Revision Tutorial

Stuart Breslin

December 2021

1 Introduction

By popular demand I'll cover the following topics

- Generalised Method of Moments
- Weak Identification
- Specification Tests
- Nonlinear Estimators and Maximum Likelihood

2 Generalised Method of Moments

To remind you gently of GMM, lets compare it to OLS. In OLS we have the regression model

$$y = X\beta + \varepsilon$$

We obtain an estimate for β by minimising the sum of squared residuals $\hat{\varepsilon}'\hat{\varepsilon}$:

$$\begin{aligned}
\frac{\partial}{\partial\hat{\beta}}(\varepsilon'\varepsilon) &= 0 \\
\frac{\partial}{\partial\hat{\beta}}((y' - \hat{\beta}X')(y - X\hat{\beta})) &= 0 \\
\frac{\partial}{\partial\hat{\beta}}(y'y - y'X\hat{\beta} - \hat{\beta}X'y + \hat{\beta}X'X\hat{\beta}) &= 0 \\
-2X'y + 2X'X\hat{\beta} &= 0 \\
\hat{\beta}_{OLS} &= (X'X)^{-1}X'y
\end{aligned}$$

One of the crucial assumptions of our data that we rely on to obtain a consistent estimate for β is weak exogeneity i.e. $E(X_i'\varepsilon_i) = 0$. This object $E(X_i'\varepsilon_i)$ is $k \times 1$, where k is the number of regressors (including the "1' term).

Let me now convince you that we don't need to do OLS, but can instead use the weak exogeneity assumption to obtain an estimate of β directly. Instead of minimising $\hat{\varepsilon}'\hat{\varepsilon}$, start with the sample moment condition:

$$\frac{1}{n}\sum_{i=1}^{n} X_i'\hat{\varepsilon}_i = 0$$

Why call this a sample moment condition? The analogous population moment condition is $E(X_i'\varepsilon_i) = 0$. But we can't see ε, but we can only see $\hat{\varepsilon}$. And we can't observe the true expectation, so we work with the sample average instead.

This can be written in matrix form (and multiply by n) to get:

$$\begin{aligned} X'\hat{\varepsilon} &= 0 \\ X'(y - X\hat{\beta}) &= 0 \\ X'y - X'X\hat{\beta} &= 0 \\ \hat{\beta}_{MM} &= (X'X)^{-1}X'y \end{aligned}$$

By some strange trickery, we get to the same estimate as OLS but without minimising the sum of squared residuals, but instead working with the sample moment condition directly. What was done here is known as "method of moments". There are k moment conditions, and these are used to inform the k parameters of the model. The basic philosophy of this is that what is true for the population moments should also hold for the sample moments.

2.1 The Instrumental Variables estimator (if weak exogeneity fails?)

Sometimes our weak exogeneity assumption is unrealistic. We may be trying to estimate the effect of education on wages, but are worried that education is endogenous. Endogeneity can arise for a few different reasons:

- Ommitted variable bias
- Reverse causality
- Measurement error
- Non-random sample selection

In this example, this could be:

- Parental income helps children get more education and helps them get a high wage job, but we don't control for this in our regression (meaning it is in the error term)
- Individuals anticipating high wage careers take on more education
- Education is self-reported and they are forgetful (or liars)

- We don't observe the wages of some of the population (because they are unemployed) or some individuals are more likely to drop out of the sample for reasons that also affect wages

Without our weak exogeneity assumption, $\hat{\beta}$ will be inconsistent. So it seems like we are hopeless here. The ideal situation would be some sort of experiment in which we randomly choose children to be put in varying educational scenarios. The randomness of such an approach would guarantee us weak exogeneity. While experimental economics is on the rise in the 21st century (particularly within development and behavioural economics), it is still unusual due to the high financial and ethical costs. The next best thing is some sort of "natural experiment". Is there some sort of random variable/event/influence that makes children more or less likely to get extra education? We call such a thing an instrumental variable. For example, Card (who recently won the Nobel prize) used proximity to college as a variable that does this. Other examples generally in economics include policy reforms, forecasting errors, weather. Ideally, an instrumental variable should be:

- Relevant (correlated "enough" with the endogenous regressor)
- Exogenous

Often the validity of an instrumental variable approach depends on these two factors. The first (which we will return to when discussing weak identification) can be tested. The second however, is generally subjective and usually relies on human intuition. For example, suppose an educational reform to improve schooling is implemented in some areas but not others. This seems like a great natural experiment; we can learn a lot by comparing wages of those who were subjected to the policy reform to their similar counterparts in the unreformed areas. But do we know that which areas implemented the reform vs. which did not are actually random? If it's non-random in a way captured by our control variables, then we are fine, but otherwise there could still be endogeneity issues. Often in instrumental variables approaches, the results are compared to OLS, as if to get a sense of the direction of OLS inconsistency, even if there is not full confidence that IV removes all the inconsistency.

In our IV estimator, we have the same number of instrumental variables as we do endogenous regressors. We still have the same notation for the regression model:

$$y = X\beta + \varepsilon,$$

where X contains all the regressors, endogenous and exogenous. We also introduce Z, which contains everything exogenous. Let's return to the example I mentioned:

1. Years of schooling would only be put in X, as we are treating it as our endogenous variable of interest
2. Control variables, such as gender, race, age etc. would be put in both X and Z

3. Proximity to college would only be put in Z, as it is not a regressor i.e. does not effect y (wages) directly, but satisfies exogeneity

We could also do IV if for example we had two endogenous regressors and two instrumental variables. More generally, there are three types of variable:

1. Endogenous regressors (put in X only)
2. Exogenous regressors or controls (put in both X and Z; known in the lectures as "included instruments")
3. Instrumental variables (put in Z only; known as "excluded instruments")

Our population moment condition is: $E(Z_i'\varepsilon_i) = 0$, which is $k \times 1$. Note that this notation means that we are not applying the weak exogeneity assumption to the endogenous regressors. Let's use the "method of moments" approach before, and work with the analogous sample moment condition:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} Z_i'\hat{\varepsilon}_i &= 0 \\ Z'\hat{\varepsilon} &= 0 \\ Z'(y - X'\hat{\beta}) &= 0 \\ Z'y - Z'X\hat{\beta} &= 0 \\ \hat{\beta}_{IV} &= (Z'X)^{-1}Z'y \end{aligned}$$

Note that in OLS, we are treating everything as exogenous, so you can see that setting $Z = X$ just gets us the OLS estimator. We now have a consistent estimator (I'll show the proof for the GMM case). Note however that IV is not technically unbiased; there is still a "small sample bias" phenomenon, I'll also discuss this for GMM case.

