Violations of OLS etc

Problems with assumed i.i.d. errors

Problems in assumptions about the error (noise) term e_i.

$$\tilde{\mathbf{y}} = \mathbf{X}\mathbf{b} + \tilde{\mathbf{e}}$$

$\tilde{\mathbf{y}}$ is random because the error $\tilde{\mathbf{e}}$ is.
When we estimate $\hat{\mathbf{b}}$ and do inference with the resulting estimate we have made a number of assumptions.

- The most important is that the error e_i of observation i is independent (or almost independent) of the error e_j of some other observation j.
- Also important is that the variance of each noise term is the same.

The most extreme assumption is that

$$e_i \sim N(0, \sigma^2)$$

(what is called i.i.d. assumption)

Deviations from i.i.d. assumptions

What worries econometricians is that these assumptions are unlikely to hold exactly.
We will therefore have to ask a couple of questions in a given estimation situation.
First, we ask whether we are close enough to the ideal assumptions that OLS and the like is sufficient for estimation and inference.
If not, we ask whether we can identify the type of deviation from the i.i.d assumption and adjust for it.
There is no "recipe" for how to do this, the more typical approach is to look for some standard types of deviations from the iid assumptions, which very often occur, and for which we have ways of adjusting.

The two best known problems are

- Heteroskedasticity – relatively independent observations, but the variance of e_i varies with observations.
- Autocorrelation – dependence between error terms, most common when we have time series observations.

What will be the consequences of using OLS in a situation with heteroskedastic errors?

1. OLS is still unbiased/consistent for b.
2. OLS is no longer efficient
3. OLS estimates of the parameter variances are no longer unbiased.

Testing for heteroskedasticity

If we think this dependence is related to the observations, it is possible to test for it.

Exercise

You are investigating the market model

$$r_{it} = a + br_{mt} + e_{it}$$

in the Norwegian Market, and apply it to the company Norsk Hydro (NHY). Collect monthly returns for NHY for the period 1980-, and monthly returns for a value weighted market index for the same period.

- Estimate the model and evaluate the results.
- You worry about the possibility of the variance of the errors varying, i.e. heteroskedasticity. To investigate this you run a regression with the squared residuals as dependent variable, and as explanatory variables a constant, the market return, and the squared market return,

$$\hat{e}_t^2 = a + b_1 r_{mt} + b_2 r_{mt}^2 + \varepsilon$$

Do you find signs of heteroskedasticity?
What can be done to remedy any problems?

Exercise Solution

Reading the data

```
> library(zoo)
> rets <- read.zoo("/../data/norway/ose_individual_stocks/r
+                  format="%Y%m%d",skip=2,header=TRUE,sep="
> Rm <- read.table("../data/norway/stock_market_indices/mar
+                  header=TRUE,sep=";");
> rNHY <- rets$Norsk.Hydro
> vw   <- Rm$VW
> data <- merge(rNHY,vw,all=FALSE)
> rNHY <- data$rNHY
> rm   <- data$vw
```

Exercise Solution

Let us first do the standard estimation

```
> reg <- lm(rNHY ~ rm)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.009918   0.003021  -3.283  0.00112 **
rm           1.134563   0.043720  25.951  < 2e-16 ***

Residual standard error: 0.0556 on 370 degrees of freedom
Multiple R-squared: 0.6454,Adjusted R-squared: 0.6444
F-statistic: 673.4 on 1 and 370 DF,  p-value: < 2.2e-16
```

Now, doing the regression that will test for heteroskedasticity

```
> rm2 <- rm^2
> e2 <- residuals(reg)^2
> het <- lm(e2~rm + rm2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0022750  0.0003128   7.273 1.84e-12 ***
rm          0.0106856  0.0040941   2.610  0.00939 **
rm2         0.0924696  0.0363757   2.542  0.01139 *

Residual standard error: 0.005232 on 405 degrees of freedom
Multiple R-squared:  0.03263,Adjusted R-squared:  0.02786
F-statistic: 6.831 on 2 and 405 DF,  p-value: 0.001208
```

