DSC 423
Assignment 2

Based on Modules 3 and 4

Ronaldlee Ejalu

StudentId: 2020637

I have completed this work independently. The solutions given are entirely my own work

Your submission must include your name and student ID. Your submission must include the honor statement: "I have completed this work independently. The solutions given are entirely my own work."

1) Short Essay (10 pts.) For each of these questions, your audience are persons that are not experts in statistics. Write with complete sentences and paragraphs. Cite any references that you use.

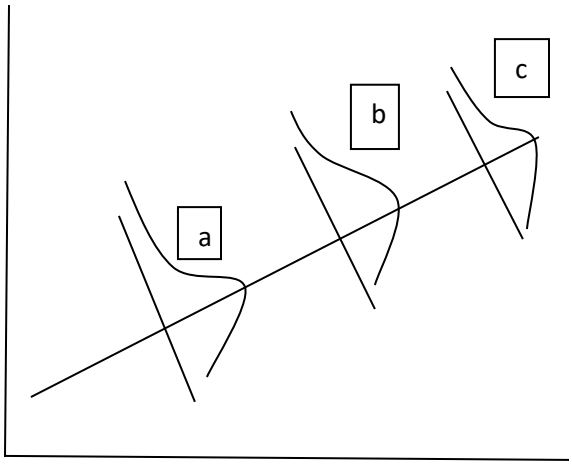
   a. (5 pts.) When building a model, you make four assumptions about the residuals. Explain what they are and how you can verify that your assumptions are correct.

      Our first assumption is that mean of the residual is zero, this should be true if we did our math correctly. The least squares regression model always produces a sum of the residual at zero.

      This can be verified when we draw a scatter plot of residuals and the y values. Y values are plotted on the vertical y – axis, and the residuals are plotted horizontally on the x-axis. If the scatter plot follows a linear pattern this shows that the sum of the residual is zero since the least squares regression model always produces a sum of the residual at zero.

      The second assumption is that the residuals are homoscedastic, this means the variance of residuals is constant throughout the independent variables. For example, looking at the image below, If we were to look at the variance of residuals that is to say a, b, and c defined below around the regression line, we would expect them to the same.

      This meets the equal variance assumption.



Our third assumption is that the residuals are normal, about half of the residuals will be above the regression line and about half below. If the residuals are not skewed, that means the assumption is satisfied.

The fourth assumption is residuals are independent. For example, one residual should not depend on another residual. If you are sampling television sets from a manufacturing line for example, two TVs next to each other in the assembly, might have a similar error. It would be better to sample items further in the assembly line.

b. (5 pts) Define 'interaction term'. From your own experience, identify an instance in which you believe an interaction term would be appropriate.

An interaction term happens when one independent variable interacts with another independent variable on a dependent variable.

For instance, let's say you were studying the effects of exercising and diet meals on weight

loss. The main effects would be the effect of exercising on weight loss, and the effect of diet meals on weight loss. The interaction effect happens when exercising and diet meals happen at the same time. This combination could either speed up weight loss or even slow it down. The synergy between exercising and having diet meals have a greater effect on weight loss than their additive individual effects, which is a special case of the interaction effect.

2) Banking (20 pts.)

a. Use the Banking dataset for this question, found under content on the D2L. This dataset consists of data acquired from banking and census records for different zip codes in the bank's current market. Such information can be useful in targeting advertising for new customers or for choosing locations for branch offices. The fields in the dataset:

   i. Median age of the population (Age)

   ii. Median years of education (Education)

