Homework - Week 11

Your name here

Questions marked with "(OS3: X.X)" are from the textbook with "X.X" as the exercise number. The answers to the odd questions (odd by book numbering that is) will be in the back of the book.

1. (OS3: 7.29) The following regression output is for predicting annual murders per million from percentage living in poverty in a random sample of 20 metropolitan areas.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-29.901	7.789	-3.839	0.001
poverty%	2.559	0.390	6.562	0.000
$s = 5.512$	$R^2 = 70.52\%$		$R^2_{adj} = 68.89\%$	

   a. Write out the linear model.

   $\hat{y} = -29.901 + 2.559x$ or $\widehat{murders} = -29.901 + 2.559(poverty\%)$

   b. Interpret the intercept.

   The intercept (-29.901) is the predicted annual murders per million when the poverty rate is 0. However, since the intercept is negative, this interpretation is meaningless, as is often the case.

   c. Interpret the slope.

   The slope (2.559) is the predicted amount the annual murders per million will increase for each unit increase (+1) of the poverty percentage.

   d. Interpret R^2.

   R^2 **(70.52) is the approximate percentage of the variation of the response variable (annual murders per million) that can be explained by the association with the predictor variable (poverty %).**

   e. Calculate the correlation coefficient.

   $r = \sqrt{R^2} = \sqrt{0.7052} = 0.8398$

2. (OS3: 7.41 a-b) Exercise 7.29 presents regression output from a model for predicting annual murders per million from percentage living in poverty based on a random sample of 20 metropolitan areas. The model output is also provided below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-29.901	7.789	-3.839	0.001
poverty%	2.559	0.390	6.562	0.000
$s = 5.512$	$R^2 = 70.52\%$		$R^2_{adj} = 68.89\%$	

   a. What are the hypotheses for evaluating whether poverty percentage is a significant predictor of murder rate?

   $H_0 : \beta_1 = 0$

   $H_a : \beta_1 \neq 0$

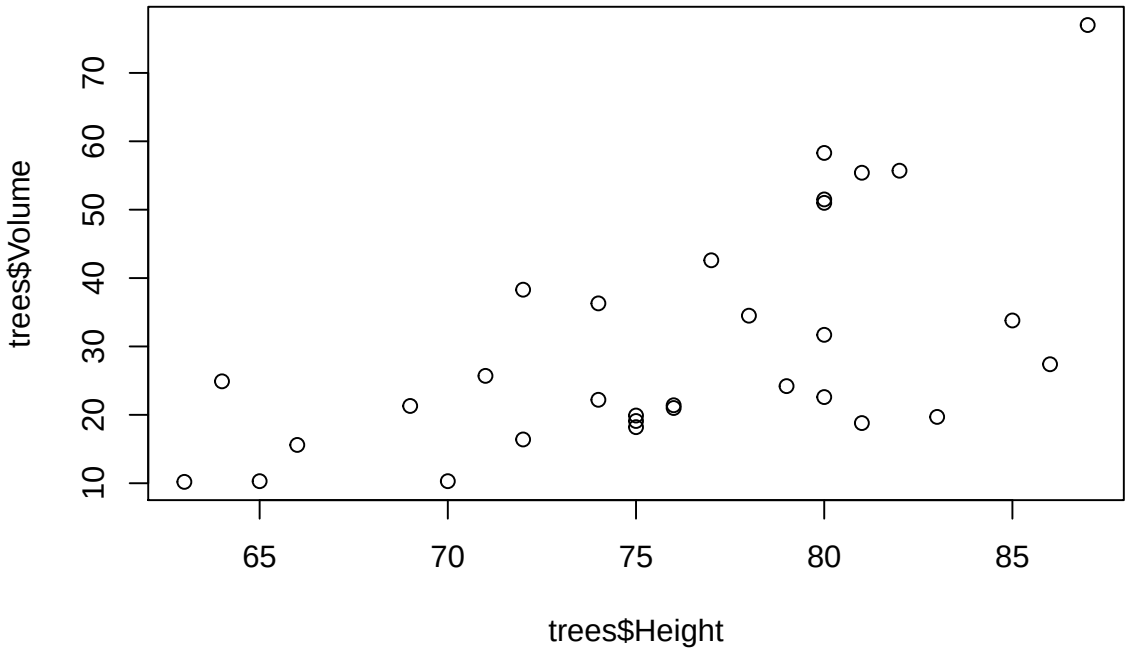
   b. State the conclusion of the hypothesis test from part (a) in context of the data.

   p < 0.05. Reject the null hypothesis. There is evidence that poverty percentage is a significant predictor of murder rate.

3. Consider the built-in dataset **trees**. It contains heights (ft), girth (in) and volumes (ft^3) of a sample of black cherry trees. Suppose we wish to be able to predict tree volume given a tree's height.

   a. Create a scatterplot of heights and volumes of trees.

```
plot(trees$Height, trees$Volume)
```



   b. What is the correlation between height and volume of trees? Is the correlation statistically significant? How would you generally describe the strength of correlation?

```
cor.test(trees$Height, trees$Volume)
```

```
##
##  Pearson's product-moment correlation
##
## data:  trees$Height and trees$Volume
## t = 4.0205, df = 29, p-value = 0.0003784
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3095235 0.7859756
## sample estimates:
##       cor
## 0.5982497
```

   $p = 0.0003784 < 0.05$. The correlation is statisically significant. A correlation of approximately 0.6 indicates a moderate correlation.

