# Regression

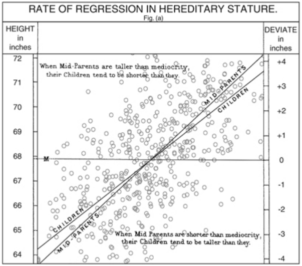
Schwartz

September 5, 2016

# The Sophomore Slump

...or sophomore jinx or sophomore jitters refers to an instance in which a second, or sophomore, effort fails to live up to the standards of the first effort. It is commonly used to refer to the apathy of students (second year of high school, college or university), the performance of athletes (second season of play), singers/bands (second album), television shows (second seasons) and films (sequels/prequels). In the United Kingdom, the "sophomore slump" is more commonly referred to as "second year blues", particularly when describing university students. In Australia, it is known as "second year syndrome", and is particularly common when referring to professional athletes who have a mediocre second season following a stellar debut. The phenomenon of a "sophomore slump" can be explained psychologically, where earlier success has a reducing effect on the subsequent effort, but it can also be explained statistically, as an effect of the regression towards the mean.

The concept of "regression" comes from genetics and was popularized by Sir Francis Galton's late 19th century publication of "Regression towards mediocrity in hereditary stature." Galton observed that extreme characteristics (e.g., height) in parents are not completely passed on to offspring, but rather the characteristics in the offspring "regress" towards a mediocre point. By measuring the heights of hundreds of people Galton was able to quantify this "regression" and in so doing invented linear regression analysis, thus laying the groundwork for much of modern statistical modeling. The term "regression" stuck.



# Objectives

- Terminology
- Least squares fit
- Normal distribution theory
- Coefficient testing
- Linear models and Multiple variables and Alternatives
- Model fit and Model selection
- Model diagnostics and Model evaluation

# Terminology

- $\mathrm{E}[Y_i] = \beta_0 + x_i\beta_1 + \epsilon_i, \epsilon_i \overset{i.i.d.}{\sim} Normal(0, \sigma^2)$

# Terminology

Coefficient

- $\mathrm{E}[Y_i] = \beta_0 + x_i\beta_1 + \epsilon_i, \epsilon_i \overset{i.i.d.}{\sim} \mathit{Normal}(0, \sigma^2)$

Intercept Error Noise

# Terminology

Coefficient

- $\mathrm{E}[Y_i] = \beta_0 + x_i\beta_1 + \epsilon_i, \epsilon_i \overset{i.i.d.}{\sim} Normal(0, \sigma^2)$

Intercept Error Noise

- $Y_i = \hat{\beta}_0 + x_i\hat{\beta}_1 + \hat{\epsilon}_i, \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n-p-1} \quad (p = \#of\ coefficients)$

# Terminology

Coefficient

- $E[Y_i] = \beta_0 + x_i\beta_1 + \epsilon_i, \epsilon_i \overset{i.i.d.}{\sim} Normal(0, \sigma^2)$

Intercept Error Noise

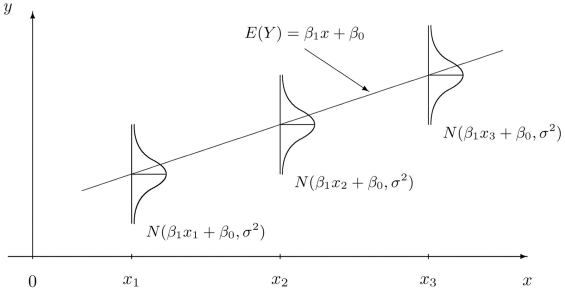
Fitted value Residual

- $Y_i = \hat{\beta}_0 + x_i\hat{\beta}_1 + \hat{\epsilon}_i, \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\hat{\epsilon}_i^2}{n-p-1} \quad (p = \#of\ coefficients)$

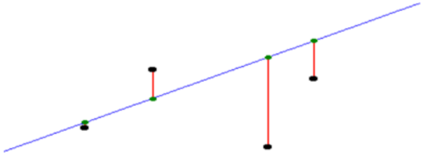
Residual Standard Error (RSE)

# Terminology

- $E[Y_i] = \beta_0 + x_i\beta_1 + \epsilon_i, \epsilon_i \overset{i.i.d.}{\sim} Normal(0, \sigma^2)$




- $Y_i = \hat{\beta}_0 + x_i\hat{\beta}_1 + \hat{\epsilon}_i, \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n-p-1} \quad (p = \#of\ coefficients)$



# Least Squares Fit

- $Y_i = \hat{\beta}_0 + x_i\hat{\beta}_1 + \hat{\epsilon}_i$

# Least Squares Fit

- $Y_i = \hat{\beta}_0 + x_i\hat{\beta}_1 + \hat{\epsilon}_i$

$$[\hat{\beta}_0, \hat{\beta}_1] = \underset{[\beta_0,\beta_1]}{argmin} \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \underset{\boldsymbol{\beta}}{argmin} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2$$

where $\mathbf{x}_i^T = [1, x_i]$ and $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$

# Least Squares Fit

- $Y_i = \hat{\beta}_0 + x_i\hat{\beta}_1 + \hat{\epsilon}_i$

$$[\hat{\beta}_0, \hat{\beta}_1] = \underset{[\beta_0,\beta_1]}{argmin} \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \underset{\beta}{argmin} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i^T \beta)^2$$

where $\mathbf{x}_i^T = [1, x_i]$ and $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$

$$\sum_{i=1}^{n} (Y_i - \mathbf{x}_i^T \beta)^2 = (\mathbf{Y} - \mathbf{x}_i\beta)^T (\mathbf{Y} - \mathbf{x}\beta)$$

where $\mathbf{Y}^T = [Y_1, Y_2, \cdots, Y_n]$ and $\mathbf{x}^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}$

# Least Squares Fit

- $Y_i = \hat{\beta}_0 + x_i\hat{\beta}_1 + \hat{\epsilon}_i$

$$[\hat{\beta}_0, \hat{\beta}_1] = \underset{[\beta_0,\beta_1]}{argmin} \textstyle\sum_{i=1}^{n} \hat{\epsilon}_i^2 = \underset{\beta}{argmin} \textstyle\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\beta)^2$$

where $\mathbf{x}_i^T = [1, x_i]$ and $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$

$$\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\beta)^2 = (\mathbf{Y} - \mathbf{x}_i\beta)^T(\mathbf{Y} - \mathbf{x}\beta)$$

where $\mathbf{Y}^T = [Y_1, Y_2, \cdots, Y_n]$ and $\mathbf{x}^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}$

$$\nabla_\beta \beta^T(\mathbf{x}^T\mathbf{x})\beta - 2\mathbf{Y}^T\mathbf{x}\beta + \mathbf{Y}^T\mathbf{Y}$$

$$= 2(\mathbf{x}^T\mathbf{x})\beta - 2\mathbf{x}^T\mathbf{Y} \qquad \text{(set to } \mathbf{0} \text{ to minimize)}$$

$$\Longrightarrow \hat{\beta} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y} \Longrightarrow \text{fitted values } \hat{\mathbf{Y}} = \mathbf{x}\hat{\beta}$$

# Least Squares Fit *bonus*

1. In simple linear regression the $\underset{\beta}{argmin}(\mathbf{Y} - \mathbf{x}\beta)^T(\mathbf{Y} - \mathbf{x}\beta)$ is

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{R_{xY} S_Y}{S_x}$$

# Least Squares Fit *bonus*

1. In simple linear regression the $\underset{\beta}{argmin}(\mathbf{Y} - \mathbf{x}\beta)^T(\mathbf{Y} - \mathbf{x}\beta)$ is

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{R_{xY} S_Y}{S_x}$$

2. Maximum likelihood estimation (MLE) is equivalent

$$\underset{\beta}{argmax} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(Y_i - \mathbf{x}_i^T \beta)^2}$$

$$= \underset{\beta}{argmax} (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{x}\beta)^T(\mathbf{Y}-\mathbf{x}\beta)}$$

$$= \underset{\beta}{argmax} - \frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{x}\beta)^T(\mathbf{Y} - \mathbf{x}\beta)$$

$$= \underset{\beta}{argmin}(\mathbf{Y} - \mathbf{x}_i\beta)^T(\mathbf{Y} - \mathbf{x}\beta)$$

