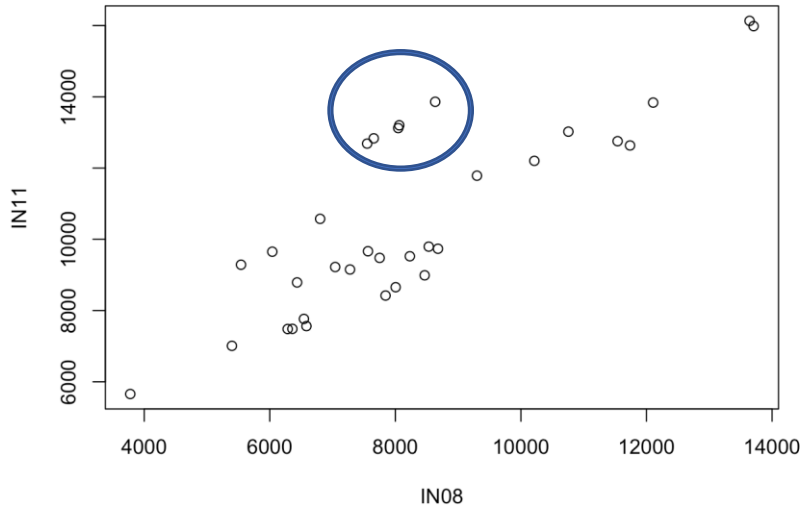
Matt Sirota HW 7

10.16

a.



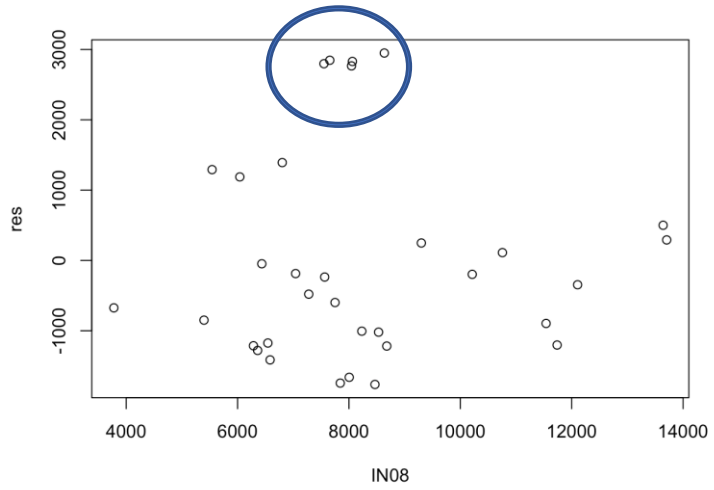
The points go from bottom left to top right of the plot. There is a positive and strong fairly linear relationship for the two sets of data. This does seem reasonable because of inflation and other factors that would cause the tuition to increase at all colleges from 2008 to 2011, which means IN11 is dependent on IN08. The outliers appear to be the five California schools.

b. $\widehat{IN11} = (slope) * IN08 + (intercept)$

$\widehat{IN11} = (0.9429) * IN08 + (2769.184)$

c.

```
> residuals(fit)
         1          2          3          4          5          6          7          8
 291.65696  -896.05773  -345.72970 -1213.20578   248.00095  2766.37666   -47.63828  2797.10392
         9         10         11         12         13         14         15         16
-1174.41256  -848.92537  -674.39725   500.00150  -197.79442   111.16281 -1006.05676  2949.01838
        17         18         19         20         21         22         23         24
 -598.52987  -236.15357 -1019.86464  -479.48893  -186.96846  1187.80064 -1202.74863 -1661.96512
        25         26         27         28         29         30         31         32
-1763.57722 -1217.46849  2846.10116  2829.29053  1391.32354 -1414.12790 -1279.92203  1291.35654
        33
-1744.16090
```



There are 5 points which are much larger than the rest (indexes 6, 8, 16, 27, 28), meaning that the actual values ended up being much larger than the expected values for these 5 colleges.

