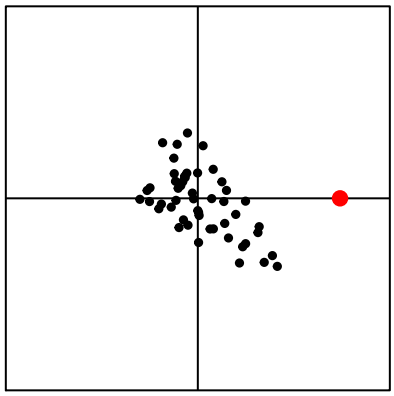
STAT 311: Homework 2
Due: Jul 8, in class

Name:

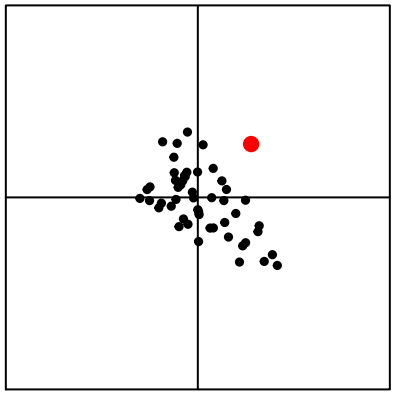
The material covered includes chapter 3 and 4 as well as the last part of chapter 2 in Mind on Statistics as well as Lectures 3, 4 and 5. In general, rounding to 2 digits is sufficient.

1 Classifying Outliers

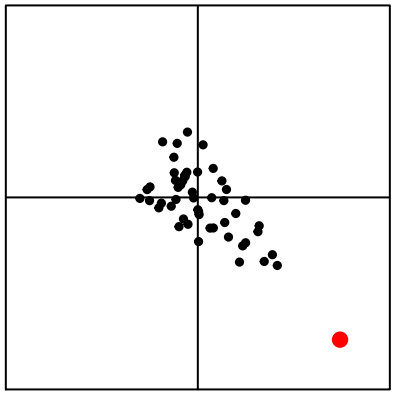
1. Draw a point which is an outlier in the X direction, but not the Y



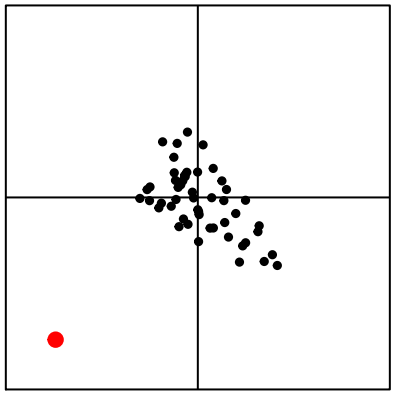
2. Draw a point which is an outlier in the joint distribution, but not any of the marginals



3. Draw a point which has large leverage, but small influence



4. Draw a point with large leverage, and large influence



2 Regression Identities

Recall the following formulas from the lecture slides for the regression coefficients.

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$\hat{b} = \frac{cov(x,y)}{var(x)} = r\frac{s_y}{s_x} \tag{1}$$

where r is the correlation coefficient.

1. If $\hat{b} = .5$ and $cov(x,y) = 2$, what is s_x?

$$\hat{b} = cov(x,y)/var(x) \Rightarrow .5 = 2/var(x)$$

$$var(x) = 4 \Rightarrow s_x = 2$$

2. If $s_y = 4$, $s_x = 2$ and $\hat{b} = 1$, what is r?

$$\hat{b} = r\frac{s_y}{s_x} \Rightarrow 1 = r\frac{4}{2}$$

$$r = \frac{1}{2}$$

3. If $s_y = 4$, $s_x = 2$ what is the largest possible value of $\hat{b}$? What is the smallest possible value of $\hat{b}$?

$$\hat{b} = r\frac{s_y}{s_x} \Rightarrow \hat{b} = r\frac{4}{2}$$

since $-1 \geq r \leq 1$, then the largest value of $\hat{b}$ is 2 and the smallest value is -2

