FA541 Applied Statistics with Applications in Finance

Humphrey De Guzman

11/5/2021

Contents

Page 602

Problem 1

10.16 Public university tuition: 2008 versus 2011 Table 10.1 shows the in-state undergraduate tuition and required fees for 33 public universities in 2008 and 2011.

(a) Plot the data with the 2008 in-state tuition (IN08) on the x axis and the 2011 tuition (IN11) on the y axis. Describe the relationship. Are there any outliers or unusual values? Does a linear relationship between the in-state tuition in 2008 and in 2011 seem reasonable?

(b) Run the simple linear regression and state the least squares regression line.

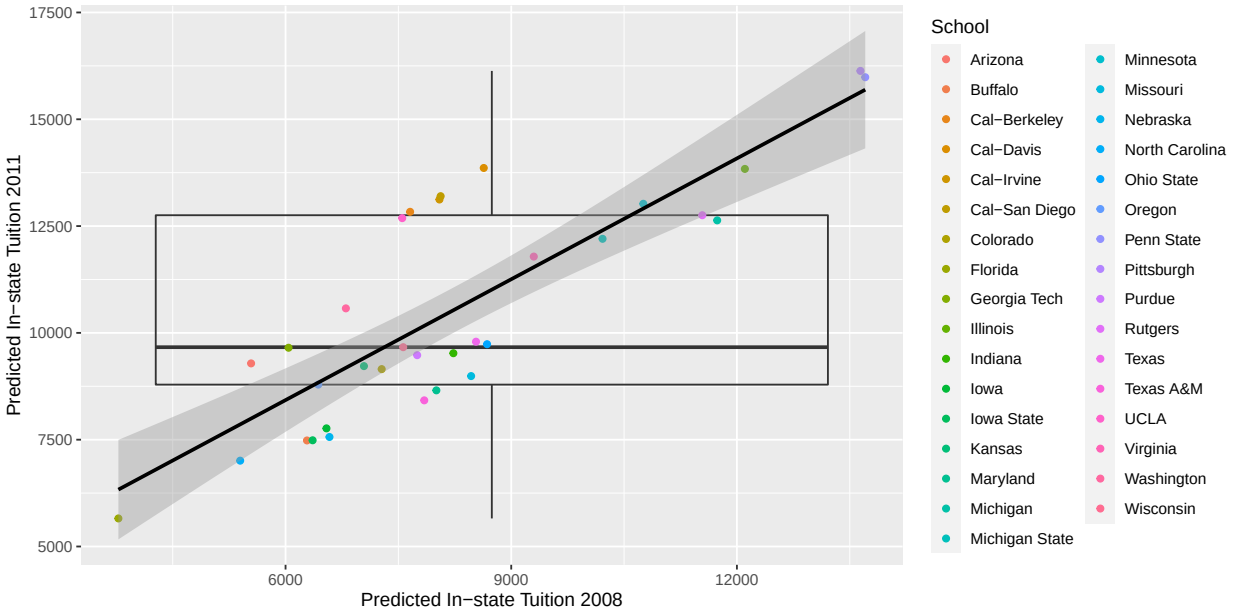
(c) Obtain the residuals and plot them versus the 2008 in-state tuition amounts. Describe anything unusual in the plot.

(d) Do the residuals appear to be approximately Normal with constant variance? Explain your answer.

(e) The 5 California schools appear to follow the same linear trend as the other schools but have higher-than predicted in-state tuition in 2011. Assume that this jump is particular to this state (financial troubles?), and remove these 5 observations and refit the model. How do the model parameters change?

(f) If you were to move forward with inference, which of these two model fits would you use? Write a short paragraph explaining your answer.

