Analytics for Observational Data (IT142IU)

Lab 5: Fitting and Hypothesis testing

1.1. Objectives

- Model fitting with PCA.

- Understanding hypothesis testing

- Apply hypothesis testing techniques to some datasets.

- Dataset sources:

o <https://www.kaggle.com/doj/federal-firearm-licensees>

o [https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature- data](https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-%20data%20)

- Programming languages: Python/Jav

1.2. Task

Part 1. Follow the instructions in [1] and describe the following steps.

|  |  |
| --- | --- |
| Dataset | Regression-Clean-Data |
| Study question | How better is the modeling after PCA? |
| Random selection of 70% of data as the training set |  |
| The results of the first regression model: Coefficients, the level of accuracy of the model. |  |
| Build 2nd regression model with principal components. |  |
| The results of the second regression model: Coefficients, the level of accuracy of the model. |  |
| Build 3rd regression model with dominant variables in significant principal components : which are variables chosen | | **Dist\_Taxi** | **Dist\_Market** | **Dist\_Hospital** | **Carpet** | **Builtup** | **Rainfall** | **Parking\_Covered** | **Parking\_No Parking** | **Parking\_Not Provided** | **Parking\_Open** | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | |
| The results of the third regression model: Coefficients, the level of accuracy of the model. |  |
| Remarks | In the first regression model, R2 is about 0.477 which indicates that 47.7% of the variation in the house price can be explained by the predictor variables. However, in this case, based on the correlation matrix, I’ve noticed that Carpet and built up have an almost perfect correlation which is mostly impractical in reality. This phenomenon is known as multicollinearity, in which there is a high correlation between predictor variables.  In the second regression model, I have tackled the demons of multicollinearity using PCA, predictor variables now no longer correlated. And there is a slight decrease in R2 (0.47).  In the last regression model, I check for dominant values based cumulative value >= 0.95. Moreover, there is a slight reduction on R2 (0.4722) which is acceptable. |

Part 2. Follow the instructions in [2] and describe the steps of hypothesis testing in your report.

|  |  |
| --- | --- |
| Datasets | Firearm Licensees |
| Study question | How to apply t-test? |
| H0 | μ0 = 2.75 |
| H1 | μ0 # 2.75 |
| Calculate a test statistic |  |
| Calculate a p-value |  |
| Make a decision and interpret your conclusions |  |

Part 3. Apply hypothesis testing to predict the land average temperature. (Significance level = 0.5

|  |  |
| --- | --- |
| Questions | Answers |
| Dataset | GlobalTemperatures.csv |
| Study question | Is the land average temperature 8.37? |
| H0 | μ = 8.37 |
| H1 | μ # 8.37 |
| Choose the sample datasets | AverageTemperature for sample dataset |
| Calculate a test statistic |  |
| Calculate a p-value |  |
| Make a decision and interpret your conclusions | P\_value > apha/2  Don’t reject the hypothesis  We can conlude that the land average temperature is 8.37 |

|  |  |
| --- | --- |
| **Questions** | **Answers** |
| Dataset | GlobalLandTemperaturesByCity.csv |
| Study question | Did Latitude affect Average Temperature? |
| H­o | Latitude did not affect Average Temperature (H0: β1=0) |
| H1 | Latitude did affect Average Temperature (H1: β1 ≠ 0) |
| Choose the sample dataset | AverageTemperature and Latitude for sample dataset |
| Calculate a test statistic |  |
| Calculate a p-value |  |
| Make a decision and interpret your conclusions | We reject the null hypothesis, which states that there is a link between latitude and average temperature (p\_value = 0), at a significant level of 0.05. |