Carson Rupp
0030633445
STAT 51200
9-22-21

STAT 51200 HW #2

1.

2.14.1

```
attach(Heights)
construct.sample = sample(1:dim(Heights)[1], floor(dim(Heights))[1]*(2/3))
```

2.14.2

The average squared residual is 6.035669
The average prediction error is 2.45676

```
m1 = lm(mheight~dheight, data=Heights, subset=construct.sample)
result <- Heights[-construct.sample,]
result['Prediction'] = predict(m1, newdata = Heights[-construct.sample,])
result['Residual'] = (result['Prediction'] - mheight)^2
n = nrow(result)
sqres <- sum(result['Residual']) / n
sqres
sqrt(sqres)
```

2.14.3

The average squared prediction error is 6.333349, it's square root is 2.516615. These averages are very similar to the ones in 2.14.2 but a little higher.

```
result['sepred'] = sd(dheight) * (1 + 1/n +
  (result['dheight'] - mean(dheight))^2)^(1/2)
sum(result['sepred'] / n)
sqrt(sum(result['sepred']) / n)
```

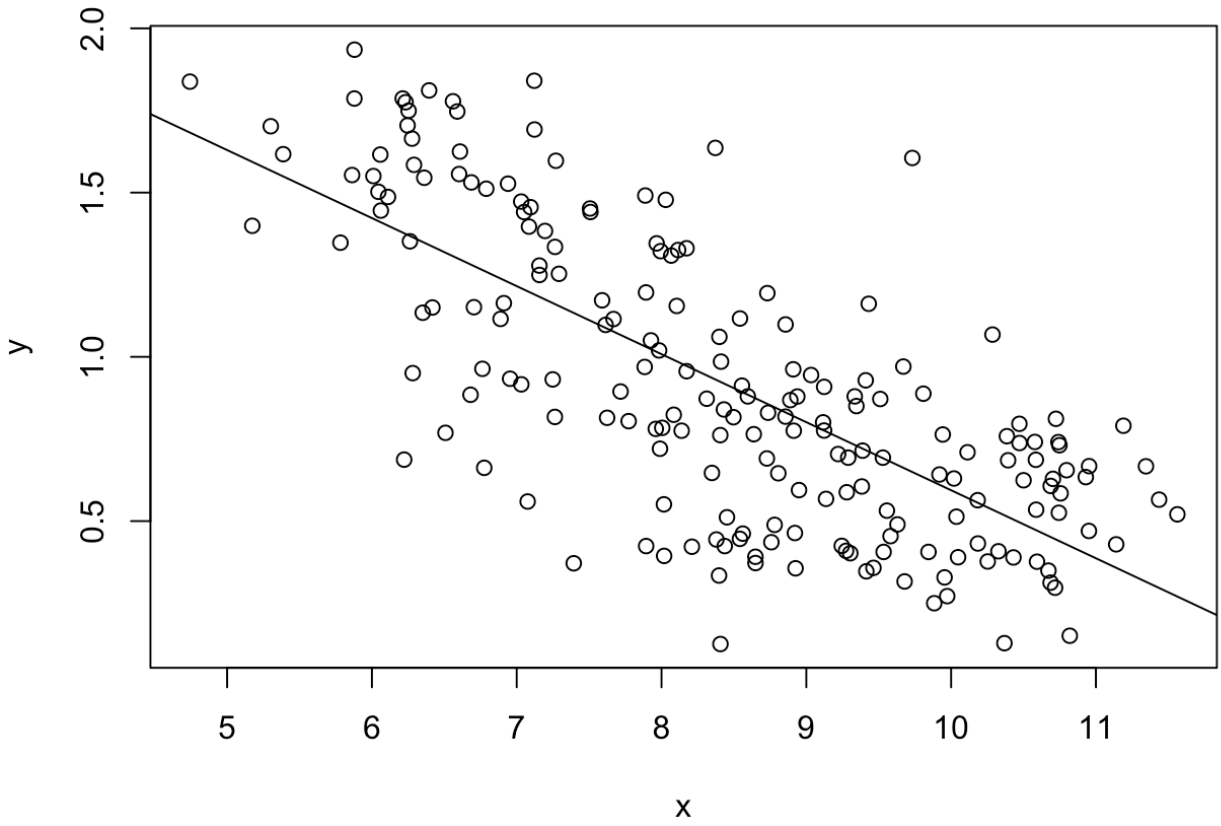
Carson Rupp
0030633445
STAT 51200
9-22-21

2.

2.16.1

attach(UN11)
x = log(ppgdp)
y = log(fertility)
line = lm(y~x, data = UN11)

2.16.2



abline(line)
plot(x,y)

2.16.3

The P-value (Significance Level) is 2e-16 and is the probability that the null hypothesis is true, which is not very probable.

res <- summary(line)
pt(coef(res)[, 3], line$df, lower = TRUE)
tstat = coef(res)[2,1] / coef(res)[2,2]
tstat

Carson Rupp
0030633445
STAT 51200
9-22-21

pt(tstat, 197, lower.tail=TRUE)
summary(line)

