GLS and related issues

November 24, 2021

Contents

1 Problems in multivariate regressions

1.1 Problems with assumed i.i.d. errors

While multicollinearity is more a problem of model selection, most of the work in econometrics is caused by 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)

2 NON-iid errors

2.1 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.

Readings (Theil, 1971, Ch 3) (Davidson and MacKinnon, 1993, 3.2, 5.5)

3 Heteroskedasticity

When we say iid we mean independent, identically distributed errors.

Heteroscedasticity is violation of the second part of the assumption, the errors are independent but the error variance is different across observations.

Example: You observe income from a number of intervals. An obvious grouping of individuals is into males and females. Suppose income from females is more variable (more part-time working etc).

This is a typical example of *heteroskedasticity*, error variance differing across observations.

3.1 Consequences of Heteroskedasticity

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.

3.2 Testing for heteroskedasticity

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

Exercise 1.

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?

Solution to Exercise 1.

Reading the data

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

Let us first do the standard estimation

```
> reg <- lm(rNHY ~ rm)
> summary(reg)
Call:
lm(formula = rNHY ~ rm)
Residuals:
      Min        1Q    Median        3Q       Max
-0.237912 -0.030945  0.000939  0.032770  0.221631

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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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)
> summary(het)

Call:
lm(formula = e2 ~ rm + rm2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.007790 -0.002331 -0.001579  0.000280  0.045714

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 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```

Now, the coefficient estimates are not significant.

Still, one can calculate so called Heteroskedasticity consistent standard errors, or "White corrected" standard errors. These typically will be slightly larger than the usual OLS errors. Doing so in this case, let us compare the two.

Consider HC or HAC 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
```

Compare the two cases

OLS	0.003020602	0.043720208
HC	0.002787477	0.055822628

3.3 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.

If we suspect the presence of heteroskedasticity, it is possible to correct for this problem. There is a standard procedure called heteroskedasticity corrected standard errors, or the "White correction."

4 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 $\widehat{\mathbf{b}}$,

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

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

The simplest possible such estimate is to replace $\mathbf{\Omega}$ with

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

althought there are alternative ways of doing such corrections.

5 Regressions involving 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.

5.1 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

5.2 How to test for autocorrelation?

Show one well known test statistic, the Durbin-Watson statistic.

5.3 The Durbin-Watson test

Suppose we run regressions involving time series data.

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

y_t and x_t are both time series, they are obervations of the same variable at different dates.

What can often happen in such data is that the errors e_t for "close" dates are related, if you have shocks that have effects over several dates of observations.

This is called "autocorrelation in errors"

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

To test for the presence of such problems we can use the Durbin Watson test, which tests for first order autocorrelation in the errors,

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

It is calculated as

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

Some intuition can be had by observing that

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

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

Exercise 2.

The Durbin-Watson test (DW) is a way to test a null hypothesis of no serial correlation in error terms. Suppose we have stated the hypothesized model at time t as:

$$y_t = \mathbf{X}_t\mathbf{b} + \varepsilon_t$$

where y_t is the dependent variable, $\mathbf{X}_t$ the explanatory variables, $\mathbf{b}$ the parameters and ε_t is the regression error.

We also allow for first-order serial correlation in the error term by adding to the above model formulation

$$\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t$$

where η_t are independently distributed random variables with mean zero and identical variances.

If we let e_t be the residual from the fitted regression, the Durbin-Watson statistic is calculated as

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

which can be expanded as

$$DW = \frac{\sum_{t=2}^{T} e_t^2 - 2\sum_{t=2}^{T} e_t e_{t-1} + \sum_{t=2}^{T} e_{t-1}^2}{\sum_{t=1}^{T} e_t^2}$$

1. Argue that the DW statistic is approximately equal to $2(1 - \hat{\rho})$, where $\hat{\rho}$ is the sample autocorrelation.
2. If you have a null hypothesis of no autocorrelation in the error terms, what is the expected value of DW?

Solution to Exercise 2.

