STAT158HW5

Oliver Shanklin

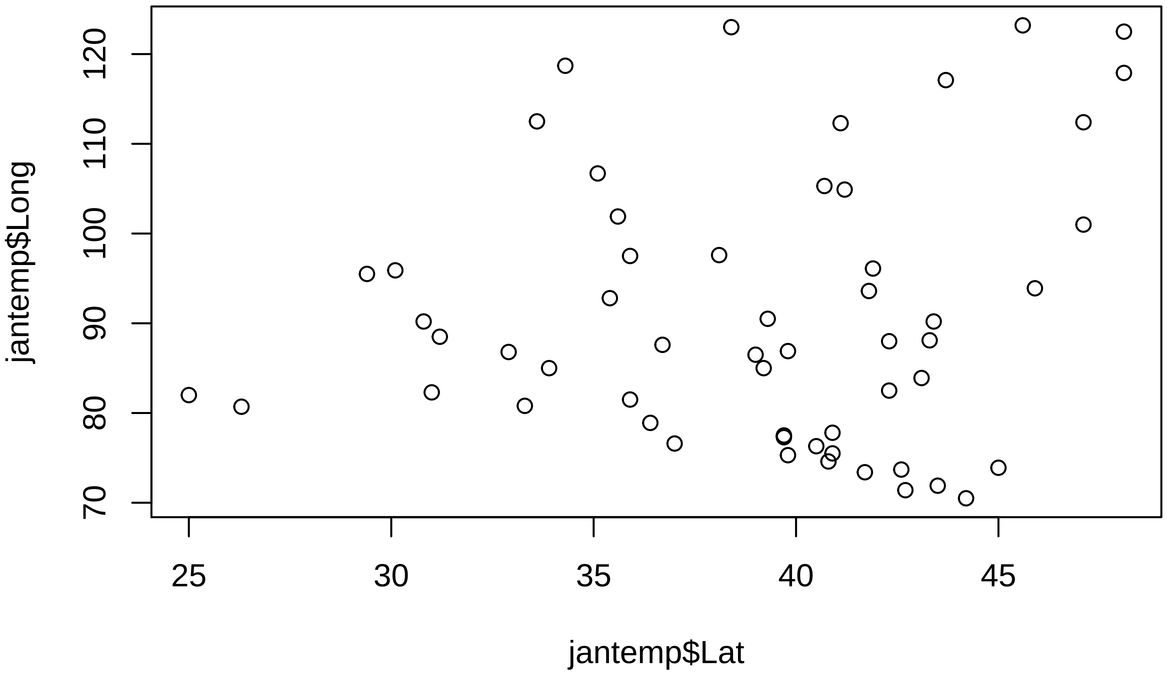
April 12, 2019

1)

a)

```
jantemp <- read.csv("january_temp.csv", header = TRUE)
attach(jantemp)

plot(jantemp$Lat, jantemp$Long)
```

