



One measure of a university's effectiveness is how many of its students graduate. Universities can try to improve rates of completion by offering services, such as advising, reduced class size, and better faculty; adjusting incentives, such as cost to students; and by selective admission, such as percent of students admitted. You can investigate some of these factors, because the federal government made this data available here is [a simplified version of the data](#)  . Which of the following best predicts the percent of university students who graduate in four years?:

Admission rate (admrate)

Average cost per year (costperyr)

Average SAT scores (satavg)

Use simple linear regression models to evaluate this question. Report **all** relevant statistical results. Calculate and compare effect sizes as is appropriate. Check assumptions. Interpret the results in light of statistical results, effect sizes, and limitations. Describe the relationship between each predictor and the outcome. Create three scatterplots one for each predictor to display the relationship between the predictor and outcome variable.

Robert Bruffey
INST314

Project 4: Linear Regression

I. Step 0:

Goal: To determine which of the following (Admission rate, Average cost per year, or average SAT scores) best predicts the percent of university students who graduate in four years.

Prediction: I believe that admission rate will have the best percent of university students who graduate in four years because if the school doesn't admit that many students then it is most likely going to accept the best of the bunch. This means they would leave out those who are more likely to not graduate in four years.

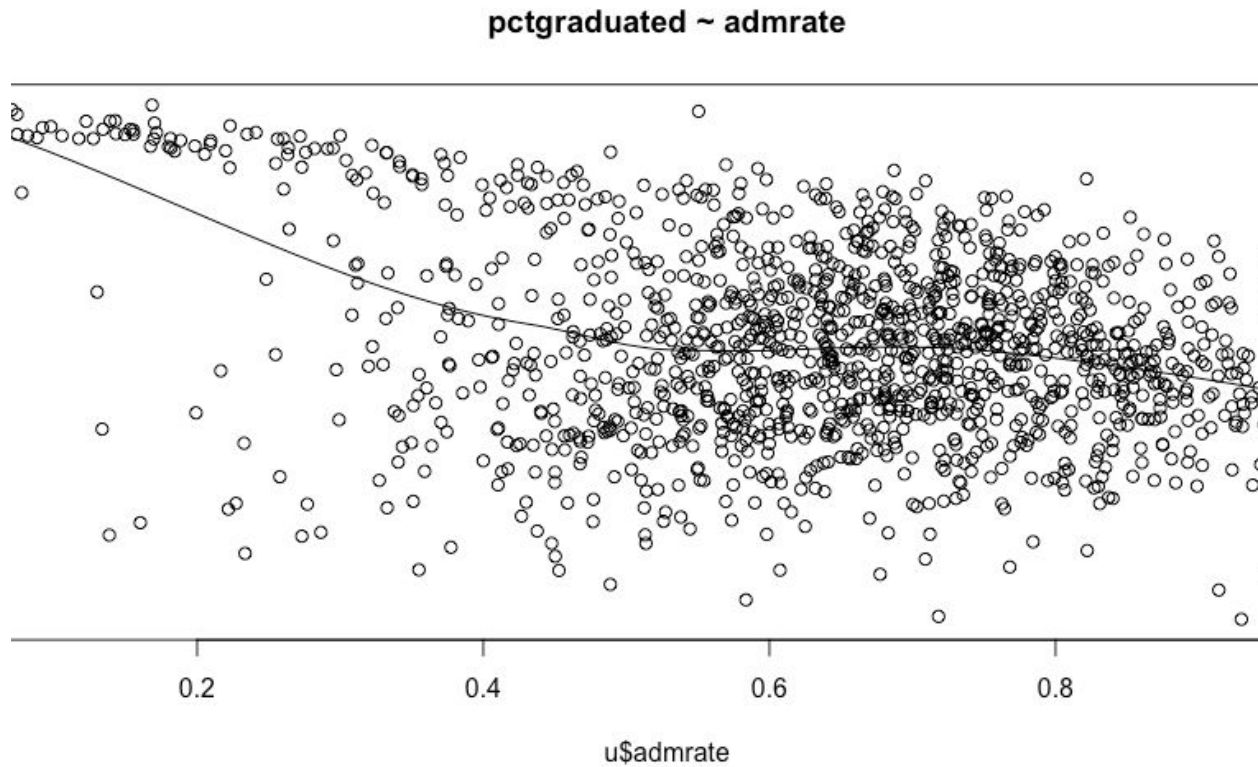
Load Data:

```
u = read.csv("~/Downloads/univdata.csv")  
View(u)
```

