

STAT 651



Lecture #20

Topics in Lecture #20

- Outliers and Leverage
- Cook's distance

Book Chapters in Lecture #20

- Small part of Chapter 11.2

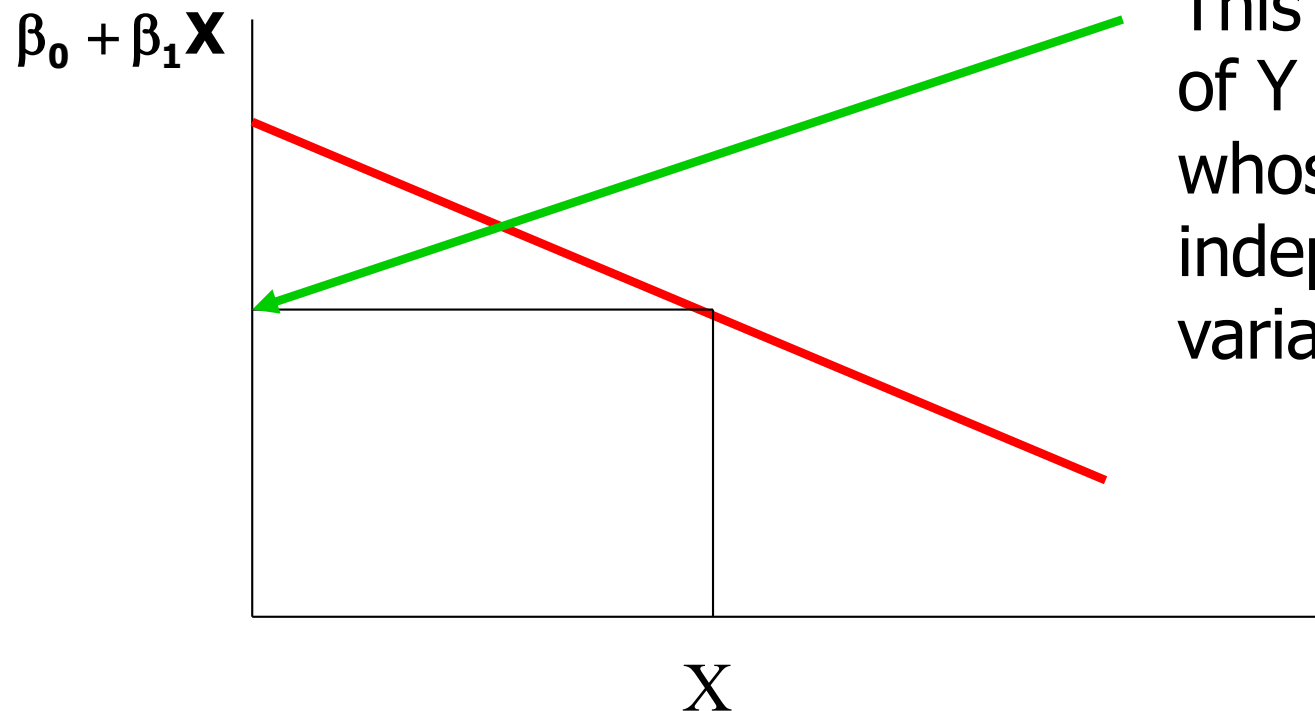
Relevant SPSS Tutorials

- Regression diagnostics
- Diagnostics for problem points

Lecture 19 Review: Population Slope and Intercept

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- If $\beta_1 < 0$ then we have a graph like this:

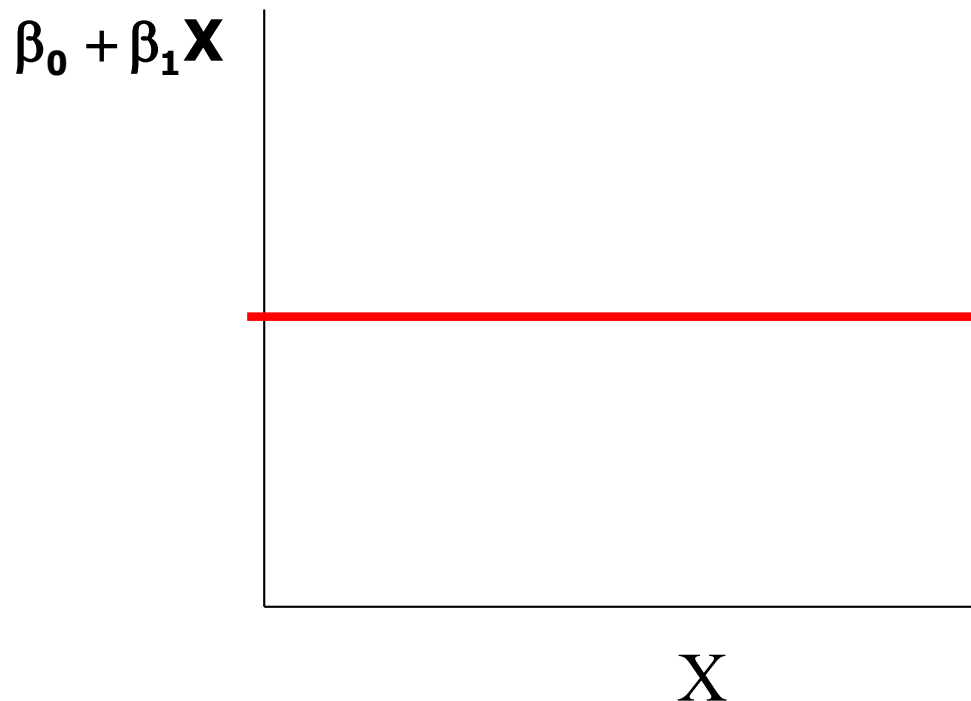


This is the mean of Y for those whose independent variable is X

Lecture 19 Review: Population Slope and Intercept

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- If $\beta_1 = 0$ then we have a graph like this:



Note how the mean of Y does not depend on X : **Y and X are independent**

Lecture 19 Review: Linear Regression

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- If $\beta_1 = 0$ then Y and X are independent
- So, we can test the null hypothesis $H_0 :$ that Y and X are independent by testing
$$H_0 : \beta_1 = 0$$
- The p-value in regression tables tests this hypothesis

Lecture 19 Review: Regression

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X} + \varepsilon$$

- The standard deviation of the errors ε is to be called σ_ε
- This means that every subpopulation who share the same value of X have
 - Mean = $\beta_0 + \beta_1 \mathbf{X}$
 - Standard deviation = σ_ε

Lecture 19 Review: Regression

- The least squares estimate $\hat{\beta}_1$ is a random variable

$$s_{\varepsilon} = \sqrt{\mathbf{MSE}}$$

- Its estimated standard deviation is

$$\mathbf{s.e.}(\hat{\beta}_1) = \frac{s_{\varepsilon}}{\sqrt{\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^2}}$$

Lecture 19 Review: Regression

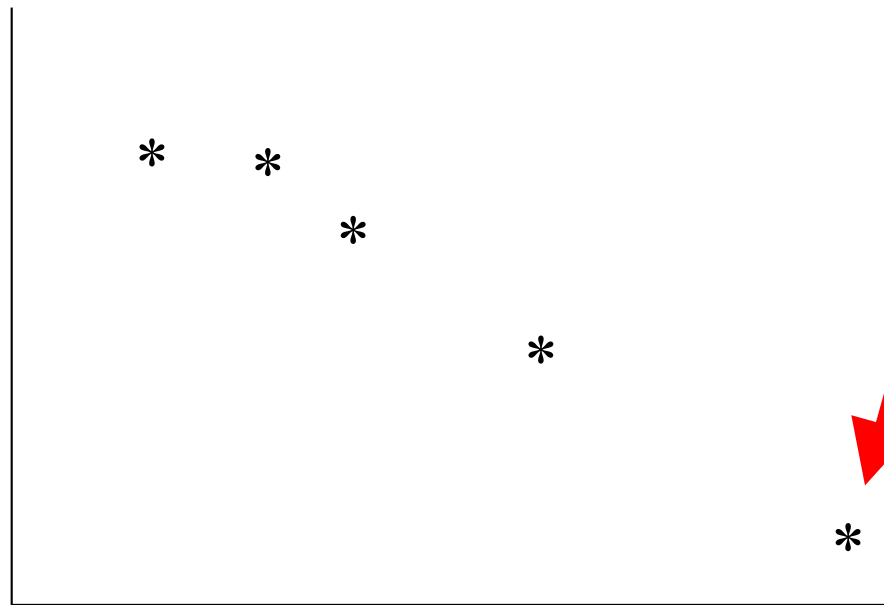
- The $(1-\alpha)100\%$ Confidence interval for the population slope is $\hat{\beta}_1 \pm t_{\alpha/2}(n-2)se(\hat{\beta}_1)$

