consistent for the simple linear regression model in (Liu, 1988) under the heterogenous variance assumption. This method was successfully used in (Härdle and Mammen, 1993) to construct hypothesis tests of parametric models using a bootstrap of Kernel regression estimators.

In addition to providing valid estimates, the Wild bootstrap is very easy to implement. Called the Wild bootstrap because it uses a single residual to estimate the sampling distribution at each observed point. Instead of resampling the residuals directly, each residual is instead multiplied by an auxiliary variable generated independent of the data. WB for confidence bands was studied by (Neumann and Polzehl, 1998) for the case of Kernel and local polynomial regression estimators. These were shown to be valid for estimating the distribution of the maximal deviation under the heterogeneous variance assumption. To the best of the author’s knowledge, there has been no work on showing the validity of the bootstrap for regression spline confidence bands. However, simulations in (Song and Yang, 2009) suggest that the WB can provide a good alternative to the asymptotic approximation for smaller sample sizes.
2.3.0.1 Algorithm

Denote $\Delta_n = \sup_{x \in [a, b]} |\sigma_{2,n}^{-1}(x) (\hat{m}_2(x) - m(x))|$, where $\sigma_{2,n}^2(x)$ is the pointwise variance function (asymptotic) of the estimator defined in (8). An exact simultaneous confidence band for $m(x)$ is determined by

$$P\{\Delta_n \leq \theta_n\} = P\{m(x) \in \hat{m}_2(x) \pm \theta_n \sigma_{n,2}(x), \forall x \in [a, b]\} = 1 - \alpha \quad (2.9)$$

In order to construct this band, we need to know both the distribution of $\Delta_n$ and an estimate of $\sigma_{n,2}^2(x)$. Since the exact finite sample distribution of $\Delta_n$ is intractable, we approximate it using bootstrapping. Further, we also use a bootstrap estimate for the pointwise variance of $\hat{m}_2(x)$. As our problem concerns data with a heterogenous variance structure, the width of the confidence band varies with $x$ through the pointwise variance function. Define the regression residuals $\hat{\epsilon}_i = Y_i - \hat{m}_2(X_i), i = 1, ..., n$. Let $B$ denote a predetermined integer and $\alpha \in (0, 1)$.

Step 1. Generate i.i.d. sample $\{\eta_{i,k}\}_{1 \leq k \leq B, i=1,...,n}$ independently of the data and residuals, where $\eta^* \sim F_{\eta}$ s.t. $E(\eta^*) \equiv 0$ and $Var(\eta^*) \equiv 1$.

Step 2. Let the $k^{th}$ bootstrap sample be defined as $\{Y_{i,k}^*, X_i\}_{i=1,...,n}$ where

$$Y_{i,k}^* = \hat{m}_2(X_i) + \hat{\epsilon}_i \eta_{i,k}^*.$$  

Write $Y_k^* = \left(Y_{1,k}^*, ..., Y_{n,k}^*\right)^T$. The linear spline operator is applied to each of the $k$ bootstrap samples to yield $k$ bootstrap estimates of the mean function at $x$:  

\[m_{2,k}^*(x) = \frac{1}{n} B_2(x)^T \left( \frac{1}{n} B_{2,n}^T B_2 \right)^{-1} B_{2,n}^T Y_k^*\]

Step 3. By 2.3.1 below, a consistent bootstrap estimate for \(\sigma_{2,n}^2(x)\) is given by

\[\sigma_{2,n}^2(x) = \frac{1}{n^2} \sum_{i=1}^{n} \left( B_2(x)^T \left( \frac{1}{n} B_{2,n}^T B_2 \right)^{-1} B_2(X_i)^T \right)^2 \hat{\epsilon}_i^2\]

Define \(\Delta_{nk}^* = \sup_{x \in [a,b]} |\sigma_{2,n}^2(x) \left( m_{2,k}^*(x) - \hat{m}_2(x) \right) |. \) In our Theorem 4.2, we show that \(\Delta_n^*\) is asymptotically equivalent to \(\Delta_n\). Therefore, to approximate the finite sample distribution of \(\Delta_n\) we use the empirical bootstrap distribution of \(\{\Delta_{n1}, \cdots, \Delta_{nB}\}\).

The Wild bootstrap \((1 - \alpha)\) simultaneous confidence bands for \(m(x)\) are then

\[\{m(x) \in \hat{m}_2(x) \pm \theta^*_\alpha \sigma_{2,n}^*(x), x \in [a,b]\}, \quad (2.10)\]

where \(\theta^*_\alpha\) is the \((1 - \alpha)\) quantile of \(\{\Delta_{n1}^*, \cdots, \Delta_{nB}^*\}\).

2.3.1 Consistency of the Bootstrap Estimators

Given the original sample, denote the conditional expectation of bootstrap estimates as \(E^*\) and the conditional probability as \(P^*\). To show consistency of the bootstrap estimates, we require the following assumption on the distribution of the \(\eta^*\).

(A5) The distribution of the bootstrap auxiliary variables \(\eta^*\) is s.t. \(E^*(\eta^*) = E(\eta^*) = 0\) and \(E^*(\text{left}(\eta^*))^2 = 1\). Further, there exists a finite positive \(\eta_0\) s.t. \(E^*|\eta^*|^3 \leq \eta_0 < \infty\).
Mirroring the error decomposition in Section 2.1, define the corresponding bootstrap quantities

\[ Y^* = (Y_1^*, ..., Y_n^*)^T = (\hat{m}_2(X_1) + \hat{\epsilon}_1 \eta_1^*, ..., \hat{m}_2(X_n) + \hat{\epsilon}_n \eta_n^*)^T = \hat{M} + E^* \]

Therefore,

\[ m_2^*(x) = \text{Proj}_{G^{(p-2)}} Y^* = \frac{1}{n} B_2(x)^T S_n B_2^T \{ \hat{M} + E^* \} \]

\[ = \hat{m}_2(x) + \frac{1}{n} \sum_{i=1}^{n} \hat{\omega}_{i,n}(x) \hat{\epsilon}_i \eta_i^* \]

Let \( t_n(x) = \hat{m}_2(x) - m_2(x) \) and its bootstrap counterpart as \( t_n^*(x) = m_2^*(x) - \hat{m}_2(x) \). Then by Lemma A.1 in Appendix A,

\[ t_n(x) = \frac{1}{n} \sum_{i=1}^{n} \hat{\omega}_{i,n}(x) \epsilon_i + (\hat{m}_2(x) - m_2(x)) = \frac{1}{n} \sum_{i=1}^{n} \hat{\omega}_{i,n}(x) \epsilon_i + O_p(n^{-2/5}) \]  \hspace{1cm} (2.11)

\[ t_n^*(x) = \frac{1}{n} \sum_{i=1}^{n} B_2(x)^T S_n B_2(X_i) \hat{\epsilon}_i \eta_i^* = \frac{1}{n} \sum_{i=1}^{n} \hat{\omega}_{i,n}(x) \hat{\epsilon}_i \eta_i^* \]  \hspace{1cm} (2.12)

Note that \( E^* (m_2^*(x)) = E (m_2^*(x)|Y, X) = \hat{m}_2(x) \) and \( E^* (t_n^*(x)) = 0 \). Denote the conditional variance of \( t_n^*(x) \) as \( \sigma_{2,n}^2(x) \). Then

\[ \sigma_{2,n}^2(x) = E^* \left( (t_n^*(x))^2 \right) = \frac{1}{n^2} \sum_{i=1}^{n} (B_2(x)^T S_n B_2(X_i))^2 \hat{\epsilon}_i^2. \]  \hspace{1cm} (2.13)
Using Lemmata 2.7.1-2.7.5 in Appendix A we have the following result for the consistency of the bootstrap variance estimator.

**Theorem 2.3.1.** Under Assumptions (A1)-(A4),

\[ \| \sigma_{2,n}^2(x) - \sigma_{2,n}^2(x) \|_\infty = O_p \left( n^{-6/5} \sqrt{\log n} \right) \quad (2.14) \]

Further, the following shows that the Wild bootstrap method yields asymptotically valid approximations to the point wise distribution of \( \hat{m}_2(x) \) as well as the supremum \( \Delta_n \). The proofs are given in Appendix B.

**Theorem 2.3.2.** Under Assumptions (A1)-(A4),

\[ \sigma_{2,n}^{-1}(x)t_n(x) \overset{D}{\rightarrow} N(0, 1). \quad (2.15) \]

\[ \sigma_{2,n}^{-1*}(x)t_n^*(x) \overset{D^*}{\rightarrow} N(0, 1). \quad (2.16) \]

\[ |\Delta_n^* - \Delta_n| = o_p(1), \quad (2.17) \]

where \( \Delta_n = \sup_{x \in [a,b]} |\sigma_{2,n}^{-1}(x)t_n(x)| \) and \( \Delta_n^* = \sup_{x \in [a,b]} |\sigma_{2,n}^{-1*}(x)t_n^*(x)| \).

### 2.3.2 Knot choice and bias correction

Finite sample bias is an important issue in confidence band coverage. The initial spline regression estimator enters into the Wild bootstrap calculations in two ways– there are the
regression residuals that are perturbed at each bootstrap iteration and the points at which the bands are centered. Theory shows that when the number of knots is chosen to be of the optimal order $\sim n^{1/5}$ the original linear spline regression estimate of the conditional mean converges to the true value in the supremum norm. However, although the spline estimator does converge to the true function, the bias goes to 0 at a rate of $n^{2/5}$, see Lemma A.1 in the Appendix. For finite samples then, the estimated function may be a poor approximation and needs to be adjusted for bias in calculating the bootstrap bands.

As can be seen from simulation results in the next section, when the original linear spline estimator is not adjusted for bias, the bands have poor coverage rates for small and moderate sized samples. However, in the course of our investigation, when we replaced the linear estimator with a higher order spline such as the quadratic or cubic, our simulated coverage rates resulted in much better coverage rates. When the unadjusted linear estimator gave band coverage rates between 30 to 50 percent when $n = 50$, a quadratic estimator had rates around 60 percent (redo these for x.eval=x.dat). Since a higher order spline is tantamount to including higher order terms in a polynomial approximation to the regression function, this indicates that bias adjustment is crucial for good band coverage.

In the literature, there are two main approaches for bias correction—either explicitly by direct estimation or implicitly by undersmoothing or oversmoothing of the original regression estimate. The latter is a much simpler solution and we follow it in our work. In our implementation, the optimal knot size $N$ is taken to be a function $N = c_1 \lfloor n^{1/5} \rfloor + c_2$, where $\lfloor \cdot \rfloor$ is the floor function and $c_1, c_2$ are constants chosen by Generalized Cross Validation (GCV). The GCV is
a model selection procedure for linear type estimators that approximates a leave-one-out cross validation estimate of model prediction error. The method relies on the mean squared error criterion and as such balances both bias and variance in the search for the best model. In order to control bias, we undersmooth the initial estimate of $\hat{m}(x)$ by using a larger number of knots equal to \( \lfloor c_1 (\log n^{1/3} \times n^{1/5}) \rfloor + c_2 \).

2.4 Examples

2.4.1 Simulation Study

To study the finite sample behaviour of the bootstrap confidence bands, we simulate data from the model in (1) with \( X \sim U[-1/2, 1/2] \) and

\[
m(x) = \sin(2\pi x), \quad \sigma(x) = \sigma_0 \frac{c - \exp(x)}{c + \exp(x)}, \quad \epsilon \sim \text{iid } N(0, 1).
\]

Under this specification, the noise to signal ratio is controlled by \( \sigma_0 \) and we use \( \sigma_0 = 0.2, 0.5, 1 \). The constant \( c \) controls the level of heteroscedasticity – the variance is nearly homoscedastic when \( c=100 \) and strongly heteroscedastic when \( c=5 \). The sample sizes are taken to be \( n = 50, 100, 200, 500 \), and the bootstrap sample size is \( B = 200 \). In addition to the bootstrap method outlined above, our simulations include the coverage rates of an oracle band that is based on using a true plug-in value for \( \sigma(x) \) to calculate the variance of \( \hat{m}(x) \). Tables 1 and 2 contain the coverage rates of 95% nominal confidence bootstrap and oracle bands as a percentage of coverage of the true curve at all data points from 500 replications.
In general, both the oracle and bootstrap bands that use an undersmoothed spline estimate \((c_0 = 1)\) for centering have much better coverage than bands with no bias adjustment \((c_0 = 1)\).

For small \(n=50\), the oracle band has coverage close to the nominal levels with bias adjustment. Further, the bootstrap band coverage for the nearly homoscedastic data goes from 0.31–0.50 to 0.70–0.75 with undersmoothing; and from 0.27–0.49 to 0.74–0.80 for the nearly heteroscedastic data. In terms of the bootstrap band performance, band coverage improves with larger sample sizes, coming close to the nominal when \(n=500\). However, for small and moderate sample sizes, the bootstrap bands cover the true function much less than the nominal rate.

Figures 3 to 5 illustrate simulated 95% confidence bootstrap and oracle bands, as well as estimates of \(\sigma_{n,2}(x)\), for data generated from model (16) with \(c = 100\) and \(n = 50, 100, 500\). Each of the figures has several components. In plots labeled ‘Mean’, the black solid lines are the true mean function \(m(x)\), the dashed red lines are the linear spline regression estimates \(\hat{m}_2(x)\), dashed green lines are the oracle bands, and the blue dashed lines are the bootstrap bands. In those labeled ‘Variance’, the green values are the oracle estimates of \(\sigma_{2,n}(x)\), the blue are the corresponding bootstrap estimates, and the vertical gray lines correspond to the interior knots. It is clear that coverage improves with increased sample size and gets close to the nominal level when \(n=500\). Further the band area declines with sample size. Still, the bands appear to be reasonable visual tools for model assessment even when \(n = 50\) for interior points, but are too narrow relative to their oracle counterparts on the boundary. As illustrated on the graphs on the right of Figure 1, the bootstrap variance estimate is below that of the oracle on
the boundary of the range of $x$. It appears then that coverage is effected by variance estimation in data sparse areas.

