Name: Pratik Premkishor Mantri

NUID:001353214

Module5R

Appendix:

|  |  |
| --- | --- |
| Index | Page no |
| Introduction | 2 |
| Data analysis | 2-10 |
| Correlation vs Regression | 11 |
| References | 11 |

**Introduction:**

**Correlation analysis:**

Correlation analysis is used to show the association between two or more variables. Moreover, correlation matrix is used to analyses the correlation between two variables at a same time.

**Methods used for correlation analysis:**

1. The Pearson-method:

The Pearson-method is used to measure the linear dependency between two variables ( which is x and y). This method can be used only when x and y are in the normal form.

2. Kendall tau and spearmen rho are rank based correlation coefficient (non-parametric).

Most of the time we use Pearson correlation method.

In Pearson method if p-value<0.05 which 5% then we can say correlation between x and y is significant.

In this assignment we have taken the Real\_estate\_evaluation dataset (reading and cleaning of data can be seen in R file)

Here I have used Pearson-method for correlation.

After cleaning the data, I have applied the Pearson method

Text

Description automatically generated with medium confidence

on the off diagonals we can see 1 because co-relation with itself is always 1.

if we look at the association with house age and distance to nearest station is it 0.02

Here we get only co-relations values and magnitude but did not get whether they statistically significant or not.

The **p-values (level of significance)** in other words

In co-relation test we see the probability and can tell how statistically significant the variables are with each other

Table

Description automatically generated

we even can tell if the p-value is lower than level of significance then it is statically significant.

If we want to see the confidence interval, then

Table

Description automatically generated

Table

Description automatically generated

Here the lower function is easier to read as it does not contain any redundancy.

Table

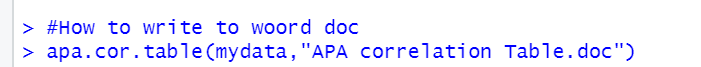
Description automatically generated with medium confidence

From the above figure we can say that:

on the off diagonals we can see 1 because co-relation with itself is always 1.

if we look at the association with house age and transaction date then that would be 0.02.

We even an **create a correlation table and export in word document** using the ApaTables function.



In the below table we can say that

Table

*Means, standard deviations, and correlations with confidence intervals.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | *M* | *SD* | 1 | 2 | 3 | 4 |
|  |  |  |  |  |  |  |
| 1. transaction\_date | 2013.15 | 0.28 |  |  |  |  |
|  |  |  |  |  |  |  |
| 2. house\_age | 17.71 | 11.39 | .02 |  |  |  |
|  |  |  | [-.08, .11] |  |  |  |
|  |  |  |  |  |  |  |
| 3. distance\_to\_nearest\_station | 1083.89 | 1262.11 | .06 | .03 |  |  |
|  |  |  | [-.04, .16] | [-.07, .12] |  |  |
|  |  |  |  |  |  |  |
| 4. number\_of\_stores\_nearby | 4.09 | 2.95 | .01 | .05 | -.60\*\* |  |
|  |  |  | [-.09, .11] | [-.05, .15] | [-.66, -.54] |  |
|  |  |  |  |  |  |  |
| 5. house\_price | 37.98 | 13.61 | .09 | -.21\*\* | -.67\*\* | .57\*\* |
|  |  |  | [-.01, .18] | [-.30, -.12] | [-.72, -.62] | [.50, .63] |
|  |  |  |  |  |  |  |

*M* and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). \* indicates *p* < .05. \*\* indicates *p* < .01.

Here one \* indicate that the value of p<0.05 which is the points are statistically significant when level of significance is 0.05. Under each number we can see confidence interval.

Moreover, \*\* indicates the p value<0.01, which is the points are statistically significant when level of significance is 0.01. Under each number we can see confidence interval.

**Part2:**

Regression:

as we know regression consists of a set of Machine learning algorithms which allow us to predict a continuous outcome variable (y) based on the value of one or multiple predictor variables (x).

