Effect of Smoking on Babies' Weight

Oscar Aguilar

December 2, 2015

Abstract

If your health isn't enough to make you quit smoking, then the health of your baby should be. Smoking during pregnancy affects you and your baby's health before, during, and after your baby is born. The nicotine (the addictive substance in cigarettes), carbon monoxide, and numerous other poisons you inhale from a cigarette are carried through your bloodstream and go directly to your baby. Smoking while pregnant will: increase your baby's heart rate, increase the risk that your baby is born prematurely and/or born with low birth weight, increase your baby's risk of developing respiratory (lung) problems, increases risks of birth defects, increases risk of Sudden Infant Death Syndrome. The more cigarettes you smoke per day, the greater your baby's chances of developing these and other health problems. There is no "safe" level of smoking while pregnant. In this project, we examine the effect of smoking on babies' weight from the statistical point of view using nonparametric methods. The data used in this project is just a portion of data from a much larger study which includes all pregnancies that occurred between 1960 and 1967 at the Kaiser Foundation Health Plan in Oakland, California. We start our analyses by estimating densities for each of continuous variables in the data. Next, we fit a nonparametric regression model without taking into account the smoking status of the mother. After that, we consider two nonparametric regression curves: one for non-smoker and another one for smoker. In order to assess the effect of smoking on babies's weight, we conduct a hypothesis test based on *wild bootstrap* to see if there was a significant difference between those curves, and the result shows that there is a significant difference between the non-smoker and smoker curves. Finally, we propose a semi-parametric partially linear model to the data.

Key words: kernel density estimation, nonparametric regression, semi-parametric regression, wild bootstrap.

Contents

1 Introduction

The methods of smoothing have been developed mainly in the last thirty years. The intensive interest in smoothing over this last three decades had two reasons: statisticians realized that pure parametric thinking in curve estimations often does not meet the need for flexibility in data analysis and the development of hardware created the demand for theory of now computable nonparametric estimates. In this project, we examine effect of smoking on babies' weight using nonparametric methods and is organized as follows: Section 2 describes the data that will be used. In Section 3, we briefly explain the nonparametric methods used for univariate density estimation and then we apply these methods to the data. In Section 4, we extend the idea of univariate density estimation to joint density estimation. Section 5 briefly explains the most common methods/techniques used in nonparametric regression such as Nayadara-Watson estimation and local linear estimation, and we then apply those methods to the data. In Section 6, we assess the effect of smoking in babies' weight. In Section 7, a semi-parametric partially linear model is proposed for the data. Finally, Section 8 wraps up this analysis with recap on interesting remarks and suggestions for further study.

2 Data Description

The data used in this project is just a portion of data from a much larger study which includes all pregnancies that occurred between 1960 and 1967 at the Kaiser Foundation Health Plan in Oakland, California. The data here are from one year of the study. It includes all 1104 male single births where the baby lived at least 28 days. The variable descriptions are given in the following table.

Table 1: Data Description

Variable	Description
Birth Weight (`bwt`)	Babys weight at birth, to the nearest ounce
Gestation (`gestation`)	Duration of the pregnancy in days
Weight (`weight`)	Mothers pre-pregnancy weight, in pounds
Smoking status (`smoke`)	Indicator for whether the mother smokes (1) or not (0)

The main purpose of analyzing this data set is to investigate is the effect of smoking status in the baby's weight. Another important aspect is to construct/build a semi-parametric model in which the baby's weight is the response and the other variables such as gestation, weight and smoking status are potential predictors.

3 Univariate Density Estimation

In this section, we provide a brief description of univariate density estimation based on kernel methods.

3.1 The Basics of Kernel Density Estimation

Let $X_1, X_2, \ldots, X_n$ denote a sample of size n from a random variable with density f. The kernel density estimate of f at the point x is given by

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{1}$$

where the kerne K satisfies $\int K(x)dx = 1$ and the smoothing parameter h is known as the bandwidth. In practice, the kernel K is generally chosen to a uni-modal probability density symmetric around zero. In this case, K satisfies the conditions

$$\int K(y)dy = 1, \quad \int yK(y)dy = 0, \quad \int y^2K(y)dy = \mu_2(K) > 0$$

One of the most popular choice for K is the Gaussian kernel; that is

$$K(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \tag{2}$$

In this project, we consider K to be the Gaussian kernel. Assuming that the underlying density is sufficiently smooth (f'' being absolute continuous and f''' being square integrable), it can be shown that if $h \to 0$ with $nh \to \infty$ as $n \to \infty$, then by Taylor series that

$$\text{Bias}\{\hat{f}(x)\} = \frac{h^2}{2}\mu_2(K)f''(x) + o(h^2) \quad \text{and} \quad \text{Var}\{\hat{f}(x)\} = \frac{R(K)f(x)}{nh} + O\left(\frac{1}{n}\right)$$

where $R(K) = \int K^2(y)dy$ is a measure of roughness of the kernel. The mean square error of the density estimate is given by

$$\text{MSE}\{\hat{f}(x)\} = \frac{1}{4}h^4(f''(x))^2\mu_2(K) + \frac{R(K)f(x)}{nh} + o(h^4) + O\left(\frac{1}{n}\right)$$

The optimal h that minimizes the leading MSE is given by

$$h = \left(\frac{f(x)R(K)}{(f''(x))^2\mu_2(K)^4}\right)^{1/5} n^{-1/5}$$

which is the optimal local bandwidth. The optimal global bandwidth is given by

$$h^* = \left(\frac{R(K)}{\mu_2(K)^2R(f'')}\right)^{1/5} n^{-1/5} \tag{3}$$

3.2 Bandwidth Selection for Kernel Density Estimates

In this section, we briefly review methods for choosing a global value of the bandwidth h.

3.2.1 Rule of Thumb

The computationally simplest method for choosing a global bandwidth h is based on replacing $R(f^{''})$, the unknown part in (3). If a Gaussian kernel is used and assuming that the underlying distribution is normal, Silverman (1986) show that (3) reduces to

$$h^* = 1.06\sigma n^{-1/5} \quad \text{and} \quad h^* = 0.79\text{IQR}n^{-1/5}$$

where σ and IQR are the standard deviation and interquartile range of X, respectively. The Silverman's rule of thumb is given by

$$h^* = 1.06\hat{\sigma} n^{-1/5} \tag{4}$$

where $\hat{\sigma} = \min\{s, \text{IQR}/1.34\}$ and s stands for sample standard deviation.

3.2.2 Cross-Validation Methods

Unlike the rule of thumb discussed in the previous section, cross-validation is an objective approach. A measure of the closeness of $\hat{f}$ and f for a given sample is the integrated squared error (ISE), which is given by

$$\text{ISE}(\hat{f}_h) = \int (\hat{f}_h(x) - f(x))^2 dx = \int \hat{f}_h(x)dx - 2\int \hat{f}_h(x)f(x)dx + \int f^2(x)dx$$

Notice that if $\text{ISE}(\hat{f}_h)$ were known, we could just find the h that minimizes $\text{ISE}(\hat{f}_h)$. However, $\text{ISE}(\hat{f}_h)$ is not known due to f. Bowman (1984) proposed choosing the bandwidth as the value that minimizes the estimate of the other two known term in $\text{ISE}(\hat{f}_h)$. That is

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{f}_{-i}(x))^2 dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(X_i) \tag{5}$$

Stone (1984) showed that the second term in (5) is an unbiased estimator of the second term in ISE. That is

$$E\left[\int \hat{f}_h(x)f(x)dx\right] = \int\int K\left(\frac{y-x}{h}\right)f(x)dxf(y)dy = E\left[K\left(\frac{Y-X}{h}\right)\right]$$

This leads to the unbiased estimate of $\int \hat{f}(x)f(x)dx$ given by

$$\frac{1}{n(n-1)}\sum_{i\neq j}\sum K\left(\frac{X_i - X_j}{h}\right) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}_{-i}(X_i)$$

Hall (1983) show that

$$\frac{1}{n}\sum_{i=1}^{n}\int(\hat{f}_{-i}(x))^2 dx = \int(\hat{f}_h(x))^2 dx + O_p\left(\frac{1}{n^2 h}\right)$$

and hence the least square cross-validation (LSCV) based criterion from (5) to

$$\mathrm{LSCV}(h) = \int (\hat{f}_h(x))^2 dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(X_i) \tag{6}$$

We next proceed to estimate the densities, using the rule of thumb to estimate the bandwidth, of the following variables: `bwt`, `gestation`, and `weight` as shown in Figure 1.

