Effect of Smoking on Babies' Weight

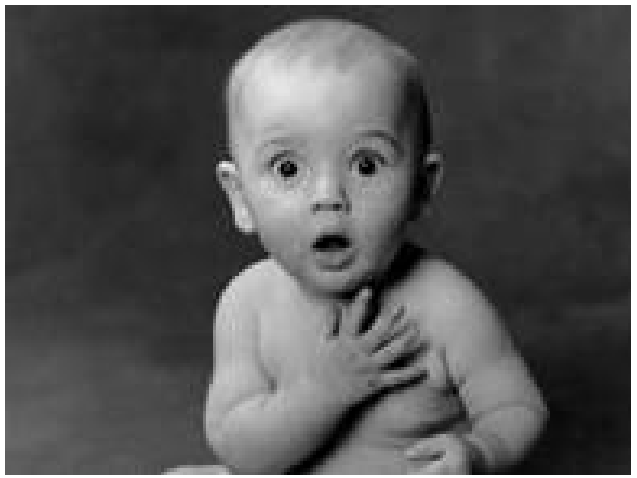
Oscar Aguilar

Department of Statistics
Iowa State University

December 2, 2015

Introduction

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.

The Data

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.

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`)	Mother smokes (1) or not (0)

Purpose of this Project

Goals

- Investigate the effect of smoking status on the baby's weight using nonparametric methods.
- 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.

Univariate Density Estimation

Bandwidth Selection

- **Rule of Thumb:**

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

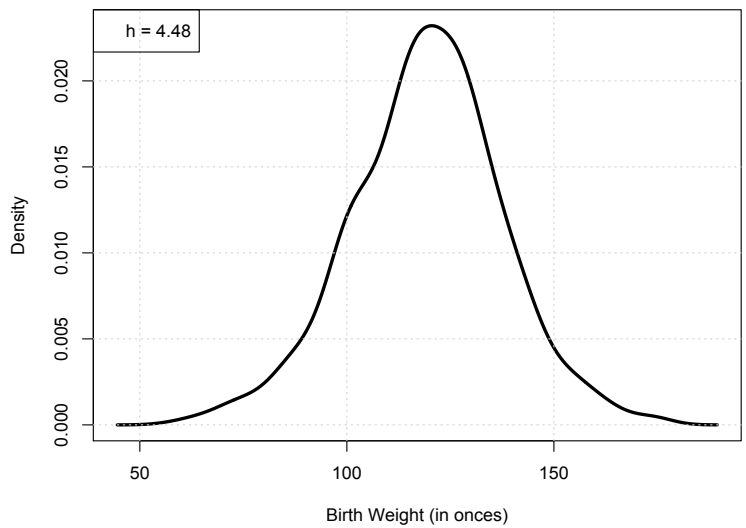
where $\hat{\sigma} = \min\{s, \text{IQR}/1.34\}$ and s stands for sample standard deviation.

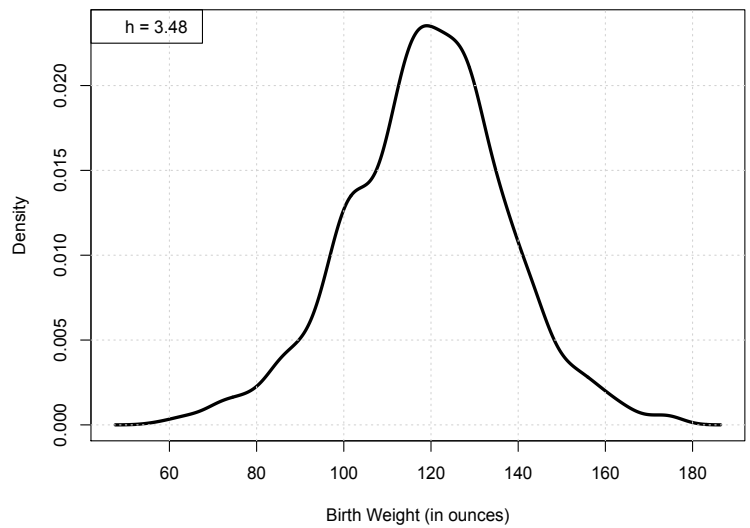
- **Cross-Validation:**

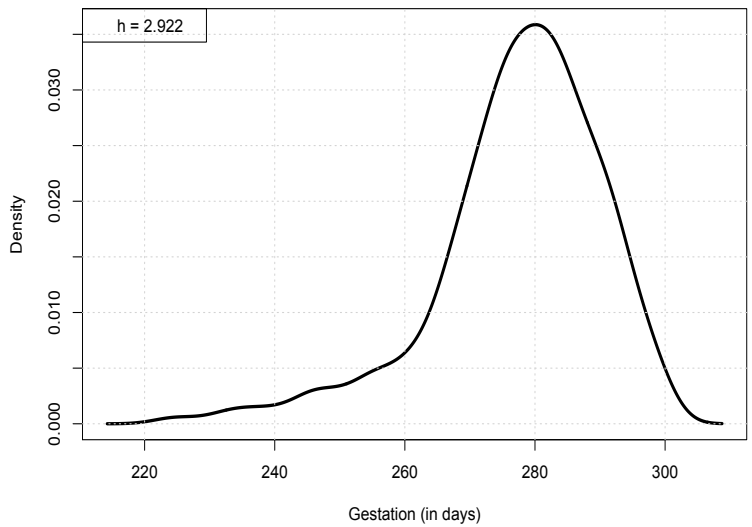
$$\text{LSCV}(h) = \int (\hat{f}_h(x))^2 dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(X_i)$$

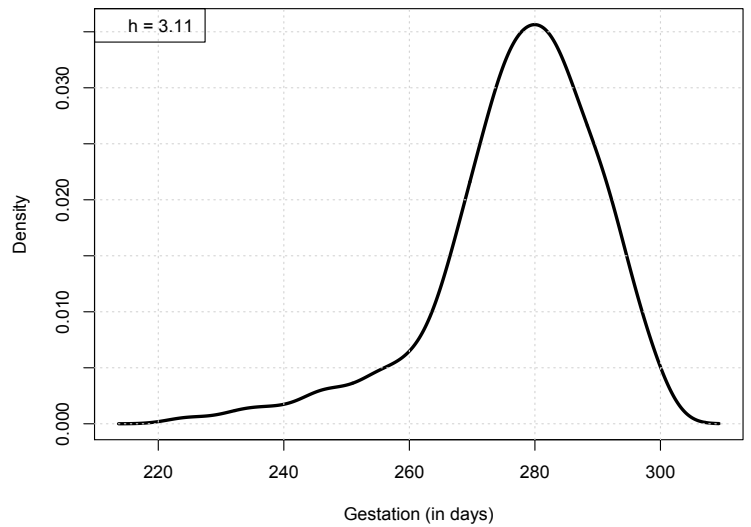
find h such that $\text{LSCV}(h)$ is minimized.