**Linear regression:**

  It is the most simple and popular technique for predicting a **continuous** variable. It shows a linear relationship between the outcome and the predictor variables.

It can be written as y = b0 + b\*x + e, where:

* b0 is the intercept,
* b is the regression weight or coefficient associated with the predictor variable x.
* e is the residual error.

**Characteristics of Independent variable:**

The independent variable can only be manipulated or changed by researchers, not by other variables in the problem. Here in the above example time is an independent variable that cannot be changed by another variable such as price (in this case).

In the below figure I have shown 1st five rows.

Table

Description automatically generated

After that I plotted a scatter plot with trend line which shows difference in x makes effect on y.

In the below plot we can say that we get perfect negative correlation as ‘distance of nearest station incresses’(X) the values of ‘house\_price’(Y) decreases.

Chart, scatter chart

Description automatically generated

Text

Description automatically generated

* **Estimate:** the intercept (b0) and the beta coefficient estimates 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 (column 2) 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.

**A low correlation (-0.2 < x < 0.2) probably suggests that much of variation of the outcome variable (y) is not explained by the predictor (x)**

So, we can do model building using linear regression.

We will be building model: housing\_price based on the distance\_to\_nearest\_station

Let’s see the summary of the model:

Text, letter

Description automatically generated

Here the R squared value is 0.4804, we can say that,

If the r^2 is close to 1 tells us that the actual values and the predicted values are close to each other. In our example is almost 48% close to each other.

In contrast, a low value lets says 0.004 tells us the distance between predicted and actual values is higher. Moreover, is r^2=1, we can say we can predict any value of y for any given value of x.

Now lets take coefficient of summary

Text

Description automatically generated with medium confidence

We can write the equation as

Housing price= 4.61-0.007\* distance\_to\_nearest\_station

From that we can take references as for distance of zero the house price would be 4.61$.

Lets say we want to calculate the house price for distance of 0 and 1000

It would be Text

Description automatically generated

**Multiple regression:**

It is used for predicting an outcome variable (y) on basis of multiple different predictor variables (x).

The formula for this is :

 y = b0 + b1\*x1 + b2\*x2 + b3\*x3

Text, letter

Description automatically generated

R^2=0.58

**We can in above example that p-value of the F-statistic is < 2.2e-16, which is highly significant. That indicate that at least one of the predictor variables is significantly related to the outcome variable.**

Coefficient of significance:

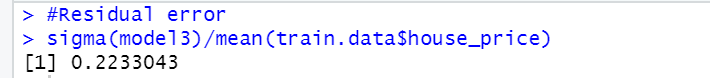
**Text

Description automatically generated**

Finally, we can write our model as:

House price=41.89-0.249\*house age-0.005\*distance to nearest station+1.415\*numbers of stores nearby

Residual error:



The RSE or model sigma, corresponds to the prediction error, which represents the average difference between the observed values and the predicted values. The lower the RSE the best the model fits to our data. Here we have 22% error which is low.

**Difference between correlation and regression:**

1.In correlation we can interchange the variables, but when it comes to regression x and y variable cannot be changed, because one is dependent and other is independent. (house price is dependent)

2. In correlation data is represented in a single point, on the other hand in regression data is represented by a line.

3. In correlation we find a value which shows the relation between the variable, let’s say the association with **house age and distance to nearest station is it 0.02.**  However, in the regression to estimate the value of a random variable based on fixed variable. Let’s say we calculate house price on the basis of fixed variable house age.

4. Correlation shows the linear regression between two variable, regression fits the best line and then estimate the one variable on basis of another variable.

References:

Visitor, J. (2018, March 11). Linear regression essentials in R. Retrieved April 06, 2021, from <http://www.sthda.com/english/articles/40-regression-analysis/165-linear-regression-essentials-in-r/>

M, S. (2005). Correlation vs regression. Retrieved April 07, 2021, from <https://onix-systems.com/blog/correlation-vs-regression>