1. Look at the expansion:

$$d = \frac{\sum_{t=2}^{T} e_t^2}{\sum_{t=1}^{T} e_t^2} - \frac{2\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=1}^{T} e_t^2} + \frac{\sum_{t=2}^{T} e_{t-1}^2}{\sum_{t=1}^{T} e_t^2}$$

   The first and third term are both sums with just one less observation of e_t, which in each case is approximately equal to one. The second term estimates $\frac{2\mathsf{COV}(e_t, e_{t-1})}{\mathsf{var}(e_t)}$ which is approximately equal to $2\hat{\rho}$. Hence,

$$d \approx 1 - 2\hat{\rho} + 1 = 2(1 - \hat{\rho})$$

2. With no autocorrelation, $\rho = 0$, expect d equal 2.

If there is no autocorrelation, $\rho = 0$, and $DW = 2(1 - 0) = 2$.
If there is no autocorrelation in the errors, DW=2.
The further it is from 2, the more likely there are problems.

5.4 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.

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

The assumed time series properties of the noise term can be used to construct the $\mathbf{\Omega}$ used in the GLS type estimation

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

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.

Exercise 3.

Consider a regression model of the standard type

$$y_t = X_t b + e_t$$

where y_t is the outcome of the dependent variable at time t, X_t the independent variables at time t, e_t the error term at time t, and b the parameters of interest. The independent variables X_t at time t may also include past values of y_t as well as other variables observed at earlier dates. We observe outcomes of y_t and X_t for a number of periods $t = 1, \ldots T$.

Suppose that the observation errors follow an AR(1) process

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

where $E[\varepsilon_t] = 0$ and var$(\varepsilon_t) = \sigma_\varepsilon^2$.

Let $\epsilon_0 = 0$ and stack the model into the matrices

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

1. Argue that the covariance matrix of the errors $\mathbf{e}$ will have the form

$$\Omega = E[ee'] = \sigma_\varepsilon^2 \begin{bmatrix} \frac{1}{1-\rho^2} & \rho & \rho^2 & \rho^3 & & & \rho^{T-1} \\ \rho & \frac{1}{1-\rho^2} & \rho & \rho^2 & & & \\ \rho^2 & \rho & \frac{1}{1-\rho^2} & \rho & & & \\ & & & \ddots & & & \\ & & & & \frac{1}{1-\rho^2} & \rho & \rho^2 \\ & & & & \rho & \frac{1}{1-\rho^2} & \rho \\ \rho^{T-1} & \rho^{T-2} & & & \rho^2 & \rho & \frac{1}{1-\rho^2} \end{bmatrix}$$

Solution to Exercise 3.

Argue for the structure of the covariance matrix.

$$y_t = x_t b + e_t$$

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

$$\varepsilon \sim (0, \sigma_\varepsilon^2)$$

Suppose $e_0 = 0$.

$$e_1 = \varepsilon_1$$

$$E[e_1 e_1'] = \sigma_\varepsilon^2$$

$$E[e_1 e_2'] = E[\varepsilon_1(\rho\varepsilon_1 + \varepsilon_2)] = E[\rho\varepsilon_1^2] = \rho E[\varepsilon_1^2] = \rho\sigma_\varepsilon^2]$$

$$E[e_2 e_2] = E[(\rho\varepsilon_1 + \varepsilon_2)(\rho\varepsilon_1 + \varepsilon_2)] = \rho^2 E[\varepsilon_1^2] + E[\varepsilon_2^2] = (\rho^2 + 1)\sigma_\varepsilon^2$$

$$E[e_3 e_3] = (\rho^4 + \rho^2 + 1)\sigma_\varepsilon^2$$

$$E[e_3 e_1] = E[(\rho e_2 + \varepsilon_3)\varepsilon_1] = \rho^2\sigma_\varepsilon^2$$

If we don't have $\varepsilon_0 = 0$, but the series start at $t = -\infty$, a bit more problematic to show, but still.

On the diagonal, terms of the form $\sigma_\varepsilon^2(1 + \rho^2 + \rho^4 + \ldots)$

This infinite sum can be shown to equal $\frac{1}{1-\rho^2}$ if $|\rho| < 1$.

Therefore end up with the given structure of the long term case.