Next, we study the robustness of our procedure to deviations from independence. Instead of white noise, we change $\varepsilon$ in (16) to be an AR(1) mean zero Gaussian process. Let $\rho$ be the lag 1 correlation parameter, then under this model, $\text{Corr}(\varepsilon_i, \varepsilon_j) = \rho^{|i-j|}$, $\text{Cov}(Y_i, Y_j) = \frac{\sigma(X_i)\sigma(X_j)}{1-\rho^2}\rho^{|i-j|}$, and $\text{Var}(Y_i) = \frac{\sigma^2(X_i)}{1-\rho^2}$. Tables 3-4 present the coverage probabilities of both the bootstrap and oracle bands for the nearly homoscedastic and strongly heteroscedastic cases for $\rho = -0.9, -0.5, 0.5, 0.9$. It is clear that the bootstrap bands have worse coverage when are

**TABLE I**

**COVERAGE PROBABILITIES FROM 500 REPLICATIONS OF THE MODEL IN (16) WITH $C=100$**

<table>
<thead>
<tr>
<th>$\sigma_0$</th>
<th>$n$</th>
<th>Oracle</th>
<th>Bootstrap</th>
<th>Oracle</th>
<th>Bootstrap</th>
<th>Average N</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>50</td>
<td>0.73 (0.38)</td>
<td>0.31 (0.32)</td>
<td>0.96 (0.86)</td>
<td>0.70 (0.65)</td>
<td>6.54</td>
</tr>
<tr>
<td>0.5</td>
<td>50</td>
<td>0.79 (0.86)</td>
<td>0.48 (0.74)</td>
<td>0.94 (1.72)</td>
<td>0.75 (1.41)</td>
<td>4.81</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>0.82 (1.70)</td>
<td>0.50 (1.44)</td>
<td>0.91 (3.57)</td>
<td>0.72 (2.74)</td>
<td>4.75</td>
</tr>
<tr>
<td>0.2</td>
<td>100</td>
<td>0.7 (0.30)</td>
<td>0.45 (0.28)</td>
<td>0.97 (0.56)</td>
<td>0.87 (0.50)</td>
<td>6.97</td>
</tr>
<tr>
<td>0.5</td>
<td>100</td>
<td>0.85 (0.68)</td>
<td>0.68 (0.63)</td>
<td>0.99 (1.19)</td>
<td>0.87 (1.09)</td>
<td>5.30</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>0.88 (1.31)</td>
<td>0.70 (1.21)</td>
<td>0.96 (2.07)</td>
<td>0.84 (1.88)</td>
<td>4.98</td>
</tr>
<tr>
<td>0.2</td>
<td>200</td>
<td>0.72 (0.24)</td>
<td>0.57 (0.23)</td>
<td>0.99 (0.40)</td>
<td>0.88 (0.39)</td>
<td>7.55</td>
</tr>
<tr>
<td>0.5</td>
<td>200</td>
<td>0.77 (0.52)</td>
<td>0.64 (0.50)</td>
<td>0.93 (0.81)</td>
<td>0.89 (0.78)</td>
<td>5.80</td>
</tr>
<tr>
<td>1</td>
<td>200</td>
<td>0.89 (0.96)</td>
<td>0.74 (0.93)</td>
<td>0.93 (1.49)</td>
<td>0.87 (1.44)</td>
<td>4.76</td>
</tr>
<tr>
<td>0.2</td>
<td>500</td>
<td>0.78 (0.17)</td>
<td>0.70 (0.17)</td>
<td>0.98 (0.28)</td>
<td>0.93 (0.28)</td>
<td>8.92</td>
</tr>
<tr>
<td>0.5</td>
<td>500</td>
<td>0.80 (0.37)</td>
<td>0.74 (0.36)</td>
<td>0.95 (0.54)</td>
<td>0.93 (0.54)</td>
<td>6.77</td>
</tr>
<tr>
<td>1</td>
<td>500</td>
<td>0.88 (0.67)</td>
<td>0.83 (0.67)</td>
<td>0.90 (0.93)</td>
<td>0.89 (0.92)</td>
<td>5.64</td>
</tr>
</tbody>
</table>
TABLE II

COVERAGE PROBABILITIES FROM 500 REPLICATIONS OF THE MODEL IN (16) WITH C=5

<table>
<thead>
<tr>
<th>$\sigma_0$</th>
<th>n</th>
<th>Oracle</th>
<th>Bootstrap</th>
<th>Oracle</th>
<th>Bootstrap</th>
<th>Average N</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>50</td>
<td>0.63 (0.26)</td>
<td>0.27 (0.22)</td>
<td>0.99 (0.63)</td>
<td>0.74 (0.48)</td>
<td>6.90</td>
</tr>
<tr>
<td>0.5</td>
<td>50</td>
<td>0.76 (0.59)</td>
<td>0.39 (0.50)</td>
<td>0.97 (1.24)</td>
<td>0.80 (1.02)</td>
<td>5.14</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>0.82 (1.13)</td>
<td>0.49 (0.97)</td>
<td>0.92 (2.14)</td>
<td>0.76 (1.76)</td>
<td>4.63</td>
</tr>
<tr>
<td>0.2</td>
<td>100</td>
<td>0.68 (0.22)</td>
<td>0.35 (0.20)</td>
<td>0.97 (0.41)</td>
<td>0.81 (0.37)</td>
<td>8.00</td>
</tr>
<tr>
<td>0.5</td>
<td>100</td>
<td>0.80 (0.48)</td>
<td>0.54 (0.45)</td>
<td>0.98 (0.82)</td>
<td>0.86 (0.74)</td>
<td>5.98</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>0.86 (0.90)</td>
<td>0.64 (0.84)</td>
<td>0.98 (1.53)</td>
<td>0.85 (1.39)</td>
<td>5.19</td>
</tr>
<tr>
<td>0.2</td>
<td>200</td>
<td>0.67 (0.17)</td>
<td>0.48 (0.16)</td>
<td>0.97 (0.31)</td>
<td>0.90 (0.30)</td>
<td>8.78</td>
</tr>
<tr>
<td>0.5</td>
<td>200</td>
<td>0.68 (0.37)</td>
<td>0.62 (0.36)</td>
<td>0.93 (0.57)</td>
<td>0.90 (0.55)</td>
<td>6.42</td>
</tr>
<tr>
<td>1</td>
<td>200</td>
<td>0.76 (0.68)</td>
<td>0.65 (0.65)</td>
<td>0.93 (1.08)</td>
<td>0.91 (1.04)</td>
<td>5.39</td>
</tr>
<tr>
<td>0.2</td>
<td>500</td>
<td>0.68 (0.12)</td>
<td>0.65 (0.12)</td>
<td>0.97 (0.21)</td>
<td>0.91 (0.20)</td>
<td>10.38</td>
</tr>
<tr>
<td>0.5</td>
<td>500</td>
<td>0.81 (0.26)</td>
<td>0.73 (0.26)</td>
<td>0.99 (0.40)</td>
<td>0.95 (0.40)</td>
<td>7.79</td>
</tr>
<tr>
<td>1</td>
<td>500</td>
<td>0.83 (0.47)</td>
<td>0.72 (0.47)</td>
<td>0.96 (0.67)</td>
<td>0.94 (0.67)</td>
<td>6.10</td>
</tr>
</tbody>
</table>

errors are correlated, especially when there is very high positive autocorrelation, $\rho = 0.9$. When $\rho = -0.9, -0.5, 0.5$ the bands’ performance is only moderately worse than for the uncorrelated case. However, when $\rho = 0.9$, band performance is very poor even when $n=500$. At this points, coverage rates are between 0.49 and 0.68. Further, even for $n=500$, the coverage rates do not approach the nominal level for all simulated values of $\rho$. Thus, the bands are useful for moderately correlated data but are not effective for analysis when high autocorrelation is suspected.

2.4.2 Canadian wage data

In this section, we illustrate the bootstrap bands with a simple regression analysis of Canadian cross-section wage data from (Pagan and Ullah, 1999). The dataset consists of 205 ran-
Figure 3. Linear spline estimates and 95% confidence bands, n=50
Figure 4. Linear spline estimates and 95% confidence bands, n=100
Figure 5. Linear spline estimates and 95% confidence bands, n=500
### Table III

Coverage Probabilities from 500 Replications of AR(1) Error Model and C=100

<table>
<thead>
<tr>
<th>$\sigma_0$</th>
<th>n</th>
<th>Oracle $\rho = -0.9$</th>
<th>Bootstrap $\rho = -0.9$</th>
<th>Oracle $\rho = -0.5$</th>
<th>Bootstrap $\rho = -0.5$</th>
<th>Oracle $\rho = 0.5$</th>
<th>Bootstrap $\rho = 0.5$</th>
<th>Oracle $\rho = 0.9$</th>
<th>Bootstrap $\rho = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>50</td>
<td>0.96</td>
<td>0.77</td>
<td>0.98</td>
<td>0.80</td>
<td>0.96</td>
<td>0.74</td>
<td>0.85</td>
<td>0.37</td>
</tr>
<tr>
<td>0.5</td>
<td>50</td>
<td>0.93</td>
<td>0.75</td>
<td>0.98</td>
<td>0.75</td>
<td>0.92</td>
<td>0.76</td>
<td>0.78</td>
<td>0.35</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>0.92</td>
<td>0.74</td>
<td>0.97</td>
<td>0.73</td>
<td>0.89</td>
<td>0.72</td>
<td>0.72</td>
<td>0.26</td>
</tr>
<tr>
<td>0.2</td>
<td>100</td>
<td>0.93</td>
<td>0.89</td>
<td>0.97</td>
<td>0.86</td>
<td>0.94</td>
<td>0.83</td>
<td>0.63</td>
<td>0.47</td>
</tr>
<tr>
<td>0.5</td>
<td>100</td>
<td>0.92</td>
<td>0.84</td>
<td>0.93</td>
<td>0.79</td>
<td>0.88</td>
<td>0.82</td>
<td>0.63</td>
<td>0.41</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>0.93</td>
<td>0.85</td>
<td>0.93</td>
<td>0.81</td>
<td>0.84</td>
<td>0.76</td>
<td>0.55</td>
<td>0.34</td>
</tr>
<tr>
<td>0.2</td>
<td>200</td>
<td>0.92</td>
<td>0.93</td>
<td>0.97</td>
<td>0.91</td>
<td>0.90</td>
<td>0.87</td>
<td>0.70</td>
<td>0.52</td>
</tr>
<tr>
<td>0.5</td>
<td>200</td>
<td>0.94</td>
<td>0.92</td>
<td>0.96</td>
<td>0.94</td>
<td>0.91</td>
<td>0.85</td>
<td>0.69</td>
<td>0.56</td>
</tr>
<tr>
<td>1</td>
<td>200</td>
<td>0.94</td>
<td>0.90</td>
<td>0.96</td>
<td>0.92</td>
<td>0.87</td>
<td>0.83</td>
<td>0.61</td>
<td>0.47</td>
</tr>
<tr>
<td>0.2</td>
<td>500</td>
<td>0.88</td>
<td>0.87</td>
<td>0.92</td>
<td>0.91</td>
<td>0.91</td>
<td>0.87</td>
<td>0.76</td>
<td>0.64</td>
</tr>
<tr>
<td>0.5</td>
<td>500</td>
<td>0.88</td>
<td>0.86</td>
<td>0.92</td>
<td>0.92</td>
<td>0.90</td>
<td>0.88</td>
<td>0.61</td>
<td>0.53</td>
</tr>
<tr>
<td>1</td>
<td>500</td>
<td>0.87</td>
<td>0.84</td>
<td>0.89</td>
<td>0.89</td>
<td>0.88</td>
<td>0.84</td>
<td>0.56</td>
<td>0.49</td>
</tr>
</tbody>
</table>

### Table IV

Coverage Probabilities from 500 Replications of AR(1) Error Model and C=5

<table>
<thead>
<tr>
<th>$\sigma_0$</th>
<th>n</th>
<th>Oracle $\rho = -0.9$</th>
<th>Bootstrap $\rho = -0.9$</th>
<th>Oracle $\rho = -0.5$</th>
<th>Bootstrap $\rho = -0.5$</th>
<th>Oracle $\rho = 0.5$</th>
<th>Bootstrap $\rho = 0.5$</th>
<th>Oracle $\rho = 0.9$</th>
<th>Bootstrap $\rho = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>50</td>
<td>0.94</td>
<td>0.73</td>
<td>0.98</td>
<td>0.74</td>
<td>0.96</td>
<td>0.72</td>
<td>0.86</td>
<td>0.40</td>
</tr>
<tr>
<td>0.5</td>
<td>50</td>
<td>0.94</td>
<td>0.77</td>
<td>0.98</td>
<td>0.75</td>
<td>0.93</td>
<td>0.72</td>
<td>0.82</td>
<td>0.40</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>0.94</td>
<td>0.74</td>
<td>0.97</td>
<td>0.73</td>
<td>0.91</td>
<td>0.76</td>
<td>0.75</td>
<td>0.28</td>
</tr>
<tr>
<td>0.2</td>
<td>100</td>
<td>0.95</td>
<td>0.87</td>
<td>0.98</td>
<td>0.87</td>
<td>0.98</td>
<td>0.88</td>
<td>0.67</td>
<td>0.47</td>
</tr>
<tr>
<td>0.5</td>
<td>100</td>
<td>0.94</td>
<td>0.88</td>
<td>0.97</td>
<td>0.85</td>
<td>0.90</td>
<td>0.81</td>
<td>0.63</td>
<td>0.45</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>0.93</td>
<td>0.86</td>
<td>0.96</td>
<td>0.81</td>
<td>0.88</td>
<td>0.81</td>
<td>0.58</td>
<td>0.35</td>
</tr>
<tr>
<td>0.2</td>
<td>200</td>
<td>0.91</td>
<td>0.91</td>
<td>0.97</td>
<td>0.91</td>
<td>0.95</td>
<td>0.90</td>
<td>0.73</td>
<td>0.56</td>
</tr>
<tr>
<td>0.5</td>
<td>200</td>
<td>0.92</td>
<td>0.94</td>
<td>0.97</td>
<td>0.95</td>
<td>0.93</td>
<td>0.88</td>
<td>0.68</td>
<td>0.57</td>
</tr>
<tr>
<td>1</td>
<td>200</td>
<td>0.94</td>
<td>0.93</td>
<td>0.97</td>
<td>0.92</td>
<td>0.89</td>
<td>0.84</td>
<td>0.67</td>
<td>0.53</td>
</tr>
<tr>
<td>0.2</td>
<td>500</td>
<td>0.91</td>
<td>0.86</td>
<td>0.95</td>
<td>0.91</td>
<td>0.91</td>
<td>0.88</td>
<td>0.73</td>
<td>0.68</td>
</tr>
<tr>
<td>0.5</td>
<td>500</td>
<td>0.88</td>
<td>0.82</td>
<td>0.90</td>
<td>0.90</td>
<td>0.92</td>
<td>0.89</td>
<td>0.62</td>
<td>0.57</td>
</tr>
<tr>
<td>1</td>
<td>500</td>
<td>0.89</td>
<td>0.84</td>
<td>0.92</td>
<td>0.90</td>
<td>0.86</td>
<td>0.83</td>
<td>0.58</td>
<td>0.50</td>
</tr>
</tbody>
</table>
domly selected observations of wage and age from the 1971 Canadian Census Public Use Tapes for male individuals having common education (grade 13). This data was used in (Hayfield and Racine, 2008) to illustrate kernel type nonparametric regression for econometric data analysis. Visual inspection of the data indicates a heterogeneous variance structure, which is in tandem with economic intuition. The log(wage) is more dispersed for the youngest workers who are just starting their careers; wages are less dispersed and do not vary much between about age 25 through 35; and wage dispersion increases with every year thereafter. We test several polynomial regression models of the logarithm of wage $y$ as a function of age $x$: $y = a_0 + \sum_{k=1}^{d} a_k x^k, \; d = 2, 3, 4, 5$.

Figures 6 and 7 present the fitted polynomial regression models along with 80% confidence bands, while the bootstrap estimate for the estimator variance in given in Figure 8. The solid green lines are the parametric fits with $d = 2, 3, 4, 5$ respectively; the dashed red line is the undersmoothed linear spline estimate of the mean functions; and the dashed blue lower and upper lines are the confidence bands. For $d=2,3,$ and even 4, the regression models fail to capture the sharp rise and then relative stagnation in wages between ages 25 though 35. The curves for $d=2$ and 3 are very easily rejected at a confidence level $\alpha = 0.20$. Further, although the curve with degree $d=4,$ seems to be a good approximation to the data for most of the age range, it still falls below the lower band at age=25. In contrast, a higher degree polynomial curve of degree $d=5$ fits the data well. The entire fitted regression curve sits within the confidence band bounds.
Overall, the bands do provide a nice visual idea of the possible shape of the unknown function over most of the range of age. Wages increase rapidly in earliest years of labor force participation, presumably because of fast acquisition of valuable skills and experience; they settle out between roughly the age of 30 and mid 50s; and they finally decline with age above 55. At the same time, the spread of the data grows with age, which is reflected in the bootstrap estimate of the variance of the spline regression fit and in turn in the increasing band width.

Figure 6. Estimates of Canadian wage mean function, d=2 and 3.

2.5 Conclusion

We developed bootstrap simultaneous confidence band for univariate conditional mean regression. The bands rely second order polynomial regression estimates for the mean function and are appropriate for independent but not necessarily identically distributed data. Simulation results show that the band have decent coverage for moderate sample sizes and are able to attain the nominal level for larger samples. Further analysis of dependent data resulted in mixed results. The bands were fairly robust to mild departures from independence. For errors that were negatively correlated or moderately positively correlated over time, coverage was only slightly worse than for the independent case. However, coverage was very poor for both small and large sample sizes when the errors were highly positively correlated.

Initial simulation studies showed that the bands suffered from significant bias when sample sizes are small. To control bias, the initial spline fit is undersmoothed for centering of the bands. The bootstrap algorithm can easily be extended to use higher order splines, which have lower bias. As of yet, however, there are no results for the asymptotic version of the confidence bands.
Figure 7. Estimates of Canadian wage mean function, d=4 and 5.
Figure 8. Canadian wage bootstrap variance estimate.
for higher order splines, in part because the B-spline functions and corresponding analysis get more complicated as the order increases. Therefore, working with the higher order bootstrap version would require more sophisticated tools than what was applied to the linear spline case. In addition, as opposed to the asymptotic bands, it is possible to modify the bootstrap algorithm to use flexible knot selection, which in turn can improve finite sample performance as well. This again, would complicate theoretical analysis even further.

Finally, although the work here focuses on univariate regression, these bands can be useful for analysis in multivariate regression. In many multivariate semiparametric models, the nonparametric component is estimated with univariate smoothing. The bands can be used to assess the adequacy of the fit just as we illustrated in examples section.