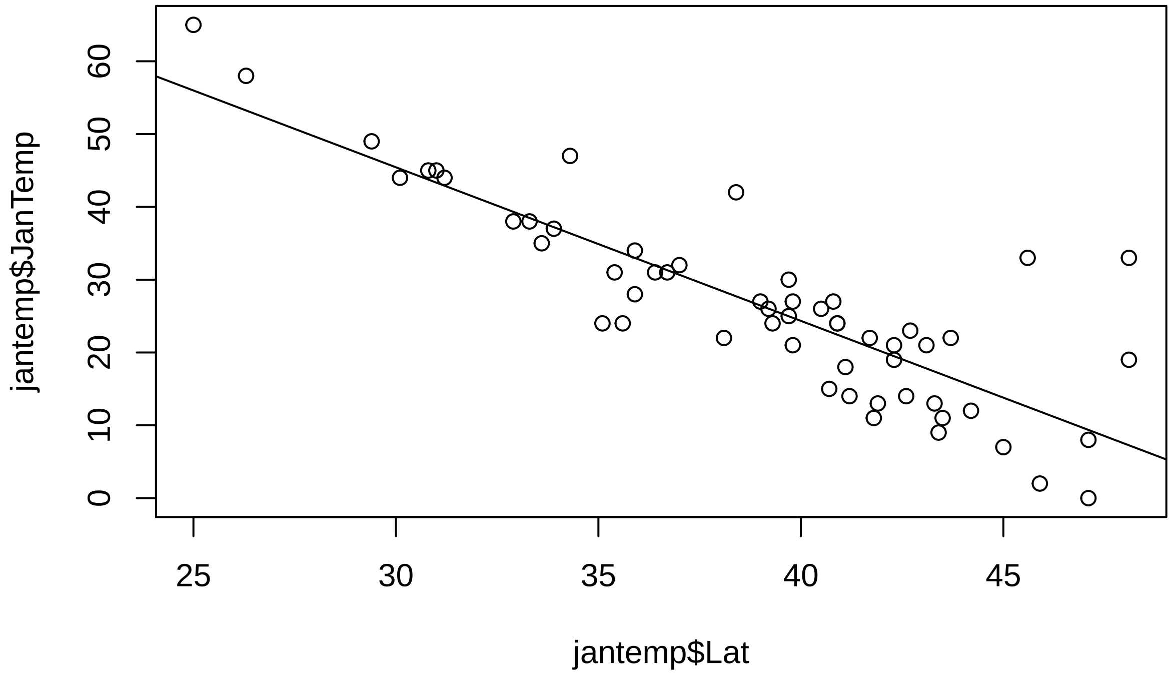


I was not sure what I was expecting for the plot.

b)

```
jantempLat <- lm(jantemp$JanTemp~jantemp$Lat)

plot(jantemp$JanTemp~jantemp$Lat)
abline(jantempLat$coefficients)
```



```r
anova(jantempLat)
```

```
## Analysis of Variance Table
##
## Response: jantemp$JanTemp
##              Df Sum Sq Mean Sq F value    Pr(>F)
## jantemp$Lat   1 7080.9  7080.9  138.28 < 2.2e-16 ***
## Residuals    54 2765.1    51.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(jantempLat)
```

```
##
## Call:
## lm(formula = jantemp$JanTemp ~ jantemp$Lat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6812  -4.5018  -0.2593   2.2489  25.7434
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 108.7277     7.0561   15.41   <2e-16 ***
## jantemp$Lat  -2.1096     0.1794  -11.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 7.156 on 54 degrees of freedom
## Multiple R-squared:  0.7192, Adjusted R-squared:  0.714
## F-statistic: 138.3 on 1 and 54 DF,  p-value: < 2.2e-16
```