II. Step 1: Plot data

- **admrate vs. pctgraduated**

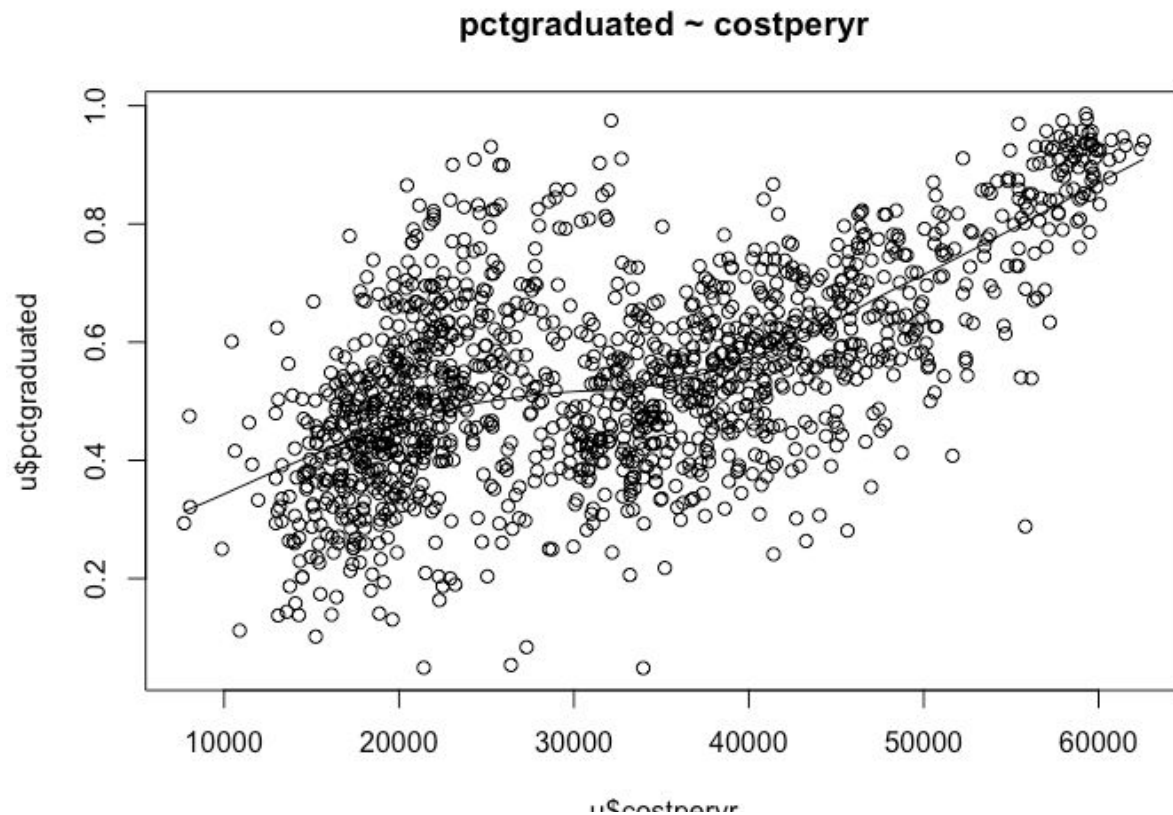
R code: `plot(u$admrate,u$pctgraduated)`



From the looks of the plot graph, we can see some kind of negative linear regression. It seems as admrate goes up, the pctgraduated decreases.

- **costperyr vs. pctgraduated**

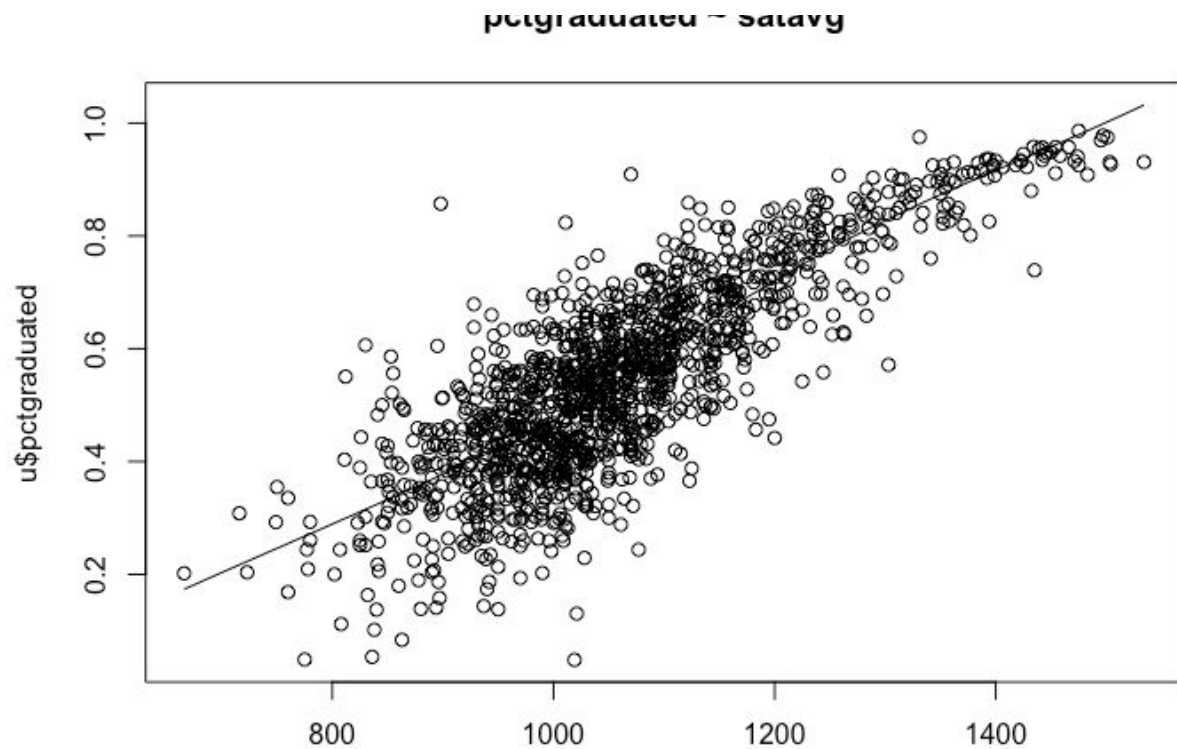
R Code: `plot(u$costperyr,u$pctgraduated)`



