Regression_Analysis

May 19, 2022

1 Regression Analysis

Estimated time needed: **30** minutes

The goal of regression analysis is to describe the relationship between one set of variables called the dependent variables, and another set of variables, called independent or explanatory variables. When there is only one explanatory variable, it is called simple regression.

1.1 Objectives

After completing this lab you will be able to:

- Import Libraries
- Regression analysis in place of the t-test
- Regression analysis in place of ANOVA
- Regression analysis in place of correlation

1.2 Import Libraries

All Libraries required for this lab are listed below. The libraries pre-installed on Skills Network Labs are commented. If you run this notebook in a different environment, e.g. your desktop, you may need to uncomment and install certain libraries.

[ ]:

```
#install specific version of libraries used in lab
#! mamba install pandas==1.3.3
#! mamba install numpy=1.21.2
#! mamba install scipy=1.7.1-y
#!  mamba install seaborn=0.9.0-y
#!  mamba install matplotlib=3.4.3-y
#!  mamba install statsmodels=0.12.0-y
```

Import the libraries we need for the lab

[1]:

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
```

Read in the csv file from the URL using the request library

[2]:
```
ratings_url = 'https://cf-courses-data.s3.us.cloud-object-storage.appdomain.
 ↪cloud/IBMDeveloperSkillsNetwork-ST0151EN-SkillsNetwork/labs/teachingratings.
 ↪csv'
ratings_df = pd.read_csv(ratings_url)
```

1.3 Lab Exercises

In this section, you will learn how to run regression analysis in place of the t-test, ANOVA, and correlation

1.3.1 Regression with T-test: Using the teachers rating data set, does gender affect teaching evaluation rates?

Initially, we had used the t-test to test if there was a statistical difference in evaluations for males and females, we are now going to use regression. We will state the null hypothesis:

- $H\_0 : \beta 1 = 0$ (Gender has no effect on teaching evaluation scores)
- $H\_1 : \beta 1$ is not equal to 0 (Gender has an effect on teaching evaluation scores)

We will use the female variable. female = 1 and male = 0

[3]:
```
## X is the input variables (or independent variables)
X = ratings_df['female']
## y is the target/dependent variable
y = ratings_df['eval']
## add an intercept (beta_0) to our model
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Print out the statistics
model.summary()
```

```
/home/jupyterlab/conda/envs/python/lib/python3.7/site-
packages/statsmodels/tsa/tsatools.py:142: FutureWarning: In a future version of
pandas all arguments of concat except for the argument 'objs' will be keyword-
only
  x = pd.concat(x[::order], 1)
```

[3]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                   eval   R-squared:                       0.022
Model:                            OLS   Adj. R-squared:                  0.020
Method:                 Least Squares   F-statistic:                     10.56
Date:                Wed, 18 May 2022   Prob (F-statistic):            0.00124
Time:                        14:36:10   Log-Likelihood:                -378.50
```

```
No. Observations:                 463   AIC:                             761.0
Df Residuals:                     461   BIC:                             769.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.0690      0.034    121.288      0.000       4.003       4.135
female        -0.1680      0.052     -3.250      0.001      -0.270      -0.066
==============================================================================
Omnibus:                       17.625   Durbin-Watson:                   1.209
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               18.970
Skew:                          -0.496   Prob(JB):                     7.60e-05
Kurtosis:                       2.981   Cond. No.                         2.47
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

Conclusion: Like the t-test, the p-value is less than the alpha ( ) level = 0.05, so we reject the null hypothesis as there is evidence that there is a difference in mean evaluation scores based on gender. The coefficient -0.1680 means that females get 0.168 scores less than men.

1.3.2 Regression with ANOVA: Using the teachers' rating data set, does beauty score for instructors differ by age?

State the Hypothesis:

- $H\_0 : 1 = 2 = 3$ (the three population means are equal)
- $H\_1 :$ At least one of the means differ

Then we group the data like we did with ANOVA

[4]:

```
ratings_df.loc[(ratings_df['age'] <= 40), 'age_group'] = '40 years and younger'
ratings_df.loc[(ratings_df['age'] > 40)&(ratings_df['age'] < 57), 'age_group']␣
 ↪= 'between 40 and 57 years'
ratings_df.loc[(ratings_df['age'] >= 57), 'age_group'] = '57 years and older'
```

Use OLS function from the statsmodel library

[5]:

```
from statsmodels.formula.api import ols
lm = ols('beauty ~ age_group', data = ratings_df).fit()
table= sm.stats.anova_lm(lm)
print(table)
```