Univariate Density Estimation Cont'd









Joint Density Estimation

Normal Reference Rule

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

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

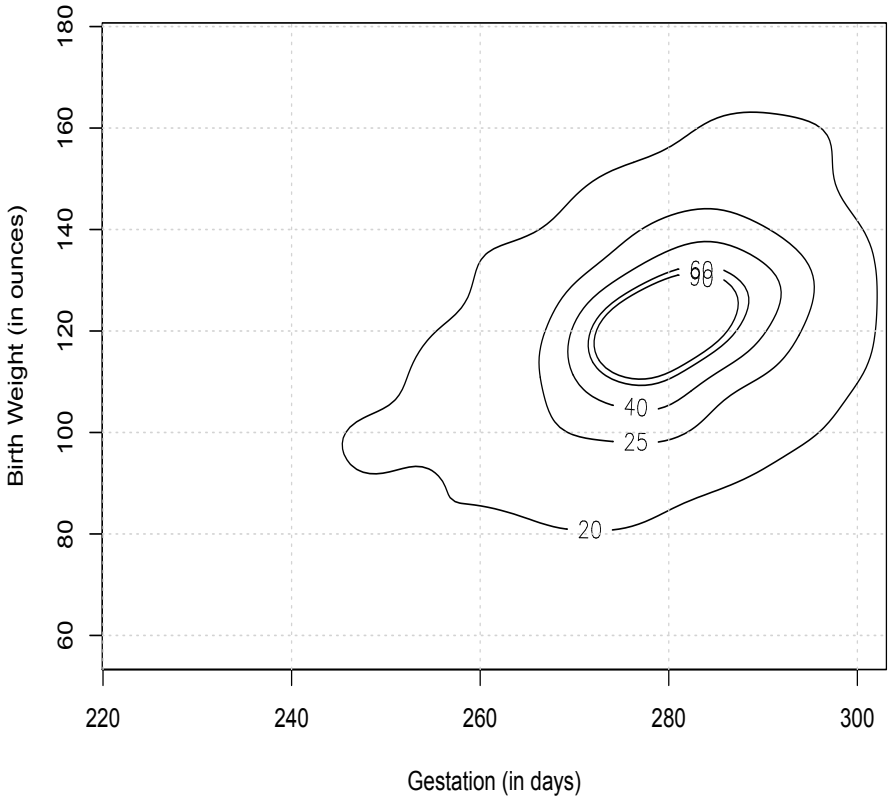
and K is the Gaussian kernel, the practical bandwidth can be chosen as

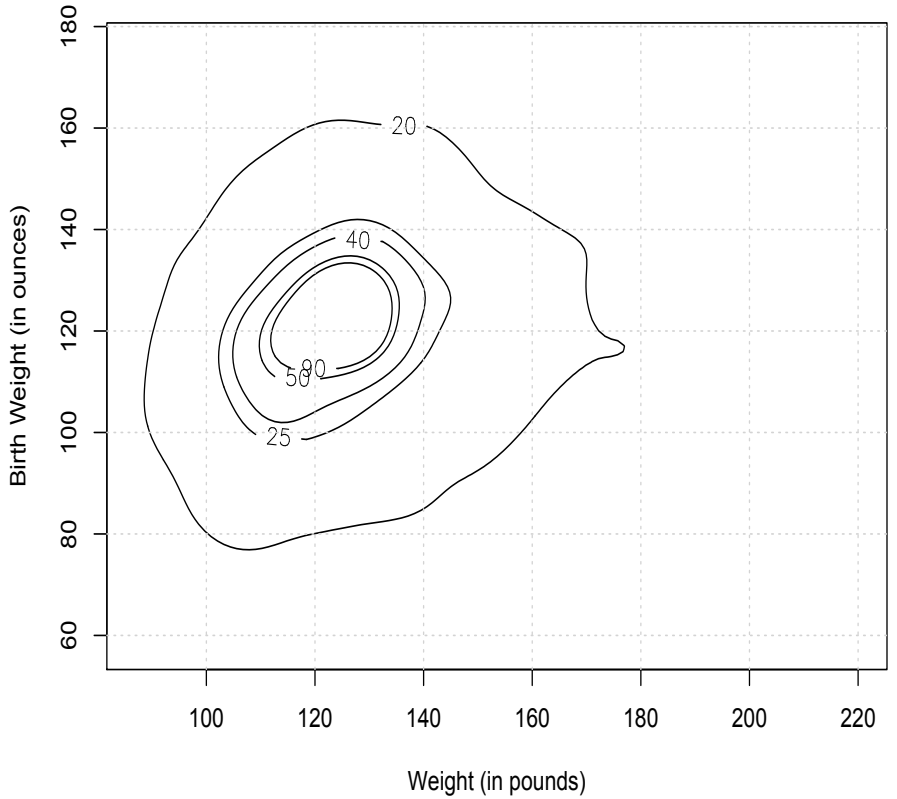
$$h_j = \left(\frac{4}{d+2} \right)^{\frac{1}{d+4}} \hat{\sigma}_j n^{-\frac{1}{d+4}}$$

