ECON 6200

Econometrics II, by Jörg Stoye

Gabe Sekeres

Spring 2025

Contents

Introduction

This course builds directly on ECON 6190, the material will not be revised. If you need to catch up on the material, especially probability and statistics, see the Hansen and Durrett textbooks. If there's one econometrics book to actually own, it's the Hansen. These notes will often reference results in either Hansen or Hayashi.

See the syllabus for how we will be assessed – homework is 30%, prelim is 30%, and the final is 40%. For homework, you are invited (and *highly* recommended) to form study groups! It's completely okay to work on the homeworks in study groups, but you will submit individual write-ups.

Roughly, in the first few weeks, we will start with the linear model in some detail. We will then generalize to IV, TSLS, and go rapidly to the Generalized Method of Moments (GMM) and extremum (or m-) estimation. We will also cover some nonparametrics as well as bootstrap.

You can think of us starting in the most specific case and moving outwards. OLS is a special case of IV, which is a special case of TSLS, which is a special case of GMM, which is a special case of extremum estimation. We will also think about Maximum Likelihood (ML) as a special case of extremum estimation, and panel data as a special case of GMM. Studying this tree will cover half of the course, and then we will have some time to cover nonparametric methods like kernel density and kernel mean, as well as bootstrapping. This may seem like we are doing some old school stuff, since the other ML (machine learning) has overtaken these methods, but theoretically they are very similar.

1 The Linear Model

We have, of course, already encountered the
Definition. *Ordinary Least Squares (OLS)* estimator:

$$\begin{aligned}
\hat{\beta} &= \left(X'X\right)^{-1} X'Y && \text{Data Matrix} \\
&= \left(\mathbb{E}_n XX'\right)^{-1} \mathbb{E}_n XY && \text{Sample Expectation} \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right) \frac{1}{n}\sum_{i=1}^{n} X_i Y_i && \text{Sample Average} \\
&= \left(\sum_{i=1}^{n} X_i X_i'\right) \sum_{i=1}^{n} X_i Y_i && \text{Clean Sample Average}
\end{aligned}$$

All of these notations are equivalent, but helpful in different concepts. Data matrix notation is extremely useful for proving results. The sample expectation is useful because it reminds us of the empirical context – this is a shorthand for sample average, which is written in two ways depending on context – essentially, when we care about the asymptotic behavior in different ways.

The interpretation of $\hat{\beta}$ depends on context:

1. In any given sample, it just projects Y onto X
2. Under weak assumptions, it converges to the population analog $\beta^\star \equiv (\mathbb{E}XX')^{-1}\mathbb{E}XY$, which is the population projection coefficient and characterizes the *best linear predictor under square loss*
3. Under stronger assumptions, it estimates a causal effect of X on Y.

We will elaborate these in order, and develop the classic theory of Least Squares estimation.

Recall that $\hat{\beta}$ can be derived as the minimization (in b) of

$$\sum_{i=1}^{n}(Y_i - X_i'b)^2 = (Y - Xb)'(Y - Xb)$$

which is of course where the name comes from. However, recall that also this minimization defined b such that Xb is the point in the span of X that is *closest to* Y in Euclidean distance. Basically, we projected Y onto X.

We can see this with the illustration in Figure 1, which uses demeaned vectors, and defines the projection $\hat{Y} \equiv X\beta = \beta_1 X_1 + \beta_2 X_2$ and the residual $\hat{\varepsilon} \equiv Y - \hat{Y}$.

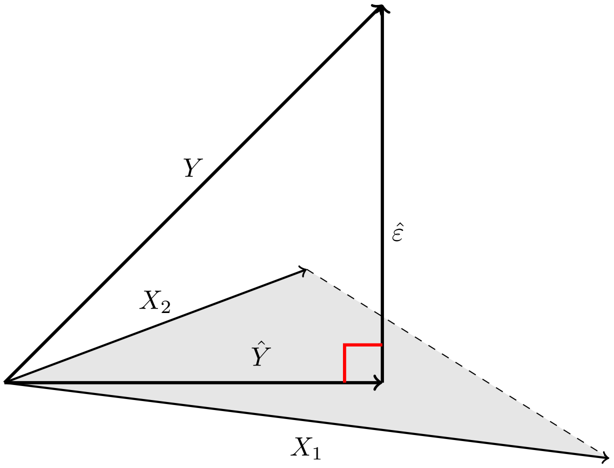

Figure 1: Illustration of OLS as Projection

Corollary 1.1. Sum of Squares Decomposition *It immediately follows from Pythagoras that*

$$Y'Y = \hat{Y}'\hat{Y} + \hat{\varepsilon}'\hat{\varepsilon}$$

Equivalently, $SST = SSE + SSR$ *or*

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} \hat{Y}_i^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

It follows immediately that $R^2 = SSE \,/\, SST \in [0,1]$. *The extreme values of* R^2 *correspond to* $\hat{Y} = 0$ *and* $\hat{\varepsilon} = 0$ *respectively. It also follows that* $\hat{\varepsilon}$ *is* by construction *orthogonal to* $\hat{Y}$.

We will recap this and basic related facts in the first homework.
Question. What happens with collinear X?

With collinear X, the set on which we project is of lower dimension. The projection $\hat{Y}$ is still unique, but the projection coefficient is not.
Definition. The *projection coefficient* $\hat{\beta}$ is defined as

$$\hat{\beta} \equiv \underset{b}{\operatorname{argmin}} \sum_{i}(Y_i - X_i'b)^2 = \underset{b}{\operatorname{argmin}}(Y - Xb)'(Y - Xb)$$

and can be characterized by the first order condition

$$\frac{\partial}{\partial b}(Y - Xb)'(Y - Xb) = -2X'Y + 2X'Xb \overset{!}{=} 0$$
$$\implies \hat{\beta} = (X'X)^{-1}X'Y$$

The fitted values and residuals equal

$$\hat{Y} = X\hat{\beta} = \underbrace{X(X'X)^{-1}X'}_{P_X\text{, the projection matrix}} Y = P_X Y$$
$$\hat{\varepsilon} = Y - \hat{Y} = (\underbrace{I_n - X(X'X)^{-1}X'}_{\text{annihilator matrix}})Y$$

Example. Frisch-Waugh(-Lovell) The projection of Y on X can be decomposed, giving rise to some important results. To fix ideas, consider projecting Y:

- on the scalar X_1 (plus a constant), getting slope coefficient $\tilde{\beta}_1$, versus
- on the scalars (X_1, X_2) (plus a constant), getting slope coefficients $(\hat{\beta}_1, \hat{\beta}_2)$.

Can we interestingly relate $\tilde{\beta}_1$ to $\hat{\beta}_1$? Yes! To do so, consider projecting X_1 on X_2, getting slope coefficient $\hat{\gamma}$. To simplify, assume all variables are demeaned. Across regressions, this leads to the first order conditions:

$$0 = \mathbb{E}_n\left(X_1\left(Y - \hat{\beta}_1X_1 - \hat{\beta}_2X_2\right)\right)$$
$$0 = \mathbb{E}_n\left(X_2\left(Y - \hat{\beta}_1X_1 - \hat{\beta}_2X_2\right)\right)$$
$$0 = \mathbb{E}_n\left(X_2\left(X_1 - \hat{\gamma}X_2\right)\right)$$

We can combine the first two, and use the third to get

$$\begin{aligned}0 &= \mathbb{E}_n(X_1 - \hat{\gamma}X_2)\left(Y - \hat{\beta}_1X_1 - \hat{\beta}_2X_2\right)\\ &= \mathbb{E}_n(X_1 - \hat{\gamma}X_2)\left(Y - \hat{\beta}_1X_1 - \hat{\beta}_1\hat{\gamma}X_2\right)\\ &= \mathbb{E}_n(X_1 - \hat{\gamma}X_2)\left(Y - \hat{\beta}_1(X_1 - \hat{\gamma}X_2)\right)\end{aligned}$$

The last line is the first order condition from regressing Y on the residuals of the regression of X_1 on X_2. Thus, $\hat{\beta}_1$ is the slope coefficient from that regression. We can extend this argument to show that Y can be residualized as well.

Remark. If we think about this as if X_1 is schooling and X_2 is family income, we are essentially including only the uncorrelated parts of those variables through this method, and don't need to worry about the correlation between the two.

When thinking about *Omitted Variable Bias*, we can similarly characterize the projection of X_2 on X_1 by

$$0 = \mathbb{E}_n(X_1(X_2 - \tilde{\gamma}X_1)) = \mathbb{E}_n\left(X_1\left(\hat{\beta}_2X_2 - \hat{\beta}_2\tilde{\gamma}X_1\right)\right)$$

We can then substitute the first order condition from above, and get

$$0 = \mathbb{E}_n\left(X_1\left(Y - \hat{\beta}_1X_1 - \hat{\beta}_2X_2\right)\right) = \mathbb{E}_n\left(X_1\left(Y - \hat{\beta}_1X_1 - \hat{\beta}_2\tilde{\gamma}X_1\right)\right)$$

So we have the first order condition from regression Y on X_1 only. We can conclude that the slope coefficient from that regression is

$$\tilde{\beta} = \hat{\beta}_1 + \tilde{\gamma}\hat{\beta}_2$$

If there is a *causal interpretation* of the projection of Y onto (X_1, X_2), then the difference term $\tilde{\gamma}\hat{\beta}_2$ is (the sample analog of) the *omitted variable bias* incurred by omitting X_2.

Remark. We haven't even really introduced the idea of random variables and expectation yet. The word *bias* here is really loose, and doesn't make sense in the world of projection. However, in the real economic word, it makes a lot of sense in most contexts – in that case, you are saying that the omitted variables systematically bias the estimator in some direction.

Example. Frisch-Waugh(-Lovell) General Statement We can now do the same thing, more generally. Partition as follows:

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \quad ; \quad \mathbb{E}XX' \equiv Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} = \begin{pmatrix} \mathbb{E}X_1X_1' & \mathbb{E}X_1X_2' \\ \mathbb{E}X_2X_1' & \mathbb{E}X_2X_2' \end{pmatrix}$$
$$\mathbb{E}XY \equiv Q_{XY} = \begin{pmatrix} Q_{1Y} \\ Q_{2Y} \end{pmatrix} = \begin{pmatrix} \mathbb{E}X_1Y \\ \mathbb{E}X_2Y \end{pmatrix}$$

and use notation $\hat{Q}_{11}$ (etc) for sample analogs. Then, it can be shown that

$$\hat{\beta} = \begin{pmatrix} \left(\hat{Q}_{11} - \hat{Q}_{12}\hat{Q}_{22}^{-1}\hat{Q}_{21}\right)^{-1}\left(\hat{Q}_{1Y} - \hat{Q}_{12}\hat{Q}_{22}^{-1}\hat{Q}_{2Y}\right) \\ \left(\hat{Q}_{22} - \hat{Q}_{21}\hat{Q}_{11}^{-1}\hat{Q}_{12}\right)^{-1}\left(\hat{Q}_{2Y} - \hat{Q}_{21}\hat{Q}_{11}^{-1}\hat{Q}_{1Y}\right) \end{pmatrix}$$

Regressing X_1 on X_2 would yield coefficients $\hat{\gamma} \equiv \hat{Q}_{22}^{-1}\hat{Q}_{21}$ and residuals $\hat{\eta} \equiv X_1 - X_2\hat{Q}_{22}^{-1}\hat{Q}_{21}$. Hence, we have that

$$\begin{aligned} \mathbb{E}_n\hat{\eta}^2 &= \mathbb{E}_nX_1^2 + \hat{Q}_{12}\hat{Q}_{22}^{-1}\mathbb{E}_nX_2^2\hat{Q}_{22}^{-1}\hat{Q}_{21} - 2\mathbb{E}_nX_1X_2\hat{Q}_{22}^{-1}\hat{Q}_{21} \\ &= \hat{Q}_{11} - \hat{Q}_{12}\hat{Q}_{22}^{-1}\hat{Q}_{21} \end{aligned}$$

By similar algebra, $\mathbb{E}_n\hat{\eta}Y = \hat{Q}_{1Y} - \hat{Q}_{12}\hat{Q}_{22}^{-1}\hat{Q}_{2Y}$. It follows that if we projected Y on the residual from regressing X_1 on X_2, we would get coefficient $\tilde{\beta}$, where

$$\tilde{\beta}_1 = \left(\hat{Q}_{11} - \hat{Q}_{12}\hat{Q}_{22}^{-1}\hat{Q}_{21}\right)^{-1}\left(\hat{Q}_{1Y} - \hat{Q}_{12}\hat{Q}_{22}^{-1}\hat{Q}_{2Y}\right) = \hat{\beta}_1$$

Remark. Verbally, the multivariate OLS coefficient on X_1 is the coefficient one would get by regressing Y on the residual from regressing X_1 on all other covariates. We can show that this statement remains true if we replace Y with the residual $Y - \mathcal{P}_{X_2}(Y)$, where $\mathcal{P}_{X_2}$ is the population projection onto X_2. This may look like a curiosity now, but is an important starting point for partially linear models and other such things. An immediate payoff is that if you recall it, you'll never forget the own-variance formula for multivariate regression coefficients.

