Lecture Notes 6

1. Autocorrelation (or Serial Correlation)

1. Definition

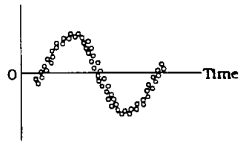
A problem usually for time series data
The error terms, u_t, are supposed to be random
However, they are related by time
The error terms are correlated
Using the covariance term, then

$$Cov\lfloor u_i u_j \rfloor \neq 0 \quad for\, i \neq j$$

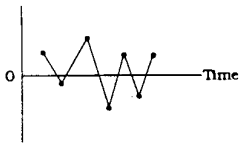
The correlation between u_t and u_{t-r} is called an autocorrelation of order r.
Usually econometricians use k, but then it is confusing
Could result from the impact of a missing x variable
Parameters estimates are unbiased, but estimated standard errors are biased.
F and t-statistics are invalid

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \cdots \beta_k X_{kt} + u_t$$

If you plot the residuals, then



Positive Correlation



Negative Correlation

Autocorrelation of degree 1 is very common

Denoted AR(1)

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \cdots \beta_k X_{kt} + u_t$$

Where $u_t = \rho u_{t-1} + v_t$

Rho, ρ, is between -1 and 1

v is error term

Covariance function – shows how two variables vary together

Correlation comes from this function

$$Cov[X,Y] = E[(X - E(X))(Y - E(Y))]$$

(i) If X = Y, then this becomes the variance for X

$$Cov[X,X] = E[(X - E(X))(X - E(X))] = E(X - E(X))^2 = Var(X)$$

(ii) Remember, E(X) is the expected value and average

(iii) Remember, $E(u_t) = E(u_{t-1}) = 0$

Autocorrelated error terms are still equal to zero on average

$$Cov[u_t, u_{t-1}] = E[(u_t - E(u_t))(u_{t-1} - E(u_{t-1}))] = E[u_t u_{t-1}]$$

I can pull out (i.e. factor) the rho from the expression and $u_t = \rho u_{t-1} + v_t$

$$Cov[u_t, u_{t-1}] = E[u_t u_{t-1}] = E[(\rho u_{t-1} + v_t)u_{t-1}] = E[\rho u_{t-1} u_{t-1}] + E[u_{t-1} v_t]$$
$$\rho E[u_{t-1}]^2 + 0 = \rho\sigma^2$$

Higher orders of Autocorrelation

Denoted AR(p)

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_p u_{t-p} + v_t$$

2. Detecting AR(1)

Use the Durbin-Watson statistic

Works for only AR(1)

Could have higher degrees of autocorrelation

Hypothesis test

$H_0 : \rho = 0$

$H_A : \rho \neq 0$

Calculate the Durbin-Watson (DW) Statistic

$$DW = \frac{\sum_{t=2}^{n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n} \hat{u}_t^2}$$

If ρ = +1, the DW = 0, i.e. positive autocorrelation

The standard errors are biased and too small

t-statistics tend to be significant

If ρ = 0, the DW = 2, i.e. no autocorrelation of AR(1)

The standard errors are not biased and t-statistics are fine.

If ρ = -1, the DW = 4, i.e. negative autocorrelation

The standard errors are biased and too large

t-statistics tend to be insignificant

Note: DW has areas where the statistic is indeterminate.

Let's not worry about it. There are other tests. We will focus on something better in a couple of weeks.

3. Fixing AR(1)

(i) First Method

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \cdots \beta_k X_{kt} + u_t \text{ and } u_t = \rho u_{t-1} + \varepsilon_t$$

Use Solver in Excel
Create a column for the residuals

$$u_t = Y_t - \beta_1 - \beta_2 X_{2t} - \beta_3 X_{3t} - \cdots \beta_k X_{kt}$$

Now create a new column and call it True Errors

$$\varepsilon_t = u_t - \rho u_{t-1}$$

Use Solver to minimize the Sum of Squared Errors (SSE) which is the True Errors column

Allow Solver to minimize by adjusting the Beta's and rho.

It does not appear to be a good method. Does a poor job for rhos with high magnitudes.

(ii) Second Method

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \cdots \beta_k X_{kt} + u_t \text{ and } u_t = \rho u_{t-1} + \varepsilon_t$$

Create lagged values for regression

$$Y_{t-1} = \beta_1 + \beta_2 X_{2t-1} + \beta_3 X_{3t-1} + \cdots \beta_k X_{kt-1} + u_{t-1}$$

Then multiply by rho,

$$\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 X_{2t-1} + \rho\beta_3 X_{3t-1} + \cdots \rho\beta_k X_{kt-1} + \rho u_{t-1}$$

Solve for pu_t

$$\rho u_{t-1} = \rho Y_{t-1} - \rho\beta_1 - \rho\beta_2 X_{2t-1} - \rho\beta_3 X_{3t-1} - \cdots \rho\beta_k X_{kt-1}$$

Substitute $u_t = \rho u_{t-1} + v_t$ into regression equation.