   iii. Median income (Income) in $

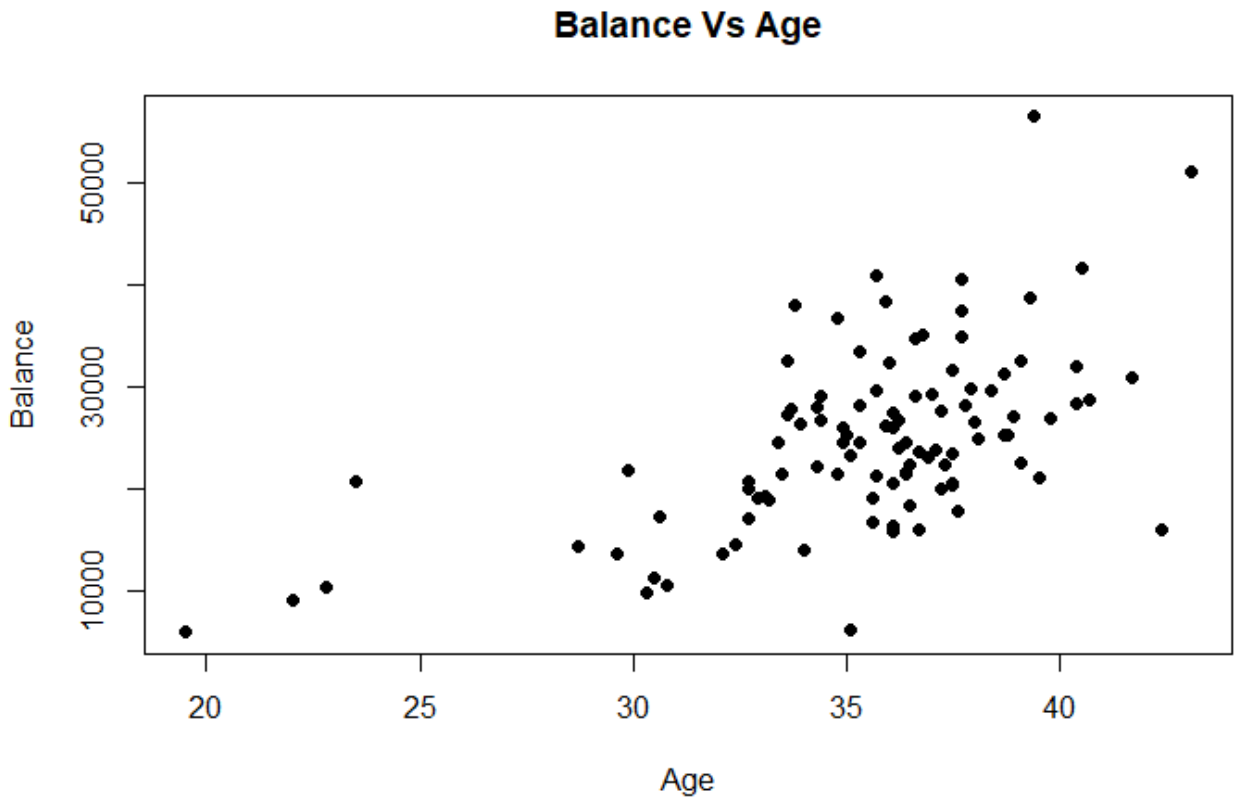
   iv. Median home value (HomeVal) in $

   v. Median household wealth (Wealth) in $

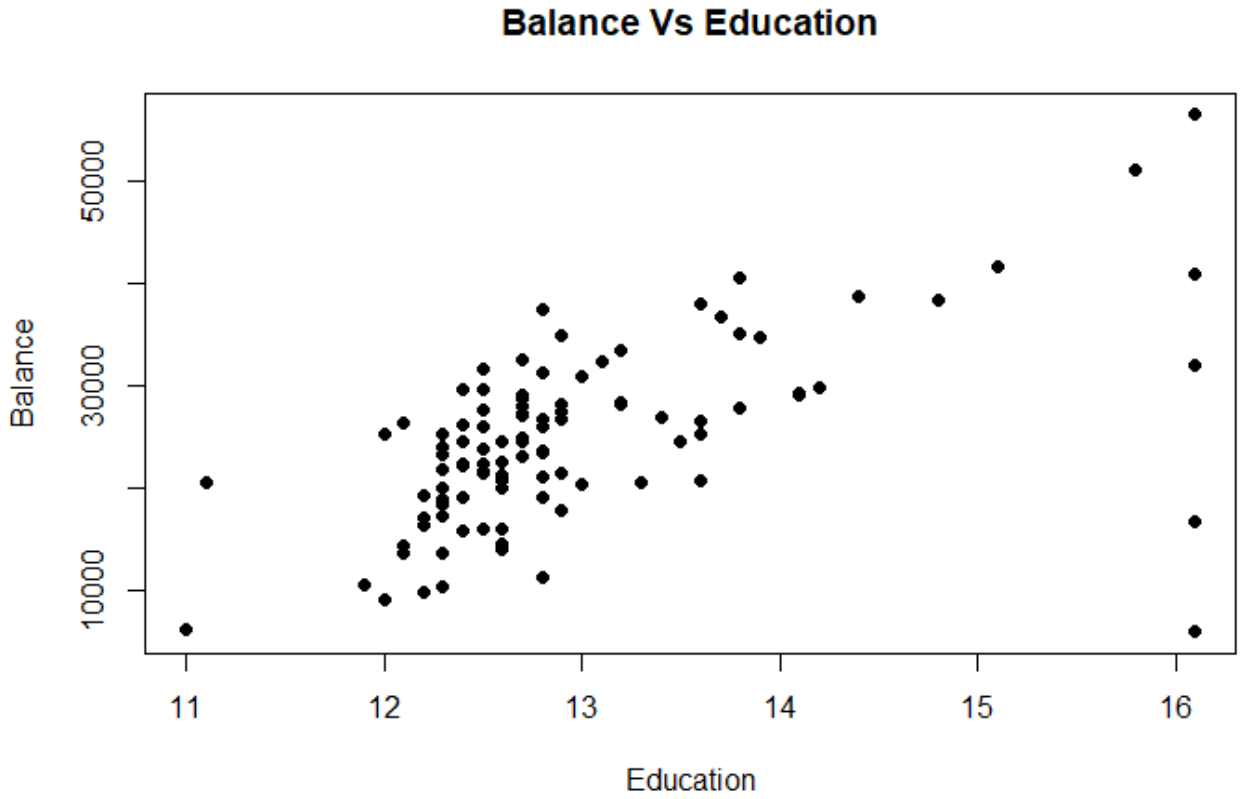
   vi. Average bank balance (Balance) in $

b. Load the data into R.