2.6 Proofs of Results

2.7 Conditional Variance of $t_n^*(x)$.

In what follows, let $C$ denote any real valued constant, $|A|$ be the elementwise max norm for any matrix $A$. We shall use the following result from Song and Yang (2009), Lemma A.1, to derive the conditional variance and distribution of our bootstrap estimators.

**Lemma 2.7.1.** Under Assumptions (A1)-(A4), there exits a constant $C_p > 0, p \geq 1$, such that for any $m \in C^{(p)}[a,b]$ and the function $\tilde{m}_p(x),$

$$\|\tilde{m}_p(x) - m(x)\|_{\infty} \leq C_p \inf_{g \in G^{(p-2)}} \|g - m\|_{\infty} = O_p(h^p) = O_p(n^{-p/(2p+1)}).$$  (2.19)
Moreover, for the function \( \tilde{\epsilon}_p(x) \),

\[
\| \tilde{\epsilon}_p(x) \|_\infty = O_p(h^p \sqrt{\log n}) = O_p(n^{-p/(2p+1)} \sqrt{\log n}).
\] (2.20)

Let \( Z_i = (Y_i - m(X_i))^2 = \epsilon_i^2 \) and \( \tilde{Z}_i = (Y_i - \hat{m}_2(X_i))^2 = \hat{\epsilon}_i^2 \). Denote the conditional variance of \( t_n^*(x) \) as \( \sigma^2_{2,n}(x) \). Then

\[
\sigma^2_{2,n}(x) = \mathbb{E} \left( (t_n^*(x))^2 \right) = \frac{1}{n^2} \sum_{i=1}^{n} (B_2(x)^T S_n B_2(X_i))^2 \tilde{\epsilon}_i^2
\]

\[
= \frac{1}{n^2} \sum_{i=1}^{n} \omega_{i,n}^2(x) \tilde{Z}_i.
\] (2.21)

Consider the next two approximations to \( \sigma^2_{2,n}(x) \). For the first, we replace the empirical inner product matrix \( S_n \) with its theoretical version.

\[
\sigma^{2(0)*}_{2,n}(x) = \frac{1}{n^2} \sum_{i=1}^{n} (B_2(x)^T S B_2(X_i))^2 \epsilon_i^2
\]

\[
= \frac{1}{n^2} \sum_{i=1}^{n} \omega_{i,n}^2(x) Z_i.
\] (2.22)

where \( \omega_{i,n}(x) = B_2(x)^T S B_2(X_i)^T \). Second, we replace the regression residuals with the unobserved errors \( Z_i = \epsilon_i^2 \),

\[
\tilde{\sigma}^{2(0)}_{2,n}(x) = \frac{1}{n^2} \sum_{i=1}^{n} \omega_{i,n}^2(x) Z_i
\] (2.23)
Lemma 2.7.2. Under Assumptions (A1)-(A4), the following terms are uniformly bounded on 
\([a, b]\)

\[\sup_{x \in [a, b]} |B_2(x)| \leq Ch^{-1/2}; \sup_{x \in [a, b]} |\omega_{i,n}(x)| \leq Ch^{-1}; \sup_{x \in [a, b]} |\tilde{\omega}_{i,n}(x)| = O_p(h^{-1}). \tag{2.24}\]

Proof. First, note that

\[
\max_j \{\|b_{j,2}\|_2^{-1}|b_{j,2}(x)|\} \leq \left[ \min_j \{\|b_{j,2}\|_2\} \right]^{-1} = \min_j \left( 2 \int_{t_j}^{t_{j+1}} K^2(\frac{x - t_{j+1}}{h}) f(x) dx \right)^{-1/2}
\]

\[
= \left( 2h \int_0^1 (1 - u)^2 f(u) du \right)^{-1/2} \leq Ch^{-1/2}
\]

Therefore,

\[
\sup_{x \in [a, b]} |B_2(x)| \leq Ch^{-1/2}
\]

Second, for any \(x \in [a, b]\), all elements of \(B_2(x) = (B_{-1,2}(x), \ldots, B_{N,2}(x))^T\) are nonnegative, at least one element of \(B_2(x)\) is nonzero and positive, and at most two elements are nonzero and positive. Using Lemma A.2 of (Wang and Yang, 2009), the nonzero elements of \(B_2(x)\) can be bounded from below as follows

\[
\min\{B_{j,2}(x)\}^{j(x)+1}_{j=j(x)-1} \geq c\min_j \{\|b_j\|_2\}^{-1} \geq ch^{-1/2}.
\]

where the location index \(j(x) = \min\{\left\lfloor \frac{x-a}{h} \right\rfloor, N\}\)
Further, by Lemma B.2 of (Wang and Yang, 2009), there exist positive $C_p > c_p > 0$ such that in probability for any $\tau \in \mathcal{R}^{N+p}$

$$c_p|\tau| \leq |S\tau| \leq C_p|\tau|.$$  

(2.25)

The above implies that for any $i = 1, \ldots, n$ the coefficients $\{\omega_{i,n}(x)\}$ are bounded in probability on $[a,b]$ by

$$|\omega_{i,n}(x)| = |B_2(x)^T S B_2(X_i)| \leq C h^{-1/2} \|S B_2(X_i)\|_\infty \leq C h^{-1/2} C_p \| B_2(X_i)\|_\infty \leq C h^{-1}$$

Finally, note that for $p=2$, $V + \tilde{B}$ is tridiagonal and symmetric. By (7) above and Lemma B.1 of (Wang and Yang, 2009),

$$S_n = (V + \tilde{B})^{-1} = (M_{N+2} + o_p(1))^{-1},$$
where
\[
M_{N+2} = \begin{pmatrix}
1 & \frac{\sqrt{2}}{4} & 0 & \cdots & 0 \\
\frac{\sqrt{2}}{4} & 1 & \frac{1}{4} & \cdots & 0 \\
0 & \frac{1}{4} & 1 & \ddots & 0 \\
& & & \ddots & \ddots \\
& & & & 1 & \frac{1}{4} & 0 \\
& & & & \frac{1}{4} & 1 & \frac{\sqrt{2}}{4} \\
& & & & 0 & \frac{\sqrt{2}}{4} & 1
\end{pmatrix}
\]

Applying again Lemma B.2 of (Wang and Yang, 2009), there exist positive \(C_p > c_p > 0\) such that in probability for any \(\tau \in \mathbb{R}^{N+p}\)

\[
c_p |\tau| \leq |S_n \tau| \leq C_p |\tau|.
\] (2.26)

Therefore, in probability,

\[
|\tilde{\omega}_{i,n}(x)| = |B_2(x)^T S_n B_2(X_i)^T| \leq C h^{-1}, \forall x \in [a, b] \ \forall i = 1, \cdots, n
\]

\[\square\]

**Lemma 2.7.3.** Under Assumptions (A1)-(A4),

\[
\|\sigma_{2,n}^{2(0)*} - \sigma_{2,n}^{2(0)}\|_{\infty} = O_p(n^{-3/2} \sqrt{\log n})
\] (2.27)
Proof. Taking the difference,

\[ \sigma_{2,n}^{2(0)*}(x) - \sigma_{2,n}^{2(0)}(x) = \frac{1}{n^2} \sum_{i=1}^{n} \omega_{i,n}(x) (\hat{Z}_i - Z_i) \]

Each of the summands can be decomposed into three parts

\[ \hat{Z}_i - Z_i = (I_i + II_i + III_i), \]

where

\[ I_i = \{m_2(X_i) - \bar{m}_2(X_i)\}^2 + \tilde{\epsilon}_2(X_i)^2 - 2\{m_2(X_i) - \bar{m}_2(X_i)\}\tilde{\epsilon}_2(X_i) \]
\[ II_i = -2\tilde{\epsilon}_2(X_i)\sigma(X_i)\varepsilon_i \]
\[ III_i = 2\{m(X_i) - \bar{m}_2(X_i)\}\sigma(X_i)\varepsilon_i \]
By Assumption A4.a, Lemma 2.7.1 and Cauchy-Schwarz inequality,

\[ |I_i| = |\{m(X_i) - \tilde{m}_2(X_i)\}^2 + \tilde{\epsilon}_2(X_i)^2 - 2\{m(X_i) - \tilde{m}_2(X_i)\}\tilde{\epsilon}_2(X_i)| \]

\[ \leq 2\{m(X_i) - \tilde{m}_2(X_i)\}^2 + 2\tilde{\epsilon}_2(X_i)^2 \]

\[ \leq 2 \times \{||m(x) - \tilde{m}_2(x)||^2_{\infty} + ||\tilde{\epsilon}_2(x)||^2_{\infty}\} \]

\[ = O_p(h^4\log n) \]

\[ |II_i| = 2 \times |\tilde{\epsilon}_2(X_i)\sigma(X_i)\varepsilon_i| \]

\[ \leq 2 \times ||\tilde{\epsilon}_2(x)||_{\infty}\sigma(X_i)||_{\infty} \times O_p(1) \]

\[ = O_p(h^2\sqrt{\log n}) \]

\[ |III_i| = 2 \times |\{m(X_i) - \tilde{m}_2(X_i)\}\sigma(X_i)\varepsilon_i| \]

\[ \leq 2 \times ||\tilde{m}_2(x) - m(x)||_{\infty}\sigma(x)||_{\infty} \times O_p(1) \]

\[ = O_p(h^2) \]

\[ \Rightarrow |\hat{Z}_i - Z_i| = O_p(h^2\sqrt{\log n}). \] (2.28)

Thus

\[ \|\sigma_{2,n}^{(0)}(x) - \tilde{\sigma}_{2,n}^{(0)}(x)\|_{\infty} = O_p(n^{-1/2}h^{-2}n^{-1/2}h^2\sqrt{\log n}) \]

\[ = O_p(n^{-3/2}\sqrt{\log n}) \] (2.29)
Lemma 2.7.4. Under Assumptions (A1)-(A4),

\[ \| \tilde{\sigma}_{2,n}^2(0) - \sigma_{2,n}^2 \|_\infty = O_p \left( n^{-6/5} \sqrt{\log n} \right) \]  \hspace{1cm} (2.30)

Proof. Fix j,l. Write

\[ \xi_i = n^{-1} \left( B_{2,j}(X_i)B_{2,l}(X_i) \epsilon_i^2 - \sigma_{jl} \right) \]

where \( \sigma_{jl} = \int B_{2,j}(u)B_{2,l}(u)\sigma^2(u)f(u)du \)
as defined in (10). Note that \( E\xi_i = 0 \). Further,

\[
E|\xi_i|^k = n^{-k}E|B_{2,j}(X_i)B_{2,l}(X_i)\epsilon_i^2 - \sigma_{jl}|^k \leq 2^{k-1}n^{-k} \left( E|B_{2,j}(X_i)B_{2,l}(X_i)\epsilon_i^2|^k + |\sigma_{jl}|^k \right) \\
\leq \left( \frac{2}{n} \right)^k E|B_{2,j}(X_i)B_{2,l}(X_i)\epsilon_i^2|^k \leq C_0^k \left( \frac{2}{nh} \right)^k \int_{t_j}^{t_{j+1}} \int_R (\sigma(u)v)^{2k} F(du, dv) \\
\leq C_1 \left( \frac{2}{nh} \right)^k h
\]

\[
E\xi_i^2 = n^{-2}E \left( B_{2,j}(X_i)B_{2,l}(X_i)\epsilon_i^2 - \sigma_{jl} \right)^2 \\
= n^{-2} \left( \int_{t_j}^{t_{j+1}} \int_R B_{2,j}^2(u)B_{2,l}^2(u)\sigma^4(u)v^4 F(du, dv) - \left( \int_{t_j}^{t_{j+1}} \int_R B_{2,j}(u)B_{2,l}(u)\sigma^2(u)v^2 F(du, dv) \right)^2 \right) \\
\geq n^{-2} \left( c_0^4 h^{-2} \int_{t_j}^{t_{j+1}} \sigma^4(u)\mu_4(u)f(u)du - C_0^4 h^{-2} \left( \int_{t_j}^{t_{j+1}} \sigma^2(u)f(u)du \right)^2 \right) \\
\geq n^{-2} h^{-2} (c_1 h - C_1 h^2) \\
\geq C_2 \left( \frac{1}{nh} \right)^2 h
\]
This implies that one can find a constant \( \alpha \geq 1/2 \) s.t. \( E|\epsilon_i|^k \leq \alpha E\epsilon_i^2 \left( \frac{2}{nh} \right)^{k-2} k! \). Applying Bernstein’s inequality, one can find a \( \delta > 0 \), s.t.

\[
P \left( \left| \sum_{i=1}^n \epsilon_i \right| > \delta (nh)^{-1/2} \sqrt{\log n} \right) \leq 2 \exp \left( -\frac{\delta^2 (nh)^{-1} \log n/2}{2nE\epsilon_i^2 + \delta (nh)^{-1} (nh)^{-1/2} \sqrt{\log n}} \right)
\]

\[
\leq 2 \exp \left( -\frac{\delta^2 (nh)^{-1} \log n/2}{2C_2 (nh)^{-1} + \delta (nh)^{-1} (nh)^{-1/2} \sqrt{\log n}} \right)
\]

\[
= 2 \exp \left( -\frac{\delta^2 \log n/2}{2C_2 + \delta (nh)^{-1/2} \sqrt{\log n}} \right)
\]

\[
\leq 2n^{-3}.
\]

\[
\sum_{n=1}^{\infty} P \left( \sup_{-1 \leq j,l \leq N} \left| \sum_{i=1}^n \left( (B_{2,j}(X_i)B_{2,l}(X_i)\epsilon_i^2 - \sigma_{jl}) \right) \right| \geq \delta (nh)^{-1/2} \sqrt{\log n} \right) \leq 2n^{-3} N \leq \sum_{n=1}^{\infty} 2n^{-2} \leq \infty.
\]

Hence,

\[
\sup_{-1 \leq j,l \leq N} \left| \sum_{i=1}^n \epsilon_i \right| = O_p \left( (nh)^{-1/2} \sqrt{\log n} \right)
\]

Finally, this implies that
\[
\left| \hat{\sigma}_{2,n}^2(x) - \sigma_{2,n}^2(x) \right| = \left| \frac{1}{n^2} \sum_{i=1}^n B_2(x)^T S B_2(X_i) B_2(X_i)^T S B_2(x) \epsilon_i^2 - \frac{1}{n} B_2(x)^T S \Sigma S B_2(x) \right|
\]
\[
= \left| \frac{1}{n} B_2(x)^T S \left( \sum_{i=1}^n \frac{1}{n} B_2(X_i) B_2(X_i)^T \epsilon_i^2 - \Sigma \right) S B_2(x) \right|
\]
\[
\leq C n^{-1} h^{-1} (nh)^{-1/2} \sqrt{\log n} = O_p(n^{-6/5} \sqrt{\log n})
\]

**Lemma 2.7.5.** Under Assumptions (A1)-(A4),

\[
\| \sigma_{2,n}^2 - \sigma_{2,n}^{2(0)*} \|_\infty = O_p \left( n^{-3/2} \sqrt{\log n} \right) \quad (2.31)
\]

**Proof.** Write

\[
\left| \sigma_{2,n}^2(x) - \sigma_{2,n}^{2(0)*}(x) \right| = \left| n^{-2} \sum_{i=1}^n (\omega_{i,n}(x) - \tilde{\omega}_{i,n}(x)) \tilde{Z}_i \right|
\]
\[
\leq n^{-2} \left| \sum_{i=1}^n (\omega_{i,n}(x) + \tilde{\omega}_{i,n}(x)) \times (\omega_{i,n}(x) - \tilde{\omega}_{i,n}(x)) \tilde{Z}_i \right|
\]
\[
\leq C n^{-2} h^{-1} |B_2^T(x)(\hat{\alpha} - \bar{\alpha})|
\]

where \( \hat{\alpha} = \{ \hat{\alpha}_1, ..., \hat{\alpha}_N \}^T = \sum_{i=1}^n (V + \tilde{B})^{-1} B_2(X_i) \tilde{Z}_i \) and \( \bar{\alpha} = \{ \bar{\alpha}_1, ..., \bar{\alpha}_N \}^T = \sum_{i=1}^n V^{-1} B_2(X_i) \tilde{Z}_i \).