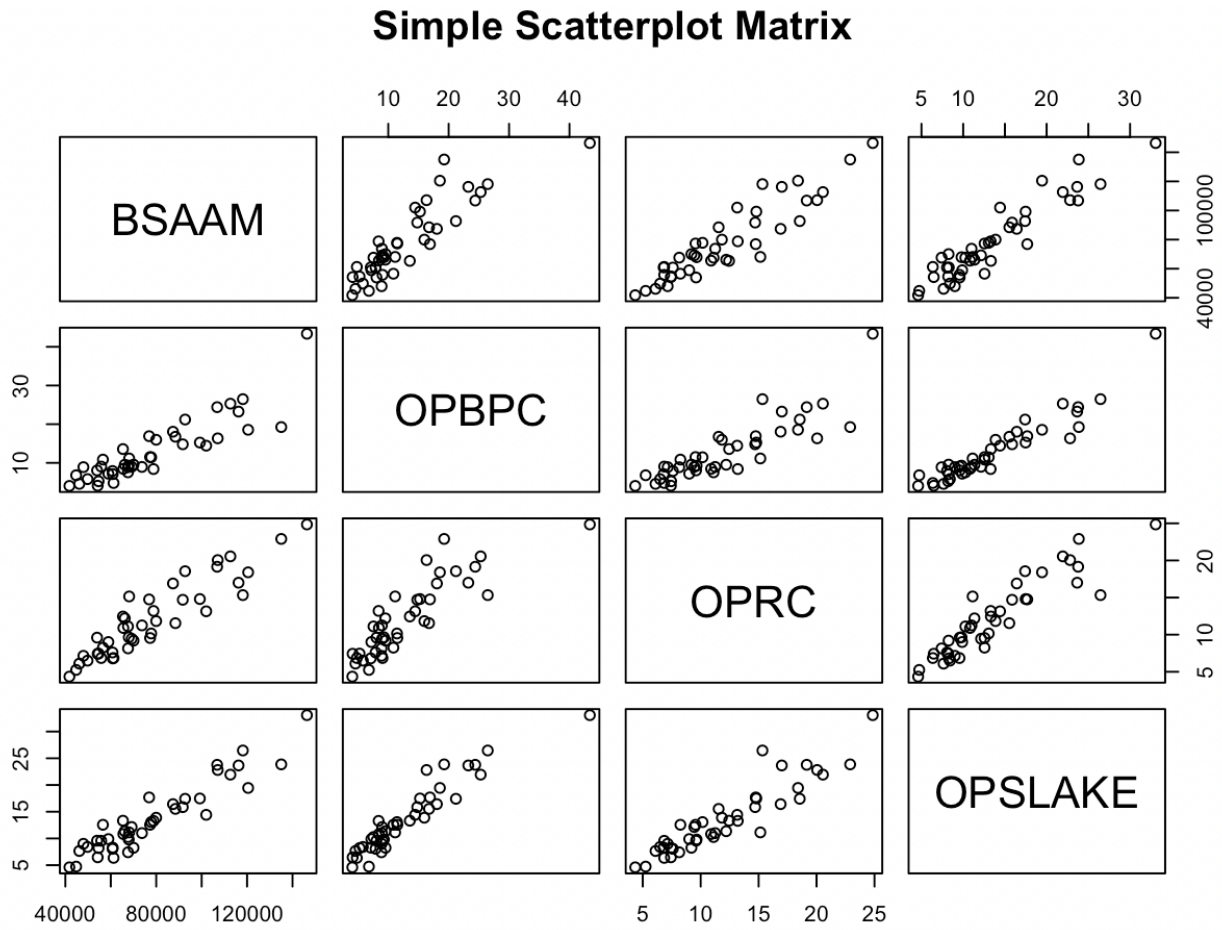
2.16.4
The R^2 (coefficient of determination) is .526
R^2 (coefficient of determination) shows how well the predictions from the regression line approximate real data points. 52.6 % of the variance in the dependent variable is explained by the independent variable in this case.
summary(line)

2.16.5
The point prediction with a generated by my model is 1.234567 for log(fertility) when ppgdp = 1000. The 95% prediction interval is (0.6258791, 1.843256). The 95% prediction interval for fertility is (1.869889, 6.31707).
mod = lm(log(fertility)~log(ppgdp), data = UN11)
dat = data.frame(ppgdp = c(1000))
#1.234567 is the predicted value when ppgdp is 1000
predict(mod, dat, interval='predict')
a = predict(mod, dat, interval='predict')[2]
b = predict(mod, dat, interval='predict')[3]
plot(ppgdp, fertility)
exp(a)
exp(b)

Carson Rupp
0030633445
STAT 51200
9-22-21

3.

3.6.1



The correlation matrices should all be positively correlated with BSAAM and OPSLAKE being the most positively correlated, OPRC being the second, and OPBPC being the third. This is in agreement with the correlation matrix calculated below:

	OPBPC	OPRC	OPSLAKE	BSAAM
OPBPC	1.0000000	0.8647073	0.9433474	0.8857478
OPRC	0.8647073	1.0000000	0.9191447	0.9196270
OPSLAKE	0.9433474	0.9191447	1.0000000	0.9384360
BSAAM	0.8857478	0.9196270	0.9384360	1.0000000

attach(water)
mod = lm(BSAAM~OPBPC+OPRC+OPSLAKE)
summary(mod)

Carson Rupp
0030633445
STAT 51200
9-22-21

```
pairs(~BSAAM+OPBPC+OPRC+OPSLAKE,data=water,
    main="Simple Scatterplot Matrix")
cor(water[, c(5,6,7,8)])
```

3.6.2

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22991.85    3545.32   6.485  1.1e-07 ***
OPBPC          40.61     502.40   0.081  0.93599
OPRC         1867.46     647.04   2.886  0.00633 **
OPSLAKE      2353.96     771.71   3.050  0.00410 **
---
```

These T-values are the test statistics calculated from our estimate/standard error and when used in a t-test, it tells us how likely our estimate is given our calculated t value under the null hypothesis (B0 = 0 in this case). When the p-value is calculated (Probability to get a test statistic that is not our calculated one) we can use this to determine how likely our estimate of B* is. In this case, our estimate of b corresponding to OPBPC is unlikely but the other two are, this signals that the relationship between OPBPC and BSAAM is likely due to chance and probably should not be included in the model.

```
summary(mod)
```