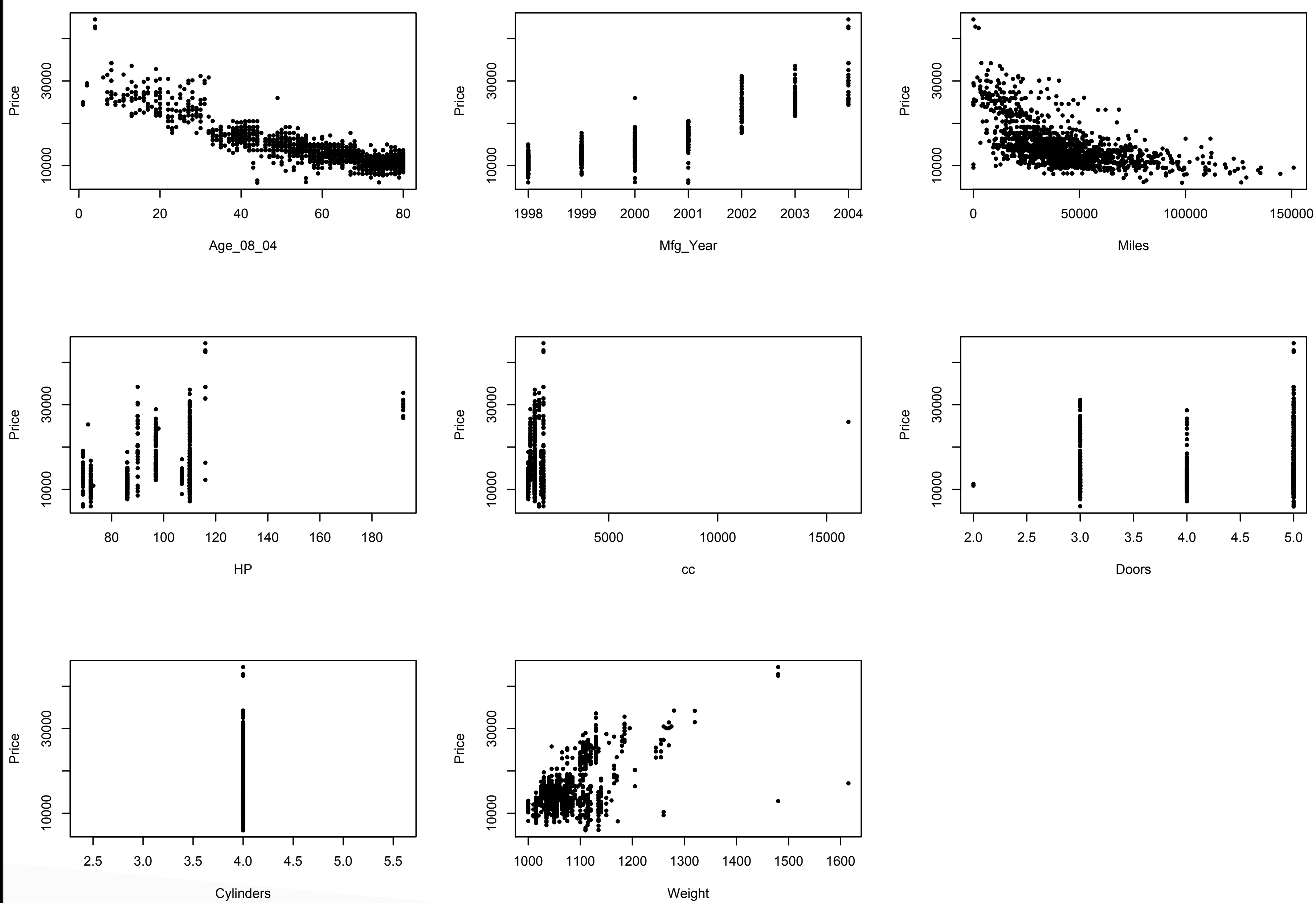
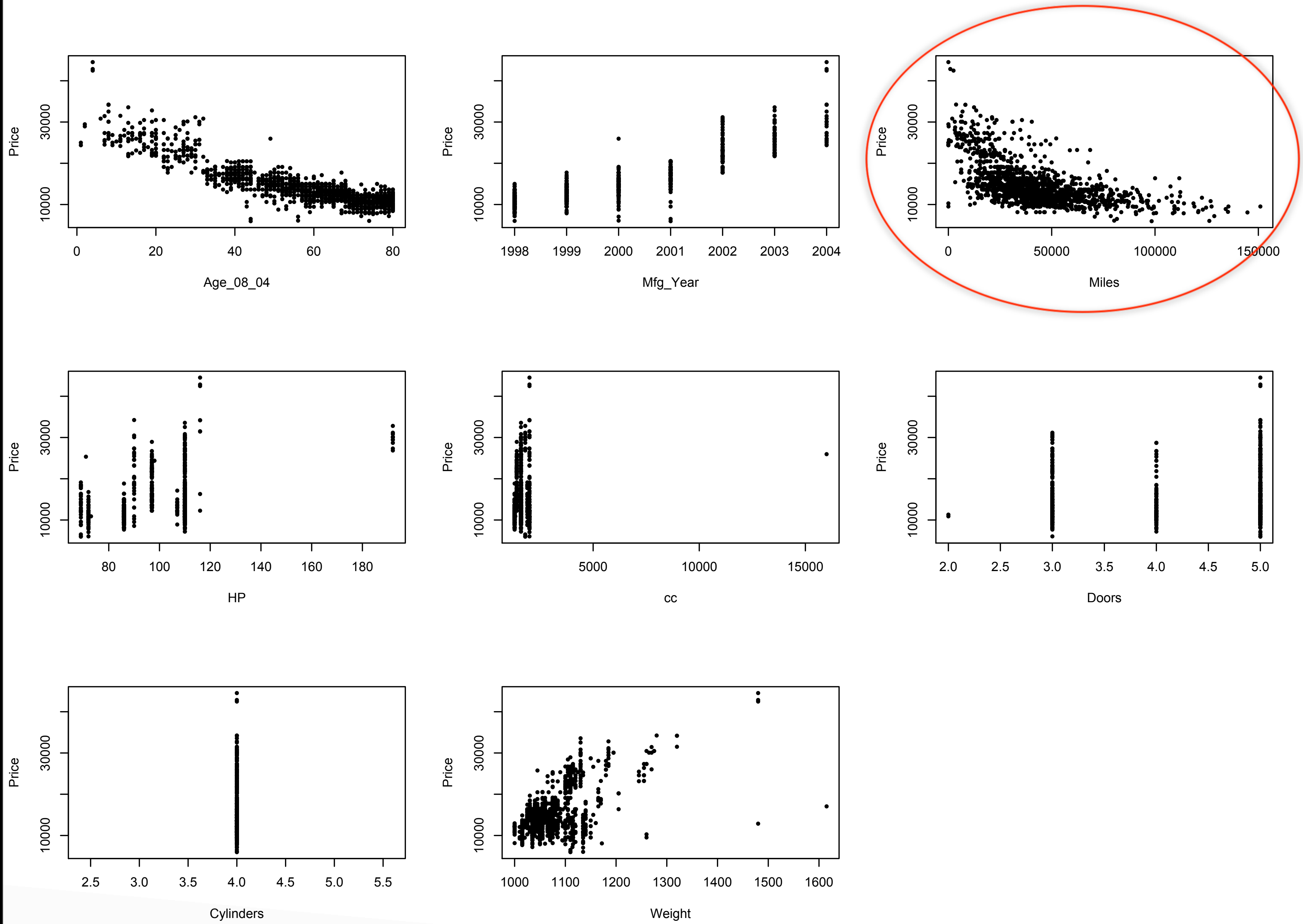