Following a similar procedure as in the proof of Lemma B.11 in Wang and Yang (2009),
we first have the equality \((V + \tilde{B})\hat{a} = V\tilde{a}\). From Lemma B.2 of the same paper, there exists a constant \(c\) such that 
\[ c|\tilde{a} - \hat{a}| \leq |V(\tilde{a} - \hat{a})| = \tilde{B} |\hat{a} - \tilde{a}| \leq |\tilde{B}| \{|\tilde{a} - \hat{a}| + |\hat{a}|\}. \]
From \((\ref{eq:lemma2.2})\), \(|\tilde{B}| = O_p(n^{-1/2}h^{-1/2}\sqrt{\log n})\). Therefore, under assumption (A4.a),

\[
n^{-1}|\tilde{a} - \hat{a}| \leq n^{-1} \left| \frac{|\tilde{B}|}{c - |\tilde{B}|} \right| \sum_{i=1}^{n} V^{-1}B_2(X_i)\tilde{Z}_i \\
\leq C_0 h^{-1/2} n^{-1/2} h^{-1/2} \sqrt{\log n} \left( n^{-1} \sum_{i=1}^{n} |\tilde{Z}_i| \right) \\
= O_p(h^{-1}n^{-1/2}\sqrt{\log n})
\]

Hence,

\[
|\sigma_{2,n}^2(x) - \sigma_{2,n}^{(0)*}(x)| \leq C n^{-2} h^{-5/2}\sqrt{\log n} \\
= O_p\left( n^{-3/2}\sqrt{\log n} \right) = O_p\left( n^{-3/2}\sqrt{\log n} \right)
\]

2.8 Conditional Distribution of \(t_n^*(x)\)

We split the proof of Theorem 2.3.1 into Lemmata 2.8.1-2.8.2 respectively.

Lemma 2.8.1. Under Assumptions (A1)-(A4),
\[
\sigma_{2,n}^{-1}(x)t_n^*(x) \implies N(0, 1).
\]

Proof. By Theorem 3.1, it is enough to show that \( \sigma_{2,n}^{-1}(x)t_n^*(x) \implies N(0, 1) \). Denote \( \xi_i^* = \frac{1}{n} \tilde{w}_{i,n}(x) \hat{\epsilon}_i \eta_i^* \). Note that \( E^* \xi_i^* = 0 \) and \( E^* |\xi_i^*|^3 < \infty \). For the linear spline, both S and \( \Sigma \) are tridiagonal nonnegative definite matrices with at most \( 3N + 4 \) nonzero entries. Therefore the variance of \( \hat{m}_2(x) \) is of the same order as

\[
\sigma_{2,n}^2(x) = n^{-1}B_2(x)^T \Sigma S B_2(x)
\]

\[
= O \left( n^{-1}h^{-3/2} \right)
\]

Further,

\[
\sigma_{2,n}^{-3}(x) \sum_{i=1}^{n} E^* |\xi_i^*|^3 = n^{-3} \sigma_{2,n}^{-3}(x) \sum_{i=1}^{n} \tilde{w}_{i,n}^3(x) |\hat{\epsilon}_i|^3 E^* |\eta_i^*|^3
\]

\[
\leq C_0 n^{-2} \left( n^{-1}h^{-3/2} \right)^{-3/2} h^{-3} \eta_0 n^{-1} \sum_{i=1}^{n} |\hat{\epsilon}_i|^3
\]

\[
\leq C_0 n^{-2} \left( n^{-1}h^{-3/2} \right)^{-3/2} h^{-3} \eta_0 C_1 n^{-1/2}
\]

\[
\leq C_2 n^{-17/20}
\]

Therefore we have that
\[
\sup_{t \in \mathbb{R}} \left| P^*[\sigma_{2,n}^{-1}(x)t_n^*(x) < t] - \Phi(t) \right| \to 0, \text{a.s.,}
\]

where \( \Phi(\cdot) \) is the standard Normal cdf. \( \square \)

**Lemma 2.8.2.** Under Assumptions (A1)-(A4), conditional on the data

\[
\|\Delta_n^* - \Delta_n\|_\infty = o_p(1), \quad (2.33)
\]

where \( \Delta_n = \sup_x |\sigma_{n,2}^{-1}(x)t_n(x)| \) and \( \Delta_n^* = \sup_x |\sigma_{n,2}^{-1}(x)t_n^*(x)| \).

*Proof.* First, for any \( x \in [a,b] \)

\[
E^*|t_n(x) - t_n^*(x)|^2 = n^{-2}E^* \left( \sum_{i=1}^n \tilde{\omega}_{n,i}(x) (\epsilon_i - \hat{\epsilon}_i \eta_i^*) \right)^2
\]

\[
\leq C_0 h^{-2} n^{-2} \left( \sum_{i=1}^n E^* (\epsilon_i^2 + \hat{\epsilon}_i^2 (\eta_i^*)^2) \right)
\]

\[
= C_0 h^{-2} n^{-2} \left( \sum_{i=1}^n \sigma_i^2(X_i) + \sum_{i=1}^n E^* \epsilon_i^2(\eta_i^*)^2 \right)
\]

\[
\leq C_0 h^{-2} n^{-1} \left( n^{-1} \sum_{i=1}^n \sigma_i^2(X_i) + n^{-1} \sum_{i=1}^n \hat{\epsilon}_i^2 \right)
\]

\[
= O_p(h^{-2}n^{-3/2}) = O_p(n^{-11/10}) \quad (2.34)
\]
Second,

\[ |t_n^*(x)| = \left| n^{-1} \sum_{i=1}^{n} \tilde{\omega}_{i,n}(x) \hat{\epsilon}_i \eta_i^* \right| \]

\[ \leq Ch^{-1} n^{-1} \sum_{i=1}^{n} |\hat{\epsilon}_i \eta_i^*| \]

Hence,

\[ h^{-2} n^{-2} E^* \left( \sum_{i=1}^{n} \hat{\epsilon}_i \eta_i^* \right)^2 = h^{-2} n^{-2} \sum_{i=1}^{n} \hat{\epsilon}_i^2 = O_p(h^{-2} n^{-3/2}) = O_p(n^{-11/10}). \tag{2.35} \]

Third,

\[ \sigma_{2,n}^2(x) = n^{-2} \sum_{i=1}^{n} \tilde{\omega}_{i,n}(x) \hat{\epsilon}_i^2 \]

\[ = O_p \left( n^{-3/2} h^{-2} \right) \]
Hence,

\[ |\sigma_{n,2}^{-1}(x) - \sigma_{n,2}^{-1}(x)| = \left| \frac{(\sigma_{n,2}^{*}(x) - \sigma_{n,2}^{2}(x))}{(\sigma_{n,2}^{*}(x)\sigma_{n,2}(x))} \right| \left( \frac{\sigma_{n,2}^{*}(x) + \sigma_{n,2}(x)}{\sigma_{n,2}(x)} \right) = O_p(n^{-6/5}\sqrt{\log n}) \left( n^{3/4}h_{n^{1/2}}h^{3/4} \right) n^{1/2}h^{3/4} = O_p(n^{-1/20}\sqrt{\log n}). \] (2.36)

Therefore by the bounds in (34)-(36) and Theorem 2.3.1,

\[ \left| \sup_x |\sigma_{n,2}^{-1}(x)t_n^{*}(x) - \sigma_{n,2}^{-1}(x)t_n(x)| \right| \leq \sup_x |\sigma_{n,2}^{-1}(x)t_n^{*}(x) - \sigma_{n,2}^{-1}(x)t_n(x)| = C_0 \left( n^{-1/20}\sqrt{\log n} n^{-11/20} + (nh^{3/2}) n^{-11/20} \right) = O_p(n^{-1/5}) \] (2.37)

\[ \Box \]

On average, 10,000 of our profit is given away in Owner discounts per Quarter. Significant effect on our bottom line—net income per quarter is now roughly close to 0. But the expansion project is EXPANSIVE (current estimate 3 Million) This severely limits our ability to seek and attain financing. We need to show lenders that the Dill Pickle is a sound financial investment. This requires both a good looking bottom line as well as showing a strong level of commitment from our Owners to think long-term.
CHAPTER 3

FREE-KNOT SPLINES FOR GENERALIZED LINEAR MODELS

3.1 Introduction

Polynomial spline regression is a popular method to estimate nonparametric models. The method consists in estimating a piecewise polynomial function of relatively low but prespecified order. Formally, let \( \tau = \{ t_j \}_{j=0}^N \) be an increasing sequence of points in \([a,b] \in \mathbb{R}\), called ‘interior knots’. A spline of order \( p \) on \([a,b]\) is a function \( g(x) \in C^{(p-1)}[a,b] \) s.t. \( g(x) = P_j(x) \) if \( x \in [t_j,t_{j+1}) \) and \( P_j(x) \) is a polynomial of order \( p \) and is differentiable at the knot points.

For conditional mean regression, if the unknown function is smooth, splines have excellent approximation and asymptotic properties ((De Boor, 2001 Rev ed), (Agarwal and Studden, 1980),(Huang, 2003)). Given proper tuning, the method is flexible enough to capture important structure of the target function without requiring many higher order terms as in usual polynomial regression. Further, splines are more efficient estimators than regular polynomial parametric models and the nonparametric kernel type regression methods. For the former comparison, spline fits are less oscillatory and possess local behaviour that is less dependent on the fit elsewhere. As for the latter, splines estimation is less computationally heavy and is easy to use for prediction.

As other nonparametric methods, the accuracy and optimality of spline regression estimation relies on careful choice of its tuning parameters. The number of knots and their placement plays
a critical role in minimizing estimator bias and variance in a balanced manner. For a spline of order $p$, the integrated mean squared error will tend to 0 asymptotically at the best possible rate of $O(n^{-2p/(2p+1)})$ if the number of knots, $N$, grows with the sample size in the order of $O(n^{1/(2p+1)})$ (Agarwal and Studden, 1980) (Lindstrom, 1999). In practice, as the number of knots determine the model dimension, a standard model selection approach can be used to choose $N$. Some popular choices rely on minimizing criterion such as AIC, BIC, or GCV to determine an appropriate number of knots (Eubank, 1999).

In addition to the number of knots, knot location is also an important factor in accuracy of spline approximations. Spline models that incorporate knot locations as additional parameters are called free-knot splines (Jupp, 1978). Properly placed knots can account locally for rapid shifts in slope and other significant changes in the target function. However, in much of the literature knot location is fixed at either uniform spacings on $[a,b]$, various percentiles of the data, or can be chosen prior to estimation by the user to account for likely local structure. This is in part due to the computational challenges involved in finding optimal solutions. Freeing up the knots, makes the objective function nonlinear. This results in many local optima that include solutions with duplicate knots ((Jupp, 1978),(Lindstrom, 1999)). This is undesirable since duplicate knots decrease the smoothness order of the fitted function. For a spline of order $p$, having $k$ identical knots at a single location induces a discontinuity of the $p - k$ order derivative at that point.

In order to prevent solutions with duplicate knots, (Jupp, 1978) first introduced what is now called the ‘Jupp Transformation’. In this approach, instead of optimizing over knot locations
directly, the algorithm searches for optimal distances between knots in such a way that solutions
with coalescent knots are less likely to be found. A more recent study in (Lindstrom, 1999)
used the Jupp transformation along with an added penalty term in the objective function that
penalized solutions that are farther away from uniform knot spacings.

Instead of direct optimization, several authors have proposed different types of approxima-
tion methods. Recently, (Molinari et al., 2004) developed an approximate solution method to
the nonlinear optimization problem. Instead of searching for a global optimum over the entire
range of the regressor \([a, b]\), the proposed algorithm searches for local optima in disjoint inter-
vals. This is done over a large number of random partitions over \([a, b]\), and the solution is
taken to be the one with the smallest mean squared error.

In contrast to trying to optimizing a nonlinear function, some authors have taken a model
selection approach to optimal knot location. These types of methods usually start with a
large number of knots distributed over the entire range of \(X\) and apply step-wise selection to
prune these down, resulting in (potentially) non uniformly spaced locations, see for example
(Kooperberg et al., 1997) and (Cox et al., 2001).

Several Bayesian methods have also been developed. Under this formulation, priors are
specified for both the number of knots and their locations. The posterior is calculated using
a Reversible Jump Markov Chain Monte Carlo, which is an MCMC algorithm specifically
designed for Bayesian model selection and averaging problems wherein the models can have
different dimensions. See for (Denison et al., 1998) and (DiMatteo et al., 2001) for a detailed
treatment.
Much of the work in free-knot splines has focused on simple linear or multivariate regression. However, nonparametric regression is very useful for estimation and inference of Generalized Linear Models (GLM). At this point, research into applying free-knot spline regression to GLM is very sparse. An study in (Bessaoud et al., 2005) considered a free-knot spline model for the logit function in a logistic regression. For a fixed number of knots and spline order, knot locations and spline coefficients are estimated uses a Quasi Newton nonlinear optimization routine. These are estimated for a series of specified knot numbers and spline orders, and the one with minimum BIC is chosen as the best fit.

In this chapter, we study the use of free-knot splines to estimate GLM regression. The main concern is to use the bootstrap to correct for finite sample bias and to construct simultaneous confidence bands. The rest of this chapter is laid out as follows. In Section 3.2, we discuss the Quasi-likelihood (QL) approach to fitting Generalized Linear Models and introduces polynomial splines for QL. In Section 3.3, we extend the model to have free-knot parameters and develop a method for estimation. We propose a penalized polynomial spline QL estimation equation for choosing optimal knot locations and estimating model parameters. To choose the optimal number of knots, we develop a step-wise selection algorithm. A bootstrap method for bias correction and inference in studied in Section 3.4. Section 3.5 presents a simulation study of the performance of the QL free-knot spline method and bootstrap procedure for small samples and Section 3.6 gives a real data example. Finally, Section 3.7 closes with a brief conclusion.
3.2 Polynomial Splines for Quasi-likelihood Estimation

In this section, we introduce polynomial spline regression for Quasi-likelihood estimation of a GLM model. Consider a generalized univariate nonparametric regression model. Suppose that the independent sample \( \{X_i, Y_i\}_{i=1}^n \) is randomly drawn from a distribution with \( E(Y_i|X_i = x) = \mu(x) \) and \( Var(Y_i|X_i = x) = \phi V(x), x \in [a, b] \). Further, assume that \( \eta(x) = g(\mu(x)) \) where \( g(\cdot) \) is a known monotone function and \( \eta(\cdot) \) is smooth. The quasi-likelihood (QL) function is defined as

\[
Q(\mu, y) = \int_y^\mu \frac{y - t}{\phi V(t)} \, dt.
\]

The unknown function \( \eta(\cdot) \) can be estimated by maximizing the empirical QL function.

\[
\hat{\eta}(x) = \arg \max_{\eta \in \mathcal{C}^{p-1}[a,b]} \sum_{i=1}^n Q \left( g^{-1}(\eta(X_i)), Y_i \right).
\]

In turn, the conditional mean estimate is \( \hat{\mu}(x) = g^{-1}(\hat{\eta}(x)) \).

The score of the QL function shares similar properties as that of the usual log likelihood (and is the main motivation for QL). Say we model the conditional mean as a function of unknown parameters \( \beta \in \mathbb{R}^d \): \( \mu(x) = \mu(x, \lambda) \). As in traditional maximum likelihood, the QL estimating equations for \( \lambda \) are obtained by differentiating Q to obtain the quasi-score function

\[
U(\lambda) = D^T V^{-1}(Y - \mu)/\phi = 0,
\]  

(3.1)

where \( D \) is a \( n \times d \) matrix of partials \( \frac{d\mu}{dx} \) and \( V = \text{diag}\{V(\mu_1), \cdots, V(\mu_n)\} \).
To define polynomial spline regression, we first fix the number of knots and locations. Let 
\( N \in \mathbb{N} \) and let \( \tau_N = \{t_j\}_{j=1}^N \) be a strictly increasing sequence of points on \([a, b] \in \mathbb{R}\), called ‘interior knots’, where
\[
a = t_0 < t_1 < \cdots < t_N < t_{N+1} = b.
\]
Divide \([a, b]\) into \((N + 1)\) subintervals \(J_0 = [a, t_1), J_j = [t_j, t_{j+1}), j = 0, ..., N - 1, J_N = [t_N, b]\). The \(\{t_j\}_{j=1}^N\) are called interior knots and can be uniform or unequally spaced. Let \(G^{(p-2)}[a, b]\) be the space of functions that are polynomials of degree \((p - 1)\) on each \(J_j\) and have \(p - 2\) continuous derivatives. It can be shown that any function in \(G^{(p-1)}\) can be uniquely represented by a linear combination of B-spline basis functions of degree \(p\) (De Boor, 2001 Rev ed).