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

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

Joint Density Estimation Cont'd





Nonparametric Regression

Nonparametric Model

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

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i, \quad i = 1, \ldots, n$$

Nonparametric Regression

Nonparametric Model

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

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i, \quad i = 1, \ldots, n$$

Nayadara-Watson Estimator

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

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

Nonparametric Regression Cont'd

Local Linear Estimator

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

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

Nonparametric Regression Cont'd

Choosing the Bandwidth

Let

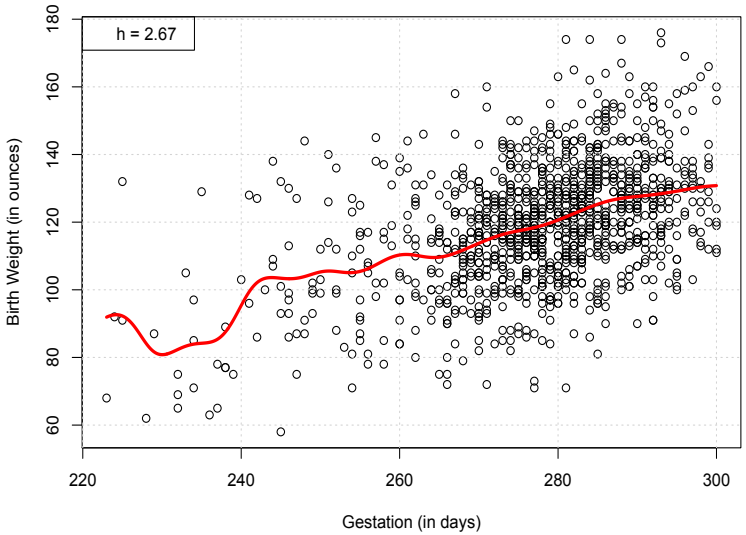
$$\text{CV}(h) = \frac{1}{n}\sum_{j=1}^{n}\left\{Y_j - \hat{m}_{h,j}(X_j)\right\}^2$$

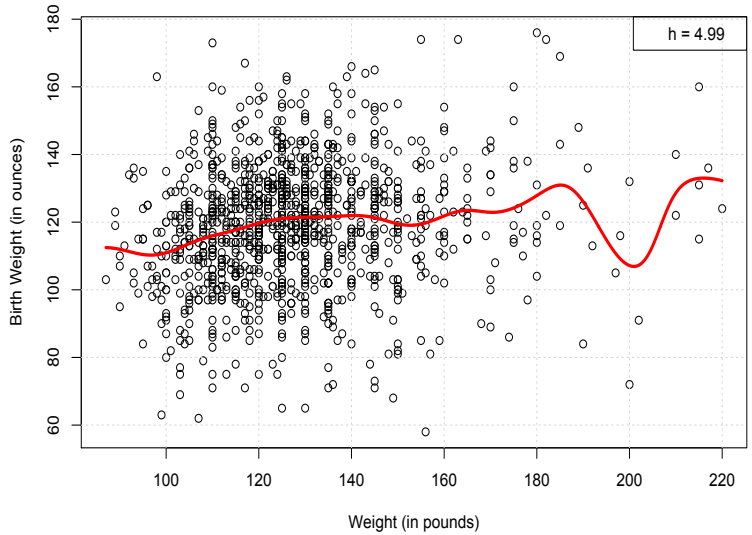
where

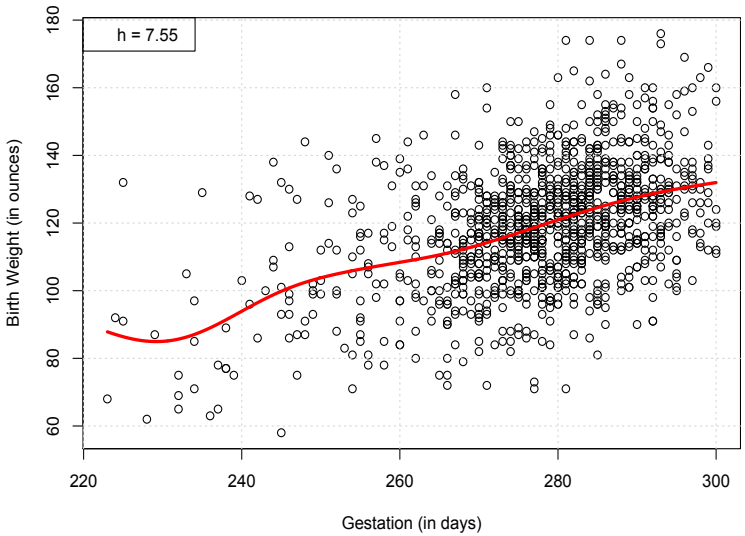
$$\hat{m}_{h,j}(x) = \frac{\sum_{i\neq j} K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{i\neq j} K\left(\frac{x-X_i}{h}\right)}$$

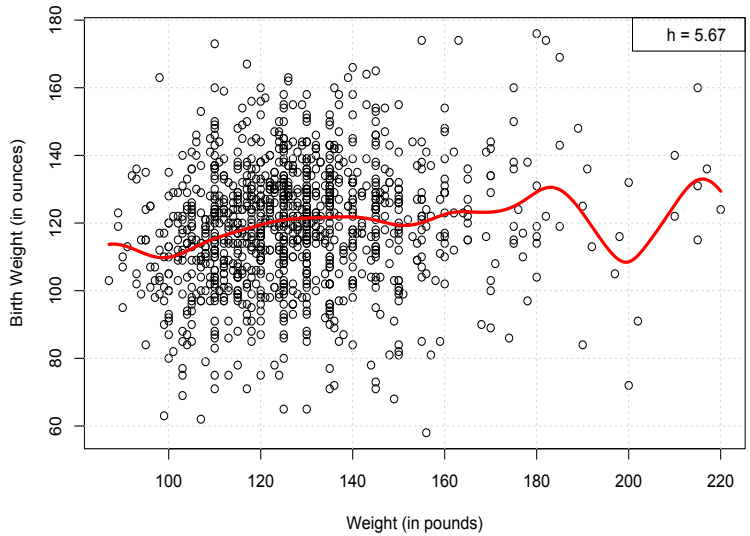
Notice that $\hat{m}_{h,j}(x)$ is the Nayadara-Watson estimator without (X_j, Y_j).

