Week 2, STAT5120, Spring 2018

1. A college admissions director selected 120 students at random to see if ACT score (X) could be used to predict GPA (Y) at the end of their first year. Open the data set CH01PR19 from ALSM.

(a) Use R to construct a **99% CI** for the slope coefficient β_1. Interpret this CI in a complete sentence, "We can be about 99% confident that..." Does the interval include 0? What is significant about the CI containing 0? The 99% CI for β_1 is (0.005385614, 0.07226864). This interval does NOT include zero as can be seen at left. The significance of the CI containing zero (if it had) would have implied that… "We can be 99% confident that for every 1-unit change in ACT scores, there will be a corresponding multiplicative change in GPA scores ranging from 0.005 to 0.072"

(b) Run a hypothesis test (with R) to see if a linear association exists between ACT score and first year GPA. State the p-value and your conclusion. We will test if a linear relationship exists between ACT Score and GPA scores by not only inspecting a scatter plot but also be proposing a hypothesis test to see if the slope is NOT equal to zero. Our null hypothesis is that the slope coefficient $\beta_1 = 0$ and the alternative hypothesis it is different from zero and thus contributing to predicting the value of our Y variable. We call a t-test in R along with a p-value discover that our t-statistic is 3.039777 and our p-value is 0.002916604, indicating the slope *is* different from 0. We reject the null hypothesis and thereby accept the alternative hypothesis that there is a linear association between ACT score and GPA.

(c) Use R to construct a **98% CI** for the intercept β_0. Interpret this CI in a complete sentence. Why is the intercept of interest? That is, what does β_0 tell you about first year GPA and ACT? The y-intercept is of interest in some cases depending on whether the model has data which include x values equal to 0. Our class text states with respect to the Toluca Company example that, "As noted earlier, the scope of the model for the Toluca Company does not extend to lot sizes of X =0. Hence the regression parameter β_0 may not have intrinsic meaning here."[1] In this case for ACT and GPA scores, no ACT scores are seen to have values = 0. We thus cannot conclude that the intercept tells us something intrincially true about GPA other than the fact that a value of 2.11 will be added to a predicted value for GPA scores given ACT scores. The 98% CI for β_0 is (1.357262, 2.870836). We can say that "We are 98% confident that the y-intercept of the regression model is between 1.36 and 2.87."

(d) Use R to construct a **95%** interval estimate of the *mean* freshman GPA for students whose ACT score is 29. Interpret this interval in a complete sentence. Our regression model is hypothesized to $\hat{Y} = 2.11 + 0.03883x$. The GPA score for a student who scored an ACT score of 29 would by 2.11 + 0.03883(29) = 3.24. To add some context to this prediction, we could be 95% confident that the mean GPA score would be between 2.835 and 3.645.

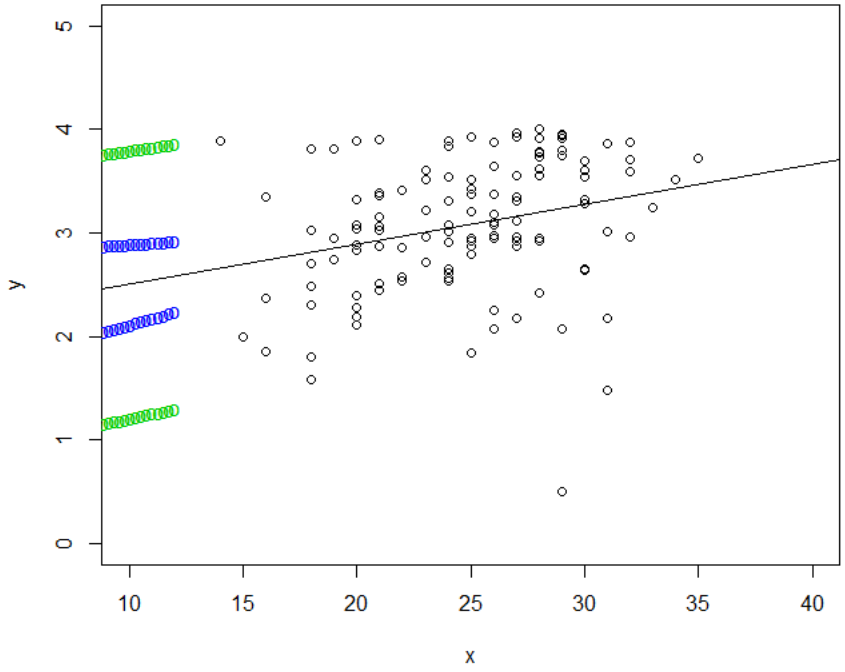
(e) Billy Billingsley scored a 29 on the ACT. Predict his freshman GPA with a 95% CI, and be sure to interpret the interval in a complete sentence. Use R of course. This predicted y-value would be 3.24 and we could be 95% confident that this GPA score would be between 1.941 and 4.539.