There are two ways to interpret OLS. Neither of them are uniquely 'right', but you should always be clear about which one you are appealing to. The *Best Linear Prediction* is an interpretation that makes sense under extremely general conditions. It's the notion of a linear model that is generalized in most predictive (notably, data science / statistical learning) applications. The *Causal (or Structural) Linear Model* is more demanding, but allows for causal interpretation. It's the notion of linear model that is generalized in most causal (for example, Instrumental Variables) applications. Note that this does not preclude predictive application of linear models as a component of causal inference. A salient example is the "first stage" in IV regression.

Best Linear Prediction. Write $Y = m(X) + \varepsilon$, where $m(x) \equiv \mathbb{E}(Y \mid X = x)$. That $\mathbb{E}(\varepsilon \mid X) = 0$ is now a tautology. We can show that $b^{\star} \equiv (\mathbb{E}XX')^{-1}\mathbb{E}(XY)$, if it exists, minimizes $\mathbb{E}(Y - X'b)^2$. That is, $\hat{Y} \equiv \mathcal{P}_X Y \equiv X'b^\star$ is the *best linear predictor under square loss*. Furthermore, under these conditions, $\mathbb{E}Xe = 0$, where $e \equiv Y - \mathcal{P}_X(Y)$, meaning that the projection error e is not correlated with X.

Theorem 1.1. *If a weak law of large numbers applies to both $\frac{1}{n}\sum_{i=1}^n X_i X_i'$ and $\frac{1}{n}\sum_{i=1}^n X_i Y_i$ and $\mathbb{E}XX'$ is nonsingular, then $b^\star$ is uniquely defined and $\hat{\beta} \xrightarrow{p} b^\star$.*

Remark. The best linear predictor $\hat{Y}$ is uniquely defined even if $\mathbb{E}XX'$ is singular. In that case, $b^\star$ is not unique. This is really important for models with lots of covariates – think in statistical learning / machine learning.

The (Causal/Structural) Linear Model Write $Y = X'\beta + \varepsilon$, where $\mathbb{E}(\varepsilon \mid X) = 0$. Equivalently, $m(x) = \mathbb{E}(Y \mid X = x) = x'\beta$. In this version, $\mathbb{E}(\varepsilon \mid X) = 0$ is *not* tautological! Our assumptions become much stronger. To be precise, we are assuming that (i) ε is mean-independent of X, and (ii) the mean of ε is zero. These are new assumptions! We pay a large cost, but there are also large benefits. This model allows for *causal interpretation* of the estimand: in expectation, a change ΔX causes a corresponding change $\Delta X'\beta$ in Y. Importantly, this difference isn't just about interpretation – some important results are only available under the stronger assumptions.

Remark. Remember that within limits, a linear model can capture nonlinearities. Some examples:

1. Polynomial expansion:
$$Y = \beta_0 + \beta_1 H + \beta_2 H^2 + \cdots + \varepsilon$$

2. Log or log-log regression:
$$\ln Y = \ln A + \alpha \ln K + (1-\alpha)\ln L + \cdots + \varepsilon$$
(Note! This changes the necessary assumption on ε, as in the primal it is now multiplied by the covariates)

3. Treatment effects with interactions:
$$Y = \beta + \delta \cdot \text{treatment} + \gamma \cdot \text{female} \cdot \text{treatment} + \cdots + \varepsilon$$

Indeed, many "big data" models are high-dimensional but linear! Think of basically any machine learning context.

OLS as a Random Variable. The central questions are: What can we say about the estimator as a random variable? Are there conditions under which it has desirable properties, notably if our objective is to learn about β (or possibly the population $\mathcal{P}_X(Y)$)?

Consider drawing n samples and taking β_i, for $i \in \{1, \ldots, n\}$ to be a random variable. If we show the histogram (in the slides) of these $\{\beta_i\}$, we can see that as n increases they tend to look normal! However, this is not necessarily a central limit theorem. Actually, the distribution of the estimators will approach the distribution of the error terms as n increases, under weaker assumptions. However, we assume normally distributed errors a lot.

Remark. We had a small aside to wonder whether we think of the vertical squared distance between each point and the line, or the shortest distance between the point and the line squared. The second is actually different from OLS – it's precisely *principal component analysis*!

Remark. What if we minimized the horizontal distance? Then we would get the projection of X onto Y rather than *vice versa* – called *reverse regression*. This is expanded on a lot in the first homework – basically, the coefficients are normalized to the variances of X and Y respectively. They are the same if and only if $\text{Var}(X) = \text{Var}(Y)$.

Remark. What if we minimize absolute values instead of squares? That would be *median regression* – at a population level, the median will solve this problem. By using other loss functions, we could tease out the other quantiles, and by using all of them we would get *quantile regression*, where we treat the quantiles as being heterogeneously treated.

For finite sample theory, we will make the following assumptions:
Assumption 1.1. *In data matrix notation, we have that*

1. $Y = X\beta + \varepsilon$ 'linearity'
2. $\mathbb{E}(\varepsilon \mid X) = 0$ 'strong exogeneity'
3. $\operatorname{rank}(X) = K$ *a.s., where* $X \in \mathbb{R}^{n\times K}$ 'rank condition' *(equivalent:* $X'X$ *nonsingular a.s.)*
4. $\mathbb{E}(\varepsilon\varepsilon' \mid X) = \sigma^2 I_n$ 'spherical error'

The first two assumptions together imply a causal linear model. Assumptions that are natural when considering OLS as the 'best linear predictor' do not suffice to attain unbiasedness of $\hat{\beta}$. These further imply that ε is zero in expectation conditional on *all* covariates – including past and future realizations. That's quite strong, but matters a lot more in time series econometrics. If we assume the data are i.i.d., this is not stronger than the earlier assumption that they are independent vector-wise.

The third assumption is an identification condition, and will fail if any covariates are linearly dependent on each other. We already talked about this previously, but note that in finite samples we are immediately excluding discrete covariates! Though the probability may be small, it doesn't meet the criterion for almost surely. We may describe $\hat{\beta}$ as 'conditionally unbiased', where we are conditioning on X.

Assumption four combines conditional uncorrelatedness and homoskedasticity of errors. The latter makes sense only in the causal model because, even if the true regression error $\varepsilon = Y - m(X)$ is homoskedastic, the projection error

$$e = Y - \mathcal{P}_X(Y) = Y - m(X) + m(X) - \mathcal{P}_X(Y) = \varepsilon + m(X) - \mathcal{P}_X(Y)$$

is not.

We have a hidden assumption that $\mathbb{E}\|X\|^2 < \infty$. This is an existence result, and not so strong generally – if we assume X is non-stochastic, nothing here relies on it. It is an assumption, however.

These assumptions give us the following theorems:
Theorem 1.2. Finite Sample Bias and Variance *Under Assumptions 1.1, we have that*

1. $\mathbb{E}\left(\hat{\beta} \mid X\right) = \beta$
2. $\operatorname{Var}\left(\hat{\beta} \mid X\right) = \sigma^2(X'X)^{-1}$

i.e. $\hat{\beta}$ *is unbiased and its variance is determined.*

Proof. We first observe that

$$\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + \varepsilon) = \beta + \underbrace{(X'X)^{-1}X'\varepsilon}_{=\text{ estimation error}}$$

Therefore, the first two claims are that

$$\mathbb{E}\left(\left(X'X\right)^{-1}X'\varepsilon \mid X\right) = 0$$
$$\operatorname{Var}\left(\left(X'X\right)^{-1}X'\varepsilon \mid X\right) = \sigma^2(X'X)^{-1}$$

We have directly that

$$\mathbb{E}\left((X'X)^{-1}X'\varepsilon \mid X\right) = (X'X)^{-1}X'\underbrace{\mathbb{E}(\varepsilon \mid X)}_{=0} = 0$$

$$\mathrm{Var}\left((X'X)^{-1}X'\varepsilon \mid X\right) = (X'X)^{-1}X'\sigma^2 I_n X'(X'X)^{-1} = \sigma^2(X'X)^{-1}$$

□

Theorem 1.3. Gauss-Markov Theorem *Under Assumptions 1.1, if an estimator $\tilde{\beta}$ is linear (in Y) and unbiased, then*

$$\mathrm{Var}(\tilde{\beta} \mid X) \geq \sigma^2(X'X)^{-1}$$

Proof. We assume that $\tilde{\beta} = CY$ for some C (that may depend on X). Define $D \equiv C - (X'X)^{-1}X'$, and we have that

$$\begin{aligned}
\beta &= \mathbb{E}\left(((X'X)^{-1}X' + D)Y \mid X\right) && \text{(Unbiased)}\\
&= \beta + \mathbb{E}\left(DY \mid X\right)\\
&= \beta + \mathbb{E}\left(D(X\beta + \varepsilon) \mid X\right)\\
&= \beta + \mathbb{E}\left(DX\beta \mid X\right) + D\underbrace{\mathbb{E}(\varepsilon \mid X)}_{=0}\\
&\implies \mathbb{E}(DX\beta \mid X) = 0
\end{aligned}$$

This result is only possible if conditional on X the expression $DX\beta$ is non-stochastic – we have that $DX\beta = 0$ for any β! This holds only if $DX = 0$. So finally, we have

$$\begin{aligned}
\mathrm{Var}(\tilde{\beta} \mid X) &= \mathrm{Var}\left(((X'X)^{-1}X' + D)Y \mid X\right)\\
&= \mathrm{Var}\left(((X'X)^{-1}X' + D)(X\beta_0 + \varepsilon) \mid X\right)\\
&= \mathrm{Var}\left(((X'X)^{-1}X' + D)\varepsilon \mid X\right)\\
&= \sigma^2((X'X)^{-1}X' + D)((X'X)^{-1}X' + D)'\\
&= \sigma^2\left((X'X)^{-1}X'X(X'X)^{-1} + DX(X'X)^{-1} + (X'X)^{-1}X'D' + DD'\right)\\
&\geq \sigma^2(X'X)^{-1} = \mathrm{Var}(\hat{\beta})
\end{aligned}$$

where the conclusion follows from the fact that DD' is positive semi-definite. □

Question. Is homoskedasticity necessary for this result?

Answer. Yes! Consider the case where $\mathbb{E}\varepsilon\varepsilon' = \Omega$ is known and diagonal but its diagonal entries are not the same. Then the Gauss-Markov assumptions apply to the transformed model

$$\Omega^{-\frac{1}{2}}Y = \Omega^{-\frac{1}{2}}X + \Omega^{-\frac{1}{2}}\varepsilon$$

so the estimator

$$\hat{\beta}_{WLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$$

is the best linear unbiased estimator. However, $\hat{\beta}_{WLS} = \hat{\beta}$ if and only if $\Omega = \sigma^2 I_n$ for some σ^2. This is called the *Weighted Least Squares* estimator, which in this case would perform better than OLS.

Some important closed-form expressions Consider a simple linear regression $Y = \alpha + \beta X + \varepsilon$. Then we have that:

$$\hat{\beta} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})X_i}$$

$$\operatorname{Var}\left(\hat{\beta} \mid X\right) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

where σ^2 is the variance of ε. In a multivariate regression, the variance of the kth component of $\hat{\beta}$ is

$$\operatorname{Var}\left(\hat{\beta}_k \mid X\right) = \frac{\sigma^2}{(1 - R_k^2) \sum_{i=1}^n (X_{ki} - \bar{X}_k)^2}$$

where R_k^2 is the R^2 of the regression of X_k on the other covariates. Why is this obvious? Frisch-Waugh-Lovell, of course! The factor $1/(1 - R_k^2)$ is sometimes called the *variance inflation factor (VIF)*.
Remark. Know the expression for the variance of the kth component of $\hat{\beta}$ by heart! You should internalize it from Frisch-Waugh-Lovell.