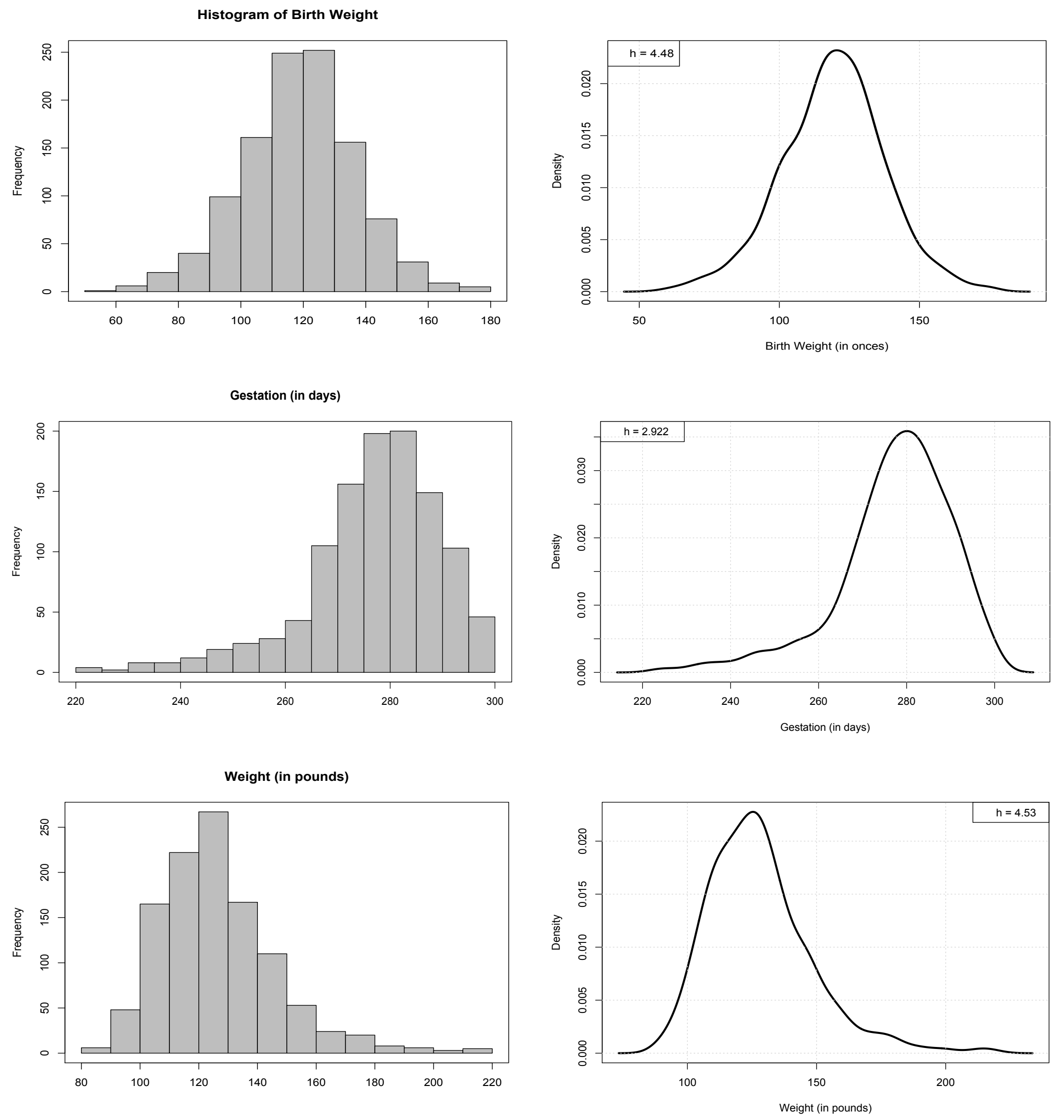

Figure 1: *Left Panels:* represent the histogram of `bwt`, `gestation` and `weight`. *Right Panels:* represent the estimated kernel densities for `bwt`, `gestation` and `weight` using the *Rule of Thumb* (4) to find h, respectively.

In Figure 1, the left panels show the histograms of `bwt`, `gestation` and `weight`. On the other hand, the right panels represent the kernel estimates of `bwt`, `gestation` and `weight`.

We used the *Rule of Thumb*, given by (4), to find the h, the bandwidth. As we can see in Figure 1, all the kernel estimates look smooth. We next proceed to estimate the densities of `bwt`, `gestation` and `weight`, but we now use the least square cross-validation to find the optimal bandwidth h as shown in Figure 2.

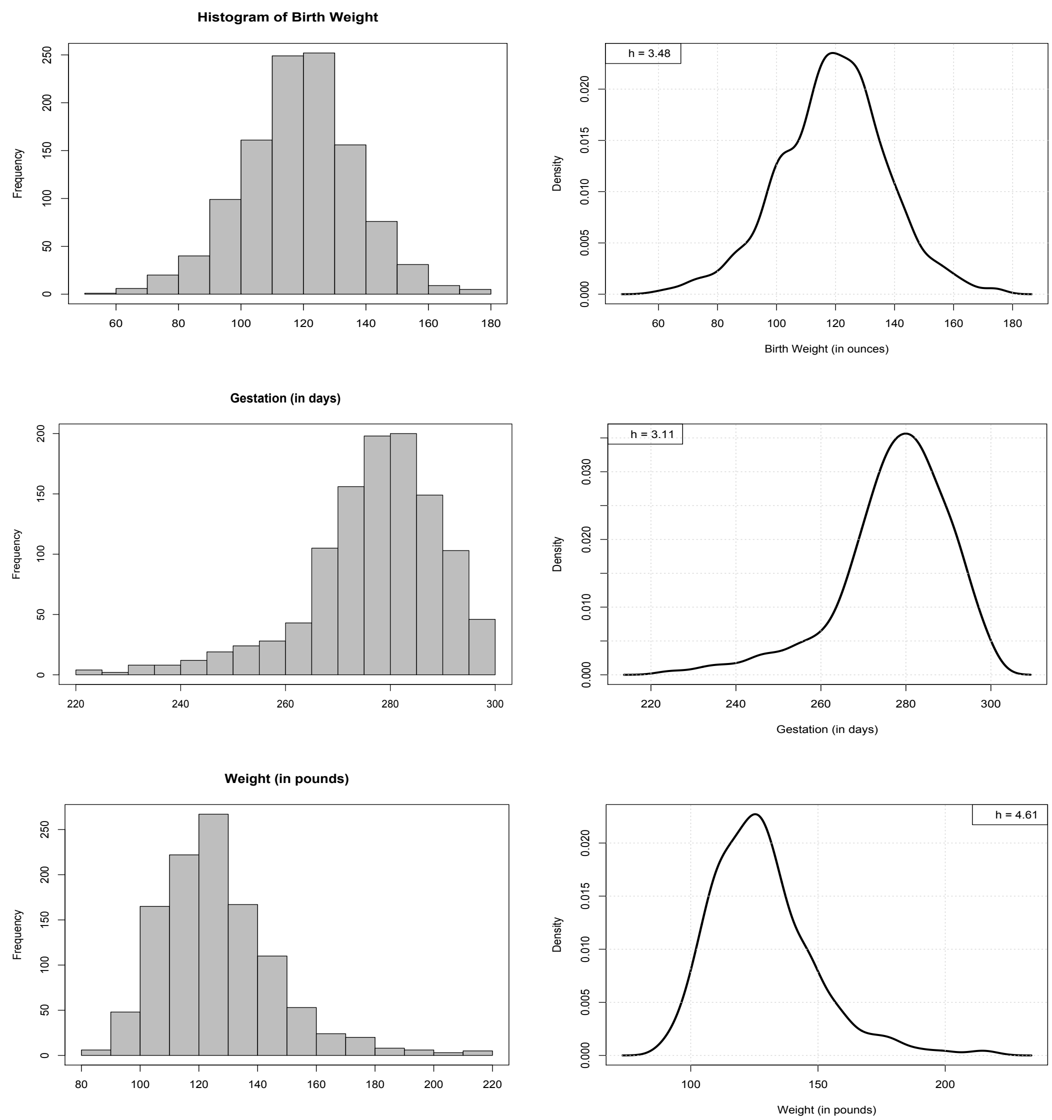

Figure 2: *Left Panels:* represent the histogram of `bwt`, `gestation` and `weight`. *Right Panels:* represent the estimated kernel densities for `bwt`, `gestation` and `weight` using the least square cross-validation (6) to find h, respectively.