To completely describe the \(p\)-th order B-spline basis, the set of interior knots \(\tau\) is augmented by boundary knots \(\{t_j\}^0_{1-p}\) and \(\{t_j\}^N_{N+1}\). The B-spline basis for the space of constant splines, \(G^{(-1)}\), \(\{B_{j,1}(x)\}_{j=0,...,N}\) is the set of indicator functions for the subintervals \(J_j\) where
\[
B_{j,1}(x) = I_{J_j}(x), j = 0, 1, \cdots, N.
\]
Let \(K(u) = (1 - |u|)_+\). The B-spline basis of \(G^{(0)}\), \(\{b_{j,2}(x)\}_{j=-1,...,N}\), is the set of functions where
\[
B_{j,2}(x) = K\left(\frac{x - t_{j+1}}{h}\right), j = -1, 0, ..., N.
\]
These can actually be generated from the constant spline by
\[
B_{j,2}(x) = \frac{(x - t_j)}{t_{j+1} - t_j}I_{[t_j, t_{j+1})}(x) + \frac{(t_{j+1} - x)}{t_{j+2} - t_{j+1}}I_{[t_{j+1}, t_{j+2})}(x), j = -1, \cdots, N
\]
In general, the \( p \)th order B-Spline basis can be generated from the \((p-1)\)th order basis as

\[
B_{j,p}(x) = \frac{(x - t_j)}{t_{j+p} - t_j} B_{j,p-1} + \frac{(t_{j+p+1} - x)}{t_{j+p+1} - t_{j+1}} B_{j+1,p-1}(x), \quad j = 1 - p, \ldots, N
\]

A \( p \)-th order polynomial spline estimator for \( \eta(x) \) is a linear combination of the B-spline basis function of the \( p \)-th order spline space \( G^{(p-1)} \),

\[
\hat{\eta}(x) = \sum_{j=1-p}^{N} \hat{\lambda}_{j,p} B_{j,p}(x), \quad (3.2)
\]

where the vector of coefficients \( \hat{\lambda}_N = \{ \hat{\lambda}_{1-p,p}, \ldots, \hat{\lambda}_{N,p} \}^T \) is determined by maximizing the quasi-likelihood

\[
\hat{\lambda}_N = \arg \max_{R^{N+p}} \sum_{i=1}^{n} Q \left( g^{-1} \left( \sum_{j=1-p}^{N} \lambda_{j,p} B_{j,p}(X_i) \right), Y_i \right). \quad (3.3)
\]

### 3.3 Free-knot Spline for GLM

In this section, we still fix the number of knots \( N \) but set the knots ‘free’ so to say. The idea is to make the spline fit even more data driven by including \( \tau \), the sequence of interior knots, as an additional set of parameters. Given \( N \), the solution in (Equation 3.3) changes to

\[
(\hat{\tau}, \hat{\lambda}) = \arg \max_{R^{N} \times R^{N+p}} \sum_{i=1}^{n} Q \left( g^{-1} \left( \sum_{j=1-p}^{N} \lambda_{j,p} B_{j,p}(X_i, \tau) \right), Y_i \right). \quad (3.4)
\]
When the knots are free, the corresponding quasi-score function is nonlinear in the knot locations. A general purpose optimization algorithm such as Newton-Raphson can be used to solve the system of corresponding estimating equations. However, direct optimization is complicated by the ‘lethargy’ property of free-knots. This property was studied in (Jupp, 1978), which established the following ‘Lethargy Theorem’ for a problem of the form:

$$
\min_{\tau,\lambda} F(\lambda, \tau) = \min_{\tau,\lambda} \| h - S_p(\tau, \lambda) \|_2^2
$$

where $S_p$ is the $p$-th order spline approximation to $h$ and $\| \cdot \|$ is convex and twice differentiable norm.

**Theorem 3.3.1.** (Lethargy Theorem) Let $S_N[a, b]$ be the simplex determined by $\{ \tau \in R^N, a < t_1 < \cdots < t_N < b \}$ and $S_N^{(k)}$ the $k$-th main open face of $S_N[a, b]$. $S_N^{(k)}$ is defined by the constraints $t_j - t_{j-1} > 0, j \neq k$ and $t_k = t_{k-1}$. Then on each $S_N^{(k)}$, $k = 2, \cdots, N$

$$
n_p^T \nabla F(\cdot, \tau) = 0,
$$

where $n_p$ is the unit outward normal to $S_N^{(k)}$ and $\nabla F(\cdot, \tau)$ is the gradient of $F$ wrt $\tau$.

This implies that the surface of the objective function has many stationary points with replicate knots and this result is independent of the data. Thus, any direct optimization algorithm will tend to have poor convergence, or 'lethargy', near $S_N^{(k)}$ and is likely to find solutions with replicate knots. Fitted splines based on replicate knots have lower order of smoothness, which effects the quality of approximations ((De Boor, 2001 Rev ed),(Huang, 2003)).
To mollify some of the adverse effects of the lethargy property, (Jupp, 1978) introduced what is now called the Jupp transformation of the knot location parameters. Define the Jupp parameters as

$$s_j = \log \left( \frac{h_{j+1}}{h_j} \right), \quad j = 1, 2, \ldots, N,$$

where \( h_j = t_j - t_{j-1} \). For the inverse transformation let \( p_j = e^{s_j}, j = 1, \ldots, N \), and \( Z = 1 + p_1 + p_1 p_2 + \ldots + p_1 p_2 \cdots p_N \). Then the spacings between knots are determined by \( h_1 = (b-a)/Z \) and \( h_{j+1} = p_j h_j, \quad j = 1, \ldots, N \).

Under this transformation, the new parameters \( s_1, \ldots, s_N \) are supported on all of \( \mathbb{R}^N \) and are unconstrained. Further, the points at the origin correspond to the uniform knot solution, while solutions with replicate knots (at points where \( h_j \to 0 \)) are mapped to negative and positive infinity. Therefore the usual Newton-Raphson type algorithms will converge to solutions without replicate knots.

To further improve computation, (Lindstrom, 1999) added a penalty on knot locations to the objective function that penalizes knots vectors that contain replicate and nearly replicate knots. The penalty is used in conjunction with the Jupp transformation and is shown to provide good convergence results in simulations. The paper further showed that for the specific types of penalty functions considered, the estimation does not suffer from lethargy. However, a drawback of the proposed penalty requires choosing several tuning parameters which further complicates the algorithm.

In the present work, we estimate a Generalized Linear Model using free-knot splines. In addition to \( \lambda_N \), we include \( \tau_N \) as additional free parameters in the QL objective function. For a
given $N$, we utilize the Jupp transformation and a simple penalty that does not require the use of tuning parameters. Although it is perhaps theoretically possible to use a direct method to find optimal knot locations, we found that without using the Jupp transformation most of the standard optimization routines we tried were unstable and for a majority of the time did not converge to a solution. Convergence (in terms of time and number of iterations) was further also improved by penalizing selections with close neighboring knots. Based on this, we propose the following penalized version of the Quasi-likelihood estimating equation for choosing $\lambda_N$ and $\tau_N$.

$$\left(\hat{\tau}_N, \hat{\lambda}_N\right) = \arg \max_{R^N \times R^{N+p}} \sum_{i=1}^{n} Q \left\{ g^{-1} \left( \sum_{j=1-p}^{N} \lambda_{j,N} B_{j,p} (X_i, \tau_N) \right), Y_i \right\} \cdot J(\tau_N)$$

$$= \arg \max_{R^N \times R^{N+p}} \tilde{Q}_n(\lambda_N, \tau_N). \quad (3.5)$$

where the penalty term is

$$J(\tau_N) = \exp \left( \min_j \left\{ \frac{h_j}{(b-a)} \right\} \prod_j \frac{h_j}{(b-a)} \right). \quad (3.6)$$

To choose the best number of knots, $N$, we use a lightly ‘greedy’ model selection approach based on minimizing BIC that searches among a pre-defined set of possible choices for $N$. The BIC adjusts estimates of model goodness-of-fit by penalizing models with a greater number of parameters or degrees of freedom. In GLM, the goodness-of-fit is measured by the scaled deviance which measures the change in likelihood from using the estimated model relative to that
of an unrestricted version where each fitted observation is equal to its observed value. Similarly, the Quasi-Deviance measures the difference in the quasi-likelihoods between the restricted and unrestricted models.

\[ D(Y; \hat{\mu}(x)) = 2 \sum_{i=1}^{n} \{ Q(Y_i, Y_i) - Q(\hat{\mu}(X_i), Y_i) \} = -2 \sum_{i=1}^{n} Q(\hat{\mu}(X_i), Y_i), \]

since the QL is zero at the observations. The BIC for the QL is given by

\[ BIC(N) = -2\phi \sum_{i=1}^{n} D(\hat{\eta}(X_i), Y_i) + 2 \cdot \log n \cdot (N + p) \]

In most cases, \( \phi \) is estimated from the data using Pearson’s residuals:

\[ \hat{\phi} = \frac{1}{n - (N + p) \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}(X_i))^2}{V(\hat{\mu}(X_i))}} \]

Our algorithm is initialized at some pre-specified minimum number of knots \( N_{\text{min}} \), the choice of which is discussed in the sequel. The procedure moves through models with increasing number of knots by comparing BIC values. It stops whenever the BIC values between two candidates does not diminish by at least \( \xi \) percent or the maximum number of knots, \( N_{\text{max}} \) has been reached. The ‘greediness’ of the algorithm is determined by \( \xi \), which is another value that needs to be set by the user. For our simulations, we found that values of \( 0.01 \leq \xi \leq 0.05 \) worked well in practice.

**Free-knot spline QL Algorithm**
Step 1. Fix \( p > 0, \xi \in (0, 1), 1 \leq N_{\text{min}} < N_{\text{max}} \ll n, N^* = N_{\text{min}} \).

Step 2. 
   i. Set \( N_0 = N_{\text{min}} \).
   
   ii. Calculate \((\hat{\tau}_{N_0}, \hat{\lambda}_{N_0}) = \arg\max \tilde{Q}_n(\lambda_{N_0}, \tau_{N_0})\)

   iii. \( N_1 = N_0 + 1 \).

   iv. Calculate \((\hat{\tau}_{N_1}, \hat{\lambda}_{N_1}) = \arg\max \tilde{Q}_n(\lambda_{N_1}, \tau_{N_1})\)

   v. If \( \text{BIC}(N_1)/\text{BIC}(N_0) < 1 - \xi \), then set \( N_0 = N_1 \) and go back to (iii); Else set 

   \[ N^* = N_0. \]

Step 3. Return \((N^*, \hat{\tau}_{N^*}, \hat{\lambda}_{N^*})\).

The input parameters \( N_{\text{min}} \) and \( N_{\text{max}} \) determine the range of possible model dimensions that the algorithm will visit. In our computations, we set

\[ N_{\text{max}} = \lceil (\log(n)n^{1/(2p+1)}) \rceil. \]

However, in practice, we found that in most of our simulations, using the BIC criterion, the algorithm hardly ever moved far away from \( N_{\text{min}} \). In fact, \( N_{\text{min}} \) was more crucial for fitting the data and construction of confidence bands. In general, for better fit \( N_{\text{min}} \) needs to increase with the sample size and decrease with spline order. However, higher starting values tended to increase the variance of our estimator, while smaller starting values increased bias. Although, we cannot provide a definitive answer on the best choice for \( N_{\text{min}} \), for this study setting \( N_{\text{min}} = \max \{2, \lfloor n^{1/(2p+1)} \rfloor \} \) appeared to have worked well in practice.

### 3.3.1 Binary Response Model

For the remained of this work, the QL free-knot spline procedure outlined in the previous sections is used to construct bootstrap confidence bands for a Logistic regression model, which
is most commonly applied to binary response data. In a binary response model, the data 
\{(X_i, Y_i), i = 1, \ldots, n\} are i.i.d. copies of \((X, Y)\) where \(X\) is in \(R^d\) and \(Y \in \{0, 1\}\). Conditional on \(X_i\), each \(Y_i\) has a Bernoulli distribution with parameter \(\mu(x) = E(Y|X = x) = g^{-1}(\eta(x))\), where \(g(\cdot)\) is a known link function and \(\eta(\cdot)\) is estimated. Given the data, the Quasi-likelihood of \(Y_i\) is

\[
Q(\mu_i, Y_i) = \sum_{i=1}^{n} Y_i \log \mu_i + (1 - Y_i) \log (1 - \mu_i),
\]

where \(\mu_i = g^{-1}(\eta(X_i))\). In binary logistic regression, the link function is set to be the logit function: \(g(x) = \logit(\mu(x)) = \log \left\{ \frac{\mu(x)}{1-\mu(x)} \right\} \); while its inverse is the expit function: \(\mu(x) = \expit(\eta(x)) = \frac{\exp\{\eta(x)\}}{1+\exp\{\eta(x)\}}\).

As discussed in the introduction, we model \(\eta(x)\) nonparametrically via free-knot spline approximation. Therefore, we assume that \(\eta(\cdot)\) is an unknown smooth function. The penalized QL function then becomes

\[
\hat{\tau}_N, \hat{\lambda}_N = \arg \max_{R_N \times R_{N+p}} \sum_{i=1}^{n} Q \left\{ \frac{\exp\{\lambda_N^T B_n(X_i)\}}{1 + \exp\{\lambda_N^T B_n(X_i)\}}, Y_i \right\} \cdot J(\tau_N),
\]

where \(B_n(X_i) = \{B_{1-p,p}(X_i, \tau_N), \ldots, B_{N,p}(X_i, \tau_N)\}\).

3.4 Bootstrap for Binary Response Model

The bootstrap is a popular nonparametric method to approximate sampling distributions of statistics. If done correctly, it can be used to estimate the bias, variance, and quantiles in cases where the distribution of the statistic of interest is unavailable or difficult to calculate. In
this work, we bootstrap the free-knot polynomial spline regression estimator to adjust for finite
sample bias and to construct simultaneous confidence bands. Define the following statistics
corresponding to the GLM spline estimator variance and maximal deviation, respectively.

\[ T_n^{(v)}(x) = \text{Var}(\hat{\eta}(x)) \]

\[ T_n^{(M)} = \sup_{x \in [a,b]} \left| \frac{\hat{\eta}(x) - \eta(x)}{\sqrt{\text{Var}(\hat{\eta}(x))}} \right|^{1/2} \]

We follow (Härdle et al., 2004) in applying a parametric bootstrap to estimate the pa-
rameters of the Binary Response Model. In this case, each bootstrap observation is sampled
from a Bernoulli distribution with mean \( \hat{\mu}(X_i) \). For a more general model with a conditional
variance assumption, the Wild bootstrap is applicable. For applications of the Wild bootstrap
for nonparametric regression see (Härdle and Mammen, 1993), (Neumann and Polzehl, 1998)
,(Claeskens and Keilegom, 2003), and (Härdle et al., 2004).

**Bootstrap Algorithm for Binary Response QL**

Step 1. Fix \( p > 0, \xi \in (0,1), 1 \leq N_{\text{min}} < N_{\text{max}} < < n, B = \text{some large integer}. \)

Step 2. For the original sample \( \{X_i, Y_i\} \), estimate the regression model using the Free-knot
Spline QL Algorithm in Section 3.3 and calculate the predicted values \( \{\hat{\mu}(X_i)\}_{i=1}^N \).

Step 3. For each \( b = 1, 2, \ldots, B \)

   i. Generate a bootstrap sample via \( Y_i^* \sim \text{Bernoulli} (\hat{\mu}(X_i)), i = 1, \ldots, n. \)
ii. For the bootstrap sample \( \{X_i, Y_i^*\} \), estimate the regression model using the Free-knot Spline QL Algorithm in Section 3.3 to obtain the bootstrap estimators \( \{\eta_i^*(\cdot), \mu_i^*(\cdot)\} \).

iii. Calculate the bootstrap statistic of interest \( T_{b,n}^{(v)*} \) or \( T_{b,n}^{(M)*} \).

Step 4. Return the bootstrap sample \( \{T_{1,n}^{(q)*}, T_{2,n}^{(q)*}, \ldots, T_{B,n}^{(q)*}\} \) where \( q \) is either \( v \) or \( M \).

Note, that to generate the bootstrap sample of the maximal deviation \( \{T_{b,n}^{(M)*}\} \), the bootstrap variance estimate needs to be calculated first. Therefore, this procedure is actually a double bootstrap algorithm.

