Keiland Pullen
ID:1977120

DSC 423 | Spring 2021
Assignment 2

Honor statement:
"I have completed this work independently. The solutions given are entirely my own work"

1a.)

First, the residuals of a regression model can be defined as the errors of a model. According to the web blog DISPLAY R (https://www.displayr.com/learn-what-are-residuals) the residuals are "the differences between observed and predicted values of data. The residuals are a measure that assists with evaluating the quality of a model.

When building a regression model, the following four assumptions are made about the residuals or errors; the mean of the errors is zero, the errors are homoscedatic, the errors are normal and the errors are independent.

The mean or average of errors is zero. A mean of zero tells us that any changes with the dependent variable are caused by the independent variables. It also tells us that there is no bias or influence placed on the dependent variable. If the mean is not zero or any number, then our model will not have accurate predictions. Homoscedatic means that the error's variance is constant. This can be visually observed on a plot of the data. The variance of the errors should not change for any data points or group of data points. For example, if the variance of the data points is spread wide on a plot, then this indicates that the model will not work well. The errors are normal means that there is a normal distribution of the errors along a regression line. For example, a plot containing a regression line would display approximately one-half of the errors above the line and one-half below the line. A normal error distribution is a good indicator that the regression model will be unbiased. The errors are independent, means that the errors are free from influence from other errors. An article on Statology.com (https://www.statology.org/linear-regression-assumptions/) provides an excellent example with stating that the errors shouldn't form any type of pattern. The article using the following example, "residuals shouldn't steadily grow larger as time goes on."

1b.)

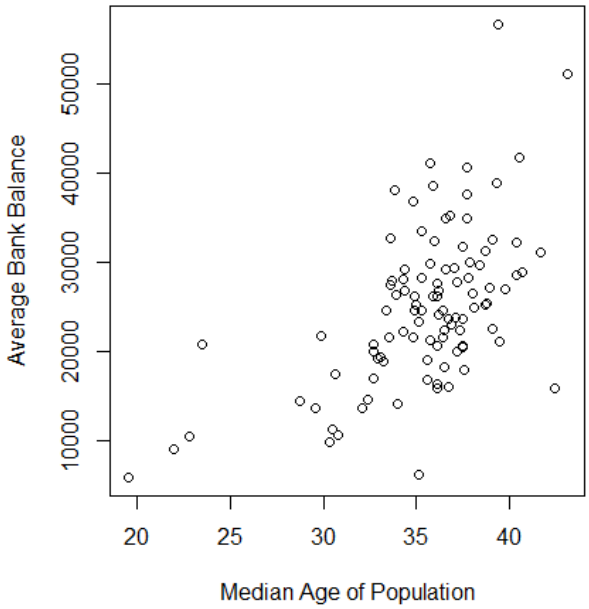
An interaction term is created when an interactive term, that has an effect on a dependent variable, is dependent on another interactive term. This definition is provided by the web-site "Stat Trek" (https://stattrek.com/multiple-regression/interaction.aspx). In simpler terminology, an interactive term can be described as the interaction of two or more independent variables.

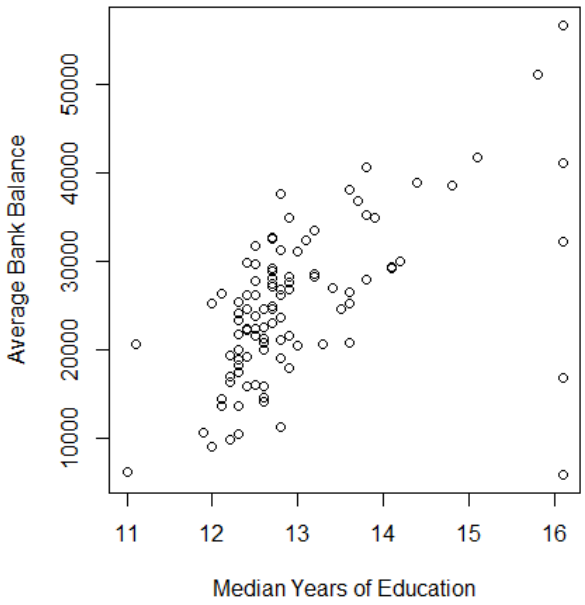
A number of examples of interactive terms can be found in a family's home medicine cabinet. If a child becomes ill with cold symptoms, the prescription may be bottle of cough syrup. The bottle dosage prescribed depends on the age of the child. In this example, the dependent term would be "cough free" or "healthy". The independent terms would be dosage and age. We know that to be "cough free" a certain amount of dosages are needed, but the amount of dosages are dependent upon the age of the child.