The exact procedure is not important, but it is possible to go from 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 $\mathbf{\Omega}$ 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 4.

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?

Solution to Exercise 4.

Reading the data

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

Doing the estimation we generate the following output: Let us first do the standard estimation

```
> reg <- lm(rNHY ~ rm)
> summary(reg)
Call:
lm(formula = rNHY ~ rm)
Residuals:
      Min        1Q    Median        3Q       Max
-0.237912 -0.030945  0.000939  0.032770  0.221631

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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```

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.
Calculate HAC (Heteroskedasticity and Autocorrelation Consistent) standard errors.
The next table shows the result of doing so for this regression, using a Newey West calculation of the HAC.
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
```

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.

6 Multicollinearity

Topics:

- Model problem, choice of independent variables.
- detection: correlation matrix

While we are at the topic of model specification testing, right place to discuss the problem of multicollinearity.

When looking at multiple regressions

$$y = a + b_1 x_1 + ... + b_k x_k + e$$

there is always a question of whether we have chosen the correct explanatory variables.

A particular problem that occurs is termed multicollinearity. If we have this problem two (or more) of the explanatory variables x are strongly related.

We will see this in the data if $\text{corr}(x_i, x_j)$ is close to one. Call these variables (near) colinear.

Practical consequences

- Estimated standard errors of the near colinear variables will be very large.
- Estimated R^2 will be very large

Testing for multicollinearity

- Very difficult in practice
- Most common is to look at correlation of explanatory variables.

Solutions to multicollinearity

- Ignore it (caught by the large standard errors)
- Drop one (or more) of the variables
- Gather more data (since this is a problem of the data, not of the model itself)

Exercise 5.

Consider the exchange rate of the US/Japan. Changes in the exchange rate is likely to be affected by changes in the interest rates in both the US and Japan. Let us limit ourselves to the US interest rates, and consider a regression

$$\Delta JPY/USD = a + b\Delta\text{US Interest rate} + e,$$

where you look at *monthly changes* in the exchange rate and interest rates.

What interest rate you want to use is not obvious, both short and long term interest rates presumably affect the exchange rate. Consider 6 month, 1 year and 10 year interest rates. Using the period 1983-2008, first estimate a univariate regression with the one year interest rate as the explanatory variable.

$$\Delta JPY/USD = a + b\Delta\text{1 y US Interest rate} + e$$

Does it seem like interest rates are related to exchange rate?

Now add two other interest rates, the 6 month and 10 year interest rates.

$$\Delta JPY/USD = a + b_1\Delta\text{6 m US Interest rate} + b_2\Delta\text{1 y US Interest rate} + b_3\Delta\text{10 y US Interest rate} + e$$

What are the results for this regression?

Can you now conclude that interest rates affect exchange rates?

What explains the results?

Solution to Exercise 5.

The data have been read in from data gathered from the federal reserve. Use log differences of monthly data.

```
> dJPY <- diff(log(mJPY))
> dM6 <- diff(log(mM6))
> dY1 <- diff(log(mY1))
> dY10 <- diff(log(mY10))
```

Take the subset of data

```
> data <- merge(dJPY['1983/2008'],dY1,all=FALSE)
```

Doing the regressions

Univariate:

```
lm(formula = dJPY ~ dY1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.153025 -0.019539  0.003541  0.020068  0.103742

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.002241   0.001809  -1.238    0.217
dY1          0.079453   0.019645   4.044 6.62e-05 ***

Residual standard error: 0.03177 on 310 degrees of freedom
Multiple R-squared: 0.05012,Adjusted R-squared: 0.04706
F-statistic: 16.36 on 1 and 310 DF,  p-value: 6.624e-05
```

Now include three explanatory variables

```
> data <- merge(dJPY['1983/2008'],dM6,dY1,dY10,all=FALSE)
```

Results

```
lm(formula = dJPY ~ dM6 + dY1 + dY10)

Residuals:
      Min        1Q    Median        3Q       Max
-0.154688 -0.019307  0.003572  0.019880  0.104358

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.002222   0.001817  -1.223    0.222
dM6          0.001829   0.042585   0.043    0.966
dY1          0.067582   0.049632   1.362    0.174
dY10         0.024114   0.048318   0.499    0.618

Residual standard error: 0.03185 on 308 degrees of freedom
Multiple R-squared: 0.0509,Adjusted R-squared: 0.04165
F-statistic: 5.506 on 3 and 308 DF,  p-value: 0.001074
```