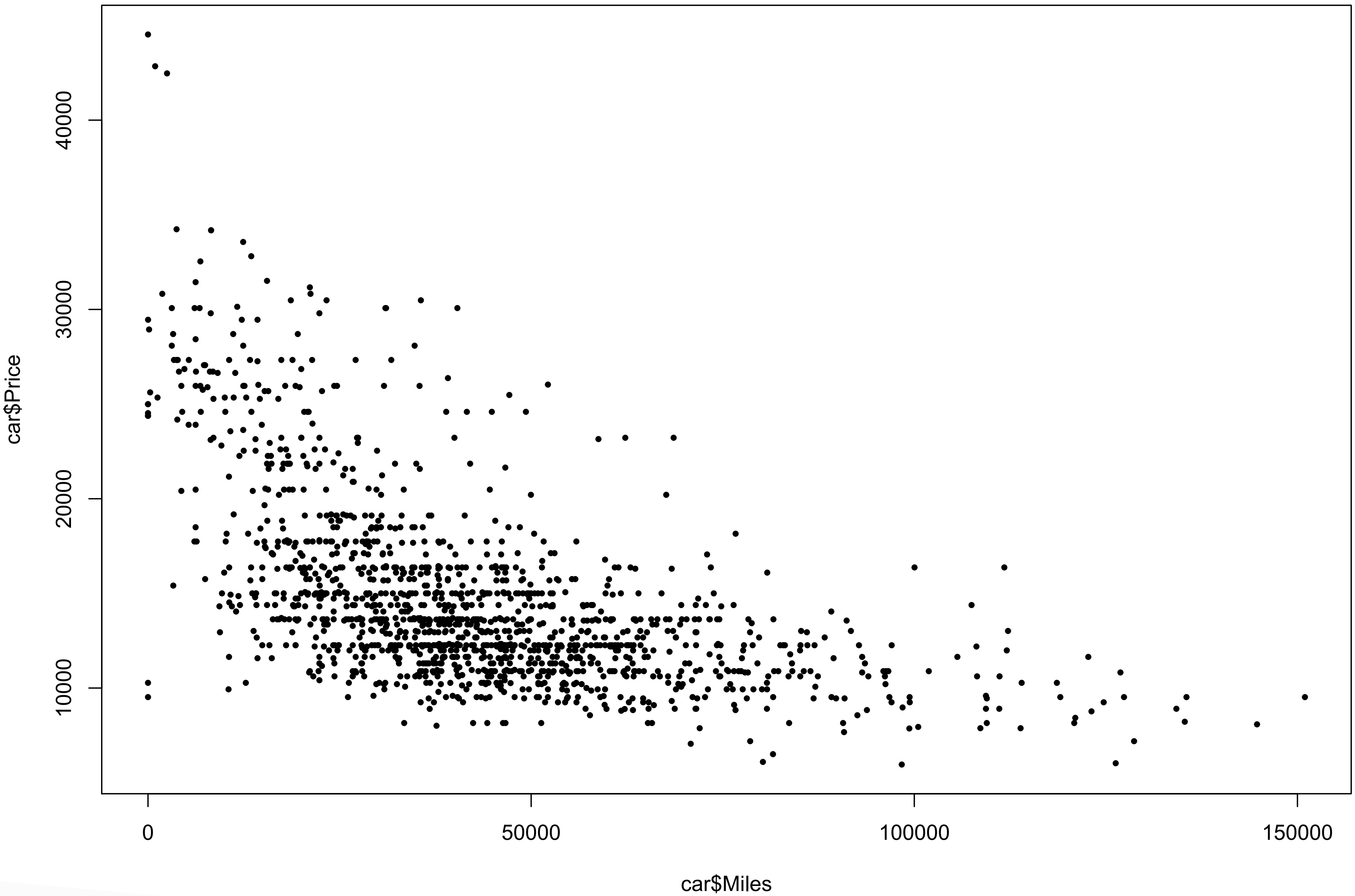
Nonlinear Regression Methods for the Cars Analysis