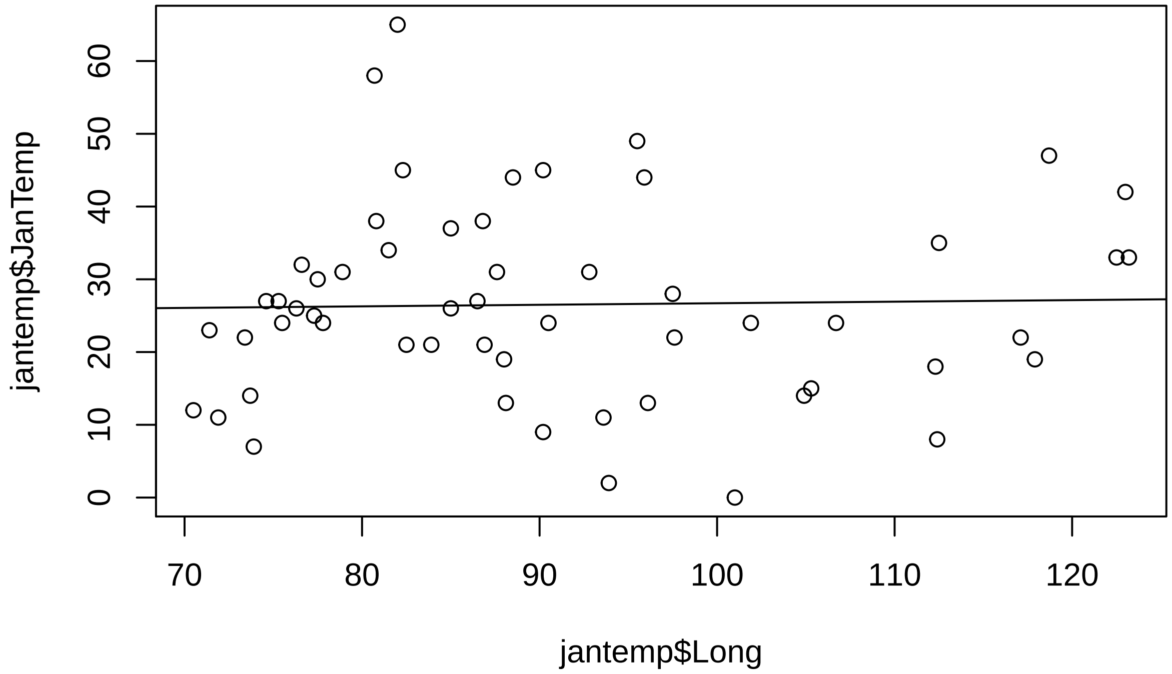
JanTemp = 108.73 - 2.11(Latitude)

R^2 is 0.7192

c)

```
jantempLong <- lm(jantemp$JanTemp~jantemp$Long)

plot(jantemp$JanTemp~jantemp$Long)
abline(jantempLong$coefficients)
```



```
anova(jantempLong)
```

```
## Analysis of Variance Table
##
## Response: jantemp$JanTemp
##               Df Sum Sq Mean Sq F value Pr(>F)
## jantemp$Long  1    5.6   5.644   0.031  0.861
## Residuals    54 9840.3 182.228
```

```
summary(jantempLong)
```

```
##
```

```
## Call:
## lm(formula = jantemp$JanTemp ~ jantemp$Long)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.733  -8.314  -1.706   6.277  38.674
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   24.5710    11.2089   2.192   0.0327 *
## jantemp$Long   0.0214     0.1216   0.176   0.8610
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.5 on 54 degrees of freedom
## Multiple R-squared:  0.0005732,  Adjusted R-squared:  -0.01793
## F-statistic: 0.03097 on 1 and 54 DF,  p-value: 0.861
```

JanTemp = 24.571 + 0.0214(Longitude)

R^2 is 0.0005732

d)

It seems that the Latitude does a better job at predicting the tempurature since the R^2 value is closer to 1.

e)

```
jantempLatLong <- lm(jantemp$JanTemp~jantemp$Lat+jantemp$Long)

anova(jantempLatLong)
```

```
## Analysis of Variance Table
##
## Response: jantemp$JanTemp
##              Df Sum Sq Mean Sq  F value  Pr(>F)
## jantemp$Lat   1 7080.9  7080.9 147.2492 < 2e-16 ***
## jantemp$Long  1  216.5   216.5   4.5014 0.03856 *
## Residuals    53 2548.6    48.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(jantempLatLong)
```

```
##
## Call:
## lm(formula = jantemp$JanTemp ~ jantemp$Lat + jantemp$Long)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9983  -3.8957   0.5577   3.7330  22.0113
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 98.64523    8.32708  11.846   <2e-16 ***
## jantemp$Lat -2.16355    0.17570 -12.314   <2e-16 ***
## jantemp$Long  0.13396    0.06314   2.122   0.0386 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.935 on 53 degrees of freedom
## Multiple R-squared:  0.7411, Adjusted R-squared:  0.7314
## F-statistic: 75.88 on 2 and 53 DF,  p-value: 2.792e-16
```

$JanTemp = 98.64 - 2.16355(Lat) + 0.13396(Long)$

$R^2 = 0.7411$

The coefficient that changed the most was the Longitude. Up from 0.02.

