DSC 423: Data Analysis and Regression

Assignment 3:

Name: Adarsh Shankar

Student ID: 2117611

Honor Statement: “I have completed this work independently. The solutions given are entirely my own work.”

Question 1. Short Essay (20 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. (10 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.

Assumption 1: Linear Relationship

The fundamental presumption of linear regression is that the independent variables, x and y, have a linear relationship to one another.

Making an x vs. y scatter plot is the simplest technique to check if this premise is true. This enables you to visually determine whether the two variables have a linear relationship. This assumption is satisfied if it appears that the plotted points might all lie along a straight line, indicating that the two variables have some kind of linear relationship.

Assumption 2: Independence

The residuals must be independent, which is the second condition made by linear regression. When working with time series data, this is mostly pertinent. In a perfect world, there shouldn't be a pattern in consecutive residuals. For instance, residuals shouldn't progressively increase with time.

Looking at a residual time series plot, which is a plot of residuals vs. time, is the simplest way to determine whether this premise is true. The majority of the residual autocorrelations should, under ideal circumstances, be contained inside the 95% confidence intervals surrounding zero, which are situated at approximately +/- 2-over the square root of n, where n is the sample size. The

Durbin-Watson test can also be used to formally determine if this presumption is true.

Assumption 3: Homoscedasticity

The residuals must have a constant variance at every level of x, according to the subsequent assumption of linear regression. We call this homoscedasticity. The residuals are considered to be heteroscedastic when this is not the case.

Regression analysis results become difficult to trust when heteroscedasticity is present. In particular, heteroscedasticity raises the variance of estimates for the regression coefficients, but the regression model misses this. This increases the likelihood that a regression model may claim that a term is statistically significant when it actually isn't.

Assumption 4: Normality

The residuals must follow a normal distribution, which is the second assumption made by linear regression.

There are two typical approaches to determine whether this premise is true:

1. Visually verify the assumption using Q-Q charts.

We can use a particular form of figure called a Q-Q plot, which stands for quantile-quantile plot, to determine whether or not the residuals of a model have a normal distribution. The normalcy assumption is satisfied if the points on the plot generally form a straight diagonal line.

2. You can also use formal statistical tests like Shapiro- Wilk, Kolmogorov-Smironov, JarqueBarre, or D'Agostino-Pearson to confirm the normality assumption. However, remember that these tests frequently draw the conclusion that the residuals are not normal because of how sensitive they are to big sample sizes.

Citation:

Zach, The Four Assumptions of Linear Regression, from

https://www.statology.org/linear- regression-assumptions/

b. (10 pts) Define ‘interaction term’. From your own experience, identify an instance in which you believe an interaction term would be appropriate.

Ans:

When one variable's effect depends on the value of another, this is known as an interaction effect. This is an example of an interaction term.

Food Condiment as an Example

We'll just include two items and two condiments in our analysis—hot dogs and ice cream— to keep things straightforward.

The example's specifics suggest that an interaction effect is not unexpected. If someone were to ask, "Which would you rather have on your dish, ketchup or chocolate sauce?" You will undoubtedly reply, "It depends on the meal type!" An interaction effect's "it depends" quality is demonstrated by this. Without knowing more details about the second variable in the interaction term—in our example, the type of food—you cannot correctly answer the question!

Question 2. BANKING (30 pts.) 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:

- Median age of the population (Age)
- Median years of education (Education)
- Median income (Income) in $
- Median home value (HomeVal) in $
- Median household wealth (Wealth) in $
- Average bank balance (Balance) in $

Ans:

banking <- read.csv("C:/Users/Adarsh/Desktop/banking.csv")

head(banking)