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \cdots \beta_k X_{kt} + u_t$$
$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \cdots \beta_k X_{kt} + \rho u_t + v_t$$

Now substitute pu_{t-1} into the equation above and re-arrange to get

$$Y_t - \rho Y_{t-1} = (1-\rho)\beta_1 + \beta_2 (X_{2t} - \rho X_{2t-1}) + \beta_3 (X_{3t} - \rho X_{3t-1})_{3t} + \cdots \beta_k (X_{kt} - \rho X_{kt-1}) + v_t$$

This is called Generalized Least Squares (GLS)

How to do it in Excel

(i) Estimate a standard regression

(ii) Using the residuals, create a column of lagged residuals

(iii) Estimate the autocorrelation, rho by the equation

$$DW \approx 2(1-\rho)$$

$$\rho = 1 - \frac{DW}{2}$$

(iv) You have an estimate for rho, plug it into regression creating new columns of variables

(v) Then estimate new regression function with rho

In theory, you keep repeating steps until rho stops changing

2. Heteroscedasticty

1. Definition

A problem usually for cross-sectional data

The error terms, u_t, are supposed to be random

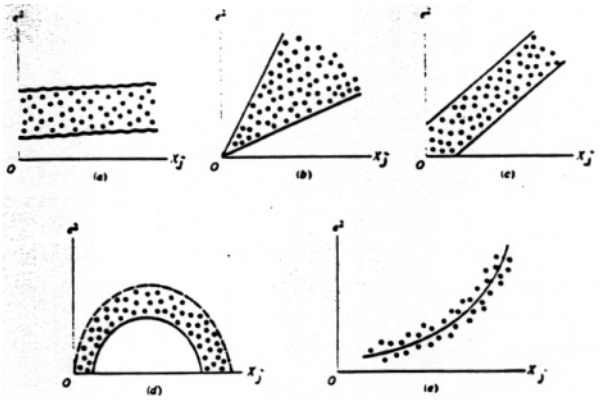
However, they are related to an X variable

Parameters estimates are unbiased, but estimated standard errors are biased.

F and t-statistics are invalid

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \cdots \beta_k X_{kt} + u_t$$

If you plot the residuals against an X variable, then



Scattergram of estimated squared residuals against X_j.

e_i^2 ARE PLOTTED AGAINST X_j

2. Detecting Heteroscedasticity – Park-Glejser Test

Estimate the standard regression

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots \beta_k X_{ki} + u_i$$

Remember you calculate the variance

$$\hat{\sigma}^2 = \frac{\sum_i \hat{u}_i^2}{n-k}$$

If there is Heteroscedasticity, the sigma is related to x variables

Square the residuals and estimate the regression

$$\ln u_i^2 = \delta_1 + \delta_2 \ln X_{2i} + \delta_3 \ln X_{3i} + \cdots \delta_k \ln X_{ki} + \varepsilon_i$$

If the F-statistic is significant, then you have heteroscedasticity

If the F-statistic is insignificant, then you do not have heteroscedasticity

Don't worry about correcting it for now.

3. Multicollinearity (Collinearity)

Definition

Starting with a standard regression

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \cdots \beta_k X_{kt} + u_t$$

Collinearity is when you have a linear relationship among the X variables

$$a_2 X_{2t} + a_3 X_{3t} + \cdots a_k X_{kt} \approx 0$$

What happens,

Least squares estimator is $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$

The matrix X^TX becomes singular

Thus, the inverse, $(X^TX)^{-1}$ explodes in values

Least squares cannot estimate the parameters

Detection

Look for the following signs:

1. Large standard errors for the parameter estimates
   Causes low t-statistics

2. Parameter estimates vary significantly from sample to sample
   Or minor changes to the model

3. Extreme correlations between the X variables

4. Performs poor forecasting

Correction

Called Ridge Regression

We calculate the matrix, X^TX

Then we manually add a small number, λ, to the diagonal
   For example, $\lambda = 0.000001$ (six zeros)

This is equivalent to $X^TX + \lambda I$

So Ridge Regression becomes, $\hat{\beta} = \left(X^T X + \lambda I\right)^{-1} X^T Y$