f)

```
jantempLatLongInt <- lm(jantemp$JanTemp~jantemp$Lat+jantemp$Long + jantemp$Lat*jantemp$Long)

anova(jantempLatLongInt)
```

```
## Analysis of Variance Table
##
## Response: jantemp$JanTemp
##                          Df Sum Sq Mean Sq  F value    Pr(>F)
## jantemp$Lat               1 7080.9  7080.9 181.4602 < 2.2e-16 ***
## jantemp$Long              1  216.5   216.5   5.5472 0.0223179 *
## jantemp$Lat:jantemp$Long  1  519.5   519.5  13.3137 0.0006109 ***
## Residuals                52 2029.1    39.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(jantempLatLongInt)
```

```
##
## Call:
## lm(formula = jantemp$JanTemp ~ jantemp$Lat + jantemp$Long + jantemp$Lat *
##     jantemp$Long)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.6738  -2.8165  -0.1268   3.4107  15.0605
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              259.48952   44.71515   5.803 3.93e-07 ***
## jantemp$Lat               -6.07039    1.08235  -5.609 7.94e-07 ***
## jantemp$Long              -1.61025    0.48139  -3.345 0.001533 **
## jantemp$Lat:jantemp$Long   0.04220    0.01156   3.649 0.000611 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.247 on 52 degrees of freedom
## Multiple R-squared:  0.7939, Adjusted R-squared:  0.782
## F-statistic: 66.77 on 3 and 52 DF,  p-value: < 2.2e-16
```

$JanTemp = 259.48952 - 6.07(Lat) - 1.61(Long) + .042(Lat)(Long)$

$R^2 = 0.7939$

The R^2 Value did not change by much.

g)

Since I ran the anova command in each part here is the list of MSE.

51.2 for b), 182.228 for c), 48.1 for e), and 39.0 for f)

The closer to 0 that the r-squared is, the higher MSE is produced, and the farther away from 0, the MSE is lower.

2)

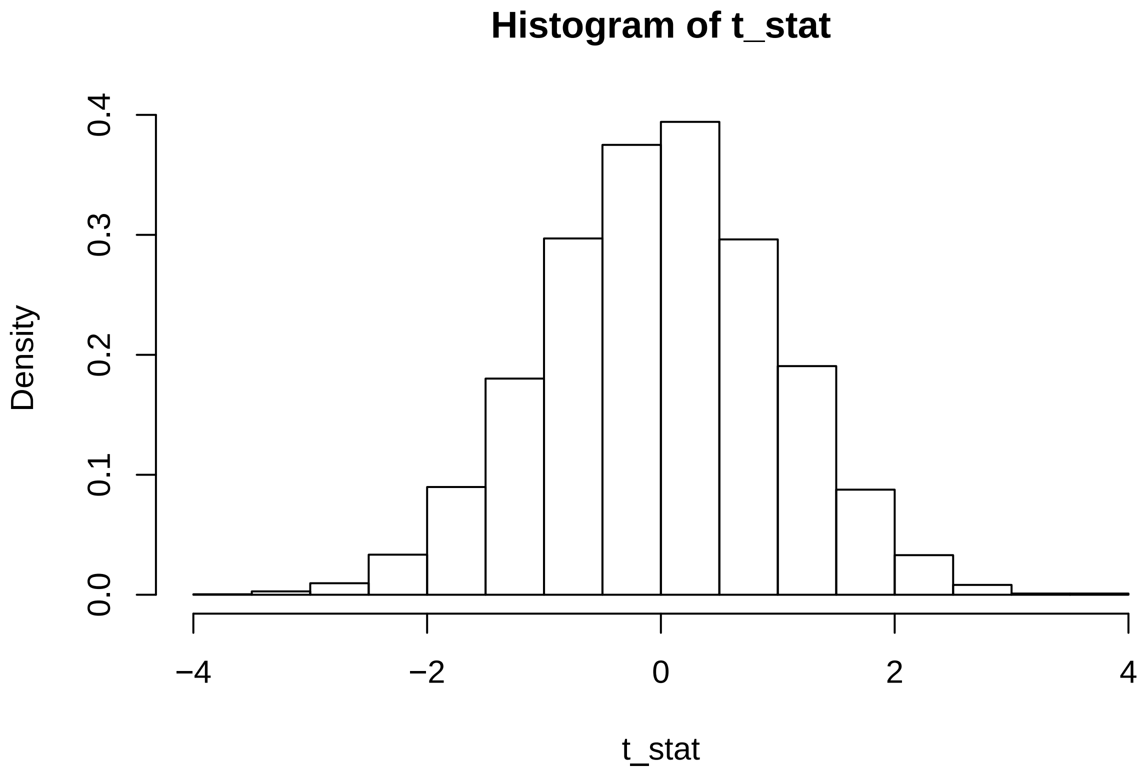
```r
t_stat <- numeric()

for(i in 1:10000){

  sample1 <- rnorm(10000,0,1)
  sample2 <- rnorm(10000,0,1)

  t_stat[i] <- t.test(sample1, sample2, var.equal = F)$statistic
}
#t_stat


hist(t_stat, freq = F)
```



