LAB2_Part2_StudentUse_Hideki_v1

February 8, 2021

1 Linear Regression with Health Datasets

Datasets: from kaggles https://www.kaggle.com/nareshbhat/health-care-data-set-on-heart-attack-possibility?select=heart.csv

2 Simple Linear Regression

2.1 1. Import libraries

```
[ ]:
```

2.2 2. Import excel data file into pandas data frame

Data Info:

Attribute Information
1) age
2) sex
3) cp = chest pain type (4 values)
4) trestbps = resting blood pressure
5) chol = serum cholestoral in mg/dl
6) fbs = fasting blood sugar > 120 mg/dl
7) restecg = resting electrocardiographic results (values 0,1,2)
8) thalach = maximum heart rate achieved
9) exang = exercise induced angina
10) oldpeak = ST depression induced by exercise relative to rest
11) slope = the slope of the peak exercise ST segment
12) ca = number of major vessels (0-3) colored by flourosopy
13) thal: 0 = normal; 1 = fixed defect; 2 = reversable defect
14) target: 0= less chance of heart attack 1= more chance of heart attack

```
[ ]:
```

```
[ ]: # to check dataframe, use display()
```

2.3 3. Data Cleaning

There might be a possiblity that the data is missing its values. use “print(df.isnull().sum))” to check if the data is ready to be processed.

[ ]:

```

```

2.4 4. Feature Selection

Now that the data is good to go, we are ready to move on to the next step of the process. As there are 14 features in the dataset, we do not want to use all of these features for training our model, because not all of them are relevant. Instead, we want to choose those features that directly influence the result (that is, prices of houses) to train the model. For this, we can use the corr() function. The corr() function computes the pairwise correlation of columns:

[ ]:

```
# we choose variables this time.
```

[ ]:

```
# #---get the top 3 features that has the highest correlation---
#select the independent variables you chose inside of ' ' to see which␣
 ↪variables has a strong correlation with age.

print(df.corr().abs().nlargest(3, 'age').index) # we put independent for you

#---print the top 3 correlation values---
print(df.corr().abs().nlargest(3, 'age').values[:,13]) # we put independent for␣
 ↪you
```

2.5 5. Create the simple linear regression graph.

hint: figure3.1 - Least squares fit

sns.regplot(x-values, y-values, order= order of variable, ci=confidence interval, scatter_kws={‘color’:‘r’,size’:9})

[ ]:

```

```

2.6 6. Find Regression Coefficient and Intercept

hint: figure 3.2
y-value will be a target value for prediction variable. In this notebook, set “age” as a prediction variable.

[ ]:

```

```

2.7 7. Create Confidence Interval: Statsmodels

hint: Confidence interval on page 67 & Table 3.1 & 3.2 - Statsmodels from pdf

[ ]:

```

```

3 Multiple Linear Regression

3.1 8. Create statsmodels

hint: Table 3.3

```
[ ]:
```

3.2 9. Create multiple linear regression on 3D plot.

hint: figure 3.5 -multiple lienear regression

```
[ ]:
```