# The *Multivariate Normal Distribution* (MVN)

$$f(\mathbf{Y}|\mathbf{x}, \beta, \sigma^2) = \prod_{i=1}^{n} f(Y_i|\mathbf{x}_i, \beta, \sigma^2)$$

# The *Multivariate Normal Distribution* (MVN)

$$
\begin{aligned}
f(\mathbf{Y}|\mathbf{x}, \beta, \sigma^2) &= \prod_{i=1}^{n} f(Y_i|\mathbf{x}_i, \beta, \sigma^2) \\
&= \prod_{i=1}^{n} N(\mathbf{x}_i^T \beta, \sigma^2)
\end{aligned}
$$

# The *Multivariate Normal Distribution* (MVN)

$$
\begin{aligned}
f(\mathbf{Y}|\mathbf{x}, \beta, \sigma^2) &= \prod_{i=1}^{n} f(Y_i|\mathbf{x}_i, \beta, \sigma^2) \\
&= \prod_{i=1}^{n} N(\mathbf{x}_i^T \beta, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(Y_i - \mathbf{x}_i^T \beta)^2}
\end{aligned}
$$

# The *Multivariate Normal Distribution* (MVN)

$$
\begin{aligned}
f(\mathbf{Y}|\mathbf{x},\beta,\sigma^2) &= \prod_{i=1}^{n} f(Y_i|\mathbf{x}_i,\beta,\sigma^2) \\
&= \prod_{i=1}^{n} N(\mathbf{x}_i^T\beta,\sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(Y_i-\mathbf{x}_i^T\beta)^2} \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{x}\beta)^T(\mathbf{Y}-\mathbf{x}\beta)} \quad = MVN(\mathbf{x}\beta,\sigma^2 I)
\end{aligned}
$$

# The *Multivariate Normal Distribution* (MVN)

$$
\begin{aligned}
f(\mathbf{Y}|\mathbf{x},\beta,\sigma^2) &= \prod_{i=1}^{n} f(Y_i|\mathbf{x}_i,\beta,\sigma^2) \\
&= \prod_{i=1}^{n} N(\mathbf{x}_i^T\beta,\sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(Y_i-\mathbf{x}_i^T\beta)^2} \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{x}\beta)^T(\mathbf{Y}-\mathbf{x}\beta)} \quad = MVN(\mathbf{x}\beta,\sigma^2 I) \\
\\
&\propto (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(\beta-(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y})^T\mathbf{x}^T\mathbf{x}(\beta-(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y})}
\end{aligned}
$$

# The *Multivariate Normal Distribution* (MVN)

$$
\begin{aligned}
f(\mathbf{Y}|\mathbf{x},\beta,\sigma^2) &= \prod_{i=1}^{n} f(Y_i|\mathbf{x}_i,\beta,\sigma^2) \\
&= \prod_{i=1}^{n} N(\mathbf{x}_i^T\beta,\sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(Y_i-\mathbf{x}_i^T\beta)^2} \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{x}\beta)^T(\mathbf{Y}-\mathbf{x}\beta)} \quad = MVN(\mathbf{x}\beta,\sigma^2 I) \\
\\
&\propto (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(\beta-(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y})^T\mathbf{x}^T\mathbf{x}(\beta-(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y})} \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(\beta-\hat{\beta})^T\mathbf{x}^T\mathbf{x}(\beta-\hat{\beta})}
\end{aligned}
$$

# The *Multivariate Normal Distribution* (MVN)

$$
\begin{aligned}
f(\mathbf{Y}|\mathbf{x},\beta,\sigma^2) &= \prod_{i=1}^{n} f(Y_i|\mathbf{x}_i,\beta,\sigma^2) \\
&= \prod_{i=1}^{n} N(\mathbf{x}_i^T\beta,\sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(Y_i-\mathbf{x}_i^T\beta)^2} \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{x}\beta)^T(\mathbf{Y}-\mathbf{x}\beta)} \quad = MVN(\mathbf{x}\beta,\sigma^2 I) \\
\\
&\propto (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(\beta-(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y})^T\mathbf{x}^T\mathbf{x}(\beta-(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y})} \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(\beta-\hat{\beta})^T\mathbf{x}^T\mathbf{x}(\beta-\hat{\beta})} \\
\implies f(\hat{\beta}|\mathbf{x},\beta,\sigma^2) &= MVN\left(\beta,\sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\right)
\end{aligned}
$$

# Testing

$$f(\mathbf{Y}|\mathbf{x}, \beta, \sigma^2) = MVN(\mathbf{x}\beta, \sigma^2 I)$$

$$\implies f(\hat{\beta}|\mathbf{x}, \beta, \sigma^2) = MVN\left(\beta, \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\right)$$

# Testing

$$f(\mathbf{Y}|\mathbf{x}, \beta, \sigma^2) = MVN(\mathbf{x}\beta, \sigma^2 I)$$
$$\implies f(\hat{\beta}|\mathbf{x}, \beta, \sigma^2) = MVN\left(\beta, \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\right)$$

For simple linear regression then

$$\hat{\beta} \sim MVN\left(\left[\begin{array}{c} \beta_0 \\ \beta_1 \end{array}\right], \frac{\sigma^2}{n\sum(x_i - \bar{x})^2}\left[\begin{array}{cc} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{array}\right]\right)$$

# Testing

$$f(\mathbf{Y}|\mathbf{x},\beta,\sigma^2) = MVN(\mathbf{x}\beta,\sigma^2 I)$$
$$\implies f(\hat{\beta}|\mathbf{x},\beta,\sigma^2) = MVN\left(\beta,\sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\right)$$

For simple linear regression then

$$\hat{\beta} \sim MVN\left(\left[\begin{array}{c}\beta_0\\ \beta_1\end{array}\right], \frac{\sigma^2}{n\sum(x_i-\bar{x})^2}\left[\begin{array}{cc}\sum x_i^2 & -\sum x_i\\ -\sum x_i & n\end{array}\right]\right)$$

where

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i-\bar{Y})(x_i-\bar{x})}{\sum_{i=1}^n (x_i-\bar{x})^2} = \frac{R_{xY}S_Y}{S_x}$$

$$\text{Var}(\hat{\beta}_1) = \sigma^2\frac{1}{\sum_{i=1}^n (x_i-\bar{x})^2} \quad \text{Var}(\hat{\beta}_0) = \sigma^2\left[\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]$$

# Testing

$$f(\mathbf{Y}|\mathbf{x},\boldsymbol{\beta},\sigma^2) = MVN(\mathbf{x}\boldsymbol{\beta},\sigma^2 I)$$
$$\implies f(\hat{\boldsymbol{\beta}}|\mathbf{x},\boldsymbol{\beta},\sigma^2) = MVN\left(\boldsymbol{\beta},\sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\right)$$

For simple linear regression then

$$\hat{\boldsymbol{\beta}} \sim MVN\left(\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix}, \frac{\sigma^2}{n\sum(x_i-\bar{x})^2}\begin{bmatrix}\sum x_i^2 & -\sum x_i\\ -\sum x_i & n\end{bmatrix}\right)$$

where

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i-\bar{Y})(x_i-\bar{x})}{\sum_{i=1}^n (x_i-\bar{x})^2} = \frac{R_{xY}S_Y}{S_x}$$

$$\text{Var}(\hat{\beta}_1) = \sigma^2\frac{1}{\sum_{i=1}^n (x_i-\bar{x})^2} \quad \text{Var}(\hat{\beta}_0) = \sigma^2\left[\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]$$

$$\text{SE}(\hat{\beta}_0) = \sqrt{\text{Var}(\hat{\beta}_0)} \quad \text{SE}(\hat{\beta}_1) = \sqrt{\text{Var}(\hat{\beta}_1)}$$

# Testing *Extra Credit I*

- What is $\mathrm{Var}(\hat{Y}_0)$? (Suppose we know $\sigma^2$)

- *Hint:* $\hat{Y}_0 = \hat{\beta}_0 + x_0\hat{\beta}_1$
- *Hint:* $Var[aX + bY] = ?$

# Testing *Extra Credit II*

- $\mathrm{Var}(Y_0)$? For a *new observation* $Y_0$ according to our model?
  (Suppose we know $\sigma^2$)