3)

a)

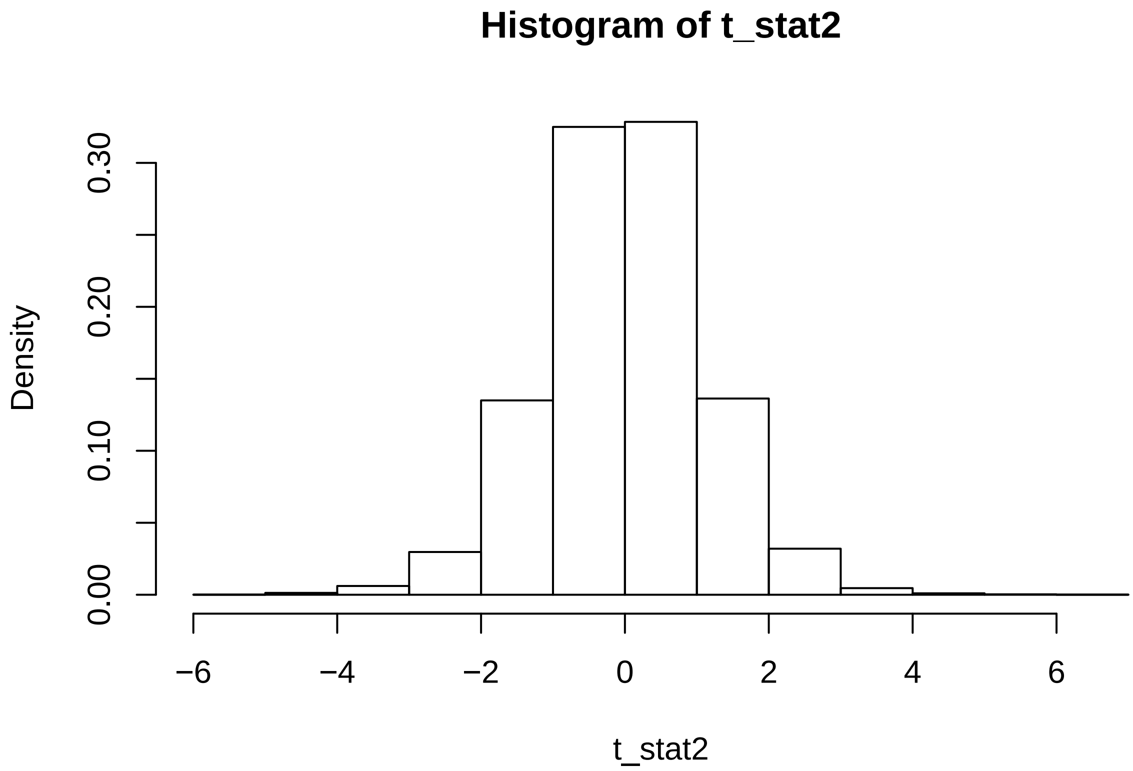
```r
t_stat2 <- numeric()

for(i in 1:10000){

  sample1 <- rnorm(100,0,1)
  sample2 <- rnorm(10,0,5)

  t_stat2[i] <- t.test(sample1, sample2, var.equal = F)$statistic
}
#t_stat


hist(t_stat2, freq = F)
```