d. The residuals are not approximately normal with constant variance because, as previously mentioned there appear to be 5 outliers due to their actual tuitions being much higher than the expected values, and due to the way the Normal quantile plot appears.

e. $\widehat{noCal11} = (0.9675) * noCal08 + (2058.3759)$

The slope increased and the intercept decreased upon the exclusion of the 5 California schools

f. It would make the most sense to use the model fit without the five California schools. This is due to the fact that they caused the rest of the data to not appear approximately normal. It is worth noting that these 5 schools are all in the same state so it appears to be a problem with state schools in California rather than state schools as a whole. If these 5 points were taken out, the data would appear more normal, and it would

provide a better view of the rest of the state schools, as California as a whole had a clear unexpected increase in tuition.

10.17

a. $H_0 : \beta_1 = 0$
$H_A : \beta_1 > 0$

$\alpha=0.05$

b. t=13.94
P=1.406 E-13

Reject null hypothesis because the p-value is so small. This means that there is sufficient evidence to conclude that the price of college will increase.

c. 0.82487 to 1.11013

I am 95% confident that the slope is between 0.82487 and 1.11013

This means that tuition increases by a rate between 0.82487 and 1.11013 percent.

d. r=0.93917
r^2=0.88204

88.204% of the variability

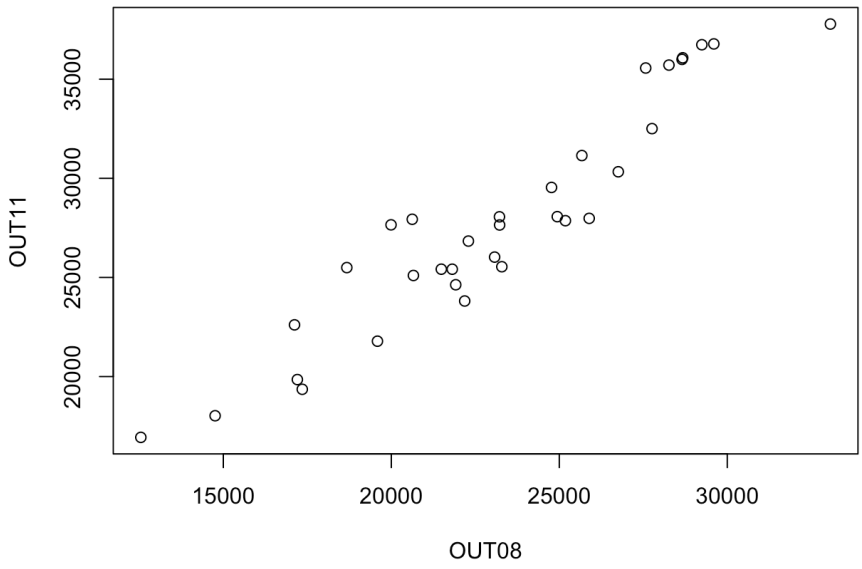
e. There were no schools with a fee of $0 in 2008, so making an inference on β_0 would be extrapolation.

10.18

a. $(0.9675) * 5100 + (2058.3759) = \$\mathbf{6,992.63}$
b. $(0.9675) * 15700 + (2058.3759) = \$\mathbf{17{,}248.13}$
c. The range of tuitions in 2008 was $3,778 to $13,706. State U (5100) is within this range so it makes sense to use the fitted equation for this school. However, Moneypit U (15700) is outside of this range, so it is not appropriate to use the equation to predict its value. It is not known if this school with a higher tuition than all of the others will follow the same pattern.

10.19

a.