Lecture 19 Review: Residuals

- You can check the assumption that the errors are normally distributed by constructing a q-q plot of the residuals

Leverage and Outliers

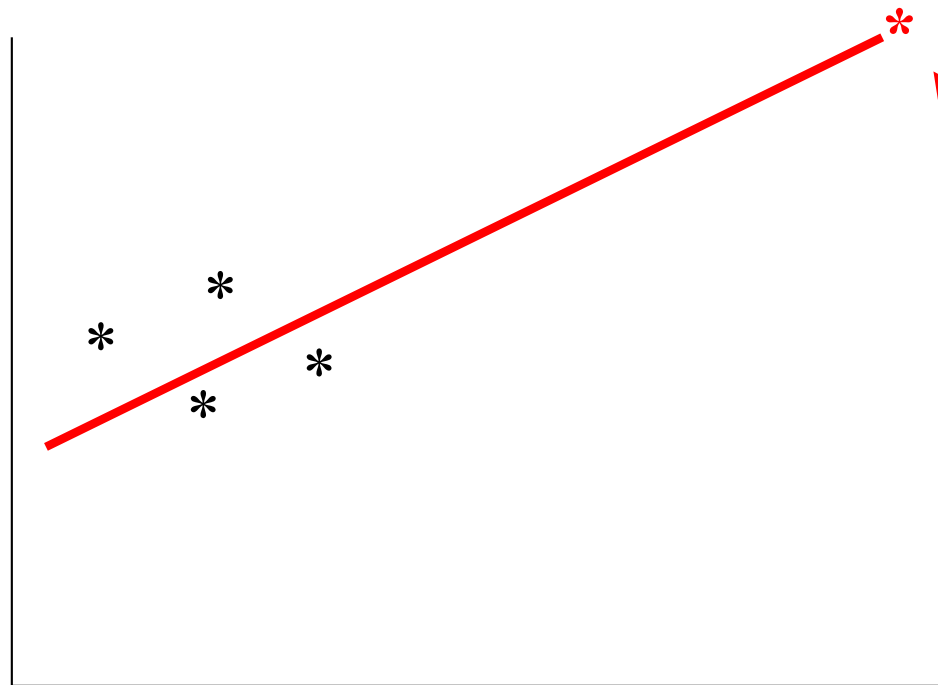
- Outliers in Linear Regression are difficult to diagnose
- They depend crucially on where X is



A boxplot of Y would think this is an outlier, when in reality it fits the line quite well

Outliers and Leverage

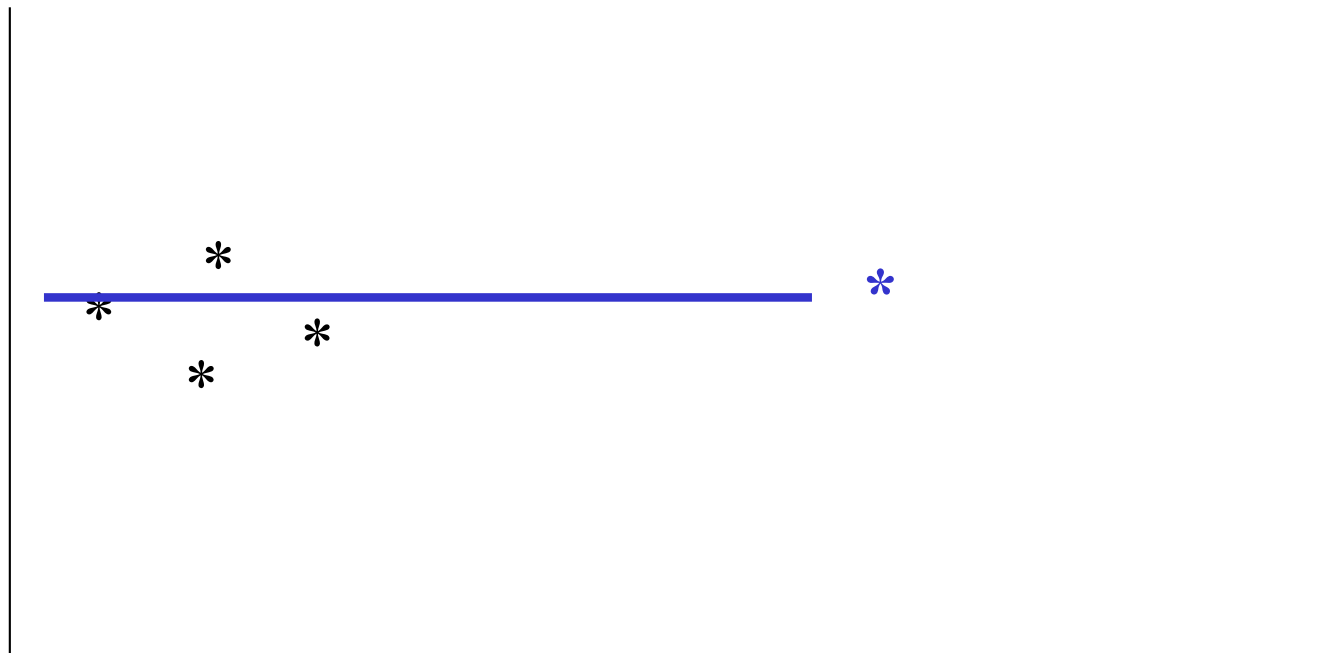
- It's also the case than one observation can have a dramatic impact on the fit



This is called a **leverage value** because its X is so far from the rest, and as we'll see, it exerts a lot of leverage in determining the line