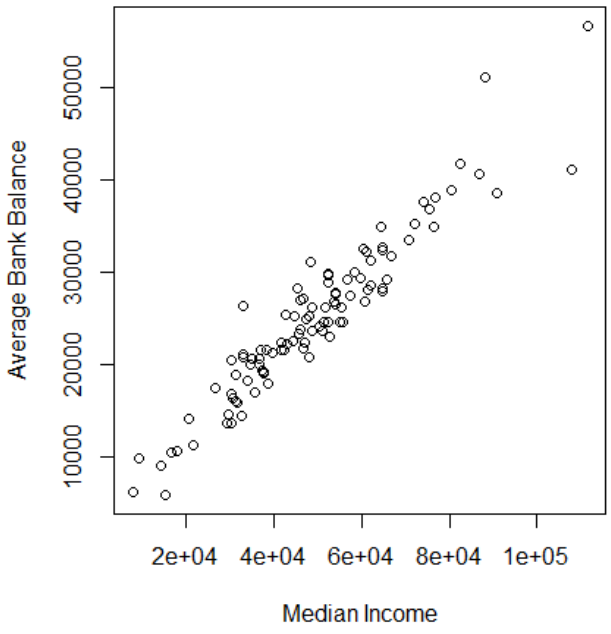
Another example, that I have experience with is weight-loss. The dependent variable is always "number of pounds lost". The independent variables are typically exercise, time and diet. The interaction terms are a combination of time and exercise. For example, I know that X amount of calories can be lost (pounds lost variable), if I run on the treadmill (exercise variable) for 30 minutes (time variable).

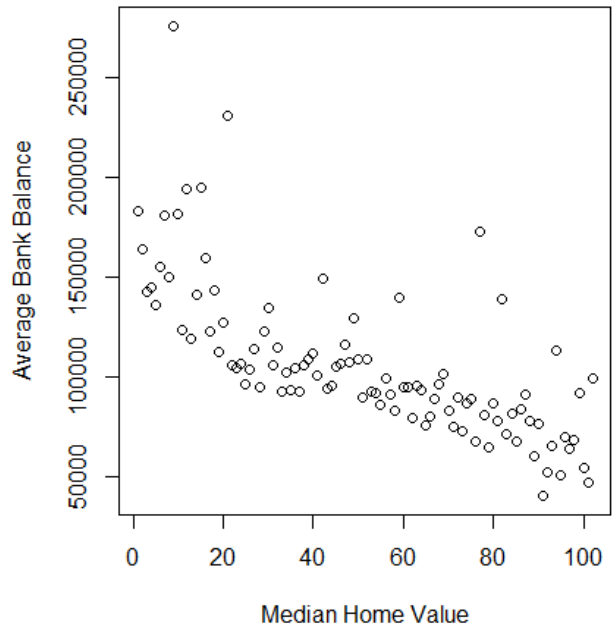
2.) BANKING

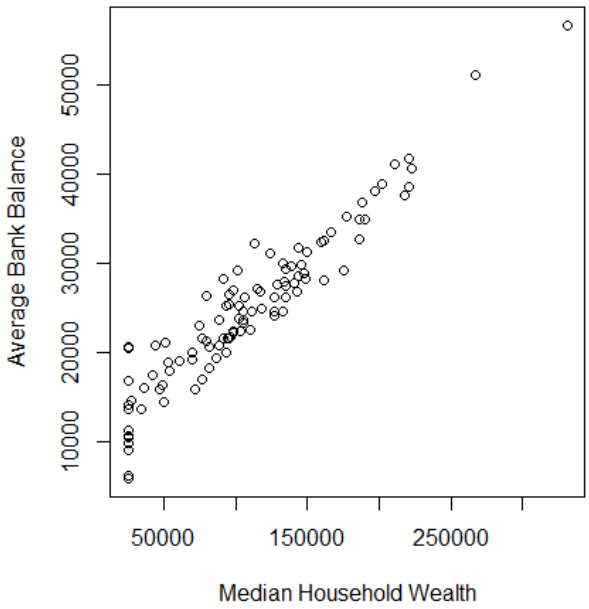
2c.) Scatterplots











The scatter plot for the “Median Age of Population” vs “Average Bank Balance” tells us that the direction is positive, the form is non-linear and the strength is moderate. This plot has several outliers as well.