(f) Is the prediction interval you obtained in part (e) wider or skinnier than the interval you obtained in part (d)? Should it be? Why? The confidence interval in part (e) is wider than in part (d) where we calculated the mean response in y. This is because the standard error calculated for a predicted individual value, as done in part (e), contains an extra "1" under the square root symbol. Chatterjee comments that this "1" is included there

[1] Kutner, Michael H., Christopher J. Nachtsheim, John Neter, and William Li, *Applied Linear Statistical Models*, 5th ed., (Chennai: McGraw Hill, 2013), 50.

because "There is greater uncertainty (variability) in predicting one observation (the next observation) than in estimating the mean response when X = x_0. The averaging that is implied in the mean response reduces the variability and uncertainty associated with the estimate."[2]

(g) Add 95% confidence interval bands for the mean response to your plot. What is the CI for mean response when X_h = 29? Also add Working-Hotelling bands for the regression line to your plot. What do these bands mean? We could be 95% confident that the mean GPA score would be between 2.835 and 3.645.



2. Why is the t-test more versatile than the F-test when testing hypotheses about β_1? Also, why is the F test a one-tailed test even though the alternative hypothesis is $\beta_1 \neq 0$? T-tests can be used to test one-tailed hypotheses in a specific direction. The F-test essentially reduces to a one-sided test but it can be in either direction; the t-test can go in a specific direction.

3. Prove the Bonferroni inequality: for any events $E_1, E_2 \ldots E_n$,

$$P\left(\bigcap_{i=1}^{n} E_i\right) \geq \sum_{i=1}^{n} P(E_i) - n + 1.$$

We have to show that $P(E_1, E_2 \ldots E_n) \geq P(E_1) + P(E_2) + P(E_n) - n + 1$

We do this by denoting the set $(E_1, E_2 \ldots E_n)$ by B. The right side of the equation then becomes

$P(BE_n) \geq P(B) + P(E_n) - n + 1$
$= P(E_1, E_2 \ldots E_{n-1}) + P(E_n) - 1$
$\geq P(E_1) + P(E_2) + P(E_{n-1}) + \ldots P(E_{n-2}) + P(E_n) - 1$
$= P(E_1) + P(E_2) + P(E_{n-1}) + P(E_{n-2}) - (n-1)$

[2] Chatterjee, Samprit, and Ali S. Hadi, *Regression Analysis by Example*, 5th ed., (Hoboken: John Wiley & Sons, 2012), 42.

4. Use the Bonferroni simultaneous inference approach to make CIs for the slope and intercept of the regression line for the Toluca Company data set CH01TA01 from ALSM. Use a family error rate of 5% (or family confidence level of 95%).

95% Cis for the slope are (2.852435, 4.287969) while 95% CI's for the y-intercept are (8.213711, 116.518)

5. Problem 2.8 from *Applied Linear Statistical Models*, by Kutner et. al: Refer to the Toluca Company (CH01TA01) example and data. A consultant advises that an increase in the amount of one lot size unit requires an increase of about **3** in the expected number of work hours for the production item.

(a) Test to see if the increase in expected number of work hours equals or differs from the consultant's opinion. The consultant suggests that an increase in 3 hours of work. We test this in R and discover our t-statistic for this test is 1.64 with a p-value of 0.943. We cannot reject the null which in this case would be that a slope of 3 would be statistically linearly significant. The consultant's opinion is therefore probably not well-founded.

