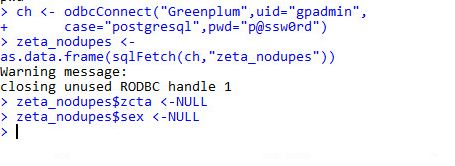
CSC570Q Data Science Essentials

Lab 6: Linear Regression

(**Linear Regression)** In this assignment, you will analyze linear regression models on different categories of data about average households in the United States.

**Gather and Prepare Data**

* In the database *training2*, create a table named *zeta\_nodupes*, which is a copy of the table *zeta* that does not have any duplicate rows of data (done in lab 3). In R, open an ODBC channel and import the *zeta\_nodupes* table.
* Remove the columns *zcta* and *sex* from the imported table.



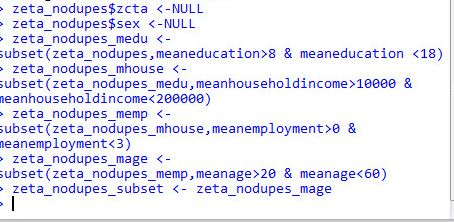
* Remove outliers by creating subsets of the original data so that:

8 < *meaneducation* < 18

10,000 < *meanhouseholdincome* < 200,000

0 < *meanemployment* < 3

20 < *meanage* < 60



* Create a variable called log\_income = log10(meanhouseholdincome).

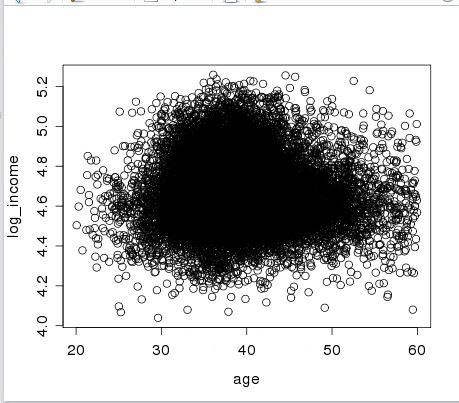
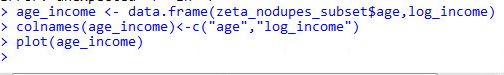


* Rename the columns *meanage*, *meaneducation*, *and meanemploymen*t as *age*, *education*, *and employment*, respectively.



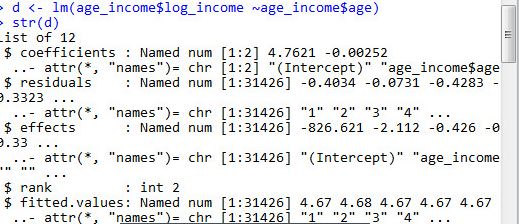
**Linear Regression Analysis**

1. We will be analyzing this data with income as the dependent variable and the other columns as independent variables. Create a scatter plot showing the effect age has on log\_income and paste it here. Do you see any linear relationship between the two variables?



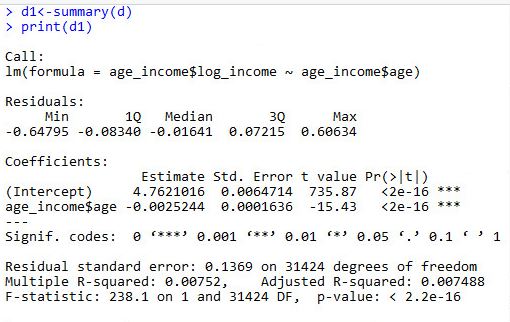
No, the linear relationship between age and log\_income is distorted and scattered.

1. Create a linear regression model between log\_income and age. What is the interpretation of the t-value? What kind of t-value would indicate a significant coefficient?



The 't' value is 735.87 of the intercept and -15.43 for the age.

1. What is the interpretation of the R-squared value? What kind of R-squared value would indicate a good fit?



The value of R-squared is 0.00752. if this value is close to 1 then it indicates a good fit

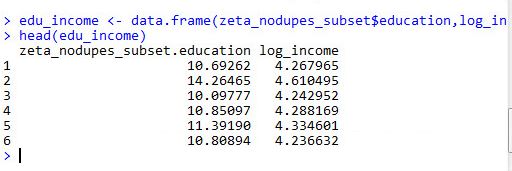
1. What is the interpretation of the F-statistic? What kind of F-statistic indicates a strong linear regression model?

The value of F-statistic is 238.1 on 1 and 31424 DF. If the linear model is just estimating same as the null model then the F-statistic should be about 1.

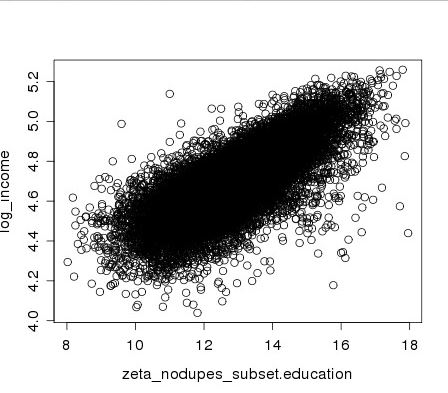
1. View a detailed summary of the previous model. What is the R-squared value? Does this suggest that the model is a good fit? Why?

The value of R-squared is 0.00752 and it's less than 1. certainly, this model is not a good fit.

1. Create a scatter plot showing the effect education has on log\_income. Do you see any linear relationship between the two variables?