We also have sample analogs and estimators for σ^2: The *sample analog* (which is a *method of moments estimator*) of σ^2 is

$$\hat{\sigma}^2 \equiv \frac{1}{n} \hat{\varepsilon}' \hat{\varepsilon} = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2$$

However, we can also show (exercise?) that $\mathbb{E}(\hat{\sigma}^2 \mid X) = \frac{n-K}{n} \sigma^2$. Why? Heuristically, the random variable $\hat{\varepsilon}$ has only $(n - K)$ degrees of freedom because it is constrained by the K equations $X'\hat{\varepsilon} = 0$. It is more common to use the unbiased

$$s^2 \equiv \frac{1}{n-K} \sum_{i=1}^n \hat{\varepsilon}_i^2$$

Estimators of the standard deviation are often important. We will call them *standard errors*. $\sqrt{s^2}$ is the standard error of the regression, and $SE(\hat{\beta}) = (s^2 \left[(X'X)^{-1}\right]_{kk})^{1/2}$ is the standard error of $\hat{\beta}_k$.
Remark. We **cannot** claim that these estimators are unbiased! Further, the use of 'standard error' is dominant in econometrics, but other disciplines[1] 'standard error' and '[sampling] standard deviation' may be synonyms. In this case, we use '*estimated* standard errors'.
Remark. The spherical error assumption was only (fully) used for the variance expressions (and Gauss-Markov). Recall that from the algebra:

$$\hat{\beta} = \beta + \underbrace{(X'X)^{-1} X' \varepsilon}_{=\text{estimation error}}$$
$$\implies \operatorname{Var}(\hat{\beta} \mid X) = (X'X)^{-1} X' \underbrace{\mathbb{E}(\varepsilon\varepsilon' \mid X)}_{=D} X (X'X)^{-1}$$

Spherical error ($D = \sigma^2 I_n$) leads to simplification, but as long as we can estimate D, it is not necessary. Recall that

$$\operatorname{Var}(\hat{\beta} \mid X) = (X'X)^{-1} \left(\sum_{i=1}^n X_i X_i' \varepsilon_i^2 \right) (X'X)^{-1}$$

Assuming only heteroskedasticity (*i.e.* D remains diagonal), we have the *oracle estimator*

$$\hat{\operatorname{Var}}_{\text{oracle}}(\hat{\beta} \mid X) \equiv (X'X)^{-1} \left(\sum_{i=1}^n X_i X_i' \varepsilon_i^2 \right) (X'X)^{-1}$$

[1] Statistics...

This is an unbiased estimator, but is not available. Plugging in $\hat{\varepsilon}_i$ leads to a plausible estimator:

$$\hat{\text{Var}}_{\text{HC0}}(\hat{\beta} \mid X) \equiv (X'X)^{-1} \left(\sum_{i=1}^{n} X_i X_i' \hat{\varepsilon}_i^2 \right) (X'X)^{-1}$$

This is biased, which motivates a degree of freedom adjustment:

$$\hat{\text{Var}}_{\text{HC1}}(\hat{\beta} \mid X) \equiv \frac{n}{n-K} (X'X)^{-1} \left(\sum_{i=1}^{n} X_i X_i' \hat{\varepsilon}_i^2 \right) (X'X)^{-1}$$

There are a ton of similar estimators. HC0 is the original (Eicker-White) heteroskedasticity robust variance estimator, and HC1 is the industry standard (*i.e.* it is the STATA default). Neither is obviously best, and Hansen has some other options. In applied work, these are dominant because homoskedasticity is rarely plausible. Using weighted least squares is technically possible, but rarely used because it requires knowing *ex ante* the structure of the heteroskedasticity. Clustered standard errors are similar to this, though we omit them for now.

From here on, we will make the following (strong) assumption:
Assumption 1.2. *ε has a normal distribution:*

$$(\varepsilon \mid X) \sim \mathcal{N}(0, \sigma^2 I_n)$$

With this, we have the following:
Theorem 1.4. *Define $s^2 = \frac{(Y-X\beta)'(Y-X\beta)}{n-K}$ and let the matrix $R \in \mathbb{R}^{r \times K}$ have maximal rank $r \leq K$. Under Assumptions 1.1 and 1.2, we then have*

$$(\hat{\beta} - \beta) \mid X \sim \mathcal{N}(0, \sigma^2 (X'X)^{-1})$$

and:

$$t\text{–ratio} = t \equiv \frac{\hat{\beta}_k - \beta_k}{\left(s^2 \left[(X'X)^{-1}\right]_{kk}\right)^{1/2}} \sim t_{n-K}$$
$$F\text{–statistic} = F \equiv \frac{(R\hat{\beta} - R\beta)'(R(X'X)^{-1}R')^{-1}(R\hat{\beta} - R\beta)}{s^2 r} \sim F_{r,n-K}$$

Thus, the null hypothesis $\mathbb{H}_0 : R\beta = r$ can be tested with exact size control by comparing

$$\frac{(R\hat{\beta} - r)'(R(X'X)^{-1}R')^{-1}(R\hat{\beta} - r)}{s^2 r}$$

to the relevant quantiles of $F_{r,n-K}$, etc.

Proof. The first part, from section: We showed earlier that $\hat{\beta} = \beta + (X'X)^{-1}X'\varepsilon$, so

$$\mathbb{E}[\hat{\beta} - \beta \mid X] = \mathbb{E}[\hat{\beta} \mid X] - \beta = \beta + (X'X)^{-1}X' \underbrace{\mathbb{E}[\varepsilon \mid X]}_{=0} - \beta = 0$$

and

$$\text{Var}(\hat{\beta} - \beta \mid X) = \text{Var}(\hat{\beta} \mid X) = \sigma^2 (X'X)^{-1}$$

Conclusion follows by also noting that the sum of normally-distributed random variables.

The rest follows directly from the definitions of the relevant distributions, and our assumptions on R. □
Remark. Recall that the t-distribution converges quickly to the standard normal as degrees of freedom

increase, but the quantiles (and thus associated critical values) can be quite different under small degrees of freedom.

2 The Linear Model in Large Samples

We will now look at large sample (asymptotic) theory; results will be weaker in that they maximally only hold *almost surely* (*i.e.* we care about stochastic convergence). However, our assumptions are also a lot weaker:

Assumption 2.1. *(X, Y) are i.i.d., $\mathbb{E}Y^2 < \infty$ and $\mathbb{E}\|X\|^2 < \infty$, and $\mathbb{E}(XX')$ is positive definite.*

Remark. These assumptions suffice for the projection coefficient to be well-defined: $b^\star \equiv (\mathbb{E}XX')^{-1}\mathbb{E}XY$. We have:

Theorem 2.1. *Under Assumptions 2.1, we have:*

$$\hat{\beta} \equiv (\mathbb{E}_n XX')^{-1}\mathbb{E}XY \xrightarrow{p} b^\star$$

Proof. By the weak law of large numbers, we have that $\mathbb{E}_n XX' \xrightarrow{p} \mathbb{E}XX'$ and $\mathbb{E}_n XY \xrightarrow{p} \mathbb{E}XY$. By the Continuous Mapping Theorem, it follows that $(\mathbb{E}_n XX')^{-1} \xrightarrow{p} (\mathbb{E}XX')^{-1}$, since as n increases $\mathbb{E}_n XX'$ is nonsingular with probability approaching 1. The claim follows directly from Slutsky's Theorem. □

We have not yet used any assumptions that set the structural linear model apart from the best linear predictor under square loss interpretation, but we've made no causal claims. However, the asymptotics apply to both!

We must assume the following to make causal claims, if the linear model $Y = X'\beta + \varepsilon$ is maintained:

Assumption 2.2. *$\mathbb{E}X\varepsilon = 0$, also called* predetermination*. Additionally, we strengthen the moment assumption: $\mathbb{E}Y^4 < \infty$ and $\mathbb{E}\|X\|^4 < \infty$.*

Remark. This is significantly weaker than the unbiasedness assumption – we only need that ε is uncorrelated with the contemporaneous regressors. This ensures that $b^\star = \beta$ (and, by implication, consistency):

$$b^\star = (\mathbb{E}XX')^{-1}\mathbb{E}XY = (\mathbb{E}XX')^{-1}\mathbb{E}X(X'\beta + \varepsilon) = \beta + (\mathbb{E}XX')^{-1}\mathbb{E}X\varepsilon = \beta$$

We could alternatively show consistency from scratch using this assumption, the way Hayashi does.

To look at the asymptotic distribution, note that

$$\begin{aligned} \hat{\beta} &= (\mathbb{E}_n XX')^{-1}\mathbb{E}_n XY = (\mathbb{E}_n XX')^{-1}\mathbb{E}_n X(Xb^\star + e) \\ \implies \hat{\beta} - b^\star &= (\mathbb{E}_n XX')^{-1}\mathbb{E}_n Xe \end{aligned}$$

where $b^\star$ is the population projection coefficient and e is the (true underlying) projection error. By predetermination, $(\mathbb{E}_n XX')^{-1}\mathbb{E}_n Xe \xrightarrow{p} 0$. However, there is the suggestion of an asymptotic result – it seems natural that

$$\sqrt{n}(\mathbb{E}_n Xe) \xrightarrow{d} \mathcal{N}(0, \Omega) = \mathcal{N}(0, \mathbb{E}(XX'e^2))$$

where the last equality just defines Ω. If that were the case, we would easily have that

$$\sqrt{n}(\hat{\beta} - b^\star) \xrightarrow{d} \mathcal{N}(0, (\mathbb{E}XX')^{-1}\Omega(\mathbb{E}XX')^{-1})$$

This derivation is basically true.[2] We just need to ensure that all terms exist. For this, our stronger Assumptions 2.2 suffice. If the kurtosis exists, we can argue that Ω is finite, by repeatedly using Cauchy-Schwartz. For any one element of Ω,

$$\begin{aligned} |\mathbb{E}(X_k X_\ell e^2)| &\le \mathbb{E}|X_k X_\ell e^2| = \mathbb{E}(|X_k||X_\ell||e^2|) \\ &\le (\mathbb{E}X_k^2 X_\ell^2)^{1/2}(\mathbb{E}e^4)^{1/2} \le (\mathbb{E}X_k^4)^{1/4}(\mathbb{E}X_\ell^4)^{1/4}(\mathbb{E}e^4)^{1/2} < \infty \end{aligned}$$

[2] 'Morally true' - Jörg

we finish by writing

$$\sqrt{n}(\hat{\beta} - b^\star) \xrightarrow{d} (\mathbb{E}XX')^{-1}\mathcal{N}(0, \Omega) = \mathcal{N}(0, (\mathbb{E}XX')^{-1}\Omega(\mathbb{E}XX')^{-1})$$

where we again use Slutsky and the properties of normal distributions. Formally, we have:
Theorem 2.2. *With Assumptions 2.1 and Assumptions 2.2, we have that*

$$\begin{aligned} \sqrt{n}(\hat{\beta} - b^\star) &\xrightarrow{d} \mathcal{N}(0, \mathrm{aVar}(\hat{\beta})) \\ \textit{where } \mathrm{aVar}(\hat{\beta}) &= (\mathbb{E}XX')^{-1}\Omega(\mathbb{E}XX')^{-1}) \\ &\quad [= Q_{XX}^{-1}\Omega Q_{XX}^{-1} = \Sigma_{XX}^{-1}\Omega\Sigma_{XX}^{-1}] \end{aligned}$$

Proof. A generalization of above. □

Remark. Note that we are *not* assuming a linear model here! We actually get this result under relatively limited assumptions, we only need that the moment conditions exist. If we want to get inference results, we need more, but as a projection result this still holds.

Definition. aVar is the *asymptotic variance*, the variance of the limiting distribution. In general, this is not necessarily the asymptotic limit of an estimator's squared variance, though the current assumptions suffice.

Remark. This theorem provides the *joint* asymptotic distribution of estimates. The information contained in joint normality is relevant for:

1. Inference on a linear combination of estimates, *e.g.* their sum or difference. This could also be achieved by reparameterization, but that's impractical.
2. Joint inference, *i.e.* confidence ellipsoids, on several coefficients
3. Inference on a known, differentiable function of β through the Delta method (conceptually straightforward, but very important in practice! See the textbook for an example worked through)
4. Conservative inference on a known, nondifferentiable function of β through projection (*i.e.* operate the function on ever b in the confidence ellipsoid). In structured cases, you may be able to improve on this – ask Jörg if this question arises in your research!

Remark. Results in this course hold pointwise as $n \to \infty$ for *given* parameter values, not *uniformly* over parameter values. How big of a problem is this? With some more effort, most results in this course are available uniformly in 'nice' cases. However, there are several cases that are not nice and are empirically relevant: (i) estimators that can be corner solutions of their problems, (ii) estimation of maxima, (iii) rare events, and (iv) post-model selection estimation and inference. In these, pointwise perspectives can be quite misleading.

Theorem 2.3. *Under Assumptions 2.1 and Assumptions 2.2, we have that*

$$\mathrm{a}\hat{\mathrm{Var}}_{HC0} \equiv (\mathbb{E}_n XX')^{-1}\hat{\Omega}(\mathbb{E}_n XX')^{-1} \xrightarrow{p} \mathrm{aVar}(\hat{\beta})$$

where $\hat{\Omega} \coloneqq \mathbb{E}_n[XX'\hat{\varepsilon}^2]$, and similarly for $HC1$, and so on. Basically, all of these are consistent.