- *Hint:* $Y_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \epsilon$

# A SERIOUSLY MAJOR TRANSITION JUST HAPPENED

And you didn't even notice

# Coefficient Testing

For $\hat{\beta} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y}$, since (under $H_0$)

$$f(\hat{\beta}|\beta, \sigma^2) = MVN\left(\beta, \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\right)$$

we have that

# Coefficient Testing

For $\hat{\boldsymbol{\beta}} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y}$, since (under $H_0$)

$$f(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}, \sigma^2) = MVN\left(\boldsymbol{\beta}, \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\right)$$

we have that

$$\frac{\hat{\beta}_i - \beta_i}{\mathrm{SE}(\hat{\beta}_i)} \sim N(0, 1)$$

# Coefficient Testing

For $\hat{\beta} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y}$, since (under $H_0$)

$$f(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}, \sigma^2) = MVN\left(\boldsymbol{\beta}, \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\right)$$

we have that

$$\frac{\hat{\beta}_i - \beta_i}{\mathrm{SE}(\hat{\beta}_i)} \sim N(0,1)$$

and if we estimate

$$\hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\beta}})}{n-p-1}$$

(where $p$ is the number of coefficients) then we have that

# Coefficient Testing

For $\hat{\beta} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y}$, since (under $H_0$)

$$f(\hat{\beta}|\beta, \sigma^2) = MVN\left(\beta, \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\right)$$

we have that

$$\frac{\hat{\beta}_i - \beta_i}{\mathrm{SE}(\hat{\beta}_i)} \sim N(0,1)$$

and if we estimate

$$\hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{x}\hat{\beta})^T(\mathbf{Y} - \mathbf{x}\hat{\beta})}{n-p-1}$$

(where $p$ is the number of coefficients) then we have that

$$\frac{\hat{\beta}_i - \beta_i}{\mathrm{SE}(\hat{\beta}_i)} \sim t_{n-p-1}$$

## Coefficient Testing

For $\hat{\beta} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y}$, since (under $H_0$)

$$f(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}, \sigma^2) = MVN\left(\boldsymbol{\beta}, \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\right)$$

we have that

$$\frac{\hat{\beta}_i - \beta_i}{\mathrm{SE}(\hat{\beta}_i)} \sim N(0,1)$$

and if we estimate

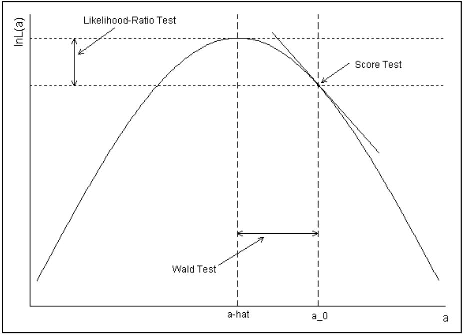
$$\hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\beta}})}{n-p-1}$$

(where $p$ is the number of coefficients) then we have that

$$\frac{\hat{\beta}_i - \beta_i}{\mathrm{SE}(\hat{\beta}_i)} \sim t_{n-p-1}$$

And this works for any number of covariates...

# More Coefficient Testing




Wald test

$$\frac{(\hat{\theta}-\theta_0)^2}{Var(\hat{\theta})} \overset{approx.}{\sim} N(0,1) \text{ under } H_0$$

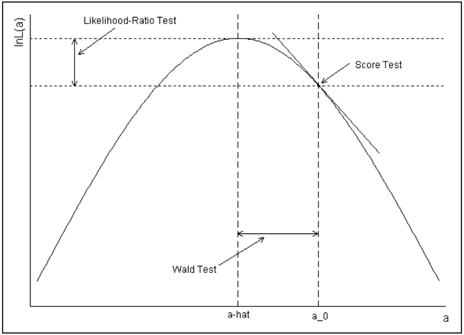
Likelihood-Ratio (LR) test

$$-2ln\left(\frac{L(\theta_0|x)}{L(\theta|x)}\right) \overset{approx.}{\sim} \chi^2_k \text{ under } H_0$$

Score test

$$\frac{\left(\frac{\partial}{\partial\theta}\log L(\theta_0|x)\right)^2}{-\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log L(\theta_0|x)\right]} \overset{approx.}{\sim} \chi^2_1 \text{ under } H_0$$

# More Coefficient Testing




Wald test

$$\frac{(\hat{\theta}-\theta_0)^2}{Var(\hat{\theta})} \overset{approx.}{\sim} N(0,1) \text{ under } H_0$$

Likelihood-Ratio (LR) test

$$-2ln\left(\frac{L(\theta_0|x)}{L(\theta|x)}\right) \overset{approx.}{\sim} \chi^2_k \text{ under } H_0$$

Score test

$$\frac{\left(\frac{\partial}{\partial\theta}\log L(\theta_0|x)\right)^2}{-\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log L(\theta_0|x)\right]} \overset{approx.}{\sim} \chi^2_1 \text{ under } H_0$$

And this works for any number of covariates...

# Multicollinearity and the Variance Inflation Factor (VIF)

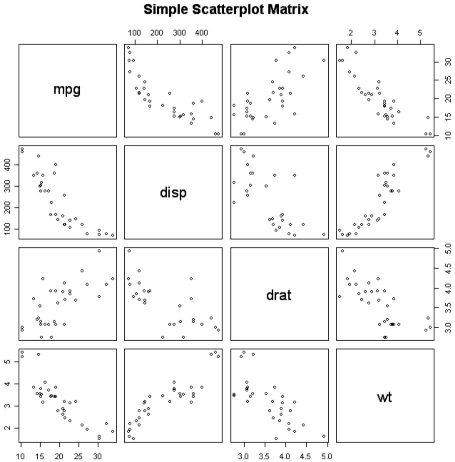
And when you have any number of covariates (features)...

$$\hat{\beta} \sim MVN\left(\beta, \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\right)$$

# Multicollinearity and the Variance Inflation Factor (VIF)

And when you have any number of covariates (features)...

$$\hat{\beta} \sim MVN\left(\beta, \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\right)$$



# Multicollinearity and the Variance Inflation Factor (VIF)

And when you have any number of covariates (features)...

$$\hat{\beta} \sim MVN\left(\beta, \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\right)$$