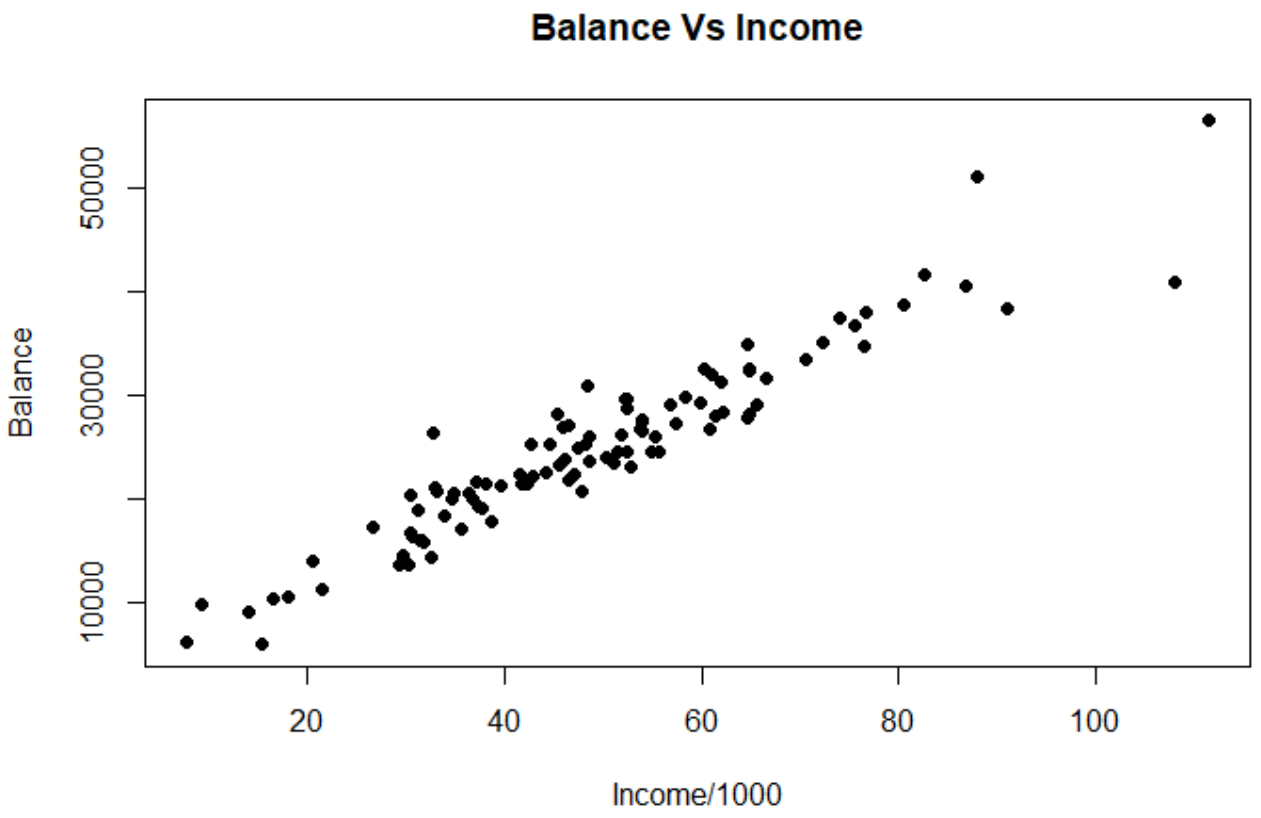
c. In R, you can create a scatterplot by using the plot command, i.e. plot(x, y). Create scatterplots to visualize the associations between bank balance and the other five variables. Paste them into your submission. Describe the relationships.