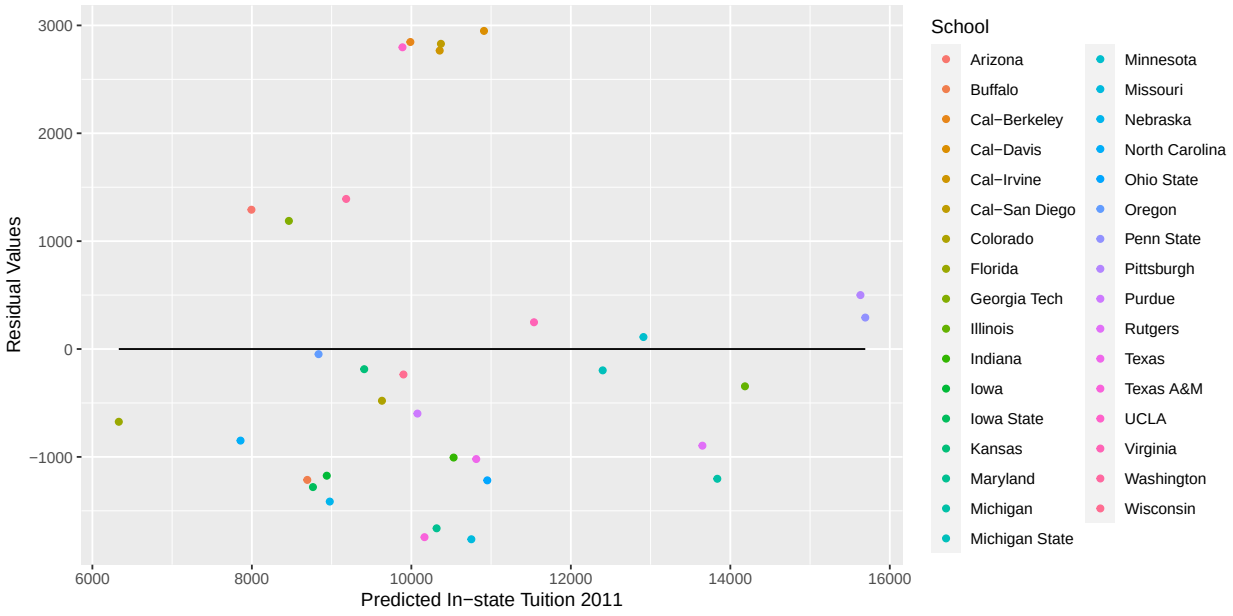
A) Below we can see the data. The graph looks like a weak linear relationship, with many points being outside of the R^2 area, especially towards the middle. While the relationship is weak, linear would be a suitable descriptor. Looking at the boxplot, we can see that there are no statistical outliers.



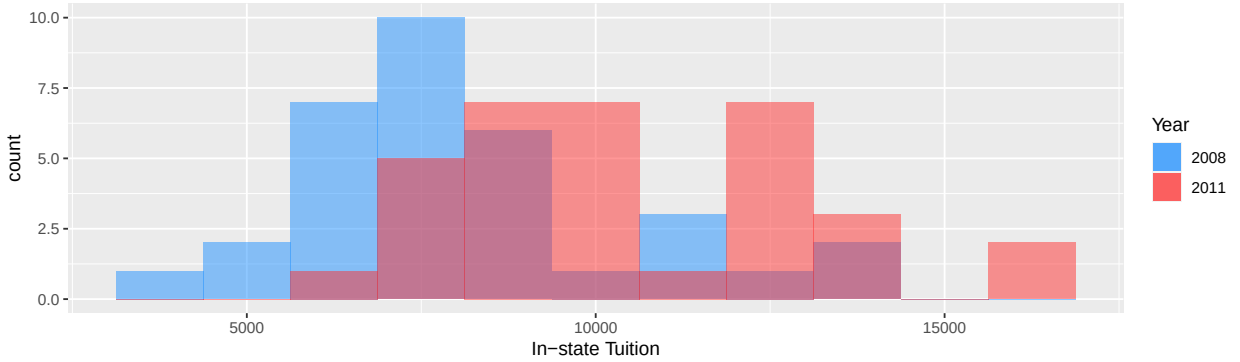
B) The regression line has a formula of $(0.9429)x_{In08} + 2769.184$, along with its summary below. Both values have significant p-values