| | DJIA | S&P 500 | Nasdaq | Canada | Mexico | Brazil | Stoxx 50 | FTSE 100 | CAC 40 | DAX | IBEX | Italy | Netherlands | Sweden | Switzerland | Nikkei | Hang Seng | Australia |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **DJIA** | | 0.97 | 0.85 | 0.57 | 0.56 | 0.52 | 0.52 | 0.48 | 0.51 | 0.56 | 0.49 | 0.50 | 0.50 | 0.42 | 0.42 | 0.09 | 0.11 | 0.07 |
| **S&P 500** | 0.97 | | 0.91 | 0.62 | 0.58 | 0.55 | 0.50 | 0.47 | 0.50 | 0.55 | 0.48 | 0.50 | 0.49 | 0.41 | 0.41 | 0.09 | 0.11 | 0.05 |
| **Nasdaq** | 0.85 | 0.91 | | 0.58 | 0.56 | 0.52 | 0.48 | 0.43 | 0.48 | 0.54 | 0.47 | 0.48 | 0.48 | 0.42 | 0.38 | 0.14 | 0.16 | 0.07 |
| **Canada** | 0.57 | 0.62 | 0.58 | | 0.53 | 0.53 | 0.42 | 0.45 | 0.41 | 0.41 | 0.42 | 0.42 | 0.39 | 0.37 | 0.35 | 0.17 | 0.22 | 0.17 |
| **Mexico** | 0.56 | 0.58 | 0.56 | 0.53 | | 0.56 | 0.42 | 0.42 | 0.44 | 0.43 | 0.43 | 0.44 | 0.39 | 0.38 | 0.38 | 0.17 | 0.25 | 0.17 |
| **Brazil** | 0.52 | 0.55 | 0.52 | 0.53 | 0.56 | | 0.33 | 0.35 | 0.32 | 0.34 | 0.34 | 0.34 | 0.29 | 0.30 | 0.28 | 0.17 | 0.22 | 0.15 |
| **Stoxx 50** | 0.52 | 0.50 | 0.48 | 0.42 | 0.42 | 0.33 | | 0.92 | 0.94 | 0.89 | 0.87 | 0.88 | 0.92 | 0.78 | 0.86 | 0.26 | 0.30 | 0.24 |
| **FTSE 100** | 0.48 | 0.47 | 0.43 | 0.45 | 0.42 | 0.35 | 0.92 | | 0.86 | 0.80 | 0.80 | 0.82 | 0.84 | 0.73 | 0.78 | 0.26 | 0.30 | 0.26 |
| **CAC 40** | 0.51 | 0.50 | 0.48 | 0.41 | 0.44 | 0.32 | 0.94 | 0.86 | | 0.89 | 0.88 | 0.89 | 0.92 | 0.78 | 0.84 | 0.28 | 0.32 | 0.25 |
| **DAX** | 0.56 | 0.55 | 0.54 | 0.41 | 0.43 | 0.34 | 0.89 | 0.80 | 0.89 | | 0.83 | 0.84 | 0.86 | 0.75 | 0.77 | 0.26 | 0.29 | 0.21 |
| **IBEX** | 0.49 | 0.48 | 0.47 | 0.42 | 0.43 | 0.34 | 0.87 | 0.80 | 0.88 | 0.83 | | 0.84 | 0.83 | 0.75 | 0.77 | 0.27 | 0.32 | 0.26 |
| **Italy** | 0.50 | 0.50 | 0.48 | 0.42 | 0.44 | 0.34 | 0.88 | 0.82 | 0.89 | 0.84 | 0.84 | | 0.85 | 0.74 | 0.78 | 0.24 | 0.29 | 0.23 |
| **Netherlands** | 0.50 | 0.49 | 0.48 | 0.39 | 0.39 | 0.29 | 0.92 | 0.84 | 0.92 | 0.86 | 0.83 | 0.85 | | 0.75 | 0.82 | 0.27 | 0.30 | 0.23 |
| **Sweden** | 0.42 | 0.41 | 0.42 | 0.37 | 0.38 | 0.30 | 0.78 | 0.73 | 0.78 | 0.75 | 0.75 | 0.74 | 0.75 | | 0.75 | 0.29 | 0.33 | 0.27 |
| **Switzerland** | 0.42 | 0.41 | 0.38 | 0.35 | 0.38 | 0.28 | 0.86 | 0.78 | 0.84 | 0.77 | 0.77 | 0.78 | 0.82 | 0.75 | | 0.29 | 0.32 | 0.29 |
| **Nikkei** | 0.09 | 0.09 | 0.14 | 0.17 | 0.17 | 0.17 | 0.26 | 0.26 | 0.28 | 0.26 | 0.27 | 0.24 | 0.27 | 0.29 | 0.29 | | 0.52 | 0.49 |
| **Hang Seng** | 0.11 | 0.11 | 0.16 | 0.22 | 0.25 | 0.22 | 0.30 | 0.30 | 0.32 | 0.29 | 0.32 | 0.29 | 0.30 | 0.33 | 0.32 | 0.52 | | 0.48 |
| **Australia** | 0.07 | 0.05 | 0.07 | 0.17 | 0.17 | 0.15 | 0.24 | 0.26 | 0.25 | 0.21 | 0.26 | 0.23 | 0.23 | 0.27 | 0.29 | 0.49 | 0.48 | |

# Multicollinearity and the Variance Inflation Factor (VIF)

And when you have any number of covariates (features)...

$$\hat{\boldsymbol{\beta}} \sim MVN\left(\boldsymbol{\beta}, \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\right)$$

$$\widehat{\text{Var}}[\hat{\beta}_j] = \frac{\hat{\sigma}^2}{(n-1)\widehat{\text{Var}}[X_j]} \cdot \frac{1}{1-R_j^2} \qquad \text{[VIF]}$$

where $R_j^2$ is the $R^2$ of $X_j$ regressed on all the other $X$'s

# Multicollinearity and the Variance Inflation Factor (VIF)

And when you have any number of covariates (features)...

$$\hat{\beta} \sim MVN\left(\beta, \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\right)$$

$$\widehat{\mathrm{Var}}[\hat{\beta}_j] = \frac{\hat{\sigma}^2}{(n-1)\widehat{\mathrm{Var}}[X_j]} \cdot \frac{1}{1-R_j^2} \qquad \text{[VIF]}$$

where $R_j^2$ is the $R^2$ of $X_j$ regressed on all the other $X$'s

Centering $X$'s can decorrelate $X$ and $X^2$...

# Multivariate Regression

That seriously major transition that just happened that you didn't even notice?

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i,\ \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

$$\boldsymbol{Y} \sim \mathrm{MVN}(\mathbf{x}\beta, \sigma^2 I)$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = MVN\left( \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ 1 & x_{13} & x_{23} & \cdots & x_{p3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \right)$$

# Multivariate Regression

That seriously major transition that just happened that you didn't even notice?

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \ \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

$$\boldsymbol{Y} \sim \mathrm{MVN}(\mathbf{x}\beta, \sigma^2 I)$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = MVN\left( \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ 1 & x_{13} & x_{23} & \cdots & x_{p3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \right)$$

- It's just a (multivariate) normal distribution

# Multivariate Regression

That seriously major transition that just happened that you didn't even notice?

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i,\ \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

$$\boldsymbol{Y} \sim \mathrm{MVN}(\mathbf{x}\beta, \sigma^2 I)$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = MVN\left( \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ 1 & x_{13} & x_{23} & \cdots & x_{p3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \right)$$

- It's just a (multivariate) normal distribution
- with a *linear model* component for the mean

# Multivariate Regression

That seriously major transition that just happened that you didn't even notice?

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \;\; \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

$$\boldsymbol{Y} \sim \mathrm{MVN}(\mathbf{x}\beta, \sigma^2 I)$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = MVN \left( \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ 1 & x_{13} & x_{23} & \cdots & x_{p3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \right)$$

- It's just a (multivariate) normal distribution
- with a *linear model* component for the mean just

# Multivariate Regression

That seriously major transition that just happened that you didn't even notice?

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i,\ \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

$$\boldsymbol{Y} \sim \mathrm{MVN}(\mathbf{x}\beta, \sigma^2 I)$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = MVN\left( \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ 1 & x_{13} & x_{23} & \cdots & x_{p3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \right)$$

- It's just a (multivariate) normal distribution
- with a *linear model* component for the mean just like

# Multivariate Regression

That seriously major transition that just happened that you didn't even notice?

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i,\ \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

$$\boldsymbol{Y} \sim \mathrm{MVN}(\mathbf{x}\beta, \sigma^2 I)$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = MVN\left( \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ 1 & x_{13} & x_{23} & \cdots & x_{p3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \right)$$

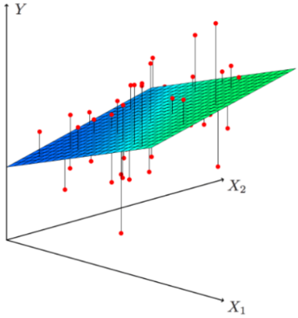
- It's just a (multivariate) normal distribution
- with a *linear model* component for the mean just like before

## Linear Models

- Linear model... that sounds too simple...

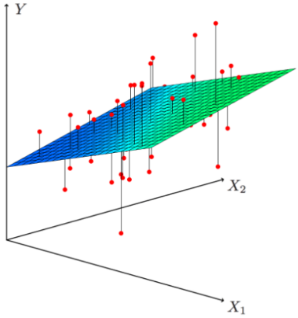
# Linear Models

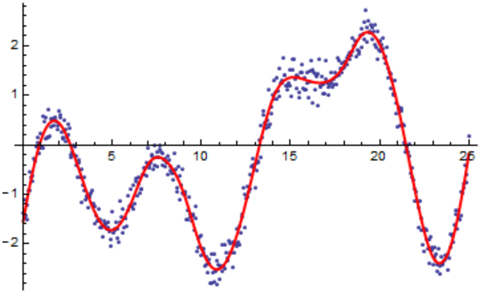
- Linear model... that sounds too simple...



# Linear Models

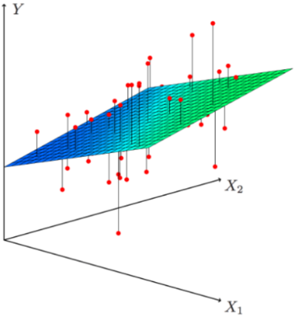
- Linear model... that sounds too simple...