The above scatterplot Balance Vs Age, the form is linear at first glance but may be a curve would serve better than a straight line. The strength is relatively weak with some noise and the direction is positive. As age increases, so as balance. This might be a good fit to test for a second order model.

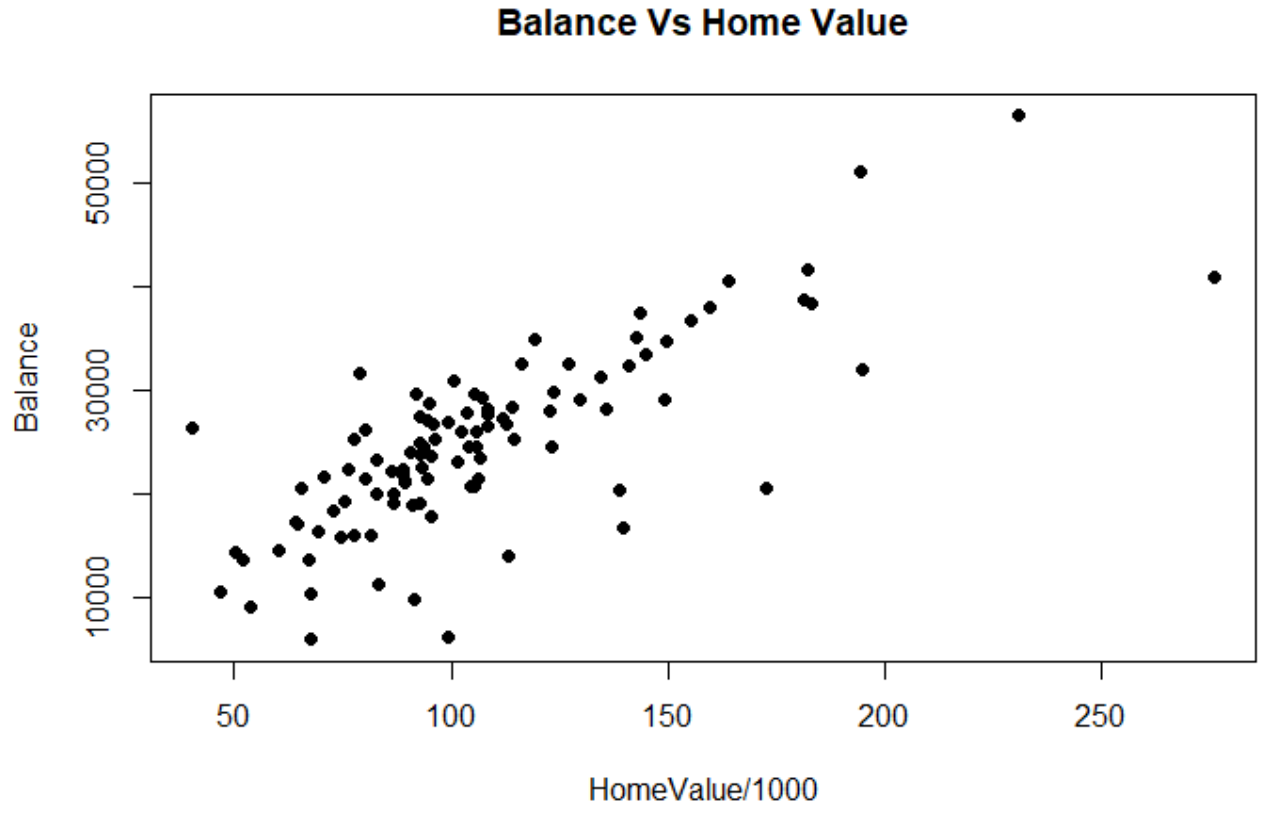