Price
Age_08_04
Mfg_Year
Miles
HP
cc
Doors
Cylinders
Weight


Price
30000
10000
0 20 40 60 80
Age_08_04
1998 1999 2000 2001 2002 2003 2004
Mfg_Year
0 50000 100000 150000
Miles
80 100 120 140 160 180
HP
5000 10000 15000
cc
2.0 2.5 3.0 3.5 4.0 4.5 5.0
Doors
2.5 3.0 3.5 4.0 4.5 5.0 5.5
Cylinders
1000 1100 1200 1300 1400 1500 1600
Weight



car$Price
40000
30000
20000
10000
0
50000
100000
150000
car$Miles

Cars Data

What is our goal with the Cars data?

1. Predict Sales Price

Why is this goal important?

- The company wants to make sure they turn a profit. If they have an accurate estimate of the sales price, they will offer a lower value for the trade-in and increase the probability of a profit.

Cars Data

How are we going to use statistics to accomplish our goal?

1. Build a regression model for predicting price.

What issues do we face statistically?

1. Non-linearity
2. Outliers
3. Greater than zero?
4. Age and year are redundant.

This unit focuses on how to do non-linear regression. To teach the methods, we'll focus on one covariate first then we'll move onto multiple covariates.

Non-linear Regression (Univariate)

<u>Option 1 (Parametric):</u> Transformations

If

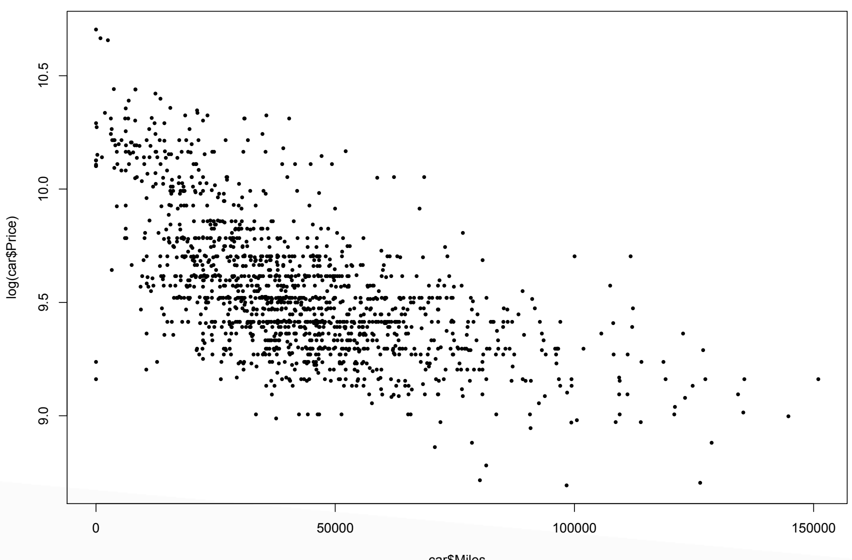
$$\log(y_i) = \beta_0 + \beta_1 x_i + \epsilon_i$$

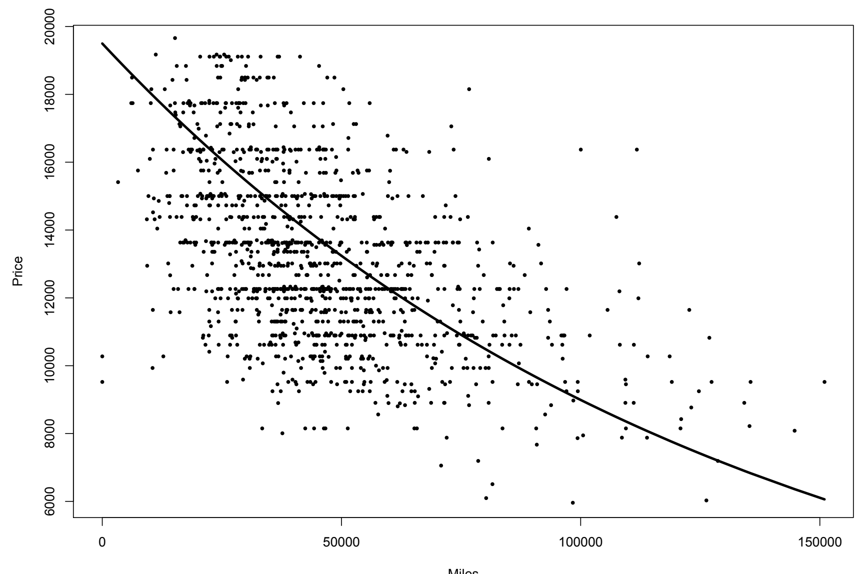
then

$$y_i = \exp(\beta_0)\exp(\beta_1 x_i)\exp(\epsilon_i)$$

which is non-linear in x_i.

Warning: May result in unsatisfactory behavior when untransformed!





Non-linear Regression (Univariate)

Option 2 (Parametric): Polynomial Regression

Let

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_P x_i^P + \epsilon_i$$

This is just a LINEAR model!

Note: when we say "linear" we mean linear in $\boldsymbol{\beta}$.

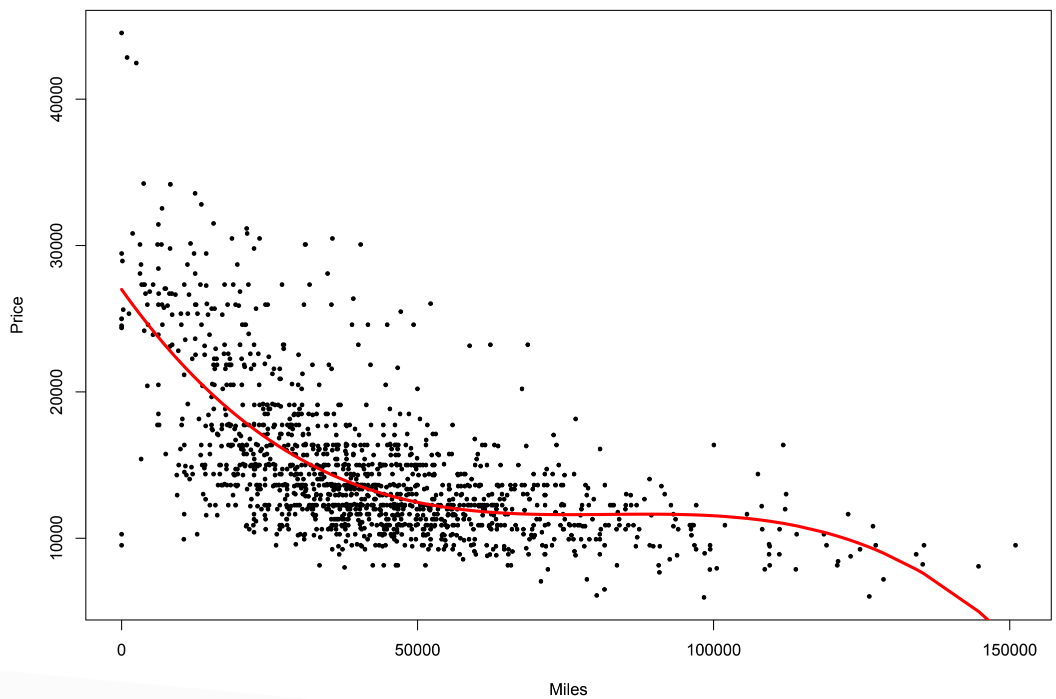
Non-linear Regression (Univariate)

Option 2 (Parametric): Polynomial Regression

Let

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_P x_i^P + \epsilon_i$$

Cubic Fit:



Non-linear Regression (Univariate)

Option 2 (Parametric): Polynomial Regression

Let

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_P x_i^P + \epsilon_i$$

Issues with polynomial regression:

1. What order do we use?
   - Variable selection, LASSO, Cross-validation
2. Collinearity
   - The higher order, the greater collinearity.
3. Scale can kill you!
   - Center and scale the predictor.
4. AWFUL behavior beyond range of data!

