ECON271: Econometrics II, linear regression

Erich Trieschman

2023 Winter quarter class notes

Contents

1 Regression models

Goal: Estimate $E[Y \mid X]$, oftentimes given $(y_i, x_i) \overset{iid}{\sim} P_\theta$ Probability theory: $P_\theta \to \mathcal{P}_n$ Statistics: $\mathcal{P}_n \to P_\theta$

1.1 Estimator properties

- **Identification:** Parameters of interest can be identified using joint distribution of observable variables and distribution assumptions. E.g., for $Y \sim N(\mu, \sigma^2), \mu = E_{\theta=(\mu,\sigma^2)}[Y]$, but for $Y \sim N(\mu_1 + \mu_2, \sigma^2)$, we can't identify μ_1, μ_2
- **Unbiased:** $E_\theta[\hat{\mu}] = \mu$
- **Admissibility:** Admissible if not inadmissible, where inadmissible means $\exists \tilde{\mu} s.t. E_\theta[(\hat{\mu} - \mu)^2] \geq E_\theta[(\tilde{\mu} - \mu)^2] \forall \theta$
- **Efficiency:** $Var_\theta(\hat{\mu}) \leq Var_\theta(\tilde{\mu}) \forall \tilde{\mu}$ unbiased
- **Consistency:** $\hat{\mu} \xrightarrow{p} \mu$
- **Asymptotic distribution:** $\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} N(0, \sigma^2)$

2 Linear regression and the OLS estimator

$$
\begin{aligned}
y &= x^T\beta + \epsilon, \text{, where} \\
E[\epsilon \mid x] = 0 &\implies E[y \mid x] = x^T\beta \text{ since } E[y \mid x] = E[x^T\beta + \epsilon \mid x] \text{(correct specification)} \\
Var(\epsilon \mid x) &= \sigma^2 \text{ (homoskedasticity)}
\end{aligned}
$$

2.1 Identification

$$
\begin{aligned}
&\beta = E[xx^T]^{-1}E[xy], \text{ since} \\
&\beta = \beta E[xx^T]^{-1}E[xx^T] = E[xx^T]^{-1}E[xx^T\beta] = E[xx^T]^{-1}E[xE[y \mid x]] = E[xx^T]^{-1}E[E[xy \mid x]] = E[xx^T]^{-1}E[xy] \\
&\beta = argmin_b E[(y - x^Tb)^2] \xrightarrow{FOC} E[2x(y - x^T\hat{\beta})] = 0 \implies E[xy] = E[xx^T]\hat{\beta}, \text{ noting this requires } E[xx^T] \text{ invertible}
\end{aligned}
$$

2.2 Estimation

$$
\begin{aligned}
\hat{\beta} =& argmin_b E_n[(y - x^Tb)^2] = argmin_b \frac{1}{n}\sum_{i=1}^{n}(y - x^Tb)^2 = argmin_b (y - X\beta)^T(y - X\beta) \\
\xrightarrow{FOC} \hat{\beta} =& \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i y_i = (X^TX)^{-1}X^Ty, \text{ again requiring } X^TX \text{ invertible}
\end{aligned}
$$

Note by construction, the first order condition is $E[x(y - x^T\beta)] = 0 = E[x\epsilon]$. This is a fact of the estimator.

2.2.1 Estimate as ratio of covariance to variance

TODO (see notes and homework)

2.3 Bias

$$
\begin{aligned}
E[\hat{\beta} \mid X] =& E[(X^TX)^{-1}X^Ty \mid X] = (X^TX)^{-1}X^TE[y \mid X] \\
=& (X^TX)^{-1}X^TX\beta = \beta \text{ when correctly specified, since } E[y \mid X] = X\beta
\end{aligned}
$$

2.4 Variance

$$\begin{aligned}Var(\hat{\beta} \mid X) &= Var((X^TX)^{-1}X^Ty \mid X) = Var((X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^TE \mid X)\\ &= (X^TX)^{-1}X^TVar(X^TE \mid X)X(X^TX)^{-1} = (X^TX)^{-1}X^TVar(x\epsilon \mid x)X(X^TX)^{-1}\\ &= (X^TX)^{-1}X^T\sigma^2X(X^TX)^{-1} = \sigma^2(X^TX)^{-1} \text{ under homoskedasticity assumption}\end{aligned}$$