In the Balance Vs Education scatterplot, there is a strong, and positive correlation between education and balance with some bit of noise. At first glance, it might look linear, but a curve would serve as a better model than a straight line. This might be a good fit to test for a second order model.



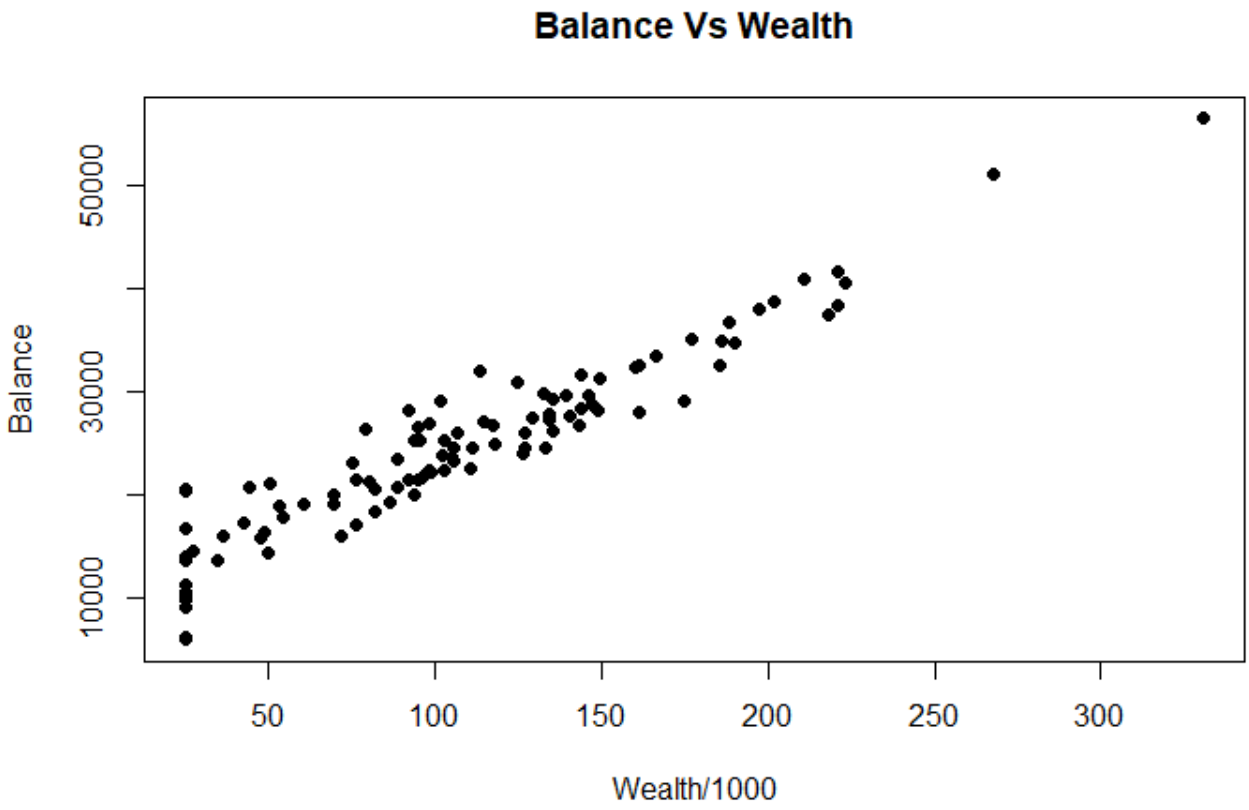
In the scatterplot above for Balance Vs Income, the correlation is strong, linear and positive.

