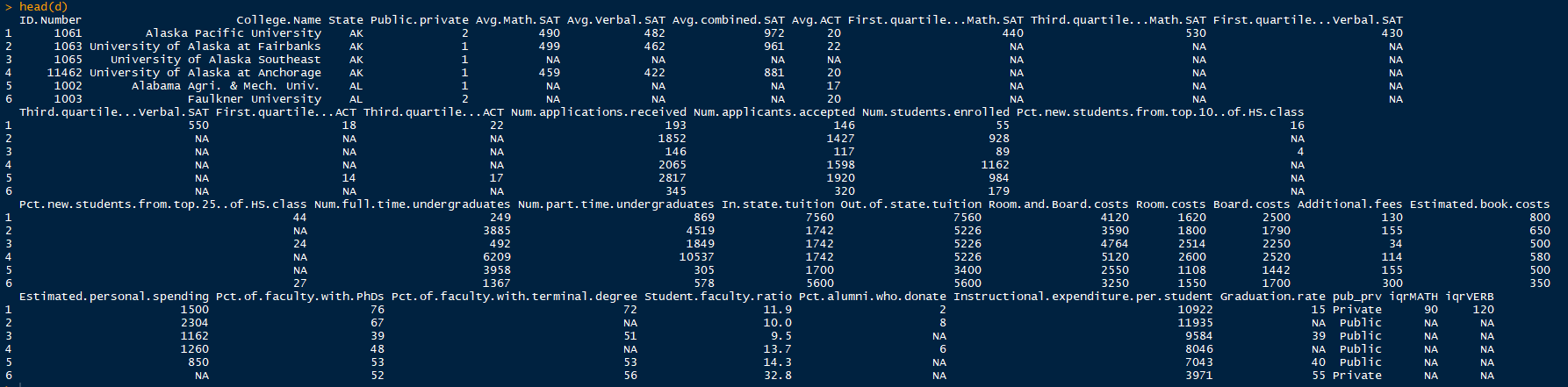
**ADAM HENDEL**

**DS 710 - Homework 9**

**R assignment**

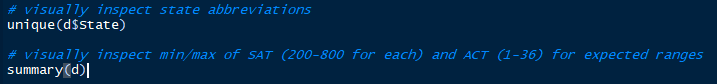
**1(a)** Read the modified data into R.  Check the first few values of each vector to ensure that they were read accurately.

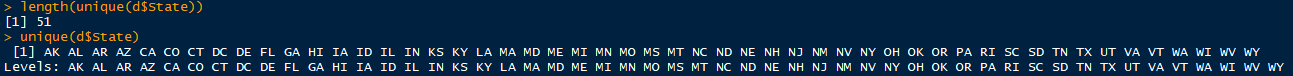


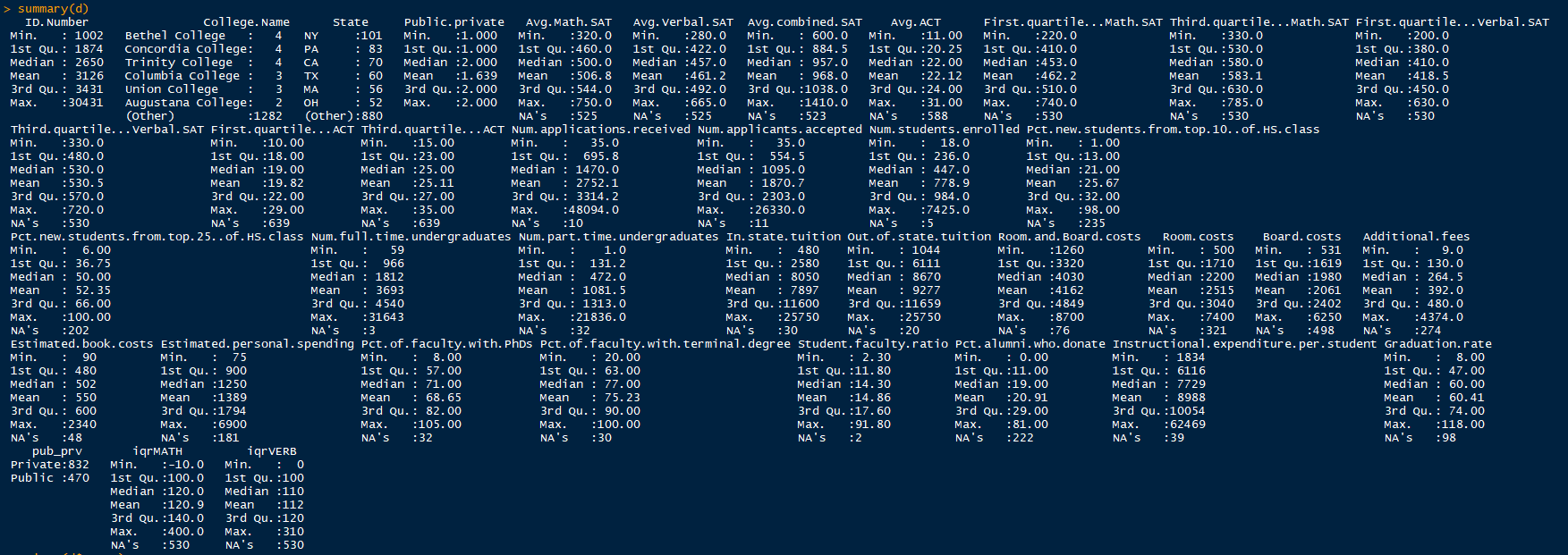


New columns came in properly and NA values reported as expected.