(b) Calculate the power of your test in the previous part if in fact the consultant's recommendation is 1/2 hour too low. Assume the standard deviation of the slope coefficient to be 0.35. In this case, the hypothesized value would be 3 hours as originally thought plue another 1/5 hours, bringing us to a value of 3.5. This is actually the slope value as predicted in the regression equation, specifically, 3.57. Power can be calculated to be 0.9983818

(c) Why is the value of the F-statistic 105.88 given in the R output irrelevant in part (a)? This is because the F-statistic does not specify the direction of a one-tailed test as is the case here. In part a we were asking if an increase of 3 labor hours was necessary. The F-test will not specify the direction of such a test as will the t-test.

(d) Repeat the power calculation when the amount the standard is actually exceeded is 1 hour and 1.5 hours. Do these power calculations seem correct? Why? This power calculation does not seem correct since it is returning a power of only .241 while the difference is greater in this hypothesis than it was in part b where the difference was not so great. As distance increases in the slope vs. the hypothesis, power should *increase* but here it does not.

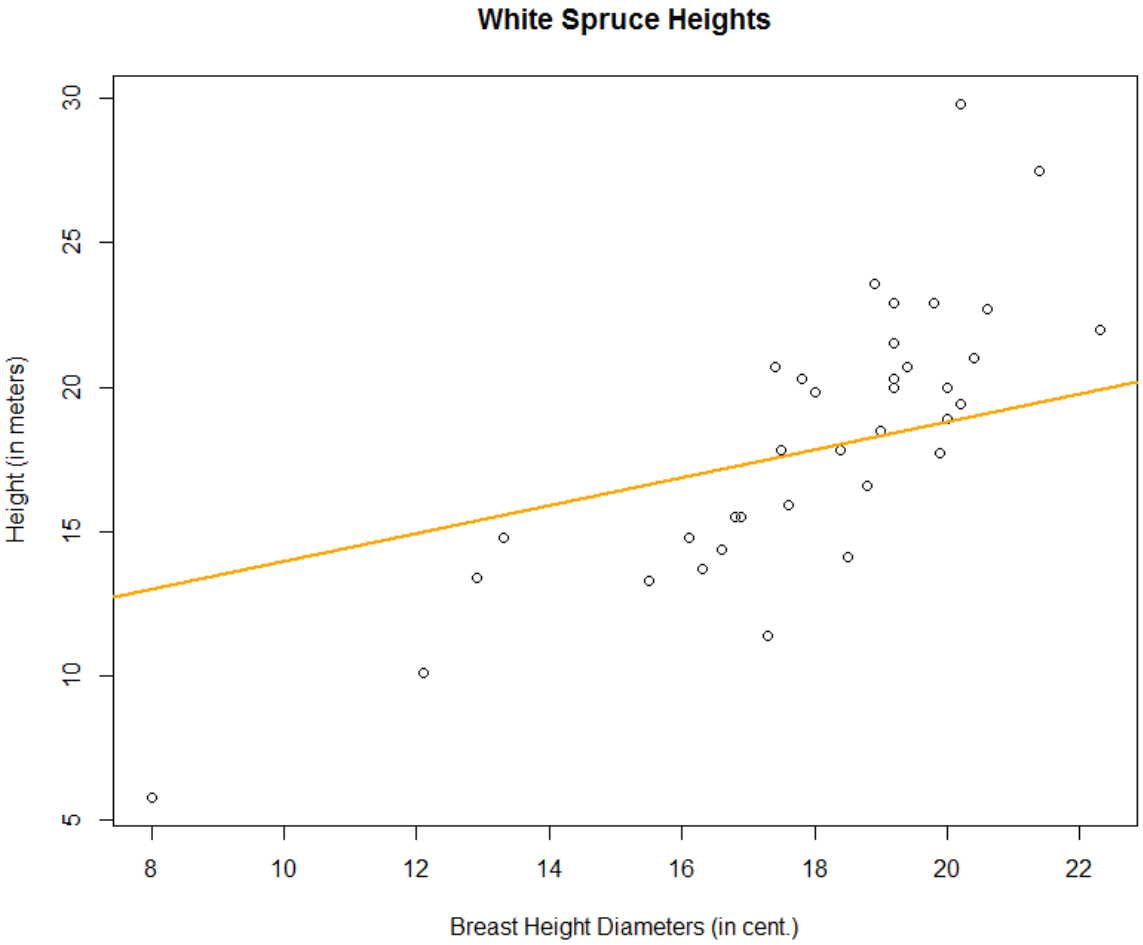
6. It turns out you can use both the Bonferroni and the Working-Hotelling approaches to make simultaneous CIs for mean response values. Sometimes the Bonferroni CIs will be tighter than the Working-Hotelling intervals, and sometimes not, but the Bonferroni CIs tend to be larger than the Working-Hotelling ones when families are large. So make both kinds of simultaneous CIs the mean work hours responses for the three lot size levels of 30, 60, and 100 for the Toluca Company data set CH01TA01 (you can read about this data set in your book btw). For the Bonferroni intervals, use a family error rate of 10% (same as family confidence level of 90%). Use 90% for the confidence level for the Working-Hotelling bands. Compare the Bonferroni intervals with the Working-Hotelling ones.