2.4.1 Asymptotic variance

$$\begin{aligned}\sqrt{n}(\hat{\beta}-\beta) =&\sqrt{n}((X^TX)^{-1}Xy-\beta), \text{ for } X \text{ data matrix of } x_i,\ y \text{ data vector of } y_i,\ (y_i,x_i) \text{ iid}\\ =&\sqrt{n}((X^TX)^{-1}Xy-(X^TX)^{-1}(X^TX)\beta) = \sqrt{n}(X^TX)^{-1}(Xy-X^TX\beta)\\ =&(X^TX)^{-1}\left(\sqrt{n}(X^T(X\beta+E))-X^TX\beta\right) = (X^TX)^{-1}\left(\sqrt{n}X^TE\right)\\ &(X^TX)\xrightarrow{p}E[xx^T] \text{ (LLN)} \implies (X^TX)^{-1}\xrightarrow{p}E[xx^T]^{-1} \text{ (continuous mapping theorem)}\\ &\sqrt{n}(X^TE-0) = \sqrt{n}(X^TE-E[E[x\epsilon \mid x]]) = \sqrt{n}(X^TE-E[x\epsilon]) \xrightarrow{d} N(0,Var(x\epsilon))\\ &\xrightarrow{d} N(0,E[xx^T]^{-1}Var(x\epsilon)E[xx^T]^{-1})\\ &\xrightarrow{d} N(0,E[xx^T]^{-1}E[x\epsilon^2x^T]E[xx^T]^{-1}) \text{ for } Var(x\epsilon) = E[(x\epsilon)(x\epsilon)^T] = E[x\epsilon^2x^T]\end{aligned}$$

Depending on correct specification and homoskedasticity, the asymptotic variance can be simplified

$$\begin{aligned}Var(x\epsilon) =&Var(E[x\epsilon \mid x]) + E[Var(x\epsilon \mid x)] = Var(xE[\epsilon \mid x]) + E[xVar(\epsilon \mid x)x^T]\\ =&0 + E[xVar(\epsilon \mid x)x^T] \text{ under correct specification}\\ =&Var(xE[\epsilon \mid x]) + \sigma^2E[xx^T] \text{ under homoskedasticity}\\ =&\sigma^2E[xx^T] \text{ under both, leading to } \sqrt{n}(\hat{\beta}-\beta)\xrightarrow{d}N(0,\sigma^2E[xx^T]^{-1})\end{aligned}$$

2.5 Efficiency of linear regression

2.5.1 Gauss-Markov Theorem

Theorem: Under assumptions below, OLS is Best Linear Unbiased Estimator (BLUE), where best is defined with respect to $Var(\hat{\beta})$

Assumptions:

- Correct specification (alternative: no omitted variable bias): $E[\epsilon_i \mid x_i] = 0$
- Homoskedasticity: $Var(\epsilon_i \mid x_i) = \sigma^2$
- No colinearity of regressors: X^TX invertible when $x_i \in \mathbb{R}^{k>1}$, or $Var(x) > 0$ when $x_i \in \mathbb{R}$

Proof sketch:

- Want to show: $Var(\hat{\beta}) \preceq Var(\tilde{\beta}) \forall \tilde{\beta}$ linear and unbiased
- Suffice to show: $Var(\tilde{\beta}) - Var(\hat{\beta}) \preceq 0 \implies Var(\tilde{\beta}) - Var(\hat{\beta}) \in S_{++}$
- Note $\tilde{\beta} = Wy \implies WX = I$ since $E[\tilde{\beta} \mid X] = \beta \implies WX\beta = \beta$
- Note $\tilde{\beta} = \hat{\beta} + W(I - X(X^TX)^{-1}X^T)y$
- Note $Cov(\hat{\beta}, W(I - X(X^TX)^{-1}X^T)y) = 0$
- Combining these observations we see $\tilde{\beta} = \hat{\beta} + S$ for $S \in S_{++}$

2.6 Incorrect specification

Even under misspecification, we can write

$E[x\epsilon] = 0$, since $E[x\epsilon] = E[x(y - x^T\beta)]$ and we define beta as$\beta \quad := argmin_bE[(y - x^Tb)^2]$ where the first order condition is $-2E[x(y -$

And we can use linear prediction as an approximation for the true underlying model. Note here that unlike for the correctly specified OLS, the estimand depends on the distribution of x, not just $E[y \mid x]$

$$E[y \mid x] \neq x^T\beta, \text{ but instead}$$
$$\beta = argmin_bE[(E[y \mid x] - x^Tb)^2] = E[xx^T]^{-1}E[xy]$$

2.6.1 Omitted variable bias

Suppose

$$\begin{aligned}
\text{True model: } & y = \beta_1^* + x\beta_2^* + u\beta_3^* + \epsilon, \text{ where } E[\epsilon \mid x, u] = 0 \\
\text{Regression: } & y = \beta_1 + x\beta_2 \\
\text{Then } \hat{\beta}_2 \text{ estimates } \beta_2^* &= \frac{Cov(y,x)}{Var(x)} = \frac{Cov(\beta_1^* + x\beta_2^* + u\beta_3^* + \epsilon, x)}{Var(x)} = \frac{Cov(\beta_1^*, x) + Cov(x\beta_2^*, x) + Cov(u\beta_3^*, x) + Cov(\epsilon, x)}{Var(x)} \\
&= \beta_2^* + \beta_3^* \frac{Cov(u,x)}{Var(x)}
\end{aligned}$$