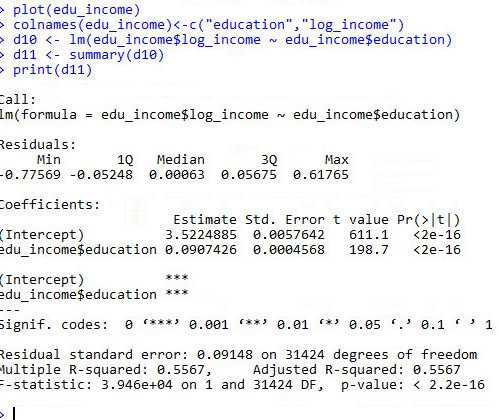


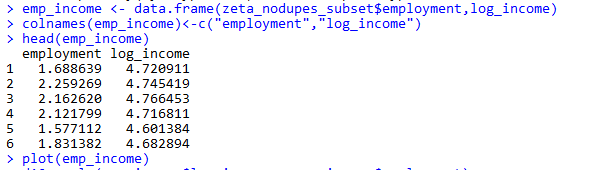
Yes, a linear relationship do exist between education and log\_income but also there are many outliers

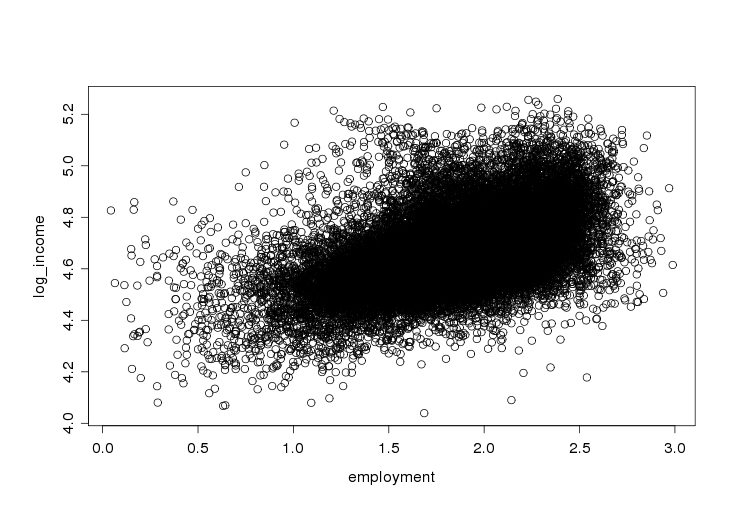
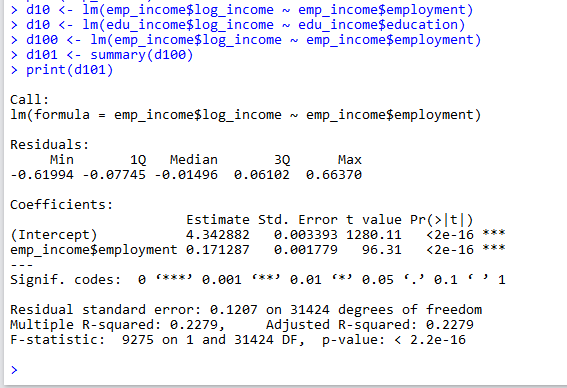


1. Analyze a detailed summary of a linear regression model between log\_incomeand education. What is the R-squared value? Is the model a good fit? Is it better than the previous model?

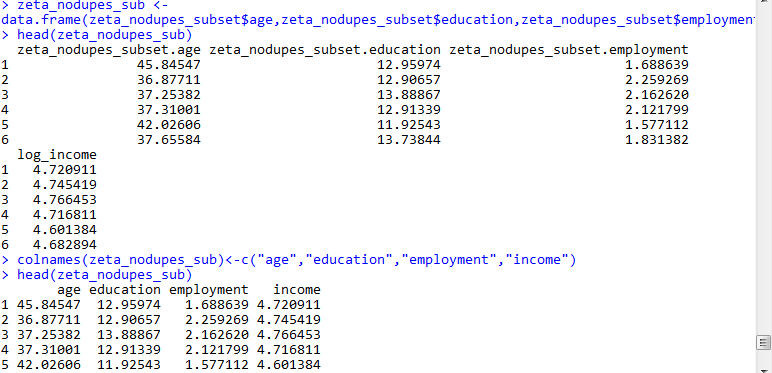
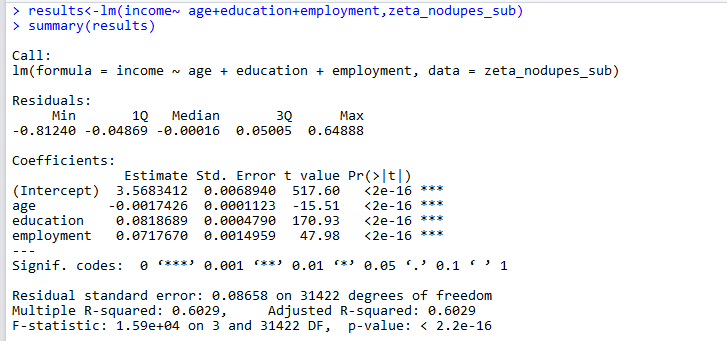
The R-squared value is 0.5567 and it's better than the previous model which was very less compared to this. Though the value is not close to 1 for an ideal fit.



1. Analyze a detailed summary of a linear regression model between the dependent variable *log\_income*, and the independent variables *age, education, and employment*. Is this model a good fit? Why? What conclusions can be made about the different independent variables?



1. Based on the coefficients of the multiple regression model, by what percentage would *income* increase/decrease for every unit of *education* completed, while all other independent variables remained constant?



For every unit of increase in the education, the income will increase by 0.0818689

1. Create a graph that contains a y = x line and uses the multiple regression model to plot the predicted data points against the actual data points of the training set.
2. How well does the model predict across the various income ranges?