Exercise Solution

To correct for potential heteroscedasticity, calculate Heteroskedasticity consistent standard errors, or "White corrected" standard errors. These typically will be slightly larger than the usual OLS errors.
HC consistent estimation from package `sandwich`.

```
> library(sandwich)
> sandwich(reg)
              (Intercept)            rm
(Intercept)  7.651440e-06 -1.607607e-05
rm          -1.607607e-05  3.003058e-03
> sqrt(diag(vcov(reg)))
(Intercept)          rm
0.003020602 0.043720208
> sqrt(diag(vcovHC(reg)))
(Intercept)          rm
0.002787477 0.055822628
```

Exercise Solution

Compare the two cases

OLS	0.003020602	0.043720208
HC	0.002787477	0.055822628

Dealing with Heteroskedasticity

1. Try to transform the problem into one with more "even" errors. E.g. take logs
2. Build a correction for heteroskedasticity into the modelling.

The White estimate of parameter covariances

Consider the regression model

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

and its OLS estimate $\widehat{\mathbf{b}}$.
$\mathbf{e}$ is the error term, and it has a distribution $\sim (\mathbf{0}, \Omega)$.
Under normality we assume that the errors are iid with common variance σ^2, which reduces Ω to $\Omega = \sigma^2\mathbf{I}$.
In more general settings $E[\mathbf{e}\mathbf{e}'] = \Omega$ is not so simple.
When one calculate the covariance matrix of $\widehat{\mathbf{b}}$,

$$V(\widehat{\mathbf{b}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

The White correction is to replace Ω with an estimate of it $\widehat{\Omega}$.

The simplest possible such estimate is to replace Ω with

$$\widehat{\Omega} = \begin{bmatrix} \hat{e}_1^2 & 0 & 0 & \\ 0 & \hat{e}_2^2 & 0 & \\ 0 & 0 & \hat{e}_3^2 & \\ & & & \ddots \end{bmatrix},$$

althought there are alternative ways of doing such corrections.

Time series

What is special about observations ordered in time, i.e, observing the same variable at different time points?
– The ordering imposed by time passing. Observations *before* fundamentally different from observations *after*.
In standard regressions with contemporaneous observations - time series nature affect error terms – observations "close" in time likely to be affected by the "same" uncertainty.
In a standard regression model

$$y_t = \mathbf{x}_t b + e_t$$

where t now indexes time, so observation is made at time t. Observation at time $t+1$ is made *after* time t, but there is no fixed timing, it may be equally spaced observations, like daily, weekly, monthly, or unevenly spaced observations, such as transaction time.

What have covered before still applies, but likely to have special dependencies among the error terms, where

$$\text{cov}(e_t, e_{t+1}) \neq 0$$

but for general j

$$\text{cov}(e_t, e_{t+j}) \to 0 \text{ as } j \text{ increases.}$$

In the limit, as $j \to \infty$, $\text{cov}(e_t, e_{t+j}) = 0$.
Is this likely to be a problem?
Most time series have this property.
It is called autocorrelation in the error terms.
Just as with heteroskedasticity, autocorrelation can be adjusted for if we know it is present.

Autocorrelation - Consequences

What are the consequences of autocorrelation in errors?

1. OLS unbiased
2. OLS inefficent
3. Estimated of paramter standard errors biased, likely to be seriously understated

The Durbin-Watson test

Regressions involving time series data.

$$y_t = a + b_1 x_t + e_t$$

Errors e_t for "close" dates are related. "autocorrelation in errors"

$$\text{cov}(e_t, e_{t+1}) \neq 0$$

Durbin Watson test: Tests for first order autocorrelation in the errors,

$$\text{corr}(e_t, e_{t+1}) = a$$

Calculated as:

$$DW = \frac{\sum(\hat{\hat{e}}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

If there is no autocorrelation in the errors, $DW = 2$.

$$DW \approx 2(1 - \hat{\rho})$$

where $\hat{\rho}$ is the estimated first order autocorrelation coefficients.