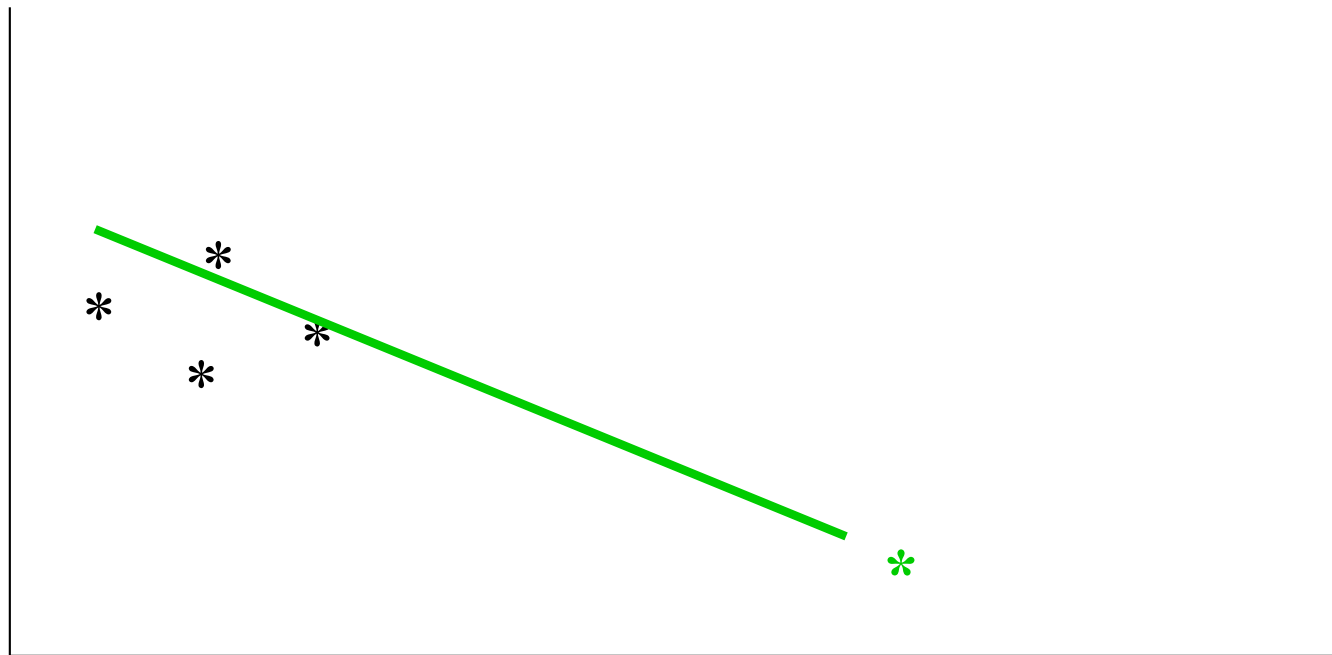
Outliers and Leverage

- It's also the case than one observation can have a dramatic impact on the fit



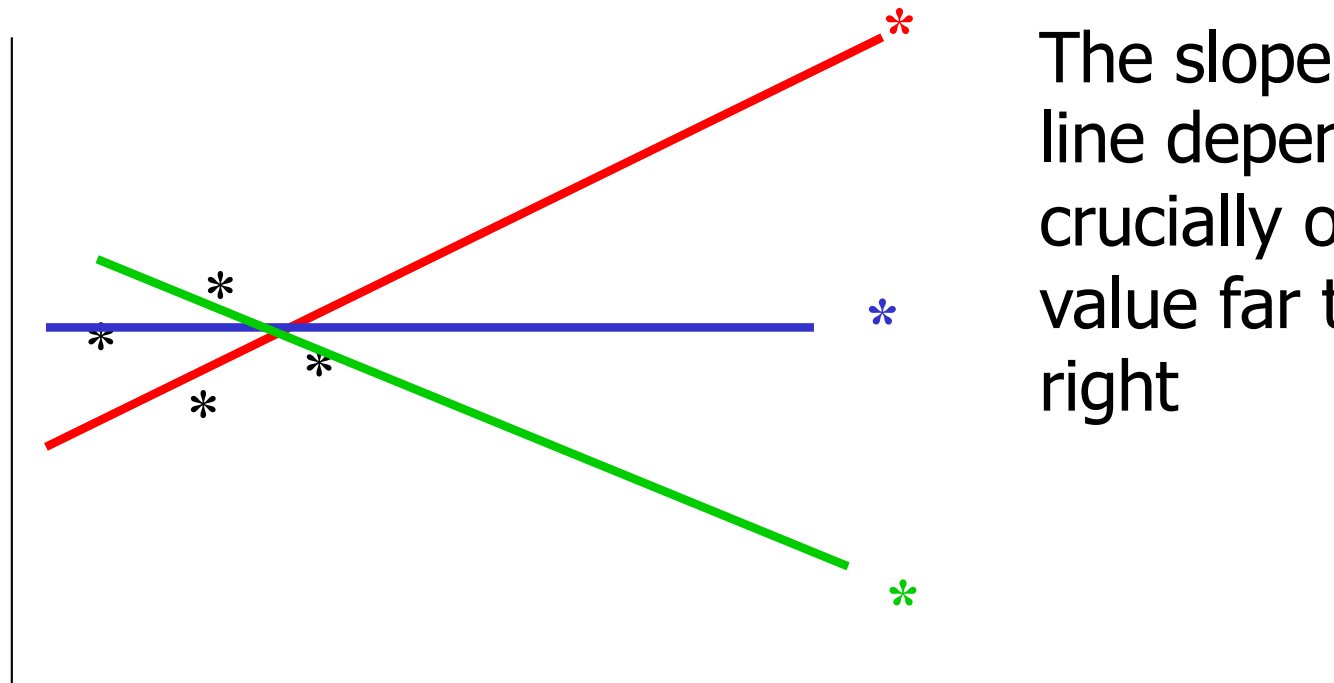
Outliers and Leverage

- It's also the case than one observation can have a dramatic impact on the fit



Outliers and Leverage

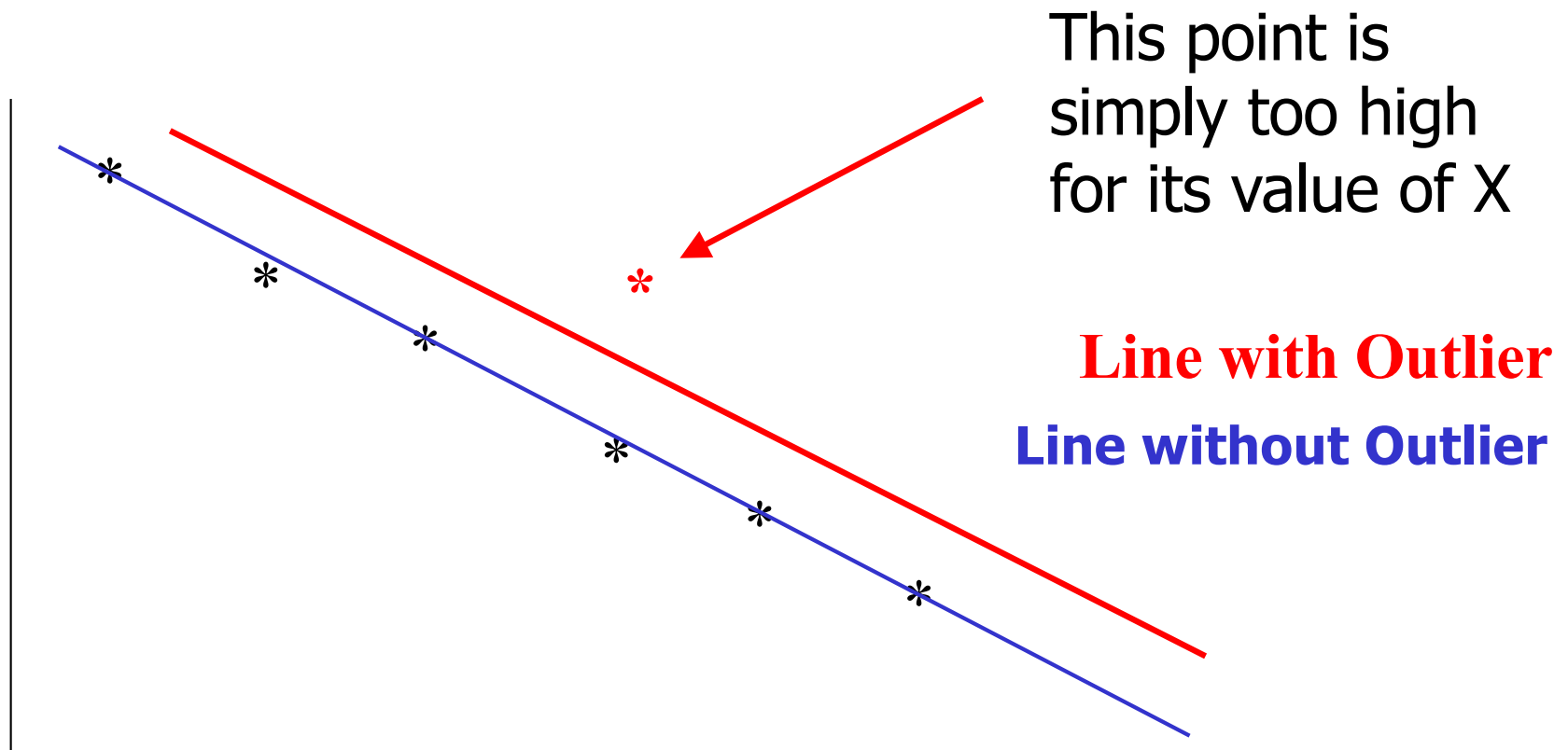
- It's also the case than one observation can have a dramatic impact on the fit