This simulation produces a histogram very tall area around the mean, unlike 2) where it looked normal.

b)

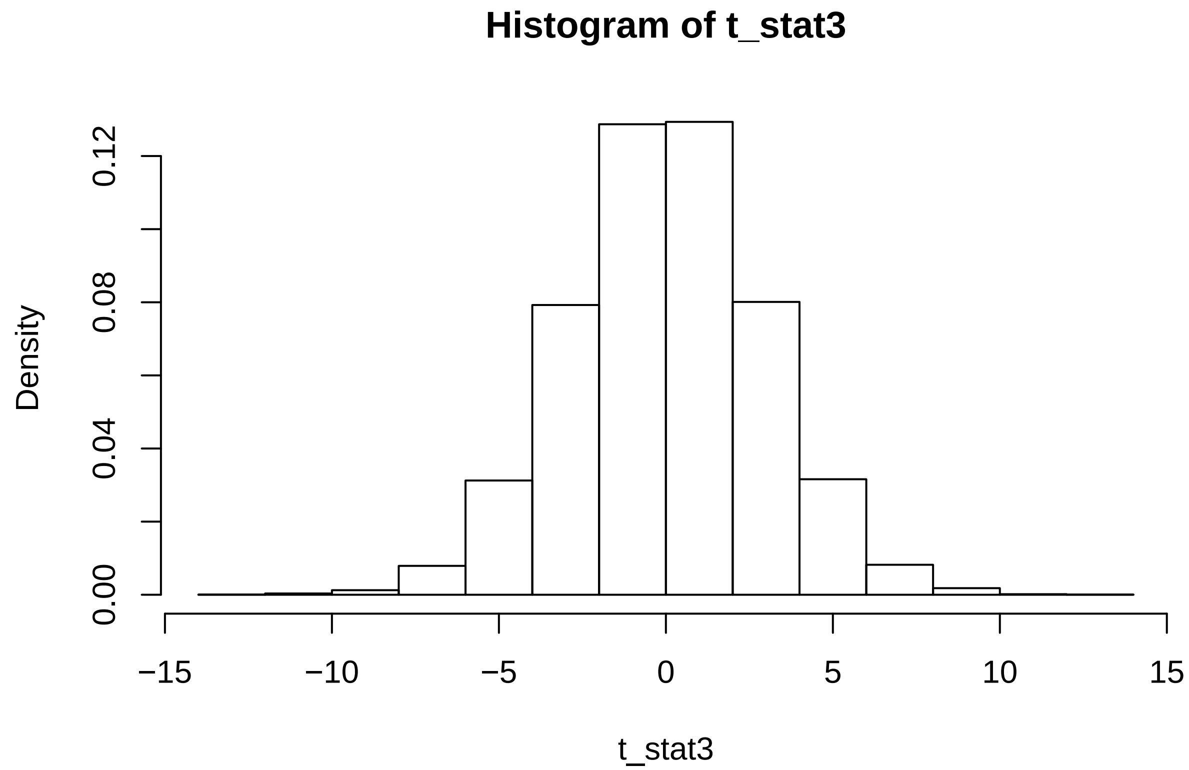
```r
t_stat3 <- numeric()

for(i in 1:10000){

  sample1 <- rnorm(100,0,1)
  sample2 <- rnorm(10,0,5)

  t_stat3[i] <- t.test(sample1, sample2, var.equal = TRUE)$statistic
}
#t_stat


hist(t_stat3, freq = F)
```



This histogram seems to be much wider with heavier tails.