From the plot graph above, we can see a linear relationship between pctgraduated and costperyr in that as the costperyr increases, the pctgraduated also seems to increase.

- **satavg vs. pctgraduated**

R Code: `plot(u$satavg,u$pctgraduated)`

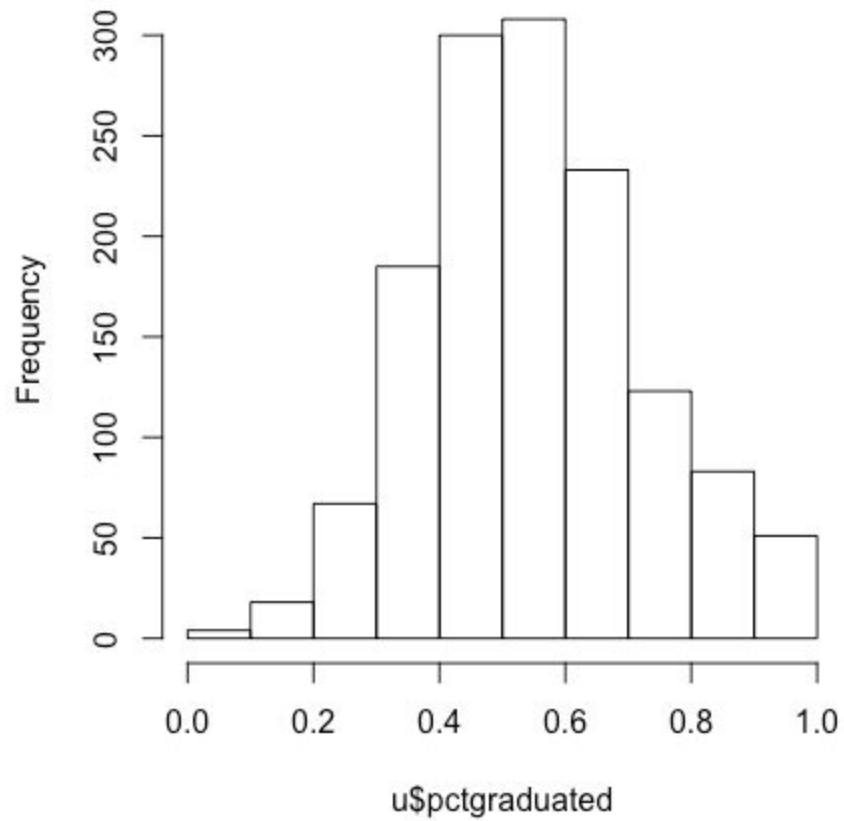


Here we can see a somewhat linear plot graph between satavg and pctgraduated. From the looks of it, as the satavg goes up the pctgraduated also goes up.