Summarize the results

	Model 1	Model 2
(Intercept)	0.00	0.00
	(0.00)	(0.00)
dY1	0.08***	0.07
	(0.02)	(0.05)
dM6		0.00
		(0.04)
dY10		0.02
		(0.05)
R^2	0.05	0.05
Adj. R^2	0.05	0.04
Num. obs.	312	312

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

To understand what is going on, look at the correlations between these variables:

```
> cor(as.matrix(cbind(dM6,dY1,dY10)))
           dM6       dY1      dY10
dM6  1.0000000 0.8792392 0.5687342
dY1  0.8792392 1.0000000 0.7167197
dY10 0.5687342 0.7167197 1.0000000
```

7 Omitted variables

What is it?

Left out variables that 1) related to one of the other explanatory variables 2) explains the dependent variable (y).

Result: Estimation of the variable correlated with omitted variable is biased.

Really a model specification issue, the omitted variable should have been included in the model.

Possible remedy: Do estimation separately for groups where it is possible to distinguish sorting by omitted variable.

Omitted variables bias

What is it? The error in estimates $\hat{b}$ of a regression when the regressor, or explanatory variable, X, are correlated with an omitted variable.

For this bias to be important, need

1. X is correlated with the omitted variable
2. The omitted variable is a determinant of the dependent variable y.

What do we know about omitted variable bias?

We can actually show its effects algebraically.

Suppose we have as the true equation

$$y = \beta_1 x_1 + \beta_2 x_2 + e$$

In estimation we omit data on x_2, and estimate

$$y = \beta_1 x_1 + e$$

The estimator of this regression, $\hat{\beta}_1$, has expectation

$$E[\hat{\beta}_1] = \beta_1 + b_{21}\beta_2$$

where b_21 is the coefficient in a regression $x_2 = b_{21}x_1 + \varepsilon$.

From this expression a couple of observations:

- If $\beta_2 = 0$, there is no omitted variables bias.
- If $b_{12} = 0$ there is no omitted variables bias.

Otherwise, the bias will depend on both

- the size of β_2 in the true model
- the correlation between x_1 and x_2 (which shows up in the estimate $\hat{b}_{12}$).

Omitted variables may cause the sign of the other estimates to switch.

Example

A well known example

Demand for food in macroeconomic relations.

Let Q_D: Demand for food per capita

P_D: A measure of the price of food.

Suppose you fit.

$$Q_D = \alpha + \beta_1 P_D + e$$

on annual data.

What sign do you expect on β_1?

– Negative, since a price increase is expected to lower demand.

Fit on annual US data, get

$$Q_D = \underset{(11.85)}{89.97} + \underset{(0.118)}{0.107} P_D$$

This could be theorized that food is a Giffen good, but it is a more likely explanation that something has been left out.

In this case it is that the disposable income Y has also changed over this period. If we include Y, per capita income, in the regression, get

$$Q_D = \alpha + \beta_1 P_D + \beta_2 Y + e$$

estimates

$$Q_D = \underset{(5.84)}{92.05} - \underset{(0.067)}{0.142} P_D + \underset{(0.031)}{0.236} Y$$

The coefficient on food prices becomes negative, higher prices, holding income constant, lowers demand.

Source: Girschick and Haavelmo (1947)

8 Generalized Least Squares. (GLS)

A bit of theory

8.1 What is generalized least squares?

8.1.1 OLS under iid assumptions.

Let us go back to the simple case of Ordinary Least Squares under it's simplest assumptions, where the model is

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

and the error terms $\mathbf{e}$ are identically and independently normally distributed, or

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

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

Under these assumptions we showed that the OLS estimator

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

has a normal distribution:

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

8.1.2 General covariance matrix.

Suppose we now relax the *iid* assumption, and allow for a more general covariance matrix $\mathbf{\Omega}$ for the error terms,

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

For example, if the error terms have different variance, but are independent (heteroskedasticity), the matrix $\mathbf{\Omega}$ would look like

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

We assume the elements of $\mathbf{\Omega}$ are known. Let us now consider estimation in this case.