The slope of the line depends crucially on the value far to the right

Outliers and Leverage

- But Outliers can occur



Outliers and Leverage

- A **leverage point** is an observation with a value of X that is outlying among the X values
- An **outlier** is an observation of Y that seems not to agree with the main trend of the data
- Outliers and leverage values can distort the fitted least squares line
- It is thus important to have diagnostics to detect when disaster might strike

Outliers and Leverage

- We have three methods for diagnosing high leverage values and outliers
- Leverage plots: For a single X , these are basically the same as boxplots of the X -space (leverage)
- Cook's distance (measures how much the fitted line changes if the observation is deleted)
- Residual Plots

Outliers and Leverage

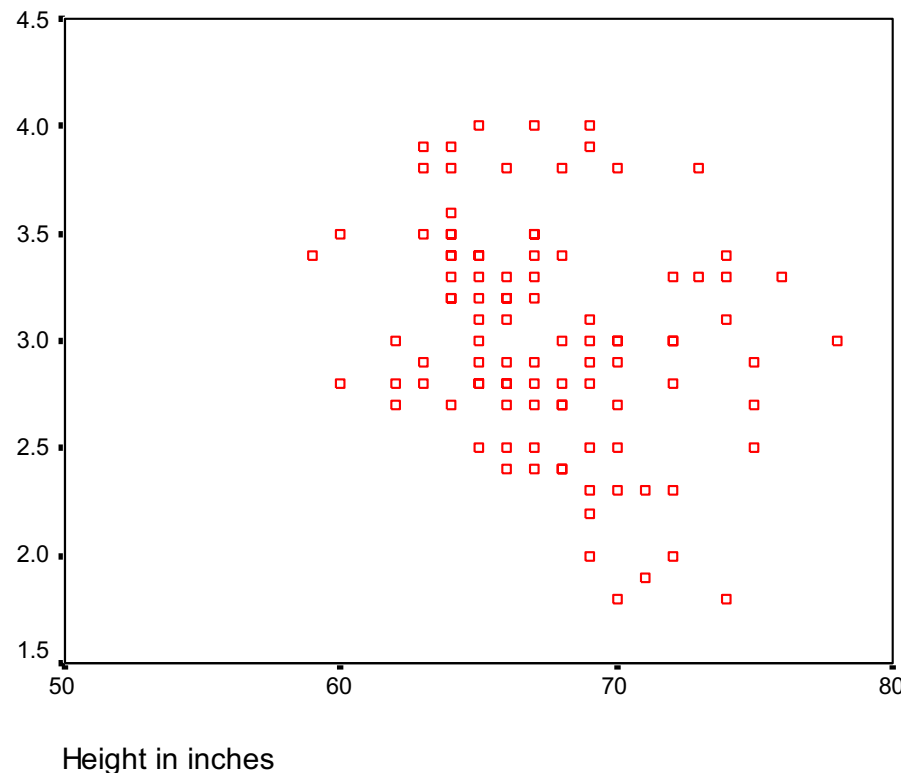
- Leverage plots: You plot the leverage against the observation number (first observation in your data file = #1, second = #2, etc.)
- Leverage for observation j is defined as

$$h_{jj} = \frac{(\mathbf{x}_j - \bar{\mathbf{x}})^2}{\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^2}$$

- In effect, you measure the distance of an observation to its mean in relation to the total distance of the X 's

Outliers and Leverage

- Remember the GPA and Height Example
- Are there any obvious outliers/leverage points?



Outliers and Leverage

- Remember the GPA and Height Example
- Are there any obvious outliers/leverage points?

Not Really!



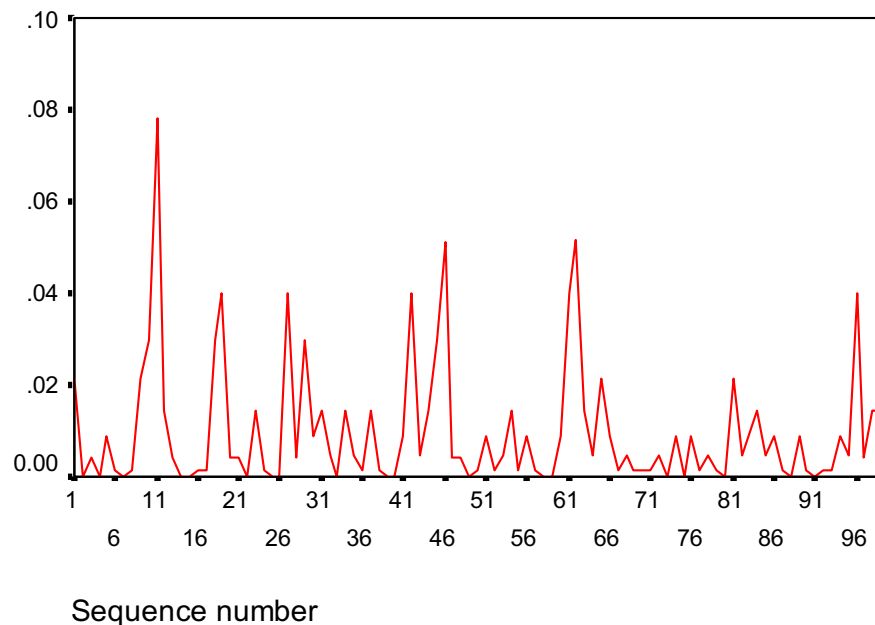
Outliers and Leverage

- The leverage plot should show nothing really dramatic

**This is just normal
Scatter. Takes
Experience to read**

Leverage Values vs Obs. Number

Y=GPA, X=Height



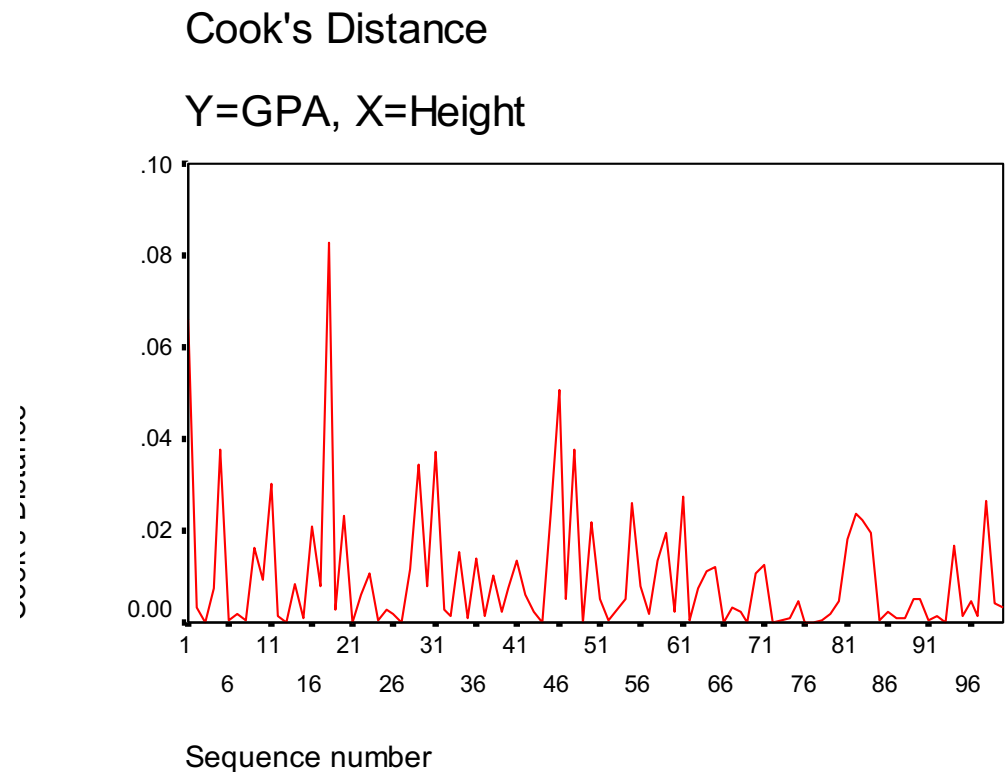
Outliers and Leverage

- The Cook's Distance for an observation is defined as follows
- Compute the fitted values with all the data
- Compute the fitted values with observation j deleted
- Compute the sum of the squared differences
- Measures how much the line changes when an observation is deleted

