Testing for autocorrelation

Pranav Gopalkrishna, 1940223

2021-11-10

# DATA SET

The following data set is simply a test data set containing generic response and regressor variable values. My goal will be to look for autocorrelation within the response variable y, given a regression model involving all the available regressors.

myData = read.csv("~/Documents/Study/computerScience/programming/r/data/justSomeData.csv")  
head(myData)

## y x1 x2 x3 x4 x5 x6 x7 x8 x9  
## 1 25.9 4.9176 1 3.472 0.998 1 7 4 42 0  
## 2 29.5 5.0208 1 3.531 1.500 2 7 4 62 0  
## 3 27.9 4.5429 1 2.275 1.175 1 6 3 40 0  
## 4 25.9 4.5573 1 4.050 1.232 1 6 3 54 0  
## 5 29.9 5.0597 1 4.455 0.988 1 6 3 56 0  
## 6 30.9 5.8980 1 5.850 1.240 1 7 3 51 1

#========================

# LINEAR REGRESSION MODEL

Creating the linear regression model, to get an idea about this model.

model = lm(y~., data = myData)  
summary(model)

##   
## Call:  
## lm(formula = y ~ ., data = myData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.8504 -1.4017 0.0929 1.7541 3.7206   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17.11351 5.88549 2.908 0.0131 \*  
## x1 2.39009 1.05740 2.260 0.0432 \*  
## x2 5.74422 4.35113 1.320 0.2114   
## x3 0.12998 0.52530 0.247 0.8087   
## x4 2.63623 4.34493 0.607 0.5553   
## x5 2.32382 1.46160 1.590 0.1378   
## x6 -1.62471 2.40137 -0.677 0.5115   
## x7 -0.09723 3.38794 -0.029 0.9776   
## x8 -0.04445 0.06212 -0.716 0.4879   
## x9 2.03656 1.97372 1.032 0.3225   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.841 on 12 degrees of freedom  
## Multiple R-squared: 0.8774, Adjusted R-squared: 0.7854   
## F-statistic: 9.539 on 9 and 12 DF, p-value: 0.0003125

As we can see, the above model does a reasonable job in explaining the observed variance in the response, with adjusted R-square being 78.54%, implying that around 79% of the variance observed in the response variable's values in the sample is explained by the model. This suggests a reasonably accurate linear relationship (based on the sample, not yet tested for the population).

#========================

# USING THE DURBIN WATSON TEST TO CHECK FOR AUTOCORRELATION

Autocorrelation is the presence of correlation between current and past values of the response variable. This can occur if past values have some impact on future values of the variable. For example, if the response variable is monthly expenditure, excess expenditure in one month may compel people to spend less the next month. Autocorrelation may also happen at a lag i.e. current values may be related to a past value before the adjacent past value. For example, current investments may have an impact on future investments after a time gap (after seeing how well the investment has performed).

## The required 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

From this, we will be using the dwtest function, which accepts the basic linear relationship of the variables as the model. For example, y~x, or y~x1+x2 etc.

dwtest(y~., data = myData)

##   
## Durbin-Watson test  
##   
## data: y ~ .  
## DW = 2.2278, p-value = 0.6585  
## alternative hypothesis: true autocorrelation is greater than 0

# CONCLUSIONS

p-value of the test statistic is 0.6585, meaning it is not significant for a 0.05 significance level (i.e. it is not unlikely enough to be considered as significant indication for autocorrelation), Hence, we may conclude that there is no autocorrelation within the response variable. This means that we may conclude that past values of y don’t affect future values. Of course, this is not 100% conclusive, and there may still be positive autocorrelation, as suggested by the DW value being 2.2278.