TESTING FOR HETEROSCEDASTICITY IN EXPENDITURE WITH RESPECT TO INCOME

Pranav Gopalkrishna, 1940223

2021-11-25

# INTRODUCTION TO HOMOSCEDASTICITY

Homoscedasticity in a model means that the error is constant along the values of the dependent variable. The best way for checking homoscedasticity is to make a scatterplot with the residuals against the dependent variable.

# DATA SET

This data set contains 30 rows of income and expenditure information, presumably of the same entity, presumably regularly collected over a fixed period of time.

setwd("/Users/pranav/Documents/Study/computerScience/programming/r/data/")  
data = read.csv("expensesAndIncome.csv")  
head(data)

## S.No. Expenditure Income  
## 1 1 10600 11000  
## 2 2 11400 12000  
## 3 3 12300 13000  
## 4 4 13000 14000  
## 5 5 13800 15000  
## 6 6 13900 16000

## Variables

Expenditure depends on income, since expenditure is made based on the received income. Making a regression model for income and expenses will aim to see the relationship between income and expenditure.

y = data$Expenditure  
x = data$Income

# CREATING LINEAR REGRESSION MODEL

We will create a linear regression model with 'Expenditure' as the response and 'Income' as the factor. We will also briefly analyze the created model.

model = lm(y~x)  
summary(model)

##   
## Call:  
## lm(formula = y ~ x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1453.94 -473.48 -73.94 483.33 1546.06   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.799e+03 7.564e+02 3.701 0.000931 \*\*\*  
## x 7.327e-01 4.798e-02 15.271 4.16e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 754.8 on 28 degrees of freedom  
## Multiple R-squared: 0.8928, Adjusted R-squared: 0.889   
## F-statistic: 233.2 on 1 and 28 DF, p-value: 4.165e-15

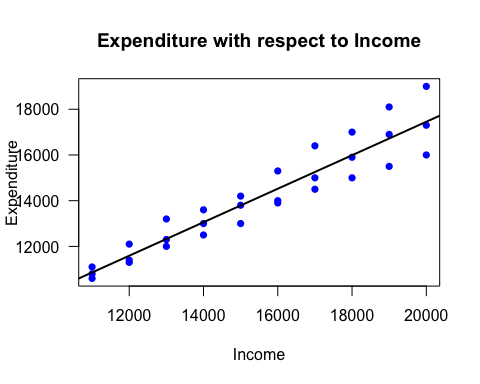
Hence, the created model is...

***Expenditure*** **= 2799 + 0.7327** • ***Income***

We see that the adjusted R-squared value is 88.9%, indicating that the model explains 88.9% of the variation observed in the response variable in the sample data. This indicates the existence of a substantial linear relationship between 'Income' and 'Expenditure'.

# VISUALIZING HOMOSCEDASTICITY

plot(x, y,  
 type = "p",  
 main = "Expenditure with respect to Income",  
 xlab = "Income",  
 ylab = "Expenditure",  
 col = "blue",  
 pch = 16,  
 las = 1)  
abline(model, lw = 2)



As can be seem, the errors noticeably increase in magnitude the larger the incomes. This indicates the presence of heteroscedasticity in the data.

# TESTING FOR HOMOSCEDASTICITY

To test the degree of heteroscedasticity in the response ‘Expenditure’, we will use the Breusch-Pagan test. The hypotheses for this test are given below...

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**H0:** *The error terms are constant.*

**H1:** *The error terms are not constant.*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

To perform this, we need to import the ‘lmtest’ package.

# install.packages("lmtest")  
library(lmtest)

## Warning: package 'lmtest' was built under R version 3.6.2

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.6.2

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

## Performing the test…

bptest(model)

##   
## studentized Breusch-Pagan test  
##   
## data: model  
## BP = 11.373, df = 1, p-value = 0.000745

## Conclusions

The test statistic for the Breusch-Pagan test is denoted by BP, and has the value 11.373 in our test. The p-value of this calculated statistic value is less than 0.05. This means that for a 0.05 significant level, we can say that the calculated test statistic obtained is significant i.e. we have sufficient evidence to reject the null hypothesis and conclude that heteroscedasticity is present in the model i.e. the differences between predicted and actual values of 'Expenditure' are not largely constant. This indicates a notable effect of extraneous random factors that affect expenditure (apart from income itself) in a relatively inconsistent manner.