How can we perform a Breusch-Pagan Test in Python?

Authored by

stats writer

April 17, 2024

The Breusch-Pagan Test is a statistical method used to determine the presence of heteroscedasticity in a dataset, which is a violation of the assumption of equal variance in linear regression models. To perform a Breusch-Pagan Test in Python, first, the data must be loaded into a Pandas DataFrame. Then, using the statsmodels library, the OLS (ordinary least squares) model must be fitted to the data. Next, the residuals from the fitted model are extracted and squared to calculate the squared residuals. Finally, the squared residuals are regressed against the independent variables and the resulting p-value is compared to a predetermined significance level to determine the presence of heteroscedasticity. This process can be automated by using the statsmodels.stats.diagnostic.het_breuschpagan() function, which returns the p-value and other relevant statistics.

Perform a Breusch-Pagan Test in Python

In regression analysis, refers to the unequal scatter of residuals. Specifically, it refers to the case where there is a systematic change in the spread of the residuals over the range of measured values.

Heteroscedasticity is a problem because ordinary least squares (OLS) regression assumes that the residuals come from a population that has *homoscedasticity*, which means constant variance.

When heteroscedasticity is present in a regression analysis, the results of the analysis become hard to trust.

One way to determine if heteroscedasticity is present in a is to use a .

This tutorial explains how to perform a Breusch-Pagan Test in Python.

Example: Breusch-Pagan Test in Python

For this example we'll use the following dataset that describes the attributes of 10 basketball players:

```
import numpy as np
import pandas as pd

#create dataset
df = pd.DataFrame({'rating': ,
'points': ,
'assists': ,
'rebounds': })

#view dataset
df
```

rating points assists rebounds
0 90 25 5 11
1 85 20 7 8
2 82 14 7 10
3 88 16 8 6
4 94 27 5 6

5 90 20 7 9
6 76 12 6 6
7 75 15 9 10
8 87 14 9 10
9 86 19 5 7

We will fit a multiple linear regression model using rating as the response variable and points, assists, and rebounds as the explanatory variables. Then we will perform a Breusch-Pagan Test to determine if heteroscedasticity is present in the regression.

Step 1: Fit a multiple linear regression model.

First, we'll fit a multiple linear regression model:

```
import statsmodels.formula.api as smf

#fit regression model
fit = smf.ols('rating ~ points+assists+rebounds', data=df).fit()

#view model summary
print(fit.summary())
```

Step 2: Perform a Breusch-Pagan test.

Next, we'll perform a Breusch-Pagan test to determine if heteroscedasticity is present.

```
from statsmodels.compat import lzip
import statsmodels.stats.api as sms

#perform Bresuch-Pagan test
names =
test = sms.het_breuschpagan(fit.resid, fit.model.exog)

lzip(names, test)
```

A Breusch-Pagan test uses the following null and alternative hypotheses:

The null hypothesis (H0): Homoscedasticity is present.

In this example, the Lagrange multiplier statistic for the test is 6.004 and the corresponding p-value is 0.1114. Because this p-value is not less than 0.05, we fail to reject the null hypothesis. We do not have sufficient evidence to say that heteroscedasticity is present in the regression model.

How to Fix Heteroscedasticity

In the previous example we saw that heteroscedasticity was not present in the regression model.

However, when heteroscedasticity actually is present there are three common ways to remedy the situation:

1. Transform the dependent variable. One way to fix heteroscedasticity is to transform the dependent variable in some way. One common transformation is to simply take the log of the dependent variable.

2. Redefine the dependent variable. Another way to fix heteroscedasticity is to redefine the dependent variable. One common way to do so is to use a *rate* for the dependent variable, rather than the raw value.

3. Use weighted regression. Another way to fix heteroscedasticity is to use weighted regression. This type of regression assigns a weight to each data point based on the variance of its fitted value. When the proper weights are used, this can eliminate the problem of heteroscedasticity.

Read more details about each of these three methods in