Outliers and Leverage

- The Cook's Distance plot should show nothing really dramatic

**This is just normal
Scatter. Takes
Experience to read**



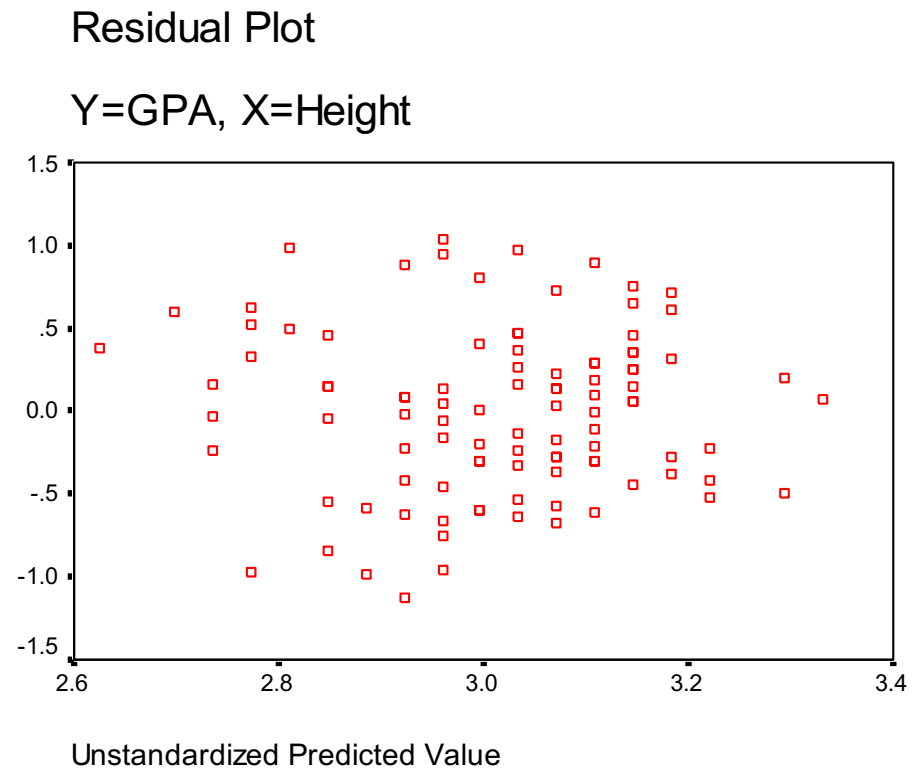
Outliers and Leverage

- The residual plot is a plot of the residuals (on the y-axis) against the predicted values (on the x-axis)
- You should look for values which seem quite extreme

Outliers and Leverage

- The residual plot should show nothing really dramatic

**This is just normal
Scatter. No massive
Outliers. Takes
Experience to read**



Outliers and Leverage

- A much more difficult example occurs with the stenotic kids

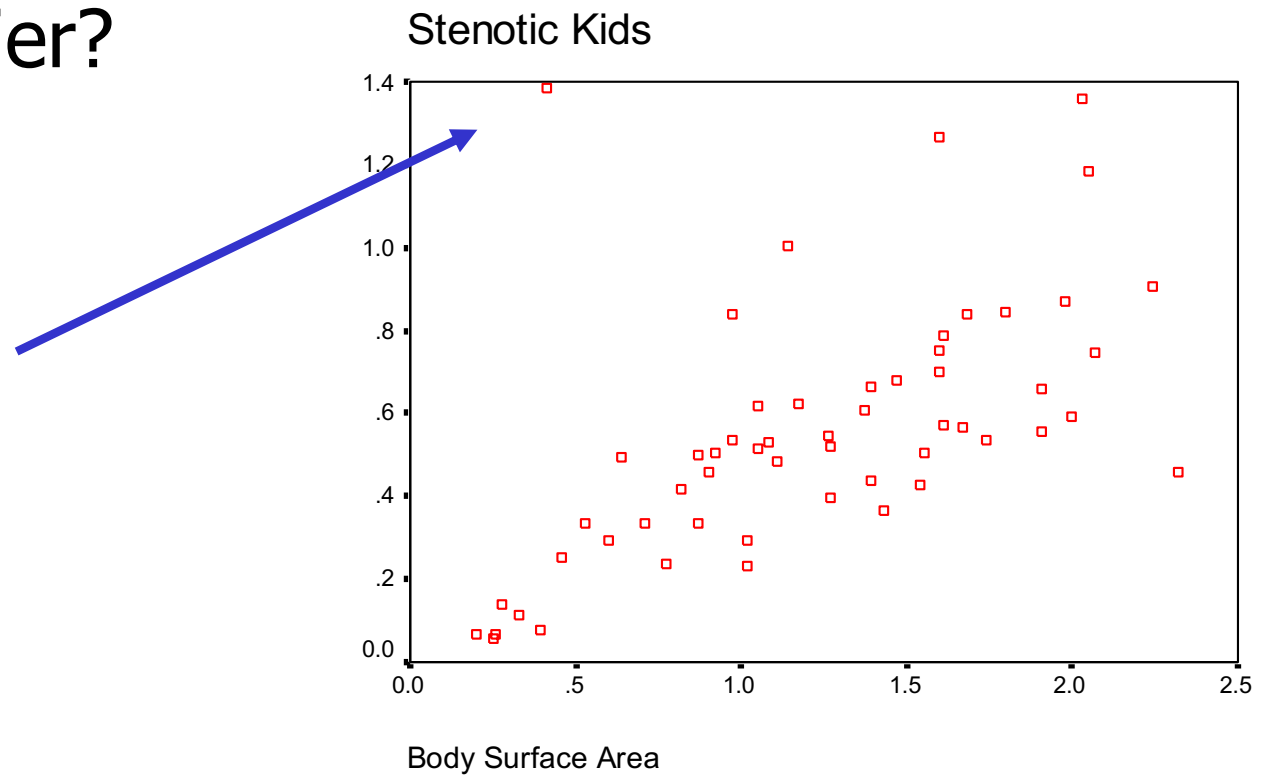
Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	.167	.079		2.099	.041	.007	.326
Body Surface Area	.319	.059	.591	5.390	.000	.200	.438

a. Dependent Variable: Log(1+Aortic Valve Area)

Outliers and Leverage

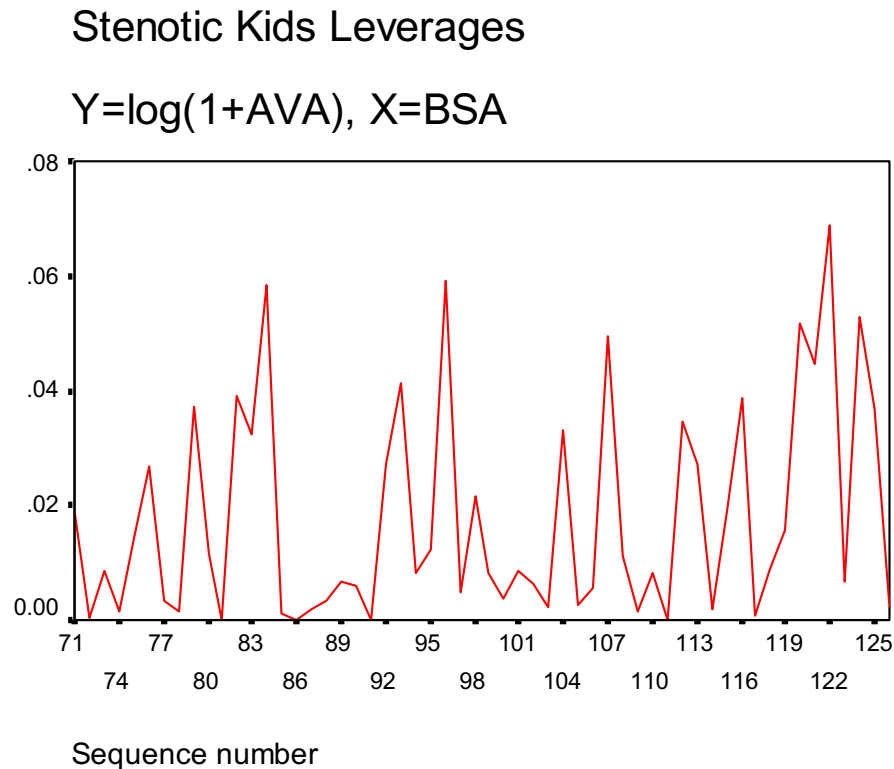
- A much more difficult example occurs with the stenotic kids
- Note: outlier?



