Autocorrelation

Dr. Kiah Wah Ong

Model Assumptions

Recall the major assumptions we have made in linear regression models

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, 2, \cdots, n$$

are

- The relationship between the response and regressors is linear.
- The error terms ϵ_i have mean zero.
- The error terms ϵ_i have constant variance σ^2 (homoscedasticity)
- The error terms ϵ_i are normally distributed.
- The error terms ϵ_i and ϵ_j are uncorrelated for $i \neq j$.
- The regressors $x_1, \cdots, x_k$ are nonrandom.
- The regressors $x_1, \cdots, x_k$ are measured without error.
- The regressors are linearly independent.

Autocorrelation

In our regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, 2, \cdots, n$$

we need the error terms ϵ_i and ϵ_j to be uncorrelated when $i \neq j$, that is

$$\text{cov}(\epsilon_i, \epsilon_j) = E[(\epsilon_i - E(\epsilon_i))(\epsilon_j - E(\epsilon_j))] = 0$$

whenever $i \neq j$.

When the correlated observations have a natural sequential order, then the correlation is referred to as **autocorrelation**.

Autocorrelation

Autocorrelation may arise when dealing with predictor and response variables that are time series, that is, the variables are time-oriented.

Autocorrelation

When autocorrelation is presence:

(1) Least squares estimates of the regression coefficients are still unbiased but not efficient in the sense that they no longer have minimum variance.

(2) The estimate of σ^2 and the standard errors of the regression coefficients may be seriously underestimated, giving a false sense of fit.

(3) The confidence intervals and various tests of significance are no longer accurate.

Durbin-Watson Test

There are various statistical test that we can employ to detect the presence of autocorrelation. Here we are going to introduce the test developed by Durbin and Watson.

The test is based on the assumption that the errors in the regression model are correlated in the sense that

$$\epsilon_i = \phi\epsilon_{i-1} + a_i$$

where a_i is normally and independently distributed random variable with mean 0 and variance σ^2 $(\mathrm{NID}(0, \sigma^2))$, and ϕ is a parameter that defines the relationship between successive values of the model errors ϵ_i, and ϵ_{i-1}, we also required that $|\phi| < 1$.

The equation above also called a first-order autoregressive process (AR(1)).

Durbin-Watson Test

Hence, a simple linear regression model with first-order autoregressive errors is given by

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{and} \quad \epsilon_i = \phi \epsilon_{i-1} + a_i$$

where y_i and x_i are the observations on the response and predictor variables at time i.

Durbin-Watson Test

For the error terms, it can be shown that:

$$
\begin{aligned}
E(\epsilon_i) &= 0 \\
\operatorname{var}(\epsilon_i) &= \frac{\sigma^2}{1-\phi^2} \\
\operatorname{cov}(\epsilon_i, \epsilon_{i-1}) &= \frac{\phi\sigma^2}{1-\phi^2} = \phi\operatorname{var}(\epsilon_i) \\
\operatorname{cov}(\epsilon_i, \epsilon_{i-s}) &= \frac{\phi^s\sigma^2}{1-\phi^2} = \phi^s\operatorname{var}(\epsilon_i), \text{ for } s > 0
\end{aligned}
$$

Durbin-Watson Test

The autocorrelation between two errors that are one period apart, or the **lag one** autocorrelation is given by

$$\begin{aligned}
\rho_1 &= \operatorname{corr}(\epsilon_i, \epsilon_{i+1}) \\
&= \frac{\operatorname{cov}(\epsilon_i, \epsilon_{i+1})}{\sqrt{\operatorname{var}(\epsilon_i)}\sqrt{\operatorname{var}(\epsilon_{i+1})}} \\
&= \frac{\phi\sigma^2\left(\frac{1}{1-\phi^2}\right)}{\sqrt{\sigma^2\left(\frac{1}{1-\phi^2}\right)}\sqrt{\sigma^2\left(\frac{1}{1-\phi^2}\right)}} \\
&= \phi
\end{aligned}$$

The lag k autocorrelation is given by $\rho_k = \phi^k$ for $k = 1, 2, \cdots$, this equation is called the **autocorrelation function**.