   c. Create a regression model of the relationship of heights and volumes of trees, with height as the predictor variable. What is the regression line equation? Is the model statistically significant?

```
summary(lm(Volume ~ Height, data=trees))
```

```
##
## Call:
## lm(formula = Volume ~ Height, data = trees)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -21.274  -9.894  -2.894  12.068  29.852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -87.1236    29.2731  -2.976 0.005835 **
## Height        1.5433     0.3839   4.021 0.000378 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 29 degrees of freedom
## Multiple R-squared:  0.3579, Adjusted R-squared:  0.3358
## F-statistic: 16.16 on 1 and 29 DF,  p-value: 0.0003784
```

$\hat{y} = -87.1236 + 1.5433x$

p = 0.0003784 < 0.05. The model is statisically significant.

d. What is the predicted volume of a tree that is 84 feet tall? Is this an appropriate prediction?

```
-87.1236 + 1.5433 * 84
```

```
## [1] 42.5136
```

```
range(trees$Height)
```

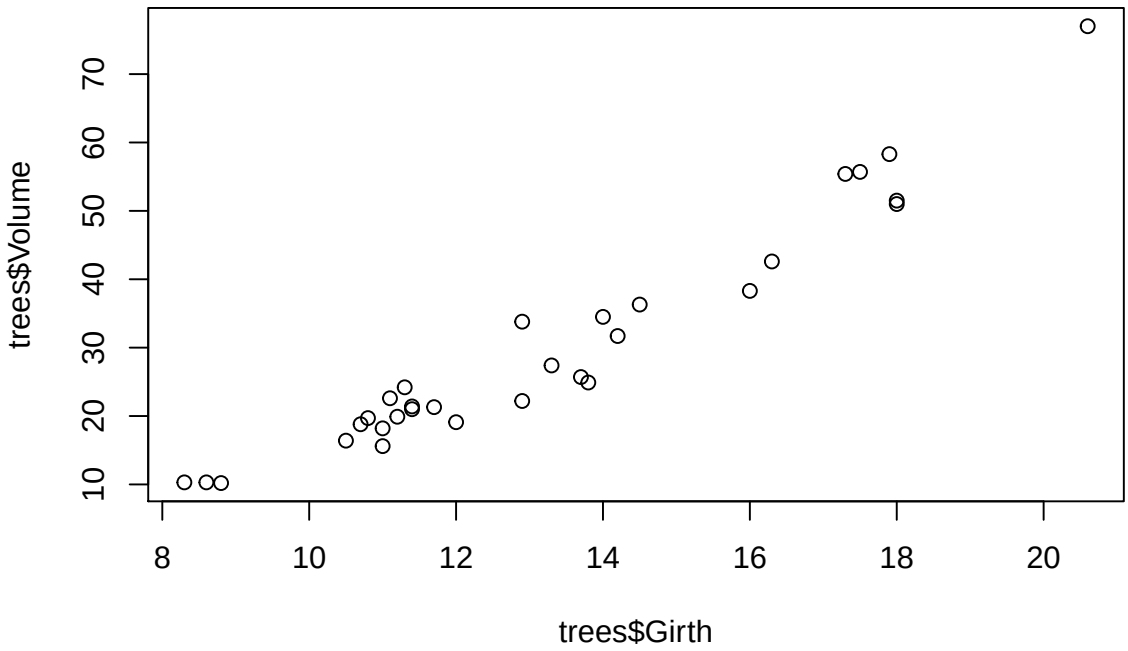
```
## [1] 63 87
```

The predicted volume for a tree 84 ft tall is 42.5136 cubic ft. The range of heights used to create the model is 63 to 87 ft, so using 84 ft for a prediction is appropriate.

4. Suppose now we wish to be able to predict tree volume given a tree's girth.

a. Create a scatterplot of girths and volumes of trees.

```
plot(trees$Girth, trees$Volume)
```



b. What is the correlation between girth and volume of trees? Is the correlation statistically significant? How would you generally describe the strength of correlation?

```
cor.test(trees$Girth, trees$Volume)
```

```
##
##  Pearson's product-moment correlation
##
## data:  trees$Girth and trees$Volume
## t = 20.478, df = 29, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9322519 0.9841887
## sample estimates:
##       cor
## 0.9671194
```

$p < 0.0001 < 0.05$. The correlation is statisically significant. A correlation of approximately 0.97 indicates a strong correlation.

c. Create a regression model of the relationship of girths and volumes of trees, with girth as the predictor variable. What is the regression line equation? Is the model statistically significant?

```
summary(lm(Volume ~ Girth, data=trees))
```

```
##
## Call:
## lm(formula = Volume ~ Girth, data = trees)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.065 -3.107  0.152  3.495  9.587
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
## Girth         5.0659     0.2474   20.48  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.252 on 29 degrees of freedom
## Multiple R-squared:  0.9353, Adjusted R-squared:  0.9331
## F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

$\hat{y} = -36.9435 + 5.0659x$

$p < 0.0001 < 0.05$. The model is statisically significant.

d. What is the predicted volume of a tree that has a girth of 23 inches? Is this an appropriate prediction?

```
-36.9435 + 5.0659 * 23
```

```
## [1] 79.5722
```

```
range(trees$Girth)
```

```
## [1]  8.3 20.6
```

The predicted volume for a tree with a 23 in girth is 79.5722 cubic ft. The range of girths used to create the model is 8.3 to 20.6 in, so using 23 in for a prediction is not appropriate.

5. Suppose you have access to both the height and girth of a tree. Based on the models developed above, which measurement should you use to make a prediction of volume? Why?

   R^2 for the height model is 0.3579 and 0.9353 for the girth model. The greater coefficient of determination for the girth model indicates that predictions from this model are more accurate.