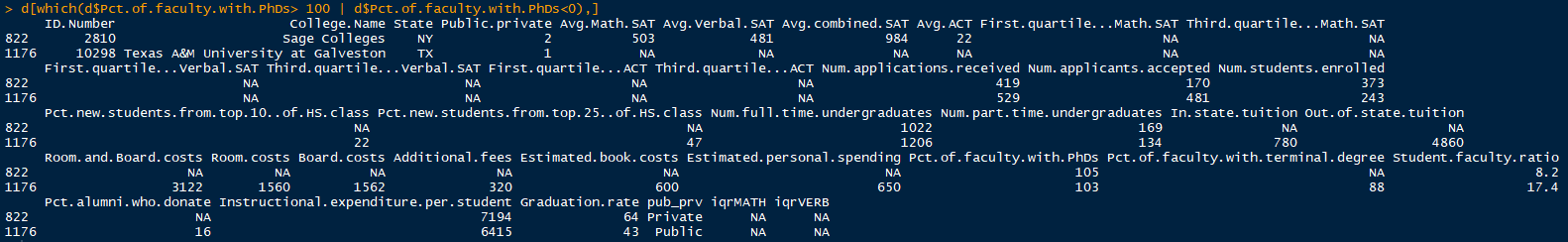
**1(b)** Examine the summary of each variable.  Identify any unrealistic values and set them to missing.  Write a sentence describing what you did, naming the colleges or universities affected.  (For example, “Listed ages less than zero (ABC University, XYZ College) were converted to missing data.”)









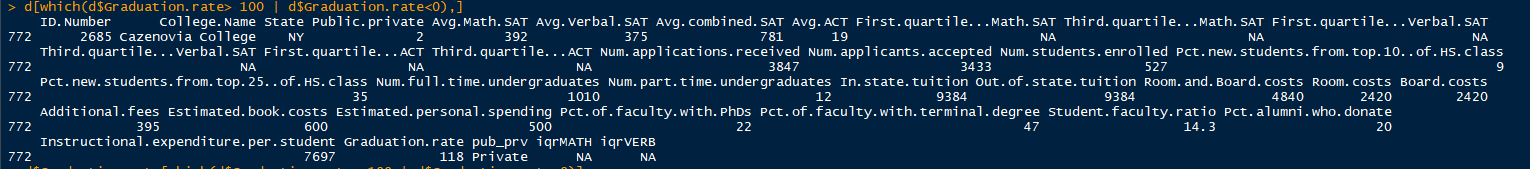


# two schools have percentages >100 (Sage Colleges, Texas A&M) for this variable---set them to NA



# grad rates > 100 or < 0 are not possible

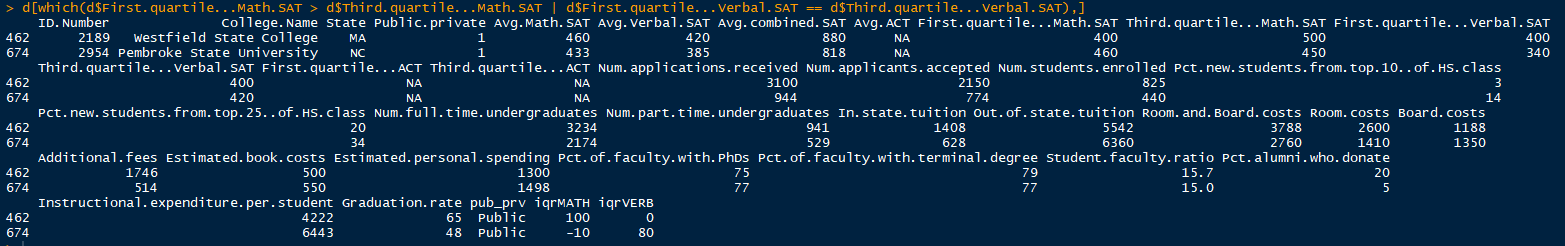




Cazenovia College has 118 percent grad rate, set it to NA



iqrMATH min value is reported at -10, so there is a record where 3QT > 1QT. Likewise, iqrVERB min is reported at 0, so 3QT==1QT in a record. We should take a look at these records.



Westfield State College has Verbal 3QT == 1QT and Pembroke State University has Math 1QT>3QT.

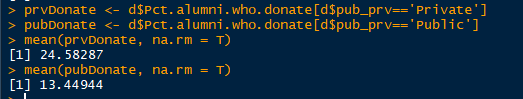
Westfield-set Verbal 1QT, 3QT and IQR to NA



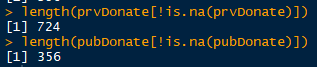
Pembroke-set Math 1QT, 3QT and IQR to NA



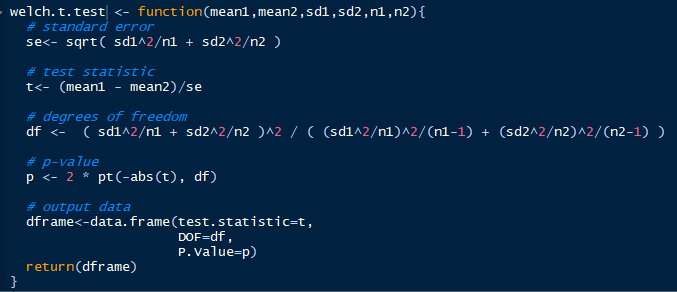
**1(c)** Find the mean percentage of alumni who donate, for private and public schools.