2.2 Two Stage Least Squares (if we have more instruments?)

What about cases where we have more instrumental variables than we do endogenous regressors? This is known as an overidentified model (the opposite case, an underidentified model is impossible to estimate). Here we will distinguish between k, the number of regressors and l, the number of exogenous variables. For example, if we had one endogenous regressor but two instrumental variables, then $l = k + 1$. To see how this works note that:

$$\begin{aligned} k &= \text{no. of endogenous regressors} + \text{no. of controls} \\ l &= \text{no. of instrumental variables} + \text{no. of controls} \end{aligned}$$

So the gap between l and k only depends on how many instrumental variables we have compared to endogenous regressors. What happens to our IV estimator?

$$\hat{\beta}_{IV} = (Z'X)^{-1}Z'y$$

This no longer exists because $Z'X$ is $l \times k$ (remember Z is $n \times l$ and X is $n \times k$) and we can only invert square matrices. Two stage least squares finds a way round this: first create an exogenous proxy for X, called $\hat{X}$, and perform OLS on $\hat{X}$. To construct the proxy, we first "regress" X on Z (note this is actually multiple regressions, but the notation is similar):

$$X = Z\lambda + \nu$$

OLS estimate:

$$\hat{\lambda} = (Z'Z)^{-1}Z'X$$

Now construct $\hat{X}$:

$$\hat{X} = Z(Z'Z)^{-1}Z'X$$

Let's use the notation $P_Z = Z(Z'Z)^{-1}Z'$, where P_Z is idempotent and symmetric.

What is the idea here? Well $\hat{X} = Z\hat{\lambda}$ is now "made out of Z", so we can treat $\hat{X}$ like it's exogenous. Note however this is not strictly true for small samples as $\hat{\lambda}$ is random and depends on X. However, as our sample gets larger and $\hat{\lambda}$ converges, $\hat{X}$ becomes like an exogenous proxy for X. Now do OLS of y on $\hat{X}$:

$$\hat{\beta} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$$

Substitute $\hat{X} = P_Z X$:

$$\hat{\beta} = (X'P_Z'P_Z X)^{-1}X'P_Z'y$$

Use the idempotent and symmetric features of the projection matrix:

$$\hat{\beta} = (X'P_Z X)^{-1}X'P_Z y$$

Now substitute $P_Z = Z(Z'Z)^{-1}Z'$ to get our 2SLS estimator:

$$\hat{\beta}_{2SLS} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y$$

Great, now we have an intuitive way to estimate overidentified models. And look at what happens if $l = k$: then $(X'Z)$ and $(Z'X)$ are both invertible. Use the rule:

$$(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$$

$$\begin{aligned}
\hat{\beta}_{2SLS:l=k} &= (Z'X)^{-1}(Z'Z)(X'Z)^{-1}X'Z(Z'Z)^{-1}Z'y \\
\hat{\beta}_{2SLS:l=k} &= (Z'X)^{-1}Z'y \\
\hat{\beta}_{2SLS:l=k} &= \hat{\beta}_{IV}
\end{aligned}$$

So IV can be thought of a special case of 2SLS when $l = k$.

While this two step procedure provides us a way of getting round the fact that $(Z'X)$ is not invertible, there is actually a different approach that will turn out to be "better".

2.3 The Generalised Method of Moments Estimator

Let's return to the population moment condition:

$$E(Z_i'\varepsilon_i) = 0,$$

and its sample analogue:

$$\frac{1}{n}\sum_{i=1}^{n} Z_i'\hat{\varepsilon}_i = 0.$$

This is a set of l equations, but we only have k unknowns (the β's), which is where our problem comes in. We cannot get the above set of equations to hold exactly. Let's imagine a generic situation where we have two equations, which look like this:

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

But suppose we are not able to get a_1 and a_2 to hit zero exactly. A different approach we can use is to penalise a_1 and a_2 for being far away from zero. The most convenient way to do this is quadratically (this is nice because the derivative of a quadratic is linear). For example:

$$\text{pentalty} = a_1^2 + a_2^2$$

Our penalty indifference curves (if you can imagine such a thing) will look like circles around the origin. But what if a_1 and a_2 are completely different in scale, so we might want to allow for different weightings e.g.

$$\text{pentalty} = w_1 a_1^2 + w_2 a_2^2,$$

where $w_1 > 0$, $w_2 > 0$. Here the indifference curves will be ellipses stretched horizontally or vertically depending on w_1/w_2. However, we can be even more general by allowing the cross term to be penalised:

$$\text{pentalty} = w_1 a_1^2 + w_2 a_2^2 + 2w_{12} a_1 a_2$$

Here, not only may our indifference curves be elliptical, but they can be be rotated depending on the sign of w_{12}. The "2" seems arbitrary, but fits with the following notation:

$$\text{penalty} = \begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

If we break this out we get $w_{11}a_1^2 + w_{22}a_2^2 + (w_{12} + w_{21})a_1 a_2$. Since w_{12} vs. w_{21} doesn't matter, we may as well force them to be equal. We must also ensure that this weighting matrix is positive definite, otherwise we may actually be reducing the penalty when me move further away from 0 in some directions. This can be generalised to:

$$\text{penalty} = a'Wa,$$

where a is an $l \times 1$ vector that we want to be close to zero and W is a $l \times l$ positive definite weighting matrix.