```
              df     sum_sq    mean_sq          F        PR(>F)
age_group    2.0  20.422744  10.211372  17.597559  4.322549e-08
```

```
Residual   460.0  266.925153   0.580272        NaN           NaN
```

Conclusion: We can also see the same values for ANOVA like before and we will reject the null hypothesis since the p-value is less than 0.05 there is significant evidence that at least one of the means differ.

1.3.3 Regression with ANOVA option 2

Create dummy variables - A dummy variable is a numeric variable that represents categorical data, such as gender, race, etc. Dummy variables are dichotomous, i.e they can take on only two quantitative values.

[6]:

```python
X = pd.get_dummies(ratings_df[['age_group']])
```

[7]:

```python
y = ratings_df['beauty']
## add an intercept (beta_0) to our model
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Print out the statistics
model.summary()
```

```
/home/jupyterlab/conda/envs/python/lib/python3.7/site-
packages/statsmodels/tsa/tsatools.py:142: FutureWarning: In a future version of
pandas all arguments of concat except for the argument 'objs' will be keyword-
only
  x = pd.concat(x[::order], 1)
```

[7]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                 beauty   R-squared:                       0.071
Model:                            OLS   Adj. R-squared:                  0.067
Method:                 Least Squares   F-statistic:                     17.60
Date:                Wed, 18 May 2022   Prob (F-statistic):           4.32e-08
Time:                        14:37:54   Log-Likelihood:                -529.47
No. Observations:                 463   AIC:                             1065.
Df Residuals:                     460   BIC:                             1077.
Df Model:                           2
Covariance Type:            nonrobust
================================================================================
=====================
                                        coef    std err          t      P>|t|
[0.025      0.975]
--------------------------------------------------------------------------------
---------------------
```

```
const                                     0.0138      0.028      0.496      0.620
-0.041       0.069
age_group_40 years and younger            0.3224      0.058      5.574      0.000
0.209       0.436
age_group_57 years and older             -0.2596      0.056     -4.621      0.000
-0.370      -0.149
age_group_between 40 and 57 years        -0.0489      0.045     -1.081      0.280
-0.138       0.040
==============================================================================
Omnibus:                       11.586   Durbin-Watson:                   0.434
Prob(Omnibus):                  0.003   Jarque-Bera (JB):               12.114
Skew:                           0.394   Prob(JB):                      0.00234
Kurtosis:                       2.913   Cond. No.                     6.90e+15
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The smallest eigenvalue is 1.35e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
"""
```

You will get the same results and conclusion

1.3.4 Correlation: Using the teachers' rating dataset, Is teaching evaluation score correlated with beauty score?

[8]:

```python
## X is the input variables (or independent variables)
X = ratings_df['beauty']
## y is the target/dependent variable
y = ratings_df['eval']
## add an intercept (beta_0) to our model
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Print out the statistics
model.summary()
```

```
/home/jupyterlab/conda/envs/python/lib/python3.7/site-
packages/statsmodels/tsa/tsatools.py:142: FutureWarning: In a future version of
pandas all arguments of concat except for the argument 'objs' will be keyword-
only
  x = pd.concat(x[::order], 1)
```

[8]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                   eval   R-squared:                       0.036
Model:                            OLS   Adj. R-squared:                  0.034
Method:                 Least Squares   F-statistic:                     17.08
Date:                Wed, 18 May 2022   Prob (F-statistic):           4.25e-05
Time:                        14:38:21   Log-Likelihood:                -375.32
No. Observations:                 463   AIC:                             754.6
Df Residuals:                     461   BIC:                             762.9
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.9983      0.025    157.727      0.000       3.948       4.048
beauty         0.1330      0.032      4.133      0.000       0.070       0.196
==============================================================================
Omnibus:                       15.399   Durbin-Watson:                   1.238
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               16.405
Skew:                          -0.453   Prob(JB):                     0.000274
Kurtosis:                       2.831   Cond. No.                         1.27
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

Conclusion: $p < 0.05$ there is evidence of correlation between beauty and evaluation scores