Outliers and Leverage

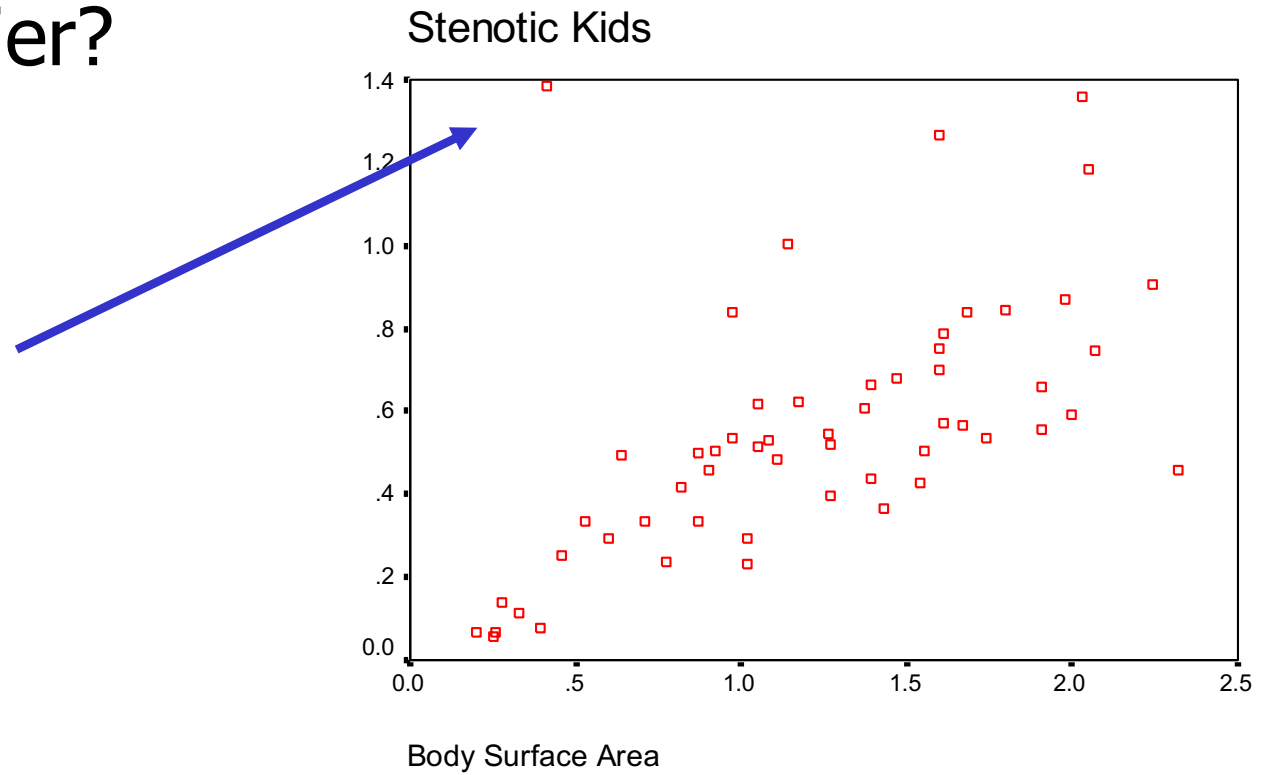
This makes sense, since the data show no unusual X-values

Scatterplot comes next



Outliers and Leverage

- A much more difficult example occurs with the stenotic kids
- Note: outlier?



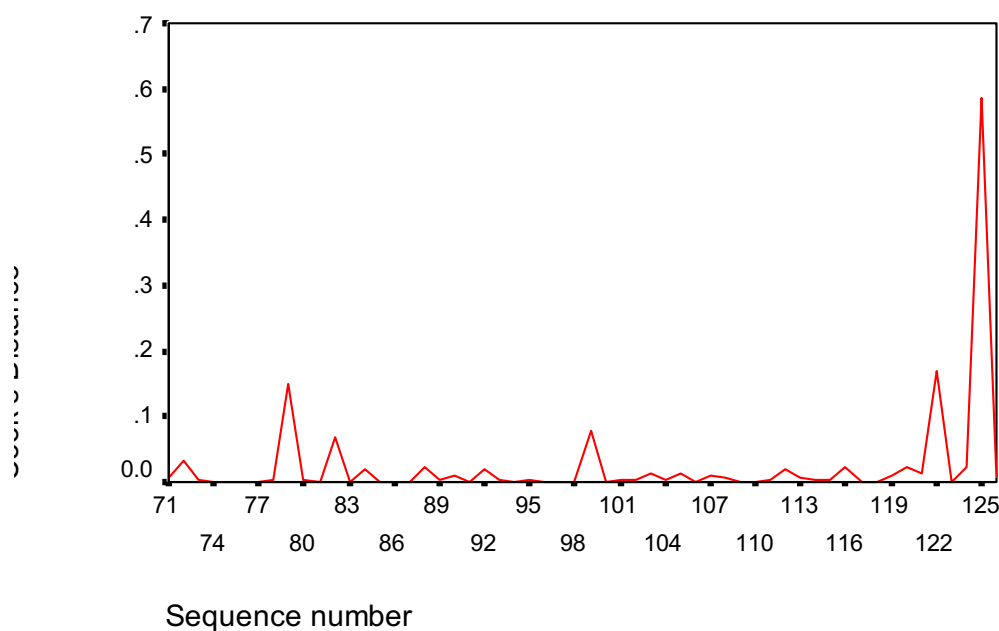
Outliers and Leverage

- Wow!