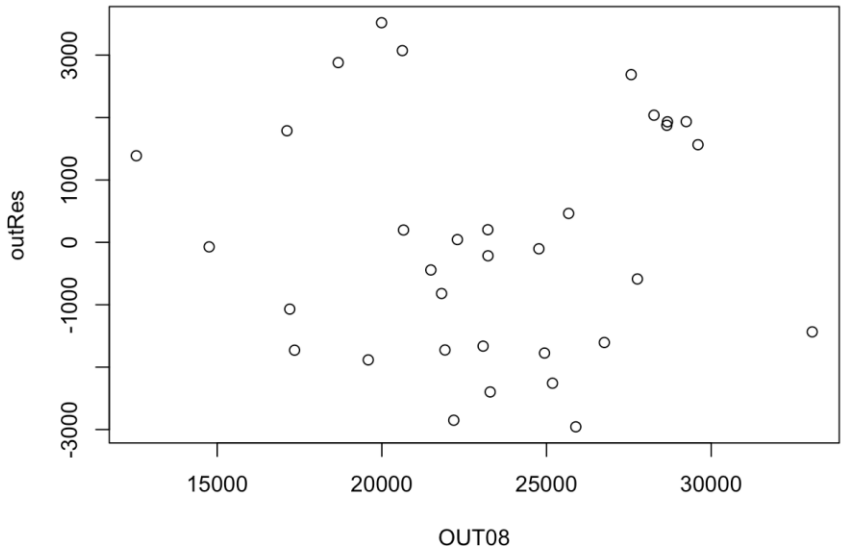


The points go from bottom left to top right and have a positive and strong and fairly linear relationship. This does seem reasonable because of inflation and other factors that would cause the tuition to increase at all colleges from 2008 to 2011, which means OUT11 is dependent on OUT08. There do not appear to be outliers.

b. $\widehat{OUT11} = (slope) * OUT08 + (intercept)$
$\widehat{OUT11} = (1.153) * OUT08 + (1075.073)$

c.

```
> residuals(out)
         1          2          3          4
2880.28960 1387.69932 2037.62439 1934.45948
         5          6          7          8
1875.80384 1931.88951 -1605.06948 3072.64701
         9         10         11         12
-2257.63988 -2956.23718  -103.29145  197.27850
        13         14         15         16
-1728.32053 1788.11164 -1663.45552 -1434.39546
        17         18         19         20
 463.20097  -72.43718 -1883.59846 -1069.92617
        21         22         23         24
  44.18558 -1724.98789 3519.43359 -1774.52046
        25         26         27         28
-2397.43352 -215.31004 -443.03190 -587.06905
        29         30         31         32
-2850.78858 2687.76752 1564.70066  202.45689
        33
 -820.03575
```



Nothing unusual appears in the plot.

d. Yes, the residuals appear to be approximately normal with constant variance. This is due to the fact that the points seem to be scattered randomly and evenly, such that there is not a group of outliers and the schools with large positive residuals are balanced out by schools with large negative residuals.

   It appears that all schools can be used for this analysis due to the fact that it appears to be approximately normal with constant variance because the normal quantile plot follows an approximate straight line.