4. If $s_y = 4$, $s_x = 2$ what is the largest possible value of $\hat{b}$? What is the smallest possible value of $\hat{b}$? Repeat

5. If $\hat{a} = 5$, $\bar{y} = 6$, $\bar{x} = 2$, $s_y = 1$ and $s_x = 1$, what is the value of r?

$$\hat{a} = \bar{Y} - \hat{b}\bar{X} \Rightarrow 5 = 6 - \hat{b}2 \Rightarrow \hat{b} = 1/2$$

$$\hat{b} = r\frac{s_y}{s_x} \Rightarrow 1/2 = r\frac{1}{1} \Rightarrow r = 1/2$$

6. If $\hat{a} = 5$, $\bar{y} = 6$, $\bar{x} = 2$, what would $\hat{y}_i$ be if $x_i = 4$? What would the residual be if $y_i = 7$?

$$\hat{a} = \bar{Y} - \hat{b}\bar{X} \Rightarrow 5 = 6 - \hat{b}2 \Rightarrow \hat{b} = 1/2$$

$$\hat{y} = 5 + \frac{1}{2}X$$

$$\hat{y} = 5 + \frac{1}{2}4 = 7$$

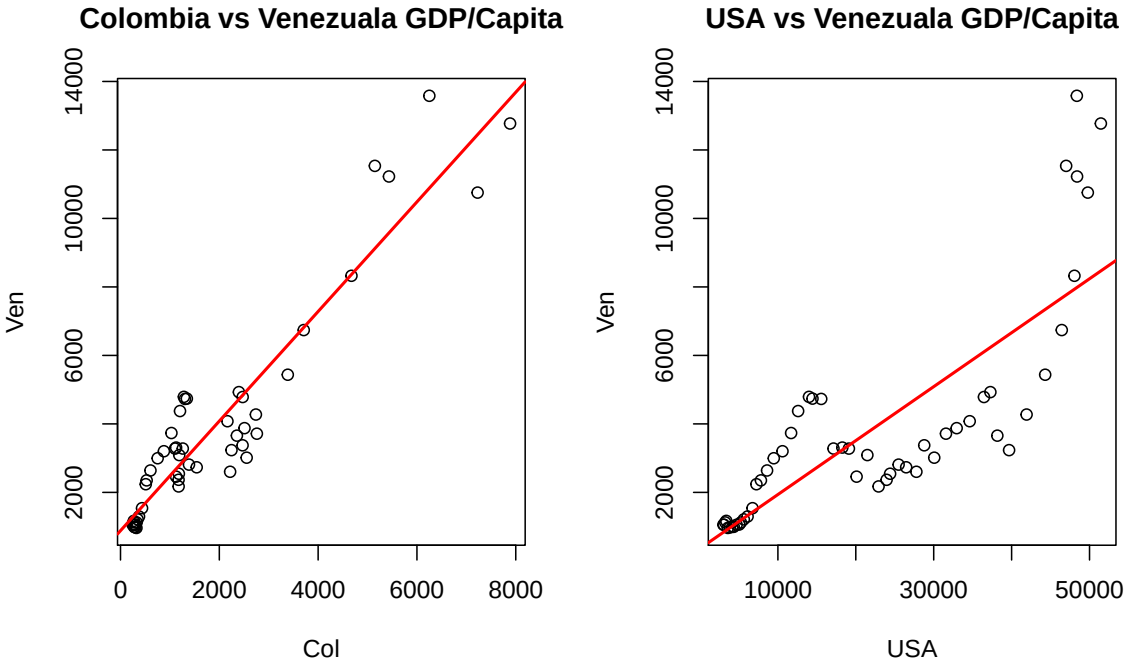
The residual is $y_i - \hat{y}$, so it is 0.