Durbin-Watson Test

Notice that when ϕ is positive, all error terms are positively correlated, but the magnitude of the correlation decreases as the errors grow further apart, since

$$\rho_k = \phi^k, \quad \text{and} \quad |\phi| < 1$$

Also, only when $\phi = 0$ are the model errors uncorrelated.

Durbin-Watson Test

The Durbin-Watson Test is a hypothesis test:

We wish to perform a hypothesis test

$$H_0 : \phi = 0 \quad \text{vs} \quad H_1 : \phi > 0$$

and the Durbin-Watson test statistic is

$$d = \frac{\sum_{i=2}^{n} (\epsilon_i - \epsilon_{i-1})^2}{\sum_{i=1}^{n} \epsilon_i^2}$$

If d is smaller than a certain critical value, we will reject the null hypothesis.

Remark: Most time series regression problems involve with positive autocorrelation. Hence our H_1 is set as $\phi > 0$.

Durbin-Watson Test

Intuition on the Durbin-Watson test statistic

$$d = \frac{\sum_{i=2}^{n}(\epsilon_i - \epsilon_{i-1})^2}{\sum_{i=1}^{n}\epsilon_i^2}$$

Notice that, if we have the i-residual ϵ_i correlates to the previous $(i-1)$-residual ϵ_{i-1}, then the two terms must "looks" alike and hence the distance $(\epsilon_i - \epsilon_{i-1})^2$ is going to be small.

Therefore if we find many lag 1 residuals, then the Durbin-Watson statistic is very small. Hence small value of d indicate higher possibility of correlation.

So if d is smaller than a certain critical value, we will reject the null hypotheses $H_0 : \phi = 0$.

Durbin-Watson Test

Unlike the usual hypothesis testing that we have seen before, in the Durbin-Watson Test, we would not be using an exact critical value for d. Instead, the decision procedure is as follows:

$$\begin{aligned} \text{If } d < d_L \text{ reject } \mathrm{H}_0 : \ \phi = 0 \\ \text{If } d > d_U \text{ do not reject } \mathrm{H}_0 : \ \phi = 0 \\ \text{If } d_L \leq d \leq d_U \text{ the test is inconclusive} \end{aligned}$$

for some lower d_L and upper d_U critical values.

Durbin-Watson Test

Remarks:

(1) The Durbin-Watson Test only check for autocorrelation with a lag of 1, hence longer lags may not be detected.

(2) It can be shown that $d \approx 2(1 - \hat{\phi})$, where $\hat{\phi}$ is the estimate for the autocorrelation parameter ϕ. Hence
   (i) when $d \approx 2$, we will have no autocorrelation (as $\hat{\phi} \approx 0$).
   (ii) As $d \to 0$, hence $\hat{\phi} \to 1$, we will have a perfect autocorrelation.
   (iii) Negative autocorrelation is rare, but we can test for it using test statistics $4 - d$ and the procedure as we have before.

Durbin-Watson Test in R

Using the dataset Autoc1.csv we run the regression and Durbin-Watson test in *R*.

```
library(lmtest)
data1<-read.csv("Autoc1.CSV", header=TRUE, sep=",")
x<-data1$x
y<-data1$y
model1=lm(y~x)
dwtest(model1)
```