Non-linear Regression (Univariate)

Option 3 (Parametric): Basis Function Expansions

Let

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_p b_p(x_i) + \epsilon_i$$

where

$$x_i : \text{predictor for } i^{th} \text{ observation}$$

$$b_p(\cdot) : p^{th} \text{ Basis Function}$$

Non-linear Regression (Univariate)

Option 3 (Parametric): Basis Function Expansions

Examples of Basis Function Expansions:

Linear Regression: $b_p(x_i) = x_i$

Polynomial Regression: $b_p(x_i) = x_i^p$

Step Functions: $b_p(x_i) = \mathbb{I}(\xi_{p-1} < x_i < \xi_p)$

where $\xi_1, \ldots, \xi_P$ are called "knots"

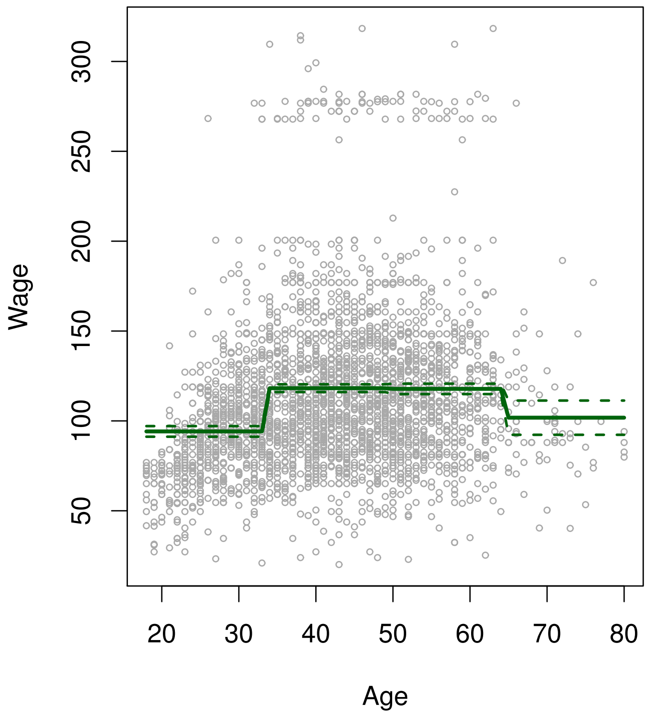
Piecewise Polynomial: $$y_i = \sum_{r=1}^{R} \sum_{p=1}^{P} \beta_{rp} b_{rp}(x_i) + \epsilon_i$$

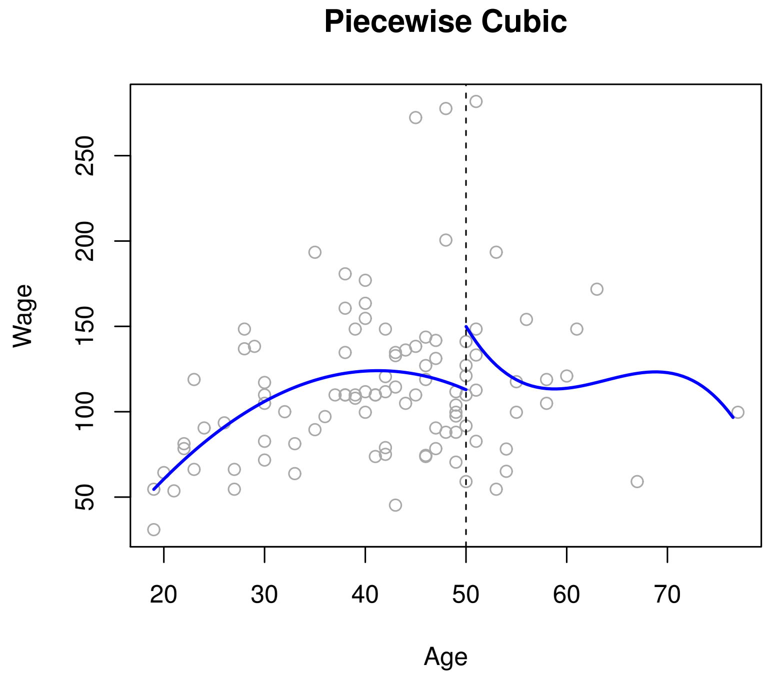
$$b_{rp}(x_i) = x_i^p \mathbb{I}(x_i \in \mathcal{R}_r)$$

where $\{\mathcal{R}_r\}$ partitions the domain of x

Non-linear Regression (Univariate)

Option 3 (Parametric): Basis Function Expansions





Piecewise regression is very flexible but fitted regression curves are not continuous at knot points.

Non-linear Regression (Univariate)

Option 3 (Parametric): Spline Basis Function Expansions

Big Ideas (the "birds eye view"):

1. Piecewise polynomial where we enforce continuity at knot points.
2. Also enforce first and second derivatives to be continuous
   - This serves to smooth out the regression curves.

Non-linear Regression (Univariate)

Option 3 (Parametric): Cubic Spline Expansion with K knots

$$y_i = \beta_0 + \sum_{p=1}^{K+3} b_p(x_i)\beta_p + \epsilon_i$$

$$b_p(x_i) = \begin{cases} x_i^p & \text{for } p \le 3 \\ (x_i - \xi_{p-3})_+^3 & \text{for } p > 3 \end{cases}$$

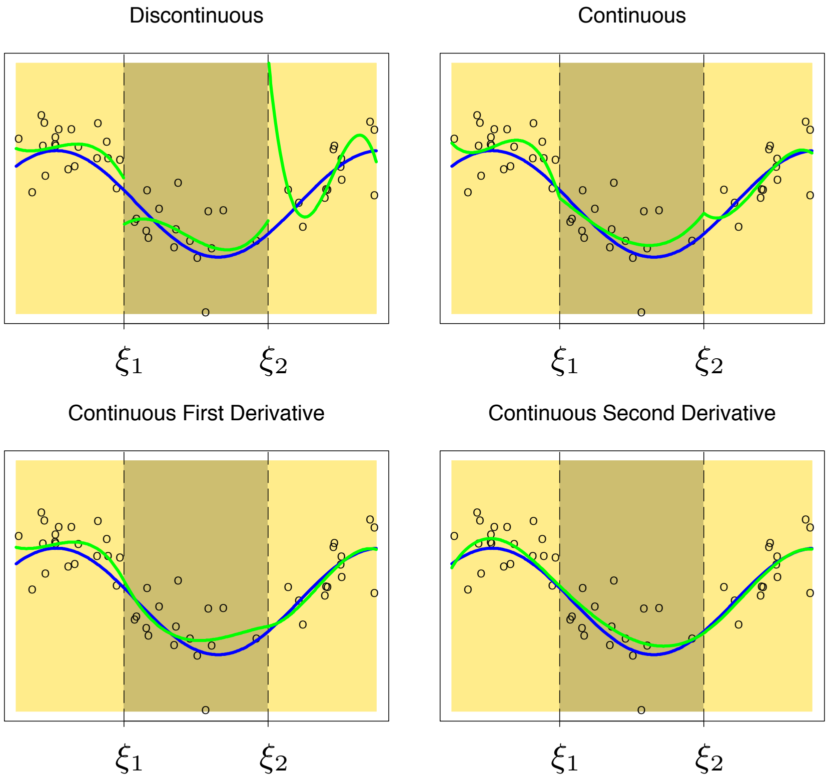
$$\text{where } (x_i - \xi_k)_+^3 = \begin{cases} (x_i - \xi_k)^3 & \text{if } x_i - \xi_k > 0 \\ 0 & \text{otherwise} \end{cases}$$