3.4.1 Confidence Bands

Rewrite the maximal deviation statistics as

\[
T_n^{(M)} = \sup_{x \in [a,b]} \left| \hat{\eta}(x) - \eta(x) \right| \left( \frac{T_n^{(v)}(x)}{\left[ T_n^{(v)}(x) \right]} \right)^{1/2}
\]

An exact simultaneous confidence band for \( \eta(x) \) is determined by

\[
P \left\{ T_n^{(M)} \leq \theta_\alpha \right\} = P \left\{ \eta(x) \in \hat{\eta}(x) \pm \theta_\alpha \left\{ T_n^{(v)}(x) \right\}^{1/2}, \forall x \in [a,b] \right\} = 1 - \alpha, \quad (3.7)
\]

where \( \theta_\alpha \) is the \( (1 - \alpha) \)th quantile of the distribution of \( T_n^{(M)} \). Since the exact finite sample distribution of the maximal deviation is unknown and beyond the scope of this project to study, we estimate it using the bootstrap.
Let $E^*$ denote the conditional expectation given the original sample and predicted values \( \{X_i, Y_i, \hat{\mu}(X_i)\}_{i=1}^n \). The bootstrap estimates for the variance and maximal deviation are given by

\[
T_n^{(v)*}(x) = \frac{1}{(B-1)} \sum_{b=1}^{B} (\eta_b^*(x) - E^*(\eta^*(x)))^2 \\
T_n^{(M)*} = \sup_{x \in [a,b]} \left| \frac{\eta^*(x) - \hat{\eta}(x)}{\left\{ T_n^{(v)*}(x) \right\}^{1/2}} \right|
\]

The Wild bootstrap \((1 - \alpha)\) simultaneous confidence bands for \( \eta(x) \) are then

\[
\left\{ \eta(x) \in \eta^*(x) \pm \theta_\alpha^* \left\{ T_n^{(v)*}(x) \right\}^{1/2} \right\}, \quad (3.8)
\]

where \( \theta_\alpha^* \) is the \((1 - \alpha)\) quantile of \( \left\{ T_n^{(M)*}, \ldots, T_n^{(M)*} \right\} \).

### 3.5 A Simulation Study

In this section, we present a simulation study of free-knot splines for QL and corresponding bootstrap confidence bands. The data is generated from a binary response model with

\[
E(Y|X = x) = \mu(x) = \expit(\eta(x)),
\]

where \( \eta(x) = x + 2 \times \exp(-16 \times (x - 0.5)^2) \). The variable X is independently drawn from a uniform \( U[-2,2] \) distribution. For all calculations we set the greediness parameter \( \xi = 0.01 \).

First, we compare the performance and ‘look’ of the free-knot estimator versus the fixed. For the fixed knots, we keep the same structure of the QL and bootstrap algorithms, but the
knots are kept fixed and uniform. First, we study how each does to fit the model with \( n = 250 \). Figures 9-10 present results for the second and third order splines respectively for one random draw of the X vectors. In general, the free-knot fit looks like a better approximation. For both orders, one of the optimal knots is set close to the bump (it is actually right on the money of the second order spline); but this is impossible to do with a uniform arrangement. Further, for the higher order spline fit, the uniform arrangement results in a very poor fit relative to the free-knot. The free knot estimate does a very good job of capturing the general shape of the unknown mean function and its logit transformation.

Now, we bootstrap each version of the QL estimator, free-knot and fixed, using a relatively small bootstrap sample size of \( B = 100 \). These are then illustrated in Figures 11-14. There are several different components to look at but the story is telling. The plots in Figures 11 and 13 give what we are calling ‘a bootstrap snapshot’ of each procedure that plot every bootstrap estimate of \( \eta^*(x) \) as well as corresponding bootstrap estimates of its standard deviation. The resulting estimates of 95% confidence bands are provided in Figures 12 and 13.

Let’s start with the results for the free-knot spline that are illustrated in Figures 11 and 12. The various components in the top panel of Figure 11 are as follows: the central solid thick black line is the original but unknown \( \eta(x) \); the set of gray lines is every estimated function \( \eta_b^* \) using free-knot splines; the central red dashed line is the estimate of \( E^*(\eta^*) \); and finally the orange vertical lines are the knot locations chosen across all of the bootstrap samples. As can be seen in this and the following panel of the estimated standard deviation, the set of optimal knot locations varies widely across the range of the observed X. However, they do tend
to cluster close to the changepoints of the bump or close to the end points of the interval. The latter leads to several issues which make this version of the bootstrap procedure a poor candidate for inference, or at least the type of inference procedures we are after. Namely, both the bootstrap estimates of the mean (red dashed line) and the standard deviation near the boundary are too large in magnitude, making it very hard to glean any important structure in those areas. In particular, this would give confidence bands that are too wide on the boundary, not very informative, and unappealing as a visualization technique.

An actual example of the bands is plotted in Figure 12, with the vertical axis cropped to show the estimated function. Although these bands cover the function perfectly, and do give a nice idea of its shape in the central part of the range, the bands explode close to the endpoints on either side. Now, these do reflect the variability in the estimation technique, namely the free-knot spline QL, but the goal is also to know something about the actual function as well. Therefore, the boundary issue requires a bit more massaging of the algorithm to make the bands more visually appealing.

In comparison, the bootstrap snapshots of the uniform knot bands in Figures 13 through 14 show a much less volatile estimator whose bootstrap average is close to the true function. There is a little variability in knot locations stemming from different knot number choices; but because of the uniform assumption the resulting locations do not move as much as in the free-knot version. There is still somewhat of a boundary issue, the estimated standard deviation does increase at both the endpoints; but the variance is much lower than the free-knot version.
The resulting confidence bands cover the function over the entire range of the observed x and
more importantly are good candidates for use to test model fit in practice.

The conclusion is somewhat different for the quadratic splines with $p = 3$. A similar set
of comparison plots for these is given in Figures 15 to 18. There is still the same pattern
in variability and bootstrap moment estimates for the free-knot spline; but it looks to give a
much better estimate of the function as well as better band coverage. The fixed knot version
now greatly underestimates the bump (it actually misses it for the most part) and the resulting
bands do not cover the function in this region. Still the free-knot bands suffer from the same
explosion on the boundaries as well as a spike in the interior region.

To test the coverage probabilities of free versus fixed knots, we ran 100 simulations of the
model with $n = 250$. Table V presents the empirical coverage rates and average band area for
the second and third order splines. Overall, both versions of the band for $p = 2$ had coverage
rates close to or higher than the nominal. However, the average area (the area within the
band) is much larger for the more volatile free-knot bands; but is reasonable for the fixed knot
bands. This further corroborates the conclusion the free-knots are much too variable to use for
confidence bands in applied data analysis. In contrast, the higher order $p = 3$ bands had much
worse coverage rates overall, albeit much narrower band areas. The smaller areas comes from
the lower variability induced in the estimates from using a lower number of knots. This suggests
that for the bump function, a higher number of knots is required. Of course, in practice, the
ture function is never known, and therefore the bands probably require some further tuning
given the apparent shape of the observed data.
TABLE V

COVERAGE RATES OF BOOTSTRAP CONFIDENCE BANDS, FREE VS. FIXED

<table>
<thead>
<tr>
<th></th>
<th>Free-knot</th>
<th></th>
<th>Fixed</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coverage</td>
<td>Area</td>
<td>Coverage</td>
<td>Area</td>
</tr>
<tr>
<td>p=2</td>
<td>0.95</td>
<td>0.92</td>
<td>148.01</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>0.99</td>
<td>1.00</td>
<td>329.4</td>
<td>1.00</td>
</tr>
<tr>
<td>p=3</td>
<td>0.95</td>
<td>0.65</td>
<td>77.69</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>0.99</td>
<td>0.33</td>
<td>52.34</td>
<td>0.19</td>
</tr>
</tbody>
</table>

3.6  **Real Example: Horseshoe Crab Data**

In this section we apply our QL procedure to estimate a logistic regression model using the horseshoe crab dataset from (Agresti, 2007). The data consists of 173 observations of nesting female horseshoe crabs. According to horseshoe crab experts, female horseshoe crabs live in quite interesting domestic arrangements, called spawning clusters. A typical female has one male that resides in her nest; and in addition, she may have other males, called satellites, that reside nearby. Behavioural research into these nesting clusters showed that the satellites are preferentially attracted to some females over others. In particular, (Brockmann, 1996) observed that "females with many satellites were larger and in better condition, but did not lay more eggs, than females with few or no satellites."

We fit a logistic model to determine the effect of the female crab’s weight on the presence or absence of male satellites. In addition, we estimate confidence bands to test whether the unknown function $\eta(x)$ is linear using the uniform knot version of the algorithm.
The estimated probability curves \( E(Y|X) \) are given in Figure 19 for spline fits of order \( p = 2 \) and \( p = 3 \). In both plots, the free-knot fits are given in solid black, fixed knot estimates in dashed blue, free-knot locations in vertical orange, and uniform knots in vertical blue. Both free and fixed uniform versions are represented, but the algorithm settled on the same uniform knot location for \( p = 2 \); while for \( p = 3 \) it settled two very close knots near the upper boundary of the weight ranges. Each fit tells the same story essentially however, the probability of having satellites increases almost linearly with weight up to roughly 30; after which point it is almost certainly that a female will have at least one satellite male.

Next, in Figure 20, a set of uniform 95% confidence bands is plotted, along with two parametric model fits. The first is an estimated line for the simple one variable linear logistic model: \( \eta(X) = \alpha + \beta X \) and the second is for a 6th degree polynomial fit. The latter is included because it is the lowest order polynomial model that lies completely inside the bands. The linear fit is in solid red; the bands are in dashed periwinkle; and the center dashed line is the second order spline fit. It is fairly clear that the simple linear model is a poor fit for the unknown link function, it misses completely the rapid shift in slope at weight 29.5. In fact, the lowest order polynomial that would fit entirely within the bands is a 6th degree polynomial, which requires estimating seven parameters versus the four with the linear spline. However, the overall impact in estimated probability is less stark. As illustrated in Figure 21, when all three estimates of the link function are transformed back to the conditional mean, all curves are close to each other. Of course prediction accuracy depends on the distribution of female weights.
3.7 Conclusion

This chapter presented a computational study of bootstrap confidence bands based on free-knot splines for the Generalized Linear Model. In free-knot spline regression, the knot locations of the usual polynomial splines enter into the model as additional parameters that need to be estimated from the data. This offers even greater flexibility and the potential to better account for rapid shifts in slope and other significant changes in the target function. However, in ‘freeing’ up the knots, the search for optimal solutions becomes very complicated. The most challenging part of free-knot computation is the existence of the ‘lethargy’ property of the objective function that causes many local optima with coalescent knot solutions (ones in which two or more knots are the same). To prevent solutions with identical knots, a penalized Quasi-likelihood estimating equation and a model selection algorithm is proposed that relies on both the Jupp transformation of knot locations and an added penalty on solutions with small minimal distances between knots. Focusing on logistic regression for binary outcome data, a parametric bootstrap is used to control for bias and to construct confidence bands for the unknown link function.

Both procedures were applied to test the usual linearity assumption of the link function in logistic regression. Simulation studies showed that regression free-knots yielded better fits than uniform knots. However, an analysis of bootstrap samples showed that ‘freeing’ up the knots adds a great deal more noise to the estimates relative to one with fixed uniform spacings. Optimal knot locations vary greatly over the range of the observed x and have the unfortunate tendency to locate close to the boundary. The latter tendency makes the bootstrap estimates
of the free-knot estimator variance explode close to the end points of the interval. This in turn makes the resulting confidence bands too wide on the perimeters of the range, which in the end obfuscates any information about the unknown function. Therefore, the bands can be used to perhaps tune the free-knot estimator to make it less variable; but are not a good candidate for use in applications to test a parametric model fit; at least at the end points. This is borne out with simulation of confidence band coverage rates, which show that although the free-knot bands have coverage close to the nominal or even higher, they are much too wide to yield any information about \( \eta(x) \). In comparison, bands based on fixed and uniformly placed knots show the same coverage rates but were much narrower on average.

Future work would include a more comprehensive simulation study with other models such as the Poisson as well as theoretical treatment of the validity of bootstrap estimates. As the QL with an identity link function is tantamount to estimating a simple conditional mean regression, the free-knot based bootstrap confidence bands extend to the model studied in Chapter 1 as well. Finally, this computational study indicates that free-knots offer greater flexibility, but come with a much higher computational cost, which is especially burdensome when the estimators have to be resampled a large number of times with the bootstrap. Further research is warranted into developing a more efficient algorithm to estimate QL free-knot spline regression that would make it a workable solution for developing bootstrap inference procedures for GLM components. In addition, further work is required in developing an optimal penalty or set of penalties to minimize estimator variance and/or to adjust for the band width on the boundary.
Figure 9. Free-knot and Fixed Spline Estimates, $p=2$
Figure 10. Free-knot and Fixed Spline Estimates, $p=3$
Figure 11. Free-knot Bootstrap Snapshot, p=2
Figure 12. Free-knot Bootstrap Confidence Bands, p=2
Figure 13. Uniform Knot Bootstrap Snapshot, p=2
Figure 14. Uniform Bootstrap Snapshot, Standard Variation, p=2
Figure 15. Free-knot Bootstrap Snapshot, p=3
Figure 16. Free-knot Bootstrap Confidence Bands, p=3
Figure 17. Uniform Bootstrap Snapshot, p=3
Figure 18. Uniform Knots Bootstrap Confidence Bands, p=3
Figure 19. Probability of Satellites vs Weight, Free and Uniform Knots
Figure 20. Uniform Bands for Crabs Link Function, p=2
Figure 21. Linear Model and Uniform Spline Estimates of Satellites Probability
CHAPTER 4

CONDITIONAL QUANTILE REGRESSION MODELS OF MELANOMA
TUMOR GROWTH CURVES

4.1 Introduction

Tumor growth curves provide a simple way to understand how tumors change over time and are a useful statistical tool in the development of successful cancer treatments. The traditional approach to fitting such curves to empirical data has been to estimate conditional mean regression functions which describe the average effect of covariates on growth. However, this method ignores the possibility that tumor growth dynamics are different for different quantiles or percentiles of the possible distribution of growth patterns. In contrast, other areas of growth analysis such as that of human children, individual growth trajectories are routinely measured against a standard group of reference curves that range over several quantiles. This type of comparison is more informative as it gives health care professionals a better understanding of where their patient is on the spectrum of 'normal' and 'abnormal' growth after controlling for important factors such as health history and family characteristics. This suggests that the conditional quantile growth curve can be an invaluable tool in the personalized approach to medical treatment and research, wherein treatments are customized to the individual in order to increase treatment response and minimize side effects.
In this paper, we take the idea of quantile reference growth curves and apply it to modeling tumor volume growth. In addition to estimating quantile curves, our work also incorporates a novel use of small sample data. In general, study designs with too few animals tend to have lower power to detect a statistically significant difference in tumor volume between treatment groups. To increase power, we pool information across several independent studies by combining the samples together and running our fitting procedure on the combined sample. This research is motivated by a series of small sample studies that investigated the effects of an experimental treatment, P28, on melanoma tumor volumes in mice. The studies differ in terms of sample size, the set of experimental treatment regimes, and melanoma cell line. However, the experimental treatments do overlap between studies, including one dosage of P28 that is included in every study. Our goal is to ascertain what effect treatment has on tumor growth trajectories for different quantiles of the population controlling for study and individual subject effects.

To construct the tumor reference growth curves, we specify a longitudinal regression model for multiple quantiles that incorporates individual subject and study fixed effects. Quantile regression methods (Koenker and Bassett Jr, 1978) extend traditional regression models for the conditional mean to conditional quantiles. Typical regression models posit a functional relationship between the average population response and a set of predictors X: $E(Y|X) = f(X, \theta)$. For example, according to the popular linear model of the conditional mean: $E(Y|X = x) = \beta_0 + x'\beta$, the population average changes linearly as a function of x. However, even for the simple linear case, it is not unrealistic to assume that the coefficients will be significantly different for subjects at the extremes of a population and those that resemble the population
average. Quantile regression (Koenker, 2005) directly models the $\tau^{th}$ population quantile as a function of covariates. This allows researchers to assess the quantile effect of various variables on the response function and that is impossible to do with the traditional conditional mean model. Furthermore, quantile regression estimates are more robust than the traditional conditional mean regression estimates, the latter being in general more sensitive to outliers and less accurate and less informative for data arising from skewed or heavy tailed distributions (See (Koenker, 2005) for a good introduction).