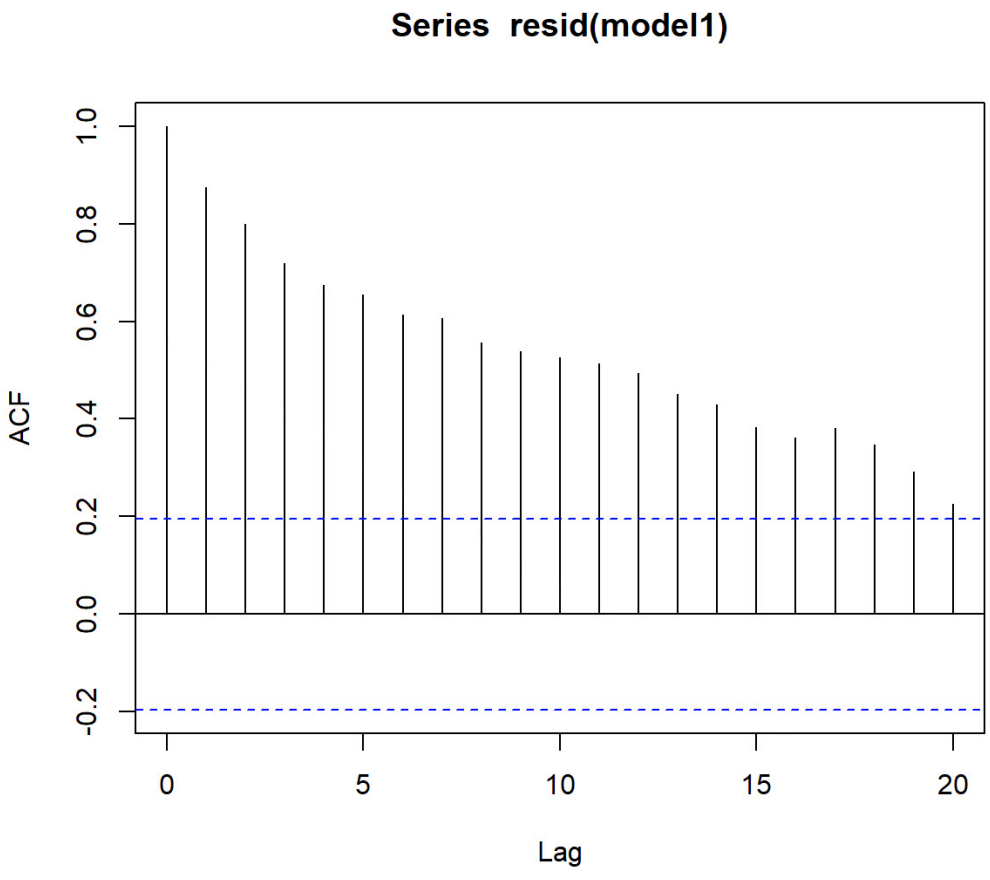
```
        Durbin-Watson test

data:  model1
DW = 0.23818, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

From the R-output, we see that the test statistics is $d = 0.23818$ and the p-value is 0. We reject $H_0 : \phi = 0$ and conclude that the errors ϵ are positively correlated.

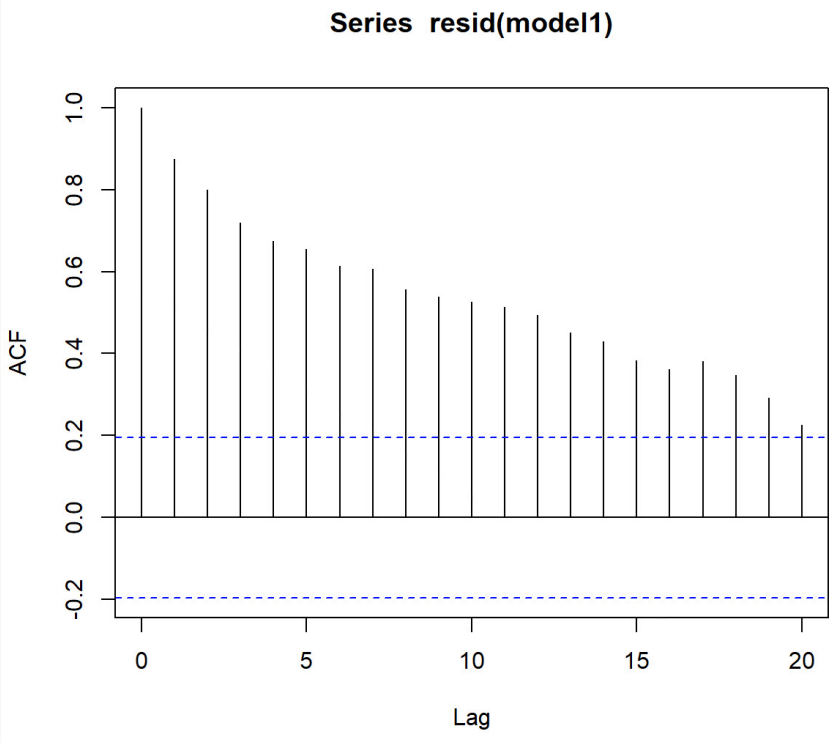
Test for Autocorrelation with the ACF Plot

Let us perform a test on the autocorrelation among residuals in the previous dataset with