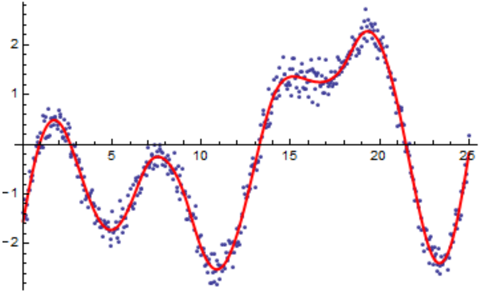


# Linear Models

- Linear model... that sounds too simple...




- "Linear" models are only linear in the coefficients

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

- The $x$'s can be pretty wild...

# Linear models that produce “non-linear” response surfaces

# Linear models that produce "non-linear" response surfaces

- Higher order terms: $X_1^{\frac{1}{2}}, X_1^2, X_1^3$

## Linear models that produce “non-linear” response surfaces

- Higher order terms: $X_1^{\frac{1}{2}}, X_1^2, X_1^3$
- Qualitative variables: $X_1 \in \{0, 1\}, X_1 \in \{a, b, c, d\}$

# Linear models that produce "non-linear" response surfaces

- Higher order terms: $X_1^{\frac{1}{2}}, X_1^2, X_1^3$
- Qualitative variables: $X_1 \in \{0, 1\}, X_1 \in \{a, b, c, d\}$

$$\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} \neq \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

# Linear models that produce "non-linear" response surfaces

- Higher order terms: $X_1^{\frac{1}{2}}, X_1^2, X_1^3$
- Qualitative variables: $X_1 \in \{0, 1\}, X_1 \in \{a, b, c, d\}$

$$\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} \neq \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

- Interactions: $X_1 \cdot X_2$

# Linear models that produce "non-linear" response surfaces

- Higher order terms: $X_1^{\frac{1}{2}}, X_1^2, X_1^3$
- Qualitative variables: $X_1 \in \{0, 1\}, X_1 \in \{a, b, c, d\}$

$$\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} \neq \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

- Interactions: $X_1 \cdot X_2$

Interpret: $\mathrm{E}[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{2^2} X_2^2 + \beta_{13} X_1 X_3$

# Linear models that produce "non-linear" response surfaces

- Higher order terms: $X_1^{\frac{1}{2}}, X_1^2, X_1^3$
- Qualitative variables: $X_1 \in \{0, 1\}, X_1 \in \{a, b, c, d\}$

$$\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} \neq \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

- Interactions: $X_1 \cdot X_2$

Interpret: $\mathrm{E}[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{2^2} X_2^2 + \beta_{13} X_1 X_3$
Is it possible to "vary one $X$ and hold all others constant"?

# Linear models that produce “non-linear” response surfaces

- Higher order terms: $X_1^{\frac{1}{2}}, X_1^2, X_1^3$
- Qualitative variables: $X_1 \in \{0,1\}, X_1 \in \{a,b,c,d\}$

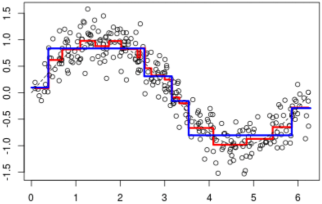
$$\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} \neq \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

- Interactions: $X_1 \cdot X_2$

Interpret: $\mathrm{E}[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{2^2} X_2^2 + \beta_{13} X_1 X_3$
Is it possible to “vary one $X$ and hold all others constant”?

Step functions

$$Y_i = \beta_j : \textit{if } a_j \leq X_i < b_j$$



# Linear models that produce "non-linear" response surfaces

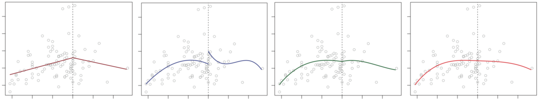
- Higher order terms: $X_1^{\frac{1}{2}}, X_1^2, X_1^3$
- Qualitative variables: $X_1 \in \{0,1\}, X_1 \in \{a,b,c,d\}$

$$\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} \neq \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

- Interactions: $X_1 \cdot X_2$

Interpret: $\mathrm{E}[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{2^2} X_2^2 + \beta_{13} X_1 X_3$
Is it possible to "vary one $X$ and hold all others constant"?

Step functions
Regression Splines



$$Y_i = \begin{cases} \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \epsilon_i & : if \;\; X_i \leq c \\ \beta_0^* + \beta_1 X_i + \beta_2^* X_i^2 + \beta_3^* X_i^3 + \epsilon_i & : if \;\; X_i > c \end{cases}$$

# Linear models that produce "non-linear" response surfaces

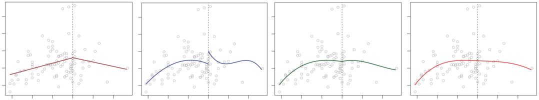
- Higher order terms: $X_1^{\frac{1}{2}}, X_1^2, X_1^3$
- Qualitative variables: $X_1 \in \{0, 1\}, X_1 \in \{a, b, c, d\}$

$$\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} \neq \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

- Interactions: $X_1 \cdot X_2$

Interpret: $\mathrm{E}[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{2^2} X_2^2 + \beta_{13} X_1 X_3$
Is it possible to "vary one $X$ and hold all others constant"?

Step functions
Regression Splines



$$h(X_i, \xi) = \begin{cases} (x - \xi)^3 & : \mathit{if}\ X_i > \xi \\ 0 & : \mathit{if}\ X_i \leq \xi \end{cases} \qquad Y_i = \begin{array}{l} \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 \\ + \beta_{s1} h(X_i, \xi_1) + \cdots + \epsilon_i \end{array}$$

# Linear models that produce "non-linear" response surfaces

- Higher order terms: $X_1^{\frac{1}{2}}, X_1^2, X_1^3$
- Qualitative variables: $X_1 \in \{0,1\}, X_1 \in \{a,b,c,d\}$

$$\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} \neq \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

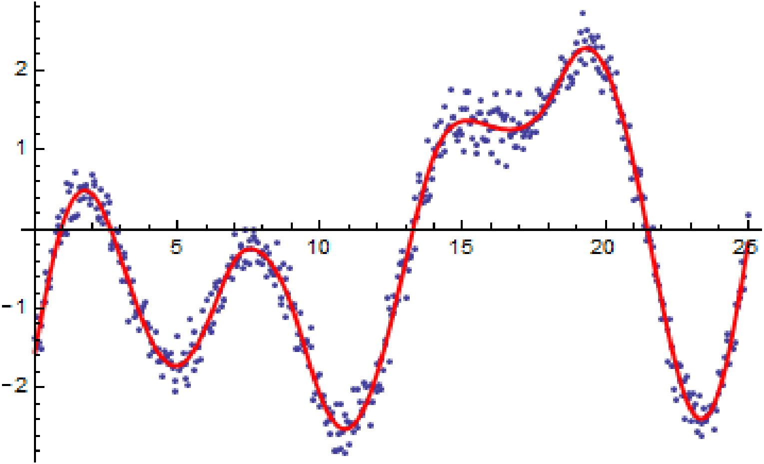
- Interactions: $X_1 \cdot X_2$

Interpret: $\mathrm{E}[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{2^2} X_2^2 + \beta_{13} X_1 X_3$
Is it possible to "vary one $X$ and hold all others constant"?