Let's now think of what a "good" choice for W is. Imagine our a_1 and a_2 as being random variables. Let's suppose a_1 has a standard deviation of 100, and a_2 has a standard deviation of 0.02. Suppose we chose W to be the identity matrix, in which case we have

$$\text{penalty} = a_1^2 + a_2^2$$

In this case, the random variation in a_1 will largely obscure any effects of a_2; so the penalty term seems unbalanced. A much more sensible approach would be the following:

$$\text{penalty} = (\frac{a_1}{100})^2 + (\frac{a_2}{0.02})^2$$

This would correspond to

$$W = \begin{bmatrix} 10,000 & 0 \\ 0 & 0.0004 \end{bmatrix}^{-1}$$

You can see here that we chose:

$$W = \begin{bmatrix} Var(a_1) & 0 \\ 0 & Var(a_2) \end{bmatrix}^{-1}$$

It seems that we could also account for the covariance by having:

$$W = \begin{bmatrix} Var(a_1) & Cov(a_1, a_2) \\ Cov(a_2, a_1) & Var(a_2) \end{bmatrix}^{-1}$$

Then our penalty looks like:

$$\text{penalty} = \frac{Var(a_2)a_1^2 + Var(a_1)a_2^2 - 2Cov(a_1, a_2)a_1a_2}{Var(a_1)Var(a_2) - Cov(a_1, a_2)^2}$$

You can see here that comparing the weights on a_1^2 and a_2^2, more weight is applied to the term with smaller variance, which seems sensible. But we also account for the cross term a_1a_2. Note that the sign on this cross term depends on the covariance of a_1 and a_2. Imagine $Cov(a_1, a_2) > 0$. Then our penalty will tend to downweigh an outcome where a_1 and a_2 are the same sign but increase the penalty if they are different signs. This makes perfect sense! If a_1 and a_2 are highly correlated, then be so surprised if they are both positive or both negative, so having a negative on the a_1a_2 term is appropriate.

Hopefully this convinces you that $W = Var(a)^{-1}$ is a sensible choice for a weighting matrix.

Let's return to our problem:

$$\frac{1}{n}\sum_{i=1}^{n} Z_i'\hat{\varepsilon}_i = 0.$$

Since we can't solve this exactly: instead we should minimise our penalty, which in the GMM context we call the J statistic:

$$J(\hat{\beta}) = (\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} Z_i'\hat{\varepsilon}_i)'W(\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} Z_i'\hat{\varepsilon}_i)$$

Note that I have multiplied the sample moment by $\sqrt{n}$. This ensures that it has a well defined asymptotic distribution. This obviously has no bearing on the minimisation problem. This means you will often see the J statistic written like:

$$J(\hat{\beta}) = n(\frac{1}{n}\sum_{i=1}^{n} Z_i'\hat{\varepsilon}_i)'W(\frac{1}{n}\sum_{i=1}^{n} Z_i'\hat{\varepsilon}_i)$$

And our sensible choice of W?

$$W = \left[Var(\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} Z_i'\hat{\varepsilon}_i)\right]^{-1},$$

AVar vs Var

or equivalently:

$$W = \left[AVar(\frac{1}{n}\sum_{i=1}^{n} Z_i'\hat{\varepsilon}_i)\right]^{-1}$$

The notation in the lectures also introduces the following:

$$\begin{aligned} g_i(\hat{\beta}) &= Z_i'\hat{\varepsilon}_i \\ \bar{g}_n(\hat{\beta}) &= \frac{1}{n}\sum_{i=1}^{n} Z_i'\hat{\varepsilon}_i, \\ S &= AVar(\bar{g}_n(\hat{\beta})) \end{aligned}$$

Which allows the J stat to be written more compactly as

$$J(\hat{\beta}) = n\bar{g}_n(\hat{\beta})'W\bar{g}_n(\hat{\beta}),$$

with sensible:

$$W = S^{-1}$$

Let's work out the GMM estimator for a given W:

$$\frac{\partial}{\partial\hat{\beta}}\big(n(\frac{1}{n}\sum_{i=1}^{n}Z_i'\hat{\varepsilon}_i)'W(\frac{1}{n}\sum_{i=1}^{n}Z_i'\hat{\varepsilon}_i)\big)=0$$

$$\frac{\partial}{\partial\hat{\beta}}\big((Z'\varepsilon)'W(Z'\varepsilon)\big)=0$$

$$\frac{\partial}{\partial\hat{\beta}}\big((y-X\hat{\beta})'ZWZ'(y-X\hat{\beta})\big)=0 \quad \text{treat ZWZ as constant}$$

$$\frac{\partial}{\partial\hat{\beta}}\big(y'ZWZ'y-y'ZWZ'X\hat{\beta}-\hat{\beta}'XZWZ'y+\hat{\beta}'X'ZWZ'X\hat{\beta}\big)=0 \quad \text{similar to OLS}$$

$$-2X'ZWZ'y+2X'ZWZ'X\hat{\beta}=0$$

$$\hat{\beta}_{GMM}=(X'ZWZ'X)^{-1}X'ZWZ'y$$

Now that we have the GMM estimator, let's check its asymptotic properties. To show consistency:

$$\hat{\beta}_{GMM}=(X'ZWZ'X)^{-1}X'ZWZ'X\beta+(X'ZWZ'X)^{-1}X'ZWZ'\varepsilon$$

$$\hat{\beta}_{GMM}=\beta+(X'ZWZ'X)^{-1}X'ZWZ'\varepsilon$$

$$\hat{\beta}_{GMM}-\beta=(X'ZWZ'X)^{-1}X'ZWZ'\varepsilon$$

$$\hat{\beta}_{GMM}-\beta=\Big(\sum_{i=1}^{n}[X_i'Z_i]\times W\times\sum_{i=1}^{n}[Z_i'X_i]\Big)^{-1}\sum_{i=1}^{n}[X_i'Z_i]\times W\times\sum_{i=1}^{n}[Z_i'\varepsilon_i]$$

$$\hat{\beta}_{GMM}-\beta=\Big(\frac{1}{n}\sum_{i=1}^{n}[X_i'Z_i]\times W\times\frac{1}{n}\sum_{i=1}^{n}[Z_i'X_i]\Big)^{-1}\frac{1}{n}\sum_{i=1}^{n}[X_i'Z_i]\times W\times\frac{1}{n}\sum_{i=1}^{n}[Z_i'\varepsilon_i]$$

By the weak law of large numbers:

$$\frac{1}{n}\sum_{i=1}^{n}X_i'Z_i\underset{p}{\to}E[X_i'Z_i]$$

$$\frac{1}{n}\sum_{i=1}^{n}Z_i'X_i\underset{p}{\to}E[Z_i'X_i]$$

$$\frac{1}{n}\sum_{i=1}^{n}[Z_i'\varepsilon_i]\underset{p}{\to}E[Z_i'\varepsilon_i]=0$$