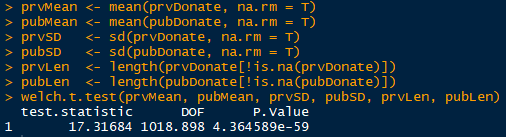


**1(d)** The two groups, public and private, have neither the same number of samples nor the same variance (as shown below)



Thus, we will use Welch’s t-test (same method from assignment 8) https://en.wikipedia.org/wiki/Welch%27s\_t-test:



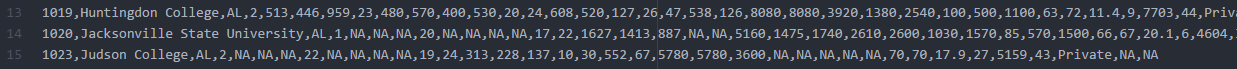


Given a *p-value­* close to zero, there is enough evidence to reject the null hypothesis that the two means are equal. The percentage of alumni that donate back to the institution is greater in private schools than in public schools.

**1(e)** Write to CSV

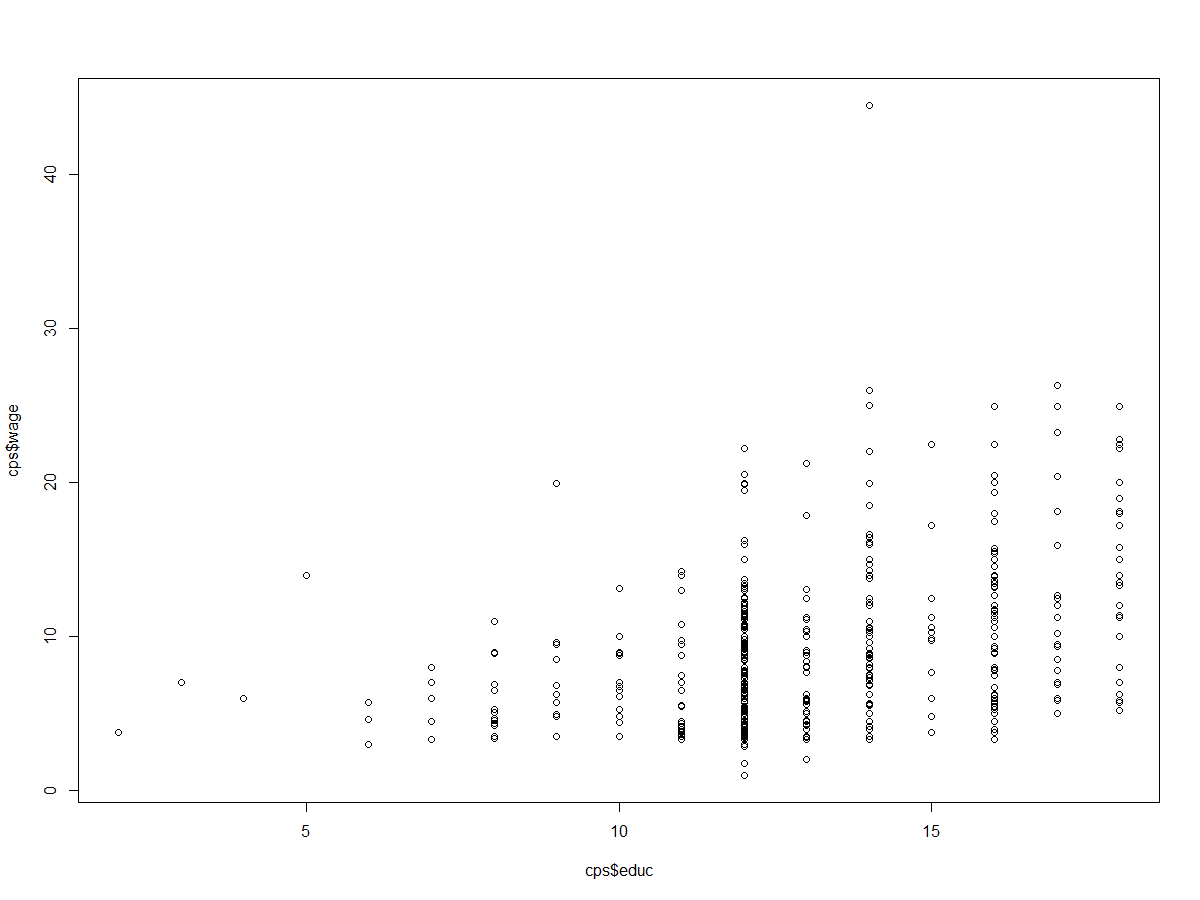
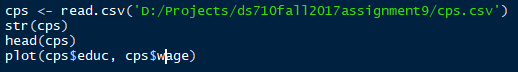


Confim w/ text editor (Atom) that quotations did not show up around entries.