Step functions
Regression Splines
Smoothing Splines

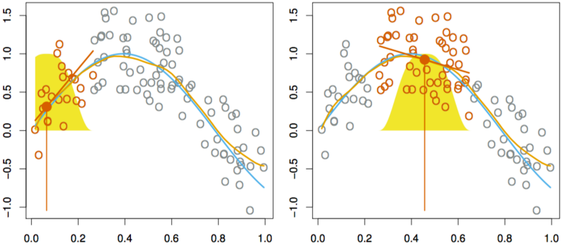
$$\min_g \sum_{i=1}^{n} (Y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

# Linear models aren't really so "linear"



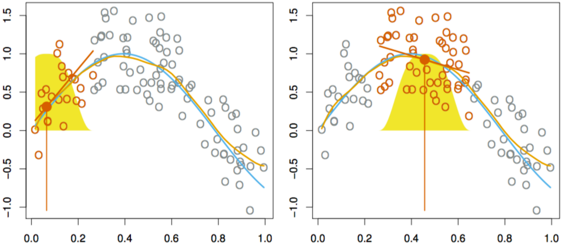
# Other ways to get "non linear" response surfaces

- Local Regression (LOESS)



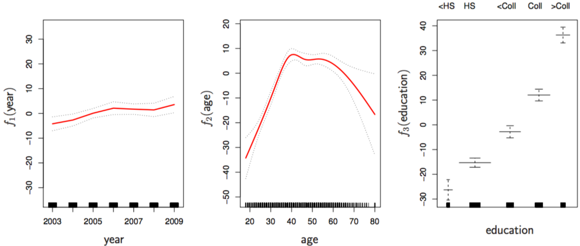
# Other ways to get "non linear" response surfaces

- Local Regression (LOESS)




- Generalized Additive Models

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i.$$



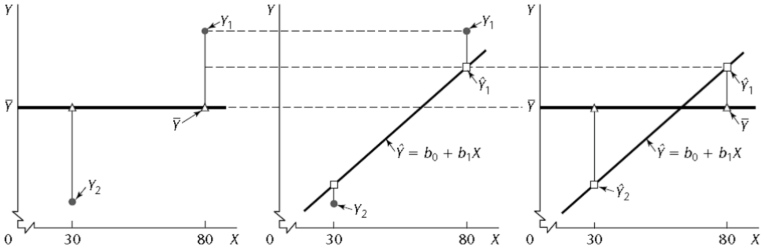
# Model Fit

| | |
|---|---|
| Residual Variation<br><br>$RSS = \sum(Y_i - \hat{Y}_i)^2 = \sum \hat{\epsilon}_i^2$ | Total Variation<br><br>$TSS = \sum(Y_i - \bar{Y})^2$<br>$= RSS + \sum(\hat{Y}_i - \bar{Y})^2$ |
| Residual Standard Error<br><br>$RSE = \sqrt{\frac{1}{n-p-1} RSS}$<br>$= \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2}{n-p-1}}$ | Proportion of Variance Explained<br><br>$R^2 = \frac{TSS-RSS}{TSS}$<br>$= 1 - \frac{RSS}{TSS}$ |

$F$-test

$$F = \frac{(TSS-RSS)/p}{RSS/(n-p-1)} = \frac{\sum(\hat{Y}_i - \bar{Y})^2/p}{\sum(Y_i - \hat{Y})/(n-p-1)}$$

# Decomposition of Total Variation




$$
\begin{aligned}
TSS &= \sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\
&= \sum(Y_i - \hat{Y}_i)^2 + 2\sum(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + \sum(\hat{Y}_i - \bar{Y})^2 \\
&= \sum(Y_i - \hat{Y}_i)^2 + 2\sum \hat{\epsilon}_i(\hat{Y}_i - \bar{Y}) + \sum(\hat{Y}_i - \bar{Y})^2 \\
& \qquad\qquad \sum \hat{\epsilon}_i = 0 \uparrow \quad \uparrow \sum \hat{\epsilon}_i \hat{Y}_i = 0 \\
&= \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2 = RSS + \sum(\hat{Y}_i - \bar{Y})^2
\end{aligned}
$$

# Hypothesis Testing for Feature Selection

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} = \frac{\sum(\hat{Y}_i - \bar{Y})^2/p}{\sum(Y_i - \hat{Y})/(n-p-1)}$$

$F \sim F_{p,n-p-1}$ (tests if any coefficient is *non-zero*)

$\frac{\hat{\beta}_i - \beta_i}{\mathrm{SE}(\hat{\beta}_i)} \sim t_{n-p-1}$ (tests if a *specific* coefficient is non-zero*)

*in the presence of all the others (this is a "last-in" test)

# Hypothesis Testing for Feature Selection

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} = \frac{\sum(\hat{Y}_i - \bar{Y})^2/p}{\sum(Y_i - \hat{Y})/(n-p-1)}$$

$$F \sim F_{p,n-p-1}$$ (tests if any coefficient is *non-zero*)

$$\frac{\hat{\beta}_i - \beta_i}{\mathrm{SE}(\hat{\beta}_i)} \sim t_{n-p-1}$$ (tests if a *specific* coefficient is non-zero*)

*in the presence of all the others (this is a “last-in” test)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.933
Model:                            OLS   Adj. R-squared:                  0.928
Method:                 Least Squares   F-statistic:                     211.8
Date:                Mon, 03 Nov 2014   Prob (F-statistic):           6.30e-27
Time:                        14:45:06   Log-Likelihood:                -34.438
No. Observations:                  50   AIC:                             76.88
Df Residuals:                      46   BIC:                             84.52
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1             0.4687      0.026     17.751      0.000         0.416     0.522
x2             0.4836      0.104      4.659      0.000         0.275     0.693
x3            -0.0174      0.002     -7.507      0.000        -0.022    -0.013
const          5.2058      0.171     30.405      0.000         4.861     5.550
==============================================================================
```

# Hypothesis Testing for Feature Selection

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} = \frac{\sum(\hat{Y}_i - \bar{Y})^2/p}{\sum(Y_i - \hat{Y})/(n-p-1)}$$

$$F \sim F_{p,n-p-1} \quad \text{(tests if any coefficient is \textit{non-zero})}$$

$$\frac{\hat{\beta}_i - \beta_i}{\mathrm{SE}(\hat{\beta}_i)} \sim t_{n-p-1} \quad \text{(tests if a \textit{specific} coefficient is non-zero*)}$$

*in the presence of all the others (this is a "last-in" test)

- Forward Selection

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.933
Model:                            OLS   Adj. R-squared:                  0.928
Method:                 Least Squares   F-statistic:                     211.8
Date:                Mon, 03 Nov 2014   Prob (F-statistic):           6.30e-27
Time:                        14:45:06   Log-Likelihood:                -34.438
No. Observations:                  50   AIC:                             76.88
Df Residuals:                      46   BIC:                             84.52
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1             0.4687      0.026     17.751      0.000         0.416     0.522
x2             0.4836      0.104      4.659      0.000         0.275     0.693
x3            -0.0174      0.002     -7.507      0.000        -0.022    -0.013
const          5.2058      0.171     30.405      0.000         4.861     5.550
==============================================================================
```

# Hypothesis Testing for Feature Selection

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} = \frac{\sum(\hat{Y}_i - \bar{Y})^2/p}{\sum(Y_i - \hat{Y})/(n-p-1)}$$

$F \sim F_{p,n-p-1}$ (tests if any coefficient is *non-zero*)

$\frac{\hat{\beta}_i - \beta_i}{\text{SE}(\hat{\beta}_i)} \sim t_{n-p-1}$ (tests if a *specific* coefficient is non-zero*)

*in the presence of all the others (this is a "last-in" test)

- Forward Selection
- Backward Selection

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | y | R-squared: | 0.933 |
| Model: | OLS | Adj. R-squared: | 0.928 |
| Method: | Least Squares | F-statistic: | 211.8 |
| Date: | Mon, 03 Nov 2014 | Prob (F-statistic): | 6.30e-27 |
| Time: | 14:45:06 | Log-Likelihood: | -34.438 |
| No. Observations: | 50 | AIC: | 76.88 |
| Df Residuals: | 46 | BIC: | 84.52 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [95.0% Conf. | Int.] |
|---|---|---|---|---|---|---|
| x1 | 0.4687 | 0.026 | 17.751 | 0.000 | 0.416 | 0.522 |
| x2 | 0.4836 | 0.104 | 4.659 | 0.000 | 0.275 | 0.693 |
| x3 | -0.0174 | 0.002 | -7.507 | 0.000 | -0.022 | -0.013 |
| const | 5.2058 | 0.171 | 30.405 | 0.000 | 4.861 | 5.550 |

# Hypothesis Testing for Feature Selection

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} = \frac{\sum(\hat{Y}_i - \bar{Y})^2/p}{\sum(Y_i - \hat{Y})/(n-p-1)}$$

$$F \sim F_{p,n-p-1}$$ (tests if any coefficient is *non-zero*)

$$\frac{\hat{\beta}_i - \beta_i}{\text{SE}(\hat{\beta}_i)} \sim t_{n-p-1}$$ (tests if a *specific* coefficient is non-zero*)