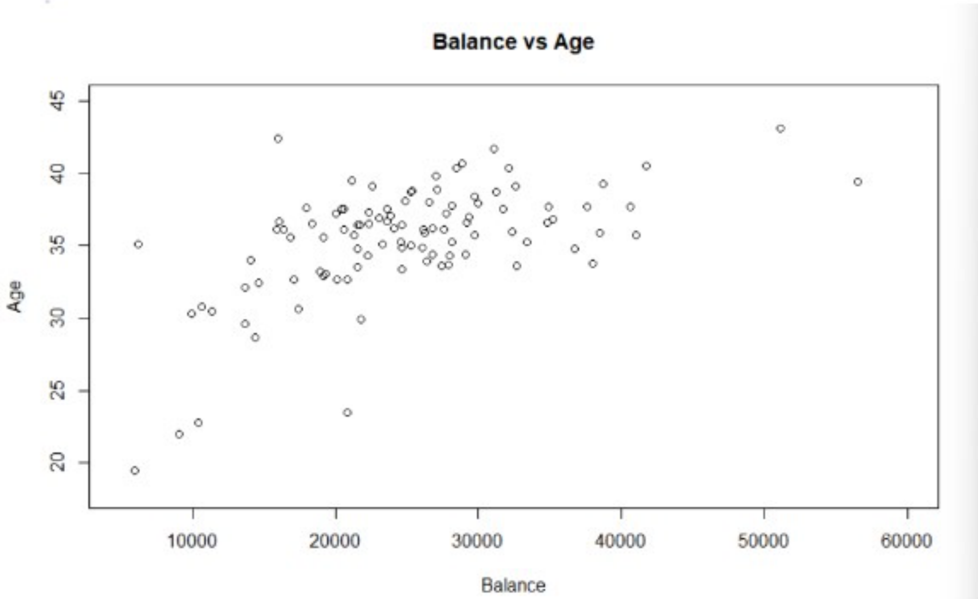
	Age	Education	Income	HomeVal	Wealth	Balance
1	35.9	14.8	91033	183104	220741	38517
2	37.7	13.8	86748	163843	223152	40618
3	36.8	13.8	72245	142732	176926	35206
4	35.3	13.2	70639	145024	166260	33434
5	35.3	13.2	64879	135951	148868	28162
6	34.8	13.7	75591	155334	188310	36708

a. (5 pts.) 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 (5 in total) into your submission. Describe the relationships.

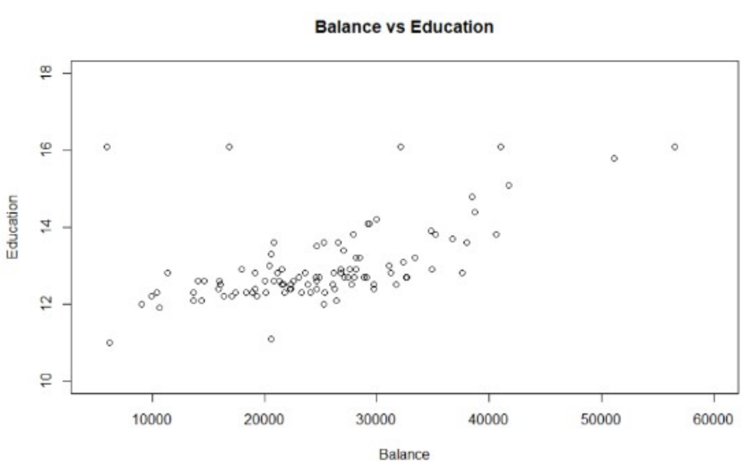
```
> summary(banking)
      Age          Education         Income          HomeVal           Wealth          Balance     
 Min.   :19.50   Min.   :11.00   Min.   :  7741   Min.   : 40313   Min.   : 24999   Min.   : 5956  
 1st Qu.:33.92   1st Qu.:12.40   1st Qu.: 35078   1st Qu.: 83017   1st Qu.: 70263   1st Qu.:20036  
 Median :36.10   Median :12.70   Median : 47656   Median : 97744   Median :102348   Median :24661  
 Mean   :35.45   Mean   :12.98   Mean   : 48811   Mean   :106845   Mean   :109026   Mean   :24888  
 3rd Qu.:37.58   3rd Qu.:13.20   3rd Qu.: 60157   3rd Qu.:121791   3rd Qu.:142518   3rd Qu.:29180  
 Max.   :43.10   Max.   :16.10   Max.   :111548   Max.   :276139   Max.   :331009   Max.   :56569  
```

Ans:

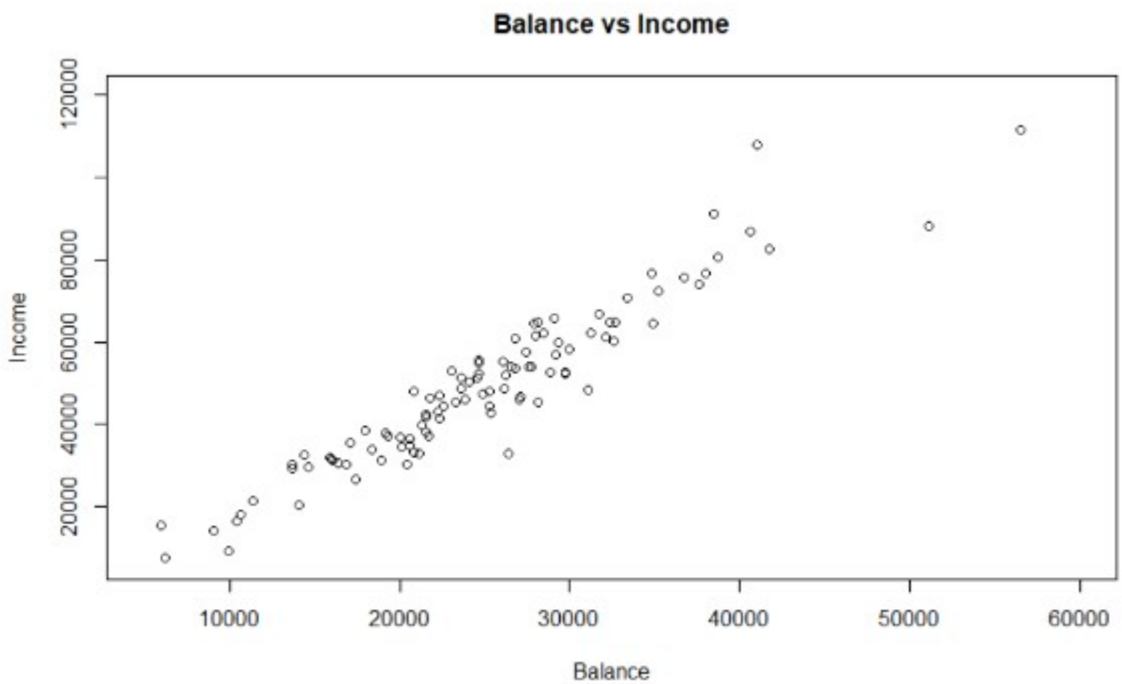
```
> plot(x = banking$Balance, y = banking$Age,
+      xlab = "Balance",
+      ylab = "Age",
+      xlim = c(5000, 60000),
+      ylim = c(18, 45),
+      main = "Balance vs Age"
+ )
```



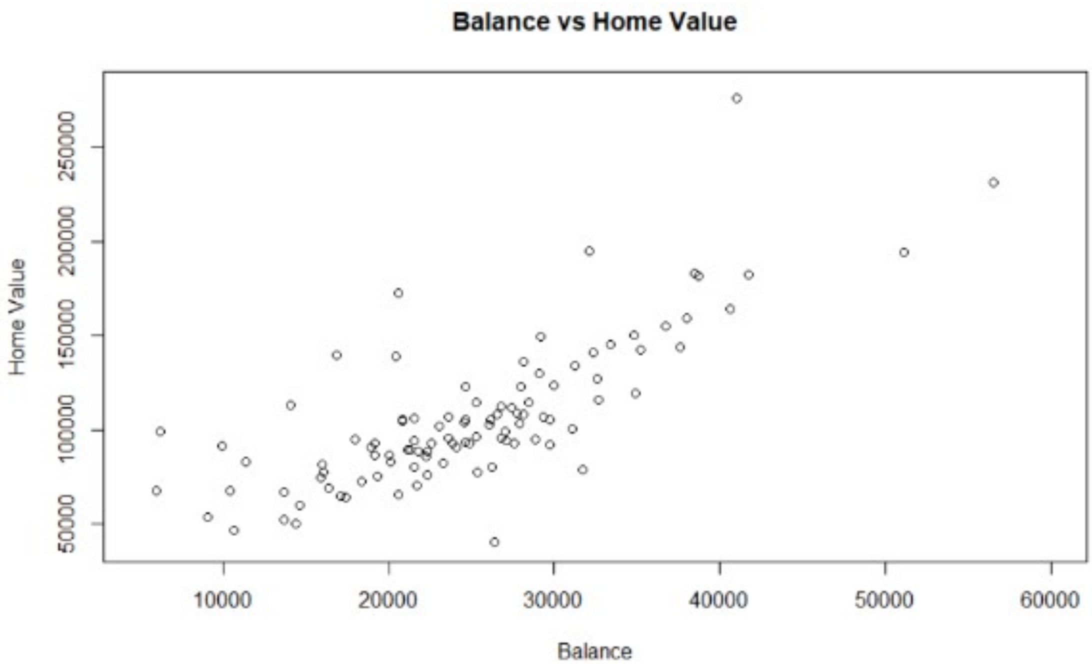

```
> plot(x = banking$Balance, y = banking$Education,
+       xlab = "Balance",
+       ylab = "Education",
+       xlim = c(5000, 60000),
+       ylim = c(10, 18),
+       main = "Balance vs Education"
+ )
```