As income increases, so does the Bank Balance.



In the above scatterplot for balance Vs HomeValue, there is a strong, linear and positive

correlation between home value. As home value increases, so does the balance.



In the above scatterplot of Balance Vs Education, there is a strong perfect, linear and positive correlation between wealth and bank balance. As Wealth increases, so does the bank balance.

d. In R, you can compute correlations between two variables by using the cor command, i.e. cor(x,y) where x and y are the names of your variables, or you can compute pair-wise correlations by using cor(D), where D is the name of your dataframe. Compute correlations found in the bank data. Interpret the correlation values. Paste them into your submission. Describe which variables appear to be strongly associated?

```
> cor(BANKING)
                Age Education    Income   HomeVal    Wealth   Balance
Age       1.0000000 0.1734611 0.4771474 0.3864931 0.4680918 0.5654668
Education 0.1734611 1.0000000 0.5731467 0.7489426 0.4681199 0.5521889
Income    0.4771474 0.5731467 1.0000000 0.7953552 0.9466654 0.9516845
HomeVal   0.3864931 0.7489426 0.7953552 1.0000000 0.6984778 0.7663871
Wealth    0.4680918 0.4681199 0.9466654 0.6984778 1.0000000 0.9487117
Balance   0.5654668 0.5521889 0.9516845 0.7663871 0.9487117 1.0000000
>
```

The correlation between age and Balance is 0.56. approximated to 0.6, which is a moderate uphill relationship

The correlation between education and Balance is 0.55 approximated to 0.6 which is moderate uphill relationship.

The correlation between Income and Balance is 0.95 which is a very strong uphill linear relationship.

The correlation between home value and Balance is 0.76 which shows a strong uphill linear relationship.

The correlation between wealth and balance is 0.95 which shows a very strong uphill linear relationship.

e. Fit a regression model of balance vs the other five variables. Present the estimated regression model and evaluate it. Recall that you can build a linear regression model by using the lm command and display the model by using the summary command.

```
> model <- lm(Balance ~ Age + Education + Income + HomeVal + Wealth, data = BANKING)
> summary(model)

Call:
lm(formula = Balance ~ Age + Education + Income + HomeVal + Wealth,
    data = BANKING)

Residuals:
    Min      1Q  Median      3Q     Max
-5365.5 -1102.6   -85.9   868.9  7746.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.033e+04  4.219e+03  -2.449 0.016160 *
Age          3.175e+02  6.104e+01   5.201 1.12e-06 ***
Education    5.903e+02  3.151e+02   1.873 0.064085 .
Income       1.468e-01  4.083e-02   3.596 0.000512 ***
HomeVal      9.864e-03  1.099e-02   0.898 0.371591
Wealth       7.414e-02  1.120e-02   6.620 2.06e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2059 on 96 degrees of freedom
Multiple R-squared:  0.9468,    Adjusted R-squared:  0.944
F-statistic: 341.4 on 5 and 96 DF,  p-value: < 2.2e-16
```

f. Which of the five predictors have a significant effect on balance? (a=.05) Explain.

Age, Income and Wealth have a significant effect on balance since their p-values from the t test look good; the p-values for the null hypothesis that Beta 1, Beta 3 and Beta 5 equals 0. We are going to reject that and accept the alternative that Beta-1, Beta-3 and Beta-5 does not equal zero, which means that the independent variables Age, Income and Wealth have an impact on the Bank Balance.