"Truncated Power Basis"

$$\Rightarrow y_i = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + x_i^3\beta_3 + \sum_{k=1}^{K}(x_i - \xi_k)_+^3\beta_{k+3} + \epsilon_i$$

Non-linear Regression (Univariate)

Option 3 (Parametric): Cubic Spline Expansion with K knots



Non-linear Regression (Univariate)

Option 3 (Parametric): Basis Function Expansions

Generalizing Cubic Splines: M-spline

$$b_p(x_i) = x_i^{p-1} \text{ for } p = 1, \ldots, M$$
$$b_{M+k}(x_i) = (x_i - \xi_k)_+^{M-1} \text{ for } k = 1, \ldots, K$$

Note: B-splines are the general class of basis functions that will generate the basis functions for any order spline.

Non-linear Regression (Univariate)

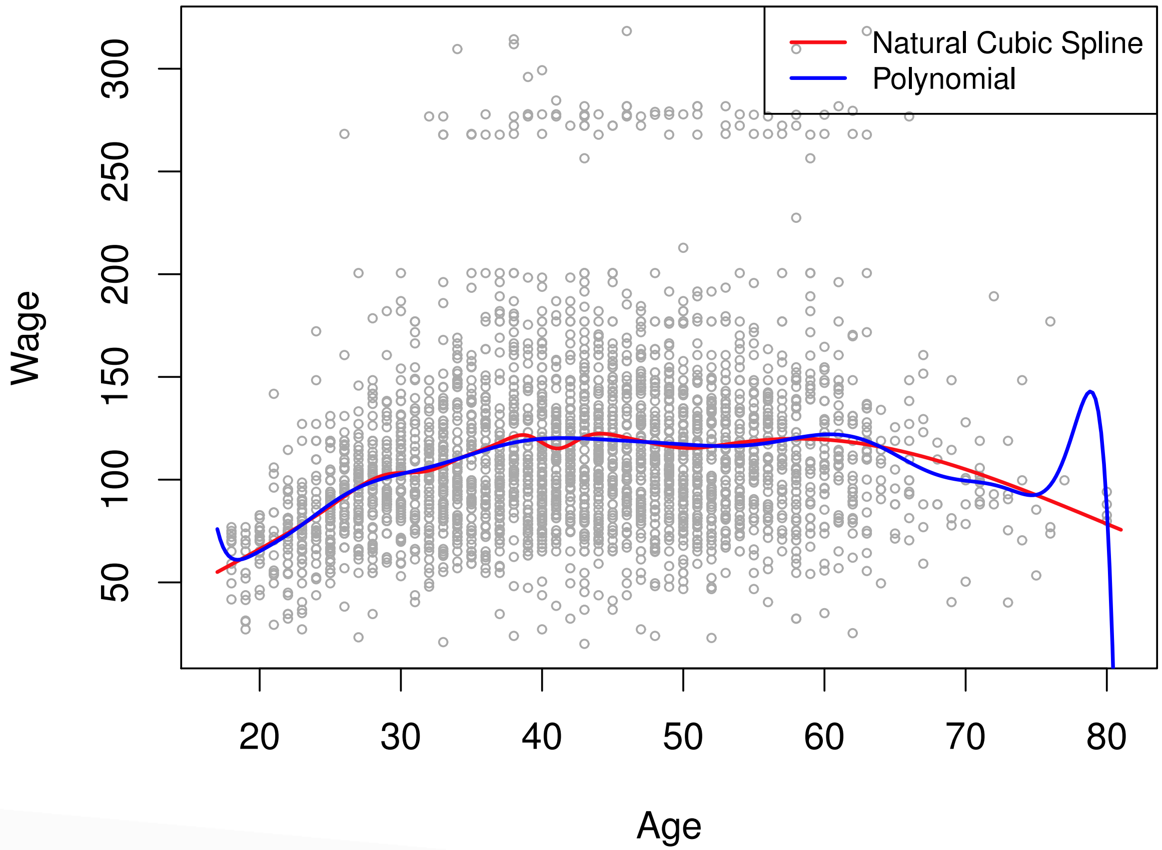
Option 3 (Parametric): Basis Function Expansions

Issues with Cubic Splines

1. AWFUL behavior beyond range of data.
   - Solution: Natural Splines
     - cubic splines but linear outside of data.
     - ugly basis functions so just use ns() in R.

Non-linear Regression (Univariate)

Option 3 (Parametric): Basis Function Expansions

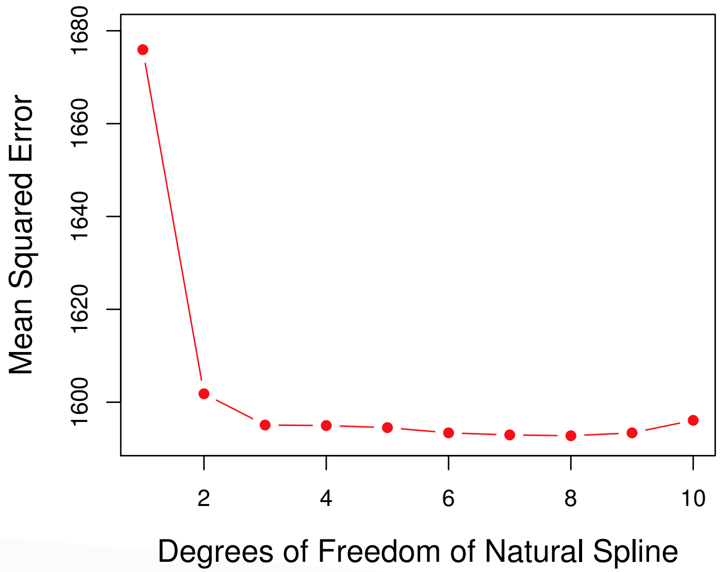


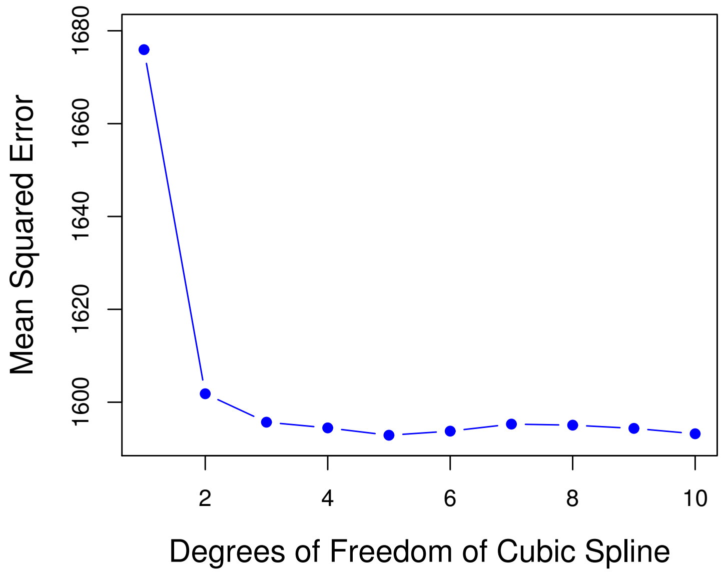
Non-linear Regression (Univariate)

