STAT 510 – HW 1

Erick Castillo

9/4/2021

Problem 1

A. I did not include the code where I install necessary packages for aesthetic purposes.

```
mod1 <- lm(wt~ht, data = htwt)
summary(mod1)
```

```
##
## Call:
## lm(formula = wt ~ ht, data = htwt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1166 -4.7744 -2.8412  0.5696 18.4581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.8759    64.4728  -0.572    0.583
## ht            0.5821     0.3892   1.496    0.173
##
## Residual standard error: 8.456 on 8 degrees of freedom
## Multiple R-squared:  0.2185, Adjusted R-squared:  0.1208
## F-statistic: 2.237 on 1 and 8 DF,  p-value: 0.1731
```

- Yes, there appears to be a relationship between the predictor and the response.
- The relationship between the variables is weak. This is given by the low R^2 and the high p-value associated with the variable ht.
- There is a positive relationship present between the variables.
- The following code calculates the predicted weight and the confidence/prediction intervals:

```
newdata = data.frame(ht = 165)

# output for prediction intervals.
predict(mod1, newdata, interval="predict")
```

```
##        fit      lwr      upr
## 1 59.16732 38.71099 79.62364
```

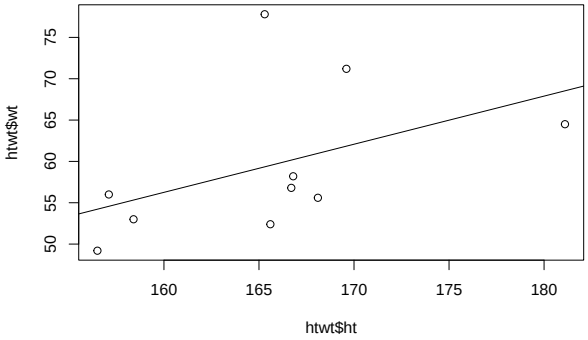
```r
# output for confidence intervals.
predict(mod1, newdata, interval="confidence")
```

```
##        fit      lwr      upr
## 1 59.16732 52.98347 65.35116
```

Note that the "fit" values are the predicted weight of a person with a height of 165 cm. The prediction and confidence intervals are $(38.71,\ 79.62)$ and $(52.98,\ 65.35)$ respectively.

B. The following is the plot of the above data.

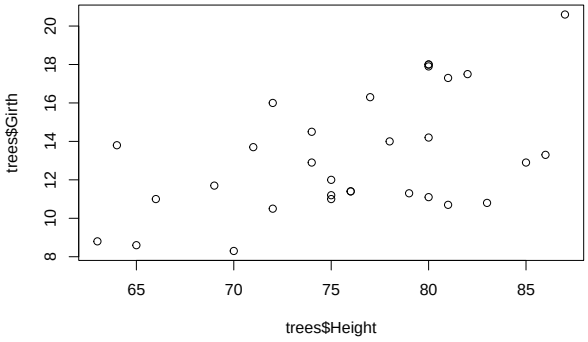
```r
plot(htwt$ht, htwt$wt)
abline(mod1)
```



Problem 2

A. Draw scatterplot of Girth and Height.

```r
plot(trees$Height, trees$Girth)
```



A simple linear regression seems appropriate in this case. Without using R to build a model, I would suggest a log() transformation on Girth because the Y_i data is very spread.

B. Compute $\bar{x}, \bar{y}, S_{xx}, S_{yy}$, and S_{xy}, along with $\hat{\beta}_1$ and $\hat{\beta}_0$. Finally draw the fitted line on the scatter plot.

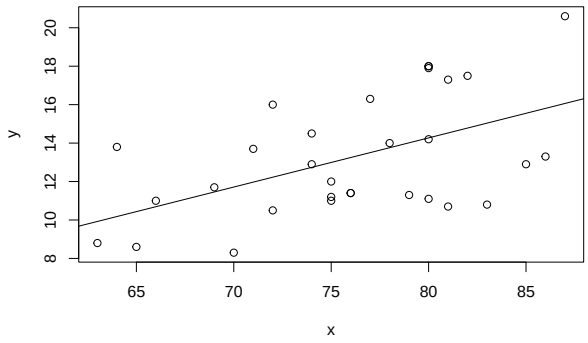
```r
x <- trees$Height
y <- trees$Girth

# calculating desired values.
xbar <- mean(x)
ybar <- mean(y)
sxx <- sum((x-xbar)^2)
syy <- sum((y-ybar)^2)
sxy <- sum((x-xbar)*(y-ybar))

# calculating parameter estimates.
b1 <- sxy/sxx
b0 <- ybar - b1*xbar

c(xbar,ybar,sxx,syy,sxy,b1,b0)
```

```
## [1]   76.0000000   13.2483871 1218.0000000  295.4374194  311.5000000
## [6]    0.2557471   -6.1883945
```

The values displayed in the above vector are $\bar{x}, \bar{y}, S_{xx}, S_{yy}, S_{xy}, \hat{\beta}_1$, and $\hat{\beta}_0$ respectively.