*in the presence of all the others (this is a "last-in" test)

- Forward Selection
- Backward Selection
- Both

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | y | R-squared: | $\boxed{0.933}$ |
| Model: | OLS | Adj. R-squared: | 0.928 |
| Method: | Least Squares | F-statistic: | 211.8 |
| Date: | Mon, 03 Nov 2014 | Prob (F-statistic): | 6.30e-27 |
| Time: | 14:45:06 | Log-Likelihood: | -34.438 |
| No. Observations: | 50 | AIC: | 76.88 |
| Df Residuals: | 46 | BIC: | 84.52 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [95.0% Conf. Int.] | |
|---|---|---|---|---|---|---|
| x1 | 0.4687 | 0.026 | 17.751 | 0.000 | 0.416 | 0.522 |
| x2 | 0.4836 | 0.104 | 4.659 | 0.000 | 0.275 | 0.693 |
| x3 | -0.0174 | 0.002 | -7.507 | 0.000 | -0.022 | -0.013 |
| const | 5.2058 | 0.171 | 30.405 | 0.000 | 4.861 | 5.550 |

# Model Selection

- Selecting statistically significant features is good

# Model Selection

- Selecting statistically significant features is good
- $R^2$ (model fit) is insufficient – more features means larger $R^2$

# Model Selection

- Selecting statistically significant features is good
- $R^2$ (model fit) is insufficient – more features means larger $R^2$
- Spuriously improving model fit to data is called *overfitting*

# Model Selection

- Selecting statistically significant features is good
- $R^2$ (model fit) is insufficient – more features means larger $R^2$
- Spuriously improving model fit to data is called *overfitting*
- We want model fits to generalize to *population* phenomenon

# Model Selection

- Selecting statistically significant features is good
- $R^2$ (model fit) is insufficient – more features means larger $R^2$
- Spuriously improving model fit to data is called *overfitting*
- We want model fits to generalize to *population* phenomenon

- Classical Model Selection Criterion

| | | |
|---|---|---|
| Mallow's $C_p$ | $\frac{1}{n}(RSS + 2p\hat{\sigma}^2)$ | |
| $AIC$ | $-2\log L + 2p$ | $D_M = -2\log f(Y \mid \hat{\theta}^{M_p}) + 2\log f(Y \mid Y)$ |
| $BIC$ | $-2\log L + p\log n$ | $D_M \sim \chi^2_{n-p}$ |
| *Adjusted* $R^2$ | $1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$ | |

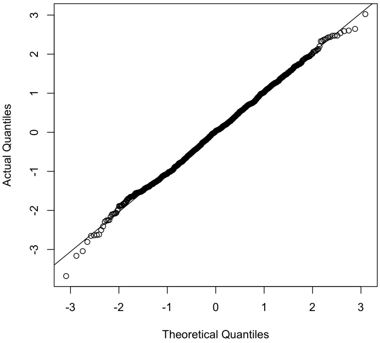
# Assumptions, violations, and remedial measures

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \ \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

# Assumptions, violations, and remedial measures

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \ \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

- Normality



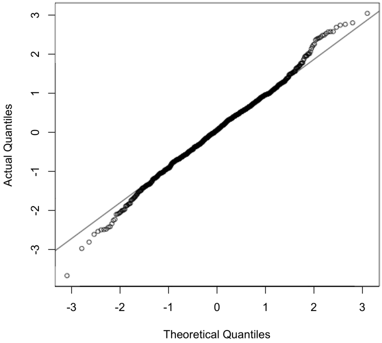

Q-Q Plot

Hypothesis testing depends on distributional assumptions

# Assumptions, violations, and remedial measures

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \ \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

- Normality



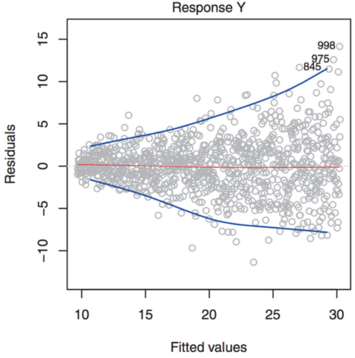

Q-Q Plot

Hypothesis testing depends on distributional assumptions

# Assumptions, violations, and remedial measures

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \ \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

- Normality
- Homoskedasticity



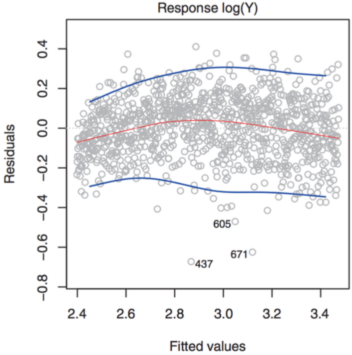

Residuals versus Fitted Values

Box-Cox transformations $\frac{Y^\lambda - 1}{\lambda}$ can help

# Assumptions, violations, and remedial measures

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i,\ \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

- Normality
- Homoskedasticity




Residuals versus Fitted Values

Box-Cox transformations $\frac{Y^\lambda - 1}{\lambda}$ can help

# Assumptions, violations, and remedial measures

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \; \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

- Normality
- Homoskedasticity
- Independence

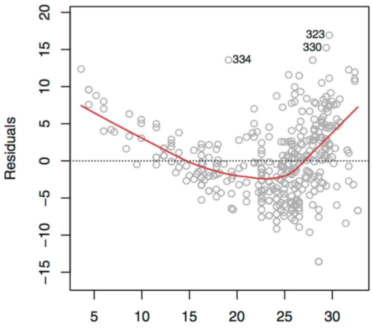
$$\mathrm{Cov}[\boldsymbol{Y} - \boldsymbol{X}\beta] = (\boldsymbol{Y} - \boldsymbol{X}\beta])^T(\boldsymbol{Y} - \boldsymbol{X}\beta])$$

$$\approx \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

# Assumptions, violations, and remedial measures

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \ \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

- Normality
- Homoskedasticity
- Independence
- Linear form




Residuals versus Feature Values

"All models are wrong, some are useful"
– George Box

# Assumptions, violations, and remedial measures

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \ \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

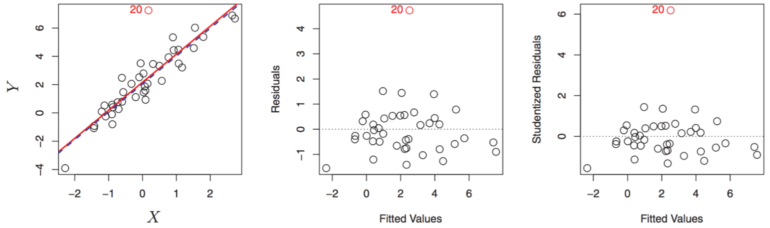
- Normality
- Homoskedasticity
- Independence
- Linear form
- Fixed $x$'s

# Concerns: Outliers

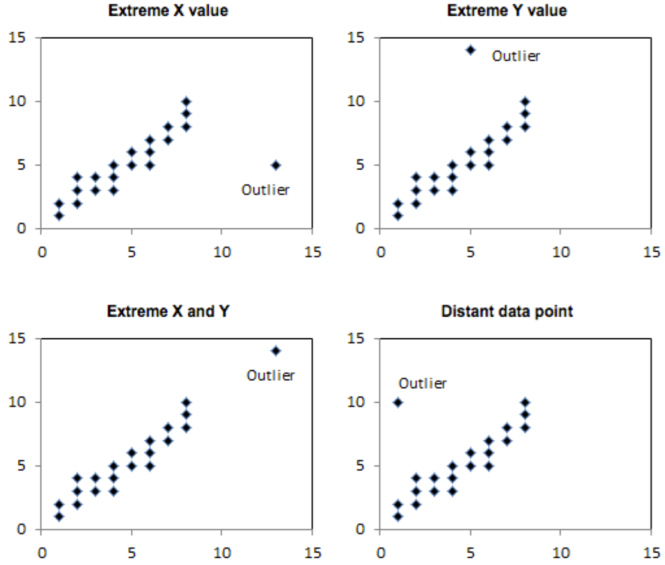
Studentized Residuals (as in "Student's t-distributioned" residuals)

$$r_i = \frac{\hat{\epsilon}_i}{\hat{s}igma^2} = \frac{\text{residual}_i}{\text{RSE}}$$