Step 2: Check for Assumptions

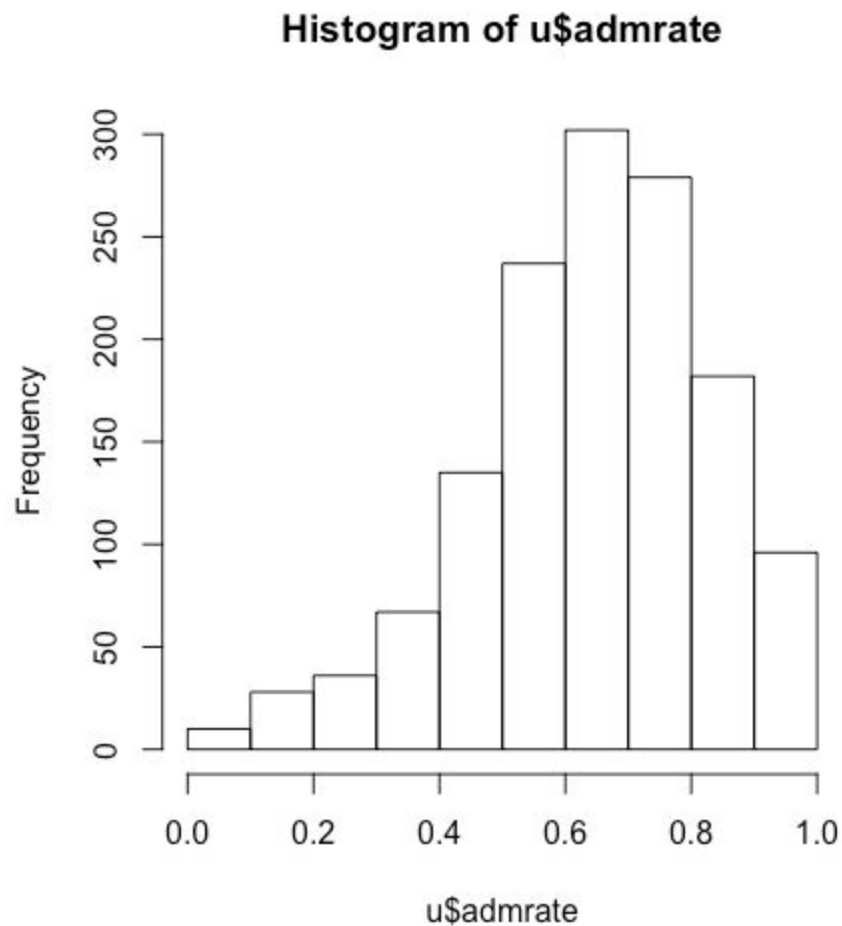
- **Admrate vs. pctgraduated**
 - Linear relationship between admrate and pctgraduated
 - Earlier we noted a somewhat linear relationship from the scatterplot
 - Admrate and pctgraduated are approximately normal
 - Pctgraduated histogram

Histogram of u\$pctgraduated



The distribution of pctgraduated appears to be normally distributed with a bell shape.

- Admrate histogram



Slightly skewed left but still holds a normal distribution or bell shape.

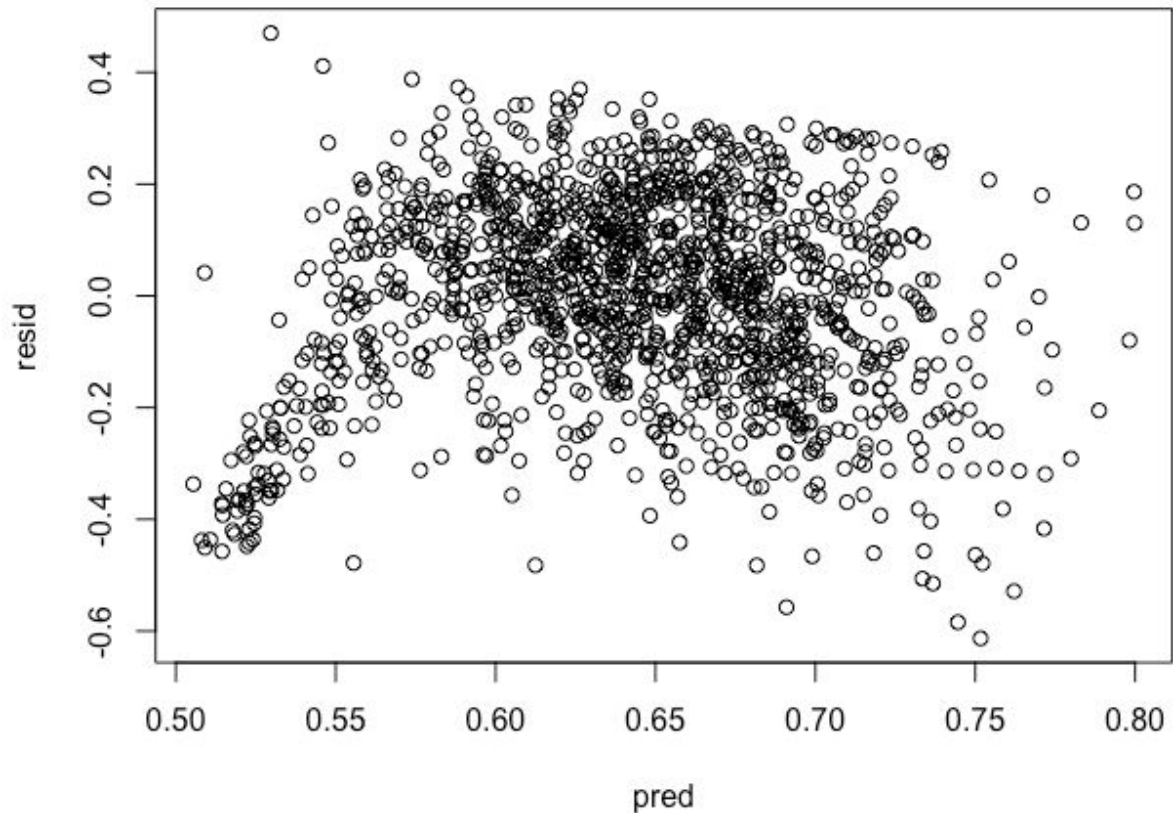
- Admrate and pctgraduated are interval or ratio
 - Admrate = ratio
 - Pctgraduated = ratio
- No outliers: there are actually 3 outliers present
 - R Code:


```
m = lm(admrate ~ pctgraduated,data = u)
pred = m$fitted.values
resid = m$residuals
resid.sd = sd(resid)
resid[abs(resid)>=3*resid.sd]
```