This is a case that there is a noticeable outlier, but **not too high leverage**

Cook's Distances, Stenotic Kids

$Y = \log(1 + \text{AVA})$, $X = \text{BSA}$

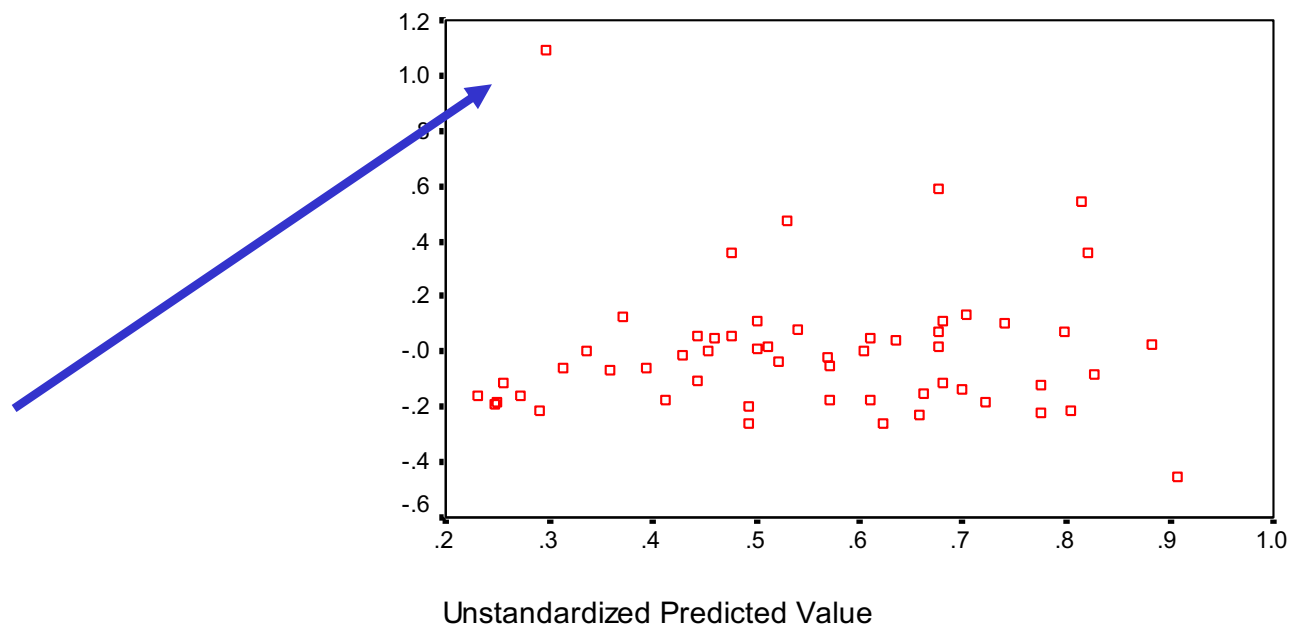


Outliers and Leverage

- Wow!

Residual plot, Stenotic Kids

$Y = \log(1 + \text{AVA})$, $X = \text{BSA}$



Outliers and Leverage: Low Leverage Outliers

Coefficients: All Stenotic Kids

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	.167	.079		2.099	.041	.007	.326
Body Surface Area	.319	.059	.591	5.390	.000	.200	.438

a. Dependent Variable: Log(1+Aortic Valve Area)

Stenotic Kids, Outlier Removed

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	8.207E-02	.065		1.260	.213	-.049	.213
Body Surface Area	.372	.048	.727	7.715	.000	.275	.468

a. Dependent Variable: Log(1 + Aortic Valve Area)

Remember: Outliers Inflate Variance!

ANOVA ^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.801	1	1.801	29.051	.000 ^a
	Residual	3.348	54	6.200E-02		
	Total	5.149	55			

a. Predictors: (Constant), Body Surface Area

b. Dependent Variable: Log(1+Aortic Valve Area)

ANOVA ^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.352	1	2.352	59.526	.000 ^a
	Residual	2.094	53	3.951E-02		
	Total	4.446	54			

a. Predictors: (Constant), Body Surface Area

b. Dependent Variable: Log(1 + Aortic Valve Area)

Outliers and Leverage

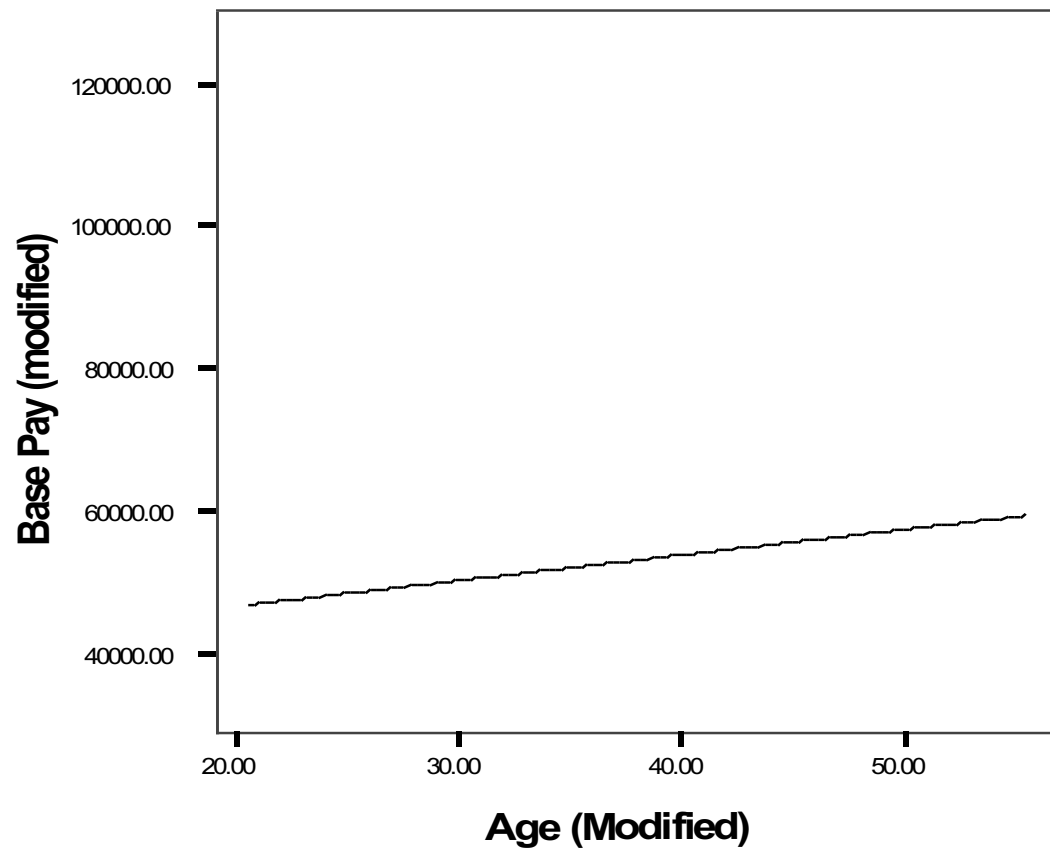
- The effect of a high leverage outlier is often to inflate your estimate of σ_{ε}^2
- With the outlier, the MSE (mean squared residual) = 0.0620
- Without the outlier, the MSE (mean squared residual) is = 0.0395
- So, a single outlier in 56 observations increases your estimate of σ_{ε}^2 by over 50%!
- This becomes important later!

Base Pay and Age in Construction

Construction Example

No outliers

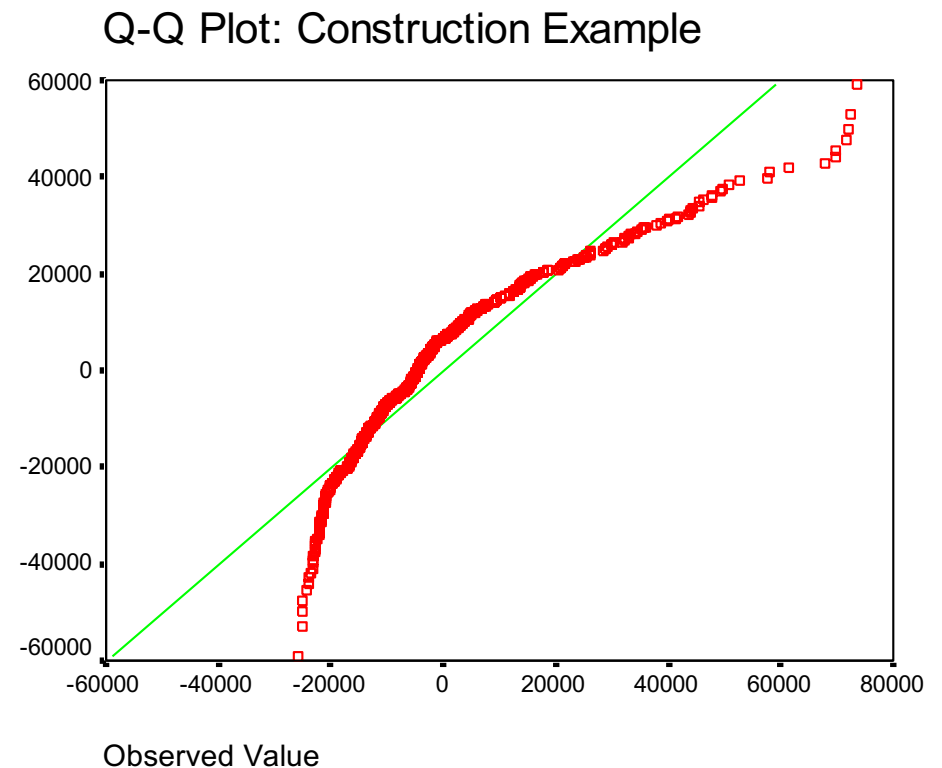
Not a strong trend, but in the expected direction