Nonparametric Regression Cont'd









Nonparametric Regression Cont'd

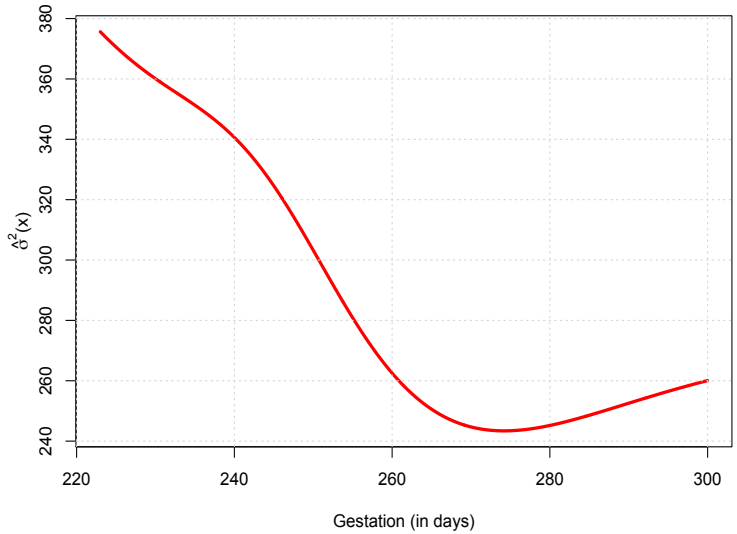
Estimating the Conditional Variance

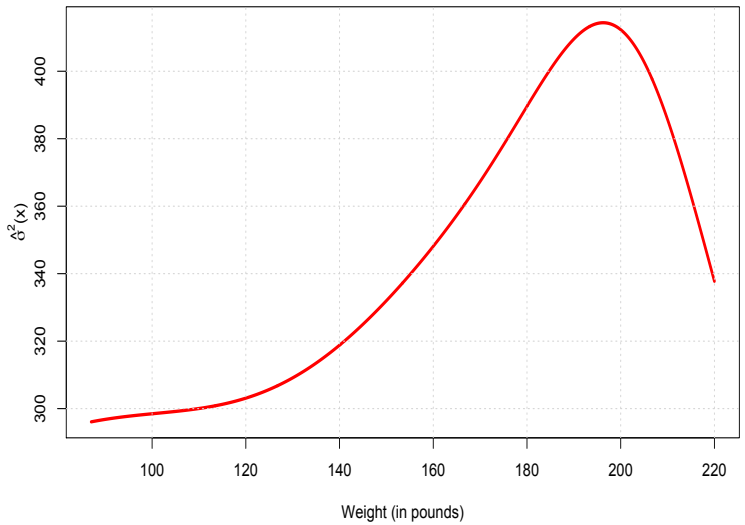
Let $\hat{m}_h$ be a kernel smoother of m based on cross-validation bandwidth h, then we estimate the conditional variance at x as

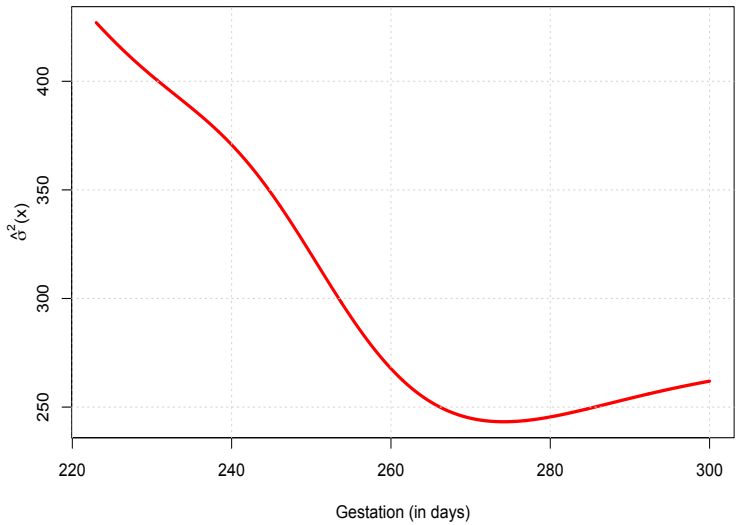
$$\hat{\sigma}^2(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_1}\right)\left\{Y_i - \hat{m}_h(X_i)\right\}^2}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_1}\right)}$$

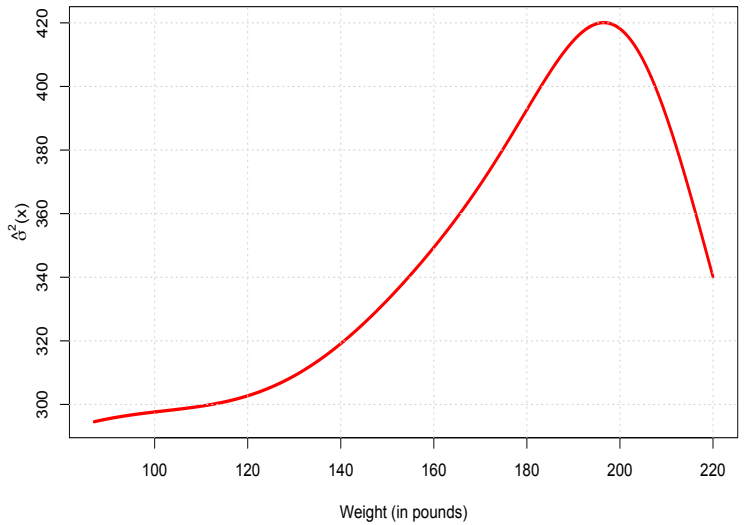
Notice that $\hat{\sigma}^2(x)$ is a Nayadara-Watson estimator on the square of residual $\{Y_i - \hat{m}_h(X_i)\}^2$.

Nonparametric Regression Cont'd









Nonparametric Regression Cont'd

Point-wise Confidence Interval

We construct point-wise confidence interval for the conditional mean estimate as follows:

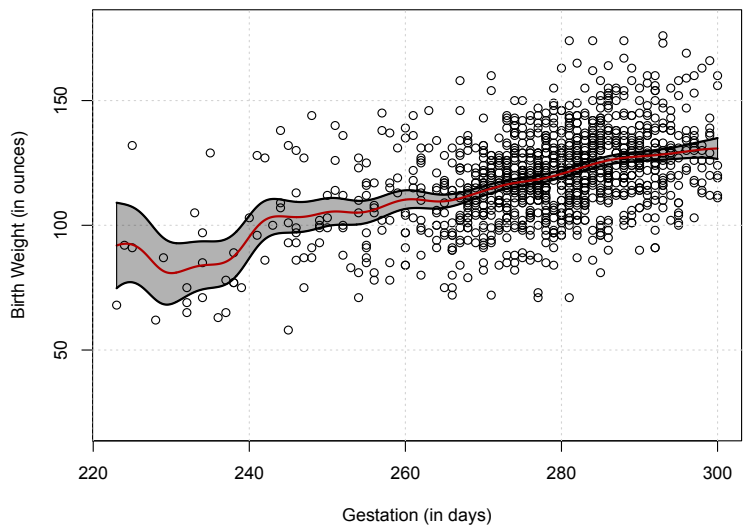
1. Compute the kernel smother $\hat{m}_h$ at distinct points $x_1, x_2, \ldots, x_k$.
2. Construct an estimate of $\sigma^2(x)$.
3. Take $Z_{1-\alpha/2}$, the $(100 - \alpha/2)$-quantile of the normal distribution and let