174	604	751
-0.5573541	-0.6130579	-0.5842643

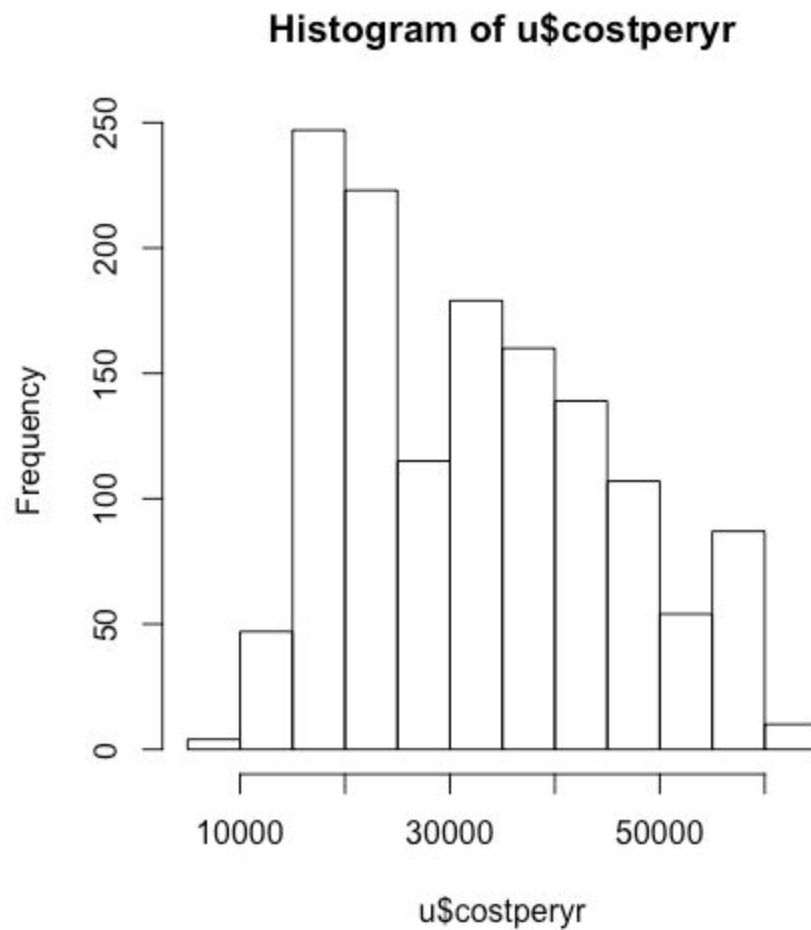
#Outliers
- Errors are constant

- plot(pred, resid)



The errors appear to be all over the place, but most tend to be in the center of the graph.

- Errors are independent
 - Every school is different and has their own admrate and pctgraduated.
- **Costperyr vs. pctgraduated**
 - Linear relationship between Costperyr and pctgraduated
 - Earlier we noted a somewhat linear relationship from the scatterplot
 - Costperyr and pctgraduated are approximately normal
 - costperyr histogram



Doesn't look too normally distributed, but the data allows for some lee-way.

- Costperyr and pctgraduated are interval or ratio
 - Costperyr = ratio

- No outliers == 1 outlier present

```
m = lm(costperyr ~ pctgraduated, data = u)
```

```
pred = m$fitted.values
```

```
resid = m$residuals
```