	fit	lwr	upr
1	347.9820	326.54494	369.4191
2	169.4719	134.36734	204.5765
3	240.8760	216.09481	265.6571
4	383.6840	358.90290	408.4652
5	312.2800	292.08026	332.4797
6	276.5780	255.14090	298.0151
7	490.7901	449.60756	531.9726
8	347.9820	326.54494	369.4191
9	419.3861	389.86150	448.9106
10	240.8760	216.09481	265.6571
11	205.1739	175.64938	234.6985
12	312.2800	292.08026	332.4797

13 383.6840 358.90290 408.4652
14 133.7699 92.58736 174.9524
15 455.0881 419.98350 490.1927
16 419.3861 389.86150 448.9106
17 169.4719 134.36734 204.5765
18 240.8760 216.09481 265.6571
19 383.6840 358.90290 408.4652
20 455.0881 419.98350 490.1927
21 169.4719 134.36734 204.5765
22 383.6840 358.90290 408.4652
23 205.1739 175.64938 234.6985
24 347.9820 326.54494 369.4191
25 312.2800 292.08026 332.4797

7. The height of a tree is often predicted based on its species and diameter at breast height. The data set whitespruce.txt contains the breast height diameters (in centimeters) and actual heights (in meters) for 36 white spruce trees. Does evidence suggest there exists a strong linear association between breast height diameter and tree height? Justify your response by making and remarking on a fitted line plot and ANOVA F-test. Be sure to state the null and alternative hypotheses and your conclusion.

A fitted line plot was constructed to examine if a linear relationship exists between Breast Height (x) and Height (y). A linear relationship seems to exist, albeit with the lowest and couple highest data points acting as high leverage points. A curvilinear relationship might also fit this data, perhaps better.



An ANOVA F-test was conducted and the following output was called:

Analysis of Variance Table

Response: whitespruce$ht

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
whitespruce$diam	1	183.245	183.245	65.101	2.089e-09	***
Residuals	34	95.703	2.815			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Finally, the null and alternative hypotheses were that there would be no linear relationship between Breast Height and Height, resulting in an F-statistic of ~1.0 while the alternative hypothesis is that we would reject the null in favor of a linear relationship. The F-test is shown above and is indeed far greater than 1.0 at a value of 65.101, prompting us to reject the null hypothesis.