g. A good model should only contain significant independent variables, so remove the variable with the largest p-value (>0.05) and refit the regression model of balance versus the remaining four predictors. Present the new regression model.

```
Call:
lm(formula = Balance ~ Age + Education + Income + Wealth, data = BANKING)

Residuals:
    Min      1Q  Median      3Q     Max
-5403.9 -1234.1   -75.0   998.6  7430.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.214e+04  3.704e+03  -3.278  0.00145 **
Age          3.242e+02  6.051e+01   5.358 5.68e-07 ***
Education    7.498e+02  2.600e+02   2.884  0.00484 **
Income       1.615e-01  3.738e-02   4.321 3.75e-05 ***
Wealth       7.265e-02  1.106e-02   6.566 2.57e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2057 on 97 degrees of freedom
Multiple R-squared:  0.9463,    Adjusted R-squared:  0.9441
F-statistic: 427.4 on 4 and 97 DF,  p-value: < 2.2e-16
```

h. Analyze if all four predictors have a significant association with balance? (a=.05) If not continue to remove one insignificant variable at a time until all of the remaining predictors are significant.

Age, Education, Income and Wealth have a significant effect on balance since their p-values from the t tests look good; the p-values for the null hypothesis that Beta-1, Beta-2, Beta-3 and Beta-4 equals 0. We are going to reject that and accept the alternative that Beta-1, Beta-2, Beta-3 and Beta-4 does not equal zero, which means that the independent variables Age, Education, Income and Wealth have an impact on the Bank Balance.

i. Interpret each of the regression coefficients for the final model.

Looking at the betas, we can define the regression line as:

Balance = -1.214e+04 + 3.242e+02X1 + 7.498e+02X2 + 1.615e-01X3 + 7.265e-02X4. This is the model that minimizes the sum of the square of errors.

Looking at the F value of 427.4 and the P-Value of 2.2e-16. The P-Value is the probability that given the null hypothesis, that all the Betas associated with the independent variables are equal to zero. We would observe the data as extreme as it is. Since the P-value is very small, so we are going to reject the null hypothesis and accept the alternative, that at least of the Betas is not equal to 0. We don't know which Beta or they are all not equal to zero. It is not what the F-TEST tells us. This is a test of the model itself, which tells me that something in my model is working.

Looking at the individual P-Value for age from the t-test, that is the p-value for the null hypothesis that Beta-1 equals zero. Because the P-Value is low, we are going to reject that null hypothesis and accept the alternative that Beta-1 is not equal to zero and then use the estimation of 3.242e+02.

Second looking at the individual P-Value for Education, that is the p-value for the null hypothesis that Beta-2 equals zero. Because the P-Value is low, we are going to reject that null hypothesis and accept the alternative that Beta-2 is not equal to zero and then use the

estimation of 7.498e+02.

Third looking at the Individual P-Value for Income, that is the p-value for the null hypothesis that Beta-3 equals zero. Because the P-Value is low, we are going to reject that null hypothesis and accept the alternative that Beta-3 is not equal to zero and then use the estimation of 1.615e-01.

Finally looking at the Individual P-Value for Wealth, that is the p-value for the null hypothesis that Beta-4 equals zero. Because the P-Value is low, we are going to reject that null hypothesis and accept the alternative that Beta-4 is not equal to zero and then use the estimation of 7.265e-02.

Because the P-Values for Age, Education, Income and Wealth are low, we are going to include them in our regression model. We are not going to trim them out.

R -Squared is 94.63%. That indicates that our model is predicting 94.63% of the variability in our y, which is our dependent variable.

j. Discuss the adj-R^2 for the final model.

The adjusted R- Squared is 94.41 percent meaning that 94.4 percent of the variability in the bank balance is explained by the model.

k. Are there any influential points in your data set? Explain what impact an influence point might have.

Yes, the influential points do exist in age, education, Income, HomeVal and Wealth.

The influence point will cause the coefficient of determination to be bigger or even sometimes smaller.

3) WATEROIL (20 pts.) In the oil industry, water that mixes with crude oil during production and transportation must be removed. Chemists have found that the oil can be extracted from the water/oil mix electrically. **Researchers at the University of Bergen (Norway) conducted a series of experiments to study the factors that influence the voltage (y) required to separate the water**

from the oil (Journal of colloid and interface science, Aug. 1995). The seven independent variables investigated in the study are listed in the table. (Each variable was measured at two levels - a "low" level and a "high" level.) Sixteen water/oil mixtures were prepared using different combinations of independent variables; then each emulsion was exposed to a high electric field. In addition, three mixtures were tested when all independent variables were set to 0. The variables are given in the table below.

