Lesson X Practical Exercise

STUDENT NAME

School of Public Affairs: Pennsylvania State University

PADM 504, Section 001: Data Analysis for Policy and Administration

Professor Lauren Azevedo

DATE

**Purpose**

Orient the reader to the purpose of your analysis. Provide the research question and how you are answering it.

**Method**

To evaluate multicollinearity, non-linearity and non-normality of the regression model the following methods are used:

**Variance Inflation Factor (VIF):**

The amount of multicollinearity in a set of independent variables is measured by VIF. If the VIF for a variable is more than 10, there is a strong chance of multicollinearity.

**Non-linearity:**

The detection non-linearity is done by a scatter plot between dependent & independent variables. Nonlinearity is commonly addressed by using polynomial factors in the regression model, such as squared, cubed, logarithm, exponential predictors.

**Non-normality:**

Q-Q plot is effective visual tool for detecting nonnormality. If a variable is normally distributed, the data points in the plot should cluster around the straight line.

**Software:**

R (version 4.2.0) is used to analyze the data and missing values are excluded before the analysis.

**Results**

Table-01: Multiple Linear Regression Model of Female Life Expectancy

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Coefficients | Estimate | Std. Error | t statistic | p-value | R square | Adjusted R square |
| Intercept | 53.9899 | 1.5544601 | 34.732 | 0.0000 \*\*\* | 0.6781 | 0.6589 |
| Urban | 0.1164 | 0.0360 | 3.2310 | 0.0016 \*\* |  |  |
| GDP | 0.0001 | 0.0003 | 0.3700 | 0.7121 |  |  |
| Radio | 0.0376 | 0.0277 | 1.3580 | 0.1775 |  |  |
| Hospbed | 0.0120 | 0.0332 | 0.3620 | 0.7182 |  |  |
| Phone | 0.0196 | 0.0911 | 0.2150 | 0.8299 |  |  |
| Docs | 0.4279 | 0.1103 | 3.8800 | 0.0002 \*\*\* |  |  |

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

F-statistic: 35.45 on 6 and 101 DF, p-value: < 0.0001

The p-value of the F statistic is less than 0.05 which imples that the fitted regression model is overall okay. From the table-01, we can see that only urban & docs variable is statistically significant at the 5% level of significance.

The intercept term is 53.9899 implies that if there is no contribution of any regressors then the female life expectancy will be 53.9899 years, on average. The p-value of this coefficient is less than 0.05 which is statistically significant at the 5% level of significance.

The coefficient of Urban is 0.1164 implies that if percentage of urbanization is increased by 1 percent then the female life expectancy will be increased by 0.1164 years, on average. The p-value of this coefficient is 0.0016 which is less than 0.05, that means, this coefficient is statistically significant at 5% level of significance.

The coefficient of GDP is 0.0001 implies that if GDP per capita increased by 1 dollars then the female life expectancy will be increased by 0.0001 years, on average. The p-value of this coefficient is 0.7121 which is greater than 0.05, that means, this coefficient statistically insignificant at 5% level of significance.

The coefficient of radio is 0.0376 implies that if radios per thousand people is increased by increased by 1 unit, then the female life expectancy will be increased by 0.0376 years, on average. The p-value of this coefficient is 0.1775 which is not statistically significant at 5% level of significance.

The coefficient of Hospbed is 0.0120 which implies that if hospital beds per 10000 people is increased by 1 unit, then the female life expectancy will be increased by 0.0120 years, on average. The p-value of this coefficient is 0.7182 which is not statistically significant at 5% level of significance.

The coefficient of phone is 0.0196. This implies that if phones per 1000 people is increased by 1 unit then the female life expectancy will be increased by 0.0196 years, on average. The p-value of this coefficient is 0.8299 which is not statistically significant at 5% level of significance.

The coefficient of docs is 0. 4279. This imples that if the doctors per 10000 people is increased by 1 unit then the female life expectancy will be increased by 0. 4279 years, on average. The p-value of this coefficient is 0.0002 which is statistically significant at 5% level of significance.

To get the best model for the above case, we have to make some transformations.

The female life expectancy is plotted against docs, gdp, hospbed, phone, radio & urban. From the plot, we can notice that all of the variable have non-linear relation with the female life expectancy. Since it is noticed from the figure that the independent variables are exponentially increases as the female life expectancy increases. To make the regression model linear, we have to make logarithm transformation of the independent variables.

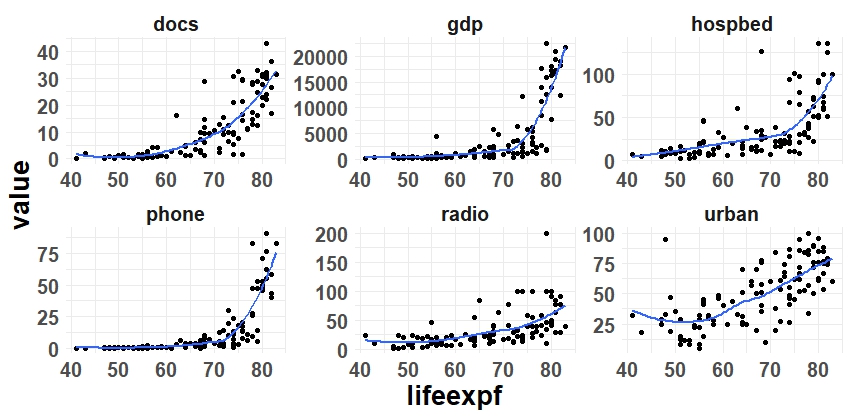


Figure: Scatter plot between response variable & regressors variables

Table-02: Variance Inflation Factor of Regressors

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Coefficients | Urban | GDP | Radio | Phone | Hospbed | Docs |
| VIF | 2.0658 | 11.2313 \* | 1.8770 | 10.8907 \* | 2.8096 | 3.9607 |