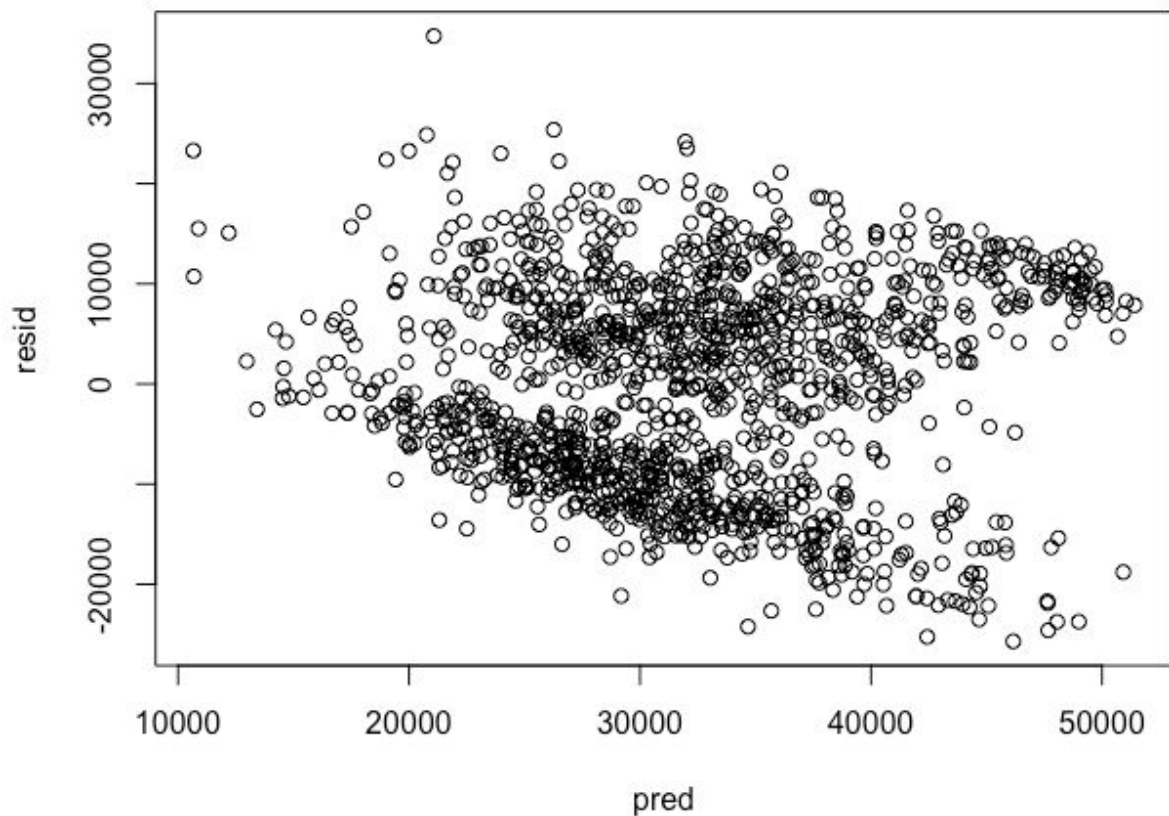
```
resid.sd = sd(resid)
```

```
resid[abs(resid)>=3*resid.sd]
```

```
105
```

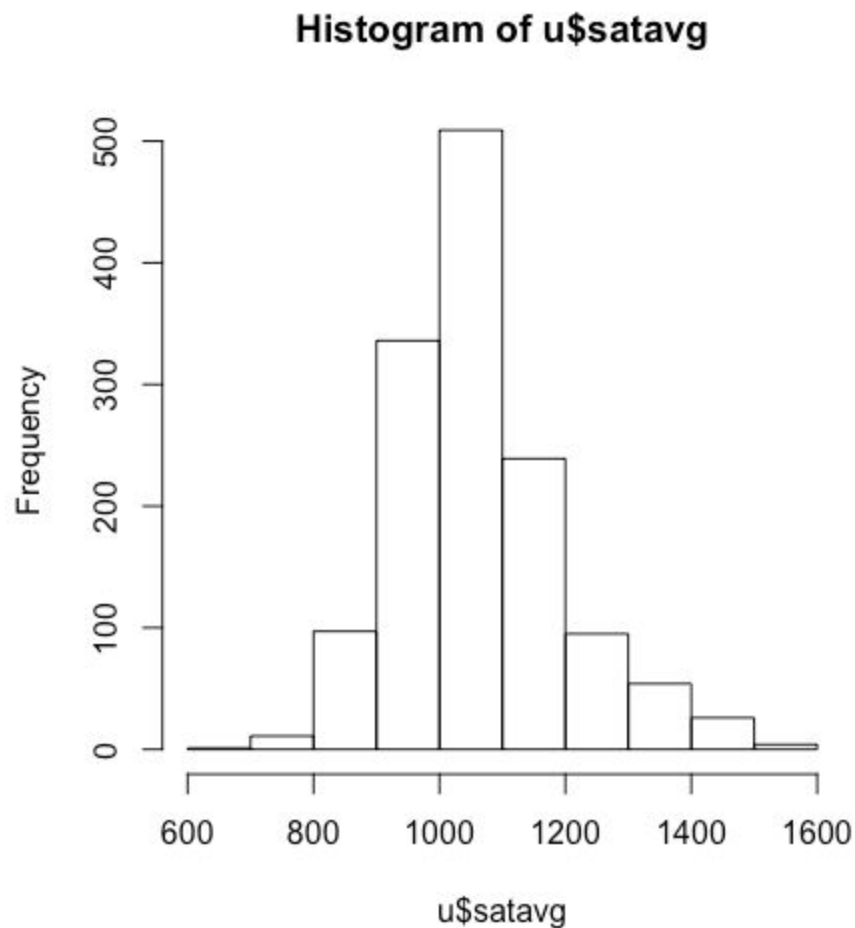
```
34734.46
```

- Errors are constant



Errors tend to be everywhere

- Errors are independent
 - Schools cant share the same costpryear and pctgraduated.
- **Satavg vs. pctgraduated**
 - Linear relationship between Satavg and pctgraduated
 - Earlier we noted a somewhat linear relationship from the scatterplot
 - Satavg and pctgraduated are approximately normal
 - Satavg histogram



Appears normally distributed and bell shaped.