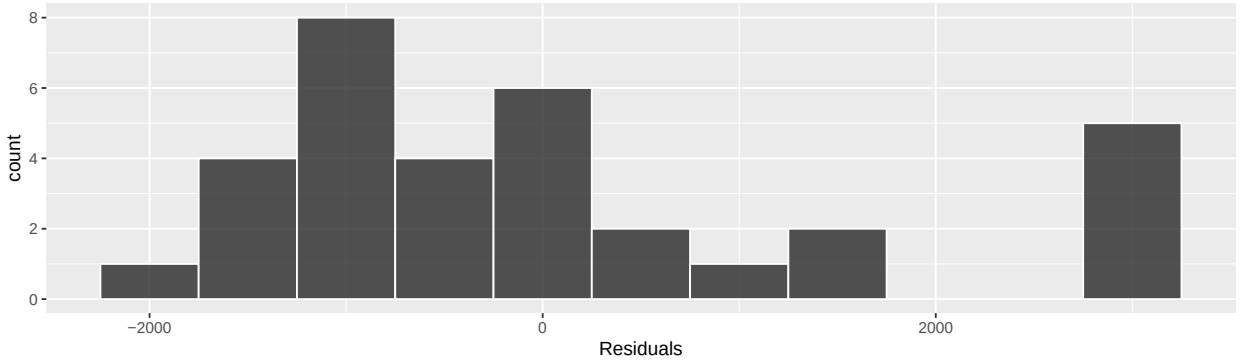
Column	Beta	Squared Error	t-statistic	p-value
(Intercept)	2769.1839888	973.8632055	2.843504	0.0078297
In08	0.9428833	0.1138333	8.283022	0.0000000

C) The residual plot looks a little odd, there is obvious asymmetry along with a lack of centering. However, there is no clear pattern which is always good to see.



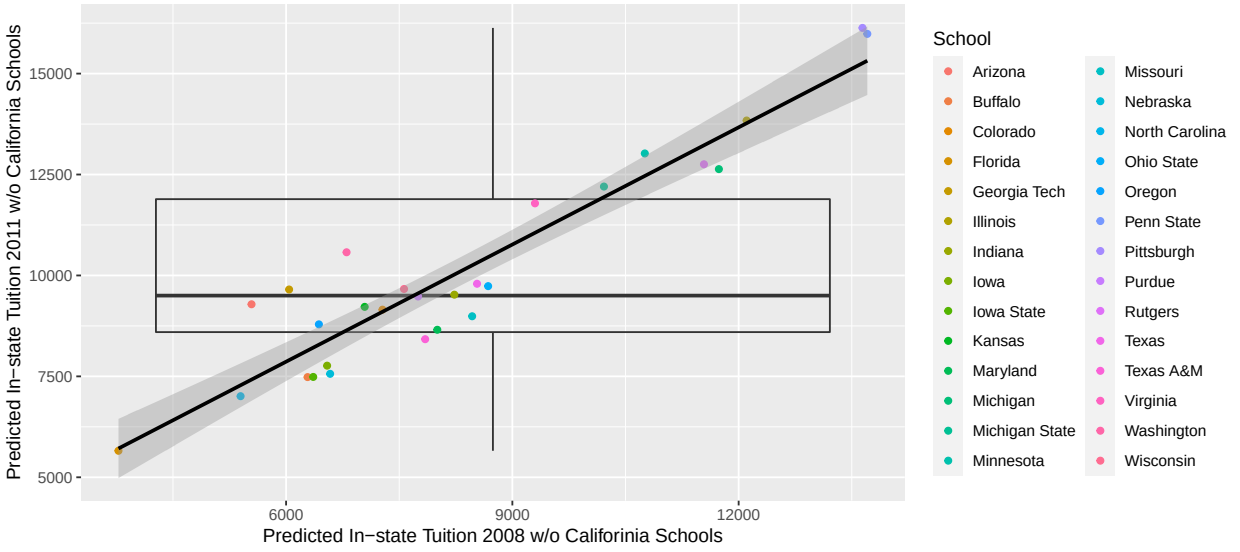
D) The residuals do not appear to be normal, nor does the variance look constant. Notice on the graph below how the graph's center and spread both change between the two histograms. It is important to note that if some values were removed, the residuals do look normal however.