Below is the plot with the regression line:

```r
plot(x,y)
abline(a = b0, b = b1)
```



C. Next obtain the estimate of σ^2 and find the estimated standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$. Then compute the t-tests for the hypotheses $\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = 0$, find the associated p-values for a two-tailed test.

```r
n <- length(x)
yhat <- b0 + b1*x # this is a vector of ALL predicted values
mse <- sum((y-yhat)^2)/(n-2)

se.b1 <- sqrt(mse/sxx)
se.b0 <- sqrt(mse*(((1/n)+(xbar^2/sxx))))

c(mse, se.b0, se.b1)
```

```
## [1] 7.4404203 5.9601994 0.0781583
```

The values in the above vector show the values for $\hat{\sigma}^2, se(\hat{\beta}_0)$, and $se(\hat{\beta}_1)$ respectively.

Next I test the hypotheses and find the associated p-values for a two-tailed test.

```
t.b1 <- (b1-0)/sqrt(mse/sxx)
t.b0 <- (b0-0)/sqrt(mse*((1/n)+(xbar^2/sxx)))

p.b1 <- 2*(1-pt(t.b1,n-2))
p.b0 <-2*(pt(t.b0, n-2))

c(t.b1,p.b1,t.b0,p.b0)
```

```
## [1]  3.272168591  0.002757815 -1.038286484  0.307716768
```

The above vector output displays the test statistic for $\hat{\beta}_1$, its corresponding p-value, the test statistic for $\hat{\beta}_0$, and $\hat{\beta}_0$'s corresponding p-value respectively.

Using these p-values we would conclude that there is good evidence to reject $H_0 : \hat{\beta}_1 = 0$ and we fail to reject $H_0 : \hat{\beta}_0 = 0$ at $\alpha = 0.05$.

D. Construct a 95% confidence interval for $\hat{\beta}_1$.

```
b1-c(-1,1)*qt(0.025,n-2)*se.b1
```

```
## [1] 0.09589546 0.41559879
```

The above output is the 95% confidence interval for $\hat{\beta}_1$.

E. Compute the variability in Girth explained by Height. Explain what this means.

```
cor(x,y)^2
```

```
## [1] 0.2696518
```

This represents the Multiple R^2 output from the lm() function. This value can be thought of as what percentage of the variation in the response variable Girth is explained by the regression model. So in this case, $\approx 27\%$ of the variability in Girth is explained by the variability in the predictor Height.

F. Compute a 95% prediction interval for the Girth of a tree which has a height of $94ft$.

```
ynew <- b0 + b1*94

# value of interest.
ynew
```

```
## [1] 17.85184
```

```
# prediction interval.
ynew + c(1,-1)*qt(0.05/2,n-2)*sqrt(mse*(1+(1/n)+(94-xbar)^2/sxx))
```

```
## [1] 11.49526 24.20841
```

The above output displays that the 95% prediction interval for the girth of a tree that has a height of $x_0 = 94$ is (11.5, 24.21).

Problem 3

A. Use the code provided in the assignment.

```
set.seed(1)
n <- 100
x = runif(n)
```

B. Use the rnorm function to generate a vector that contains 100 observations from a $N(0, 0.25)$ distribution.

```
eps <- rnorm(n,0,sqrt(0.25))
```

C. Use the above vectors to create a vector y according to the model

$$Y = -1 + 0.5x + \epsilon$$

. What are the values of β_0 and β_1 in this linear model?

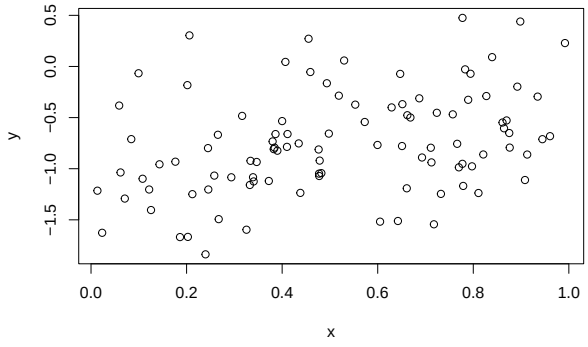
```
y <- -1+0.5*x+eps
length(y)
```

```
## [1] 100
```

The length of the above vector is 100. $\beta_0 = -1$ and $\beta_1 = 0.5$.