Option 3 (Parametric): Basis Function Expansions

Issues with Cubic Splines

2. Selecting the number of knots
   - Solutions:
     - Cross Validation





Non-linear Regression (Univariate)

Option 3 (Parametric): Basis Function Expansions

Issues with Cubic Splines

2. Selecting the number of knots
   - Solutions:
     - Cross Validation
     - Specify "Degrees of Freedom"
       - ns(): num.knots = df-1
       - bs(): num.knots = df-degree

Non-linear Regression (Univariate)

Option 3 (Parametric): Basis Function Expansions

Issues with Cubic Splines

3. Selecting where to put the knots
   - Solutions:
     - Evenly spaced over the domain
     - Bayesian: estimate them

Non-linear Regression (Univariate)

Option 3 (Parametric): Basis Function Expansions

Issues with Cubic Splines

4. Splines Bases are not orthogonal
   - Solution: Use wavelets instead

Non-linear Regression (Univariate)

Option 3 (Parametric): Basis Function Expansions

Wavelet Basis Expansions:

1. Start with a "father" wavelet:

$$w(x) : x \in [0, 1]$$

2. Translate and dilate the father wavelet:

$$w_{jk}(x) = 2^{j/2} w(2^j x - k)$$

3. Fit the model:

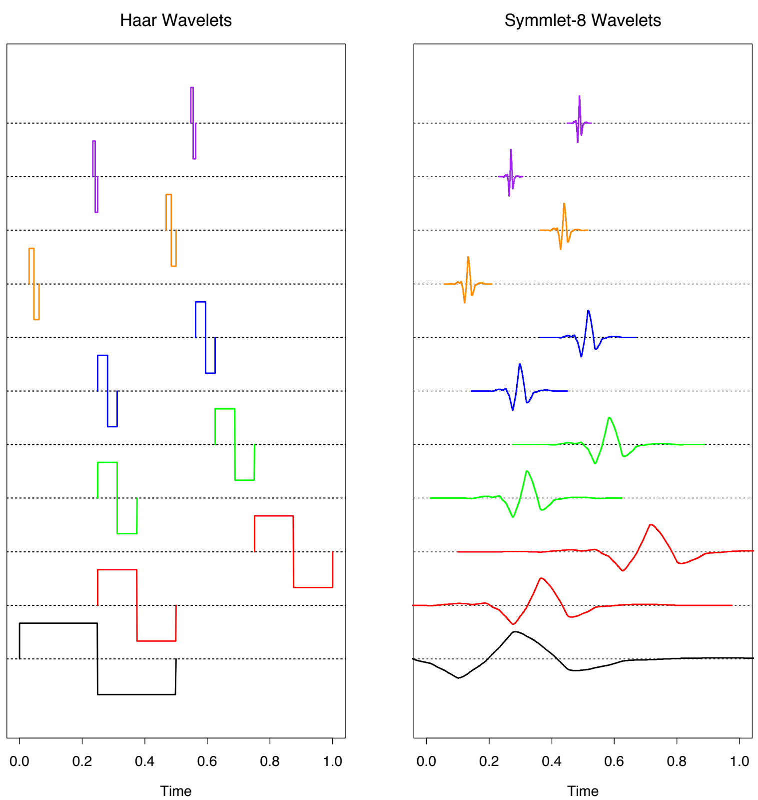
$$y_i = \sum_j \sum_k \beta_{jk} w_{jk}(x) + \epsilon_i$$

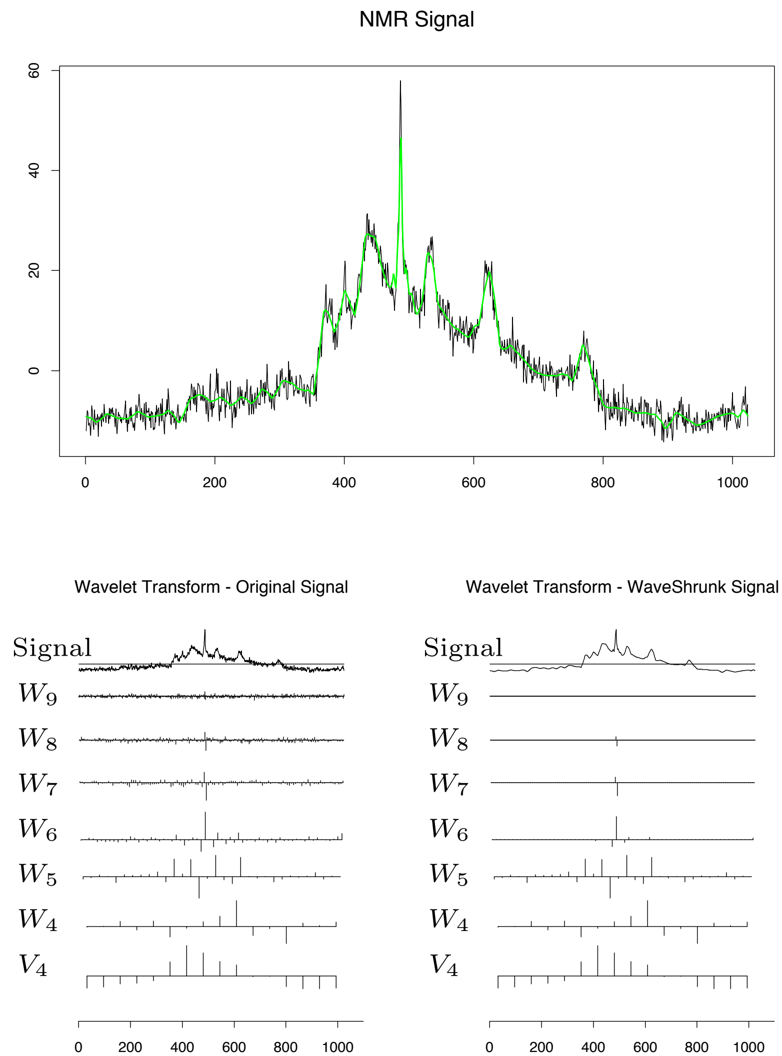
4. LASSO (shrink) the coefficients.