The scatter plot for the “Median Years of Education” vs “Average Bank Balance” has a positive direction, the form is linear and the strength is moderate and it contains several outliers.

The scatter plot for the “Median Income” vs “Average Bank Balance” has a positive direction in addition to being linear. The strength is strong and there are a small number of outliers.

The scatter plot for the “Median Home Value” vs “Average Bank Balance” has a negative direction. Along with being linear, it has a strength that is moderate to weak.

The scatter plot for the “Median Household Wealth” vs “Average Bank Balance” is linear with a positive direction and its strength is quite strong.

2d.)

```
> d <- BANKING[,]   # d is our data frame
> cor(d)
```

	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 variables that are strongly related are HomeVal and Education with a correlation coefficient or R value of 0.7489426. The variables Income and HomeVal also have a strong relationship with an R value of 0.8953552. Variables Income and Wealth display the strongest relationship with an R value of .9466654. The variables Wealth and HomeVal could also be included as it has an R value of 0.6984778.

2e.)

```
Call:
lm(formula = Balance ~ ., data = d)

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
```

The estimate regression model has a p-value of 2.2e-16 which is very small and tells us that we should reject the null hypothesis and accept the alternative that at least one Beta will be equal to zero. The t-test for Education is not good, because Education could be considered a categorical variable in this case. The values appear to be number of years of education. The t-tests look good for three of the variables; Age, Income, and Wealth. The adjusted R-squared is 94% which means that this model explains 94% of the variability in y.

2f.)
The predictors of Age, Income and Wealth seem to have the most significance on Balance. The t-tests for Wealth and Age are very close to 0 and the t-test for Income isn't bad. Also, the asterisks indicate the each variable is significant.

2g. & 2h.)

```
> model2 <- lm(Balance ~. - Education, data = d)
> summary(model2)

Call:
lm(formula = Balance ~ . - Education, data = d)

Residuals:
   Min      1Q  Median      3Q     Max
-6003.8 -1068.6  -175.8  1027.7  8037.9

Coefficients:
             Estimate      Std. Error    t value      Pr(>|t|)
(Intercept) -3.358e+03    2.011e+03     -1.669       0.098275 .
Age          2.975e+02     6.088e+01     4.888       4.04e-06 ***
Income       1.566e-01     4.102e-02     3.819       0.000236 ***
HomeVal      2.147e-02     9.192e-03      2.336       0.021551 *
Wealth       7.115e-02     1.123e-02      6.337       7.38e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2085 on 97 degrees of freedom
Multiple R-squared:  0.9448,  Adjusted R-squared:  0.9425
F-statistic: 415.1 on 4 and 97 DF,  p-value: < 2.2e-16
```

For this model, the t-tests look good and the adjusted R-squared tells us that this model explains 94.25% of the variability of y. The Age, Income and Wealth variables each look good. The HomeVal variable will be removed as it has the worse significance on Balance.

2i.)

```
> model2 <- lm(Balance ~. - Education - HomeVal, data = d)
> summary(model2)

Call:
lm(formula = Balance ~ . - Education - HomeVal, data = d)

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
```

In this final model, the remaining three variables each pass the t-test, the p-value looks good and each appears to have a significant impact on the dependent variable. The variable with the most impact appears to be Wealth.

2j.)

The adjusted R-squared value of the final model is 93.99 percent. However, it should be noted that this value is lower than the R-squared values for each of the previous models. At 94%, this model is considered a good-fit. This model explains 93.99% of the variability in y.

2k.)

Each of the scatter plots displayed several data points that were outliers. Each of those outliers can be considered as influential points. These influential points can increase the variability in the data and thus weaken the model. Influential points can be removed if they in error, but in some cases their values can be substituted with the mean of the respective data set.

3a. & 3b.)

In this model, there were originally 8 variables in the dataset. Interaction terms were created along with second order terms. The number of variables expanded to 36.