And by the continuous mapping theorem and Slutsky's theorem, we may mul-

CMT: talk in weak identification

tiply these together and apply the inverse:

$$\hat{\beta}_{GMM} - \beta \underset{p}{\to} \Big(E[X_i'Z_i] \times W \times E[Z_i'X_i]\Big)^{-1} E[X_i'Z_i] \times W \times E[Z_i'\varepsilon_i]$$
$$\hat{\beta}_{GMM} - \beta \underset{p}{\to} 0$$

Great, so we know $\hat{\beta}_{GMM}$ is consistent. What about it's asymptotic distribution? root n is placed at zero-mean term, then can use CLT

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta) = \Big(\frac{1}{n}\sum_{i=1}^{n}[X_i'Z_i] \times W \times \frac{1}{n}\sum_{i=1}^{n}[Z_i'X_i]\Big)^{-1} \frac{1}{n}\sum_{i=1}^{n}[X_i'Z_i] \times W \times \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}[Z_i'\varepsilon_i]$$

from CMT:
if element in a function converges,
the function also converges

Notice the strategic placement of $\sqrt{n}$.

First deal with the left chunk. By WLLN, CMT and Slutsky:

$$\Big(\frac{1}{n}\sum_{i=1}^{n}[X_i'Z_i] \times W \times \frac{1}{n}\sum_{i=1}^{n}[Z_i'X_i]\Big)^{-1} \frac{1}{n}\sum_{i=1}^{n}[X_i'Z_i] \times W \underset{p}{\to} \Big(\Sigma_{ZX}'W\Sigma_{ZX}\Big)^{-1}\Sigma_{ZX}'W,$$

where $\Sigma_{ZX} = E(Z_i'X_i)$.

Now for the right part. By the central limit theorem:

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}[Z_i'\varepsilon_i] \underset{d}{\to} N(0, AVar(\frac{1}{n}\sum_{i=1}^{n}[Z_i'\varepsilon_i])) = N(0, S)$$

Combining both with Slutsky's theorem, remember:

$$\begin{aligned} \text{if} \quad & \hat{a} \underset{p}{\to} a \\ \text{and} \quad & \hat{b} \underset{d}{\to} n(0, v_b) \\ \text{then} \quad & \hat{a}\hat{b} \underset{d}{\to} N(0, av_ba') \end{aligned}$$

This means:

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta) \underset{d}{\to} N\Big(0, \big(\Sigma_{ZX}'W\Sigma_{ZX}\big)^{-1}\Sigma_{ZX}'WSW\Sigma_{ZX}\big(\Sigma_{ZX}'W\Sigma_{ZX}\big)^{-1}\Big)$$

Very messy! But if we choose the sensible W, $W = S^{-1}$, this simplifies greatly:

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta) \underset{d}{\to} N\Big(0, \big(\Sigma_{ZX}'S^{-1}\Sigma_{ZX}\big)^{-1}\Big)$$

Note: We can only use this simplification because W is chosen efficiently. If we do 2SLS (where $W = (Z'Z)^{-1}$) we should generally use the messier form.

do not know S?

2.4 The two stage feasible efficient estimator

We have an equation for $\hat{\beta}_{GMM}$ and we know its asymptotic properties, but this assumed we knew $S = AVar(\frac{1}{n}\sum_{i=1}^{n} Z_i'\varepsilon_i)$. In practice, we don't know this, but can estimate S: $\hat{S}$.

First obtain the two stage least squares estimator to get a consistent estimate of β, and collect the error terms $\hat{\varepsilon}$.

Then estimate

$$\hat{S} = A\hat{V}ar(\frac{1}{n}\sum_{i=1}^{n} Z_i'\varepsilon_i),$$

using the appropriate covariance estimator (classical, HC, CR or HAC).

Then estimate $\hat{\beta}_{FGMM}$ using $\hat{W} = \hat{S}^{-1}$:

$$\hat{\beta}_{FGMM} = (X'Z\hat{W}Z'X)^{-1}X'Z\hat{W}Z'y,$$

and we have our asympototic variance estimate:

$$A\hat{V}ar(\hat{\beta}_{FGMM}) = \left(S_{ZX}'\hat{S}^{-1}S_{ZX}\right)^{-1},$$

where $S_{ZX} = \frac{1}{n}\sum_{i=1}^{n} Z_i'X_i$ is our sample estimate of Σ_{ZX}. A special case should be noted. If we choose the classical estimate:

$$\hat{S} = \hat{\sigma}(Z'Z)^{-1},$$

then we end up with the 2SLS estimator. This means that under conditional homoskedasticity, the 2SLS turns out to be efficient! However, conditional homoskedasticity is generally a bad assumption, so we can use either HC, CR or HAC estimates for S, depending on whether we have cross section, panel or time series, and obtain asymptotic efficiency gains for our GMM estimator over the 2SLS estimator. Re-iterated GMM estimators are also possible; we could take the error terms and re-estimate $\hat{S}$, plug it back in... until the estimator meets some sort of numerical convergence, but the asymptotics remain the same.

3 Weak Identification

Take a close look at

$$\hat{\beta}_{GMM} = (X'ZWZ'X)^{-1}X'ZWZ'y$$

Generally, $(X'ZWZ'X)^{-1}$ will be invertible, barring some sort of perfect multicolinearity in the data. But consider the asymptotics used (WLLN+CMT):

$$(X'ZWZ'X)^{-1} \underset{p}{\to} (\Sigma_{Zx}'W\Sigma_{Zx})^{-1}.$$ if it is not inverstible (singular)?

The continuous mapping theorem only works if the function is continuous at the point of convergence. If $f(.)$ is not continuous at $f(E(a_i))$, then $f(\frac{1}{n}\sum_{i=1}^{n} a_i)$

is not guaranteed to converge to $f(E(a))$. Our function is the inversion of a matrix. This function is not continuous if the matrix being inverted is singular.

