**Assignment 03 (423: Data Analysis and Regression)**

**Saransh Thakur**

**Id: 2020070**

**Problem 1**

1. Residual is the difference between an observed value of the response variable and the value predicted by the regression line.

The assumption about residual: -

1. **Normality of residual: -** It tells if a dataset is well modeled by a normal distribution. It means to say residual should be normally distributed, then regression line is a good fit.
2. **Unbiased: -** It is assumed that the residuals have an average value of zero means the error terms are normally distributed around zero.
3. **Homoscedastic: -** Means “same stretch”, the spread of the residuals should be the same.
4. **Independent error assumption:** - Residual terms are independent of each other. This means that there is no correlation between the residuals and the predicted values.
5. An interaction term is represented as the product of two or more independent variables. An interaction occurs when an independent variable has a different effect on the outcome depending on the values of another independent variable.

ŷ = b0 + b1X1 + b2X2 + **b3X1X2**

**b3X1X2**is an interaction term in the above equation.

Example: -

1. Interaction between adding sugar to coffee and stirring the coffee. Neither of the two individual variables has many effects on sweetness but a combination of the two does.
2. Interaction between smoking and inhaling asbestos fibers: Both raise lung carcinoma risk, but exposure to asbestos multiplies the cancer risk in smokers and non-smokers. Here, the joint effect of inhaling asbestos and smoking is higher than the sum of both effects.

**Problem 2**

Chart, scatter chart

Description automatically generated

The scatter plot tells us that it has a positive direction and has linear form, but it has weakstrengthand has a lot of outliers.

Chart, scatter chart

Description automatically generated

The scatter plot tells us that it has a positive direction and has linear form, but it has a moderatepositive relationship and has a lot of outliers.

Chart, scatter chart

Description automatically generated

The scatter plot tells us that it has a positive direction as well as a linear form, and it has a strong positive correlation with no outliers.

Chart, scatter chart

Description automatically generated

The scatter plot tells us that it has a linear form and positive direction and has a moderate positive correlation and has outliers.

Chart, scatter chart

Description automatically generated

The scatter plot tells us that it has a positive direction as well as linear form and has a strong positive correlation with no outliers.

A picture containing text

Description automatically generated

1. Balance and Income appear to be strongly associated. They both correlate 95.2%.
2. Wealth and balance are also strongly associated. They both correlate 94.8%.
3. Wealth and Income appear to be strongly associated. They both correlate 94.6%.
4. Income and homeval correlate 79.5%.
5. Balance and homeval correlate at 76.6%

A screenshot of a computer

Description automatically generated with low confidence

F-test has a very low P-value, this means that we reject the null hypothesis i.e. (all the Betas are equal to zero) and accept the alternative hypothesis that at least one is not equal to zero.

It tells us that something in the model is working.

To check, the significance of (a=.05) effect on balance we will use a t-test.

Ho (Null Hypothesis) – B1 = 0

H1 (Alternative Hypothesis) – B1=/= 0

For, **Wealth** we reject the null hypothesis and accept the alternative hypothesis, so we are going to use the prediction of 7.414e-02

For, **HomeVal** we fail to reject the null hypothesis and we will remove HomeVal from our model.

For, **Income** we reject the null hypothesis and accept the alternative hypothesis, so we are going to use the prediction of 1.468e-01

For, **Education** we fail to reject the null hypothesis and we will remove HomeVal from our model.

For, **Age** we reject the null hypothesis and accept the alternative hypothesis, so we are going to use the prediction of 3.175e+02

1. **We need to remove the home value from the model because the home value has the largest p-value.**

**A close-up of a document

Description automatically generated with low confidence**

After removing the home value from our model, the t-test for the remaining four variables is under a significant value of 0.05 (5%), which means we need not remove any other variable from our model.

1. The regression coefficients for the final model are: -

B0 (y-intercept) is -1.214e+04

B1 (Age) is 3.242e+02

B2 (Education) is 7.498e+02

B3 (Income) is 1.615e-01

B4 (Wealth) is 7.265e-02

A screenshot of a computer

Description automatically generated with low confidence

The adjusted R-squared is 0.9441 which means 94.41% variability in y is explained by our model.

At last, we may conclude that this is a very good model with a low P-value and a high R-squared value.