In a longitudinal study, a random sample of individual subjects is repeatedly measured at different points in time. Although the subjects are independent, usually they are not homogeneous, and care must be taken in controlling for the individual idiosyncratic factors. Finally, observations within subjects can also be correlated, which although does not effect the consistency and robustness of parameter estimates ((Wei and He, 2006)), it may potentially effect efficiency. The aim of longitudinal analysis is to tease out the general population level response from the individual subject deviations and possible correlations. Several different approaches exist in the literature on applying the quantile regression method to longitudinal data.

Much work exists on methods that control for within subject correlation and attempt to minimize efficiency of parameter estimates. Some proposed methods borrow from the generalized linear model (GLM) approach to conditional mean regression by specifying a parametric or semiparametric model for the correlation structure of either the responses or of the estimating equations used to fit models. These include a generalized estimating equations (GEE) type model formulation, (see for example ((Jung, 1996), (Lipsitz et al., 1997), (Fu and Wang,
and a quadratic inference function (QIF) type method (Leng and Zhang, 2012). In a similar spirit to linear mixed effects models, (Geraci and Bottai, 2007) and (Liu and Bottai, 2009) formulated a mixed effects model quantile regression model that incorporate subject specific random effects. In this approach, the responses are assumed to come from an asymmetric Laplace distribution that depends on the target quantile to be estimated. Further the random effects are assumed to follow a multivariate distribution that needs to be specified. Alternatively, (Karlsson, 2008) accounts for any possible within subject correlation by using a weighted sum of the usual check function, where the weights are estimated from the correlation of residuals from an initial unweighted median regression estimate.

As we are only concerned with illustrating how the quantile approach can be applied to model tumor growth, considerations on parameter efficiency are outside the scope of the current work. We do, however, aim to construct reference growth curves and thus require a method that can be used to estimate a set of different quantile regressions simultaneously. To the best of our knowledge, the proposed models described above are limited to estimating a single quantile at a time and most require specifying either a distribution for the errors or a correlation matrix for the regression estimating equations.

A fully distribution free method that can be used to model multiple quantiles simultaneously is available in (Koenker, 2004). Called Penalized Fixed Effects Quantile Regression (PFEQR), this method extends the original quantile regression model to longitudinal data by incorporating fixed subject specific model parameters that are penalized to prevent overfitting and decrease model dimension. This method was applied in (Wei and He, 2006) and (Wei et al., 2006) to
estimate a semiparametric quantile regression model for children's conditional growth charts that includes autoregressive terms. In the former paper, this model is shown to be robust for deviations from an IID error structure wherein there is no within subject correlation. Further, and of great importance for researchers, the empirical results in (Wei et al., 2006) showed that the magnitude of the relationship between current height and past height varies at different points of the conditional distribution of height. In particular, infants at the lower tail of the height distribution had much higher estimated AR coefficients than those in the upper tails—a result that is consistent with a catch-up hypothesis for very small infants but is impossible to test with a conventional conditional mean model.

In our work, we illustrate the potential of quantile regression to study tumor growth. We use PFEQR to estimate a set of quantile exponential growth curves of tumor volume from mice grafted with human melanoma cells. The data arises from multiple independent studies with small sample sizes. We aim to improve the variability of the estimates by combining the data for estimation. The rest of the paper is laid as follows. In Section 2, we give an overview of the PFEQR method for estimating simultaneous quantile regression curves from longitudinal and provide some discussion on tuning parameter choice. In Section 3, we present a quantile regression analysis of treatment efficacy for melanoma tumor growth using an exponential quantile model with fixed study and fixed subject effects.

4.2 Penalized Fixed Effects Quantile Regression for Longitudinal Data

The quantile tumor growth curves presented in this paper rely on small sample longitudinal studies. Longitudinal data or repeated measurements data is one of the most common types
of data collected in pharmacological research. In such studies, samples of individual subjects are followed over time to assess the efficacy of experimental treatments. More formally, let $Y_{ij}, i = 1, ..., N, j = 1, ..., t_i$ be the $i^{th}$ subject's response at time $j$ and $x_{ij}$ a p element vector of observed covariates. Assume that each $Y_{ij}$ is distributed according to some unknown cumulative distribution function $F(x)$, whose form may depend on a set of covariates $x \in \mathbb{R}^p$. Following (Koenker, 2005), a linear conditional quantile function can then be defined as

$$Q_{Y_{ij}}(\tau | x_{ij}) = x_{ij}^T \beta(\tau),$$

(4.1)

where $\tau \in (0, 1), Q_{Y_{ij}}(\cdot) = F_{Y_{ij}}^{-1}$, and $\beta \in \mathbb{R}^p$ is unknown and depends on $\tau$.

Note that if there is no between-subject heterogeneity, then $F$ is the same for all $Y_{ij}$. This is not a realistic assumption for longitudinal data, wherein different subjects have their own idiosyncratic responses to the same experimental conditions. Moreover, although subjects may be independent of each other, observations within subjects tend to be correlated. The aim of longitudinal analysis is to tease out the general population level response from the individual subject deviations. A popular approach for the conditional mean models is the use of mixed models. These include subject specific fixed and random effects that account for distributional differences as well as within subject correlation. For quantile regression, an analogous approach has been introduced in (Koenker, 2004). The specification in (1) is augmented as

$$Q(Y_{ij}|\tau, x) = \alpha_i + x_{ij}^T \beta(\tau),$$

(4.2)
where $\alpha_i$ is the $i^{th}$ subject’s individual fixed effect. Arguing from the fact that estimating the usual conditional mean random effects mixed model is equivalent to estimating a penalized fixed effects regression, the author in (Koenker, 2004) proposed an analogous estimating procedure for quantile longitudinal regression. Define the piecewise linear ”check function” $\rho_\tau(u) = (\tau - 1\{u < 0\})u$. Further let $(\tau_1, ..., \tau_K)$ be the quantiles of interest, $(\omega_1, ..., \omega_K)$ weights that control the influence of individual quantile curve estimates on the overall solution, and $\lambda$ a shrinkage parameter that controls the size of the estimated subject specific fixed effects. Then the estimates to (2) are defined as solutions to the following regularization problem that imposes a lasso penalty on the subject fixed effects coefficients:

$$
\min_{\beta \in \mathbb{R}^K} \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{j=1}^{t_i} \omega_k \rho_{\tau_k}(y_{ij} - \alpha_i - x_{ij}^T \beta(\tau_k)) + \lambda \sum_{i=1}^{N} |\alpha_i|.
$$

Both the weights and $\lambda$ must be specified before estimation. Following the suggestion in (Koenker, 2004), in this work we use symmetric weights that are higher for central quantiles and decrease at the upper and lower tails. In particular, to estimate the five quantiles $\tau = (0.05, 0.25, 0.50, 0.75, 0.95)$ we use $w = (0.050, 0.225, 0.450, 0.225, 0.050)$.

Under the model in (2) the fixed effects do not depend on $\tau$ and account for a subject specific location shift. It is possible to model the $\alpha_i$’s as functions of $\tau$, which would account for subject specific distributional shifts. However, even the simpler ‘location’ fixed effects parametrization in (2) includes $N$ extra parameters, which if estimable, tends to increase the
variance of estimates. Accounting for distributional subject specific effects via fixed effects would reduce the efficiency of estimates even further if not make the estimation intractable.

4.2.1 Selection of tuning parameter $\lambda$

For the choice of the shrinkage parameter, several methods have been proposed in the literature for penalized quantile regression of non-correlated and single time series data but little has been done for longitudinal data specifically. In general, shrinkage or regularization parameters are chosen to minimize some sort of optimality criterion that are functions of model accuracy and prediction error. Under the assumption of random subject specific effects, (Lamarche, 2010) derived a form for the tuning parameter that minimizes the asymptotic variance of the estimated quantile fixed effects. In that work, the optimal value of $\lambda$ depends on plug-in estimators of both the density of the random subject effects and the sparsity function of the data. A more simpler approach relies on the Schwartz Information Criterion (SIC) to choose the tuning parameter. Originally proposed in (Koenker et al., 1994) for penalized quantile regression models for non-longitudinal data, this was further studied in (Li and Zhu, 2008) for linear quantile regression models with many predictors and in (Li et al., 2007) for nonparametric quantile regression. However to our knowledge, no rigorous work has been done in using the SIC for the penalized panel quantile regression method.

In addition, nonparametric methods such as the bootstrap (Hastie et al., 2009) are often used in penalty selection for nonparametric regression. However, their use is complicated for the longitudinal data set-up. Standard selection methods usually rely on 'accurate' estimates of out-of-sample or the testing sample model prediction error and therefore require out-of-
sample prediction of subject-specific fixed effects. In addition, the within subject correlation structure precludes the use of simple cross-validation techniques and simple bootstrapping of observations across time. Some work in this area has been done in (Kapetanios, 2008) on panel data bootstrapping but the study did not examine applications to model selection. One alternative can be to take a finite sample approach to the asymptotic selection criterion of (Lamarche, 2010) and take $\lambda$ as a minimizer of a bootstrap estimate of (total) parameter variance. Possible resampling schemes can be the cross-sectional subject panel bootstrap of (Kapetanios, 2008) or the wild bootstrap as described in (Feng et al., 2011). However, the study of bootstrap and other model selection criteria methods for panel quantile regression is beyond the scope of this paper and we present results based on a SIC type approach. We do use the cross-sectional subject bootstrap studied in (Kapetanios, 2008) for model inference in the Section 3 of this paper.

For the analysis of the melanoma data presented here, we chose the SIC to set the penalty term. The SIC measures relative goodness of fit and prediction accuracy. Define $\hat{f}(x, \tau_k, \lambda)$ as the $\tau_k^{th}$ conditional quantile estimate from a model fitted with penalty $\lambda$. Let the ‘elbow’ set be the set of all zero-value residuals:

$$ g_\lambda = \{ \hat{e}_{ijk} \mid \hat{e}_{ijk} = y_{ij} - \hat{f}(x_{ij}, \tau_k, \lambda) = 0 \}. $$

Then in the context of longitudinal quantile regression, the SIC is defined as
\[ SIC(\lambda) = \log \left( \frac{1}{KT} \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{j=1}^{t_i} \rho_{\tau_k}(y_{ij} - \hat{f}(x_{ij}, \tau_k, \lambda)) \right) + |g_\lambda|, \]

where \( T = \sum_{i=1}^{N} t_i \) is the total number of observations and \( |g_\lambda| \) is the number of elements in \( g_\lambda \). Lower values of SIC indicate better model fit and lower prediction error. A model selection procedure would choose the model fit with \( \lambda^* \) where

\[ \lambda^* = \arg \min_{\lambda} SIC(\lambda) \]

In the quantile regression definition of \( SIC(\lambda) \), \( |g_\lambda| \) is a measure of the effective degrees of freedom (edf) of the regression fit. In ordinary linear regression, the degrees of freedom is equal to the number of parameters, and in general is used to quantify the complexity of a fitting procedure. Overly complex and overfitted models tend to have worse prediction accuracy and higher parameter variances. The SIC penalizes model complexity through the edf term. (Li et al., 2007) and (Li and Zhu, 2008) showed that the size of the elbow set \( |g_\lambda| \) is a valid estimate for quantile regression edf. However, in our experience with modeling the tumor data, \( |g_\lambda| \) proved to be zero or very small for most values of the penalty parameter (we tried \( \lambda \) ranging from 0 to 10000). Therefore, we modified the criterion for the elbow set to be

\[ g_\lambda = \{ \hat{e}_{ijk} \mid |\hat{e}_{ijk}| = |y_{ij} - \hat{f}(x_{ij}, \tau_k, \lambda)| \leq \epsilon \}, \]
where $\epsilon$ is a very small positive number. We ran the model selection for values of $\epsilon = 0, 10^{-16}, 10^{-14}, 10^{-12}$, on an equally spaced grid of $\lambda$ values. For the grid spacing, we looked at interval sizes equal to 1, 0.5, and 0.1. In all cases, the optimal $\lambda^*$ was small ranging from 0.3 to 2.5. The estimation results presented here are based on $\lambda = 1$.

4.3 Modeling Quantile Tumor Growth Curves with Combined Studies Data

4.3.1 Data

The quantile growth curve method is illustrated using data from xenograft mouse model studies of P28. As opposed to conventional chemotherapies, which destroy cancer and noncancer cells alike, P28 is a novel chemotherapy treatment that preferentially enters a variety of human cancer cells and increases their level of the tumor suppressor protein P53. The latter is an important factor in the body’s natural anti-cancer arsenal– it works by inhibiting cell growth and triggering cell death. Its deficiency has been associated with increased cancer incidence in humans and mice. For more on P53 and melanoma see (Kichina et al., 2003) for a technical treatment.

For our analysis here, we use data from six independent studies that tested P28 on immunosuppressed mice grafted with human melanoma (MEL) cells that had the P53 gene. The data is comprised of 4 weeks of semi-daily measurement of tumor height, length, and width for each mouse. For each study, the mice were randomly assigned to one of several different chemotherapies, including different doses of P28, which were administered on a daily or semi-daily (3 times per week) basis. The studies differ in terms of sample size, the set of experimental treatments, and melanoma cell line. However, the experimental treatments do overlap between
studies, including one dosage of P28 that shows up in every design. A summary of the subject counts by study and treatment groups is found in Table VI. In addition to P28, different studies included variations to P28– Azurin which was a precursor to P28, and P28 PEG which a biochemical variant with different physiological effects. Finally, five of the six studies included DTIC, which currently is the only FDA approved chemotherapy treatment for melanoma. Due to the mechanism by which P28 fights melanoma, it is potentially less toxic than the standard treatment, and therefore one question for analysis is whether it is as effective in reducing tumor growth as DTIC.

Each of the six studies represents an independent small sample experiment of how well P28 works in suppressing tumor growth relative to untreated tumors. Since the individual study designs used a random assignment of mice to treatment groups, each study should yield an unbiased estimate of the baseline control growth rate and a treatment effect. However, the sample sizes are quite small, ranging between a total of 35 to 60 mice, with each divided among multiple treatment groups. Estimates based on very small sample sizes tend to have higher variances and are thus less informative than those from larger samples. The higher variances lead to a less powerful hypothesis tests and larger confidence regions. This is especially so in the (sparse) panel data context found here, wherein the number of parameters is of the same order as the number of subjects.

For more informative regression estimates, we pool information between studies to estimate the growth curves. For the melanoma data, although the designs are different between studies, there is an overlap in some of the treatments. In particular, all three studies included a group
### TABLE VI

**MICE COUNTS BY STUDY AND TREATMENT. MEL2, MEL29, AND MEL23 ARE THREE DIFFERENT MELANOMA CELL LINES WITH P53 EXPRESSION.**

<table>
<thead>
<tr>
<th>Cell Line</th>
<th>Studies (s)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEL2</td>
<td>Control</td>
<td>12</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>9</td>
<td>10</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>Azurin 10mg/kg daily</td>
<td>8</td>
<td>.</td>
<td>10</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>P28 5mg/kg daily</td>
<td>.</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>.</td>
<td>.</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>P28 10mg/kg daily</td>
<td>8</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>6</td>
<td>6</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>P28 20mg/kg daily</td>
<td>.</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>6</td>
<td>6</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>DTIC 4mg/kg daily</td>
<td>.</td>
<td>.</td>
<td>10</td>
<td>10</td>
<td>6</td>
<td>.</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>DTIC 5mg/kg daily</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>P28 10mg/kg, 3 times per week</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>P28 PEG 5mg/kg, 3 times per week</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>6</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>P28 PEG 10mg/kg, 3 times per week</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>8</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>Margin Total</td>
<td>36</td>
<td>50</td>
<td>60</td>
<td>60</td>
<td>35</td>
<td>40</td>
<td>281</td>
<td></td>
</tr>
</tbody>
</table>
that received a medium daily dose of P28 (10mg/kg). Instead of estimating separate quantile regression curves by study, the studies are combined and a single quantile curve is fitted to all the data. The estimated coefficients are then used to construct curves and to assess treatment efficacy.

4.3.2 Tumor Growth Model

Popular models of tumor development posit that volume growth follows an S shaped curve like that of populations in environments with limited resources (Gerlee, 2013). During early stages of tumor growth, the collection or population of cells that make up a tumor grows exponentially. However, the rate at which cell numbers and in turn tumor volume grows declines with tumor size, until volume plateaus at what is called the 'carrying capacity' or maximum volume that can be sustained given the limited resources of the organ or physiological system. Common models of this type include the Gompertz and Logistic curves ((Gerlee, 2013), (Demidenko, 2006)).