- Satavg and pctgraduated are interval or ratio
 - Satavg = interval
- No outliers == 8 outliers

```
m = lm(satavg ~ pctgraduated, data = u)
```

```
pred = m$fitted.values
```

```
resid=m$residuals
```

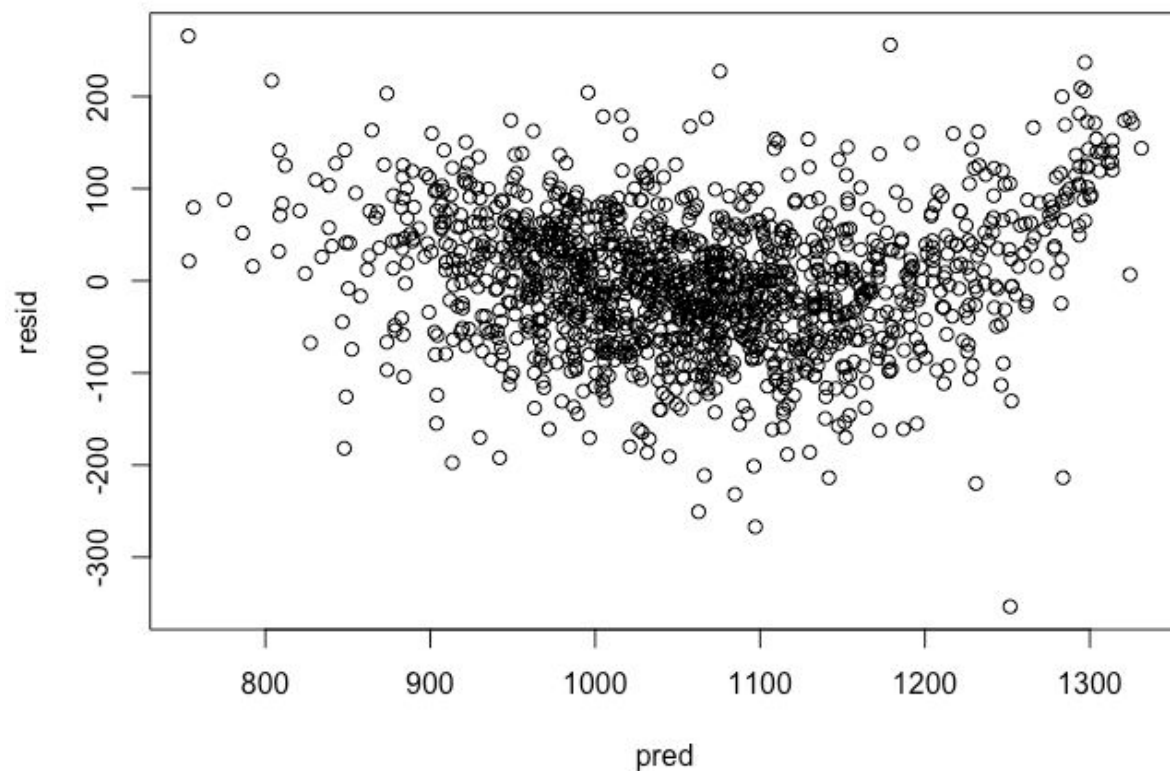
- > resid.sd = sd(resid)
- > resid[abs(resid)>=3*resid.sd]

```

51      137      162      261      766      809
236.9811 -267.0981 -250.7097 265.7238 227.4101 256.0601
853     1276
-231.7109 -353.6608

```

- Errors are constant



Errors seem to have equal residuals throughout

- Errors are independent
- One error can't be for more than satavg for a school.

