**Name: Pratik Premkishor Mantri**

**NUID:001353214**

**Module6R**

**Appendix:**

|  |  |
| --- | --- |
| Index | Page no |
| Introduction | 2 |
| Data analysis | 3-11 |
| References | 12 |

**Introduction to Dummy Variable:**

**Categorical variables** are variables that classify observations into groups. They have a limited number of different values, called levels. For example, In the given assignment we have Geographic as a categorical variable. the Geography can be splinted and categorized into different states b grouping. Such as 1 for Washingtons and 0 for Nevada.

Here in Regression analysis requires numerical variables. So, when a Someone wishes to include a categorical variable in a regression model, some steps need to be done to make the results interpretable.

In these steps, the categorical variables are recoded in separate binary variables. This is called “dummy coding” and leads to the creation of a table called **contrast matrix** This is done automatically by statistical software, such as R.

**Categorical variables with two levels**

Here we know the formula for the regression equation, y = b0 + b1\*x. b0 and `b1 are the regression beta coefficients, representing the intercept and the slope, respectively.

Based on the geography variable, we can create a new dummy variable that takes the value:

* 1 if a state is Washington.
* 0 if a state is Nevada.

and use this variable as a predictor in the regression equation.

* b0 + b1 if state is washington.
* bo if state is Nevada.

The coefficients can be interpreted as follow:

1. b0 is the average Death rate among nevada,
2. b0 + b1 is the average death rate among washington.
3. and b1 is the average difference in death rate between washington and Nevada.

Text

Description automatically generated

From the above observation we can say that the

**P=0.0023<0.05** so the data is highly significant between Washington and nevada.

* **Estimate:** the intercept (b0) and the beta coefficient estimate which are associated to each predictor variable
* **Std.Error**: the standard error of the coefficient estimates. This shows the accuracy of the coefficients. The larger the std error, the less confident we are about the estimate.
* **t-value**: the t-statistic, is a coefficient estimate divided by the standard error of the estimate.
* **Pr(>|t|):** The p-value which is corresponding to the t-statistic. The smaller the value of p, the more significant the estimate is.

To investigate the relationship between variables we use correlation. The correlation coefficient is denoted by **' r ' (between -1 and +1)** and quantifies the linear association's direction and strength between the variables. The correlation can be positive or negative between two variables. The sign of the 'r' indicates the direction of the association, and the magnitude of the 'r' indicates the strength.

Its value ranges between -1 (perfect negative correlation: when x increases, y decreases)

A value **closer to 0 suggests a weak relationship between the variables.**

The R squared value is 0.43.

**Subset analysis:**

Here we have taken a subset with Texas as a geography.

Text

Description automatically generated

Here the P-value is 0.01612, so the data between Target deathrate and poverty percent is almost significant to each other.

The scatterplot shows us the effect of predictor on the outcome variable.

Chart, scatter chart

Description automatically generated

The R-squared value is very low, So there is not that much effect on the outcome variable.

Text

Description automatically generated

There are **four** scenarios:

**1) low R-square and low p-value (p-value <= 0.05):**

model does not say much of variation of the data, but it is significant.

**2) low R-square and high p-value (p-value > 0.05):**

: model does not say much of variation of the data and it is not significant (worst scenario)

**3) high R-square and low p-value:**

explains a lot of variation within the data and it is significant (best scenario)

**4) high R-square and high p-value:**

model explains a lot of variation within the data, but it is not significant (model is worthless)

In this the solution is the **worst case** scenario among all.Becasue the p>0.05 and R value is not above 0.02.

**Visualize the data:**

Chart, scatter chart

Description automatically generated

This shows the horizontal line which mean there is almost no relation between x (predictor) and Y (outcome variable)

**Multiple regression:**

Text

Description automatically generated

* **Estimate:** the intercept (b0) and the beta coefficient estimate which are associated to each predictor variable
* **Std.Error**: the standard error of the coefficient estimates. This shows the accuracy of the coefficients. The larger the std error, the less confident we are about the estimate.
* **t-value**: the t-statistic, is a coefficient estimate divided by the standard error of the estimate.
* **Pr(>|t|):** The p-value which is corresponding to the t-statistic. The smaller the value of p, the more significant the estimate is.

After seeing the P<0.05 but R is also low so we can say that model does not say much of variation of the data, but it is significant.

Here we can see the difference between the percentage of white and black people vs death rate.

To investigate the relationship between variables we use correlation. The correlation coefficient is denoted by **' r ' (between -1 and +1)** and quantifies the linear association's direction and strength between the variables. The correlation can be positive or negative between two variables. The sign of the 'r' indicates the direction of the association, and the magnitude of the 'r' indicates the strength.

Its value ranges between -1 (perfect negative correlation: when x increases, y decreases)

Text, letter

Description automatically generated

Chart, scatter chart

Description automatically generated

Here we can say that the target date rate by percentage is more in white people as compared to the black people, Here we do not know the reason as in every state there is high percentage of white people have a high target death rate. This is **an observation** to look at.

**References:**

Kassambara, M., Visitor, P., & Mann, T. (2018, March 11). Regression with categorical variables: Dummy coding essentials in r. Retrieved April 11, 2021, from <http://www.sthda.com/english/articles/40-regression-analysis/163-regression-with-categorical-variables-dummy-coding-essentials-in-r/>

M, S. (2005). Dealing with missing values. Retrieved April 11, 2021, from <https://uc-r.github.io/missing_values>