D. Create a scatter plot displaying the relationship between x and y.

```
plot(x,y)
```



It appears that there is a somewhat positive relationship between x and y. This relationship is mostly obscured because of the spread of the points.

E. Fit a LS model to compare β_0 and β_1 to $\hat{\beta}_0$ and $\hat{\beta}_1$.

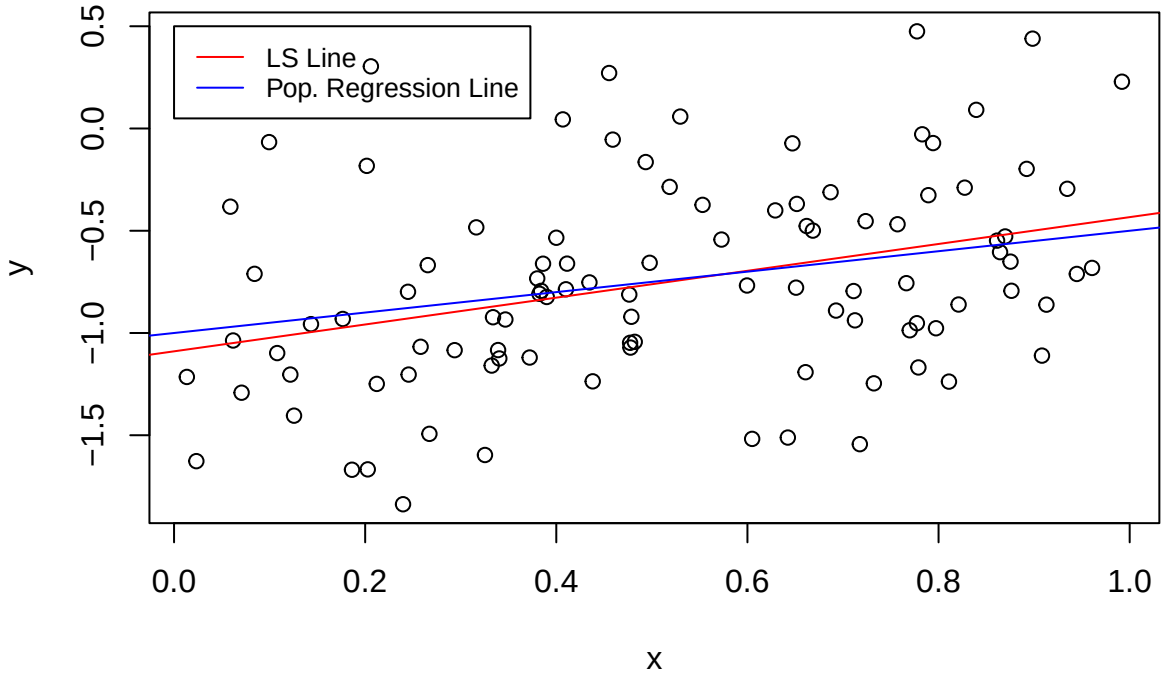
```
mod2 <- lm(y~x)
summary(mod2)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92489 -0.28111 -0.04353  0.26214  1.25830
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.0897     0.1029 -10.589  < 2e-16 ***
## x             0.6562     0.1767   3.713 0.000341 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4705 on 98 degrees of freedom
## Multiple R-squared:  0.1233, Adjusted R-squared:  0.1144
## F-statistic: 13.79 on 1 and 98 DF,  p-value: 0.0003405
```

Note that $\beta_1 = 0.5$ while $\hat{\beta}_1 \approx 0.656$ and $\beta_0 = -1$ while $\hat{\beta}_0 \approx -0.9411$. This means that the values are not exactly what were declared to be in the model, but they are close.

F. Display the LS line on the scatter plot from D. Draw the population regression line on the plot in a different color. Use the legend() command to create the appropriate legend.

```
plot(x,y)
abline(mod2,col = 'red')
abline(a = -1, b = 0.5, col = 'blue')

legend(0, 0.5, legend=c("LS Line", "Pop. Regression Line"),
       col=c("red", "blue"), lty = 1, cex=0.8)
```


LS Line
Pop. Regression Line
y
x
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.2
0.4
0.6
0.8
1.0