Codes: ‘\*’ presence of multicollinearity

Since VIF of GDP & Phone is greater than 10, these two variables are responsible for multicollinearity.

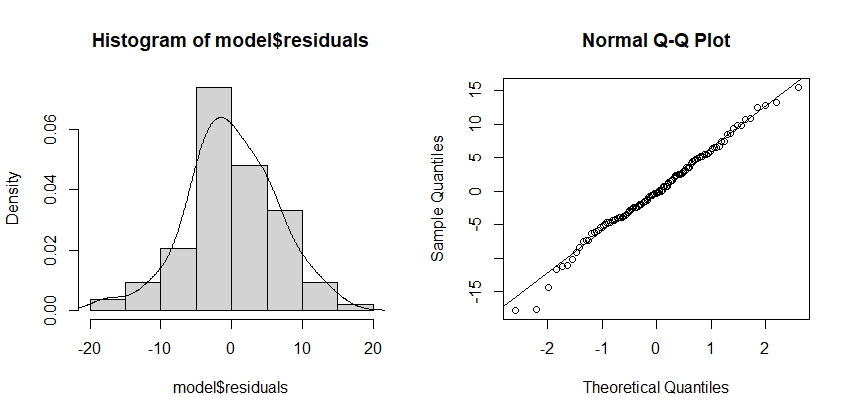


Figure: Q-Q plot of residuals

From the histogram of the residuals it can be seen that it looks approximately normal. From the normality test, it can be noticed from the qqplot of the residuals is that approximately all the points falls on the reference line which indicate that the residuals are normally distributed. Thus, the normality assumption is not violated.

To remove the multicollinearity from the regression model, we can omit the variables which are responsible for multicollinearity. Similarly, To remove the non-linearity from the regression model, we can make logarithm transformation of some independent variables.

Table-03: Regression Model of Female Life Expectancy

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Coefficients | Estimate | Std. Error | t statistic | p-value | R square | Adjusted R square |
| Intercept | 33.9718 | 4.1887 | 8.1100 | 0.0000 \*\*\* | 0.7312 | 0.7207 |
| Log(urban) | 3.5233 | 1.1762 | 2.9950 | 0.0034 \*\* |  |  |
| Log(radio) | 2.9029 | 0.8129 | 3.5710 | 0.0005 \*\*\* |  |  |
| Log(hospbed) | 2.1963 | 1.0051 | 2.1850 | 0.0311 \* |  |  |
| Docs | 0.3594 | 0.0827 | 4.3450 | 0.0000 \*\*\* |  |  |

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

F-statistic: 70.04 on 4 and 103 DF, p-value: < 0.0001

Since the p-value of the F statistic is less than 0.05, the overall regression model okay.

The intercept term is 33.9718 implies that if there is no contribution of any regressors then the female life expectancy will be 33.9718 years, on average. The coefficient of Log(urban) is 3.5233 implies that if the logarithm of percentage of urbanization is increased by 1 unit then the female life expectancy will be increased by 3.5233 years, on average. The coefficient of Log(radio) is 2.9029 implies that if the logarithm of radios per 1000 people is increased by 1 unit then the female life expectancy will be increased by 2.9029 years, on average. The coefficient of Log(hospbed) is 2.1963 implies that if the logarithm of hospital beds per 10000 people is increased by 1 unit then the female life expectancy will be increased by 2.1963 years, on average. The coefficient of docs is 0.3594 implies that if the doctors per 10000 people is increased by 1 unit then the female life expectancy will be increased by 0.3594 years, on average.

In the above regression model, p-value of all the coefficient is less than 0.05, then all the coefficient is statistically significant at 5% level of significance.

The value of R square is 0.7312 implies that the approximately 73.12% variation in the female life expectancy can be explained by Log(urban), Log(radio), Log(hospbed), docs.

Table-04: VIF of the transformed variables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients | Log(urban) | Log(radio) | Log(hospbed) | Docs |
| VIF | 1.8131 | 1.8708 | 2.6945 | 2.7216 |

Codes: ‘\*’ presence of multicollinearity

After logarithm transformation of the independent variables and omission of the multicollinear variables, the multicollinearity has been removed from the regression model.

Thus, the table-03 contains the best regression model for the female life expectancy.

**Discussion and Policy Implications/Recommendations**

Most of the practical exercises ask you to offer policy implications or recommendations based on your findings. First discuss the overall meaning of your analysis and then present your policy implications and recommendations to the reader.

**References** (if used)