E) Here we can see a much better fitting regression, especially seen in the middle. The formula is now $(0.9675)x_{noCali} + 2058.376$

Column	Beta	Squared Error	t-statistic	p-value
(Intercept)	2058.3759444	600.2582395	3.429151	0.0020298
In08	0.9674987	0.0693883	13.943249	0.0000000



F) Moving forward I would use the second model with the removal of the five California schools. This is mostly due to the fact that because California had experienced external troubles that were not captured by the features we had studied, it serves as an "outlier" to our data. To be more specific, it introduces an exogenous break in our model whose information is not captured but is a covariate to our features (latent variable). That being said, the model would not be able to predict California schools well and should be stipulated in its use.

Problem 2

10.17 More on public university tuition. Refer to the previous exercise. We'll now move forward with inference using the model fit you chose in part (f) of the previous exercise.

(a) Give the null and alternative hypotheses for examining the linear relationship between 2008 and 2011 in-state tuition amounts.

(b) Write down the test statistic and P-value for the hypotheses stated in part (a). State your conclusions.

(c) Construct a 95% confidence interval for the slope. What does this interval tell you about the annual percent increase in tuition between 2008 and 2011?

(d) What percent of the variability in 2011 tuition is explained by a linear regression model using the 2008 tuition?

(e) Explain why inference on b0 is not of interest for this problem.

A) This is simply $H_0 : \beta = 0$ and $H_a : \beta > 0$, or to put simply, the tuition increases from the year 2008 to 2011.

B) From the table earlier we saw a t-statistic of roughly 13.94 and a p-value $<$ 10e7, which is extremely significant. This means we can reject the null hypothesis and accept that tuition increases.

C) We use a degree of freedom of 27 for our calculation. This would be a range of $0.9675 \pm 2.052(0.06939)$. This translate into a change of (-17.49%, 10.99%) over the course of three years, dividing by three we get (-5.83%,3.66%) annual change. This is taken from values of 0.8251 and 1.1099, note that if we multiply by a number less than 1, we are actually reducing our value.

D) This is simply the R^2 value which is, 88.2%, if we use the adjusted R^2 then it is 87.75%.

E) Inference of B_0 is not important as this simply tells us the minimum value of a tuition jump from 2008 to 2011 given tuition is \$0 USD, which does not really translate well as this is not a feasible occurrence.

Page 603

Problem 3

10.18 Even more on public university tuition. Refer to the previous two exercises

(a) The in-state tuition at State U was \$5100 in 2008. What is the predicted in-state tuition in 2011?

(b) The in-state tuition at Moneypit U was \$15,700 in 2008. What is its predicted in-state tuition in 2011?

(c) Discuss the appropriateness of using the fitted equation to predict tuition for each of these universities.