In Figure 2, the left panels show the histograms of `bwt`, `gestation` and `weight`. On the other hand, the right panels represent the kernel estimates of `bwt`, `gestation` and `weight`. We used the least squares cross-validation, given by (6), to find the h, the bandwidth. As we can see in Figure 2, all the kernel estimates look smooth; however, the kernel estimate for `bwt` looks a little bumpy. On the other hand, the other two kernel estimates look very similar to ones shown in Figure 1. Table 2 shows the bandwidths for each of the variables using the two considered approaches.

Table 2: Bandwidths

Variable	*Rule of Thumb*	LSCV
`bwt`	4.48	3.48
`gestation`	2.92	3.11
`weight`	4.48	4.61

From Table 2, we see that the major difference between the bandwidths is for `bwt`; that is why the kernel estimates look different for `bwt`. After finding kernel estimates for each of the variables, we next proceed to find joint kernel estimates for the response, `bwt`, and each of the potential predictors, `gestation` and `weight`.

4 Joint Density Estimation

Kernel density estimation can be easily generalized from univariate to multivariate data, in theory if not always in practice. Let $X_1, X_2, \ldots, X_n$ be iid sample in $\mathbb{R}^d$ with pdf $f : \mathbb{R}^d \to \mathbb{R}$. The general form of the kernel density estimator is given by

$$\hat{f}(x) = \frac{1}{n|H|} \sum_{i=1}^{n} \mathbb{K}\left(H^{-1}(x - X_i)\right) \tag{7}$$

where H is a $d \times d$ nonsingular bandwith matrix, $\mathbb{K} : \mathbb{R}^d \to \mathbb{R}$ is a kernel satisfying

$$\int_{\mathbb{R}^d} \mathbb{K}(x)dx = 1, \quad \int_{\mathbb{R}^d} x\mathbb{K}(x)dx = 0, \quad \int_{\mathbb{R}^d} xx^T\mathbb{K}(x)dx = \begin{pmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_d^2 \end{pmatrix}$$

One of the most popular choice for K is the Gaussian kernel; that is

$$\mathbb{K}(x) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}x^T x\right) = K(X_1) \cdots K(X_d) \tag{8}$$

where $K(X_i)$ for $i = 1, \ldots, d$ are univariate Gaussian kernel estimates (2). It can be shown that the bias and variance of (2) are given by

$$\text{Bias}\{\hat{f}(x)\} = \frac{1}{2}h^2 \text{Trace}(\nabla^2 f(x)) + o(h^2) \quad \text{and} \quad \text{Var}\{\hat{f}(x)\} = \frac{R(K)f(x)}{nh^d} + O\left(\frac{1}{n}\right)$$

where $R(K) = \int_{\mathbb{R}^d} \mathbb{K}^2(x)dx$. Notice that the kernel density estimate uses the same h in each component/direction of data as in the univariate case.

4.1 Normal Reference Rule to Choose Bandwidth

Assume that f is the pdf of $N_d(0, \Sigma)$, where

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_d^2 \end{pmatrix}$$

and K is the Gaussian kernel, the practical bandwidth can be chosen as

$$h_j = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} \hat{\sigma}_j n^{-\frac{1}{d+4}} \tag{9}$$

Notice that for $d = 2$, h_j reduces to

$$h_j = \hat{\sigma}_j n^{-\frac{1}{6}} \tag{10}$$

We next proceed to estimate the joint densities of the response `bwt` with each of the potential predictors: `gestation`, and `weight` as shown in Figure 3.

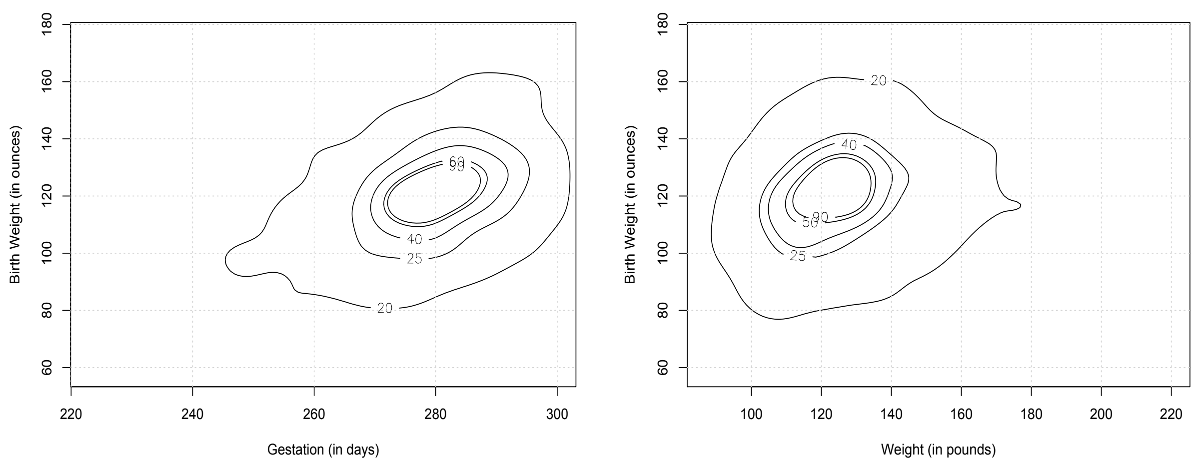

Figure 3: *Left Panel:* represents the contour-plot of the joint density estimate of `gestation` and `bwt` using the normal reference rule (10) to find the bandwidths . *Right Panel:* represents the contour-plot of the joint density estimate of `weight` and `bwt` using the normal reference rule (10) to find the bandwidths.

5 Smoothing for Nonparametric Models

Smoothing of a data set $\{(X_i, Y_i)\}_{i=1}^n$ involves the approximation of the mean response curve m in the regression relationship given by

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i, \quad i = 1, \dots, n \tag{11}$$

where $E(\varepsilon|X=x)=0$, $\mathrm{Var}(\varepsilon|X=x)=1$, $m(x)=E(Y|X=x)$ and $\sigma^2(x)=\mathrm{Var}(Y|X=x)$. $m(x)$ is the conditional mean function and $\sigma^2(x)$ is the conditional variance function. Notice that both $m(x)$ and $\sigma^2(x)$ are unknown. We next consider two estimators for m: the Nayadara-Watson estimator and local linear regression estimator.

5.1 Nadayara-Watson Estimator

Let m be the conditional expected value of Y conditioned on X. That is

$$\begin{aligned} m(x) &= E(Y|X=x) \\ &= \int y \frac{f(x,y)}{f(x)} dy \end{aligned}$$

Note that the kernel estimate of the conditional pdf of y given x is given by

$$\hat{f}(y|x) = \frac{1}{h_y} \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_x}\right) K\left(\frac{y-Y_i}{h_y}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_x}\right)}$$

We then can estimate $m(x)$ with $\hat{m}(x)$ as follows:

$$\begin{aligned} \hat{m}(x) &= \int y \hat{f}(y|x) dy \\ &= \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_x}\right) Y_i}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_x}\right)} \end{aligned}$$

$\hat{m}(x)$ is often referred as the *Nayadara-Watson* estimator. It can be shown that

$$\mathrm{Bias}\{\hat{m}(x)\} = \frac{1}{2} h^2 \frac{m''(x) f(x) + 2m'(x) f'(x)}{f(x)} \sigma_K^2 + o(h^2) \tag{12}$$

$$\mathrm{Var}\{\hat{m}(x)\} = \frac{\sigma^2(x) R(K)}{nhf(x)} + o\left(\frac{1}{nh}\right) \tag{13}$$

We next consider another estimator for m.