8.2 OLS

We could try using the OLS estimator, just like in the iid case.

What we want to consider are the questions whether the estimator is *consistent* and *efficient*. To check consistency we check if the expectation of the OLS estimator $\widehat{\mathbf{b}}^{ols}$ equals the true parameter $\mathbf{b}$.

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

According to the model,

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

plug this $\mathbf{y}$ into the OLS estimator:

$$\begin{aligned} \widehat{\mathbf{b}}^{ols} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\left(\mathbf{X}\mathbf{b} + \mathbf{e}\right) \\ &= \mathbf{b} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{e} \end{aligned}$$

Take expected values:

$$\begin{aligned} E\left[\widehat{\mathbf{b}}^{ols}\right] &= \mathbf{b} + E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\right] \\ &= \mathbf{b} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E\left[\mathbf{e}\right] \\ &= \mathbf{b} \end{aligned}$$

Let us next calculate the covariance matrix of this:

$$\begin{aligned} & E\left[(\widehat{\mathbf{b}}-\mathbf{b})(\widehat{\mathbf{b}}-\mathbf{b})'\right] \\ &= E\left[\left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}-\mathbf{b}\right\}\left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}-\mathbf{b}\right\}'\right] \\ &= E\left[\left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\mathbf{b}+\mathbf{e})-\mathbf{b}\right\}\left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\mathbf{b}+\mathbf{e})-\mathbf{b}\right\}'\right] \\ &= E\left[\left\{(\mathbf{b}+(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{e})-\mathbf{b}\right\}\left\{(\mathbf{b}+(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{e})-\mathbf{b}\right\}'\right] \\ &= E\left[\left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\right\}\left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\right\}'\right] \\ &= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E\left[\mathbf{e}\mathbf{e}'\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

If you compare this to OLS in the iid case, where the covariance matrix is $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, using OLS in this case may not be the best thing to do, since this covariance matrix seems to be more complicated, which means it may be larger than the covariance matrix of the "optimal" estimator.

8.3 Linear transform.

If we think about it, the simplest thing to try would be a linear transform of some kind. In this case, a linear transform is a multiplication with *some* matrix $\mathbf{C}^{-1}$. For now, let us not think about the form of this matrix, and just look at the consequences of doing the transform. (We will see later where $\mathbf{C}$ is coming from.)

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

Remember how we found the OLS estimates as the minimum of the SSE function:

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

The first order conditions for this minimization problem is:

$$-2\left(\mathbf{C}^{-1}\mathbf{X}\right)'\left(\mathbf{C}^{-1}\mathbf{y}-\mathbf{C}^{-1}\mathbf{X}\mathbf{b}\right)$$

Simplify this expression and solve for $\mathbf{b}$:

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

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

Let us first show that this estimator is unbiased by checking that its expectation equals $\mathbf{b}$.

$$\begin{aligned} E[\widehat{\mathbf{b}}] &= E[\left(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{y} \\ &= E[\left(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\left(\mathbf{X}\mathbf{b}+\mathbf{e}\right) \\ &= E[\left(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X}\right)\mathbf{b}] + (\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}E[\mathbf{e}] \\ &= \mathbf{b} + (\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{0} \\ &= \mathbf{b} \end{aligned}$$

So this is also a consistent estimator, no matter what the form of $\mathbf{C}$ is. The question is then to find the form of the matrix $\mathbf{C}$ that is "good" in some sense. Since we want an estimate that is precise, we want our estimator $\widehat{\mathbf{b}}$ to have a low variance. Let us calculate the covariance matrix of $\widehat{\mathbf{b}}$ for any general $\mathbf{C}$.