Proof. (Sketch) The bottleneck is the consistency of $\hat{\Omega}$. Write:

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i \hat{e}_i^2 = \underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i X_i' e_i^2}_{\xrightarrow{p}\Omega} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i X_i (\hat{e}_i^2 - e_i^2)}_{\xrightarrow{p}0}$$

That the rightmost term disappears can be shown through (tedious) repeated use of Cauchy-Schwartz and Hölder. □

Proof. (Assuming Homoskedasticity) If we assume that $\mathbb{E}(e^2 \mid X) = \sigma^2$, then we have the simplification

$$\mathrm{aVar}(\hat{\beta}) = (\mathbb{E}XX')^{-1}\sigma^2$$

and showing the consistency of

$$\hat{\mathrm{aVar}}(\hat{\beta}) = (\mathbb{E}XX')^{-1}s^2$$

is simple, and requires weaker assumptions (specifically, second moments suffices). However, recall that this makes sense only for a structural linear model (and is still restrictive). $\square$

Inference. Let $r : \mathbb{R}^k \to \mathbb{R}$ be a continuously differentiable function with $\nabla r(\cdot) = R(\cdot)$ (the easiest example is where r extracts a component of β). Define $\theta = r(\beta)$ and $\hat{\theta} = r(\hat{\beta})$. By Delta Method, standard convergence results, and Slutsky, we have

$$\begin{aligned}
\sqrt{n}(\hat{\theta} - \theta) &\xrightarrow{d} \mathcal{N}(0, \mathrm{aVar}(\hat{\theta})) \\
\mathrm{aVar}(\hat{\theta}) &= R(\beta)\,\mathrm{aVar}(\hat{\beta})R(\beta)' \\
\hat{\mathrm{aVar}}(\hat{\theta}) \equiv R(\hat{\beta})\,\mathrm{aVar}(\hat{\beta})R(\hat{\beta})' &\xrightarrow{p} R(\beta)\,\mathrm{aVar}(\hat{\beta})R(\beta) \\
\implies t(\theta) \equiv \frac{\hat{\theta} - \theta}{SE(\hat{\theta})} \equiv \frac{\sqrt{n}(\hat{\theta} - \theta)}{\left(R(\hat{\beta})\hat{\mathrm{aVar}}(\hat{\beta})R(\hat{\beta})'\right)^{1/2}} &\xrightarrow{d} \mathcal{N}(0, 1)
\end{aligned}$$

This is the asymptotic t-statistic. For this to hold, we need that $\mathrm{aVar}(\hat{\theta})$ is finite. A sufficient condition is that $R(\beta) \neq 0$ and that $\mathrm{aVar}(\hat{\beta})$ has full rank.
Definition. Dividing by the standard error like this is called *studentization*. It ensures that the asymptotic distribution does not depend on unknown parameters, ensuring that the t-statistic (and others) are *asymptotic pivots*.

The previous result lets us create hypothesis tests and confidence intervals where the asymptotic sizes converge. Let $\Phi(\cdot)$ denote the standard normal cdf and define the quantiles $\Phi^{-1}(1 - \alpha) \coloneqq c_\alpha$. Then

$$\begin{aligned}
\mathbb{P}\{|t(\theta)| \leq c_{\alpha/2}\} &\xrightarrow{p} 1 - \alpha \\
\mathbb{P}\{t(\theta) \in CI_\alpha(\theta)\} &\xrightarrow{p} 1 - \alpha \\
CI_\alpha(\theta) &= \left[\hat{\theta} - c_{\alpha/2} \cdot SE(\hat{\theta}), \hat{\theta} + c_{\alpha/2} \cdot SE(\hat{\theta})\right]
\end{aligned}$$

and we compute one-sided confidence intervals similarly. If $r(\beta) = p'\beta$ for some known vector p, then $R(\cdot) = p'$ and the t-statistic simplifies to

$$t(\theta) \equiv \frac{\hat{\theta} - \theta}{SE(\hat{\theta})} = \frac{\sqrt{n}(\hat{\theta} - \theta)}{\left(p'\hat{\mathrm{aVar}}(\hat{\beta})p\right)^{1/2}}$$

If p is a basis vector, this further simplifies the t-statistic for an individual coefficient. Choices like $p = (0, 1, -1, 0, \ldots, 0)$ lets us test the equality of two coefficients.

Under the causal linear model, another application is to $p = x$, in which case $\theta = \mathbb{E}(Y \mid X = x)$, which is called a *regression interval*. Note that the standard error depends on x and will be smaller for more central values of x.
Remark. The regression interval is *not* a forecast confidence interval! For forecast intervals, we must take ε_t into account.

We can generalize this to $\theta = r(\beta)$ where $r : \mathbb{R}^k \to \mathbb{R}^q$ is vector-valued. With this, $R(\cdot)$ is now the Jacobian of r, we have that

$$\begin{aligned}
\sqrt{n}(\hat{\theta}-\theta) &\xrightarrow{d} \mathcal{N}(0, \mathrm{aVar}(\hat{\theta})) \\
\mathrm{aVar}(\hat{\theta}) &= R(\beta)\,\mathrm{aVar}(\hat{\beta})R(\beta)' \\
\mathrm{a}\hat{\mathrm{V}}\mathrm{ar}(\hat{\theta}) \equiv R(\hat{\beta})\,\mathrm{aVar}(\hat{\beta})R(\hat{\beta})' &\xrightarrow{p} R(\beta)\,\mathrm{aVar}(\hat{\beta})R(\beta) \\
\implies W(\theta) \equiv \sqrt{n}(\hat{\theta}-\theta)'(\mathrm{a}\hat{\mathrm{V}}\mathrm{ar}(\hat{\theta}))^{-1}\sqrt{n}(\hat{\theta}-\theta) &\xrightarrow{d} \chi^2_q
\end{aligned}$$

This is the *Wald* statistic. It can induce confidence regions similarly to the confidence intervals above – however these are ellipsoids. These will be appropriate if one is genuinely interested in simultaneous inference of several scalars, and their projections onto the axes will be valid but (possibly very) conservative confidence intervals.
Remark. A sufficient condition for this to hold is that both $R(\beta)$ and $\mathrm{aVar}(\hat{\beta})$ are full-rank, which would mean that the hypotheses must be (locally) linearly independent at the truth – for linear hypotheses, this is easy to check and a global property.

3 Instrumental Variables

Remark. Instrumental variables are of huge importance across economics, and are one of the greatest contributions of econometrics to empirical methods across science. They are actively used in causal inference across disciplines, specifically in biostatistics. Their appeal is that they allow for causally interpretable estimates if we think that (i) the linear model (or generalizations) is structural so β has causal interpretation, but (ii) ε correlates with X, because it absorbs relevant omitted variables. Of course, this remarkable result requires strong assumptions.
Model. For simplicity, consider simple linear regression

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where we cannot assets that $\mathbb{E}X\varepsilon = 0$. The interpretation is that β_1 has a causal interpretation, but we could not estimate all relevant covariates. Now suppose that we have a random variable Z with the following properties:

$$\underbrace{\operatorname{cov}(Z, X) \neq 0}_{\text{Relevance}} \qquad \text{and} \qquad \underbrace{\operatorname{cov}(Z, \varepsilon) = 0}_{\text{Validity}}$$

This implies that

$$\frac{\operatorname{cov}(Z, Y)}{\operatorname{cov}(Z, X)} = \frac{\beta_1 \operatorname{cov}(Z, X) + \operatorname{cov}(Z, \varepsilon)}{\operatorname{cov}(Z, X)} = \beta_1$$

Example. (Ginburgh & van Ours (2003)) Y is the career success of a classical musician, X is their placement in a prestigious competition for young musicians, and Z is the order of appearance at the competition. The effect of X on Y is interesting, but 'talent' impacts both. However, appearing late in the competition (verifiably) predicts success! Since order of appearance is random, it serves as an *instrumental variable* (or just *instrument*).
Definition. *Relevance* requires that $\operatorname{cov}(Z, X) \neq 0$. We need there to be some instrument-induced variation to play around with. Otherwise, we could pay a research assistant to flip coins all day and use that as an instrument.

Relevance must be **testable**. The covariance is consistently estimated with its sample analog. Indeed, it is standard practice to report the F-statistic from a 'first-stage regression' of X on Z.
Definition. *Validity* requires that $\operatorname{cov}(Z, \varepsilon) = 0$. We need the instrument-induced variation to be exogenous. Otherwise, we could just use X as an instrument for itself.

Validity is **not testable**. This will change once we have more instruments than regressors.
Definition. We can illustrate causal models using *Directed Acyclic Graphs (DAG)*. Arrows read as 'causes...'. The traditional OLS graph is Figure 2.

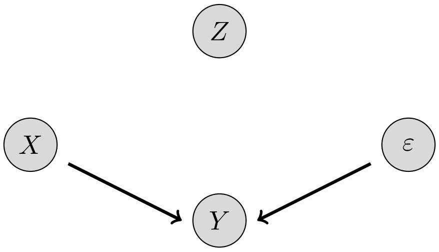

Figure 2: OLS DAG. The r.v. Z plays no role, and $\operatorname{cov}(X, \varepsilon) = 0$.

If X is *endogenous*, the DAG may look like Figure 3.

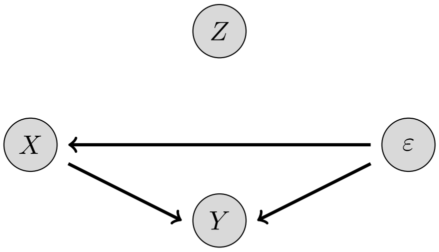

Figure 3: OLS DAG. The r.v. Z plays no role, but $\text{cov}(X, \varepsilon) \neq 0$.

The typical DAG representation of instrumental variables is Figure 4.

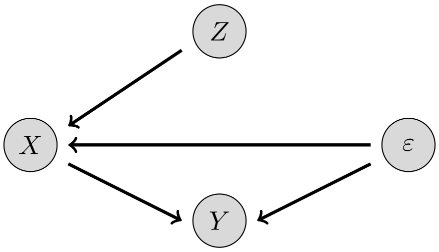

Figure 4: IV DAG. $\text{cov}(Z, X) \neq 0$, $\text{cov}(Z, \varepsilon) = 0$, and $\text{cov}(X, \varepsilon) \neq 0$.

Example. Instrumental variables in medical statistics: *Mendelian randomization* is an instrumental variables technique. The idea is that, in our terminology, genetic variation exogenous. Think about Y = coronary disease, X = HDL cholesterol level, and ε = confounders (of which there are many – diet and lifestyle heavily affect HDL cholesterol level). If we have some genetic variant Z that increases HDL cholesterol level, we can use that as an instrument. The bottleneck assumption is that Z affects *only* HDL cholesterol level, not coronary disease.

Example. Some famous[3] examples:

1. Wright (1928) was the first example – he wanted to estimate supply and demand elasticities for vegetable oils, but realized that quantities and prices are equilibrium quantities and prices, so endogenous. He used demand shifters (price changes of substitutes) and supply shifters (weather) as instruments. He averaged the estimators, which is something we would do now.

2. Angrist & Krueger (1991) aim to estimate the returns to compulsory schooling, but selection into schooling is correlated. They instrument it by quarter of birth, where a different quarter may lead to another year 'exposed' to compulsory schooling.

3. Angrist & Evans (1998) want to estimate the effect of having children on female labor force participation, but the decision to have children is endogenous. They instrument with gender of children – this is clearly random, but some parents prefer to have children of both genders so if they have two of the same, they are more likely to have a third than if the two are different genders.

4. Card (1993,5) looks to estimate the returns to college education, but college attendance is endogenous. He instruments college attendance with distance from a college growing up, which may be pivotal in attendance decisions.