For the melanoma studies, preliminary examination of the data indicates that the study durations were far too short for the sample growth curves to exhibit the S shape reviewed above. Rather tumor volumes increased exponentially over the course of four weeks with no apparent inflection point or asymptotic limit to size. Therefore, we model the initial four week tumor growth curves using an exponential model. A selected comparison by study of the observed tumor volumes for the P28, 10 mg daily regimen and Control groups is given in Figures 23 through 25. In each plot, the blue colored components are control group values and red components are treatment group values.
Figure 22. Observed tumor volumes, Control and P28, 10mg daily groups, MEL2 and MEL 29 cell lines.
Figure 23. Observed tumor volumes, Control and P28, 10mg daily groups, MEL23 cell lines.
Figure 24. Observed tumor volumes, Control and P28, 10mg daily groups, All Studies.

Let \( Y = \log V(t) \), be the natural logarithm of tumor volume at time \( t \). Define the following indices for the observed data: \( s = 1, \ldots, 6 \) is an index for the different studies of melanoma, \( j = 0, \ldots, M \) enumerates the treatment groups, and \( i = 1, \ldots, N_{sj} \) the subject index within groups. Further let \( \tau \in (0, 1) \). Given the observations above, an adequate model for the \( \tau^{th} \) quantile curve for the natural logarithm of tumor volume across time is

\[
Q_{Y_{isj}}(\tau; t) = \alpha_i + \beta_0(\tau) + \beta_s(\tau) + \{\beta_C(\tau) + \beta_j(\tau)\delta_{ij}\} \times t,
\]

where \( \alpha_i \) is the baseline subject fixed effect which is constant across quantiles, \( \beta_s(\tau) \) is the quantile study effect, \( \beta_C(\tau) \) is the quantile change in log volume of control group tumors, \( \beta_j(\tau) \)
is the quantile treatment effect on the growth rate from the $j^{th}$ treatment, and $\delta_{ij}$ is set to 1 if the subject received the $j^{th}$ treatment group and is 0 otherwise.

### 4.3.3 Results

Estimates from a Penalized Fixed Effects Quantile Regression estimation procedure for five quantiles are presented in Tables Table VII and Table VIII. Analogous estimates from a Linear Mixed Effects model for the conditional mean are included as well. For PFQR, the finite sample distribution of parameter estimates is unknown. To assess statistical significance of the regression coefficients, we used a stratified subject bootstrap to construct confidence intervals. For our bootstrap procedure, we resample subjects with replacement by study and treatment group. This is similar to the cross-section bootstrap for panel data as studied in (Kapetanios, 2008) with the addition of stratification to account for the original study designs.

Our results indicate that the growth rate of untreated tumors increased with higher quantiles. The estimated growth rates for Control group tumors are 6.5% per day for the 5$^{th}$ quantile, 7.8% for the median, and 8.7% for the 95$^{th}$. All the estimated control rates are statistically significant at an $\alpha = 0.05$ level. Here significance is based on whether the bootstrap confidence intervals contains 0. As for treated tumors with P28, only the 10 mg dose of the experimental drug had a statistically significant effect on growth rates and only at the upper quantiles. For tumors treated with the medium daily dose of P28, the 75$^{th}$ quantile growth rate was 1.5 percentage points lower than the same value from the Control distribution, and the 95$^{th}$ rate was 2.7 lower. This yields upper tail estimates for the P28,10mg tumor growth rate on par with the lower tail of the control untreated tumors. In contrast, the standard DTIC chemotherapy
had a statistically significant effect on tumor growth in the lower tail and central parts of the distribution. The DTIC effect was not statistically significant at the 95th percentile.

The resulting fitted growth curves for P28–10 mg/kg daily dose, DTIC, and the Control are plotted in Figure 26. The curves correspond to the 0.05, 0.25, 0.50, 0.75, and 0.95 quantiles, with higher curves corresponding to higher quantiles. The blue dashed lines are the control group curves and the red solid lines are one of the two treatment groups. The results for P28 are shown in the left panel and those for DTIC in the right. As can be seen, for both treatments, the reference lines are shifted downwards relative to Control. Further, the DTIC curves are lower for the lower quantiles compared to P28; but the latter is more effective in lowering the 95th quantile curve.

Figure 25. Estimated Quantile Growth Curves for Control, P28,10 mg/kg, and DTIC.
### TABLE VII
ESTIMATED QUANTILE TREATMENT EFFECTS

<table>
<thead>
<tr>
<th>Quantile</th>
<th>Intercept</th>
<th>Control</th>
<th>Azurin</th>
<th>DTIC, 4 mg</th>
<th>P28, 10mg</th>
<th>DTIC, 5 mg</th>
<th>P28, 10, weekly</th>
<th>P28PEG, 10 mg</th>
<th>P28PEG, 5 mg</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>-5.594</td>
<td>0.065</td>
<td>0.015</td>
<td>-0.033,0.008</td>
<td>-0.011</td>
<td>0.033</td>
<td>0.033</td>
<td>0.031</td>
<td>0.029</td>
</tr>
<tr>
<td>0.25</td>
<td>-5.196</td>
<td>0.075</td>
<td>0.000</td>
<td>-0.028</td>
<td>-0.004</td>
<td>0.004</td>
<td>0.002</td>
<td>0.021</td>
<td>0.008</td>
</tr>
<tr>
<td>0.50</td>
<td>-4.785</td>
<td>0.078</td>
<td>-0.010</td>
<td>-0.027</td>
<td>-0.006</td>
<td>-0.015</td>
<td>-0.010</td>
<td>-0.012</td>
<td>0.000</td>
</tr>
<tr>
<td>0.75</td>
<td>-4.259</td>
<td>0.081</td>
<td>0.010</td>
<td>-0.025,0.002</td>
<td>-0.014</td>
<td>0.005</td>
<td>-0.008</td>
<td>-0.015</td>
<td>0.000</td>
</tr>
<tr>
<td>0.95</td>
<td>-3.630</td>
<td>0.087</td>
<td>0.020</td>
<td>-0.038,0.004</td>
<td>-0.023</td>
<td>0.005</td>
<td>-0.011</td>
<td>-0.027,0.002</td>
<td>0.013</td>
</tr>
</tbody>
</table>

**Mean (LME):**

- Intercept: (-5.878,-5.293)
- Control: (0.051,0.082)
- Azurin: (-0.009,0.04)
- DTIC, 4 mg: (-0.033,-0.008)
- P28, 10mg: (-0.026,0.007)
- DTIC, 5 mg: (0.017,0.052)
- P28, 10, weekly: (0.005,0.053)
- P28PEG, 10 mg: (0.004,0.024)
- P28PEG, 5 mg: (0.008,0.021)

### TABLE VIII
ESTIMATED QUANTILE GROWTH RATES BY TREATMENT

<table>
<thead>
<tr>
<th>Quantile</th>
<th>Control</th>
<th>Azurin</th>
<th>DTIC</th>
<th>P28, 10mg</th>
<th>DTIC</th>
<th>P28, 10, weekly</th>
<th>P28PEG, 10 mg</th>
<th>P28PEG, 5 mg</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.065</td>
<td>0.080</td>
<td>0.045</td>
<td>0.054</td>
<td>0.049</td>
<td>0.029</td>
<td>0.021</td>
<td>0.031</td>
</tr>
<tr>
<td>0.25</td>
<td>0.075</td>
<td>0.075</td>
<td>0.047</td>
<td>0.071</td>
<td>0.065</td>
<td>0.008</td>
<td>0.021</td>
<td>0.008</td>
</tr>
<tr>
<td>0.50</td>
<td>0.078</td>
<td>0.077</td>
<td>0.047</td>
<td>0.072</td>
<td>0.074</td>
<td>0.000</td>
<td>0.012</td>
<td>0.000</td>
</tr>
<tr>
<td>0.75</td>
<td>0.081</td>
<td>0.071</td>
<td>0.051</td>
<td>0.067</td>
<td>0.076</td>
<td>-0.012</td>
<td>0.000</td>
<td>-0.012</td>
</tr>
<tr>
<td>0.95</td>
<td>0.087</td>
<td>0.067</td>
<td>0.059</td>
<td>0.064</td>
<td>0.072</td>
<td>-0.011</td>
<td>0.000</td>
<td>-0.011</td>
</tr>
</tbody>
</table>

**Mean (LME):**

- Control: (0.069,0.084)
- Azurin: (0.069,0.084)
- DTIC: (0.069,0.084)
- P28, 10mg: (0.069,0.084)
- P28PEG, 10 mg: (0.069,0.084)
- P28PEG, 5 mg: (0.069,0.084)

- Control: (0.069,0.084)
- Azurin: (0.069,0.084)
- DTIC: (0.069,0.084)
- P28, 10mg: (0.069,0.084)
- P28PEG, 10 mg: (0.069,0.084)
- P28PEG, 5 mg: (0.069,0.084)


<table>
<thead>
<tr>
<th>Quantile</th>
<th>0.05</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>0.95</th>
<th>Mean (LME)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTIC</td>
<td>15.403</td>
<td>14.748</td>
<td>13.591</td>
<td>11.748</td>
<td>8.252</td>
<td>10.830</td>
</tr>
</tbody>
</table>

Looking at the corresponding estimates from the conditional mean model, the coefficient values appear to be a rough average of the quantile estimates and mask the differences in growth rates across quantiles. For p28, 10 mg dose, the estimated mean effect is half the estimated effect for the 95th quantile. Similarly, the DTIC mean estimate is again half the value of its effect for lower and central quantiles.

A useful measure of relative tumor growth is the doubling time, $T_D$, which is the number of days that a tumor takes to double in size. Under the exponential model, this is a fairly easy calculation:

$$T_D = \frac{\ln(2)}{\gamma},$$

where $\gamma$ is an estimated growth rate. The quantile doubling times are presented in Table 4. Looking across quantiles, the Control doubling times decline with quantiles. Slow growing tumors of the 5th quantile take 10.7 days to double, those with a median growth trajectory take
8.9 days, and those in the extreme 95th quantile double in 8 days. For mice treated with the medium dose of P28, the doubling times increased by 1.8 days for the 75th quantile and 3.1 days for the 95th over that of the control, with the result that the larger and faster growing tumors treated with P28 take longer to double in size than the smaller and slower growing untreated control tumors.

As a final step, we constructed quantile growth tumor growth curves for the estimates in Table VII. Curves for the two significant treatments–p28 10mg/day and DTIC–are illustrated in Figure Figure 25. In both panels, the dashed blue lines are the Control group curves and the red solid lines are treated tumor curves. Examining the curves, shows that the two treatments have different quantile effects. The first panel compares the P28 quantiles to that of the control group. P28 appears to have little effect for the lower quantile trajectories, but it is very effective for the 75th and 95th quantiles. The 75th quantile growth of tumors treated with P28 is on par with the median of untreated tumors, and the 95th quantile curve is on par with the 75th quantile curve of the untreated tumors. The second panel gives a similar comparison for DTIC. In contrast to P28, DTIC appears to have an opposite effect on suppressing growth. The DTIC curves are lower than control for the lower tail, median, and the 75th percentile. However, DTIC appears to not suppress growth of the larger and fast growing tumors in the 95th percentile.

These results have an important implication for melanoma treatment development and personalized patient treatment regimes. Unlike the conclusions that can be drawn from least squares regression, the quantile growth curves show that both the conventional and experimental
treatments did not work the same across the entire population of subjects. In particular, mice with faster growing tumors fared better from treatment with the less toxic P28 and not the conventional chemotherapy in the first month of tumor development, while those with smaller tumors did well with DTIC. Thus, the conditional quantile growth curve, and quantile regression in general, can be an invaluable tool in the individualized approach to medical treatment and drug development, wherein treatments are customized to the individual in order to increase treatment response and minimize side effects.

4.4 Conclusion

In this chapter, we study the use quantile regression to estimate quantile growth models for small sample longitudinal studies of tumor growth. This approach is demonstrated using data from three studies of xenograft mice models of melanoma cell growth. Due to the short time range of the observations, a simple exponential growth model is specified with added fixed study effects to control for study differences and fixed subject effects to account for mouse heterogeneity. The model is estimated using the Penalized Fixed Effects Quantile Regression (PFEQR) approach of (Koenker, 2004). This method extends the original quantile regression model to longitudinal data by incorporating fixed subject specific model parameters that are penalized to prevent overfitting and decrease model dimension. This approach has several advantages. One it is fully nonparametric in the sense that it does not require specifying any distributional or moment assumptions on the unobserved error structure; which heretofore is required for other longitudinal quantile regression procedures. Second, and in part due to less stringent distributional assumptions, it allows simultaneous estimation of several quantiles.
However, as of yet, very little work has been done to develop valid inference procedure for PFEQR. Therefore, for inference and hypothesis testing, we used a stratified bootstrap in which we resample data by study and treatment group.

Our estimates indicate that experimental treatment P28 has a different effect on growth rates at different quantiles, with corresponding quantile growth curves much lower than their Control group counterparts at the upper tail but are similar to Control at the central and lower tail points of the tumor distribution. Similarly, the conventional treatment for melanoma, DTIC, also has a different effect on suppressing tumor growth, but is opposite to that of P28. The conventional treatment is more effective in abating growth among smaller tumors, but has no significant effect for larger tumors.

These results are really the tip of the iceberg of how quantile regression can be applied in research into cancer treatments. Further steps would look at more realistic extensions of the model to incorporate within subject correlation. As discussed in the introduction, many of the other longitudinal quantile regression methods incorporate within subject correlation; but at the cost of needing to specify or assume something about the distribution of the data. Under the PFEQR formulation, it is possible to capture correlation through model parameters allowing for a parsimonious way to account for shifts in within subject correlation across quantiles. In fact, the model used by (Wei and He, 2006) and (Wei et al., 2006) to estimate quantile regression models for childrens conditional growth charts do incorporate autoregressive terms. Alternatively, the PFEQR estimating equations can be weighted using some initial estimate of the unknown correlation structure as in (Karlsson, 2008).
Finally, the concept of combining independent studies of the same phenomenon to generate a combined regression estimate calls to mind the idea of a meta-quantile regression, of which very little exists in the literature. Further, since the PFEQR procedure incorporates a lasso style penalty it would be very interesting to study meta-quantile regression wherein several penalized estimates are combined. One central question would be how close the meta-analysis of individual penalized estimates would be to a penalized estimate from a single combined dataset.
CITED LITERATURE


VITA

Ella Revzin

| Education | M.S., Mathematics, Concentration in Statistics  
University of Illinois at Chicago, Chicago, IL  
Dec 2011 |
|-----------|--------------------------------------------------|
|           | B.A., Economics  
University of Chicago, Chicago, IL  
June 2003 |

| Experience | Coyote Logistics, Chicago, IL  
Researcher  
2014 to Present |
|------------|--------------------------------------------------|
|            | University of Illinois at Chicago  
Graduate Teaching Assistant  
2009 to 2013 |
|            | University of Illinois at Chicago  
Research Assistant  
2009 to 2013 |
|            | University of Illinois at Chicago  
Statistician  
2009 to 2013 |
|            | Soochow University, Suzhou, China  
NSF EAPSI Fellow  
Summer 2012 |
|            | RCF Economic and Financial Consulting, Chicago, IL  
Economist  
2003-2009 |
Presentations

‘An Assessment of Mathematics Placement Policy Changes at UIC’
Statistics Seminar
Department of Mathematics, Statistics, and Computer Science
University of Illinois-Chicago, IL
January 2012

‘Bootstrap Confidence Bands For Regression’
China-U.S. Young Scientist Forum
Beijing, China
August 2012

‘Bootstrap Confidence Bands for Regression Curves Using Polynomial Splines’
Joint Statistical Meetings 2013
Montreal, Quebec
August 2013

‘Modeling Conditional Quantile Tumor Growth Curves By Combining Independent Small Sample Study Data’
10th International Conference on Health Policy Statistics
Chicago, IL
October 2013

‘Bootstrap Confidence Bands for Regression Curves Using Polynomial Splines’
Graduate Statistics Seminar
Department of Mathematics, Statistics, and Computer Science
University of Illinois-Chicago, IL
November 2013

Affiliations
American Statistical Association
Institute of Mathematical Statistics

Languages
English, Russian