$$\begin{aligned}
&E\left[(\widehat{\mathbf{b}}-\mathbf{b})(\widehat{\mathbf{b}}-\mathbf{b})'\right] \\
&= E\left[\left\{(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{y}-\mathbf{b}\right\}\left\{(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{y}-\mathbf{b}\right\}'\right] \\
&= E\left[\left\{(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}(\mathbf{X}\mathbf{b}+\mathbf{e})-\mathbf{b}\right\}\left\{(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}(\mathbf{X}\mathbf{b}+\mathbf{e})-\mathbf{b}\right\}'\right] \\
&= E\left[\left\{(\mathbf{b}+(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{e})-\mathbf{b}\right\}\left\{(\mathbf{b}+(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{e})-\mathbf{b}\right\}'\right] \\
&= E\left[\left\{(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{e}\right\}\left\{(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{e}\right\}'\right] \\
&= E\left[(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{e}\mathbf{e}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1}\right] \\
&= (\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}E\left[\mathbf{e}\mathbf{e}'\right]\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{\Omega}\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1}
\end{aligned}$$

So far the $\mathbf{C}^{-1}$ terms do not seem to do any good. But what if it were the case that

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

Then, since

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

the above would reduce to:

$$\begin{aligned}
&E\left[(\widehat{\mathbf{b}}-\mathbf{b})(\widehat{\mathbf{b}}-\mathbf{b})'\right] \\
&= (\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{\Omega}\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{\Omega}\mathbf{\Omega}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X})^{-1} \\
&= \left(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\right)\left(\mathbf{X}'\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{X}\right)^{-1} \\
&= \left(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\right)^{-1}
\end{aligned}$$

Using the matrix $\mathbf{C} = \mathbf{\Omega}^{\frac{1}{2}}$ to premultiply the data:

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

is termed the *generalized least squares* estimator.

The covariance matrix of $\widehat{\mathbf{b}}$ is:

$$\text{var}\left(\widehat{\mathbf{b}}^{gls}\right) = \left(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\right)^{-1}$$

Compare this to the covariance matrix of the OLS estimator:

$$\text{var}\left(\widehat{\mathbf{b}}^{ols}\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

8.4 Showing optimality of GLS

We want to compare the GLS covariance matrix

$$\text{var}\left(\widehat{\mathbf{b}}^{gls}\right) = \left(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\right)^{-1}$$

with the OLS covariance matrix.

$$\text{var}\left(\widehat{\mathbf{b}}^{ols}\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

We need to show that

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

is positive definite.

To do this, we use the following useful linear algebra result (See e.g. (Davidson and MacKinnon, 1993, App 2) for a proof)

Consider the matrix difference

$$\mathbf{A} - \mathbf{B}$$

Result: $\mathbf{A} - \mathbf{B}$ *is positive definite if and only if* $\mathbf{B}^{-1} - \mathbf{A}^{-1}$ *is positive definite.* It should be clear where this will be used.

We want to show that

$$\mathbf{V}(\widehat{\mathbf{b}}^{ols}) - \mathbf{V}(\widehat{\mathbf{b}}^{gls})$$

is positive definite. According to the above result, this is equivalent to showing that

$$\mathbf{V}(\widehat{\mathbf{b}}^{gls})^{-1} - \mathbf{V}(\widehat{\mathbf{b}}^{ols})^{-1}$$

is positive definite.

Calculate this quantity

$$\left(\left(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\right)^{-1}\right)^{-1} - \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right)^{-1}$$

and simplify

$$\begin{aligned} &= \mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{\Omega}\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} \\ &= \mathbf{X}'\left(\mathbf{\Omega}^{-1} - \mathbf{X}(\mathbf{X}'\mathbf{\Omega}\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{X} \\ &= \mathbf{X}'\mathbf{\Omega}^{-\frac{1}{2}}\left(\mathbf{I} - \mathbf{\Omega}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{\Omega}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{\frac{1}{2}}\right)\mathbf{\Omega}^{-\frac{1}{2}}\mathbf{X} \end{aligned}$$

We want to show that this is positive definite.