3 Decomposing the sum of squared errors

Recall the decomposition of the Sum of Squares for Y from the lecture slides-

$$\underbrace{\sum_i (y_i - \bar{y})^2}_{SS_{total}} = \underbrace{\sum_i (\hat{y}_i - \bar{y})^2}_{SS_{reg}} + \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{SS_{error}} \tag{2}$$

Suppose we are trying to predict Venezuala's GDP per capita. For our explanatory variable, we have either US GDP per capita or Colombia's GDP per capita. We have data across 53 years.



For the Venezualan data, we know that-

$$SS_{total} = \sum_i (y_i - \bar{y})^2 = 499,062,880$$

and for the regression of "Venezuala $\sim$ Colombia", we have

$$SS_{error} = \sum_i (y_i - \hat{y}_i)^2 = 55,968,556$$

and for the regression of "Venezuala $\sim$ USA," we have

$$SS_{regression} = \sum_i (y_i - \hat{y}_i)^2 = 323,320,383$$

(a) What is the standard deviation of the Venezuala observations?

$$s_x = \sqrt{\frac{1}{N-1} SS_{total}} = \sqrt{\frac{1}{N-1} 499,062,880} = 3097.96$$

(b) What is $SS_{regression}$ for the regression using Colombia as the explanatory variable?

$$SS_{total} = SS_{regression} + SS_{error} \Rightarrow 499062880 = SS_{regression} + 55,968,556$$

$$SS_{regression} = 443,094,324$$

(c) What is the correlation between Colombia and Venezuala GDP/capita?

$$r^2 = \frac{SS_{regression}}{SS_{total}} = \frac{443094324}{499062880}$$

$$r = .94$$

(d) If the slope of the regression is 1.6, what is $s_{Colombia}$?

$$\hat{b} = r\frac{s_y}{s_x} \Rightarrow 1.6 = .94\frac{3097.96}{s_x}$$

$$s_x = 1819.5$$

(e) What is SS_{error} for the regression using USA as the explanatory variable?

$$SS_{total} = SS_{error} + SS_{regression} \Rightarrow 499062880 = SS_{error} + 323,320,383$$

$$SS_{error} = 175,742,497$$

(f) What is the correlation between USA and Venezuala GDP/capita?

$$r^2 = \frac{SS_{regression}}{SS_{total}} = \frac{323320383}{499062880}$$

$$r = .80$$

(g) If the slope of the regression is .16, what is s_{USA}?

$$\hat{b} = r\frac{s_y}{s_x} \Rightarrow .16 = .80\frac{3097.96}{s_x}$$

$$s_x = 15489.8$$

(h) What would be a better predictor of Venezuala's GDP per capita, the USA GDP/capita or Colombia's GDP/capita?

Since the r^2 value for Columbia is higher, it would probably be a better predictor than the USA.

4 Discrete Data

The TV show American Ninja Warrior travels to several cities around the US and challenges contestants in each city to complete an obstacle course. Competitors who complete the course, or those who go further than others in a shorter amount of time, are invited to continue on to the next round. So far, in the 2016 season, American Ninja Warrior has visited 5 cities. The breakdown of many individuals have completed the course by city is shown below. The data is from Wikipedia.

	Atlanta	Indianapolis	LA	OKC	Philadelphia	Total
Complete	27	25	17	15	9	93
Did Not Complete	3	5	13	15	21	57
Total	30	30	30	30	30	150

(a) What proportion of all the contestants are from Atlanta?

30 / 150 = .2

(b) What proportion of all the contestants are from Atlanta or Indianapolis?

(30 + 30) / 150 = .4

(c) What proportion of all the contestants completed the course?

93 / 150 = .62

(d) What proportion of all the contestants completed the course and are from Atlanta?

27 / 150 = .18

(e) What proportion of the Philadelphia contestants completed the course?

9 / 30 = .3

(f) What proportion of the contestants who completed the course are from Philadelphia?

9 / 93 = .10