Definition. We will derive the IV estimator as a *Generalized Method of Moments (GMM)* estimator. Recall

[3] Note! Famous $\neq$ good!

the method of moments: If we assume

$$Y = X'\beta + \varepsilon, \mathbb{E}X\varepsilon = 0 \Longrightarrow \mathbb{E}(X(Y - X'\beta)) = 0$$

then a natural idea for estimating β is to solve the empirical moment condition

$$\mathbb{E}_n(X(Y - X'\hat{\beta})) = 0$$

which clearly yields the OLS estimator (as the above is the FOC of the minimization problem). However, if we assume that $\mathbb{E}X\varepsilon \neq 0$, then OLS will be inconsistent for β because it will estimate the projection coefficient

$$b^\star = (\mathbb{E}XX')^{-1}\mathbb{E}XY = \beta + (\mathbb{E}XX')^{-1}\mathbb{E}X\varepsilon$$

However, if we also observe a random k-vector Z with $\mathbb{E}Z\varepsilon = 0$ and $\text{rank}(\mathbb{E}ZX') = k$, then we have the moment condition

$$\mathbb{E}(Z(Y - X'\beta)) = 0 \Longrightarrow \beta = (\mathbb{E}ZX')^{-1}\mathbb{E}ZY$$

where the assumptions assume that this is well-defined. Naturally, the estimator is

$$\hat{\beta}_{IV} = (\mathbb{E}_n ZX')^{-1}\mathbb{E}_n ZY$$

Note that

$$\hat{\beta}_{IV} = (\mathbb{E}_n ZX')^{-1}\mathbb{E}_n ZY = (\mathbb{E}_n ZX')^{-1}\mathbb{E}_n Z(X'\beta + \varepsilon) = \beta + \underbrace{(\mathbb{E}_n ZX')^{-1}\mathbb{E}_n Z\varepsilon}_{= \text{ estimation error}}$$

From here, consistency follows essentially directly as $\mathbb{E}_n Z\varepsilon \xrightarrow{p} 0$. However, we cannot claim unbiasedness, even if we assume the stronger that $\mathbb{E}(\varepsilon \mid Z) = 0$. We will defer asymptotic theory until later. Note that some components of Z may also appear in X, because if some elements of X are unassociated with ε, they can act as their own instruments. By setting $Z = X$, we consider OLS as the special case of X instrumenting itself.

In the simple linear model, when X and Z are scalars, the IV slope estimator can be expressed as

$$\frac{\sum_{i=1}^n (Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^n (Z_i - \bar{Z})(X_i - \bar{X})}$$

The data matrix expression is, naturally, $\hat{\beta}_{IV} = (Z'X)^{-1}Z'Y$. For an application, consider regressing X on Z and then Y on $\hat{X}$. In data matrix notation, we have

$$\begin{aligned}
\tilde{\beta} &= (\hat{X}'\hat{X})^{-1}\hat{X}'Y \\
&= ((Z(Z'Z)^{-1}Z'X)'(Z(Z'Z)^{-1}Z'X))^{-1}(Z(Z'Z)^{-1}Z'X)'Y \\
&= (X'Z(Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y \\
&= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y \\
&= (Z'X)^{-1}Z'Z(X'Z)^{-1}X'Z(Z'Z)^{-1}Z'Y \\
&= (Z'X)^{-1}Z'Y = \hat{\beta}_{IV}
\end{aligned}$$

Remark. The IV estimator can be thought of as a two-stage model where we project X onto Z and then project Y onto the fitted values $\hat{X}$. This gives some straightforward intuition – we exploit only the variation in X that is due to[4] variation in Z. This interpretation is closely related to the DAG interpretation, and it's why the regression of X on Z is often called the *first-stage regression*. It is usually reported and should be highly significant – a rule of thumb is that $F \geq 10$ for the overall first-stage regression.

[4] In a correlative sense, we have not made causal claims. We will return to this when thinking about machine learning and optimal instruments.

Also note that this does not require Z and X to be the same length – we can use more instruments than regressors. However, a caveat:
Definition. If $Z'X$ is not square (but assuming it still has maximal rank!) the algebraic deviation becomes

$$\begin{aligned}\tilde{\beta} &= (\hat{X}'\hat{X})^{-1}\hat{X}'Y \\ &= ((Z(Z'Z)^{-1}Z'X)'(Z(Z'Z)^{-1}Z'X))^{-1}(Z(Z'Z)^{-1}Z'X)'Y \\ &= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y = \hat{\beta}_{TSLS}\end{aligned}$$

where we can go no further than the last step. This is the *two-stage least squares (TSLS)* estimator.
Remark. A major difference in the estimators is that for IV, $Z'\hat{\varepsilon} = Z'(Y - X\hat{\beta}) = 0$ by construction. Can this also be true in TSLS? No! With $\ell > k$ instruments, this is ℓ linear equations in k unknowns. This has some major implications: the estimator cannot set $Z'\varepsilon = 0$, but we can show that the estimator minimizes (in b) $\|Z'(Y - Xb)\|$ for some norm, not necessarily the best norm. However, the validity of instruments becomes testable – in large samples, we should have that $\frac{1}{n}Z'\hat{\varepsilon} \approx 0$. These considerations lead us from two-stage least squares to the generalized method of moments. Detailed asymptotic characterization of TSLS follows from GMM.

We can think about asymptotics of the simple instrumental variable case, where

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where (Y, X, z) are scalars, and we make the further assumption of homoskedasticity.[5] From previous algebra, we get that

$$\begin{aligned}\sqrt{n}(\hat{\beta}_1^{IV} - \beta_1) &= \frac{\sqrt{n}\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z})\varepsilon_i}{\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z})(X_i - \bar{X})} \\ &= \frac{\sqrt{n}\frac{1}{n}\sum_{i=1}^n (Z_i - \mathbb{E}Z)\varepsilon_i}{\frac{1}{n}\sum_{i=1}^n (Z_i - \mathbb{E}Z)(X_i - \mathbb{E}X)} + o_P(1) \\ &\xrightarrow{d} \frac{\mathcal{N}(0, \sigma_Z^2\sigma^2)}{\rho\sigma_Z\sigma_X} = \mathcal{N}\left(0, \frac{\sigma^2}{\rho^2\sigma_X^2}\right)\end{aligned}$$

We can compare this directly to

$$\sqrt{n}(\hat{\beta}_1^{OLS} - \beta_1) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2}{\sigma_X^2}\right)$$

So the asymptotic variance of the IV estimator has the first stage explained variation in the denominator.
Remark. This is not a result about finite sample variance! In fact, the finite sample variance of the IV estimator does not exist.

A corollary of this result is that the IV estimator has higher asymptotic variance, as a trade-off to bias. If $\rho^2 = 1$, the two asymptotic variances coincide (in fact, the estimators algebraically coincide – this is equivalent to using X as an instrument for itself). The asymptotic variance diverges to ∞ as $\rho \to 0$, which suggests some intuition about weak instruments.
Definition. A *weak instrument* is an instrument Z where $\operatorname{corr}(Z, X) \approx 0$, meaning that the instruments explain very little of the variation in X.
Example. (Following Staiger & Stock (1997)) To formally model a weak instrument, set $\rho_n = \frac{\rho}{\sqrt{n}}$. Asymptotic approximation is powerful – it allows us to invoke CLT and other results. However, in a pointwise perspective as it trivializes the problem as previous asymptotics hold for any $\rho \neq 0$. *Parameter drift* allows us to invoke asymptotic approximations without approximating away the problem. Compare to Pitman Drift for analyzing the local power of hypothesis tests. The idea is *not* that parameters actually change with n.

[5] We make this assumption because (i) the asymptotic variance is instructive, and (ii) it allows us to formally characterize weak instruments.

Rather, the idea is to internalize the intuition that whether an instrument is weak depends on ρ in relation to n. See Goldberger on micronumerosity for more on this.

We will develop this for scalar (X, Z). See Hayashi for a general statement. The first- and second stage regressions are

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \varepsilon \\ X &= \gamma_0 + \frac{\gamma_1}{\sqrt{n}} Z + \eta \end{aligned}$$

Then

$$\hat{\beta}_1 - \beta_1 = \frac{\sqrt{n}\mathbb{E}_n(Z - \bar{Z})\varepsilon}{\sqrt{n}\mathbb{E}_n(Z - \bar{Z})(X - \bar{X})}$$

but

$$\begin{aligned} \sqrt{n}\mathbb{E}_n(Z - \bar{Z})(X - \bar{X}) &= \sqrt{n}\mathbb{E}_n(Z - \bar{Z})\left(\gamma_0 + \frac{\gamma_1}{\sqrt{n}} Z + \eta\right) \\ &= \underbrace{\gamma_1 \mathbb{E}_n(Z - \bar{Z})Z}_{\xrightarrow{p} \gamma_1 \sigma_Z^2} + \sqrt{n}\mathbb{E}_n(Z - \bar{Z})\eta \end{aligned}$$

Assuming CLT applies, we attain that

$$\hat{\beta}_1 - \beta_1 \xrightarrow{d} \frac{a}{\gamma_1 \sigma_Z^2 + b} \qquad \text{where} \qquad \begin{pmatrix} a \\ b \end{pmatrix} \sim \mathcal{N}(0, \Omega)$$

where Ω is the variance-covariance matrix of $(Z\varepsilon, Z\eta)$. This estimator is inconsistent! Indeed, it converges to a distribution. Observe that the extent of this problem scales inversely with $|\gamma_1|$ – as $|\gamma_1| \to \infty$, the problem vanishes.

Remark. One of the examples referenced above, Angrist & Krueger (1991), has an instrument that is arguably weak, though n is large. Bound, Jaeger, & Baker (1993) replicated some of their tables with a random 'instrument' they added to the data. This spawned a large literature in weak instruments. Here at Cornell, Pepe has done a lot of work on weak instruments.

Remark. We might think about using many instruments instead of one strong instrument. In fact, if we let the number of instruments grow as n grows, Hansen has a formalization of the fact that that estimator is inconsistent.

Remark. We can also encounter issues with 'too-strong' instruments. These do not exist in theory, but consider the example where (i) we think that X is endogenous, and (ii) $\text{corr}(Z, X) = 0.99$. Theoretically there's no issue with this, as long as $\text{cov}(Z, \varepsilon) = 0$. However, that's not testable, and intuitively it would be very strange if Z and X were so correlated and Z would not at all be covariant with the errors. So in practice, we often want $F \leq 25$, even though in theory higher F is better.

4 Generalized Method of Moments

The Generalized Method of Moments (GMM) and its relatives like the Method of Simulated Moments and Indirect Inference are of great importance in applied work. This statement (and the name!) are due to Hansen (1982), and contributed to his Nobel Prize. There were precursors to a large part of this theory, but we will develop a fairly general statement, though we restrict attention to linear moment conditions. Nonlinear GMM will be developed as a special case of extremum estimation.

We can think of GMM as extending TSLS in several ways:

1. Since we cannot set sample moments exactly to zero, we must choose a norm to minimize. Is there a best norm?
2. We allow for heteroskedasticity also in the estimation stage (Heteroskedasticity robust standard errors for TSLS are straightforward and standardly used, but we will see that in the estimation stage, TSLS can be argued to presume homoskedasticity.)
3. We consider testing instrument validity
4. It will become clear that restricting to linear moment conditions simplifies the math but is not essential

Remark. The *generalized* in GMM refers to the fact that we allow for (and explore the implications of!) overidentification, when we have more moment equations than parameters.
Definition. We know that

$$\mathbb{E}g(Y, X, Z; \theta) = 0$$

where $\theta \in \mathbb{R}^k$ and $g(\cdot)$ is a known smooth function mapping into $\mathbb{R}^\ell$, with $\ell \geq k$. The case of $\ell > k$ will be called *overidentified*. We will assume $g(\cdot)$ is linear. This is not essential.
Remark. This yields OLS, IV, TSLS, and (after generalizing to multiple outcomes) seemingly unrelated regression equations (SURE) and simple panel data estimators as special cases. For future reference, consider also *probabilistic regression (probit)* where $\mathbb{E}[X(Y - \Phi(X'\beta))] = 0$, and best-response conditions such as Euler equations (this was the original application of GMM, from Hansen & Singleton (1983)).

The *GMM estimator* is

$$\begin{aligned}\hat{\theta}(W) &= \underset{\theta}{\operatorname{argmin}}\, J_n(\theta)\\ J_n(\theta) &= n\bar{g}_n(\theta)'W\bar{g}_n(\theta)\\ \bar{g}_n(\theta) &\equiv \frac{1}{n}\sum_{i=1}^{n} g(\theta; \cdot)\end{aligned}$$

where W is a weight matrix defining the norm that we minimize. Recall that if we are overidentified, we cannot set $J_n(\theta)$ to zero. We therefore have a family of estimators, and will think about how we optimally choose W. The scale factor in $J_n(\cdot)$ is for convenience, ensuring that it converges to a non-degenerate limit.
Example. We begin with the linear case, where

$$g(\beta, \cdot) = Z(Y - X'\beta)$$

This covers all of the estimators we've seen so far, where we might have $Z = X$. In data matrix notation, the estimator minimizes

$$(Z'Y - Z'X\beta)'W(Z'Y - Z'X\beta)$$

so the first order condition is:

$$\begin{aligned} -2X'ZW(Z'Y - Z'X\hat{\beta}) &= 0 \\ \implies X'ZWZ'X\hat{\beta} &= X'ZWZ'Y \\ \implies \hat{\beta} &= (X'ZWZ'X)^{-1}X'ZWZ'Y \\ \Big[&= (S'_{XZ}WS_{XZ})^{-1}S'_{XZ}Ws_{XY}\Big] \end{aligned}$$

where the last line is the same in Hayashi's notation. More instructively, if we set $\mu = Z'Y$ and $G = Z'X$, we want to minimize

$$(\mu - G\beta)'W(\mu - G\beta)$$

which looks exactly like weighted least squares with k regressors, ℓ observations, and weights W, so that

$$\hat{\beta} = (G'WG)^{-1}G'W\mu$$

We can additionally directly compare

$$\begin{aligned} \hat{\beta}_{GMM}(W) &= (X'ZWZ'X)^{-1}X'ZWZ'Y \\ \hat{\beta}_{TSLS} &= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y \end{aligned}$$

so the TSLS estimator is GMM with weights $(Z'Z)^{-1}$ This raises two immediate questions: Since we have a family of weight matrices, which one is the best? When, if ever, is that best weight $(Z'Z)^{-1}$?