Step 3: Run the test

- **Admrate vs. pctgraduated**

```
m = lm(u$pctgraduated ~ u$admrate)
```

```
summary(m)
```

Call:

```
lm(formula = u$pctgraduated ~ u$admrate)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.49167	-0.11627	0.00348	0.12156	0.45941

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.72429	0.01618	44.77	<2e-16
u\$admrate	-0.27459	0.02416	-11.37	<2e-16

(Intercept) ***
u\$admrate ***

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1671 on 1370 degrees of freedom

Multiple R-squared: 0.08619, Adjusted R-squared: 0.08552

F-statistic: 129.2 on 1 and 1370 DF, p-value: < 2.2e-16

cor.test(u\$admrate,u\$pctgraduated)

Pearson's product-moment correlation

data: u\$admrate and u\$pctgraduated

t = -11.367, df = 1370, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.3412003 -0.2444546

sample estimates:

cor

-0.293579

- Costperyr vs. pctgraduated

cost = lm(costperyr ~ pctgraduated,data = u)

summary(cost)

Call:

lm(formula = u\$pctgraduated ~ u\$costperyr)

Residuals:

Min	1Q	Median	3Q	Max
-0.51224	-0.09485	-0.00492	0.09298	0.44003

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.878e-01	1.032e-02	27.89	<2e-16 ***

```
u$costperyr 8.033e-06 2.964e-07 27.10 <2e-16 ***
```

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.141 on 1370 degrees of freedom

Multiple R-squared: 0.349, Adjusted R-squared: 0.3485

F-statistic: 734.5 on 1 and 1370 DF, p-value: < 2.2e-16

```
cor.test(u$costperyr,u$pctgraduated)
```

Pearson's product-moment correlation

data: u\$costperyr and u\$pctgraduated

t = 27.102, df = 1370, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.5552140 0.6241846

sample estimates:

cor

0.5907775

- Satavg vs. pctgraduated

```
sat = lm(satavg ~ pctgraduated,data = u)
```

```
summary(sat)
```

Call:

```
lm(formula = u$pctgraduated ~ u$satavg)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.45332	-0.06331	0.00564	0.06559	0.48790

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-6.143e-01	2.195e-02	-27.98	<2e-16 ***
u\$satavg	1.095e-03	2.053e-05	53.34	<2e-16 ***

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.09964 on 1370 degrees of freedom
Multiple R-squared: 0.6749, Adjusted R-squared: 0.6747
F-statistic: 2845 on 1 and 1370 DF, p-value: < 2.2e-16

```
cor.test(u$satavg,u$pctgraduated)
      Pearson's product-moment correlation
data: u$satavg and u$pctgraduated
t = 53.335, df = 1370, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8035644 0.8380352
sample estimates:
      cor
0.8215492
```

Summary:

1) Admrate Results

- a) From our linear model, we get an R-Squared value of 0.08619 and an Adjusted R-squared value of 0.08552. Typically, these values are low and that means there is little variation in the data for admrate to pctgraduated. Next, we have a F-statistic of 129.2 for the admrate, which is the lowest out of the bunch. The std. Error for admrate is 0.02416 and the std. Error for the intercept is .01618, which are both low, but still the highest out of the other two independent variables. Also, the pearson corellation for our data is -0.293579, which is a weak negative correlation for our data. As, the admrate goes up, the pctgraduated tends to decrease. Finally, our p-value is <2e-16, which is well below our significance level of .05, so our data is considered significant and that the means of admrate and pctgraduated are not equal. Overall, admrate is not the best way to predict the percent of university students who graduate in four years.

2) Costperyr Results

- a) From our linear model, we get an R-Squared value of .349 and an Adjusted R-squared value of .3485. These values are not too low and that means there is some variation in the data for costperyr to pctgraduated. Next, we have a F-statistic of 734.5 for the costperyr, which is the middle value out of the bunch. The std. Error for costperyr is 2.964e-07 and the std. Error for the intercept is 1.032e-02, which are both low and the lowest out of all the

independent variables. Also, the pearson correlation for our data is .5907775 which is a somewhat strong positive correlation for our data. As, the costperyr goes up, the pctgraduated tends to increase. Finally, our p-value is $<2e-16$, which is well below our significance level of .05, so our data is considered significant and that the means of costperyr and pctgraduated are not equal. Overall, costperyr is not the best way to predict the percent of university students who graduate in four years.

3) Satavg Results

- a) From our linear model, we get an R-Squared value of .6749 and an Adjusted R-squared value of .6747. These values are the highest out of the other independent variables tested, which means satavg has the highest variation in its data. Next, we have a F-statistic of 2845 for the satavg, which is the highest value out of the bunch. The std. Error for satavg is $2.053e-05$ and the std. Error for the intercept is $2.195e-02$, which are both low. Also, the pearson correlation for our data is .8215492 which is a strong positive correlation for our data. As, the satavg goes up, the pctgraduated tends to increase. Finally, our p-value is $<2.2e-16$, which is well below our significance level of .05, so our data is considered significant and that the means of satavg and pctgraduated are not equal. **Overall, I would say that satavg is our best choice to choose that best predicts the percent of university students who graduate in four years.**

Limitations

- Overall, there aren't too many limitations for this data set. One limitation could be that the admrate for a school could be high, but they don't get many applications so in reality they are not taking in that many students, which could cause them to have a low pctgraduated to a high admrate. Also, some schools are private and others are public and they can be extremely different settings. If we could categorize our data into a private school section and then a public school section that would give us a better understanding of our data. Also, another independent variable that we could have tested that I believe has an effect would be GPAs.