```
library(tseries)
data1<-read.csv("Autoc1.CSV", header=TRUE, sep=",")
x<-data1$x
y<-data1$y
model1=lm(y~x)
dwtest(model1)
acf(resid(model1))
```



Test for Autocorrelation with the ACF Plot



The x-axis of the plot corresponds to the different lags of the residuals, while the y-axis shows the correlation of each lag, and the blue line represents the significance level. Notice that the first vertical bar (lag-0) shows the correlation of a residual with itself and therefore is always taken as one.

In the absence of autocorrelation, the subsequent vertical bars would quickly drop to almost zero, or at least between or near the dashed blue line. Hence in our case, we see that autocorrelation is clearly present.

Remediation of Autocorrelation

- Look for an improved model. Look for missing predictor variables, since these variables often show up as autocorrelated residuals.
- Improve the measurement/experiment to remove the dependence of time, space, and order.
- Preforming a suitable change of variables.

First-Order Autoregressive Correction

Recall our simple linear regression model with first-order autoregressive errors given by

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{and} \quad \epsilon_i = \phi \epsilon_{i-1} + a_i$$

where y_i and x_i are the observations on the response and predictor variables at time i and $a_i \sim \text{NID}(0, \sigma^2)$.

Let us define new variable y' and x' as follows:

$$y_i' = y_i - \phi y_{i-1}, \quad x_i' = x_i - \phi x_{i-1}$$

We will show how these transformations lead to a new model that would not have autocorrelation.

First-Order Autoregressive Correction

With $y_i' = y_i - \phi y_{i-1}, x_i' = x_i - \phi x_{i-1}$ and $\epsilon_i = \phi \epsilon_{i-1} + a_i$, we see that

$$\begin{aligned} y_i' &= y_i - \phi y_{i-1} \\ &= \beta_0 + \beta_1 x_i + \epsilon_i - \phi(\beta_0 + \beta_1 x_{i-1} + \epsilon_{i-1}) \\ &= \beta_0(1-\phi) + \beta_1(x_i - \phi x_{i-1}) + \epsilon_i - \phi\epsilon_{i-1} \\ &= \beta_0' + \beta_1 x_i' + a_i \end{aligned}$$

where $\beta_0' = \beta_0(1-\phi)$. This gives a better transformed model.

However, there is one catch, namely, we do not have the information on ϕ.

This can be fixed by using $r = \rho_1 = corr(\epsilon_i, \epsilon_{i+1})$ as an approximation for ϕ.

First-Order Autoregressive Correction

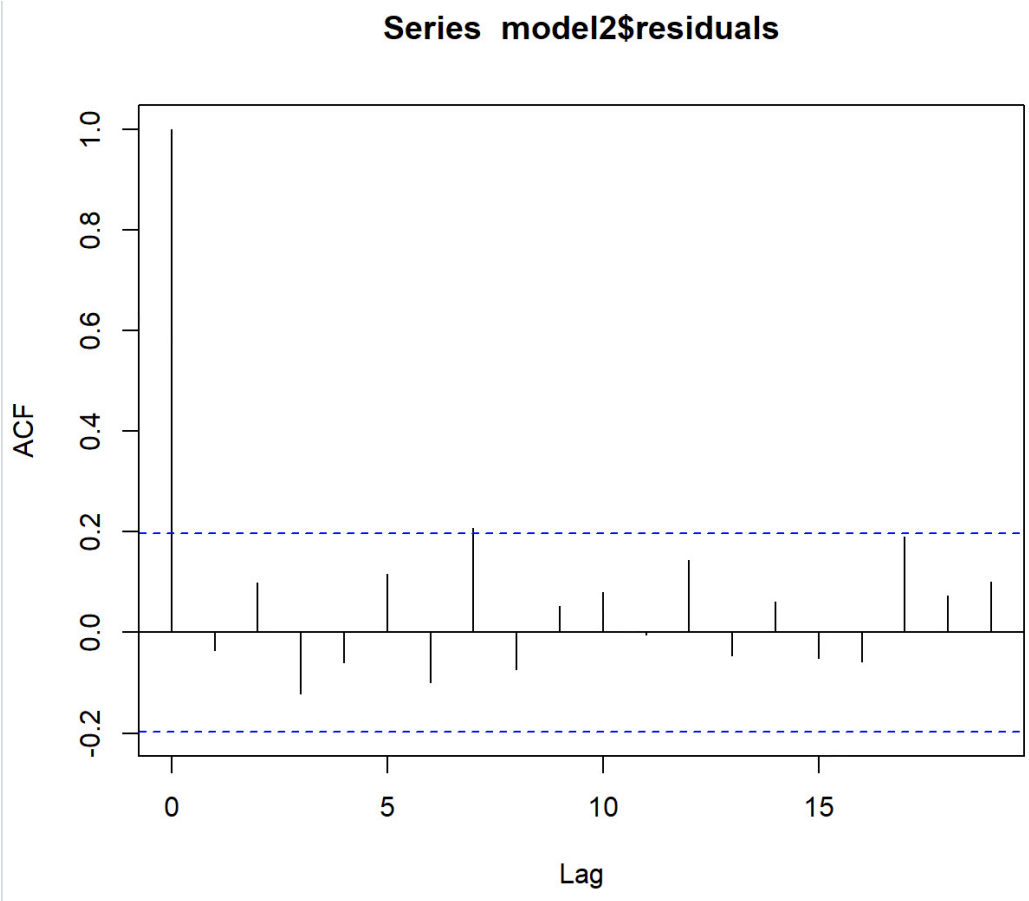
With $y_i' = y_i - ry_{i-1}, x_i' = x_i - rx_{i-1}$, we will do an "ordinary" regression between y_i' and x_i'.