Dealing with autocorrelation

What to do when detecting problems?

1. Small change of model, such as differencing.
2. Modelling the problem explicitly, time series analysis, of which more later.
3. Stay with the model, but work with robust estimates of standard errors.

Let us look in more detail at the last choice, calculating robust errors.
If we know exactly the type of autocorrelation we can see how this affects the covariance terms of the errors.

$$\mathbf{e} \sim N(0, \Omega)$$

The assumed time series properties of the noise term can be used to construct the Ω used in the GLS type estimation

$$\hat{\mathbf{b}}^{wls} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}(X'\Omega^{-1}\mathbf{y})$$

For example, suppose we assume

$$e_t = \rho e_{t-1} + \varepsilon_t$$

Thes auto regressive assumption of order 1.
If we were to *know* that this is the true relation between errors, we can use that to calculate the whole covariance matrix.

Need assumptions about the relationship between errors, i.e. how how much auto correlation falls as the "distance" between the errors increases, ie. $\text{cov}(e_t, e_{t+j}) \to 0$ as j increases.
In practice we don't know the exact lag structure, one therefore typically in cases like this, consider what as called autocorrelation corrected errors. This is often called the HAC correction.
What is being done is calculating the covariance matrix Ω under assumptions about the lag structure, where one sets a maximum number of lags with possible nonzero autocovariances. I.e. if k is the max number of lags,

$$\text{cov}(e_t, e_{t+k+j}) = 0 \text{ if } j > 0$$

Most econometric computer packages will implement procedures of this type, and call it "robust" standard errors, or "HAC" corrected errors, whete HAC stands for Heteroskedasticity Autocorrelation Corrected standard errors.
This is an important procedure when doing estimation in financial economics settings, many of the relationships we test use time series data.

Exercise

You are investigating the market model

$$r_{it} = a + br_{mt} + e_{it}$$

in the Norwegian Market, and apply it to the company Norsk Hydro (NHY). Collect monthly returns for NHY for the period 1980-, and monthly returns for a value weighted market index for the same period.
After having estimated the model you worry that the errors in the estimation may be autocorrelated. Calculate a statistic that informs you about this.
What can be done to offset any problems due to autocorrelation of errors?

Exercise Solution

Reading the data

```
> library(zoo)
> rets <- read.zoo("../../../../data/norway/ose_individual_stock
+                  format="%Y%m%d",skip=2,header=TRUE,sep=",")
> Rm <- read.table("../../../../data/norway/stock_market_indices
+                  header=TRUE,sep=";");
> rNHY <- rets$Norsk.Hydro
> vw   <- Rm$VW
> data <- merge(rNHY,vw,all=FALSE)
> rNHY <- data$rNHY
> rm   <- data$vw
```

Exercise Solution

First do the standard estimation

```
lm(formula = rNHY ~ rm)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.009918   0.003021  -3.283  0.00112 **
rm           1.134563   0.043720  25.951  < 2e-16 ***

Residual standard error: 0.0556 on 370 degrees of freedom
Multiple R-squared: 0.6454,Adjusted R-squared: 0.6444
F-statistic: 673.4 on 1 and 370 DF,  p-value: < 2.2e-16
```

Exercise Solution

Calculating the DW test

```
> library(car)
> durbinWatsonTest(reg)
 lag Autocorrelation D-W Statistic p-value
   1       0.9968775             0       0
 Alternative hypothesis: rho != 0
```

We see strong signs of significant first order autocorrelation.