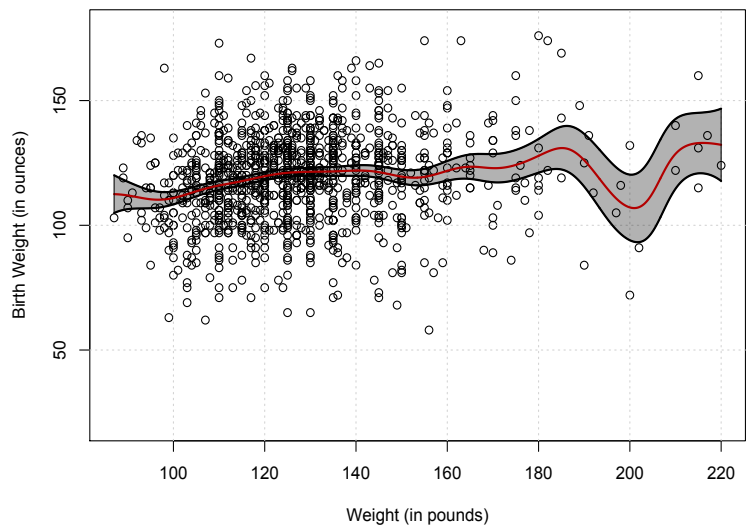
$$\text{Lower} = \hat{m}_h - \frac{Z_{1-\alpha/2} R(K)^{1/2} \hat{\sigma}(x)}{(nh\hat{f}(x))^{1/2}}$$

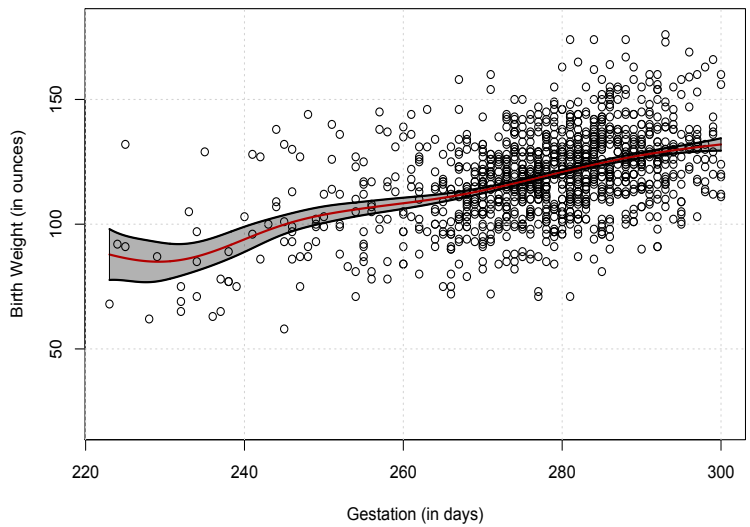
$$\text{Upper} = \hat{m}_h + \frac{Z_{1-\alpha/2} R(K)^{1/2} \hat{\sigma}(x)}{(nh\hat{f}(x))^{1/2}}$$

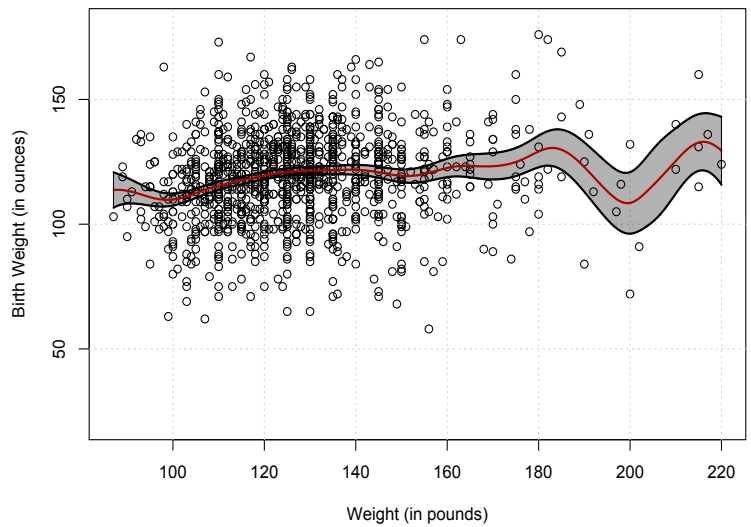
4. Draw the interval [Lower,Upper] around $\hat{m}_h(x)$ at the k distinct points $x_1, x_2, \ldots, x_k$.

Nonparametric Regression Cont'd









Nonparametric Regression Cont'd

Bootstrap Confidence Bands

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. 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

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

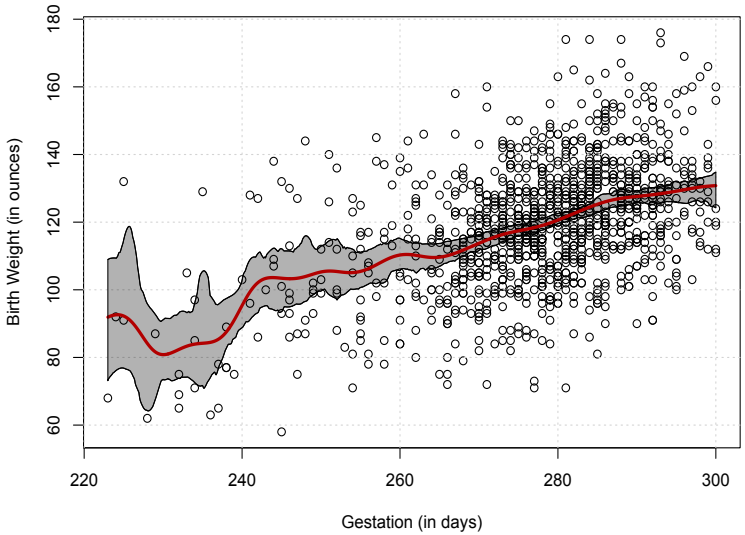
3. Resample $\{\varepsilon_i^*\}_{i=1}^n$ from $\hat{G}_i$