We can think about the WLS example – what are the best weights in WLS? With i.i.d. data[6] they are the inverse standard deviation. More generally, the ideal W is the variance-covariance matrix of errors ε.

This raises the idea of estimating the optimal W. But didn't we have earlier than generalized least squares is rarely used in practice? Yes! But this example has limitations. The variance-covariance matrix of ε is $n \times n$, and estimating it comes with a number of additional assumptions (and/or nonparametric convergence rates). In our WLS example, the error is $\eta \equiv \mu - G\beta$, where the variance-covariance matrix is $\ell \times \ell$. Estimating that is plausible.

This result motivates *two-state (efficient) GMM*:

1. Compute a preliminary estimate of β to get residuals
2. Use residuals to estimate the optimal weight matrix $\hat{W}$
3. Report a final estimate $\hat{\beta}_{GMM}(\hat{W})$

We will show that this procedure minimizes asymptotic variance, and if we assume homoskedasticity the TSLS weights $(Z'Z)^{-1}$ will indeed be optimal.

Assumption 4.1. *We have the following assumptions for GMM:*

1. *We observe i.i.d. realizations* $(Y_i, X_i, Z_i), i = 1, 2, \dots$
2. $\mathbb{E}(Z(Y - X'\beta)) = 0$
3. $\mathbb{E}|Y^4| < \infty$, $\mathbb{E}\|X\|^4 < \infty$, $\mathbb{E}\|Z\|^4 < \infty$
4. $Q \equiv \mathbb{E}(ZX')$ *has full rank* k
5. W *is positive definite*
6. $\Omega \equiv \mathbb{E}(ZZ'\varepsilon^2)$ *is positive definite.*

Remark. Assumptions 4.1 are also the assumptions for TSLS, which will emerge as a special case.

[6] Note: excludes panel estimators!

The GMM estimator has the following asymptotic distribution (a generalization of the algebra from OLS):

$$\begin{aligned}
\hat{\beta}_{GMM}(W) &= (X'ZWZ'X)^{-1}X'ZWZ'Y \\
&= \beta + (X'ZWZ'X)^{-1}X'ZWZ'\varepsilon \\
&= \beta + \left(\frac{1}{n}X'ZW\frac{1}{n}Z'X\right)^{-1}\frac{1}{n}X'ZW\frac{1}{n}Z'\varepsilon \\
&= \beta + \left(\mathbb{E}(XZ')W\mathbb{E}(ZX')\right)^{-1}\mathbb{E}(XZ')W\frac{1}{n}Z'\varepsilon + o_P(1) \\
&= \beta + (Q'WQ)^{-1}Q'W\frac{1}{n}Z'\varepsilon + o_P(1) \\
\implies \hat{\beta}_{GMM}(W) - \beta &\xrightarrow{p} 0 \\
\sqrt{n}(\hat{\beta}_{GMM}(W) - \beta) &\xrightarrow{d} \mathcal{N}(0, (Q'WQ)^{-1}Q'W\Omega WQ(Q'WQ)^{-1})
\end{aligned}$$

Remark. This only requires second moments, not fourth moments.

The matrix $\Omega = \mathbb{E}(ZZ'\varepsilon^2)$ is really the variance-covariance matrix of the moment conditions. Intuitively, a condition under which Ω has larger variance across the diagonal is noisier. In fact, in the WLS example from before, Ω parameterizes the heteroskedasticity in our regression with ℓ observations and k parameters. This suggests Ω^{-1} as the efficient weighting matrix – which we do not know, but can estimate using residuals from a preliminary regression. Furthermore, the earlier results hold if $\hat{W} \xrightarrow{p} W$.

Theorem 4.1. *Let Assumptions 4.1 hold, and let $\hat{W} \xrightarrow{p} W^\star \equiv \Omega^{-1}$. Then:*

1. *The asymptotic variance becomes*

$$V^\star = (Q'W^\star Q)^{-1}Q'W^\star\Omega W^\star Q(Q'W^\star Q)^{-1}$$

   which simplifies to

$$V^\star = (Q'\Omega^{-1}Q)^{-1}$$

2. *$V^\star$ as defined in (1) is the best asymptotic variance: $V \geq V^\star$ for any other estimator (meaning that $V - V^\star$ is positive semi-definite)*[7]

3. *$V^\star$ is only attained by estimators that are asymptotically equivalent to $\hat{\beta}(\Omega^{-1})$.*

Proof. We will show that (i) with efficient weighting, V simplifies to $V^\star$, (ii) $V \geq V^\star$, and (iii) the inequality is strict unless the estimators are (asymptotically) equivalent. First, write:

$$\begin{aligned}
V &= A'\Omega A, \text{ where } A = WQ(Q'WQ)^{-1} \\
V^\star &= B'\Omega B, \text{ where } B = \Omega^{-1}Q(Q'\Omega^{-1}Q)^{-1}
\end{aligned}$$

and observe that

$$B'\Omega A = (Q'\Omega^{-1}Q)^{-1}Q'\Omega^{-1}\Omega WQ(Q'WQ)^{-1} = V^\star = B'\Omega B \implies B'\Omega(A - B) = 0$$

Thus,

$$V = A'\Omega A = (B + (A - B))'\Omega(B + (A - B)) = \underbrace{B'\Omega B}_{V^\star} + \underbrace{(A - B)'\Omega B}_{0} + \underbrace{B'\Omega(A - B)}_{0} + \underbrace{(A - B)'\Omega(A - B)}_{\text{p.s.d.}}$$

□

[7] This ordering is very strong! The efficient estimator will also minimize the variance of any $p'\theta$, through linearity of the quadratic form.

Remark. This motivates the efficient (two-stage) GMM estimator:

$$\begin{aligned}\hat\beta_{TSGMM} &\equiv \hat\beta(\hat W)\\ \hat W &\equiv \left(\mathbb{E}_n(ZZ'\hat\varepsilon^2)\right)^{-1}\\ \hat\varepsilon &= Y - X\hat\beta\end{aligned}$$

where $\hat\beta$ is any consistent estimator of β, for example a GMM estimator with any reasonable weighting matrix. The industry standard is TSLS.
Remark. We could also use the *centered estimator*

$$\hat W = \left[\mathbb{E}_n\left((Z\hat\varepsilon - \mathbb{E}_n(Z\hat\varepsilon))(Z\hat\varepsilon - \mathbb{E}_n(Z\hat\varepsilon))'\right)\right]^{-1}$$

which literally estimates the variance, rather than the uncentered second moment of Z. The two are the same if $\mathbb{E}Z\varepsilon = 0$, but the centered estimator is consistent for variance even if the model is misspecified.
Remark. If we further assume homoskedasticity, so that

$$\mathbb{E}(\varepsilon^2 \mid Z) = \sigma^2 \implies \Omega = \mathbb{E}(ZZ'\varepsilon^2) = \sigma^2\mathbb{E}(ZZ')$$

the estimator with the ideal weighting matrix simplifies:

$$\begin{aligned}\hat\beta_{GMM}(\Omega^{-1}) &= (X'Z\sigma^{-2}(\mathbb{E}ZZ')^{-1}Z'X)^{-1}X'Z\sigma^{-2}(\mathbb{E}ZZ')^{-1}Z'Y\\ &= (X'Z(\mathbb{E}ZZ')^{-1}Z'X)^{-1}X'Z(\mathbb{E}ZZ')^{-1}Z'Y\end{aligned}$$

but $\mathbb{E}ZZ'$ can be estimated by $\mathbb{E}_nZZ' = \frac{1}{n}Z'Z$. Because $\frac{1}{n}$ cancels from the expression, we can succinctly write

$$\hat\beta_{GMM}(\Omega^{-1}) = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y = \hat\beta_{TSLS}$$

So under homoskedasticity, the efficient estimator is the two-stage least squares estimator!
Question. Should we *always* do Efficient GMM?

The case for efficient GMM is more compelling than for FGLS, and some results (notably specification testing) require efficient GMM. In practice, efficient GMM is quite common. However, there are some caveats. In finite sample, estimating Ω introduces noise. Monte Carlo simulations suggest that with small samples and moderate heteroskedasticity, TSLS may perform better. On the other hand, estimation of Ω could in principle be iterated, which (under conditions that guarantee convergence at least asymptotically) removes path dependency. Implementations for this exist, but asymptotic analysis does not suggest a gain and we might be concerned about error propagation.
Definition. With modern computing power, we could also directly compute

$$\hat\beta_{cGMM} \equiv^{8} \operatorname*{argmin}_{\beta} \bar g_n(\beta)'\hat W \bar g_n(\beta)$$

where we optimize over the criterion function and the weighting matrix at the same time. This is called *Continuous Updating GMM*, and while not numerically the same estimator has the same asymptotic distribution.
Question. Should we use all the instruments we can think of?

Now that we can use more instruments than regressors, we might be tempted to use all instruments we can think of, including ‘technical instruments’ (polynomials of instruments, etc). Intuitively, we shouldn’t do that. But what is the formal argument against? First, every instrument must be justified, so the ‘cost of assumptions’ will increase (especially the exclusion restriction). Second, estimating another weight matrix introduces more finite sample noise. Third, even without weighting it can be shown that TSLS is inconsistent

[8] Unique result under conditions, which we just assume here.

if there are many instruments in the sense that $\ell_n/n \to \alpha > 0$. For this, the instruments don't even need to be weak. Actually, in practice we should only invoke asymptotics if $n \gg \ell$, limiting the number of instruments we can use.

Definition. What is qualitatively true in overidentified ($\ell > k$) models is that the model itself can be tested! We can actually introduce the *joint validity of moments test*. This works because we can (empirically) test the assumption that the instruments are orthogonal to the sample.

Theorem 4.2. *Under Assumptions 4.1, assuming the model is overspecified, then*

$$J_n \equiv J(\hat{\beta}_{TSGMM}) \xrightarrow{d} \chi^2_{\ell-k}$$

Remark. The intuition here is that we try and set an ℓ-vector to zero but only have k free parameters to do so. This means we have a residual with $\ell - k$ degrees of freedom. If the model is well specified, the residual is of order $O(n^{-1/2})$. If (and only if!) we use the efficient weighting matrix, it is further asymptotically multivariate standard normal in a certain $(\ell - k)$-subspace. Then its square is is $\chi^2_{\ell-k}$.