Step 1: Show

$$\left(\mathbf{I} - \mathbf{\Omega}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{\Omega}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{\frac{1}{2}}\right)$$

is an idempotent matrix.

$$
\begin{aligned}
&\left(\mathbf{I}-\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{\frac{1}{2}}\right)\left(\mathbf{I}-\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{\frac{1}{2}}\right)' \\
&= \left(\mathbf{I}-\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{\frac{1}{2}}\right)\left(\mathbf{I}-\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{X}'(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}\boldsymbol{\Omega}^{\frac{1}{2}}\right)' \\
&= \mathbf{I} \\
&\quad -\left(\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{\frac{1}{2}}\right) \\
&\quad -\left(\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{X}'(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}\boldsymbol{\Omega}^{\frac{1}{2}}\right) \\
&\quad +\left(\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{\frac{1}{2}}\right)\left(\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{X}'(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}\boldsymbol{\Omega}^{\frac{1}{2}}\right) \\
&= \mathbf{I} \\
&\quad -\left(\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{\frac{1}{2}}\right) \\
&\quad -\left(\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{X}'(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}\boldsymbol{\Omega}^{\frac{1}{2}}\right) \\
&\quad +\left(\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}')(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}\boldsymbol{\Omega}^{-\frac{1}{2}}\right) \\
&= \mathbf{I} \\
&\quad -\left(\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{\frac{1}{2}}\right) \\
&\quad -\left(\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{X}'(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}\boldsymbol{\Omega}^{\frac{1}{2}}\right) \\
&\quad +\left(\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}\boldsymbol{\Omega}^{-\frac{1}{2}}\right) \\
&= \mathbf{I} \\
&\quad -\left(\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{X}'(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}\boldsymbol{\Omega}^{\frac{1}{2}}\right)
\end{aligned}
$$

which is the matrix we started with. We have thus shown that it is idempotent

Step 2: Argue that all idempotent matrices are positive definite.
proof: If $\mathbf{A}$ is idempotent,

$$\omega'\mathbf{A}\omega = \omega'\mathbf{A}\mathbf{A}\omega = \omega'\mathbf{A}\mathbf{A}\omega = (\omega'\mathbf{A})(\mathbf{A}\omega)$$

This last is a quadratic norm, which we know to be non-negative.

Step 3: Since what we have is a quadratic form in an idempotent matrix, it is positive definite.

8.5 Alternatively

A simpler way to see that the GLS is optimal, is to 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

$$
\begin{aligned}
\mathbf{y}^* &= \mathbf{C}^{-1}\mathbf{y} \\
\mathbf{X}^* &= \mathbf{C}^{-1}\mathbf{X} \\
\mathbf{e}^* &= \mathbf{C}^{-1}\mathbf{e},
\end{aligned}
$$

To see this, realize that since

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

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.

Exercise 6.

Consider the standard model: $y_t = x_t b + e_t$, which hold for each t, and we assume that $e_t \sim (0, \sigma^2), \quad t = 1, \ldots, T$, each error term e_t has mean 0 and a constant variance σ^2.

Observations of this model has been collected by several groups. Each group has only reported the sample mean of their observations and the number of observations used to generate the mean.

The problem is that we do not observe all (say) T observations, but the observations are grouped into (say) N*groups*, and we only know the sample means of the grouped data, and how many observations (n_i) were used in generating this sample mean.

The next figure illustrates the situation. The underlying data is listed on the left, and the data we know on the right.

observations	n_i	$\bar{x}$	$\bar{y}$
$x_{11}\ x_{12}\ \cdots\ x_{1n_1}$	n_1	$X_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{1j}$	$Y_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} y_{1j}$
$x_{21}\ x_{22}\ \cdots\ x_{2n_2}$	n_2	$X_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2j}$	$Y_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} y_{2j}$
	$\vdots$		
$x_{N1}\ x_{N2}\ \cdots\ x_{N,n_N}$	n_N	$X_N = \frac{1}{n_N}\sum_{j=1}^{n_N} x_{Nj}$	$Y_N = \frac{1}{n_N}\sum_{j=1}^{n_N} y_{Nj}$

The number of observations in each group (n_i) may be different, but they sum to the total number of observations, $\sum_{i=1}^{N} n_i = T$.

1. How should this be optimally estimated using GLS?

Solution to Exercise 6.