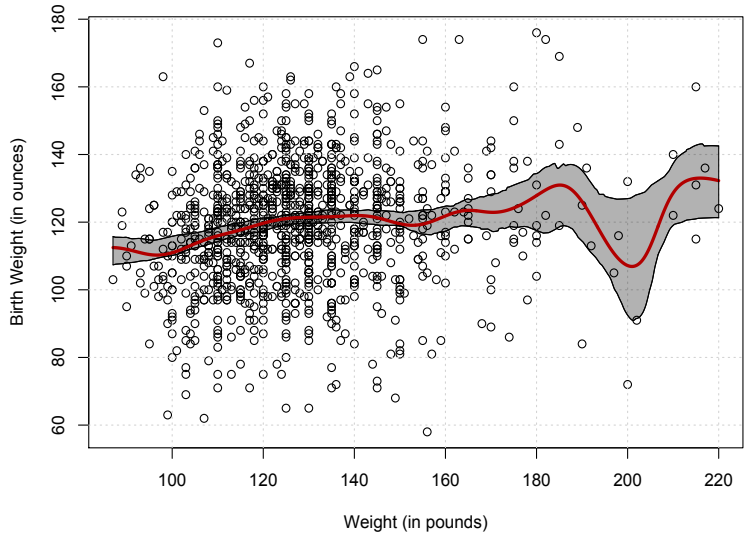
Nonparametric Regression Cont'd

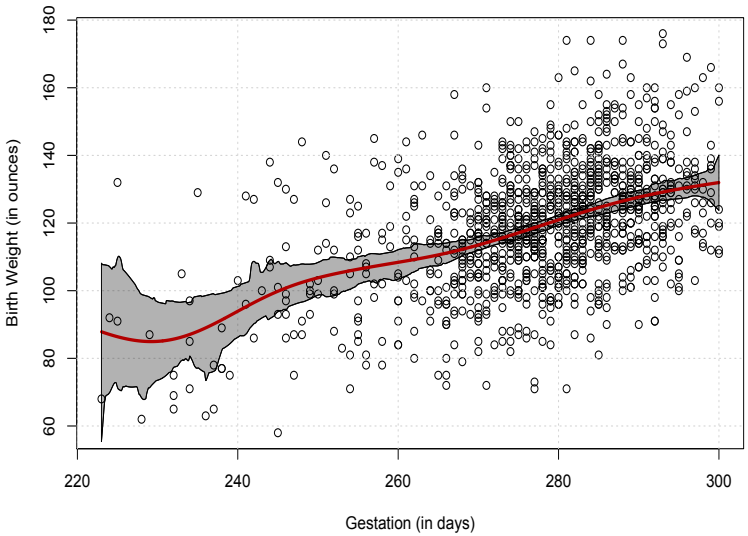
Bootstrap Confidence Bands Cont'd

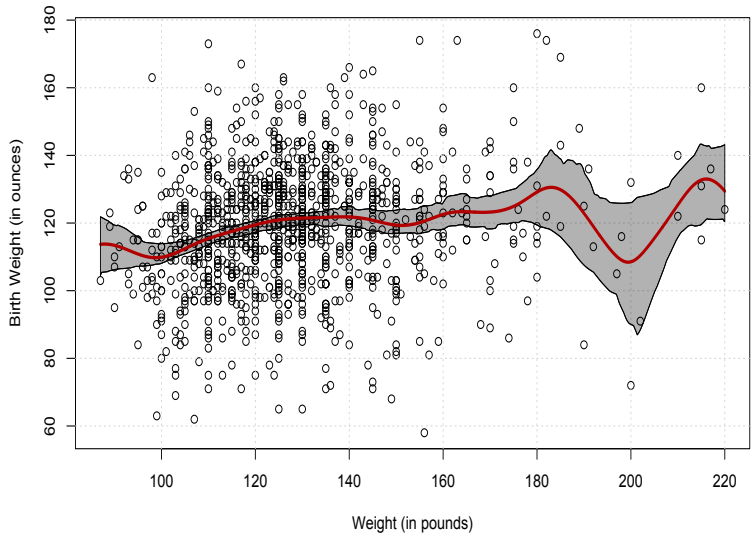
4. Compute $Y_i^* = \hat{m}_h(X_i) + \varepsilon_i^*$
5. Compute the kernel smoother $\hat{m}_h^*$ from the data set $\{(X_i, Y_i^*)\}_{i=1}^n$.
6. Repeat steps 3 to 5 B times, where B is the number of bootstrap samples.
7. Find Lower* as the $\alpha/2$ empirical quantile of the B bootstrap estimates $\hat{m}_h^*$. 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 $x_1, x_2, \ldots, x_k$.

Nonparametric Regression Cont'd









Nonparametric Regression Cont'd

Testing Significance of Predictors

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

$$H_0 : m(X_j) = C$$
$$H_a : m(X_j) \neq C$$

where C is a constant. The test statistic is given by

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

Nonparametric Regression Cont'd

Testing Significance of Predictors Cont'd

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 using $\hat{C} = \overline{Y}$.
3. Resample $\{\varepsilon_i^*\}_{i=1}^n$ using *wild bootstrap* as previously explained.
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^* 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 $T_n > T_{n,(1-\alpha)}^*$