Proof. Previous results imply that $\frac{1}{n}Z'\varepsilon = O_P(n^{-1/2})$, so we write[9]

$$\begin{aligned} J_n &= n\left(\frac{1}{n}Z'\hat{\varepsilon}\right)'\hat{\Omega}^{-1}\left(\frac{1}{n}Z'\hat{\varepsilon}\right) \\ &\approx n\left(\frac{1}{n}Z'\hat{\varepsilon}\right)'\Omega^{-1}\left(\frac{1}{n}Z'\hat{\varepsilon}\right) \\ &= n\left(C'\frac{1}{n}Z'\hat{\varepsilon}\right)'(C'\Omega C)^{-1}\left(C'\frac{1}{n}Z'\hat{\varepsilon}\right) \\ &= n\left(C'\frac{1}{n}Z'\hat{\varepsilon}\right)'\left(C'\frac{1}{n}Z'\hat{\varepsilon}\right) \end{aligned}$$

where $\Omega^{-1} = CC' \iff \Omega(C')^{-1}C^{-1}$, meaning that C is the Cholesky Root of Ω^{-1}. Next, we have that

$$\begin{aligned} \left(C'\frac{1}{n}Z'\hat{\varepsilon}\right) &= C'\frac{1}{n}Z'(\varepsilon - X(\hat{\beta} - \beta)) \\ &= C'\frac{1}{n}Z'\left[\varepsilon - X\left(\left(\frac{1}{n}X'Z\right)\hat{\Omega}^{-1}\left(\frac{1}{n}Z'X\right)\right)^{-1}\left(\left(\frac{1}{n}X'Z\right)\hat{\Omega}^{-1}\left(\frac{1}{n}Z'\varepsilon\right)\right)\right] \\ &= \left[I_\ell - C'\left(\frac{1}{n}Z'X\right)\left[\left(\left(\frac{1}{n}X'Z\right)\hat{\Omega}^{-1}\left(\frac{1}{n}Z'X\right)\right)^{-1}\left(\left(\frac{1}{n}X'Z\right)\hat{\Omega}^{-1}\left(\frac{1}{n}Z'C\right)\right)\right]\right]C'\frac{1}{n}Z'\varepsilon \\ &\approx \left[I_\ell - \underbrace{C'\left(\frac{1}{n}Z'X\right)}_{:=\hat{R}}\left[\left(\left(\frac{1}{n}X'Z\right)CC'\left(\frac{1}{n}Z'X\right)\right)^{-1}\left(\left(\frac{1}{n}X'Z\right)CC'\left(\frac{1}{n}Z'C\right)\right)\right]\right]C'\frac{1}{n}Z'\varepsilon \\ &= \left(I_\ell - \hat{R}(\hat{R}'\hat{R})^{-1}\hat{R}'\right)C'\frac{1}{n}Z'\varepsilon \\ &\approx \left(I_\ell - R(R'R)^{-1}R'\right)C'\frac{1}{n}Z'\varepsilon, \text{ where } R \equiv C'\mathbb{E}(ZX') \end{aligned}$$

We so far have

$$J_n \approx n\left(C'\frac{1}{n}Z'\hat{\varepsilon}\right)'\left(C'\frac{1}{n}Z'\hat{\varepsilon}\right) \qquad \text{and} \qquad \left(C'\frac{1}{n}Z'\hat{\varepsilon}\right) \approx \left(I_\ell - R(R'R)^{-1}R'\right)C'\frac{1}{n}Z'\varepsilon$$

[9] In this proof, $\approx$ means that we drop an $o_P(1)$ term.

and observe that

$$\sqrt{n}C'\left(\frac{1}{n}Z'\varepsilon\right)\xrightarrow{d}\mathcal{N}(0,C'\Omega C)=\mathcal{N}(0,C'(C')^{-1}C^{-1}C)=\mathcal{N}(0,I_\ell)$$

Define the random variable $u\sim\mathcal{N}(0,I_\ell)$, then it follows that

$$J_n\xrightarrow{d}\left(I_\ell-R(R'R)^{-1}R'\right)'\left(I_\ell-R(R'R)^{-1}R'\right)u\sim\chi^2_{\ell-k}$$

because $I_\ell-R(R'R)^{-1}R'$ projects u onto a lower dimensional subspace. It remains to show that the subspace is of dimension $\ell-k$. This follows from the fact that $I_\ell-R(R'R)^{-1}R'$ is the annihilator matrix associated with $R=C'\mathbb{E}(ZX')$, which has rank k and null space of rank $\ell-k$. (You could use the rank-nullity condition of idempotent matrices to prove this statement directly). □

Example. Visualization of the above theorem. Consider a situation with $k=1$ endogenous regressor and $\ell=2$ instruments, where:

$$\Omega=\begin{bmatrix}4&0\\0&1\end{bmatrix}\implies\Omega^{-1}=\begin{bmatrix}1/4&0\\0&1\end{bmatrix},C=\begin{bmatrix}1/2&0\\0&1\end{bmatrix}$$

$$\mathbb{E}ZX'=\begin{bmatrix}1\\1\end{bmatrix}\implies R=\begin{bmatrix}1/2\\1\end{bmatrix}$$

The first moment condition has four times the variance of the second, so in the WLS analogy it should be given half of the weight. The column space of R is the line spanned by $\begin{bmatrix}1/2&1\end{bmatrix}'$, so its null space is the line spanned by $\begin{bmatrix}-1&1/2\end{bmatrix}'$. Visually, we have Figure 5.

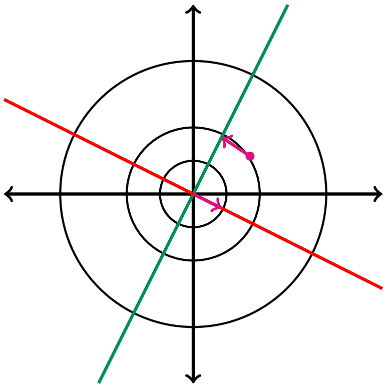

Figure 5: Column Space and Null Space of R in a simple example. An element is projected onto the column space, leaving residuals in the null space that are also standard normal.

Definition. Let $\theta=r(\beta)$ and $\hat\theta=r(\hat\beta)$ for some function $r:\mathbb{R}^k\to\mathbb{R}^q$, where $r\in\mathbf{C}^1$. Suppose also that $R(\beta)=\frac{\partial r(\beta)}{\partial\beta'}$ has full rank of q at the true value of β. Then we can construct a *Wald test*:

$$W\equiv n(\hat\theta-\theta)\left(R(\hat\beta)'\hat V_\beta R(\hat\beta)\right)^{-1}(\hat\theta-\theta)\xrightarrow{d}\chi^2_q$$

We will later develop other tests in a more general setting.

Remark. We can extend this analysis to multiple equations, where

$$\begin{aligned}Y_m&=X_m\beta_m+\varepsilon_m&&m=1,\ldots,M\\\mathbb{E}Z_m\varepsilon_m&=0&&m=1,\ldots,M\end{aligned}$$

The X_m may be the same (like in seemingly unrelated regression equations), or they may overlap (like in panel data with some time invariant regressors) – coefficients are not restricted across equations in a general setting, but they are in many applications.
Example. In *Seemingly Unrelated Regressions* (this example from Griliches, 1976), we may have that

$$\begin{aligned} \mathit{LW69} &= \alpha_1 + \beta_1 \cdot \mathit{schooling69} + \gamma_1 \cdot \mathit{IQ} + \delta_1 \cdot \mathit{experience69} + \varepsilon_1 \\ \mathit{KWW} &= \alpha_2 + \beta_2 \cdot \mathit{schooling69} + \gamma_2 \cdot \mathit{IQ} + \varepsilon_2 \end{aligned}$$

where *LW69* is the logged wage, and *KWW* is a measure of ability (as are *IQ* and *experience69*). We can think of this as regressing the two outcomes on the same regressor but assuming *a priori* that $\delta_1 = 0$. That alone makes this model overidentified, and jointly estimating both equations leads to a more efficient estimate (*if we believe the assumptions!*). We can construct a (fictitious) version of this with *panel data*, and get

$$\begin{aligned} \mathit{LW69} &= \alpha_1 + \beta_1 \cdot \mathit{schooling69} + \gamma_1 \cdot \mathit{IQ} + \delta_1 \cdot \mathit{experience69} + \varepsilon_1 \\ \mathit{LW80} &= \alpha_2 + \beta_2 \cdot \mathit{schooling80} + \gamma_2 \cdot \mathit{IQ} + \delta_2 \cdot \mathit{experience80} + \varepsilon_2 \end{aligned}$$

We can define

$$\bar{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_M \end{bmatrix} \qquad \bar{X} = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_M \end{bmatrix}$$

$$\bar{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_M \end{bmatrix} \qquad \bar{Z} = \begin{bmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_M \end{bmatrix}$$

and we have the moment condition

$$\mathbb{E}(\bar{Z}(\bar{Y} - \bar{X}'\bar{\beta})) = 0$$

and estimator

$$\hat{\beta}(W) = \left(\mathbb{E}_n(\bar{X}\bar{Z}')W\mathbb{E}_n(\bar{Z}\bar{X}')\right)^{-1}\left(\mathbb{E}_n(\bar{X}\bar{Z}')W\mathbb{E}_n(\bar{Z}\bar{Y})\right)$$

The estimator is the same as before, except that the data matrix notation becomes unwieldy. Assuming that $(\bar{Y}, \bar{X}, \bar{Z})$ fulfill Assumptions 4.1, we have that

$$\begin{aligned} \sqrt{n}(\hat{\beta}(W) - \beta) &\xrightarrow{d} \mathcal{N}(0, V_\beta) \\ V_\beta &= (\bar{Q}'W\bar{Q})^{-1}\bar{Q}'W\bar{\Omega}W\bar{Q}(\bar{Q}'W\bar{Q})^{-1} \\ \bar{Q} &= \mathbb{E}(\bar{Z}\bar{X}') \\ \bar{\Omega} &= \mathbb{E}(\bar{Z}\varepsilon\varepsilon\bar{Z}') \end{aligned}$$

and results on efficient GMM are also the same as before.
Remark. This is extremely powerful! We just derived a collection of historically distinct estimators. A small caveat is that we must think very carefully about what the assumptions on the new objects actually mean. We've made the assumption that $(X_1, \ldots, X_M, Y_1, \ldots, Y_M, Z_1, \ldots, Z_M)$ are *all* i.i.d., which is stronger than assuming that equation-by-equation.
Remark. This is the same as estimating the equations separately (*i.e.* will lead to the same result) if (i) everything is just identified *or* (ii) W is block diagonal, where its blocks correspond to equations. It can be

instructive to write out the estimator for $M = 2$:

$$\begin{aligned}\hat{\beta}(\hat{W}) &= \left[\begin{bmatrix}\mathbb{E}_n(X_1Z_1') & 0\\ 0 & \mathbb{E}_n(X_2Z_2')\end{bmatrix}\begin{bmatrix}W_{11} & W_{12}\\ W_{21} & W_{22}\end{bmatrix}\begin{bmatrix}\mathbb{E}_n(X_1Z_1') & 0\\ 0 & \mathbb{E}_n(X_2Z_2')\end{bmatrix}\right]^{-1}\\ &\cdot \begin{bmatrix}\mathbb{E}_n(X_1Z_1') & 0\\ 0 & \mathbb{E}_n(X_2Z_2')\end{bmatrix}\begin{bmatrix}W_{11} & W_{12}\\ W_{21} & W_{22}\end{bmatrix}\begin{bmatrix}\mathbb{E}_n(Z_1Y_1)\\ \mathbb{E}_n(Z_2Y_2)\end{bmatrix}\\ &= \begin{bmatrix}\mathbb{E}_n(X_1Z_1')W_{11}\mathbb{E}_n(Z_1X_1') & \mathbb{E}_n(X_1Z_1')W_{12}\mathbb{E}_n(Z_2X_2')\\ \mathbb{E}_n(X_2Z_2')W_{21}\mathbb{E}_n(Z_1X_1') & \mathbb{E}_n(X_2Z_2')W_{22}\mathbb{E}_n(Z_2X_2')\end{bmatrix}\cdot\begin{bmatrix}\mathbb{E}_n(X_1Z_1')W_{11}\mathbb{E}_n(Z_1Y_1)+\mathbb{E}_n(X_1Z_1')W_{12}\mathbb{E}_n(Z_2Y_2)\\ \mathbb{E}_n(X_2Z_2')W_{21}\mathbb{E}_n(Z_1Y_1)+\mathbb{E}_n(X_2Z_2')W_{22}\mathbb{E}_n(Z_2Y_2)\end{bmatrix}\end{aligned}$$

What happens if $W_{12} = W_{21} = 0$? We get the simplification:

$$\begin{aligned} &= \begin{bmatrix}\mathbb{E}_n(X_1Z_1')W_{11}\mathbb{E}_n(Z_1X_1') & 0\\ 0 & \mathbb{E}_n(X_2Z_2')W_{22}\mathbb{E}_n(Z_2X_2')\end{bmatrix}\cdot\begin{bmatrix}\mathbb{E}_n(X_1Z_1')W_{11}\mathbb{E}_n(Z_1Y_1)\\ \mathbb{E}_n(X_2Z_2')W_{22}\mathbb{E}_n(Z_2Y_2)\end{bmatrix}\\ &= \begin{bmatrix}(\mathbb{E}_n(X_1Z_1')W_{11}\mathbb{E}_n(Z_1X_1'))^{-1}\mathbb{E}_n(X_1Z_1')W_{11}\mathbb{E}_n(Z_1Y_1)\\ (\mathbb{E}_n(X_2Z_2')W_{22}\mathbb{E}_n(Z_2X_2'))^{-1}\mathbb{E}_n(X_2Z_2')W_{22}\mathbb{E}_n(Z_2Y_2)\end{bmatrix}\\ &= \begin{bmatrix}\hat{\beta}_1(W_{11})\\ \hat{\beta}_2(W_{22})\end{bmatrix}\end{aligned}$$

So when we assume that the cross-equation weights are zero, the multi-equation GMM just stacks the single equation estimators!