Exercise Solution

Calculate HAC (Heteroskedasticity and Autocorrelation Consistent) standard errors.
Use HAC consistent estimation from package `sandwich`, and dompare it to *just* the HC consistent estimate

```
> library(sandwich)
> sandwich(reg)
              (Intercept)            rm
(Intercept)  7.651440e-06 -1.607607e-05
rm          -1.607607e-05  3.003058e-03
> sqrt(diag(vcov(reg)))
(Intercept)          rm
0.003020602 0.043720208
> sqrt(diag(vcovHC(reg)))
(Intercept)          rm
0.002787477 0.055822628
 > sqrt(diag(vcovHAC(reg)))
 (Intercept)          rm
 0.002927377 0.058289906
```

Exercise Solution

Compare the three cases

OLS	0.003020602	0.043720208
HC	0.002787477	0.055822628
HAC	0.002927377	0.058289906

Typical outcome, the HAC consistent estimates of standard errors being larger, but note that the OLS estimate for the constant actually is larger.

GLS - some theory

OLS under iid assumptions.

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e},$$

$$\mathbf{e} \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I}),$$

Here σ is a (known) constant, and $\mathbf{I}$ is the identity matrix. Under these assumptions

$$\mathbf{b}hat^{ols} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

has a normal distribution:

$$\mathbf{b}hat^{ols} \sim \mathcal{N}\left(b, \sigma(\mathbf{X}'\mathbf{X})^{-1}\right)$$

GLS - violates this

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e},$$

$$\mathbf{e} \sim (0, \Omega),$$

Example: Heteroscedasticity

$$\Omega = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & \\ \vdots & & \ddots & \\ & \cdots & & \sigma_n^2 \end{bmatrix}$$

GLS replaces OLS with alternative linear estimator by using a matrix $\mathbf{C}$ transforming original problem

$$\mathbf{C}^{-1}\mathbf{y} = \mathbf{C}^{-1}\mathbf{X}b + \mathbf{C}^{-1}\mathbf{e}$$

and calculate the ols estimator of this problem

$$\widehat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{y}$$

The choice of $\mathbf{C}$ from

$$\Omega = \mathbf{C}\mathbf{C}$$

Implying

$$(\Omega)^{-1} = (\mathbf{C}\mathbf{C})^{-1} = \mathbf{C}^{-1}\mathbf{C}^{-1}$$

$$\widehat{\mathbf{b}}^{GLS} = \left(\mathbf{X}'\Omega^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Omega^{-1}\mathbf{y}$$

Optimality GLS

To see that the GLS is optimal, realize that the transform $\mathbf{C}^{-1}$ translates the original model

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

into a *iid* estimation case

$$\mathbf{y}^* = \mathbf{X}^*\mathbf{b} + \mathbf{e}^*,$$

where

$$\mathbf{y}^* = \mathbf{C}^{-1}\mathbf{y}$$

$$\mathbf{X}^* = \mathbf{C}^{-1}\mathbf{X}$$

$$\mathbf{e}^* = \mathbf{C}^{-1}\mathbf{e},$$

Since

$$E[\mathbf{e}^*\mathbf{e}^{*\prime}] = \mathbf{C}^{-1}E[\mathbf{e}\mathbf{e}']\mathbf{C}^{-1} = \mathbf{C}^{-1}\Omega\mathbf{C}^{-1} = \mathbf{I},$$

the errors of the transformed model are *iid*, and the Gauss-Markov theorem can be applied to this, stating that the GLS estimator is BLUE.

Unknown covariance matrix

So far we have assumed that e.g.

$$\mathbf{y}_t = \mathbf{x}_t\mathbf{b} + \mathbf{e}_t$$

$$E[\mathbf{e}'\mathbf{e}] = \Omega$$

where Ω is known. What if we don't know the matrix Ω? One obvious way to proceed is then a two-step procedure.

1. Estimate $\mathbf{b}$ by some suboptimal, but consistent procedure, such as for example OLS in the linear case.
2. Use residuals from this first step estimation to estimate Ω, for example

$$\widehat{\Omega} = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$$

3. Use this estimated $\widehat{\Omega}$, just like Ω above to run GLS. For example, use the transform $\mathbf{C} = \hat{\Omega}^{\frac{1}{2}}$ in the case above.