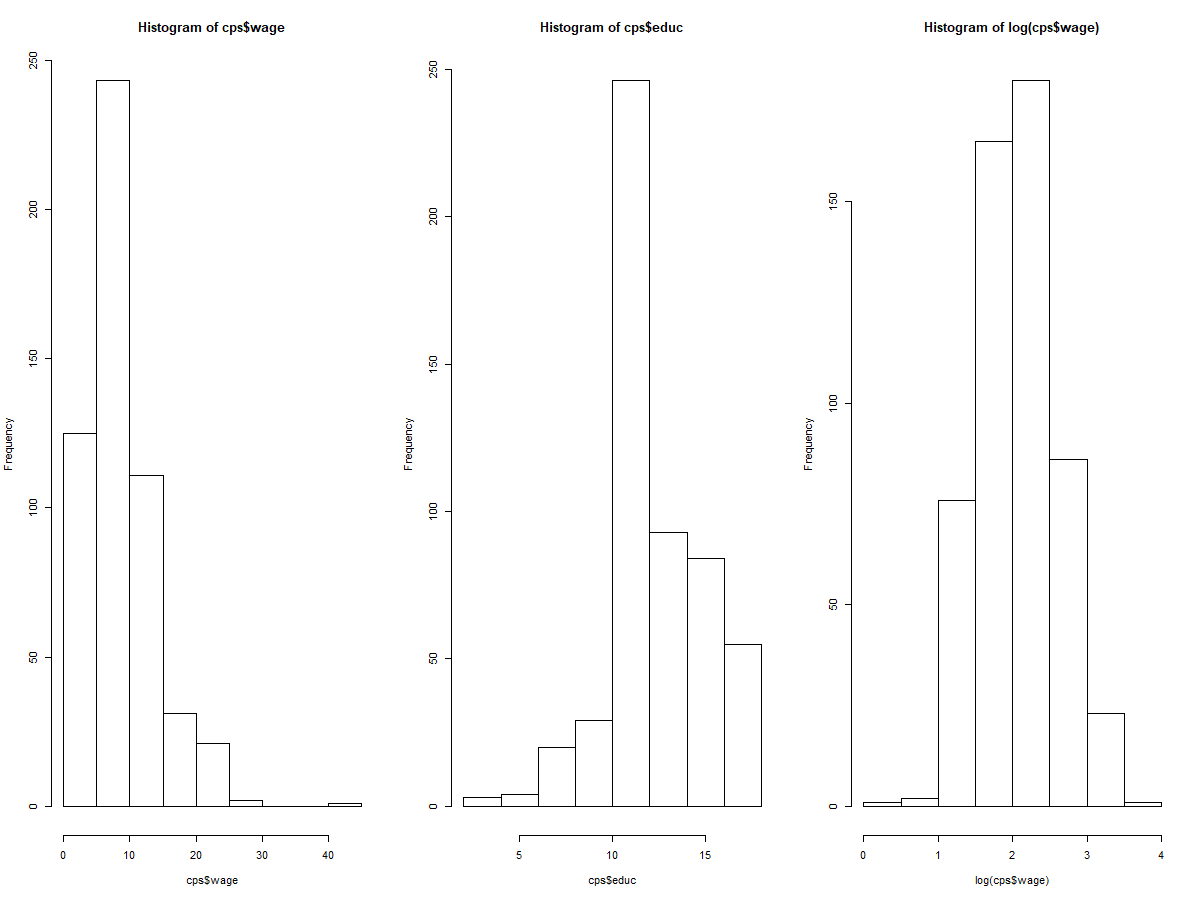
**2(a)**

Read the data into R and plot wages versus education.

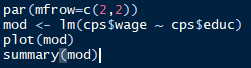
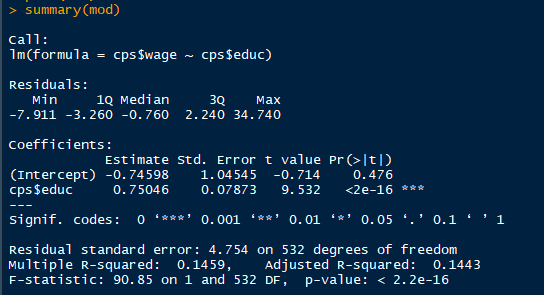


There appears to be a relationship between wages and education. Low education seems to have lower wages, but also lower observations. The min wage of higher education levels seems to be higher than lower educations as well. A linear regression might help explain the significant of the linear relationship.

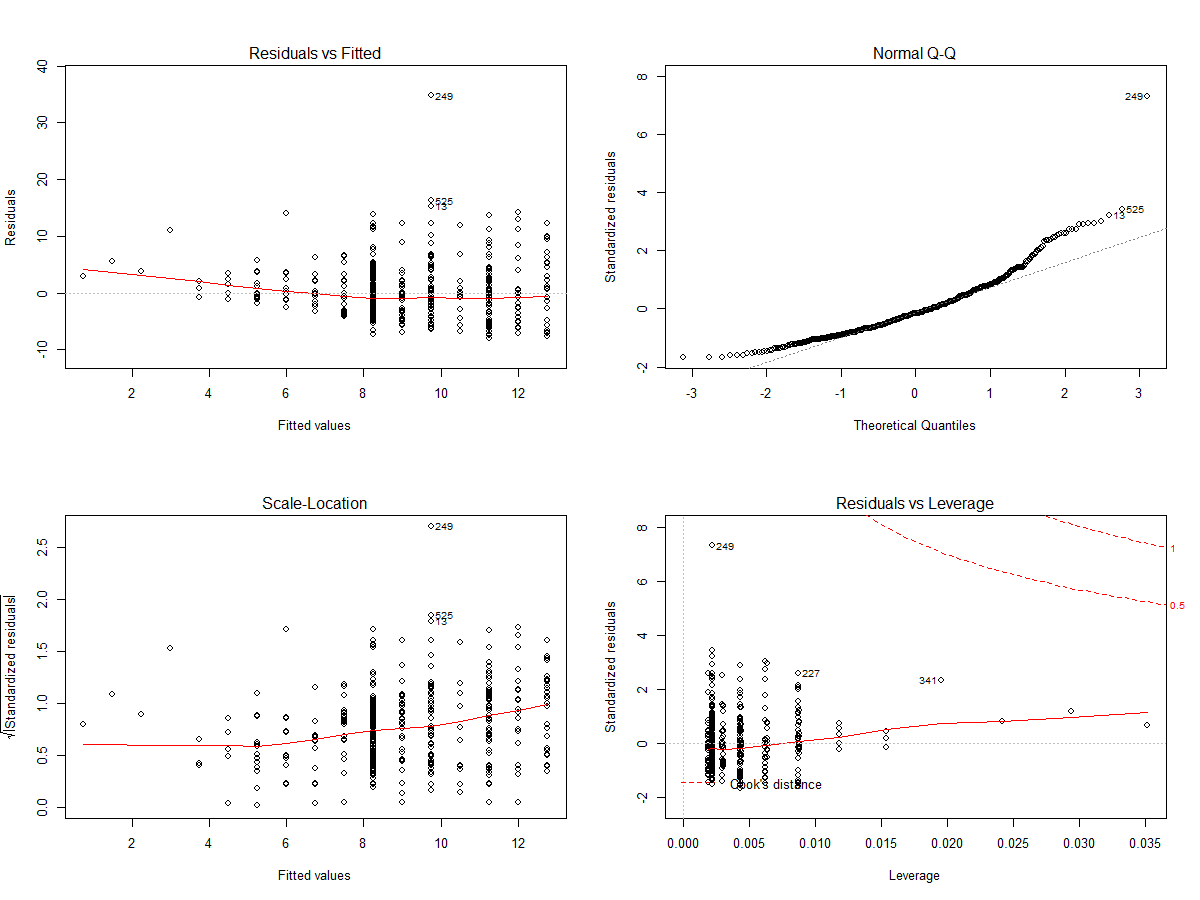
We observe right skewness on the wage variable, thus a log transformation might help our analysis.