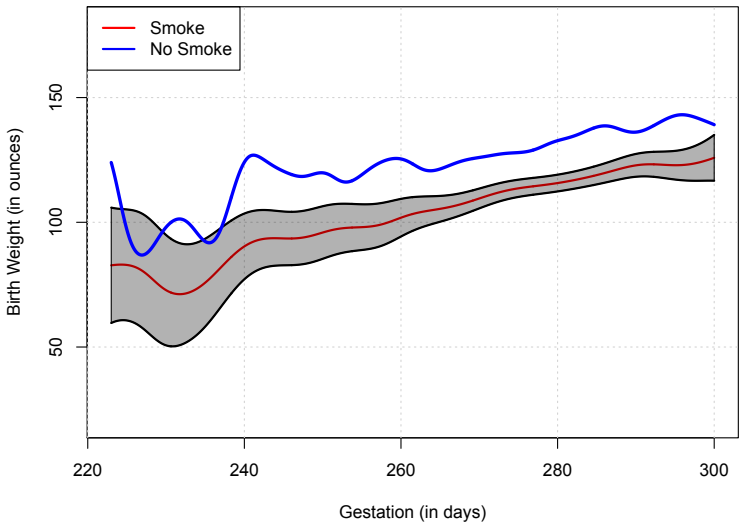
Nonparametric Regression Cont'd

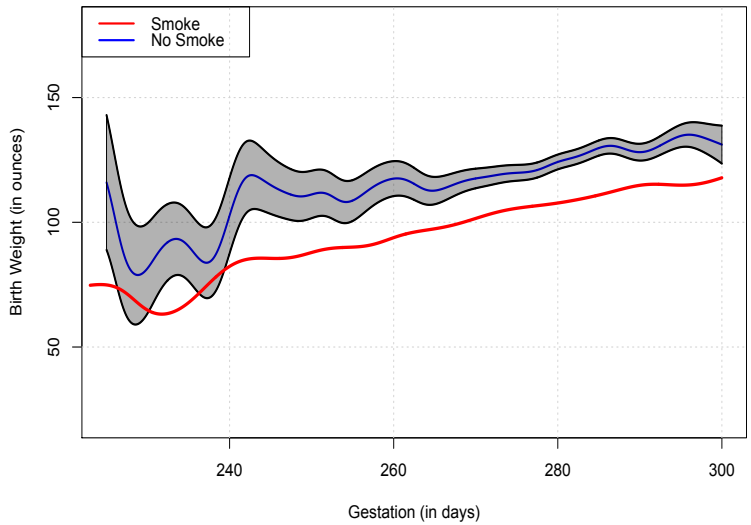
Using the above procedure/algorithm, we check for the significance of `gestation` and `weight`.The result of the hypothesis tests based on 1000 bootstraps are shown in the below table.

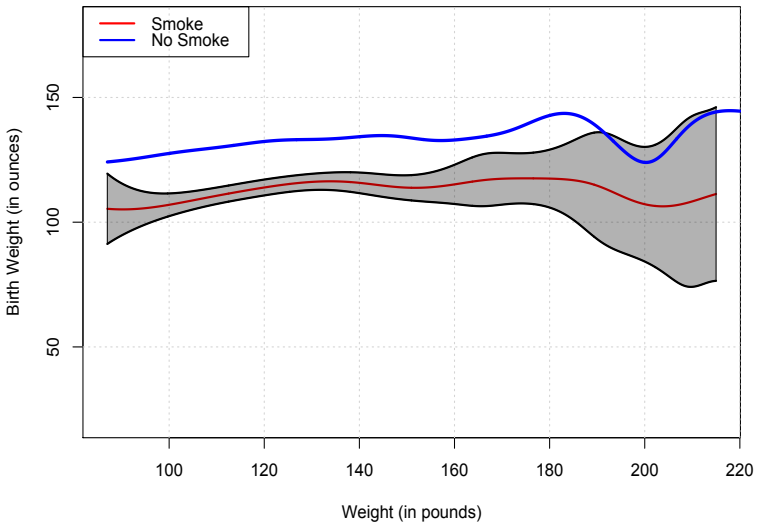
Table : Significance of Predictors

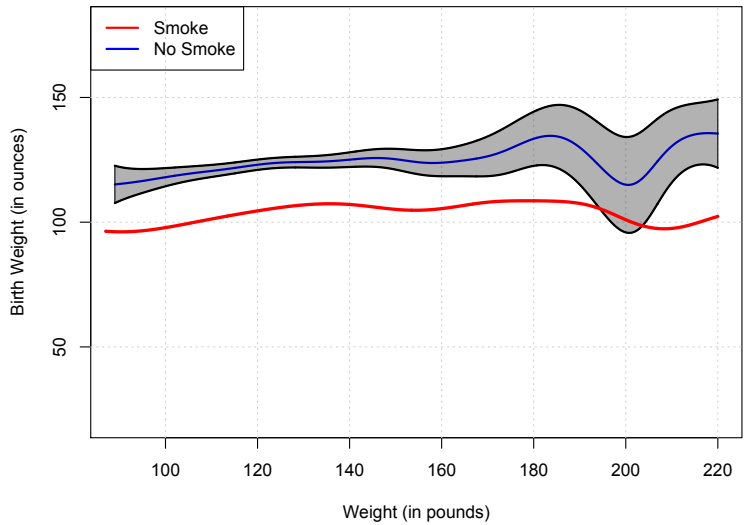
Variable	T_n	$T^*_{n,0.95}$
`gestation`	21956.1	19876.3
`weight`	38722.67	35321.46

Assessing the Effect of Smoking on Babies' Weight









Assessing the Effect of Smoking on Babies' Weight Cont'd

Comparing Regression Curves

We consider the approach 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

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

Assessing the Effect of Smoking on Babies' Weight Cont'd

Comparing Regression Curves Cont'd

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

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

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

where

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

Assessing the Effect of Smoking on Babies' Weight Cont'd

Comparing Regression Curves Cont'd

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

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

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

Assessing the Effect of Smoking on Babies' Weight Cont'd

Comparing Regression Curves Cont'd

Neumeyer and Dette (2003) proposed a resampling procedure based on wild bootstrap 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 $K_N^{(i)} = \sup_{t \in [0,1]} |\hat{R}_N^{(i)}(t)|$, $i = 1, 2$.
4. Resample $\{\varepsilon_i^*\}_{i=1}^n$ using *wild bootstrap* as previously explained.
5. Compute $Y_i^* = \hat{m}_h(X_i) + \varepsilon_i^*$

Assessing the Effect of Smoking on Babies' Weight Cont'd

Comparing Regression Curves Cont'd

6. Compute the corresponding marked empirical processes

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

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

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 $K_n^{(i)} > K_{n,(1-\alpha)}^{*(i)}$