1.4 Practice Questions

1.4.1 Question 1: Using the teachers' rating data set, does tenure affect beauty scores?

- Use = 0.05

[9]:

```python
### insert code here
## put beauty scores in a list
y = ratings_df['beauty']
## add an intercept (beta_0) to our model
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)
```

```python
# Print out the statistics
model.summary()
```

```
/home/jupyterlab/conda/envs/python/lib/python3.7/site-
packages/statsmodels/tsa/tsatools.py:142: FutureWarning: In a future version of
pandas all arguments of concat except for the argument 'objs' will be keyword-
only
  x = pd.concat(x[::order], 1)
```

```
[9]: <class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                 beauty   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 1.010e+34
Date:                Wed, 18 May 2022   Prob (F-statistic):               0.00
Time:                        14:39:12   Log-Likelihood:                 16160.
No. Observations:                 463   AIC:                        -3.232e+04
Df Residuals:                     461   BIC:                        -3.231e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       8.674e-18   7.84e-18      1.107      0.269   -6.73e-18    2.41e-17
beauty         1.0000   9.95e-18   1.01e+17      0.000       1.000       1.000
==============================================================================
Omnibus:                       35.667   Durbin-Watson:                   0.445
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               41.953
Skew:                           0.729   Prob(JB):                     7.76e-10
Kurtosis:                       3.221   Cond. No.                         1.27
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

Double-click **here** for a hint.

Double-click **here** for the solution.

1.4.2 Question 2: Using the teachers' rating data set, does being an English speaker affect the number of students assigned to professors?

- Use "allstudents"
- Use = 0.05 and = 0.1

[10]:

```
## insert code here
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Print out the statistics
model.summary()
```

```
/home/jupyterlab/conda/envs/python/lib/python3.7/site-
packages/statsmodels/tsa/tsatools.py:142: FutureWarning: In a future version of
pandas all arguments of concat except for the argument 'objs' will be keyword-
only
  x = pd.concat(x[::order], 1)
```

[10]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                 beauty   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 1.010e+34
Date:                Wed, 18 May 2022   Prob (F-statistic):               0.00
Time:                        14:39:46   Log-Likelihood:                 16160.
No. Observations:                 463   AIC:                        -3.232e+04
Df Residuals:                     461   BIC:                        -3.231e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       8.674e-18   7.84e-18      1.107      0.269   -6.73e-18    2.41e-17
beauty         1.0000   9.95e-18   1.01e+17      0.000       1.000       1.000
==============================================================================
Omnibus:                       35.667   Durbin-Watson:                   0.445
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               41.953
Skew:                           0.729   Prob(JB):                     7.76e-10
Kurtosis:                       3.221   Cond. No.                         1.27
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

Double-click **here** for a hint.

Double-click **here** for the solution.

1.4.3 Question 3: Using the teachers' rating data set, what is the correlation between the number of students who participated in the evaluation survey and evaluation scores?

- Use "students" variable

[12]:

```
## insert code here
X = ratings_df['students']
y = ratings_df['eval']
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Print out the statistics
model.summary()
```

```
/home/jupyterlab/conda/envs/python/lib/python3.7/site-
packages/statsmodels/tsa/tsatools.py:142: FutureWarning: In a future version of
pandas all arguments of concat except for the argument 'objs' will be keyword-
only
  x = pd.concat(x[::order], 1)
```

[12]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                   eval   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                 -0.001
Method:                 Least Squares   F-statistic:                    0.5806
Date:                Wed, 18 May 2022   Prob (F-statistic):              0.446
Time:                        14:41:35   Log-Likelihood:                -383.46
No. Observations:                 463   AIC:                             770.9
Df Residuals:                     461   BIC:                             779.2
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.9823      0.033    119.689      0.000       3.917       4.048
students       0.0004      0.001      0.762      0.446      -0.001       0.002
==============================================================================
Omnibus:                       15.259   Durbin-Watson:                   1.198
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               16.283
Skew:                          -0.456   Prob(JB):                     0.000291
Kurtosis:                       2.888   Cond. No.                         74.8
==============================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

Double-click **here** for a hint.

Double-click **here** for the solution.

1.5 Authors

Aije Egwaikhide is a Data Scientist at IBM who holds a degree in Economics and Statistics from the University of Manitoba and a Post-grad in Business Analytics from St. Lawrence College, Kingston. She is a current employee of IBM where she started as a Junior Data Scientist at the Global Business Services (GBS) in 2018. Her main role was making meaning out of data for their Oil and Gas clients through basic statistics and advanced Machine Learning algorithms. The highlight of her time in GBS was creating a customized end-to-end Machine learning and Statistics solution on optimizing operations in the Oil and Gas wells. She moved to the Cognitive Systems Group as a Senior Data Scientist where she will be providing the team with actionable insights using Data Science techniques and further improve processes through building machine learning solutions. She recently joined the IBM Developer Skills Network group where she brings her real-world experience to the courses she creates.

1.6 Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2020-08-14	0.1	Aije Egwaikhide	Created the initial version of the lab