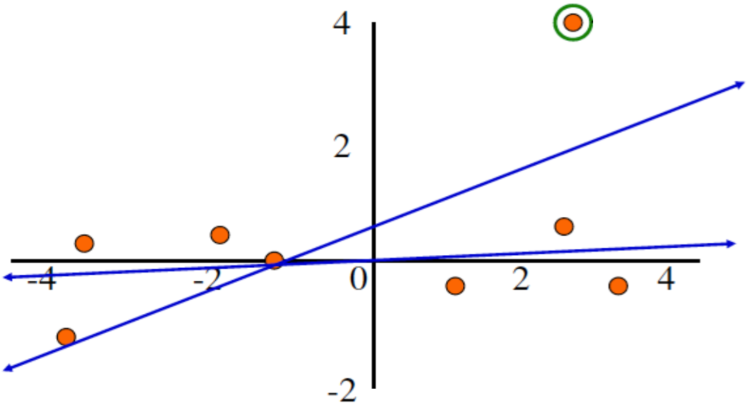
have a t-distribution



# Concerns: Outliers



# Concerns: Outliers



# Concerns: High Leverage Points

The *hat* matrix $H$ "puts the hat on $\boldsymbol{Y}$": $H$ projects $\boldsymbol{Y}$ onto $\hat{\boldsymbol{Y}}$ – the (least squares) closest vector (to $\boldsymbol{Y}$) in the column space of $x$

$$
\begin{aligned}
H &= \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T \\
\hat{\boldsymbol{Y}} &= \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y} \\
&= H\mathbf{Y}
\end{aligned}
$$

# Concerns: High Leverage Points

The *hat* matrix $H$ "puts the hat on $\boldsymbol{Y}$": $H$ projects $\boldsymbol{Y}$ onto $\hat{\boldsymbol{Y}}$ – the (least squares) closest vector (to $\boldsymbol{Y}$) in the column space of $x$

$$
\begin{aligned}
H &= \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T \\
\hat{\mathbf{Y}} &= \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y} \\
&= H\mathbf{Y}
\end{aligned}
$$

- Diagonal element $H_{ii} \in [0, 1]$, and $\sum_{i=1}^{n} H_{ii} = rank(\mathbf{x})$

# Concerns: High Leverage Points

The *hat* matrix $H$ "puts the hat on $\boldsymbol{Y}$": $H$ projects $\boldsymbol{Y}$ onto $\hat{\boldsymbol{Y}}$ – the (least squares) closest vector (to $\boldsymbol{Y}$) in the column space of $x$

$$\begin{aligned} H &= \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T \\ \hat{\mathbf{Y}} &= \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y} \\ &= H\mathbf{Y} \end{aligned}$$

- Diagonal element $H_{ii} \in [0, 1]$, and $\sum_{i=1}^{n} H_{ii} = rank(\mathbf{x})$
- $H_{ii}$ shows much $\hat{Y}_i$ depends on $Y_i$, however

# Concerns: High Leverage Points

The *hat* matrix $H$ "puts the hat on $\boldsymbol{Y}$": $H$ projects $\boldsymbol{Y}$ onto $\hat{\boldsymbol{Y}}$ – the (least squares) closest vector (to $\boldsymbol{Y}$) in the column space of $x$

$$
\begin{aligned}
H &= \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T \\
\hat{\mathbf{Y}} &= \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y} \\
&= H\mathbf{Y}
\end{aligned}
$$

- Diagonal element $H_{ii} \in [0,1]$, and $\sum_{i=1}^{n} H_{ii} = rank(\mathbf{x})$
- $H_{ii}$ shows much $\hat{Y}_i$ depends on $Y_i$, however

  $H_{ii}$ actually measures "extremeness" of $x_i$

# Concerns: High Leverage Points

The *hat* matrix $H$ "puts the hat on $\boldsymbol{Y}$": $H$ projects $\boldsymbol{Y}$ onto $\hat{\boldsymbol{Y}}$ – the (least squares) closest vector (to $\boldsymbol{Y}$) in the column space of $x$

$$\begin{aligned} H &= \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T \\ \hat{\mathbf{Y}} &= \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y} \\ &= H\mathbf{Y} \end{aligned}$$

- Diagonal element $H_{ii} \in [0, 1]$, and $\sum_{i=1}^{n} H_{ii} = rank(\mathbf{x})$
- $H_{ii}$ shows much $\hat{Y}_i$ depends on $Y_i$, however

  $H_{ii}$ actually measures "extremeness" of $x_i$
- Relative comparison of $H_{ii}$'s id.'s "high leverage observations"

# Concerns: High Leverage Points

The *hat* matrix $H$ "puts the hat on $\boldsymbol{Y}$": $H$ projects $\boldsymbol{Y}$ onto $\hat{\boldsymbol{Y}}$ – the (least squares) closest vector (to $\boldsymbol{Y}$) in the column space of $x$

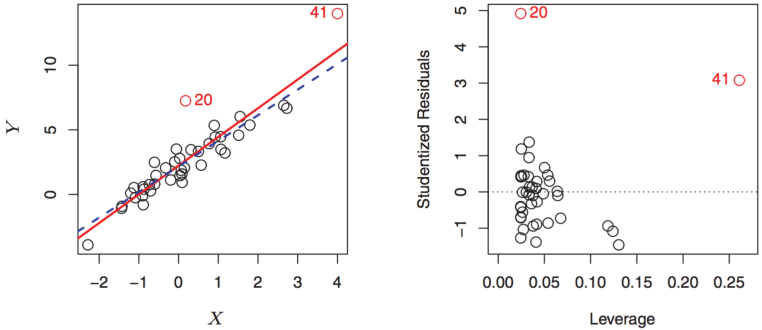
$$
\begin{aligned}
H &= \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T \\
\hat{\mathbf{Y}} &= \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y} \\
&= H\mathbf{Y}
\end{aligned}
$$

- Diagonal element $H_{ii} \in [0, 1]$, and $\sum_{i=1}^{n} H_{ii} = rank(\mathbf{x})$
- $H_{ii}$ shows much $\hat{Y}_i$ depends on $Y_i$, however

  $H_{ii}$ actually measures "extremeness" of $x_i$
- Relative comparison of $H_{ii}$'s id.'s "high leverage observations"

  $H_{ii}$ is called the *leverage* of observation $i$

# Concerns: High Leverage Points



# Concerns: Highly Influential Points – *Cook's distance*

$$D_i \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(-i)}\right)^2}{\hat{\sigma}^2 p} = \frac{\hat{\epsilon}_i}{\hat{\sigma}^2 p} \frac{h_i}{(1 - h_i)^2}$$

Influential data points $i$ might have $D_i$

# Concerns: Highly Influential Points – *Cook's distance*

$$D_i \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_{j(-i)}\right)^2}{\hat{\sigma}^2 p} = \frac{\hat{\epsilon}_i}{\hat{\sigma}^2 p}\frac{h_i}{(1-h_i)^2}$$

Influential data points $i$ might have $D_i$

$> 3 \times \bar{D}$

$> 1$

$> 4/n$

$> F_{p,n-p}^{1-\alpha}$

# Concerns: Highly Influential Points – *Cook's distance*

$$D_i \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(-i)}\right)^2}{\hat{\sigma}^2 p} = \frac{\hat{\epsilon}_i}{\hat{\sigma}^2 p} \frac{h_i}{(1-h_i)^2}$$

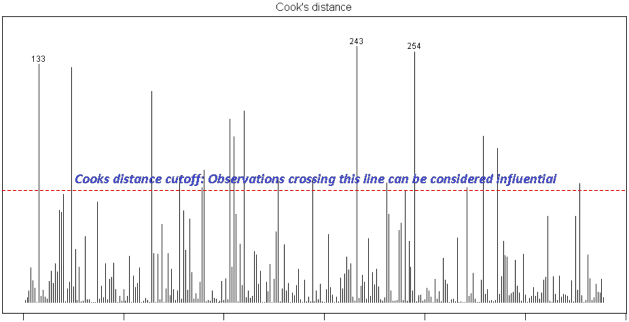
Influential data points $i$ might have $D_i$

$> 3 \times \bar{D}$

$> 1$

$> 4/n$

$> F^{1-\alpha}_{p,n-p}$



# Concerns: Highly Influential Points – *Cook's distance*