Take a simpler example. Suppose we just have a scalar. What happens to

$$(\frac{1}{n}\sum_{i=1}^{n} a_i)^{-1}$$

Usually we would just say this converges to $(E(a_i))^{-1}$. But what if $E(a_i) = 0$. Then it doesn't converge at all, it will blow up to positive or negative infinity depending on what side it happens to land on. The analogous problem for our GMM estimator is that if $\Sigma'_{Zx} W \Sigma_{Zx}$ is not invertible, then our GMM estimator will fail to converge. This is actually equivalent to saying $Rank(\Sigma_{ZX}) < k$:

$$\begin{aligned} &\text{iff} \quad Rank(\Sigma_{ZX}) < k : \\ &\Sigma'_{Zx} W \Sigma_{Zx} \quad \text{is singular} \end{aligned}$$

columns are correlated-->Z X are correlated

The "rank condition" is therefore $Rank(\Sigma_{ZX}) = k$, which guarantees $\Sigma'_{Zx} W \Sigma_{Zx}$ is invertible.

What does this mean in practice? The rank condition holds if the instrumental variables are relevant i.e. correlated with the endogenous regressors. In our case mentioned before, imagine proximity to college had no effect on schooling. Then it would be completely useless as a natural experiment. Subsequently, our IV/2SLS/GMM estimators will fail to converge.

3.1 Testing for Weak Identification

In the case where there is just one endogenous regressor, the procedure is relatively straightforward. Let x_e denote the endogenous regressor. Run OLS:

$$x_e = Z\lambda + \nu$$

Now remember Z includes everything exogenous; controls and excluded instruments! Now test the null-hypothesis that the coefficients **on the excluded instruments only** are all equal to zero by running an F-test on:

why? 2 reasons

$$H_0 : \quad \lambda_g = 0 \quad \forall g \in \text{excluded instruments}$$

The traditional rule of thumb is to say that we have weak instruments if $F < 10$. Why do we include the controls in the regression but not the test? For the excluded instruments to be relevant, they must correlate with the endogenous regressor in a way not already captured by the controls. Hence why we regress it on all of Z but only test the excluded instruments.

What if there is more than one endogenous regressor? Could we just do the test on each regressor? Unfortunately, it would be possible for both regressors to pass the test, but still have a weak identification problem. It is not only necessary for the excluded instruments to be relevant to the regressors, but they must do so in a manner that is independent. For example, suppose we have

two endogenous regressors, x_1, x_2, and two instruments z_1, z_2, no controls, no constant, and consider how the endogenous regressors relate to the instruments:

$$x_1 = \lambda_{11} z_1 + \lambda_{12} z_2 + \nu_1$$
$$x_2 = \lambda_{21} z_1 + \lambda_{22} z_2 + \nu_2$$

do not have enough instruments

But now suppose λ_{12} and λ_{22} were zero, so only z_1 helped explain the instruments. Each regressor would pass the test, but the model would be underidentified since only one of the instruments is actually relevant. Or what if:

$$x_1 = z_1 + 3z_2 + \nu_1$$
$$x_2 = 2z_1 + 6z_2 + \nu_2$$

Then we run into a similar problem: if we think of it as 2SLS, then $\hat{x}_1$ and $\hat{x}_2$ will end up becoming perfectly correlated as our sample gets large. Another way the rank assumption would fail is if the instruments were perfectly correlated with each other. Testing for the multiple regressor case can be done with

- Anderson Canonical Correlation Test
- Cragg-Donald test use minimum eigen values
- Kleibergen-Paap test

These aren't covered in this course, but remember the names. The first two rely on the classical error assumptions, whereas the third can use more robust error behaviour.

3.2 The Anderson-Rubin Test

not to test WI; test parameters

The Anderson-Rubin Test is not a test for weak identification. It is a hypothesis test regarding regression parameters that is robust to weak idenfitication.

if there is WI, can we modify model, such that estimator is consistent?

In this course we cover the Anderson-Rubin test for the case where we have one endogenous regressor of interest. Notation: $X_1 = Z_1$ the controls, x_2 the exogenous regressor, Z_2: the excluded instruments. Our reduced form model:

$$y = X_1 \beta_1 + x_2 \beta_2 + \varepsilon$$
$$x_2 = X_1 \Pi_1 + Z_2 \Pi_2 + \nu$$

Remember: reduced form is when we write every endogenous variable (this includes y and any endogenous regressors) on the left hand side of separate regressions. The right hand sides for the endogenous regressors include all the exogenous stuff. It's what we would do in the first stage of 2SLS.

Now suppose I have a null hypothesis: $H_0: \quad \beta_2 = \beta^*$. How would we test this? The standard way would be to obtain $\hat{\beta}_{GMM}$, obtain the standard error for $\hat{\beta}_2$, and then run a t-test (or a Wald test but this should be the same for large samples). However, the asymptotics we derived for $\hat{\beta}_{GMM}$ are all wrong if the rank condition fails!

The Anderson-Rubin test provides a way to test the null hypothesis that is still valid in the presence of weak identification. However, we use a different null:

$$H_0: \quad \beta_2 = \beta^* \quad \text{and} \quad E(Z_i'\varepsilon_i) = 0$$

If we reject the null, it could be because $\hat{\beta}_2 \neq \beta^*$ and/or weak exogeneity fails.

To see how the test works, first subtract $x_2\beta^*$ from either side of the regression function:

$$y - x_2\beta^* = X_1\beta_1 + x_2(\beta_2 - \beta^*) + \varepsilon$$

This procedure may look strange, but we set up the term $\beta_2 - \beta^*$ which should be zero under the null. The left hand side we may denote $\tilde{y}$:

$$\tilde{y} = X_1\beta_1 + x_2(\beta_2 - \beta^*) + \varepsilon$$

Is it OK to regress $\tilde{y}$ on X_1 and x_2? No, because x_2 is endogenous. So instead, substitute the reduced form for x_2:

$$\tilde{y} = X_1\beta_1 + (X_1\Pi_1 + Z_2\Pi_2 + \nu)(\beta_2 - \beta^*) + \varepsilon$$

Now rearrange the terms:

$$\begin{aligned}
\tilde{y} &= X_1\theta_1 + Z_2\theta_2 + \eta, \quad \text{where :} \\
\tilde{y} &= y - x_2\beta^*, \\
\theta_1 &= \beta_1 + \Pi_1(\beta_2 - \beta^*) \\
\theta_2 &= \Pi_2(\beta_2 - \beta^*) \\
\eta &= \nu(\beta_2 - \beta^*) + \varepsilon
\end{aligned}$$

The above regression is "ok", because X_1 and Z_2 are both exogenous. Now let's think about the parameter θ_2. Under the null hypothesis θ_2 should be zero. There are two reasons for this to not be the case: either the true $\beta_2 \neq \beta^*$ or the endogeneity conditions are invalid, which would cause inconsistency in the parameter estimates. Why do we say this is robust to weak identification? If there is weak identification, then $\Pi_2 = 0$, which should make our θ_2 small. This should cause us more often to fail to reject when there is weak identification i.e. we avoid making type 1 errors because of incorrect inference due to weak identification.