Base Pay and Age in Construction

**Not even close to
normally
distributed**

**Cries out for a
transformation**

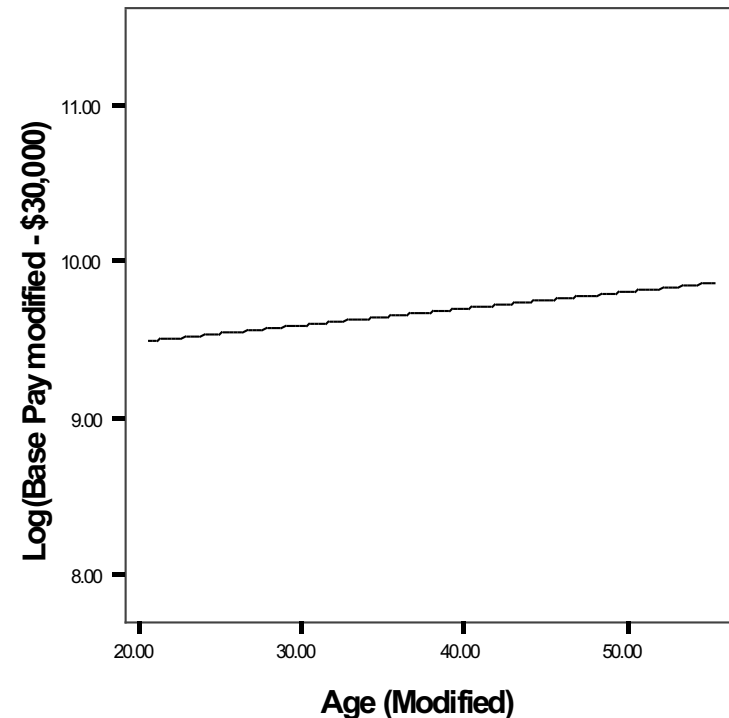


Log(Base Pay) and Age in Construction

Expected trend, but weak

Odd data structure: salaries were rounded in clumps of \$5,000

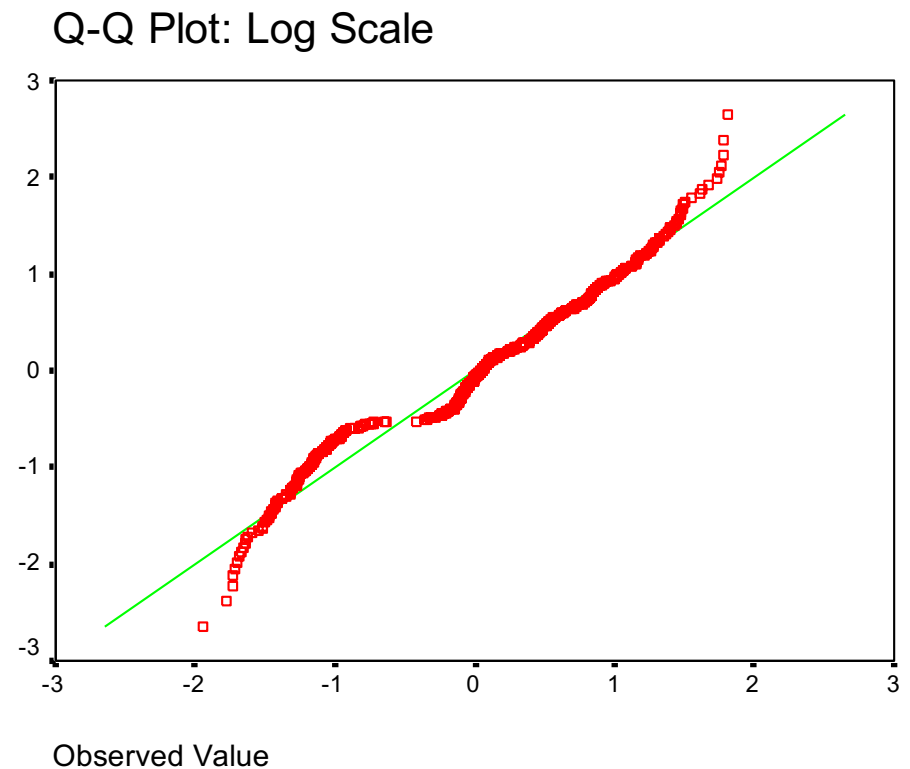
Construction Example: Log Scale



Log(Base Pay) and Age in Construction

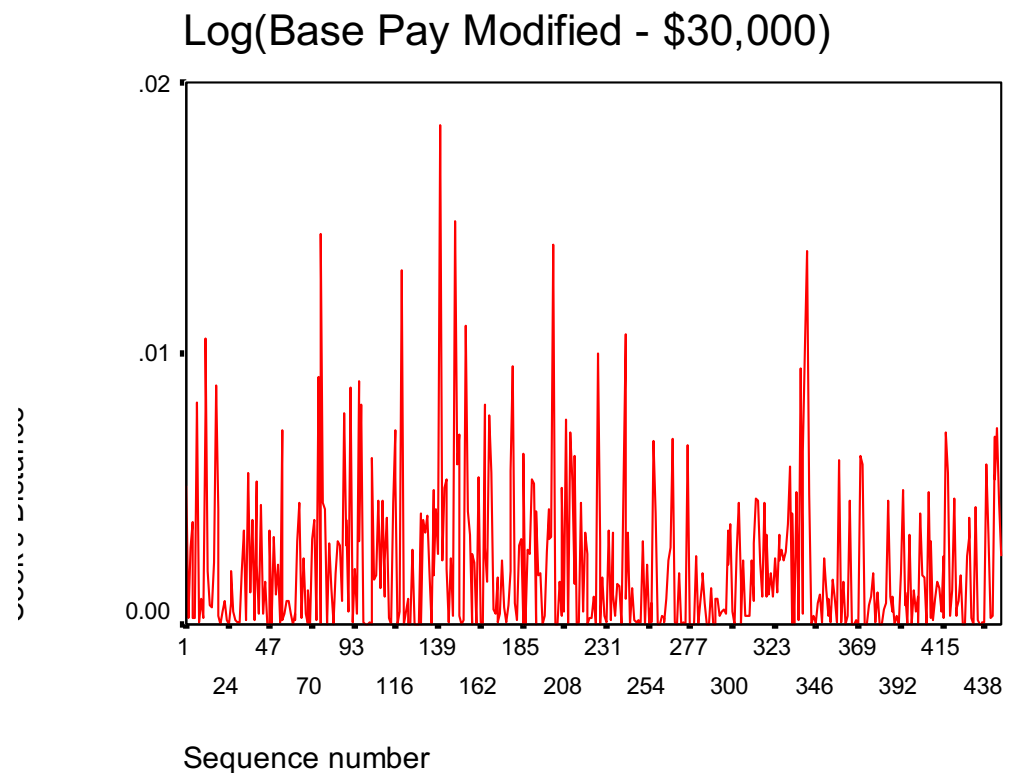
Much better residual plot

Good time to remember why we want data to be normally distributed



Log(Base Pay) and Age in Construction

No real massive influential points, according to Cook's distances



Log(Base Pay) and Age in Construction

Note the statistically significant effect: do we have 99% confidence?

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.057	1	5.057	6.459	.011 ^a
	Residual	348.368	445	.783		
	Total	353.425	446			

a. Predictors: (Constant), Age (Modified)

b. Dependent Variable: Log(Base Pay modified - \$30,000)

Log(Base Pay-\$30,000) and Age in Construction

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	9.277	.164		56.689	.000
Age (Modified)	1.073E-02	.004	.120	2.542	.011

a. Dependent Variable: Log(Base Pay modified - \$30,000)