APPENDIX: R SCRIPTS

Question 1:

```
> library(xlsx)
> ch01pr19 <- read.xlsx("c:/Users/allen.baumgarten/Documents/CH01PR19.xlsx", sheetIndex = "Sheet1")

# Wrote the following script ("sls_regression_outputs.r")and then called it via the source() function
x <- ch01pr19$ACT_Score  # ch01pr19$ACT_Score; ch01ta01$Labor_Hours;  used_cars$age
y <- ch01pr19$GPA     # ch01pr19$GPA; ch01ta01$Lot_Size; used_cars$price
n <- nrow(ch01pr19)
df <- n-2
ci <- "CI% used: 98%"
alpha <- .975   #Assign alpha-level: 90% CI=.950; 95% CI=.975; 98% CI=.990; 99% CI=.995
x_val <- 29

# Regression parameters and outputs
regression <- lm(y~x)        # regression function
b1 <- cor(x,y) * sd(y)/sd(x)       # calc b1 by hand
b0 <- mean(y) - (b1*mean(x))  # calc b0 by hand
fits <- b1*x + b0                  # fitted values
residuals <- y - fits              # residuals
SSE <- sum(residuals^2)            # SSE
MSE <- SSE/(df)                    # MSE
residuals_sd <- sqrt(MSE)          # se of the residuals

se.predicted <- sqrt(MSE*(1+.10 +((x_val-mean(x))^2)/sum((x-mean(x))^2)))
lpt.pred <- x_val*b1+b0 - qt(alpha, df) * se.predicted     # 95% CI for predicted value of Y (lower)
rpt.pred <- x_val*b1+b0 + qt(alpha, df) * se.predicted     # 95% CI for predicted value of Y (upper)

se.mnresponse <- sqrt(MSE*(.10+((x_val-mean(x))^2)/sum((x-mean(x))^2)))
lpt.mn <- x_val*b1+b0 - qt(.975, df) * se.mnresponse     # 95% CI for the mean response in Y (lower)
rpt.mn <- x_val*b1+b0 + qt(.975, df) * se.mnresponse     # 95% CI for the mean response in Y (upper)


sd.b1 <- sqrt(MSE/sum((x-mean(x))^2))   # se of b1
sd.b0 <- sqrt(MSE) * sqrt(1/n + (mean(x) ^2)/sum((x-mean(x))^2))   # se of b0
left.ci.b1 <- b1 - qt(alpha,df)*sd.b1      # lower CI for b1
right.ci.b1 <- b1 + qt(alpha,df)*sd.b1   # upper CI for b1
left.ci.b0 <- b0 - qt(alpha,df)*sd.b0      # lower CI for b0
right.ci.b0 <- b0 + qt(alpha,df)*sd.b0   # upper CI for b0

# Hypothesis tests
#    B1 One-sided test (less/greater than):
hyp_b1_1 <- -10      # Input a value to test the hypothesis of b1. eg, -10 or 0.  One-sided test only:  less than?
ts.beta1_1 <- (b1 - (hyp_b1_1))/sd.b1
pval_ts.beta1_1 <- pt(ts.beta1_1,df)

#    B1 Two-sided test (differs from):
hyp_b1_2 <- 0
ts.beta1_2 <- (b1 - (hyp_b1_2))/sd.b1
pval_ts.beta1_2 <- 2*pt(-abs((b1-hyp_b1_2)/sd.b1),df)

#   B0 Two-sided test (differs from):
hyp_b0 <- 0
pval_ts.beta_0 <- 2*pt(-abs((b0-hyp_b0)/sd.b0),df)
```

```
yhat <- b0 + b1*x_val

# Print Outputs:
print(summary(regression))
writeLines("")
print(ci)
writeLines("")
writeLines("CI of b0:")
print(b0)
writeLines("")
print(left.ci.b0)
print(right.ci.b0)
writeLines("")
writeLines("CI of b1:")
print(b1)
writeLines("")
print(left.ci.b1)
print(right.ci.b1)
plot(x,y)
abline(regression, lwd=2, col="orange")
source("sls_regression_outputs.r")

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q      Median      3Q      Max
-2.74004 -0.33827  0.04062  0.44064  1.22737

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
x            0.03883    0.01277   3.040  0.00292 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6231 on 118 degrees of freedom
Multiple R-squared:  0.07262,        Adjusted R-squared:  0.06476
F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917


[1] "CI% used: 99%"

CI of b0:
[1] 2.114049
[1] 1.273903
[1] 2.954196

CI of b1:
[1] 0.03882713
[1] 0.005385614
[1] 0.07226864

hyp_b1_2 <- 0
ts.beta1_2 <- (b1 - (hyp_b1_2))/sd.b1
```

```
pval_ts.beta1_2 <- 2*pt(-abs((b1-hyp_b1_2)/sd.b1),df)

b1 <- cor(x,y) * sd(y)/sd(x)  # calc b1 by hand
b0 <- mean(y) - (b1*mean(x))  # calc b0 by hand
fits <- b1*x + b0           # fitted values
residuals <- y - fits       # residuals
SSE <- sum(residuals^2)      # SSE
MSE <- SSE/(df)              # MSE
residuals_sd <- sqrt(MSE)    # se of the residuals
se <- sqrt(MSE*(.10+((x_val-mean(x))^2)/sum((x-mean(x))^2)))
lpt <- x_val*b1+b0 - qt(.975, df) * se  # 95% CI for predicted Y (lower)
rpt <- x_val*b1+b0 + qt(.975, df) * se  # 95% CI for predicted Y (upper)
> lpt
[1] 2.835119
> rpt
[1] 3.644953
> nd <- data.frame(y=29)
> predict(regression, nd, interval="confidence", level=.95)
se.predicted <- sqrt(MSE*(1+.10 +((x_val-mean(x))^2)/sum((x-mean(x))^2)))
lpt.pred <- x_val*b1+b0 - qt(alpha, df) * se.predicted # 95% CI for predicted value of Y (lower)
rpt.pred <- x_val*b1+b0 + qt(alpha, df) * se.predicted # 95% CI for predicted value of Y (upper)
p_conf1 <- predict(regression, interval="confidence", level=.95)
p_pred1 <- predict(regression, interval="prediction", level=.95)
nd <- data.frame(x=seq(0,12,length=51))
p_conf2 <- predict(regression, interval="confidence", newdata=nd, level=.95)
p_pred2 <- predict(regression, interval="prediction", newdata=nd, level=.95)
plot(y~x, data=data, ylim=c(0,5), xlim=c(10,40))
plot(data$x ~ data$y, data=data, ylim=c(0,6), xlim=c(-1,40))
abline(regression)
matlines(nd$x, p_conf2[,c("lwr","upr")], col=4, lty=1, type="b", pch="o")
matlines(nd$x, p_pred2[,c("lwr","upr")], col=3, lty=2, type="b", pch="o")
```