We want to estimate $\mathbf{b}$ given the data

$$\{(X_1, Y_1, n_1), (X_2, Y_2, n_2), \ldots, (X_N, Y_N, n_N)\}$$

When you take the mean of both y and x, you preserve the linear relationship:

$$y_t = x_t b + e_t,$$

since

$$\frac{1}{n_i}\sum_{j=1}^{n_i}(x_{ij}b) = \left(\frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}\right) b = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

Thus, the model still applies to the stratified means:

$$Y_i = X_i b + \varepsilon_i,$$

but we no longer have *iid* errors as long as the number of observations in each group is different. This is because

$$\mathsf{var}(X_i) = \mathsf{var}\left(\frac{1}{n_i}\sum_{j=1}^{n_1} x_{ij}\right) = \left(\frac{1}{n_i}\right)^2 \sum_{j=1}^{n_i} \mathsf{var}(x_{ij}) = \frac{1}{(n_i)^2} n_i \sigma^2 = \frac{1}{n_i}\sigma^2$$

Then the matrix $\mathbf{\Omega} = E\left[\varepsilon\varepsilon'\right]$, the covariance matrix of the errors used in generating the sample estimates, looks like the following:

$$\mathbf{\Omega} = \sigma^2 \begin{bmatrix} \frac{1}{n_1} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & 0 & \cdots & 0 \\ & & \ddots & & \\ 0 & & & & \frac{1}{n_N} \end{bmatrix}$$

Now, remember that we want the matrix $\mathbf{C}$ that is defined by $\mathbf{CC} = \mathbf{\Omega}$, or $\mathbf{C} = \mathbf{\Omega}^{\frac{1}{2}}$. In this case it is easy to see that

$$\mathbf{C} = \mathbf{\Omega}^{\frac{1}{2}} = \sigma \begin{bmatrix} \frac{1}{\sqrt{n_1}} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{n_2}} & 0 & \cdots & 0 \\ & & \ddots & & \\ 0 & & & & \frac{1}{\sqrt{n_N}} \end{bmatrix}$$

Thus, to run GLS on this problem, we transform the data to be of the following form:

$$\mathbf{C}^{-1}\mathbf{Y}_i = \mathbf{C}^{-1}\mathbf{X}_i b + \mathbf{C}^{-1}\varepsilon$$

The matrix $\mathbf{C}^{-1}$ is found as:

$$\mathbf{C}^{-1} = \frac{1}{\sigma}\begin{bmatrix} \sqrt{n_1} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{n_2} & 0 & \cdots & 0 \\ & & \ddots & & \\ 0 & & & & \sqrt{n_N} \end{bmatrix}$$

We observe that we want to transform each observation X_i by multiplying by a factor of $\sqrt{n_i}$.

8.6 Unknown covariance matrix.

So far we have assumed that e.g.

$$\mathbf{y}_t = \mathbf{x}_t\mathbf{b} + \mathbf{e}_t$$
$$E[\mathbf{e}'\mathbf{e}] = \mathbf{\Omega}$$

where $\mathbf{\Omega}$ is known.

What if we don't know the matrix $\mathbf{\Omega}$?

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 $\mathbf{\Omega}$, for example

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

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

This type of multi-step procedure will often produce good estimates, in particular if we can parameterize $\mathbf{\Omega}$ by a few parameters.

This type of multi-step procedure is called *feasible GLS*. An alternative to this is to use Maximum Likelihood.

Readings. (Davidson and MacKinnon, 1993, 9.2) discusses the known covariance matrix case, and (Davidson and MacKinnon, 1993, 9.6) the unknown covariance matrix case. (Davidson and MacKinnon, 1993, 9.6) how to do feasible GLS.

(White, 1984, Ch 1) is a more general discussion of GLS.

References

Russel Davidson and James G MacKinnon. *Estimation and Interference in Econometrics*. Oxford University Press, 1993.

M A Girschick and T Haavelmo. Statistical analysis of the demand for food. *Econometrica*, April 1947.

Henri Theil. *Principles of econometrics*. Wiley, 1971.

Halbert White. *Asymptotic theory for econometricians*. Academic Press, 1984.