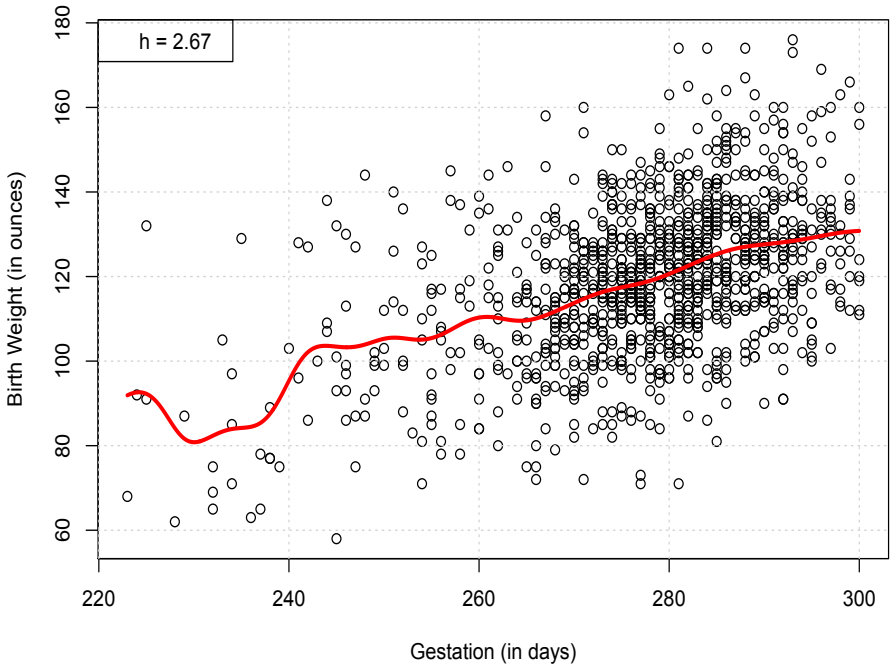
Assessing the Effect of Smoking on Babies' Weight Cont'd

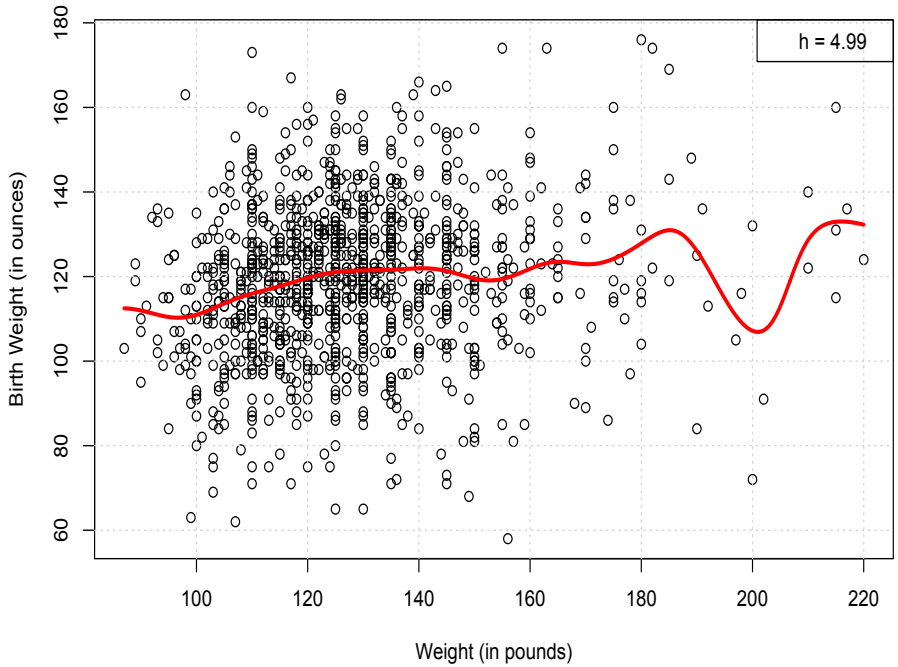
Many of the technical details of the hypothesis of equal regression curves have been skipped. For more details for the above procedure/algorithm see Neumeyer and Dette (2003). Using the above procedure/algorithm, we check for the equality of regression curves. The result of the hypothesis tests based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

Table : Testing Equality of Regression Curves

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

Semi-Parametric Model





Semi-Parametric Model Cont'd

Semi-Parametric Partially Linear Model

We then consider a semi-parametric partially linear single index model as

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

where $\boldsymbol{\beta}$ is the vector of unknown parameters and g is an unknown function, $\mathbf{z} = (\texttt{gestation}, \texttt{smoke})$, and $X = \texttt{weight}$. 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

Semi-Parametric Model Cont'd

Semi-Parametric Partially Linear Model Cont'd

$$\tilde{Y}_i = Y_i - \sum_j W_{nj}(X_i) Y_j$$

$$\tilde{\mathbf{z}}_i = \mathbf{z}_i - \sum_j W_{nj}(X_i) \mathbf{z}_j$$

$$\tilde{\varepsilon}_i = \varepsilon_i - \sum_j W_{nj}(X_i) \varepsilon_j$$

Then the estimated β is given by

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

Semi-Parametric Model Cont'd

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

Table : Estimated Coefficient of Semi-parametric Model

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

Semi-Parametric Model Cont'd

Goodness of Fit

We formally want to test

$$H_0 : m = m_\theta \quad \text{versus} \quad H_a : m \neq m_\theta$$

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

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

Semi-Parametric Model Cont'd

Goodness of Fit Cont'd

We approximate T_n as follows

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

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.
3. Resample $\{\varepsilon_i^*\}_{i=1}^n$ using *wild bootstrap* as previously explained.
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^*, using $\hat{m}_h^*$.

Semi-Parametric Model Cont'd

Goodness of Fit Cont'd

7. Repeat steps 3 to 6 B times, where B is the number of bootstrap samples.
8. Reject H_0 if $T_n > T^*_{n,(1-\alpha)}$

Semi-Parametric Model Cont'd

Using the above procedure/algorithm, we check for appropriateness of the proposed model. The result of the hypothesis tests based on 100 bootstraps are shown in the below table.

Table : Goodness of Fit of Semi-Parametric Partially Linear Model

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

Conclusion & Further Study

- Smoking has an effect on babies' weight.
- The presented results are based on 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 a computationally further study will be to develop more efficient efficient procedures/algorithms than the ones presented in this project.
- If more covariates were available, it will be interesting to see if there are other factors than may affect babies's weight/health.

Thanks!
Questions?