Non-linear Regression (Univariate)

Option 3 (Parametric): Basis Function Expansions

Wavelet Basis Expansions:




NMR Signal
60
40
20
0
0
200
400
600
800
1000
Wavelet Transform - Original Signal
Wavelet Transform - WaveShrunk Signal
Signal
W_9
W_8
W_7
W_6
W_5
W_4
V_4
0
200
400
600
800
1000

Non-linear Regression (Univariate)

Option 4 (Non-parametric): Kernel Smoothing

Big Idea: Just fit a function "locally" to prediction points.

$$\hat{y}(x_0) = \frac{\sum_{i=1}^{N} K_\delta(x_0, x_i) y_i}{\sum_{i=1}^{N} K_\delta(x_0, x_i)} = \sum_{i=1}^{N} w_i(x_0, x_i \mid \delta) y_i$$

$$\delta \equiv \text{ Bandwidth Parameter}$$

Note: This is a "non-generative" (recursive) model and is just used as a predictive tool.

Non-linear Regression (Univariate)

Option 4 (Non-parametric): Kernel Smoothing

Issues:

1. What kernel do you choose?

Window: $K_\delta(x_0, x_i) = \mathbb{I}(\|x_i - x_0\| \leq \delta)$

Gaussian: $K_\delta(x_0, x_i) \propto \exp\{\frac{1}{2\delta^2}(x_i - x_0)^2\}$

Epanechnikov: $K_\delta(x_0, x_i) = \begin{cases} \frac{3}{4}\left(1 - \left(\frac{|x_i - x_0|}{\delta}\right)^2\right) & \text{if } \frac{|x_i - x_0|}{\delta} < 1 \\ 0 & \text{otherwise} \end{cases}$

Bisquare: $K_\delta(x_0, x_i) = \begin{cases} \left(1 - \left(\frac{|x_i - x_0|}{\delta}\right)^2\right)^2 & \text{if } \frac{|x_i - x_0|}{\delta} < 1 \\ 0 & \text{otherwise} \end{cases}$

Non-linear Regression (Univariate)

Option 4 (Non-parametric): Kernel Smoothing

Issues:

1. What kernel do you choose?

$$\text{Tricube: } K_\delta(x_0, x_i) = \begin{cases} \left(1 - \left(\frac{|x_i - x_0|}{\delta}\right)^3\right)^3 & \text{if } \frac{|x_i - x_0|}{\delta} < 1 \\ 0 & \text{otherwise} \end{cases}$$

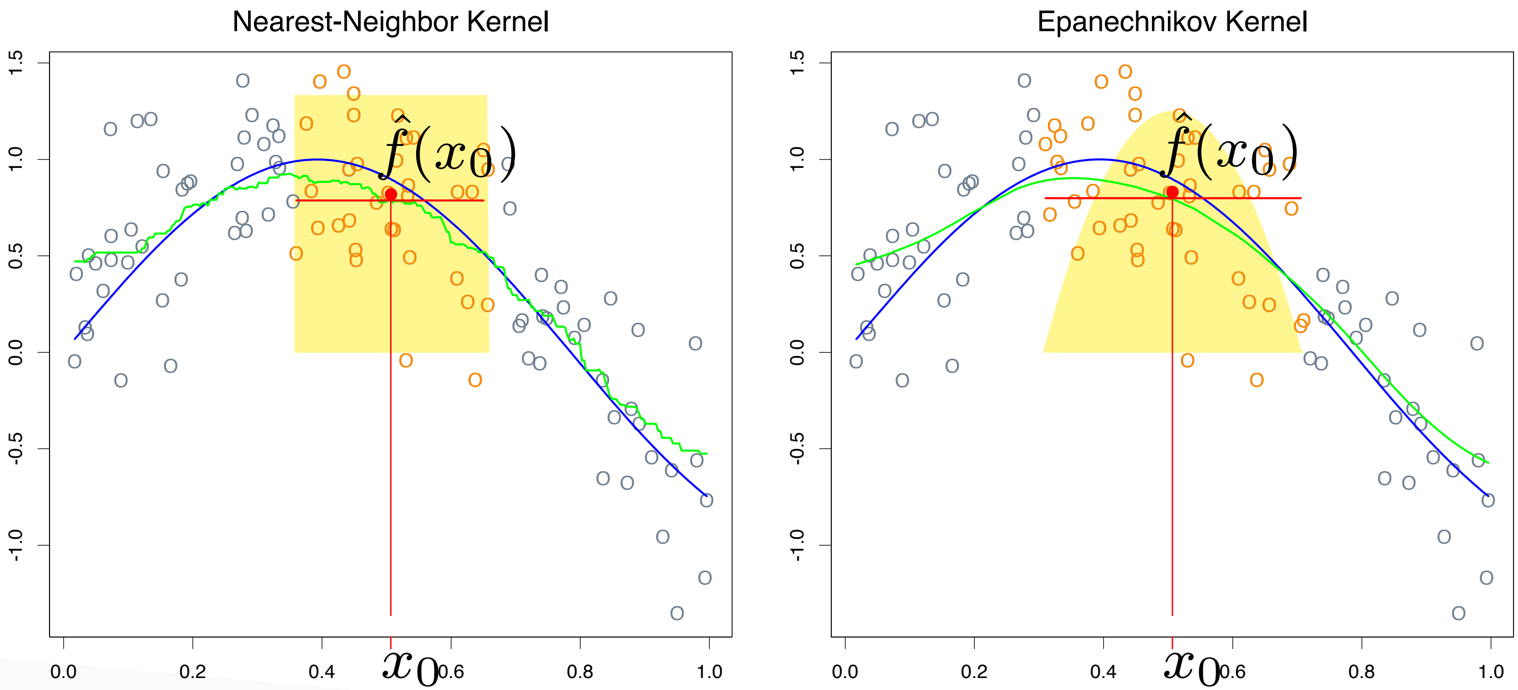
$$\text{Wendland: } K_\delta(x_0, x_i) = \begin{cases} (1-d)^6(35d^2 + 18d + 3)/3 & \text{if } d = \frac{|x_i - x_0|}{\delta} < 1 \\ 0 & \text{otherwise} \end{cases}$$

Non-linear Regression (Univariate)

Option 4 (Non-parametric): Kernel Smoothing

Issues:

1. What kernel do you choose?



Non-linear Regression (Univariate)

Option 4 (Non-parametric): Kernel Smoothing

Issues:

2. How do you choose the bandwidth parameter?
   - Cross-validation

Non-linear Regression (Univariate)

Option 4 (Non-parametric): Kernel Smoothing

Issues:

3. How would you estimate σ^2?

Under the normality assumption:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}$$

But,

$$\hat{y}_i = \frac{\sum_{i=1}^{N} K_\delta(x_0, x_i) y_i}{\sum_{i=1}^{N} K_\delta(x_0, x_i)}$$

so we use the observation to predict the observation!