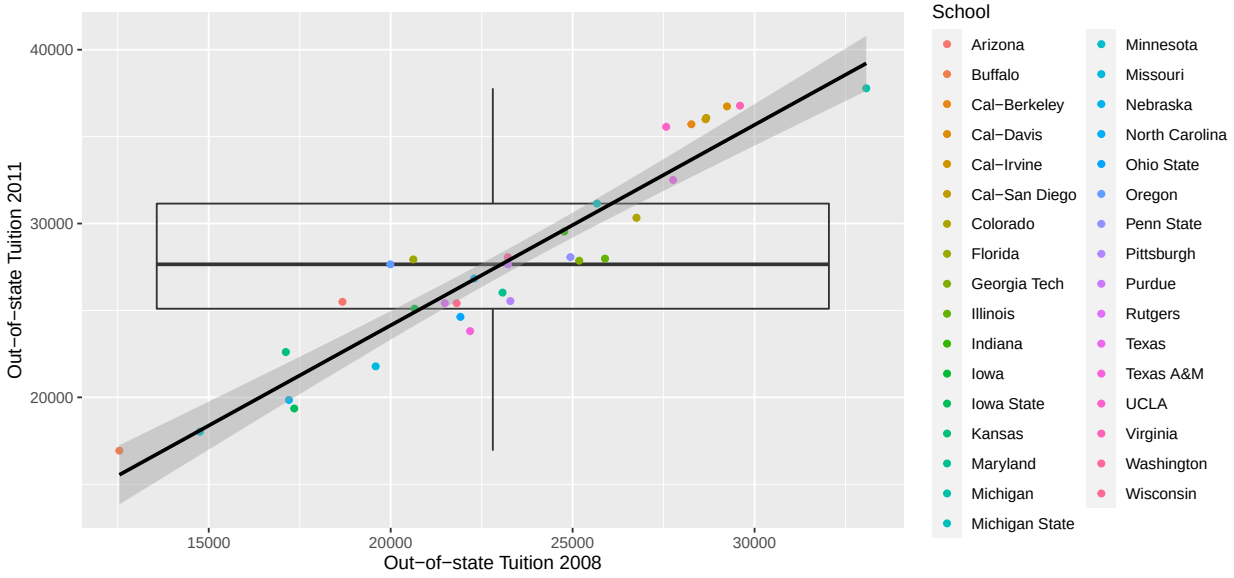
A using our model, the predicted 2011 tuition is to be $6992.62 roughly. We find this by simply plugging the value of 5100 into our linear equation.

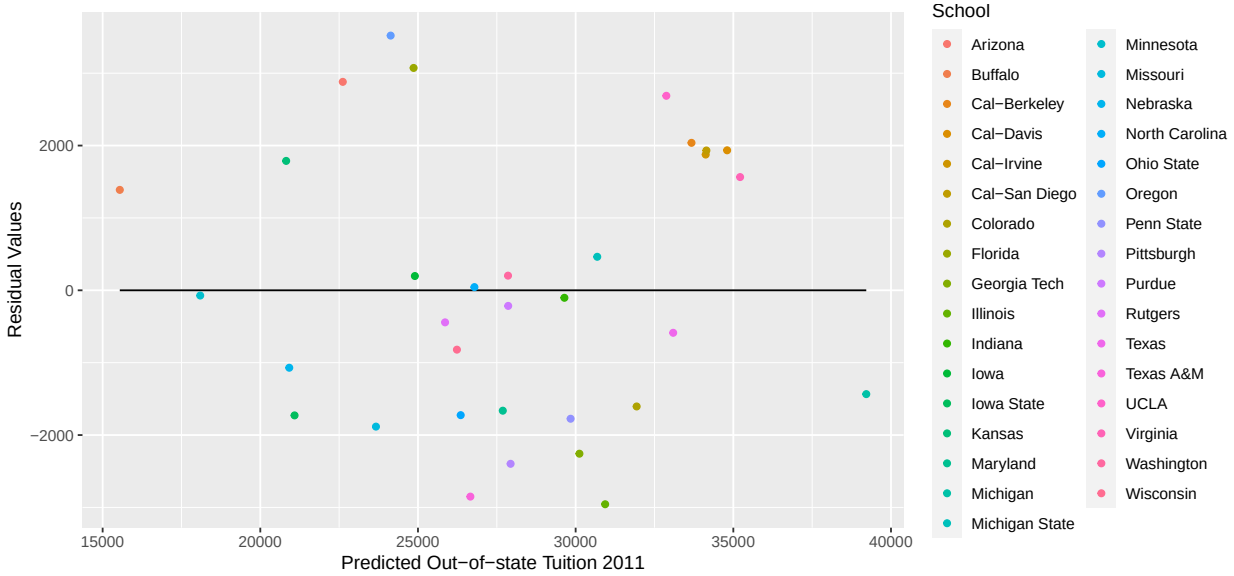
B Again we do something similar, the value this time is $17248.11 rounded up.

C The previous values calculated are not the true values of the future tuition and are at the will of the model's variance. More importantly, if any of the two hypothetical universities were from the state of California, the model would actually do a poor job in predicting said tuition(s). If the two universities are from the United States but not specifically from the state of California, I believe this model would accurately predict the tuition given some variance.