Experiment number

y: voltage (kw/cm)

x1: disperse phase volume (%)

x2: salinity (%)

x3: temperature (^{0}C)

x4: time delay (hours)

x5: surfactant concentration (%)

x6: span:triton

x7: solid particles (%)

a. Use R to perform a regression analysis on the WATEROIL dataset (found on the D2L). Consider interaction terms and second-order terms. Evaluate the t-tests, F-Test and adj-R^2 accordingly.
b. (5 pts.) Paste your final model into your submission.

```
> model1 <- lm(Voltage ~ Volume + Salinity + VolumeSQ +  Surfactant + VolSurfactant, data=WATEROIL)
> summary(model1)

Call:
lm(formula = Voltage ~ Volume + Salinity + VolumeSQ + Surfactant +
    VolSurfactant, data = WATEROIL)

Residuals:
   Min     1Q Median     3Q    Max
-0.775 -0.235 -0.005  0.235  0.875

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.0666667  0.2404625   4.436 0.000672 ***
Volume        -0.0910625  0.0264504  -3.443 0.004368 **
Salinity       0.1700000  0.0694155   2.449 0.029267 *
VolumeSQ       0.0008880  0.0003383   2.625 0.020995 *
Surfactant     1.1800000  0.3292669   3.584 0.003334 **
VolSurfactant -0.0120000  0.0052062  -2.305 0.038305 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4165 on 13 degrees of freedom
Multiple R-squared:  0.7815,    Adjusted R-squared:  0.6975
F-statistic:   9.3 on 5 and 13 DF,  p-value: 0.0006019

>
```

c. (15 pts.) Describe your model. Assume your audience is a fellow DSC423 student. Your description should begin by reporting basic facts about your model; but should also include an analysis of the findings.

We can define the regression line as Voltage = 1.0667 – 0.091X1 + 0.17X2 + 0.0009X1^2 + 1.18X3 - 0.012X1X3. This is the model that minimizes the sum of the square of errors.

Looking at the F-value of 9.3 and P-value of 0.0006. The P-value is the probability that given the null hypothesis, that all the Betas associated with the independent variables are equal to zero. We would observe the data as extreme as it is. Since the P-value is very small, so we are going to reject the null hypothesis and accept the alternative, that at least of the Betas is not equal to 0. We don't know which Beta or they are all not equal to zero. It is not what the F-TEST tells us. This is a test of the model itself, which tells me that something in my model is working.

The adjusted R-Squared is 69.75 percent meaning that 69.75 percent of the variability in Voltage is explained by our model.

Looking at the individual P-Value for volume from the t-tests, that is the p-value for the null hypothesis that Beta-1 equals zero. Because the P-Value is low, we are going to reject that null hypothesis and accept the alternative that Beta-1 is not equal to zero and then use

the estimation of -0.091.

Looking at the individual P-Value for Salinity from the t-tests, that is the p-value for the null hypothesis that Beta-2 equals zero. Because the P-Value is low, we are going to reject that null hypothesis and accept the alternative that Beta-2 is not equal to zero and then use the estimation of 0.17

Looking at the individual P-Value for the second order term (VolumeSQ) from the t-tests, that is the p-value for the null hypothesis that Beta-3 equals zero. Because the P-Value is low, we are going to reject that null hypothesis and accept the alternative that Beta-3 is not equal to zero and then use the estimation of 0.0008.

Looking at the individual P-Value for Surfactant from the t-tests, that is the p-value for the null hypothesis that Beta-4 equals zero. Because the P-Value is low, we are going to reject that null hypothesis and accept the alternative that Beta-4 is not equal to zero and then use the estimation of 1.18.

Lastly looking at the individual P-Value for the interaction term, VolSurfactant, from the t-tests, that is the p-value for the null hypothesis that Beta-5 equals zero. Because the P-Value is low, we are going to reject that null hypothesis and accept the alternative that Beta-5 is not equal to zero and then use the estimation of -0.012.

Because the P-Values for Volume, Salinity, Second Order term of Volume (VolumeSQ), Surfactant and the interaction term of volume and Surfactant are low, we are going to include them in our regression model. We are not going to trim them out.