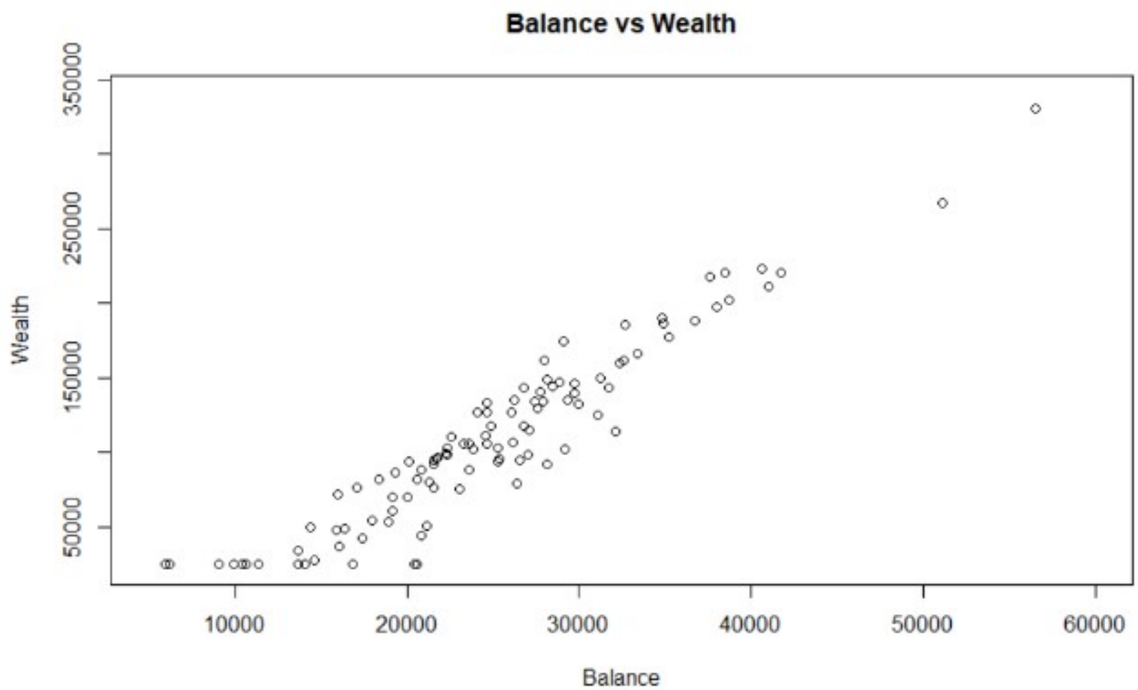
```
> plot(x = banking$Balance, y = banking$Income,
+       xlab = "Balance",
+       ylab = "Income",
+       xlim = c(5000, 60000),
+       ylim = c(7000, 120000),
+       main = "Balance vs Income"
+ )
```




```
> plot(x = banking$Balance, y = banking$HomeVal,
+       xlab = "Balance",
+       ylab = "Home Value",
+       xlim = c(5000, 60000),
+       ylim = c(40000, 280000),
+       main = "Balance vs Home Value"
+ )
```




```
> plot(x = banking$Balance, y = banking$Wealth,
+       xlab = "Balance",
+       ylab = "Wealth",
+       xlim = c(5000, 60000),
+       ylim = c(24000, 340000),
+       main = "Balance vs Wealth"
+ )
```



b. (5 pts.) 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 for the bank data. Paste them into your submission. Describe which variables appear to be strongly associated? Interpret any correlation values you deem important.