Question 4:

```
# source("sls_regression_outputs.r")

data <- ch01ta01
x <- data$Labor_Hours
y <- data$Lot_Size
n <- nrow(data)
df <- n-2
ci <- "CI% used: 95%"
alpha <- .975   #Assign alpha-level: 90% CI=.950; 95% CI=.975; 98% CI=.990; 99% CI=.995
x_val <- 3

# Regression parameters and outputs
regression <- lm(y~x)        # regression function
b1 <- cor(x,y) * sd(y)/sd(x)  # calc b1 by hand
b0 <- mean(y) - (b1*mean(x))  # calc b0 by hand
fits <- b1*x + b0           # fitted values
residuals <- y - fits       # residuals
SSE <- sum(residuals^2)      # SSE
MSE <- SSE/(df)              # MSE
residuals_sd <- sqrt(MSE)    # se of the residuals
```

```
se.predicted <- sqrt(MSE*(1+.10 +((x_val-mean(x))^2)/sum((x-mean(x))^2)))
lpt.pred <- x_val*b1+b0 - qt(alpha, df) * se.predicted # 95% CI for predicted value of Y (lower)
rpt.pred <- x_val*b1+b0 + qt(alpha, df) * se.predicted # 95% CI for predicted value of Y (upper)

se.mnresponse <- sqrt(MSE*(.10+((x_val-mean(x))^2)/sum((x-mean(x))^2)))
lpt.mn <-~ x_val*b1+b0 - qt(.975, df) * se.mnresponse  # 95% CI for the mean response in Y (lower)
rpt.mn <- x_val*b1+b0 + qt(.975, df) * se.mnresponse  # 95% CI for the mean response in Y (upper)


sd.b1 <- sqrt(MSE/sum((x-mean(x))^2))   # se of b1
sd.b0 <- sqrt(MSE) * sqrt(1/n + (mean(x) ^2)/sum((x-mean(x))^2))   # se of b0
left.ci.b1 <- b1 - qt(alpha,df)*sd.b1   # lower CI for b1
right.ci.b1 <- b1 + qt(alpha,df)*sd.b1  # upper CI for b1
left.ci.b0 <- b0 - qt(alpha,df)*sd.b0   # lower CI for b0
right.ci.b0 <- b0 + qt(alpha,df)*sd.b0  # upper CI for b0

# Hypothesis tests
#    B1 One-sided test (less/greater than):
hyp_b1_1 <- 3     # Input a value to test the hypothesis of b1. eg, -10 or 0.  One-sided test only:  less than?
ts.beta1_1 <- (b1 - (hyp_b1_1))/sd.b1
pval_ts.beta1_1 <- pt(ts.beta1_1,df)

#    B1 Two-sided test (differs from):
hyp_b1_2 <- 0
ts.beta1_2 <- (b1 - (hyp_b1_2))/sd.b1
pval_ts.beta1_2 <- 2*pt(-abs((b1-hyp_b1_2)/sd.b1),df)

#   B0 Two-sided test (differs from):
hyp_b0 <- 0
pval_ts.beta_0 <- 2*pt(-abs((b0-hyp_b0)/sd.b0),df)

yhat <- b0 + b1*x_val

# Print Outputs:
print(summary(regression))
writeLines("")
print(ci)
writeLines("")
writeLines("CI of b0:")
print(b0)
writeLines("")
print(left.ci.b0)
print(right.ci.b0)
writeLines("")
writeLines("CI of b1:")
print(b1)
writeLines("")
print(left.ci.b1)
print(right.ci.b1)
plot(x,y)
abline(regression, lwd=2, col="orange")
b1left.end.pt <- b1 - qt(.975, df)*sd.b1
b1right.end.pt <- b1 + qt(.975, df)*sd.b1
b1left.end.pt
[1] 2.852435
```

```
b1right.end.pt
[1] 4.287969
> b0left.end.pt <- b0 - qt(.975, df)*sd.b0
> b0right.end.pt <- b0 + qt(.975, df)*sd.b0
> b0left.end.pt
[1] 8.213711
> b0right.end.pt
[1] 116.518
```