**2 (b)**

At first glance, it seems like the linear model could be a good start in the explanation of the relationship in our data (low p-value, but low adj R sqr)

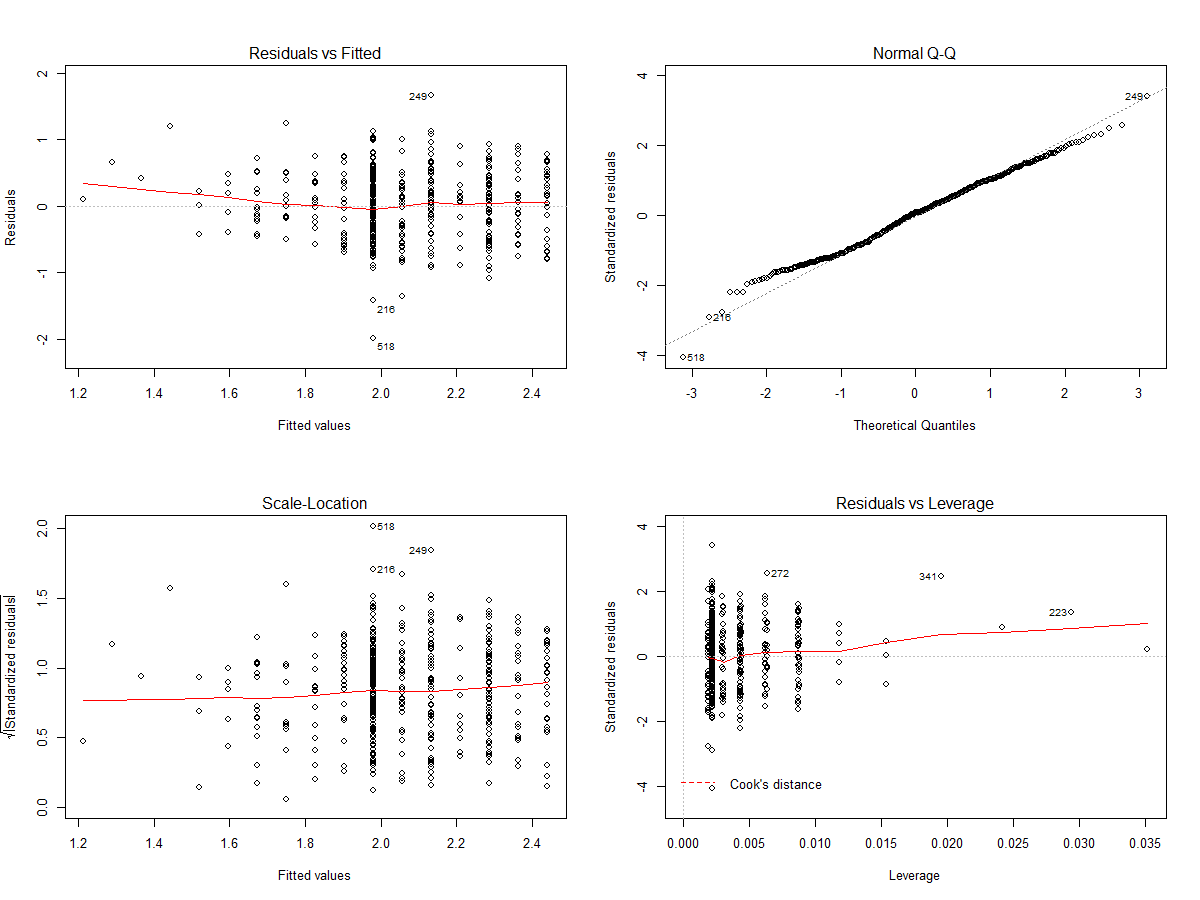
When we examine diagnostic plots of the linear model we can see higher theoretical quantiles trail upwards on (QQ plot), thus a log transform of wage might help the residuals fall into a normal distribution.