Non-linear Regression (Univariate)

Option 4 (Non-parametric): Kernel Smoothing

Issues:

3. How would you estimate σ^2?

Alternative estimate:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_{i,-i})^2}{N}$$

$$\hat{y}_{i,-i} \equiv \text{Prediction for } y_i \text{ leaving out } y_i$$

Note: Because this is just a predictive tool, it may just be better to report the hold-out MSE.

Non-linear Regression (Univariate)

Option 4 (Non-parametric): Kernel Smoothing

Issues:

4. How would we get prediction intervals?
   - Bootstrap

$$\hat{y}^b(x_0) \sim \mathcal{N}\left(\sum_{i=1}^{N} w_i^{(b)} y_i^{(b)}, \hat{\sigma}^2_{(b)}\right)$$

$$(\hat{y}^{(0.025)}(x_0), \hat{y}^{(0.975)}(x_0))$$

Non-linear Regression (Univariate)

Option 4 (Non-parametric): Kernel Smoothing

Issues:

5. What do you do at the boundaries?
   - Variance of predictions will be bigger and probably biased.
   - Solution: Use Local Regression

Non-linear Regression (Univariate)

Option 5 (Non-parametric): Local Regression

Find $\hat{\beta}_0(x_0), \hat{\beta}_1(x_0)$ that minimize

$$\min_{\hat{\beta}_0(x_0),\hat{\beta}_1(x_0)} \sum_{i=1}^{n} K_\delta(x_0, x_i)(y_i - \hat{\beta}_0(x_0) - \hat{\beta}_1(x_0)x_i)^2$$

then $\hat{y}(x_0) = \hat{\beta}_0(x_0) + \hat{\beta}_1(x_0)x_0$.

Notes:

1. This is just a weighted least squares problem (which we already know how to solve).
2. LOESS smoothing – local regression using a quadratic model rather linear model.

$$\min_{\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2} \sum_{i=1}^{n} K_\delta(x_0, x_i)(y_i - \hat{\beta}_0(x_0) - \hat{\beta}_1(x_0)x_i - \hat{\beta}_2(x_0)x_i^2)^2$$

Non-linear Regression (Univariate)

Option 6 (Non-parametric): Smoothing Splines

We want:

1. a function $g(x_i)$ so that $\sum_{i=1}^{n}(y_i - g(x_i))^2$ is small.
2. $g(x_i)$ to be smooth (not overfit the data).

A smoothing spline finds $g(x_i)$ that solves

$$\underbrace{\sum_{i=1}^{n}(y_i - g(x_i))^2}_{\text{Loss}} + \underbrace{\lambda \int g''(x)^2 dx}_{\text{Smoothness Penalty}}$$

Non-linear Regression (Univariate)

Option 6 (Non-parametric): Smoothing Splines

A few notes on smoothing splines:

1. Second derivative is a smoothness penalty (how fast is my derivative changing).
2. Solution is a that $g(x_i)$ is a natural cubic splines with knots at each of $x_1, \ldots, x_n$ then coefficients shrunk back towards zero.

Non-linear Regression (Univariate)

Option 6 (Non-parametric): Smoothing Splines

Issues to consider with smoothing splines:

1. No knot selection but how do we choose λ?
   - Cross-validation – leave one out cross validation is computationally cheap.

Non-linear Regression (Multiple Variables)

How do we extend these methods from "simple" to "multiple"?

Generalized Additive Models (GAMs):

$$y_i = \beta_0 + \sum_{p=1}^{P} f_p(x_{ip}) + \epsilon_i$$

$f_p(x_{ip})$: Function for p^{th} variable

Non-linear Regression (Multiple Variables)

How would we fit a GAM with a smoothing spline?

Fitting GAMs using backfitting:

- Fit each function separately after adjusting for the others. That is, fit

$$y_i - \beta_0 - \sum_{j \neq p} f_j(x_{ij}) = f_p(x_{ip}) + \epsilon_i$$

for $p = 1, \ldots, P$ until it converges.

Non-linear Regression (Multiple Variables)

How would we fit a GAM with a smoothing spline?

Fitting GAMs in R:

- The book tells you to use the "gam" library in R but I think the "mgcv" is better.
- The "mgcv" package follows same format as "gam" in book but has options to calculate estimates of uncertainty.
- Example Code for Predictive Interval:

```
gam.obj <- gam(y~x)
pred <- predict.gam(gam.obj,newdata,se.fit=TRUE)
pred.se <- sqrt(pred$se.fit^2+gam.obj$sig2)
low <- pred$fit – qt(0.975,df=gam.obj$df.residual)*pred.se
up <- pred$fit + qt(0.975,df=gam.obj$df.residual)*pred.se
```

Non-linear Regression (Multiple Variables)

How would you do model selection for GAMs?

Model Selection for GAMs

$$y_i = \beta_0 + \sum_{p=1}^{P} f_p(x_{ip}) + \epsilon_i$$

We need to treat each function separately. So either include a function or don't. "Function selection"

To do this, we also need to specify what each function is. Justify based on univariate fits.

No canned R function to do this (that I know of).

Expectations for Cars Analysis

1. Description of Cars data and the problems associated with it (particularly you need to mention non-linearity).

2. Write out a GAM where you state what each function is. Be sure to give intuition for terms that may be unfamiliar to those who will be in your seat next year (e.g. briefly describe what a spline is).
   - Justify your choice of function. For example, justify a spline by showing me a picture and using CV to choose how many knots to use (univariate justification is fine). Or, justify a smoothing spline by stating its advantages.
   - I am not necessarily going to look for model selection but it would be great if you did something (hard to code).
   - Strongly encourage use of spline, smoothing spline or local regression because these have canned R functions.

3. Fit your GAM model and provide graphs of how price relates to each variable (you may be able to report betas). Show me a residual plot to justify assumptions.

4. Give some measure of how accurate predictions are under your model (e.g. coverage and predictive interval width).

My Cars Analysis

1. Description of Cars data and the problems associated with it.

2. I would use a natural spline for miles, linear model for Weight and the other variables I would treat as categorical.
   - I'd also throw out cylinders, fix cc, and use year instead of age (people usually ask the year not necessarily the age in months of a car).
   - NOTE: Some of the categories have very few observations so you may need to combine categories (e.g. doors, color, and cc).

3. Justify the number of knots for my natural spline based on cross-validation. Evenly space the knots because I can't tell where to put them based on the scatterplot.

4. Do a forward "function selection" based on cross-validated mean square error.

5. For my final model, report coverage and predictive interval width of hold-out sample.