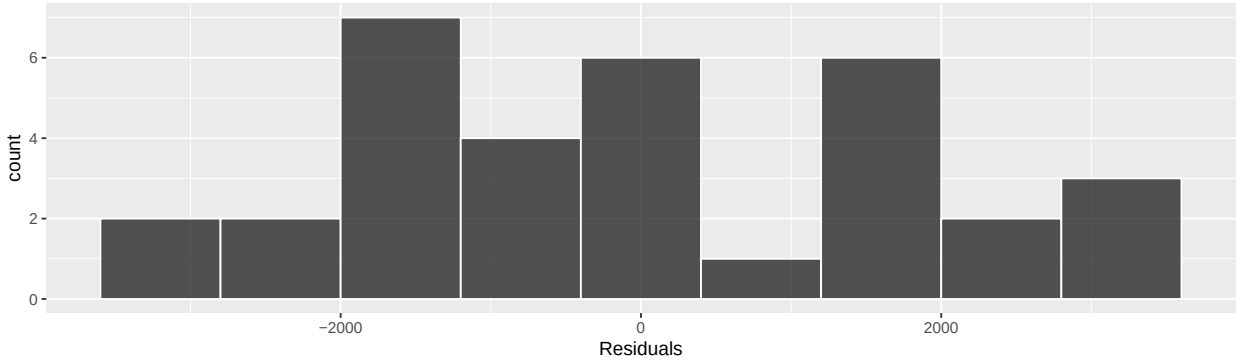
Problem 4

10.19 Out-of-state tuition. Refer to Exercise 10.16. In addition to in-state tuition, out-of-state tuition for 2008 (OUT08) and 2011 (OUT11) was also obtained. Repeat parts (a) through (d) of Exercise 10.16 using these tuition rates. Does it appear we can use all the schools for this analysis or are there some unusual observations? Explain your answer.





Column	Beta	Squared Error	t-statistic	p-value
(Intercept)	1075.073002	1699.8594200	0.6324482	0.5317316
Out08	1.153386	0.0717496	16.0751546	0.0000000



Answer) From the visualization there does not seem to be any schools that need to be removed from the pool, with the residual plot looking roughly centered around the middle. Investigating closer, we can see that our residuals do not look normally distributed and have several peaks. If we also look at the table, we can notice that our intercept is not significant, though that is not much of a problem. Again our B_x has a low p-value, which means this model *can* be used and has some merit.

Problem 5

10.20 More on out-of-state tuition. Refer to the previous exercise

(a) Construct a 95% confidence interval for the slope. What does this interval tell you about the annual percent increase in out-of-state tuition between 2008 and 2011?

(b) In Exercise 10.17(c) you constructed a similar 95% confidence interval for the annual percent increase in in-state tuition. Suppose that you want to test whether the increase is the same for both tuition types. Given the two slope estimates b1 and standard errors, could we just do a variation of the two-sample t test from Chapter 7? Explain why or why not.

A) This is equal to roughly $1.1534 \pm (2.03693)0.071745$ using $df = 32$. I found this using R's `qtnorm()` function. Translating this range, it is equal to (1.00726,1.29954) which means that in a three year interval, we are 95% confident that tuition will increase by 0.726% - 29.95%. For an annual base. we can simply divide by three to get a range of (0.242%, 9.98%) annual increase in tuition.