Question 5:

```
hyp_b1_1 <- 3
ts.beta1_1 <- (b1 - (hyp_b1_1))/sd.b1
pval_ts.beta1_1 <- pt(ts.beta1_1,df)
> pt(5 + qt(.05,23),23)
> pt(1 + qt(.05,23),23)
```

Question 6:

```
p_conf1 <- predict(regression, interval="confidence", level=.95)
       fit      lwr      upr
1  347.9820 326.54494 369.4191
2  169.4719 134.36734 204.5765
3  240.8760 216.09481 265.6571
4  383.6840 358.90290 408.4652
5  312.2800 292.08026 332.4797
6  276.5780 255.14090 298.0151
7  490.7901 449.60756 531.9726
8  347.9820 326.54494 369.4191
9  419.3861 389.86150 448.9106
10 240.8760 216.09481 265.6571
11 205.1739 175.64938 234.6985
12 312.2800 292.08026 332.4797
13 383.6840 358.90290 408.4652
14 133.7699  92.58736 174.9524
15 455.0881 419.98350 490.1927
16 419.3861 389.86150 448.9106
17 169.4719 134.36734 204.5765
18 240.8760 216.09481 265.6571
19 383.6840 358.90290 408.4652
20 455.0881 419.98350 490.1927
21 169.4719 134.36734 204.5765
22 383.6840 358.90290 408.4652
23 205.1739 175.64938 234.6985
24 347.9820 326.54494 369.4191
25 312.2800 292.08026 332.4797
```

Question 7:

```
> whitespruce_reg <- lm(whitespruce$ht ~ whitespruce$diam)
> plot(whitespruce$ht,whitespruce$diam,xlab="Breast Height Diameters (in cent.)",ylab="Height (in meters)",main="White Spruce
Heights")
> abline(whitespruce_reg, lwd = 2, col = "orange")
> anova(whitespruce_reg)
Analysis of Variance Table
```

```
Response: whitespruce$ht
               Df  Sum Sq Mean Sq F value    Pr(>F)
whitespruce$diam  1 183.245 183.245  65.101 2.089e-09 ***
Residuals        34  95.703   2.815
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```