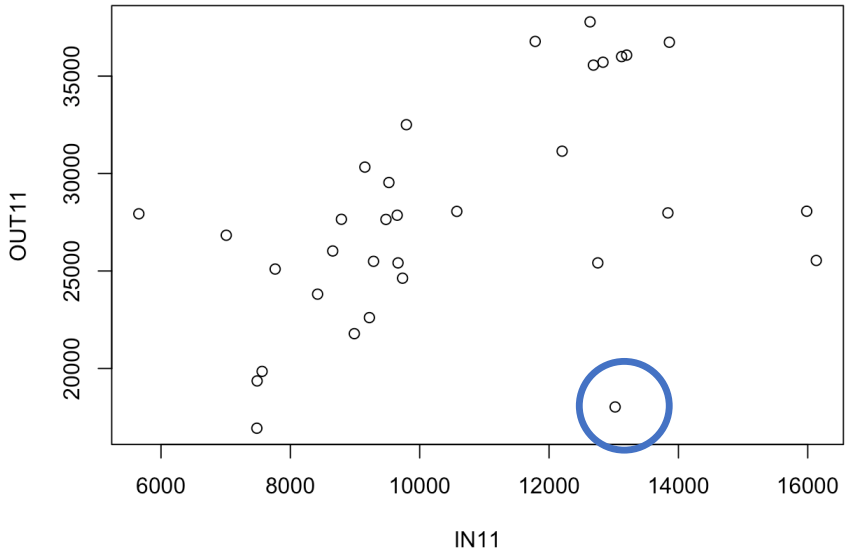
10.20

a. 1.0071 to 1.30

   I am 95% confident that the slope is between 1.0071 and 1.30. This means that the tuition will increase at a rate between 1.0071 and 1.30 percent

b. The two sample t test would not work because this would require two independent samples. These samples in this problem are not independent.

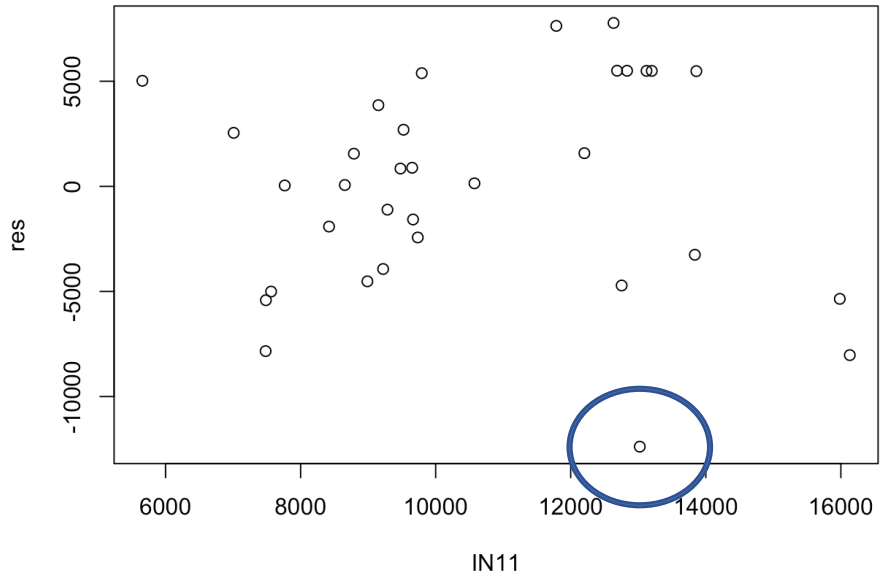
10.21



The points don't follow a defined straight line and the relationship is weak and linear, and positive. Minnesota is an outlier.

$$\widehat{OUT11} = (1.017) * IN11 + (17159.716)$$

```
> res
         1          2          3
-1109.13087 -7838.16116  5497.95973
         4          5          6
 5480.34613  5493.01556  5491.67652
         7          8          9
 3861.16954  5020.16907   884.58591
        10         11         12
-3255.27619  2692.78332    40.98051
        13         14         15
-5416.22983 -3932.03217    62.70166
        16         17         18
 7771.39318  1576.80943 -12383.26771
        19         20         21
-4519.03220 -5004.55171  2544.95895
        22         23         24
-2431.83897  1553.40125 -5352.11710
        25         26         27
-8028.65785   845.57301 -4716.66689
        28         29         30
 5384.14816 -1914.28121  5500.50048
        31         32         33
 7631.95101   142.75771 -1575.63726
```



The only unusual thing about the plot is the outlier, Minnesota. The normal quantile plot also looks like it follows an approximately normal distribution.

In conclusion, the relationship between IN11 and OUT11 is a weak linear and positive relationship. The equation of predicted values is seen to equal 1.017, which is the slope, multiplied by a given IN11 value, and then added to the intercept, which is 17159.716. It evident that the residuals are approximately normal with constant variance from the IN11 vs residual plot, as well as the normal plot. Minnesota is the outlier of the set of data, but other than that there does not seem to be any problems with the data.