Build the Linear Model using a log transformation on wages



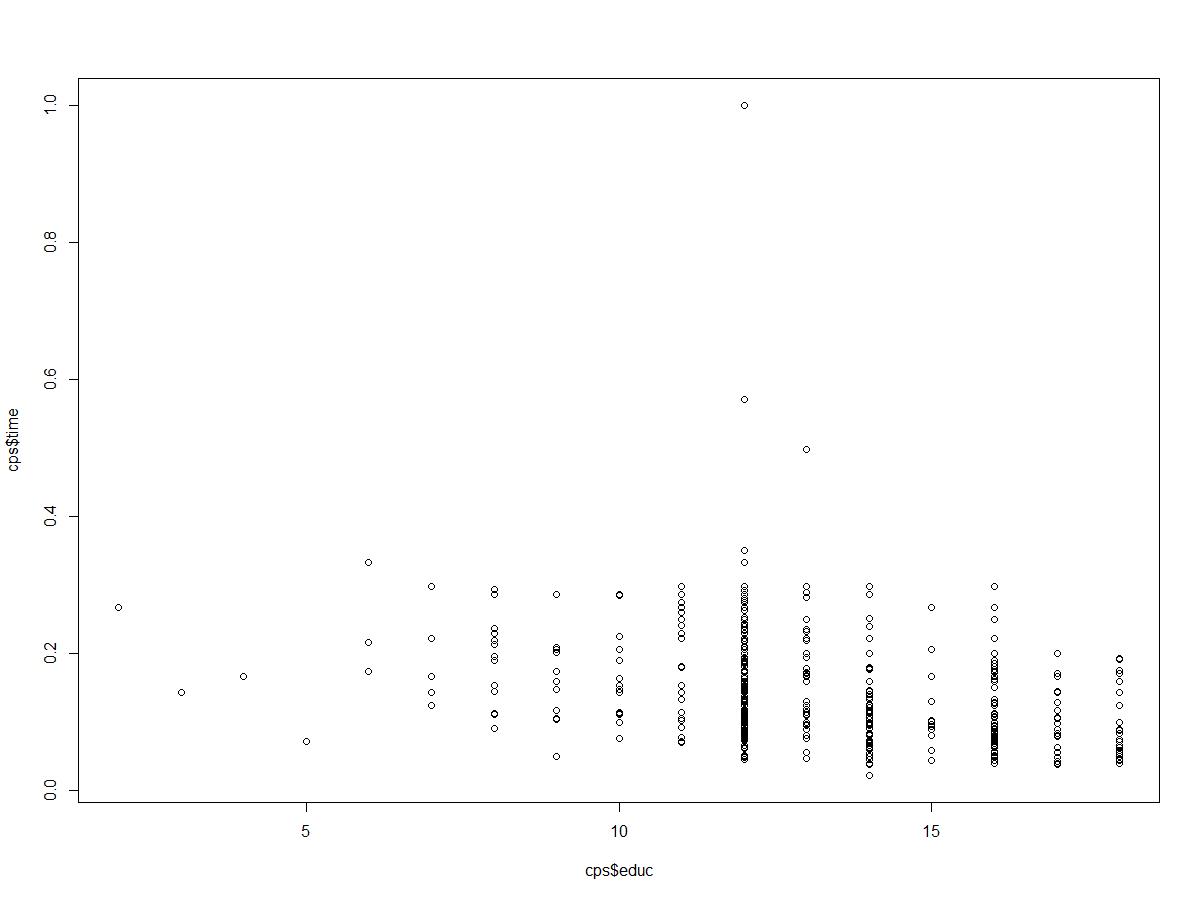
The residuals follow a normal distribution much better under a log transformation.



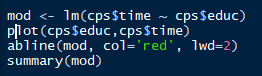
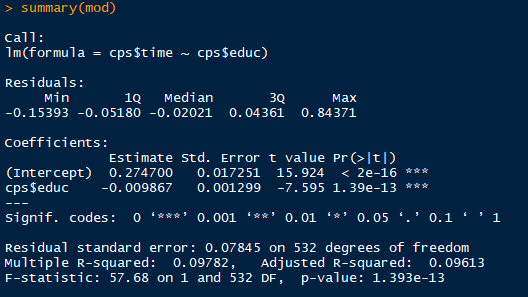
**2(c)** Create a new variable which is the inverse of wage. i.e. the amount of time (hours) it takes to earn a single dollar.