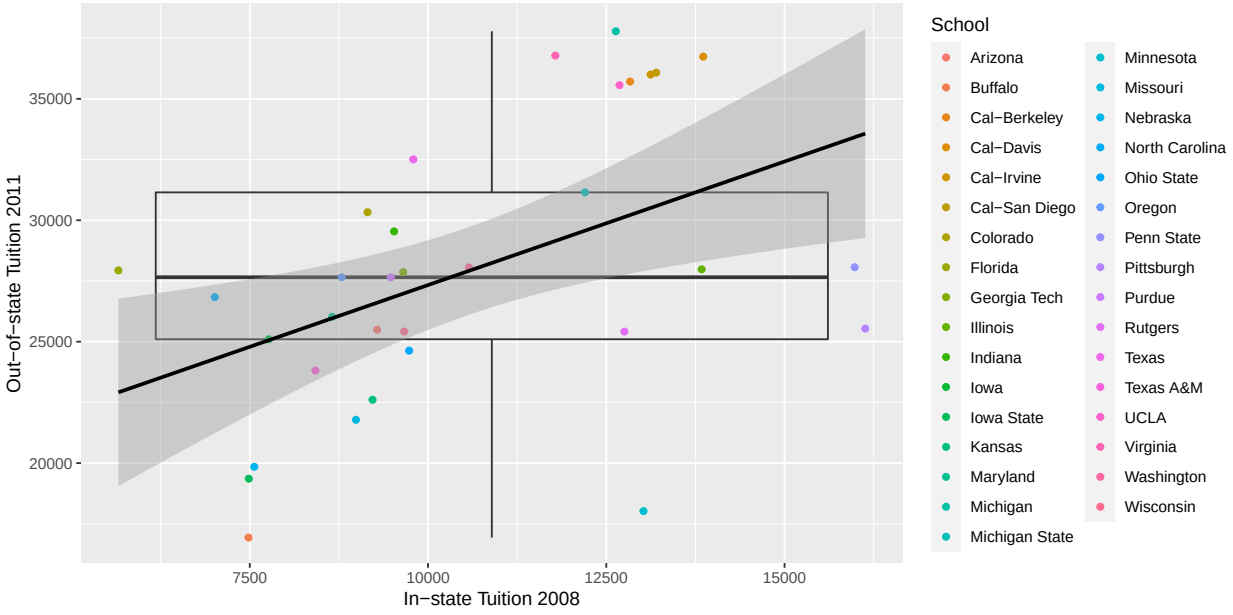
B) While we should be able to calculate a t-test on some normally distributed data and that the difference between two normally distributed variables retains normality, there must be some large assumptions to make in order for this to work. Specifically, because we are using some standard error rather than standard deviation, we have to make the assumption that the standard error calculated is in some way indicative of the population standard deviation for the distributions we are using in our t test. So to put simply, it is possible only when our standard error is reflective of the true population standard deviation.