To summarise the procedure:

- Construct $\tilde{y} = y - x_2\beta^*$

- Regress $\tilde{y}$ on X_1 and Z_2 to get $\hat{\theta}_1$, $\hat{\theta}_2$
- Test $\theta_2 = 0$
- If fail to reject: $\beta_2 = \beta^*$ seems reasonable and exogeneity assumptions valid
- If reject: either $\beta \neq \beta^*$ or exogeneity assumptions fail

To construct an "A-R confidence interval":

- Construct a grid of potential β_2
- For each gridpoint, test that value for β^*
- Each fail to reject gridpoint (say at the 5 % significance level) is part of the 95% confidence interval
- Each reject gridpoint is not part of the confidence interval

Weak identification should cause the confidence intervals to blow up.

4 Specification Tests

Often when we study statistical models, we make tests about the parameters of the model. We are inherently interested in the relationships between economic variables and how it affects our understanding of the real world, and our theories.

But statistical models themselves have assumptions embedded within them. These are not always testable. But sometimes they are.

4.1 Heteroskedasticity

The homoskedasticity assumption in linear regression models is testable. For the moment, compare $\hat{S}_{Classical}$ to $\hat{S}_{HC}$:

$$\hat{\sigma}^2 \frac{1}{n} \sum_{i=1}^{n} X_i' X_i \quad vs. \quad \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 X_i' X_i$$ variance depends on i

Heteroskedasticity is actually only relevant to our inference if it is correlated with any second order term of X. By this I mean heteroskedasticity could exist in a strange manner that doesn't result in any second order correlation, but we wouldn't really care about it. So a comprehensive way to test for heteroskedasticity is to do the following. Do OLS as normal and collect the error terms $\hat{\varepsilon}$. Now take the squares of the error terms $\hat{\varepsilon}_i^2$ and regress it on all the cross terms. E.g. if we have a constant, an x_1 and an x_2, we would regress:

$$\hat{\varepsilon}^2 = \lambda_0 + \lambda_1 x_1 + \lambda_2 x_2 + \lambda_{11} x_1^2 + \lambda_{22} x_2^2 + \lambda_{12} x_1 x_2 + \nu$$

Care must be taken in setting up this regression so as to not repeat terms. We would then test the null that all the λ's (excluding λ_0) are equal to zero.

The problem with this in practice is that there can be many parameters in this regression $(k(k+1)/2)$.

Some simpler tests which aren't as comprehensive, but nicer for smaller samples:

- Just regress $\hat{\varepsilon}^2$ on the same stuff you regressed y on initially (k) e.g., X
- The same as above, but include $\hat{y}^2$ to capture some second order information more likely to be relevant $(k+1)$ e.g., X + yhat squared
- Same as above, but also include all the square terms (but not cross terms) $(2k)$

4.2 Non-linearities pretend it is linear, then...

The "RESET" is a simple way to detect any obvious non-linearities where we have ran a linear regression model. After initially running the linear regression, take $\hat{y}$ and run the regression:

$$y = X\beta + \hat{y}^2\gamma_2 + \hat{y}^3\gamma_3 + ... + \nu$$

We then test the null that the γ's are equal to zero. If this is rejected, then we reject the linear model.

4.3 Serial Correlation

Breusch-Godfrey test for serial correlation. In a panel-data or time series setting, we can test whether the errors are serially correlated (this creates endogeneity problems for models that have lagged dependent variables). In a times-series setting, take the estimated error terms from the initial estimation and run the regression:

$$\hat{\epsilon}_t = \rho_1\hat{\epsilon}_{t-1} + \rho_2\hat{\epsilon}_{t-2} + ... + \rho_1\hat{\epsilon}_{t-p} + \nu$$

Generally we only test up to a set level p depending on the length of the data. Note there is no constant: we do not need one as they are all mean zero anyway. We test the null that all the ρ's are zero. If we reject, then this is evidence of serial correlation.

4.4 The Hausman Test

The Hausman statistic can be used generally in GMM frameworks to test restrictions in an overidentified model. Often we use the Hausman stat in panel data settings to test whether the Random Effects assumption is true. Remember, the Random Effects is more restricted than Fixed Effects because we are adding the assumption that the entity effects are uncorrelated with the regressors.

$$H = n(\hat{\beta}_{FE} - \hat{\beta}_{RE})'(V(\hat{\beta}_{FE}) - V(\hat{\beta}_{RE}))^{-1}(\hat{\beta}_{FE} - \hat{\beta}_{RE})$$

H has an asymptotic chi-squared distribution. The number of degrees of freedom is the number of additional assumptions being imposed. Here there are $k-1$ (each regressor is being assumed to be exogenous, although we wouldn't count the constant). A large H rejects the random effects moment restrictions i.e. rejects that the entity effects are uncorrelated with the regressors. In practice, we use classical estimates of V to ensure that the difference in the variance is positive definite in small samples. This means the test is not robust to failures of this assumption however.

4.5 The J-statistic

In overidentified models, the J stat acts as a measure of how close we are to getting the sample moment conditions to line up to zero. A large J stat suggests something has gone wrong. Under the null that all the population moment conditions are valid, the J stat will have an asympototic Chi-squared distribution. The number of degrees of freedom is $l-k$. Rejecting the null because the J is generally taken to mean that at least one of our restrictions is wrong. This can occur because we have treated a regressor incorrectly by denoting it the wrong type:

- Endogenous regressor
- Exogenous regressor (or control, or included instrument)
- Instrumental variable (or excluded instrument)

5 Non-linear Estimators and Maximum Likelihood

Our linear regression and its extensions to GMM and further applications in panel and time series settings all have analytical solutions to the parameters e.g.