5.2 Local Polynomial Regression

Consider

$$R(\boldsymbol{\beta}) = \sum_{i=1}^n \left\{ Y_i - \beta_0 - \beta_1(x-X_i) - \cdots \beta_p (x-X_i)^p \right\}^2 K\left(\frac{x-X_i}{h}\right)$$

The design matrix is given by

$$X_x = \begin{pmatrix} 1 & x-X_1 & \cdots & (x-X_1)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x-X_n & \cdots & (x-X_n)^p \end{pmatrix}$$

and the weight matrix is given by

$$W_x = \frac{1}{h}\text{diag}\left[K\left(\frac{x-X_i}{h}\right), \ldots, K\left(\frac{x-X_i}{h}\right)\right]$$

The goal is to find $\boldsymbol{\beta}$ such that $R(\boldsymbol{\beta})$ is minimized. It can be shown that $\hat{\boldsymbol{\beta}}$, the vector parameter that minimizes $R(\boldsymbol{\beta})$, is given by

$$\hat{\boldsymbol{\beta}} = (X_x^T W_x X_x)^{-1} X_x W_x Y$$

Let

$$S_\ell(x) = \frac{1}{nh}\sum_{i=1}^{n}(x-X_i)^\ell K\left(\frac{x-X_i}{h}\right)$$

Then the local linear regression estimate of m is given by

$$\hat{m}_1(x) = \frac{1}{nh}\sum_{i=1}^{n}\frac{[S_2(x)-S_1(x)(x-X_i)]K\left(\frac{x-X_i}{h}\right)Y_i}{S_2(x)S_0(x)-S_1^2(x)} \tag{14}$$

Assuming that $f(x) > 0$ and $f(\cdot)$, $m^{(2)}(\cdot)$ and $\sigma^2(\cdot)$ are continuous in a neighborhood of x and $h \to 0$ and $nh \to \infty$. Then, for an interior point of x

$$E\{\hat{m}_1(x)\} = \frac{1}{2}h^2\frac{m^{''}(x)f(x)+2m^{'}(x)f^{'}(x)}{f(x)}\sigma_K^2 + o(h^2) \tag{15}$$

$$\text{Var}\{\hat{m}_1(x)\} = \frac{\sigma^2(x)R(K)}{nhf(x)} + o\left(\frac{1}{nh}\right) \tag{16}$$

5.3 Choosing the Bandwidth

In the previous two sections, we briefly explained two different approaches to estimate m: the Nayadara-Watson estimator and local linear estimator. After discussing those estimators, we know that they depend on the bandwidth. In the next section, we briefly explain how to choose the bandwidth using cross-validation methods.

5.3.1 Cross-Validation

Cross-validation is an objective method to choose the bandwidth because the goal is to find the bandwidth h such the objective function is minimized. Let

$$\text{CV}(h) = \frac{1}{n}\sum_{j=1}^{n}\left\{Y_j - \hat{m}_{h,j}(X_j)\right\}^2 \tag{17}$$

where

$$\hat{m}_{h,j}(x) = \frac{\sum_{i\neq j} K\left(\frac{x-X_i}{h}\right)Y_i}{\sum_{i\neq j} K\left(\frac{x-X_i}{h}\right)}$$

Notice that $\hat{m}_{h,j}(x)$ is the Nayadara-Watson estimator without (X_j, Y_j). We choose the bandwidth, h, that minimizes (17). Now that we know how to choose the bandwidth, we next proceed to conduct two nonparametric regression. In the first regression, `bwt` is the response and `gestation` is the predictor; on the other hand, in the second regression, `bwt` is the response and `weight` is the predictor. The kernel smothers $\hat{m}$ for each regression is shown in Figure 4.

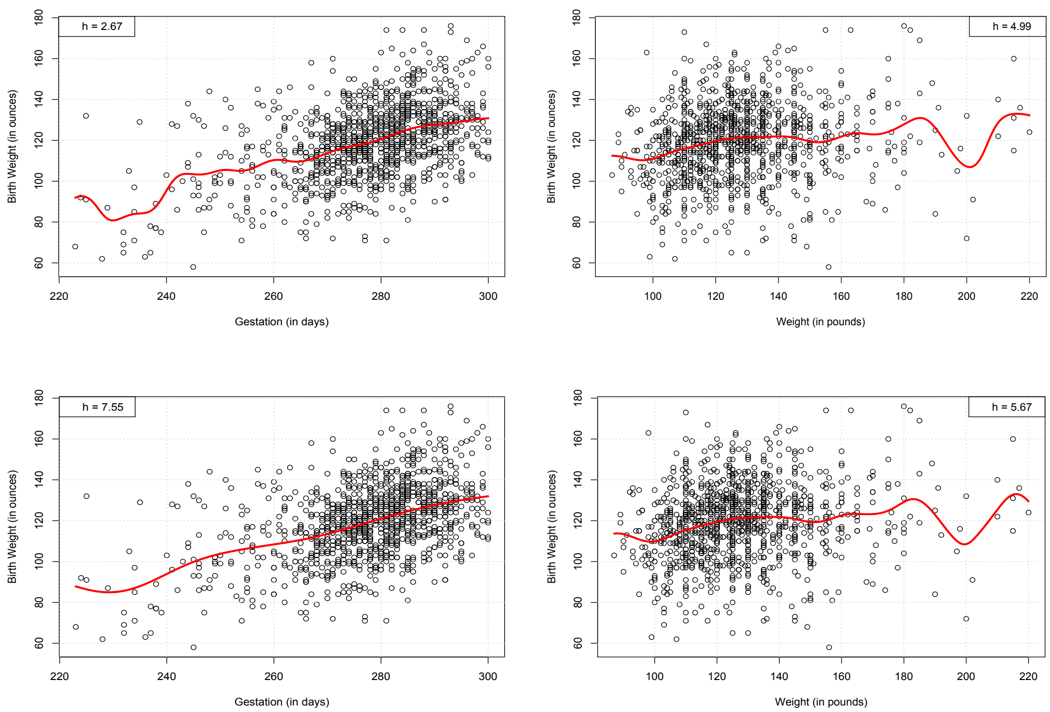

Figure 4: *Top Left Panel:* the red curve represents the Nayadara-Watson estimate of m when `bwt` is the response and `gestation` is the predictor. *Top Right Panel:* the red curve represents the Nayadara-Watson estimate of m when `bwt` is the response and `weight` is the predictor. *Bottom Left Panel:* the red curve represents the local linear estimate of m when `bwt` is the response and `gestation` is the predictor. *Bottom Right Panel:* the red curve represents the local linear estimate of m when `bwt` is the response and `weight` is the predictor.

In Figure 4, all the conditional mean estimates were estimated using the bandwidth, h, that minimized the CV function given by (17). The top left and bottom left panels represent the Nayadara-Watson and local linear estimates of the nonparametric model where `bwt` is the response and `gestation` is the predictor, respectively. If we compare those estimates, the local linear estimate looks smoother than the Nayadara-Watson estimate. On the other hand, the top right and bottom right panels represent the Nayadara-Watson and local linear estimates of the nonparametric model where `bwt` is the response and `weight` is the predictor, respectively. If we compare those estimates, Nayadara-Watson estimate looks smoother than the local linear estimate. The next step is to construct confidence bands for the conditional mean estimates, but before constructing confidence bands for the conditional mean estimates, we briefly explain how to estimate the conditional variance.

5.4 Estimating the Conditional Variance

Let $\hat{m}_h$ be a kernel smoother of m based on cross-validation bandwidth h, then we estimate the conditional variance at x as

$$\hat{\sigma}^2(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_1}\right)\left\{Y_i - \hat{m}_h(X_i)\right\}^2}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_1}\right)} \tag{18}$$

Notice that $\hat{\sigma}^2(x)$ is a Nayadara-Watson estimator on the square of residual $\{Y_i - \hat{m}_h(X_i)\}^2$ and h_1. The most common approach to choose h_1 is by cross-validation as explained in Section 5.3.1. In Figure 5, we show the estimated conditional variance for each of the conditional mean estimates shown in Figure 4.