Question. When should we estimate equations jointly?

Answer. If we naively interpret this result, we could estimate all of the equations in the world jointly – if they're actually unrelated, the weighting matrix will pick that up. This seems intuitively untrue, but how do we show it formally? Well, we face an escalating number of nuisance parameters (entries of W). More importantly, model misspecification is contagious – the estimator's probability limit equals

$$\text{plim}\ \hat{\beta}(W) = \beta + \left(\mathbb{E}(\bar{X}\bar{Z}')W\mathbb{E}(\bar{Z}\bar{X}')\right)^{-1}\mathbb{E}(\bar{X}\bar{Z}')W\mathbb{E}(\bar{Z}\varepsilon)$$

If any one entry of $\mathbb{E}(\bar{Z}\varepsilon)$ is nonzero, then every entry of the matrix product is nonzero. This holds unless W is block diagonal in the equations. In that case (and *only* that case!) the joint estimation is efficient.

Example. *Multiple Regression vs. Seemingly Unrelated Regression* If $X_1 = \cdots = X_M = Z_1 = \cdots = Z_M$, then this is just multiple regression – that is, we regress different $Y_1, \ldots, Y_M$ on the same exogenous regressors. We can verify that the estimator just stacks OLS in this case.

Alternatively, suppose that some regressors are dropped from some equations, meaning that we think their coefficients are zero. However, we still consider them exogenous in all equations, so we can use them as overidentifying instruments. Formally, let $Z_1 = \cdots = Z_M = \bigcup_{m=1}^{M} X_m$ and the X_m are not all the same. This is the *Seemingly Unrelated Regression (SUR)* estimator.[10]

Remark. As we add assumptions to the basic model, we can recover some classic estimators that were developed independently. These assumptions typically allow us to simplify expressions. See Hayashi for more detail.

- If we assume homoskedasticity, we get *Full Information Instrumental Variables Efficient (FIVE) estimation* (from Brundy & Jorgenson, 1971)
- If we additionally assume that $Z_1 = \cdots = Z_M$, we get *Three-Stage Least Squares (3SLS)*[11] (from Zellner & Theil, 1962)
- SUR is the final specialization, where we set $Z_1 = \cdots = Z_M = \bigcup_{m=1}^{M} X_m$

[10] Historically, it would also require homoskedasticity, but that's actually a distinct issue.

[11] In modern terminology, this is a two-stage estimator. However, it could be expressed as TSLS with 'pre-pre-estimation' of cross-equation correlation of errors, hence the name.

Remark. Next, we can think about common coefficients. Consider the following specification:

$$Y_m = X_m\beta + \varepsilon_m, \qquad \mathbb{E}Z_m\varepsilon_m = 0\ \forall\ m = 1,\dots,M$$

The main intuition here is that an essentially the same (but not constant) covariate is observed across equations. We do not need to revisit every covariate in every equation: some components of X_m could be zero almost surely. If we define

$$\bar{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_M \end{bmatrix} \quad ; \quad \bar{X} = \begin{bmatrix} X_1 & \cdots & X_M \end{bmatrix} \quad ; \quad \bar{Z} = \begin{bmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_M \end{bmatrix}$$

then we have moment condition $\mathbb{E}(\bar{Z}(\bar{Y} - \bar{X}'\beta)) = 0$ and estimator

$$\hat{\beta}(W) = \left(\mathbb{E}_n(\bar{X}\bar{Z}')W\mathbb{E}_n(\bar{Z}\bar{X}')\right)^{-1}\mathbb{E}_n(\bar{X}\bar{Z}')W\mathbb{E}_n(\bar{Z}\bar{Y})$$

This looks like what we had before, but the different definition of $\bar{X}$ changes identification: that $\mathbb{E}(\bar{Z}\bar{X}')$ has full rank is implied by $\mathbb{E}(Z_mX_m')$ having full rank for each $m = 1,\dots,M$.

Common coefficients allow for estimation of many parameters that would not otherwise be identifiable. An important application is panel data:

- If we impose the assumptions that characterized SUR before, we get the *Random Effects* estimator
- If we furthermore assume that ε_m is uncorrelated across m, then this simplifies to *Pooled OLS*

The difference between those estimators is about efficiency, not identification. We will revisit random effects and pooled OLS soon.

Remark. Efficient GMM vaguely resembles WLS or *Feasible Generalized Least Squares (FGLS)*. As a reminder, weighted least squares minimizes

$$(Y - X\beta)'W(W - X\beta)$$

where the weighting matrix W gives differential weights to different observations. If we know that $\mathbb{E}(\varepsilon\varepsilon' \mid X) = \sigma^2 \cdot \Omega$ for known Ω, then setting $W = \Omega^{-1}$ is variance minimizing, and the resulting estimator is the BLUE by Gauss-Markov. To see the analogy to WLS as undergraduates learn it, note that we can equivalently minimize

$$(C(Y - X\beta))'(C(Y - X\beta))$$

where C is the Cholesky root of W. In particular, if observations are uncorrelated,

$$\Omega = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \quad ; \quad \Omega^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_n^2 \end{bmatrix} \quad ; \quad C = \begin{bmatrix} 1/\sigma_1 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_n \end{bmatrix}$$

and the minimand can be expressed as $\sum_{i=1}^n((y_i - x_i'\beta)/\sigma_i)^2$.

Question. What do we do if we don't know Ω? We could estimate it, but estimating an $n \times n$ matrix from n data points seems absurd.

There's two possible paths: (i) if we have a specific parameter model for how $\mathbb{E}(\varepsilon^2 \mid X = x)$ changes with x, we could estimate it; and (ii) we could run a general flexible regression of $\hat{\mu}_i^2$ on X. The latter regressions are often run to construct Breusch-Pagan or other heteroskedasticity tests. However, (i) is rare and for (ii) we incur a lot of effort (and error propogation!) for reweighting that only matters if weights are very different than even. This explains why the latter approach has a name (Feasible Generalized Least Squares) but is extremely uncommon in practice.

So what does this have to do with GMM? We can write the GMM estimator as the weighted OLS estimator in a fictitious regression, where we set $\mu = Z'Y$ and $G = Z'X$, then we want to minimize

$$(\mu - G\beta)'W(\mu - G\beta)$$

and this looks like WLS with k regressors, ℓ observations, and weight matrix W. The closed-form WLS estimator

$$\hat{\beta}_{WLS} = (G'WG)^{-1}G'W\mu$$

precisely recovers the GMM estimator (once we substitute back the transformations). So if GMM is similar to FGLS, why is it so much more common? The analogy is useful but has limitations. Every moment condition in the original problem is an 'observation' in the fictitious regression. In our WLS analogy, the 'error' is $\eta \equiv \mu - G\beta$, with variance-covariance matrix of size $\ell \times \ell$, so we are estimating a much smaller matrix, which becomes a lot more precise as n grows! We can also see this in the objective functions:

$$(Y - X\beta)' \underbrace{W}_{n\times n}(Y - X\beta) \qquad \text{vs.} \qquad (\mu - G\beta)' \underbrace{W}_{\ell\times\ell}(\mu - G\beta)$$

5 Panel Data

Remark. An *extremely quick* and non-exhaustive treatment.
Definition. *Panel data* are data that come in a two-dimensional array, most commonly in 'time' and 'units of observation'. We have *sample size* n and *number of waves* T. With a *balanced panel*, we have nT observations.[12] We call a *short panel* one where in practice $T \ll n$, and the asymptotics require that we fix T and let $n \to \infty$. Alternatively, we have a *long panel*, where we fix n and let $T \to \infty$. This is really a multivariate time series. Some applications require analyses where we send $n \to \infty$ and $T \to \infty$ at the same or different rates. Here, we consider only short panels.
Remark. All estimators in this section are variations on multiple equation common coefficients GMM. We will develop from specific to general here, starting with

$$Y_{it} = X_{it}'\beta + \varepsilon_{it} \qquad \text{where } \mathbb{E}(X_{it}\varepsilon_{it}) = 0$$

From there, we get:
Definition. The *pooled OLS estimator* is defined as

$$\hat{\beta}_{pool} \equiv \left(\sum_{i=1}^{n} X_i'X_i\right)^{-1} \sum_{i=1}^{n} X_i'Y_i = (X'X)^{-1}X'Y \xrightarrow{p} \beta$$

This is unbiased and consistent under the assumptions that confirm that previously (where X and Y are pooled data matrices). It also has the same asymptotics if we make the same assumptions as above. However, the homoskedasticity assumption is *incredibly* strong here. At the very least, we need to use a *cluster-robust variance estimator*

$$\hat{V}_{pool} = (X'X)^{-1}\left(\sum_{i=1}^{n} X_i'\hat{\varepsilon}_i\hat{\varepsilon}_i X_i\right)(X'X)^{-1}$$

We had previously omitted cluster-dependent errors, but note the analogy to FGLS: the central term estimates the $(T \times T)$-matrix $\mathbb{E}(X_i'\hat{\varepsilon}_i\hat{\varepsilon}_i X_i)$. However, we could do even better! As previously, the variance matrix is 'small', which raises the possibility of using an FGLS-like approach for estimation. In fact, we are precisely doing efficient (multi-equation) GMM.
Definition. A popular model is the *Random Effects Model*, where we assume cross-sectional homoskedasticity but with the specific structure

$$\mathbb{E}(\varepsilon_i\varepsilon_i' \mid X_i) = \begin{bmatrix} \sigma_u^2 + \sigma_\epsilon^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 + \sigma_\epsilon^2 & \cdots & \sigma_u^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 + \sigma_\epsilon^2 \end{bmatrix}$$

which has only two degrees of freedom. The assumption here is that $\varepsilon_{it} = u_i + \epsilon_{it}$ where u_i and ϵ_{it} are uncorrelated. In other words, our structural equation becomes the model

$$Y_{it} = X_{it}'\beta + u_i + \epsilon_{it} \text{ where } \mathbb{E}(X_{it}u_i) = \mathbb{E}(X_{it}\epsilon_{it}) = 0$$

Remark. The random effects estimator can be thought of as the feasible GLS estimator for this model, where we pre-estimate σ_u and σ_ϵ. We will get more details later, under some theory yet to come. Note that Hayashi defines the random effects estimator as a two-stage GMM assuming homoskedasticity, which is similar to three-stage least squares. This treatment has more degrees of freedom.

What's important is that (i) random effects has the same identifying assumptions as pooled OLS, and (ii) it adds a FGLS/TSGMM step, and since within-unit correlation of ε is both salient and easily modeled (note:

[12] We will consider only balanced panels.

in short panels only!), this is often desirable.
Remark. In the random effects model, the condition that $\mathbb{E}X_{it}u_i = 0$ is very restrictive. If we think of u_i as unobserved, time-invariant covariates, we are saying that these cannot at all be correlated with the observed. If this fails, do we have anything else to estimate?

Consider the *Fixed Effects (FE)* equation

$$Y_{it} - Y_{i,t-1} = (X_{it} - X_{i,t-1})'\beta + \epsilon_{it} - \epsilon_{i,t-1}$$

When can we estimate this by OLS? We need that (i) ϵ_{it} is uncorrelated with past and future ϵ_t, and (ii) that $(X_{it} - X_{i,t-1})$ fulfills a rank condition (this will fail with time-invariant regressors). This is the basic idea of fixed effects estimation, where we can 'difference away' the fixed effects in one of three ways: (i) first differencing (*between estimator*), (ii) demeaning (*within estimator*), and (iii) adding an indicator of each unit (*dummy variable regression*). Demeaning and dummy variables are numerically the same, and correspond to the 'classic' fixed effects estimator. From the point of view of identification, the methods are the same, but they imply different weighting matrices. If the weighting matrix is pre-estimated, they are asymptotically the same.

We additionally have that:

1. The implied weighting matrix $M = 1(1'1)^{-1}1'$ is efficient if the idiosyncratic error ϵ_{it} is homoskedastic and uncorrelated. (for the between estimator, we need that ϵ_{it} follows a random walk in direction t. This is less salient)
2. We can actually show that FE equals TSLS (or really SUR), thinking of the cross-equation restrictions in
$$\mathbb{E}X_{is}\epsilon_{it} = 0 \;\forall\; s, t$$
as overidentifying restrictions / instruments
3. The estimator necessarily has higher (asymptotic) variance than pooled OLS. This is because OLS algebra applies, but demeaning reduces the sum of squared deviations of any random variable
4. For variance estimation, a degrees of freedom adjustment of $T(T-1)$ is not negligible for realistic T and is therefore recommended for including under asymptotic justification
5. In the formal model that motivates the random effects model, the FE estimator can be used to estimate σ_ϵ^2. Doing this first and then backing out σ_u^2 is the standard approach