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y$$

There are many ways a regression function could be non-linear e.g.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Here we have a quadratic term in x, so the function is non-linear in x. However, from the econometricians point of view, this is still within the realms of linear regression: we just treat x^2 as a separate regressor from x. What really matters is whether the function is *linear in the parameters*. For example, take:

$$y = \beta_0 + \beta_1 x + \beta_2 x^{\beta_3} + \varepsilon$$

Now this is a different class of problem from OLS due to the fact that β_3 enters the function non-linearly. If, like before, we choose our parameters to minimise

the sum of squares, this is known as "non-linear least squares". We generally won't have analytical solutions for such problems, but instead characterise the solution, e.g.

$$\hat{\beta} = \underset{\beta_0}{argmin} \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

More generally:

$$\hat{\beta} = \underset{\beta_0}{argmin} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(X_i, \beta_0))^2,$$

start from a point may get wrong estimators

where $f(X_i, \beta)$ is non-linear in β.

5.1 Non-Linear GMM

Similarly to non-linear least squares, we also have the analogous non-linear GMM. This occurs when our moment conditions are non-linear, meaning we characterise $\hat{\beta}$ by:

$$\hat{\beta} = \underset{\beta_0}{argmin} \quad n\bar{g}_n(\beta_0)' W \bar{g}_n(\beta_0),$$

where $\bar{g}_n(\beta_0)$ is non linear in β_0. As you can see, like linear GMM, we still choose the estimator to minimise the J statistic. How does the non-linear GMM estimator behave? Let's take a fixed W. It is first useful to think about the "true β. It would be characterised in the following manner:

need WLLN to converge expectation, but expectation may not exist, heavy tail

$$\beta = \underset{\beta_0}{argmin} \quad nE[g(\beta_0)]' W E[g(\beta_0)]$$

i.e. we can think of the true β as being the value that minimises our J statistic if we could observe the whole population (hence the expectations rather than sample averages). To get consistency, we need the objective function to be well behaved (the regularity conditions are a little complicated) basically the above equation has to have a unique solution for β_0, and $\bar{g}_n(\beta_0)$ must be guaranteed to converge to $E[g(\beta_0)]$. For both our asymptotic inference to work, we also need the objective to be twice differentiable at the true value. Remember from before:

Non-efficient Linear GMM:

$$AVar(\hat{\beta}_{GMM}) = (\Sigma_{ZX}' W \Sigma_{ZX})^{-1} \Sigma_{ZX}' W S^{-1} W \Sigma_{ZX} (\Sigma_{ZX}' W \Sigma_{ZX})^{-1}$$

Efficient Linear GMM:

$$AVar(\hat{\beta}_{GMM}) = (\Sigma_{ZX}' W \Sigma_{ZX})^{-1}$$

Our asymptotic variance has analogoues for the non-linear case. Note that in the linear case, $E[g_n(\beta)] = E[Z_i'(y_i - X_i\beta)]$. We can see that:

$$E[\frac{\partial}{\partial \beta} g_n(\beta)] = -\Sigma_{ZX}$$

and by definition:

$$S = Var(g_n(\beta))$$

This means (not going to show the derivation, this is a shortcut) we can get our analogoues for the non-linear case by replacing Σ_{ZX} with $E[-\frac{\partial}{\partial\beta}g_n(\beta)]$ and be careful to define S properly.

Like linear GMM, we can also do a two step process, by which we estimate using an arbitrary positive definite W (there may be an analogy to 2SLS available to use). After the first step we estimate:

$$\hat{S} = AVar(g_n(\hat{\beta}))$$

Just like before, $\hat{S}$ may be modified depending on our assumptions behind the data. Then we use $\hat{W} = \hat{S}^{-1}$, plug back in and re-estimate $\hat{\beta}$.

if y is not continuous?

5.2 Maximum Likelihood

The most common application of non-linear estimators is to maximum likelihood problems. Sometimes our dependent variable doesn't follow a continuous structure, so OLS seems a bit inappropriate, because the "distance" from $\hat{y}$ to y isn't meaningful. Data can take on lots of different types. Examples are

- Binary outcomes e.g. y describes whether or not an individual is in the labour force or not. For this we typically use probit or logit regression
- Categorical outcomes e.g. y describes the religion of an individual. For this we typically use multinomial logistic regression or a variant such as hierarchical models
- Ordered outcome models e.g. y describes whether a country is non-democratic, mixed, or democratic. Interesting because we can put the categories in an order, but don't want to impose a cardinal value. Typically use ordinal regression
- Count data e.g. y describes how many bedrooms a house has. Typically use poisson or negative binomial regression
- Censored data; the dependent variable is continuous in some regions but clustered on an edge. E.g. y is weekly expenditure on tobacco; continuous for those who smoke but lots of zero values present. Typically use a Tobit regression

Whatever the model, the goal of the estimator is to maximise the likelihood function

$$L(\theta) = \prod_{i=1}^{n} L_i(\theta)$$

The product comes from the fact that we think of the i outcomes as independent. What $L_i(\theta)$ represents depends on the context; if y_i is a discrete outcome then

it is simply the probability mass. But if it is continuous, we use the probability density. Most of the time we are really doing *conditional* maximum likelihood, because we model the probability of y_i conditional on x, but make no model for the probability of x. For example in a probit model:

$$L_i(X, y, \theta) = \Phi(X_i\theta) \quad \text{if} \quad y_i = 1,$$
$$L_i(X, y, \theta) = 1 - \Phi(X_i\theta) \quad \text{if} \quad y_i = 0,$$

normal CDF
(cumulative density function)

Note I wrote $L_i(X, y, \theta)$, not $L(X_i, y_i, \theta)$. If dealing with non-iid data, we may want to explicitly model the L's as having some sort of dependency on data from multiple time periods, yet still treat the L's themselves as being independent (especially in a time series setting). We can also throw Z's into the mix if we want to be explicit about some sort of exogeneity idea, doesn't really matter.