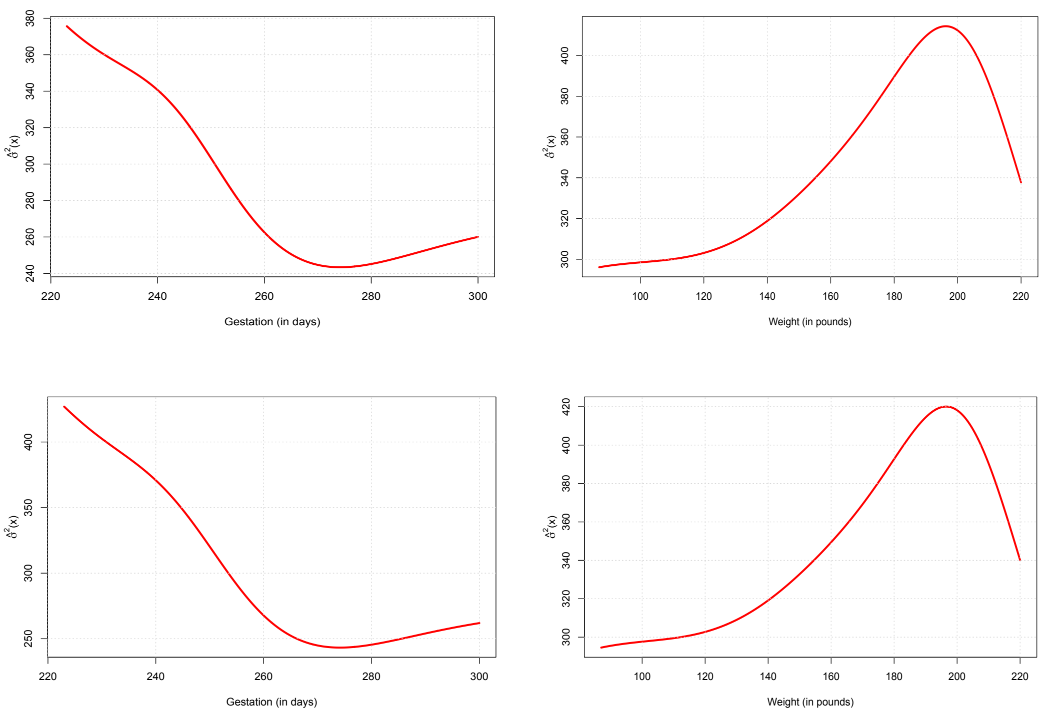

Figure 5: *Top Left Panel:* the red curve represents the estimated conditional variance from the *top left panel* in Figure 4. *Top Right Panel:* the red curve represents the estimated conditional variance from the *top right panel* in Figure 4 . *Bottom Left Panel:* the red curve represents the estimated conditional variance from the *bottom left panel* in Figure 4. *Bottom Right Panel:* the red curve represents the estimated conditional variance from the *bottom right panel* in Figure 4.

In Figure 5, in each of the conditional variance graphs, we see that there are some regions where the estimated variance is higher. These regions represent the regions where we have less data points than other regions. Now that we have estimates for the conditional mean estimates, we can construct point-wise confidence intervals for the conditional mean estimates as we will in the Section 5.5.

5.5 Point-wise Confidence Interval

The aim of this section is to provide a procedure/algorithm that can be used to construct point-wise confidence interval for the conditional mean estimates. We construct point-wise confidence interval for the conditional mean estimate as follows:

1. Compute the kernel smother $\hat{m}_h$ at distinct points $x_1, x_2, \ldots, x_k$.
2. Construct an estimate of $\sigma^2(x)$ as explained in Section 5.4.
3. Take $Z_{1-\alpha/2}$, the $(100-\alpha/2)$-quantile of the normal distribution and let

$$\text{Lower} = \hat{m}_h - \frac{Z_{1-\alpha/2} R(K)^{1/2} \hat{\sigma}(x)}{(nh\hat{f}(x))^{1/2}} \quad \text{and} \quad \text{Upper} = \hat{m}_h + \frac{Z_{1-\alpha/2} R(K)^{1/2} \hat{\sigma}(x)}{(nh\hat{f}(x))^{1/2}}$$

4. Draw the interval [Lower,Upper] around $\hat{m}_h(x)$ at the k distinct points $x_1, x_2, \ldots, x_k$.

Using the above procedure/algorithm, we construct point-wise confidence interval for the conditional mean estimates shown in Figure 4.

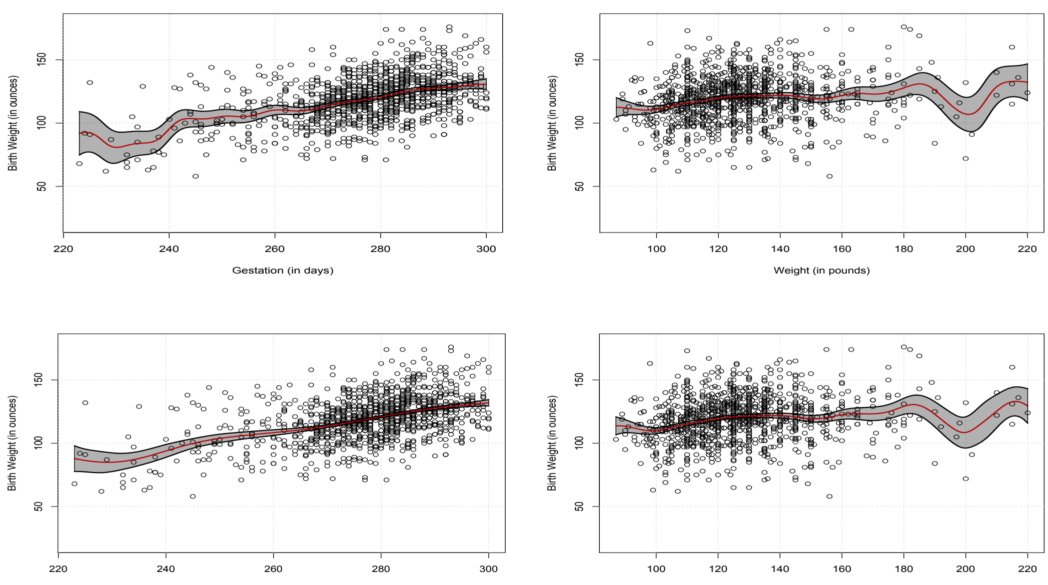

Figure 6: *Top Left Panel:* the gray shaded area represents the 95% confidence bands of the estimated conditional mean from the *top left panel* in Figure 4. *Top Right Panel:* the gray shaded area represents the 95% confidence bands of the estimated conditional mean from the *top right panel* in Figure 4 . *Bottom Left Panel:* the gray shaded area represents the 95% confidence bands of the estimated conditional mean from the *bottom left panel* in Figure 4. *Bottom Right Panel:* the gray shaded area represents the 95% confidence bands of the estimated conditional mean from the *bottom right panel* in Figure 4.

In Figure 6, in each of the 95% confidence bands, we see that there are some regions where the 95% confidence bands is wider. These regions represent the regions where we have less data points than other regions. Next we consider another approach to construct confidence bands for the conditional mean estimates based on bootstrap procedures.

5.6 Bootstrap Confidence Bands

Another method of constructing confidence bands is based on bootstrap procedures. The bootstrap is a resampling technique that prescribes taking "bootstrap samples" using the same random mechanism that generated the data. In the case of fixed design models with homoscedastic error structure we can construct bootstrap confidence bands as follows:

1. Compute the kernel smoother $\hat{m}_h$ from the data set $\{(X_i, Y_i)\}_{i=1}^n$.
2. Compute $\{\hat{\varepsilon}_i\}_{i=1}^n = Y_i - \hat{m}_h(X_i)$.
3. Resample with replacement $\{\hat{\varepsilon}_i^*\}_{i=1}^n$ from $\{\hat{\varepsilon}_i\}_{i=1}^n$.
4. Compute $Y_i^* = \hat{m}_h(X_i) + \hat{\varepsilon}_i^*$
5. Compute the kernel smoother $\hat{m}_h^*$ from the data set $\{(X_i, Y_i^*)\}_{i=1}^n$.
6. Repeat steps 3 to 5 B times, where B is the number of bootstrap samples.
7. Find Lower* as the $\alpha/2$ empirical quantile of the B bootstrap estimates $\hat{m}_h^*$. Similarly, find Upper* as the $1 - \alpha/2$ empirical quantile of the B bootstrap estimates.
8. Draw the interval [Lower*,Upper*] around $\hat{m}_h(x)$ at the k distinct points $x_1, x_2, \ldots, x_k$.