Model 1:

Call:

lm(formula = Voltage ~ . + timeTemp + salTemp + surfTemp + spanTemp + solidTemp + timeSal + timeSurf + timeSpan + timeSolid + salSurf + salSpan + salSolid + salDel + volSal + volTemp + volDel + volSurf + volSpan + volSolid + surfSpan + surfSolid + vol2 + sal2 + temp2 + del2 + surf2 + span2 + solid2, data = d)

Residuals:

1	2	3	4	5	6	7	8	9	10	11	12	13
0.05500	-0.05500	-0.05500	0.05500	0.05500	-0.05500	-0.05500	0.05500	-0.05500	0.05500	0.05500	-0.05500	-0.05500

14	15	16	17	18	19
0.05500	0.05500	-0.05500	0.01333	0.01333	-0.02667

Coefficients: (20 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.0666667	0.0741370	14.388	0.000728 ***
Volume	-0.0474921	0.0054964	-8.641	0.003261 **
Salinity	0.6400000	0.0676775	9.457	0.002506 **
Temperature	0.0521352	0.0227382	2.293	0.105680
Delay	0.0809839	0.0203360	3.982	0.028335 *
Surfactant	1.1800000	0.1015163	11.624	0.001368 **
SpanTriton	1.1105263	0.4451813	2.495	0.088129 .
SolidPart	-4.0906671	0.5294711	-7.726	0.004508 **
volSal	-0.0078333	0.0010701	-7.320	0.005266 **
volTemp	0.0002895	0.0001690	1.713 0	.185178
volDel	0.0003368	0.0001352	2.492	0.088325 .
volSurf	-0.0120000	0.0016051	-7.476	0.004956 **
volSpan	-0.0160000	0.0064205	-2.492	0.088325 .
volSolid	0.0660667	0.0088768	7.443 0	.005021 **
timeTemp	-0.0081197	0.0012411	-6.542	0.007259 **
salTemp	NA	NA	NA	NA
surfTemp	NA	NA	NA	NA
spanTemp	0.0273684	0.0135167	2.025	0.136027
solidTemp	NA	NA	NA	NA
timeSal	NA	NA	NA	NA
timeSurf	NA	NA	NA	NA
timeSpan	NA	NA	NA	NA
timeSolid	NA	NA	NA	NA
salDel	NA	NA	NA	NA
salSurf	NA	NA	NA	NA
salSpan	NA	NA	NA	NA
salSolid	NA	NA	NA	NA
surfSpan	NA	NA	NA	NA
surfSolid	NA	NA	NA	NA
vol2	NA	NA	NA	NA
sal2	NA	NA	NA	NA
temp2	NA	NA	NA	NA
del2	NA	NA	NA	NA
surf2	NA	NA	NA	NA
span2	NA	NA	NA	NA
solid2	NA	NA	NA	NA

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1284 on 3 degrees of freedom

Multiple R-squared: 0.9952, Adjusted R-squared: 0.9712

F-statistic: 41.53 on 15 and 3 DF, p-value: 0.005279

Model 4:

```
Call:
lm(formula = Voltage ~ . - Temperature - Delay - SpanTriton -
    SolidPart - Salinity - timeTemp - volSal - volTemp - volSurf -
    volSolid - volDel - volSpan - salTemp - surfTemp - spanTemp -
    solidTemp - timeSal - timeSurf - timeSpan - timeSolid - salSurf -
    salSpan - salSolid - salDel - surfSpan - surfSolid - vol2 -
    sal2 - temp2 - del2 - surf2 - span2 - solid2, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-0.86384 -0.25726  0.01945  0.25781  1.35945

Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  1.01041    0.26326     3.838    0.001451 **
Volume       -0.02208   0.00512     -4.313   0.000536 ***
Surfactant   0.42836    0.10240     4.183    0.000703 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.516 on 16 degrees of freedom
Multiple R-squared:  0.5872,  Adjusted R-squared:  0.5356
F-statistic: 11.38 on 2 and 16 DF,  p-value: 0.000843
```

3c.)

For this first model, there were originally 8 variables. Different combinations of the variables were combined to create interaction variables. Each of the variables was squared to create second-order term variables. The first model has an Adjusted R-square value of 97% suggests that this model may be a good fit for the data. This means that this model explains 97% of the variability in Y. The p-value is less than .05 and several of the variables passed the t-test. A number of the interactive variables returned a value of N/A. This means that the variables are linearly related to each other. As such, those variables will need to be dropped.

After dropping the insignificant variables and those with a value of N/A, a second model was built. The p-values, t-tests and adjusted R-squared values were analyzed. The values of each seemed to worsen with an Adjusted R-square value of 71.85% and p-value of 0.006375. The insignificant terms were dropped again, for the third model. Again, the Adjusted R-square value decreased and the p-value lowered to 0.003077. Finally, a fourth model was built by removing the last of the insignificant terms – which left two lone terms. The Adjusted R-squared value was 53.56%, which means that this model has a little over a 50% chance of being a good fit for the data. The p-value was 0.000843 and the t-tests were good.

In this case, I would perform tests with both models and compare their outputs to determine which would be the best fit for the data.