```
library(tseries)
library(lmtest)
data1<-read.csv("Autoc1.CSV", header=TRUE, sep=",")
x<-data1$x
y<-data1$y
model1=lm(y~x)
dwtest(model1)

rho=acf(resid(model1))
acf(model1$residuals)

r=rho[1]$acf[1,,]

n=length(model1$residuals)
yprime=rep(0,n-1)
xprime=rep(0,n-1)
for(i in 1:n-1){
 yprime[i]=data1$y[i+1]-r*data1$y[i]
 xprime[i]=data1$x[i+1]-r*data1$x[i]
}
plot(xprime,yprime)

model2=lm(yprime~xprime)

dwtest(model2, alternative="greater")
```

First-Order Autoregressive Correction

The Durbin-Watson Test on this new model gives

```
        Durbin-Watson test

data:  model2
DW = 2.0407, p-value = 0.5999
alternative hypothesis: true autocorrelation is greater than 0
```



The new ACF plot shows no sign of auto-correlation.

First-Order Autoregressive Correction

The correct estimate of the intercept and the slope of the original model y vs x is then calculated as

$$\beta_0 = \frac{\beta_0'}{1-r} \quad \text{and} \quad \beta_1 = \beta_1'$$

```
Call:
lm(formula = yprime ~ xprime)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9869 -1.0792  0.0061  1.1475  4.0715

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.05656    0.14276   0.396    0.693
xprime       1.15696    0.11428  10.123   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.42 on 97 degrees of freedom
Multiple R-squared:  0.5137,    Adjusted R-squared:  0.5087
F-statistic: 102.5 on 1 and 97 DF,  p-value: < 2.2e-16
```