The bootstrap of residuals can also be applied in the stochastic design setting. A slightly different approach would be to resample the residuals from a two-point distribution which has mean zero, variance equal to the square of the residual, and third moment equal to the cube of the residual. This procedure is called *Wild Bootstrapping*

1. Compute the kernel smoother $\hat{m}_h$ from the data set $\{(X_i, Y_i)\}_{i=1}^n$.
2. Let $\{\hat{\varepsilon}_i\}_{i=1}^n = Y_i - \hat{m}_h(X_i)$ be the observed residual at point X_i. Then we define $\hat{\varepsilon}_i^*$ to be a random variable having a two-point distribution, $\hat{G}_i$, where

   $$\hat{G}_i = pa + (1-p)b$$

   is defined through the three parameters a, b and p. Some algebra reveals that the parameters a, b and p are given by

   $$a = \hat{\varepsilon}_i(1-\sqrt{5})/2 \quad b = \hat{\varepsilon}_i(1+\sqrt{5})/2 \quad p = (\sqrt{5}+1)/2\sqrt{5}$$

3. Resample $\{\varepsilon_i^*\}_{i=1}^n$ from $\hat{G}_i$
4. Compute $Y_i^* = \hat{m}_h(X_i) + \varepsilon_i^*$
5. Compute the kernel smoother $\hat{m}_h^*$ from the data set $\{(X_i, Y_i^*)\}_{i=1}^n$.
6. Repeat steps 3 to 5 B times, where B is the number of bootstrap samples.
7. Find Lower* as the $\alpha/2$ empirical quantile of the B bootstrap estimates $\hat{m}_h^*$. Similarly, find Upper* as the $1 - \alpha/2$ empirical quantile of the B bootstrap estimates.
8. Draw the interval [Lower*,Upper*] around $\hat{m}_h(x)$ at the k distinct points $x_1, x_2, \ldots, x_k$.

Using the above procedure/algorithm, we construct wild-bootstrap based confidence bands for the conditional mean estimates shown in Figure 4.

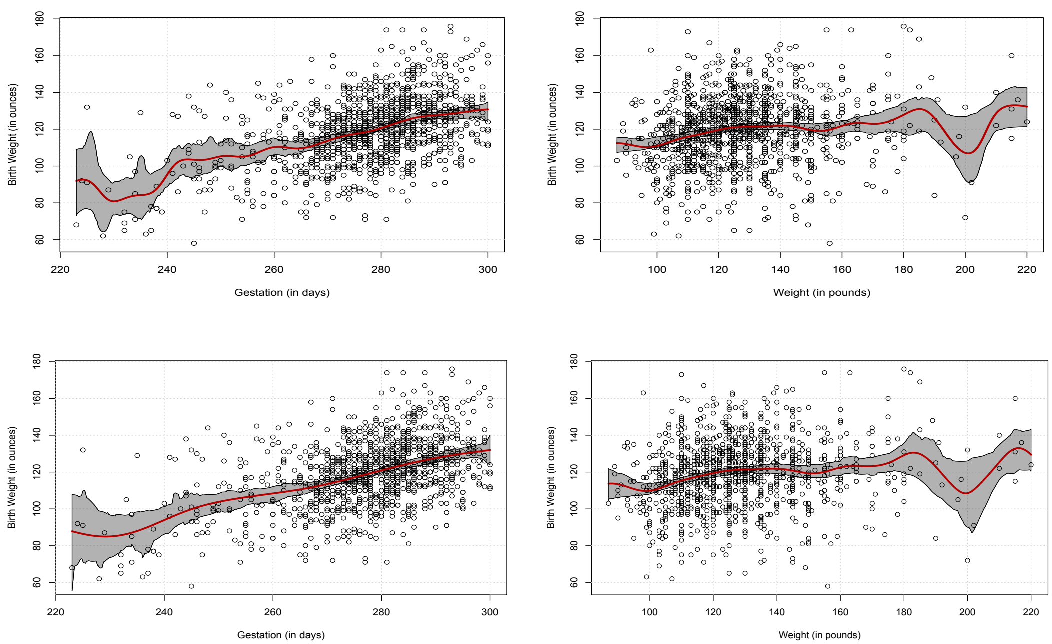

Figure 7: *Top Left Panel:* the gray shaded area represents the 95% wild-bootstrap based confidence bands of the estimated conditional mean from the *top left panel* in Figure 4. *Top Right Panel:* the gray shaded area represents the 95% wild-bootstrap based confidence bands of the estimated conditional mean from the *top right panel* in Figure 4. *Bottom Left Panel:* the gray shaded area represents the 95% wild-bootstrap based confidence bands of the estimated conditional mean from the *bottom left panel* in Figure 4. *Bottom Right Panel:* the gray shaded area represents the 95% wild-bootstrap based confidence bands of the estimated conditional mean from the *bottom right panel* in Figure 4.

In Figure 7, in each of the 95% wild-bootstrap based confidence bands, we see that there are some regions where the 95% confidence bands is wider. These regions represent the regions where we have less data points than other regions. These behaviors agrees with what it shown in Figure 6. So far, we have only considered two nonparametric regression models: one having `bwt` as the response and `gestation` as the predictor, and another having `bwt` as the response and `weight` as the predictor. By analyzing those two models, it seems that there is a relationship between the response and each of the potential predictors. Before we consider a nonparametric model having `bwt` as the response and `gestation` and `weight` as the predictors, we need to conduct a hypothesis test to see if there is a relationship between the response and each of the potential predictors.

5.7 Testing Significance of Predictors

In order to build a good nonparametric regression model, we may only consider predictors that have a relationship with response. In other words, we need to conduct a hypothesis test to check if there is a relationship between the response and a potential predictor. Consider the following nonparametric model

$$Y_j = m(X_j) + \varepsilon_j$$

where Y is the response and X is the predictor. We formally want to test

$$\begin{aligned} H_0 &: m(X_j) = C \\ H_a &: m(X_j) \neq C \end{aligned}$$

where C is a constant. Notice that if we fail to reject the null hypothesis, we conclude that there is not relationship between X and Y, so X should not be included in the nonparametric model. On the other hand, if we reject the null hypothesis, we conclude that there is a relationship between X and Y, so X should be included in the nonparametric model. In order to conduct the hypothesis test, we consider a L_2-distance test statistics. The considered statistics is given by

$$T_n = nh^{1/2} \int (\hat{m}_h(X_j) - \hat{C})^2 \pi(X_j) dX_j \tag{19}$$

where $\pi(X_j)$ is a weight function. We can approximate T_n as follows

$$T_n = nh^{1/2} \frac{1}{n} \sum_{j=1}^{n} (\hat{m}_h(X_j) - \hat{C})^2 \tag{20}$$

We test the significance of a predictor as follows:

1. Compute the kernel smoother $\hat{m}_h$ from the data set $\{(X_i, Y_i)\}_{i=1}^n$.
2. Compute T_n, given by (20), using $\hat{C} = \overline{Y}$.
3. Resample $\{\varepsilon_i^*\}_{i=1}^n$ using *wild bootstrap* as explained in Section 5.6.

4. Compute $Y_i^* = \hat{m}_h(X_i) + \varepsilon_i^*$
5. Compute the kernel smoother $\hat{m}_h^*$ from the data set $\{(X_i, Y_i^*)\}_{i=1}^n$.
6. Compute T_n^*, given by (20), using $\hat{m}_h^*$ and $\hat{C}^* = \overline{Y}^*$.
7. Repeat steps 3 to 6 B times, where B is the number of bootstrap samples.
8. Reject H_0 if $T_n > T_{n,(1-\alpha)}^*$

Using the above procedure/algorithm, we check for the significance of `gestation` and `weight`.The result of the hypothesis tests based on 1000 bootstraps are shown in Table 3.

Table 3: Significance of Predictors

Variable	T_n	$T_{n,0.95}^*$
`gestation`	21956.1	19876.3
`weight`	38722.67	35321.46

Based on the results shown in Table 3, we conclude that there is a relationship between the response, `bwt`, and each of the predictors `gestation` and `weight`. Therefore, both `gestation` and `weight` should be included in a nonparametric model in which `bwt` is the response.

