Ximan Liu
1935858
DSC 423
HW3

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

[Question 1]
1a)
We make four assumptions about the residuals when building a model.
Firstly, linearity. We assume that there exists a linear relationship between the independent variable X and the dependent variable Y. To verify that, we can draw a scatter plot of value X and value Y. If the plot seems like can fall along a straight line like linear pattern and not a curve, then that the assumption is true.
Secondly, independence. We assume that the residuals are independent, because we do not want a following continuously pattern on residuals. To verify that, we need to check out residual time series plot, as known as a plot of residuals and time. Most of residual falls within 95% confidence brands around aero, which located +/-2 over the square root of n, n is the size of sample.
Thirdly, normality. We assume that residuals are normal distributed. To verify that, we use quantile-quantile plot (as known as Q-Q plot). It indicates whether the residuals follow normal distribution. If the points in the plot look like a straight diagonal line, the assumption is approved.
Lastly, homoscedasticity. We assume that residuals have constant variance at every level of X. If residuals do not follow this assumption, then the situation is heteroscedasticity. it is quite bad because it increases the variance of the regression coefficient estimates, but the model does not notice that. To verify assumption, we can draw a scatter plot of fitted value versus residuals. If the residuals become more spread out as the fitted values getting larger, then there is a cone shape showed up as a typical sign of heteroscedasticity.

Citation:
Zach, V. (2021). The Four Assumptions of Linear Regression - Statology. Retrieved 5 July 2021, from https://www.statology.org/linear-regression-assumptions/

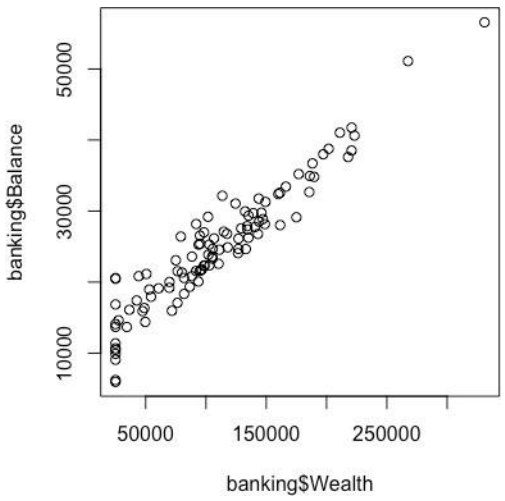
1b)
When the values of one independent variable affect the outcome of another independent variable, this is called an interaction term.
For example, imagine we have to do experiments to determine whether erythromycin is more effective, or penicillin is more effective. At this time, we cannot arbitrarily say which antibiotic is better, but should judge which is more effective based on the type of bacteria. For example, erythromycin may not be sensitive to type A bacteria, but it is more sensitive to type B bacteria. Penicillin has a significant effect on type A bacteria but has little effect on type B bacteria. In contrast, the antibiotics are greatly affected by another variable, that is,

bacteria. Therefore, the outcome of the test result, which shows which antibacterial drug is effective, also appears the interaction situation.

Citation:
Frost, J. (2021). Understanding Interaction Effects in Statistics - Statistics By Jim. Retrieved 5 July 2021, from https://statisticsbyjim.com/regression/interaction-effects/

[Question 2]
2a)
The graph indicates positive linear relationship between Balance and other variables (Age, Education, Income, HomeVal and Wealth). The relationship between Balance and Income, and the relationship between Balance and Wealth represent strong associations.
There are possible potential outliers at the top right corner.
The relationship of Age and Balance is weak association because there is a cloud in the graph.
The relationship of Education and Balance is weak association because there is a cloud in the graph.
The relationship of HomeVal and Balance is strong association. There are possible potential outliers in the plot.

banking$Education
banking$HomeVal
banking$Age
banking$Income



2b)
cor(banking$Age, banking$Balance)
0.5654668

cor(banking$Education, banking$Balance)
0.5521889

cor(banking$Income, banking$Balance)
0.9516845

cor(banking$HomeVal, banking$Balance)
0.7663871

cor(banking$Wealth, banking$Balance)
0.9487117

The higher the correlation r value (between 0 and 1), the stronger the linear relationship. cor(banking$Income, banking$Balance) = 0.9516845 and cor(banking$Wealth, banking$Balance) = 0.9487117 represent highly strong perfect positive association between Balance and Income, and Balance and Wealth.

cor(banking$HomeVal, banking$Balance) = 0.7663871 represents a relatively strong association between Balance and Homeval.

cor(banking$Age, banking$Balance) = 0.5654668 and cor(banking$Education, banking$Balance) = 0.5521889 represent a moderate weak association between Balance and Age, and Balance and Education.

2c)

```
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
```

Balance = (-1.033e+04) + (3.175e+02) * Age + (5.903e+02) * Education + (1.468e-01) * Income + (9.864e-03) * HomeVal + (7.414e-02) * Wealth + e
The adjusted R-squared value is 0.944 which is a quite high number.

2d)
If we use the t-tests on the parameters for alpha = 0.05, then Age, Income and Wealth have significant effects on Balance. The p-value of all of them are either less than 0.0001 or less than alpha = 0.05.

2e)

```
# Drop HomeVal & Education
m2 <- lm(Balance ~ Age + Income + Wealth, data = banking)
summary(m2)
```

After excluding the Education and HomeVal from the model, we need to fit the regression model once again.

The model equation is Balance = (-3.115e+03) + (3.019e+02) *Age + (2.119e-01) * Income + (6.381e-02) * Wealth + e

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

Residuals:
    Min      1Q  Median      3Q     Max
-4991.0 -1201.0  -166.8  1059.5  7281.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.115e+03  2.054e+03  -1.517    0.133
Age          3.019e+02  6.222e+01   4.852 4.61e-06 ***
Income       2.119e-01  3.425e-02   6.188 1.42e-08 ***
Wealth       6.381e-02  1.102e-02   5.789 8.52e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2132 on 98 degrees of freedom
Multiple R-squared:  0.9417,    Adjusted R-squared:  0.9399
F-statistic: 527.7 on 3 and 98 DF,  p-value: < 2.2e-16
```

2f)

The value of adj-R2 (0.9399 or 93.99%) coefficient indicates the quantity of the variation in Balance explained by the regression line. At this time, Adj-R2 (0.9399 or 93.99%) of the variation in Balance is explained by Age, Income and Wealth.

A high value of R2 does not necessarily mean that the regression model is a good fit for the data because R2 will always take a high value even if the variables have no effect on balance. However, Adj-R2 (0.9399 or 93.99%) does not increase with the addition of a x-variable that does not improve the regression model. A higher Adj-R2 (0.9399 or 93.99%) typically indicates a better model.