Ans:

```
> correlation <- cor(banking)*100
> correlation
                Age Education    Income   HomeVal    Wealth   Balance
Age       100.00000  17.34611  47.71474  38.64931  46.80918  56.54668
Education  17.34611 100.00000  57.31467  74.89426  46.81199  55.21889
Income     47.71474  57.31467 100.00000  79.53552  94.66654  95.16845
HomeVal    38.64931  74.89426  79.53552 100.00000  69.84778  76.63871
Wealth     46.80918  46.81199  94.66654  69.84778 100.00000  94.87117
Balance    56.54668  55.21889  95.16845  76.63871  94.87117 100.00000
> |
```

To make the correlation values easier to read, I increased them by 100.

The strength of the linear link increases with greater correlation r values. The ones in bold have a clear linear relationship.

Homeval and balance have a moderately significant correlation.

Others who score below 70 have a poor correlation between them.

c. (5 pts.) Fit a single 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.

Ans:

```
> model1 <- lm( banking$Balance ~ banking$Age + banking$Education +banking$Income + banking$Homeval + banking$Wealth )
> summary(model1)

Call:
lm(formula = banking$Balance ~ banking$Age + banking$Education +
    banking$Income + banking$Homeval + banking$Wealth)

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 *
banking$Age        3.175e+02  6.104e+01   5.201 1.12e-06 ***
banking$Education  5.903e+02  3.151e+02   1.873 0.064085 .
banking$Income     1.468e-01  4.083e-02   3.596 0.000512 ***
banking$Homeval    9.864e-03  1.099e-02   0.898 0.371591
banking$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

>
```

d. (5 pts.) Which of the five predictors have a significant (a=.05) effect on balance? Explain.

Ans:

Age, wealth, and income all have a big impact on balance. If the P-Value is less than 0.05 and the test has a 95% confidence interval or a 5% level of significance, we reject the null hypothesis. We reject the null hypothesis when the P-value is less than the test's level of significance.

A P-value of 2.0 X 10-09, 0.000512, and 1.1 X 10-06 is much lower than a P-value of 0.05 when doing a test with 95% confidence or 5% significance. The null hypothesis will be disproved as a result. When using the P-Value technique to testing hypotheses, If the P-value is less than the level of significance, which is Alpha = 0.05, the judgment rule is to reject the null hypothesis.

e. (5 pts.) 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. Analyse 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 the remaining predictors are significant. Present the final regression model.

Ans:

After dropping home Value

```
> model3 <- lm( banking$Balance ~ banking$Age + banking$Education +banking$Income + banking$Wealth )
> summary(model3)

Call:
lm(formula = banking$Balance ~ banking$Age + banking$Education +
    banking$Income + banking$Wealth)

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 **
banking$Age        3.242e+02  6.051e+01   5.358 5.68e-07 ***
banking$Education  7.498e+02  2.600e+02   2.884  0.00484 **
banking$Income     1.615e-01  3.738e-02   4.321 3.75e-05 ***
banking$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
```

f. (5 pts.) Interpret each of the regression coefficients for the final model. Discuss the adjR 2 for the final model. Is this a good model? Explain.

Ans:

The proportion of variation in the dependent variable (outcome) that can be explained by the predictor variables (factors) included in a regression model is quantified by the adjusted R-squared (adj-R2), a statistical metric. It is a modified version of the standard R-squared (R2) that considers the sample size and the number of predictor variables and is frequently used as a measure of a regression model's goodness-of-fit.

The adj-R2 score of 0.9441, or 94%, indicates the significance of variation and indicates that the predictor variables used in the regression model can account for 94% of the variance in the dependent variable. This reveals that the predictor variables are crucial in explaining the variation in the dependent variable since it shows that the model can explain a significant percentage of the variability in the outcome variable.