6 Assessing the Effect of Smoking on Babies' Weight

In the previous section, we conclude that there is a relationship between the response, `bwt`, and each of the predictors `gestation` and `weight`. However, up to this point, we have not taken into account the smoking status in previous our analysis. In this section, we study the effect of smoking on babies's weight. In other words, we want to check there is a difference in the weights of babies, who mother is a smoker, and babies who mother is not a smoker. We start by comparing the nonparametric regression curves taking into account the smoking status as shown in Figure 8.

In Figure 8, in all the panels the general behavior is that the conditional mean estimate of smoker moms does not fall in the 95% confidence bands of the conditional mean estimate of non-smoker moms and vice-versa. However, there are a couple regions where the conditional mean estimate of smoker moms does fall in the 95% confidence bands of the conditional mean estimate of non-smoker moms and vice-versa; these regions are the regions where we have less data values. After exploring the plots shown in Figure 8, we may conclude that there is a difference between the smoker and non-smoker regression curves; however, we as statisticians we need to make conclusions based on statistical analysis/procedures. In other words, we need to conduct a hypothesis test to see if there is a significance difference between the two regression curves.

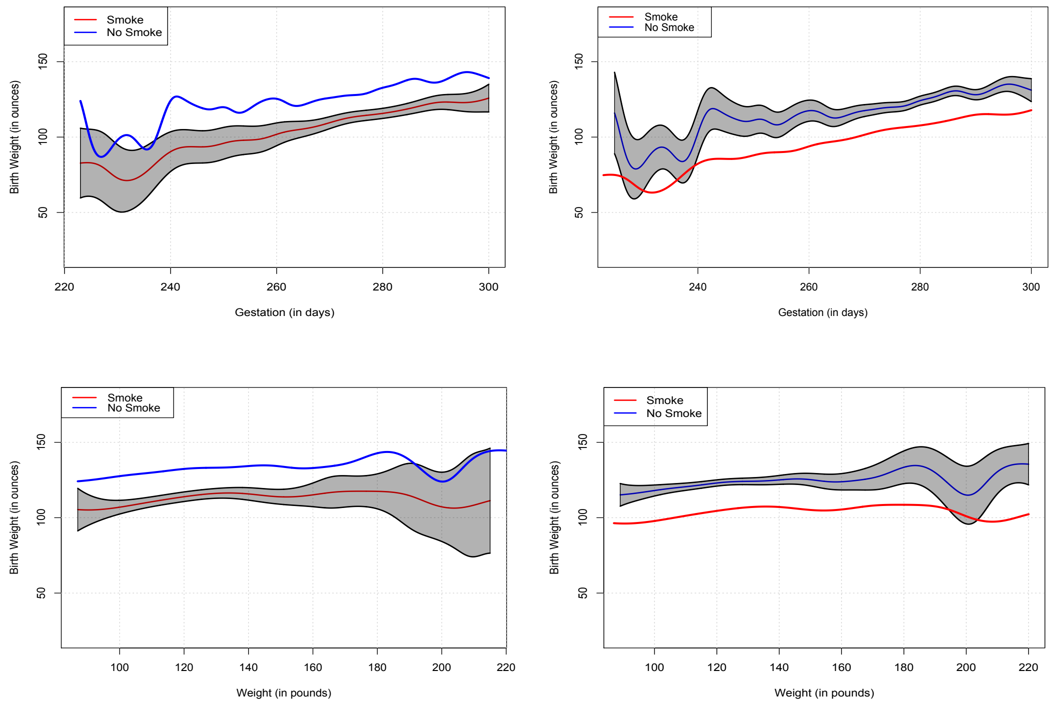

Figure 8: *Top Left Panel:* the red curve represents the conditional mean estimate of smoker mom, and the gray shaded area represents its 95% confidence bands. The blue curve represents the conditional mean estimate of non-smoker moms. *Top Right Panel:* the blue curve represents the conditional mean estimate of non-smoker mom, and the gray shaded area represents its 95% confidence bands. The red curve represents the conditional mean estimate of smoker moms. *Bottom Left Panel:* the red curve represents the conditional mean estimate of smoker mom, and the gray shaded area represents its 95% confidence bands. The blue curve represents the conditional mean estimate of non-smoker moms. *Bottom Right Panel:* the blue curve represents the conditional mean estimate of non-smoker mom, and the gray shaded area represents its 95% confidence bands. The red curve represents the conditional mean estimate of smoker moms.

We consider the procedure proposed by Neumeyer and Dette (2003). Let

$$Y_{ij} = f_i(X_{ij}) + \sigma_i(X_{ij})\varepsilon_{ij}, \quad j = 1, \ldots, n_i, \;\; i = 1, 2$$

where $X_{ij}(j = 1, \ldots, n_i)$ are independent observations with positive density r_i on the interval [0,1] and ε_{ij} are independent identically distributed random variables with mean 0 and variance 1. We want to test

$$H_0 : f_1 = f_2$$
$$H_a : f_1 \neq f_2$$

It is assume that f_1, f_2, and the densities r_1 and r_2 are suppose to be $d(\geq 2)$ times continuous differentiable. Note that

$$\hat{r}_i(x) = \frac{1}{n_i h} \sum_{j=1}^{n_i} K\left(\frac{x - X_{ij}}{h}\right)$$

denotes the density estimator of the i-th sample and

$$\hat{r}(x) = \frac{n_1}{N}\hat{r}_1(x) + \frac{n_2}{N}\hat{r}_2(x)$$

where $N = n_1 + n_2$ and $\hat{r}(x)$ is the density estimator of the combined sample. The Nayadara-Watson estimator of the combined sample is given by

$$\hat{f}(x) = \frac{1}{Nh}\sum_{i=1}^{2}\sum_{j=1}^{n_i} K\left(\frac{x - X_{ij}}{h}\right) Y_{ij}\frac{1}{\hat{r}(x)} = \frac{(n_1/N)\hat{r}_1(x)\hat{f}_1(x) + (n_2/N)\hat{r}_2(x)\hat{f}_2(x)}{\hat{r}(x)}$$

where

$$\hat{f}_i(x) = \frac{1}{n_i h}\sum_{j=1}^{n_i} K\left(\frac{x - X_{ij}}{h}\right) Y_{ij}\frac{1}{\hat{r}_i(x)}$$

Note that under the null hypothesis of equal regression curves we have $f_1 = f_2 = f$. For $i = 1, 2$ we define residuals

$$e_{ij} = \frac{n_{3-i}}{N}(Y_{ij} - \hat{f}(X_{ij}))\hat{r}(X_{ij})\hat{r}_{3-i}(X_{ij})$$
$$f_{ij} = \frac{N}{n_i}(Y_{ij} - \hat{f}(X_{ij}))/\hat{r}_i(X_{ij})$$

and consider the marked empirical processes

$$\hat{R}_N^{(1)}(t) = \frac{1}{N}\sum_{j=1}^{n_1} e_{1j} I(X_{1j} \le t) - \frac{1}{N}\sum_{j=1}^{n_2} e_{2j} I(X_{2j} \le t) \tag{21}$$

$$\hat{R}_N^{(2)}(t) = \frac{1}{N}\sum_{j=1}^{n_1} f_{1j} I(X_{1j} \le t) - \frac{1}{N}\sum_{j=1}^{n_2} f_{2j} I(X_{2j} \le t) \tag{22}$$

Neumeyer and Dette (2003) proposed a resampling procedure based on wild bootstrap to conduct the hypothesis of equal regression curves as follows:

1. Compute the kernel smoother $\hat{f}_g(X_{ij})$, where $\hat{f}_g(X_{ij})$ denotes the Nayadara-Watson estimator of the total sample using bandwidth g.
2. Compute $\hat{\varepsilon}_{ij} = Y_{ij} - \hat{f}_g(X_{ij})$.
3. Compute $K_N^{(i)} = \sup_{t\in[0,1]} |\hat{R}_N^{(i)}(t)|$, $i = 1, 2$, where $\hat{R}_N^{(1)}(t)$ is given by (21) and $\hat{R}_N^{(2)}(t)$ is given by (22).
4. Resample $\{\varepsilon_i^*\}_{i=1}^n$ using *wild bootstrap* as explained in Section 5.6.
5. Compute $Y_i^* = \hat{m}_h(X_i) + \varepsilon_i^*$