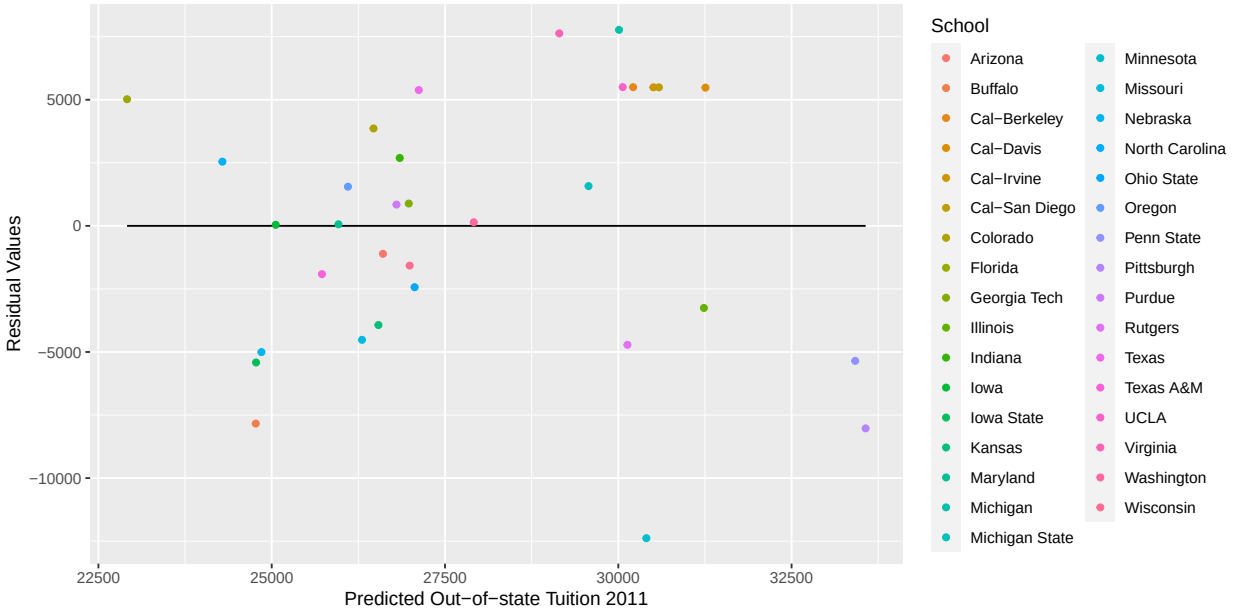
Problem 6

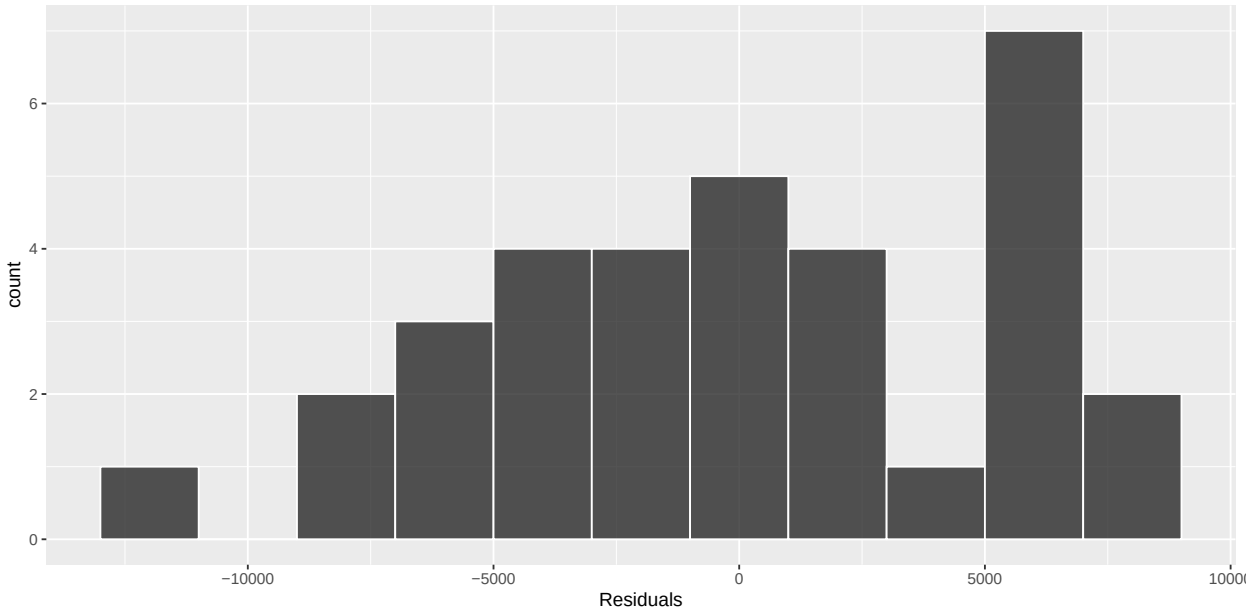
10.21 In-state versus out-of-state tuition. Refer to the previous five exercises. We can also investigate whether there is a linear association between the in-state and out-of-state tuition. Perform a linear regression analysis using the 2011 data, complete with scatterplots and residual checks, and write a paragraph summarizing your findings.

Answer) The relationship is indeed linear. We see that the formula is $(1.01717)x + 17159.7158$ with both coefficients having significant p-values. The residual plot looks quite poor, with there being a clear pattern forming a negative quadratic relationship. We also do not see the residuals being normally distributed, with again a spike towards the higher end of the histogram. If these values were to be removed we may see a difference in the regression/residuals. The histogram is of the residuals is on the following page.

Column	Beta	Squared Error	t-statistic	p-value
(Intercept)	17159.715811	3713.8325050	4.620487	0.0000635
In11	1.017167	0.3420526	2.973716	0.0056524






count
6
4
2
0
−10000
−5000
0
5000
Residuals