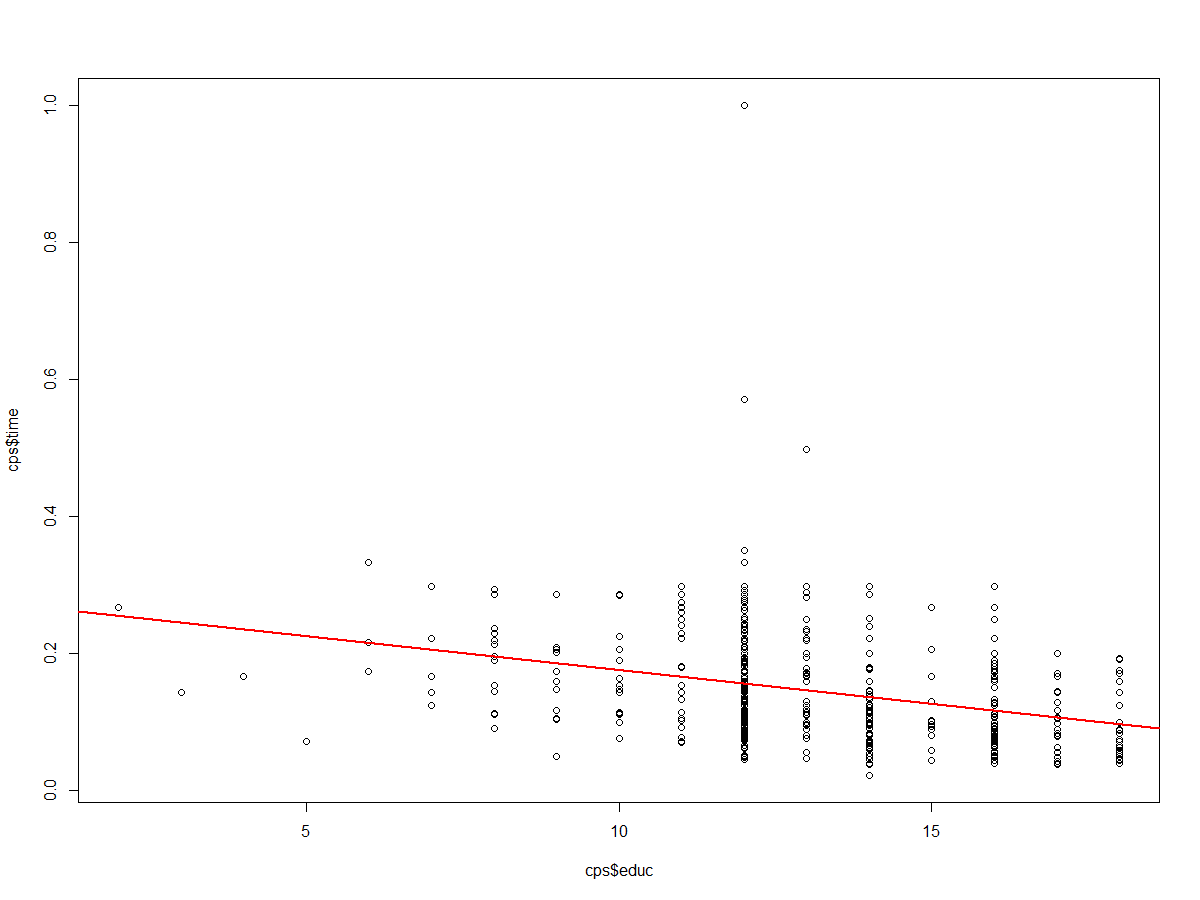
**2(d)** Linear regression is appropriate with these variables, since the dependent variable is continuous, and we have a numerical independent variable.



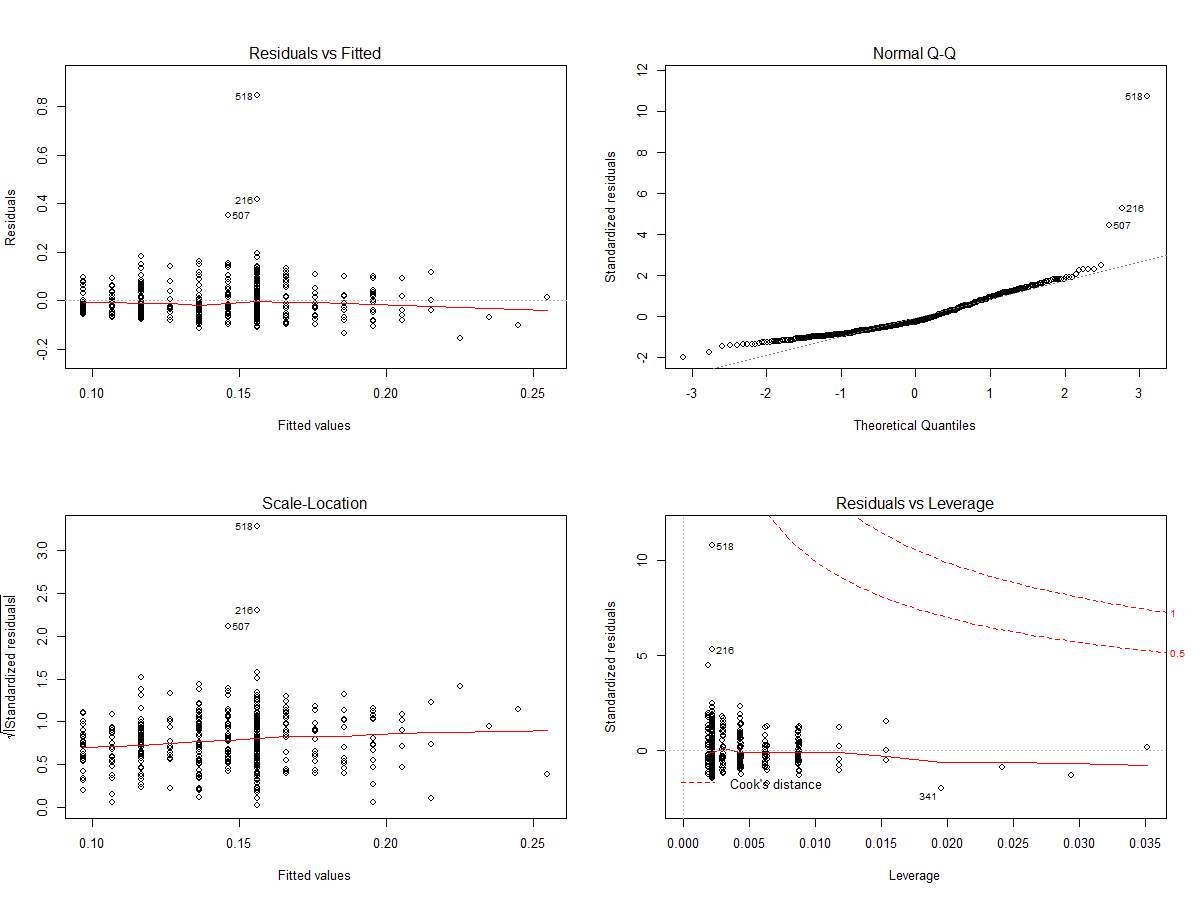
**2(e)** Perform linear regression:

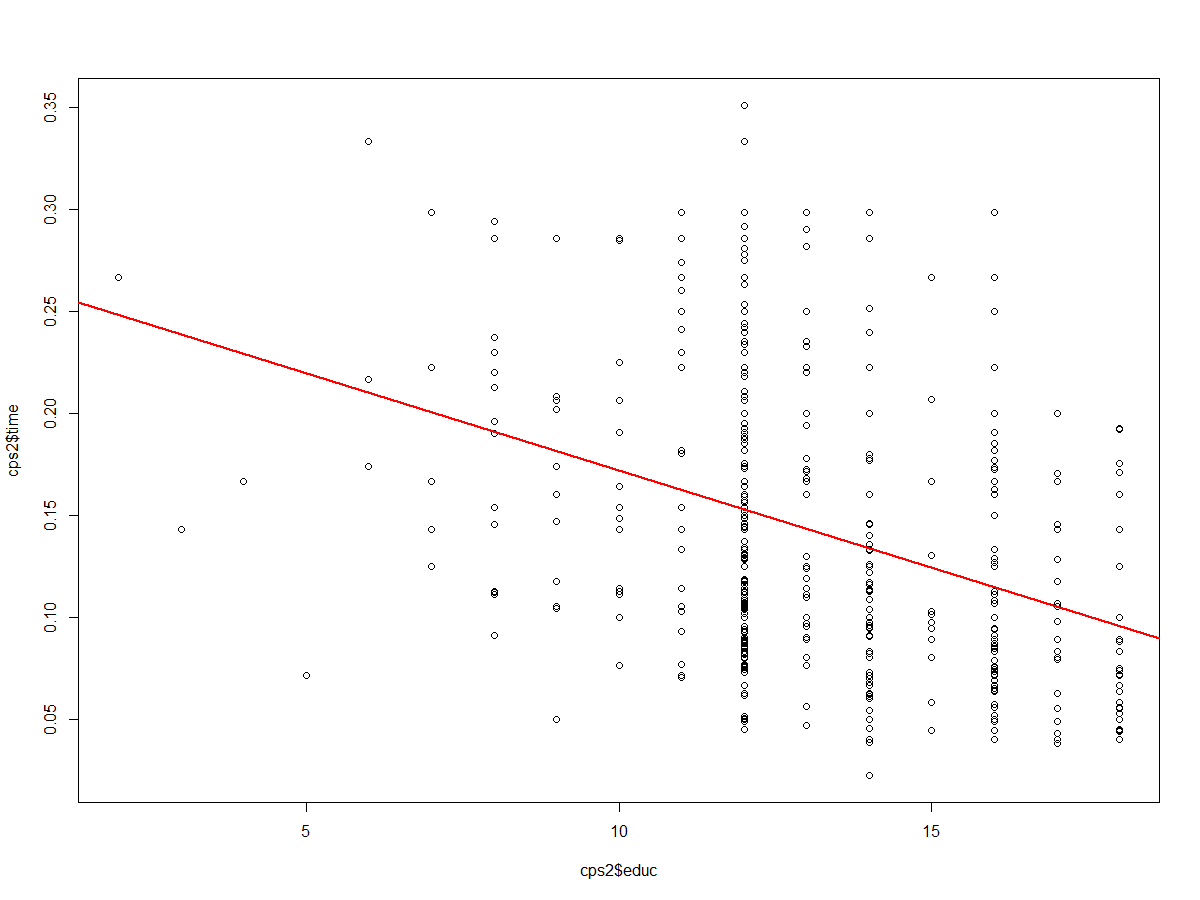
A linear model does describe the general relationship between amount of the hours it takes to earn one dollar and the years of educations. A p-value very close to zero provides enough evidence to reject the null hypothesis that the variable *time* is not related to the variable *education.* There is quite a bit of variability that the linear model does not account for though, as indicated by a low Adjusted R-squared. Generally, more years of education seem to be related to a shorter amount of time to earn one dollar. I am not exactly convinced by this data that simply attending grad school will equate to more earnings (though I am happy with my decision to attend grad school for many other reasons).

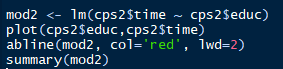
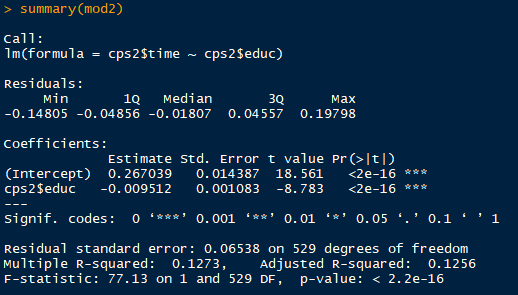


**2(f)** By inspection of the diagnostic plots we can see that records 216, 507 and 518 are outliers.



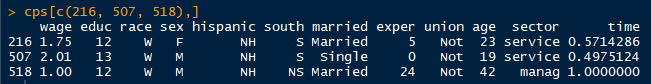
We can try removing these from the plot to see how our analysis changes.



Removing the outliers further supports the relationship between years of education and time to earn a dollar. By removing the outliers from the analysis our Adjusted R-Squared increased to 0.13, which means there is still quite a bit of variance that the linear model is not explaining. The analysis doesn’t really address whether attending graduate school specifically increases wages over not attending grad school. To answer this question, I would focus the analysis more on earnings of individuals with >16 years of education. This could be framed as a decision to stop education at 16 years (i.e. bachelor degree) or continue to grad school to go >16 years of education.

I am curious to know more about the outliers that we removed:



These were very low wage earners with 12-13 years of education, i.e. neither a bachelor nor master’s degree. While removing these outliers does improve the metrics of the linear model, it does not help address whether attending **grad school** (rather than stopping education at a bachelors degree)equates to increased earnings.