So how does the generic maximum likelihood estimation work? Given the $L_i(\theta)$ that we have, we can characterise the estimater as follows:

$$\hat{\theta} = \underset{\theta}{argmax} \prod_{i=1}^{n} L_i(\theta).$$

It's much nicer to write this as a sum, not a product. To do so, introduce $l_i(\theta) = logL_i(\theta)$, and note from the rule of logs:

$$log(\prod_{i=1}^{n} L_i(\theta)) = \sum_{i=1}^{n} l_i(\theta)$$

Furthermore, since logs are an increasing function, it won't affect the maximisation procedure. Let's also chuck in a $\frac{1}{n}$ for good measure:

$$\hat{\theta} = \underset{\theta}{argmax} \frac{1}{n} \sum_{i=1}^{n} l_i(\theta).$$

Similar to non-linear GMM, we can characterise the "true" θ_0:

$$\theta_0 = \underset{\theta}{argmax} E[l_i(\theta)].$$

To get consistency and our asymptotic normality, we again need regularity assumptions: there must be a unique solution to the above, $\frac{1}{n}\sum_{i=1}^{n} l_i(\theta)$ must converge to $E[l_i(\theta)]$ and $l(\theta)$ must be twice differentiable at $\theta = \theta_0$. What about the asymptotic variance of the MLE estimator? To think about this, try

to see the analogues when jumping from OLS (iid) to MLE.

OLS: heteroskedasticity robust standard errors:

$$AVar(\hat{\beta}) = S_{XX}^{-1}\hat{S}S_{XX}^{-1},$$
$$\text{where} \quad \hat{S} = AVar(\frac{1}{n}\sum_{i=1}^{n} X_i'\epsilon_i)$$

OLS: classical errors, simplifies to:

$$AVar(\hat{\beta}) = \sigma^2\Sigma_{XX}^{-1}$$

In OLS, we are trying to minimise

$$\text{penalty} = \frac{1}{n}\sum_{i=1}^{n}(y_i - X_i\beta)^2,$$

whereas in MLE, we are trying to *minimise*:

$$\text{penalty} = -\frac{1}{n}\sum_{i=1}^{n} l_i(\theta).$$

(note the negative to turn it into a minimisation). Let's consider the first and second derivates of the penalty function in OLS:

$$\frac{\partial}{\beta}(\text{penalty}) = -2\frac{1}{n}\sum_{i=1}^{n} X_i'(y_i - X_i\beta)$$
$$\frac{\partial^2}{\partial\beta\partial\beta'}(\text{penalty}) = 2\frac{1}{n}\sum_{i=1}^{n} X_i'X_i$$

Let's rewrite these slightly:

$$-\frac{1}{2}\frac{\partial}{\beta}(\text{penalty}) = \frac{1}{n}\sum_{i=1}^{n} X_i'\epsilon_i$$
$$\frac{1}{2}\frac{\partial^2}{\partial\beta\partial\beta'}(\text{penalty}) = \frac{1}{n}\sum_{i=1}^{n} X_i'X_i$$

Finally, the trick: take AVar on both sides to get the definition of S and use definition of S_{XX}

$$AVar(-\frac{1}{2}\frac{\partial}{\beta}(\text{penalty})) = \hat{S}$$
$$\frac{1}{2}\frac{\partial^2}{\partial\beta\partial\beta'}(\text{penalty}) = S_{XX}$$

We now have a way to analogise the AVar of OLS to the AVar of MLE. Instead of penalty, write the negative of the log-likelihood function, and do some manipulation (using $\sim$ to mean similar):

$$\begin{aligned}\frac{1}{4}AVar(\frac{\partial}{\theta}(l(\theta))) &\sim \hat{S} \\ -\frac{1}{2}\frac{\partial^2}{\theta\theta'}(l(\theta)) &\sim S_{XX}\end{aligned}$$

We have special names for the expressions that arise here: the score:

$$\begin{aligned}s_i(\theta) &= \frac{\partial}{\partial\theta}l_i(\theta) \\ \bar{s}_n(\theta) &= \frac{1}{n}s_i(\theta)\end{aligned}$$

and the Hessian:

$$\begin{aligned}H_i(\theta) &= \frac{\partial^2}{\partial\theta\partial\theta'}l_i(\theta) \\ \bar{H}_n(\theta) &= \frac{1}{n}s_i(\theta)\end{aligned}$$

This means, returning to the earlier equation, our analogies are:

$$\begin{aligned}\frac{1}{4}AVar(\bar{s}_n(\theta)) &\sim \hat{S} \\ -\frac{1}{2}\bar{H}_n &\sim S_{XX}\end{aligned}$$

Let's plug this into the general AVar formula for OLS:

$$AVar(\hat{\theta}) = (-\frac{1}{2}\bar{H}_n)^{-1}(\frac{1}{4}AVar(\bar{s}_n(\theta)))(-\frac{1}{2}\bar{H}_n)^{-1}$$

Simplifying:

$$AVar(\hat{\theta}) = \bar{H}_n^{-1}AVar(\bar{s}_n(\theta))\bar{H}_n^{-1}$$

Great, now we have a formula for the asymptotic variance of a MLE. But we can go further, the same way the OLS AVar simplifies under classical errors. In the special case that the model is *correctly* specified i.e. the likelihood function reflects the true likelihood of the data, then $AVar(\bar{s}_n(\theta)) = -\bar{H}_n$. In this case, the asymptotic variance simplifies to:

$$AVar(\hat{\theta}) = -\bar{H}_n,$$

Note that we also talk about the "information matrix" which is just the population analogue to $\bar{H}_n$ (but negative).

It seems like we have two different formulas for AVar. This is the same way we have multiple formulas for AVar in OLS depending on our assumptions. In MLE, if we are confident the model is correctly specified, then the latter is fine. But generally, the former is more robust. Sometimes, bootstrapping is a useful way of checking for unusual behaviour.

So to recap, to get AVar:

- Obtain the MLE estimate $\hat{\theta}$ based on numerically maximising the log-likelihood function $\frac{1}{n}\sum_{i=1}^{n} l_i(\theta)$
- Calculate the AVar of the score: $s(\theta) = \frac{1}{n}\sum_{i=1}^{n}[\frac{\partial}{\partial\theta}l_i(\theta)][\frac{\partial}{\partial\theta}l_i(\theta)]'$
- Calculate the Hessian: $H(\theta) = \frac{1}{n}\frac{\partial^2}{\partial\theta\partial\theta'}l_i(\theta)$
- Use the robust AVar: $H^{-1}AVar(s)H^{-1}$
- Or use the correctly specified AVar: $-H^{-1}$