6. Compute the corresponding marked empirical processes

$$\hat{R}_N^{(1)*}(t) = \frac{1}{N}\sum_{l=1}^{2}\sum_{j=1}^{n_l}(-1)^{l-1}(Y_{lj}^* - \hat{f}_h^*(X_{lj}))\hat{r}_h(X_{lj})\frac{n_{3-l}}{N}\hat{r}_{3-l,h}(X_{lj})I(X_{lj} \leq t)$$

$$\hat{R}_N^{(2)*}(t) = \frac{1}{N}\sum_{l=1}^{2}\sum_{j=1}^{n_l}(-1)^{l-1}(Y_{lj}^* - \hat{f}_h^*(X_{lj}))\frac{N}{n_l}\frac{1}{\hat{r}_{l,h}(X_{lj})}I(X_{lj} \leq t)$$

7. Compute $K_N^{*(i)} = \sup_{t\in[0,1]} |\hat{R}_N^{*(i)}(t)|$, $i = 1, 2$.

8. Repeat steps 4 to 7 B times, where B is the number of bootstrap samples.

9. Reject H_0 if $K_n^{(i)} > K_{n,(1-\alpha)}^{*(i)}$

Note that many of the technical details of the hypothesis of equal regression curves have been skipped. For more details for the above procedure/algorithm see Neumeyer and Dette (2003). Using the above procedure/algorithm, we check for the equality of regression curves. The result of the hypothesis tests based on 1000 bootstraps are shown in Table 4.

Table 4: Testing Equality of Regression Curves

Variable	$K_n^{(2)}$	$K_{n,0.95}^{*(2)}$
`gestation`	1096.15	987.34
`weight`	2087.67	1957.64

Based on the results shown in Table 4, we conclude that the regression curve for non-smoker moms is different from smoker moms. This results show that smoking affect the baby's weight. Thus, from the statistics point of view, we should include the smoking status in our model. In the next section, we consider a semi-parametric model in which the smoking status is included.

7 Smoothing for Semi-parametric Models

In the top left panel of Figure 4, we see that there is positive relationship between `gestation` and `bwt`. On the other hand, in the top right panel of Figure 4, there is no a clear relationship between `weight` and `bwt`; it seems that there is some sort of nonlinear relationship. Based on analyses presented in the previous sections, we then consider a semi-parametric partially linear single index model as

$$Y_i = \mathbf{z}_i^T\boldsymbol{\beta} + g(X_i) + \varepsilon_i \tag{23}$$

where $\boldsymbol{\beta}$ is the vector of unknown parameters and g is an unknown function, $\mathtt{z} = (\mathtt{gestation}, \mathtt{smoke})$, and $X = \mathtt{weight}$. We estimate the vector of parameters, β, and the unknown function, g, using the procedure given in class. Let

$$W_{nj}(x) = \frac{K\left(\frac{x-X_j}{h}\right)}{\sum_j K\left(\frac{x-X_j}{h}\right)}$$

We then center X_i, Y_i and ε_i as follows

$$\tilde{Y}_i = Y_i - \sum_j W_{nj}(X_i)Y_j$$
$$\tilde{\mathbf{z}}_i = \mathbf{z}_i - \sum_j W_{nj}(X_i)\mathbf{z}_j$$
$$\tilde{\varepsilon}_i = \varepsilon_i - \sum_j W_{nj}(X_i)\varepsilon_j$$

Then the estimated $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum_i \tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^T\right)^{-1} \frac{1}{n}\sum_i \tilde{\mathbf{z}}_i\tilde{Y}_i \quad \text{and} \quad \hat{g}(x) = \sum_i W_{ni}(x)\{Y_i - \mathbf{z}_i^T\hat{\boldsymbol{\beta}}\}$$

Using the above procedure, the estimated $\boldsymbol{\beta}$ is shown in Table 5.

Table 5: Estimated Coefficient of Semi-parametric Model

	`gestation`	`smoke`
$\hat{\boldsymbol{\beta}}$	0.619	-8.15

7.1 Assessing Goodness of Fit of the Semi-Parametric Partially Linear Model

After proposing a semi-parametric model and estimating the vector of coefficients, we need to conduct a hypothesis test to check if the proposed semi-parametric model is appropriate for the data. We formally want to test

$$H_0 : m = m_\theta$$
$$H_a : m \neq m_\theta$$

where m_θ is give by (23). The test statistics is given by

$$T_n = nh^{d/2}\int\{\hat{m}_h(x) - \tilde{m}_{\hat{\theta}}(x)\}^2 w(x)dx$$

where h is the bandwidth of the nonparametric regression, $\hat{m}_h(x)$, $\hat{m}_{\hat{\theta}}$ is the fitted semi-parametric model, $w(x)$ is a weight function, d is the dimension of X_i, and

$$\tilde{m}_{\hat{\theta}}(x) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\hat{m}_{\hat{\theta}}}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)}$$

Note that $\tilde{m}_{\hat{\theta}}(x)$ is a Nayadara-Watson smoother of $\hat{m}_{\hat{\theta}}$. We approximate T_n as follows

$$T_n = h^{d/2} \sum_{i=1}^{2} \{\hat{m}_h(X_i) - \tilde{m}_{\hat{\theta}}(X_i)\}^2 \tag{24}$$

Then, using the procedure/algorithm from class notes, we conduct the hypothesis test as follows

1. Compute the kernel smoother $\hat{m}_{\hat{\theta}}$ from the data set $\{(X_i, Y_i)\}_{i=1}^{n}$.
2. Compute T_n, given by (24).
3. Resample $\{\varepsilon_i^*\}_{i=1}^{n}$ using *wild bootstrap* as explained in Section 5.6.
4. Compute $Y_i^* = \hat{m}_{\hat{\theta}}(X_i) + \varepsilon_i^*$
5. Compute the kernel smoother $\hat{m}_{\hat{\theta}}^*$ from the data set $\{(X_i, Y_i^*)\}_{i=1}^{n}$.
6. Compute T_n^*, given by (24), using $\hat{m}_h^*$.
7. Repeat steps 3 to 6 B times, where B is the number of bootstrap samples.
8. Reject H_0 if $T_n > T_{n,(1-\alpha)}^*$

Using the above procedure/algorithm, we check for appropriateness of the proposed model (23). The result of the hypothesis tests based on 100 bootstraps are shown in Table 6.

Table 6: Goodness of Fit of Semi-Parametric Model

	T_n	$T_{n,0.95}^*$
value	15255.42	20264.69

Based on the results shown in Table 6, we conclude that the proposed semi-parametric partially linear model, given by (23), is appropriate for the data.

8 Conclusions and Further Study

In this project, we examine the effect of smoking on babies' weight by the means of nonparametric methods. We first start by conducting density estimation of each of the continuous variables in the data. After that, we proceed to estimate the joint density of the response, `bwt`, with each of the predictors, `gestation` and `weight`. Then, we proceed to check if each of the predictors were significant or not; the results shows that both predictors are significant. We then consider two nonparametric regression curves: one regression curve for non-smoker mothers and another regression curve for smoker mothers. We formally compare those curves using the test proposed by Neumeyer and Dette H. (2003). The results shows that there is a significant difference between the two regression curves. We then proposed a semi-parametric partially linear model for the data in which the smoking status is included.

Note that we only analyzed a small subset of a much larger data set. It will be interesting to analyze the entire data set and check if we reach the same conclusion. Nonparametric methods are computationally intensive, so another further study will be to develop more efficient efficient procedures/algorithms than the ones presented in this project. Finally, if more covariates were available, it will be interesting to see if there are other factors than may affect babies's weight/health.

References

[1] Chen, S. X. (2015). *Nonparametric Statistics Lecture Notes.*

[2] Hall, P. (1983). *Large Sample Optimality of Least-Square Cross-Validation in Density Estimation.* Ann. Statist. 11 1156-1174.

[3] Neumeyer, N., Dette H. (2003). *Nonparametric Comparison of Regression Curves: An Empirical Process Approach.* Ann. Statist. 31 880 - 920

[4] Sheather, S. J. (2004). *Density Estimation.* Statistical Science 19 No 4, 588 - 597.

[5] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

[6] Simonoff, J. S. (1996). *Smoothing Methods in Statistics.* Springer, New York.

[7] Stone, C.J. (1984). *An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates.* Ann. Statist. 12 1285 - 1297