3 Maximum likelihood estimation (MLE)

Estimation technique where we find the parameter that maximizes the likelihood of our data:

$$\hat{\theta} = argmax_\theta f_\theta(z_1, \ldots, z_n) = \prod_{i=1}^{n} f_\theta(z_i) \text{ for } z_i \text{ i.i.d.}$$

Oftentimes, we maximize the log-likelihood instead because it i) simplifies calculations, i) provides numerical stability, and iii) has ties to the information inequality ($\theta_0 = argmax_\theta E[\log f_\theta(x)]$)

3.1 Conditional maximum likelihood

When we focus on conditional maximum likelihood, we don't always need to estimate all parameters. In fact, the log helps us drop extraneous ones.

$$\begin{aligned}
\text{Given: } z = (y,x), \;\; y \mid x \sim f_\beta(y \mid x), \;\; x \sim g_\phi(x) &\implies f_\theta(x) = f\beta(y \mid x) g_\phi(x) \\
\log L(\theta) &= \sum_{i=1}^{n} \log(f_\theta(z_i)) = \sum_{i=1}^{n} \log(f_\beta(y_i \mid x_i)) + \log(g_\phi(x_)) \\
\frac{\partial}{\partial \beta} \log L(\theta) &= \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \log(f_\theta(z_i)) + 0
\end{aligned}$$

3.2 Generalized linear models

Linear prediction ($\nu = x^T\beta$) with a link function ($E[y \mid x] = g^{-1}(\nu) = \mu$). Common family is the linear exponential family of densities ($f_\mu(y) = \exp a(\mu) + b(y) + c(\mu)y$)

Distribution	**Linear exponential density**	$E[y]$	$Var(y)$
Normal (σ^2 known)	$\exp(\frac{-u^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2) - \frac{y^2}{2\sigma^2} + \frac{\mu}{\sigma^2}y)$	$\mu = \mu$	σ^2
Bernoulli	$\exp(\ln(1-p) + \ln(\frac{p}{1-p})y)$	$\mu = p$	$\mu(1-\mu)$
Exponential	$\exp(\ln(\lambda) - \lambda y)$	$\mu = \frac{1}{\lambda}$	μ^2
Poisson	$\exp(-\lambda - \ln(y!) + y\ln\lambda)$	$\mu = \lambda$	μ

3.3 Extremum estimators

Extremum estimators (also called M-estimators) solve $\hat{\theta} = argmax_\theta \hat{Q}_n(\theta)$. Under regularity conditions (including uniform convergence of $\hat{Q}_n(\theta)$ to $Q_0(\theta)$), we have that $\hat{\theta} \xleftarrow{p} \theta_0$ (consistency).

Clearly, the MLE is an extremum estimator: $\frac{1}{n}\sum_{i=1}^{n} \log(f_\theta(z_i)) = \hat{Q}_n(\theta) \longrightarrow Q_0(\theta) = E_{\theta_0}[\log(f_\theta(z))]$ with $\theta_0 = argmax Q_0(\theta)$. Hence, MLE is consistent

3.4 Asymptotic normality

We say that $\hat{\theta}$ is asymptotically linear with influence function $\psi(z)$ if

$$\begin{aligned}
&\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_i \psi(z_i) + o_P(1) \text{ with } E[\psi(z)] = 0 \text{ and finite variance} \\
&\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, E[\psi(z)\psi(z)^T]) \text{ by CLT}
\end{aligned}$$

Consider the FOC of the MLE

$$\sum_i s_{\hat{\theta}}(z_i) = 0 \text{ where } s_\theta = \partial/\partial\theta \log f_\theta(z)$$

$$s_{\hat{\theta}}(z_i) \approxeq s_{\theta_0}(z_i) + \partial/\partial\theta s_{\theta_0}(z_i)(\hat{\theta} - \theta_0)$$

$$s_{\hat{\theta}}(z_i) = s_{\theta_0}(z_i) + \partial/\partial\theta s_{\overline{\theta}}(z_i)(\hat{\theta} - \theta_0) \text{ by mean-value theorem for } \left\|\overline{\theta} - \theta_0\right\|_x \leq \left\|\hat{\theta} - \theta_0\right\|_x$$

$$0 = \sum_i s_{\hat{\theta}}(z_i) = \sum_i s_{\theta_0}(z_i) + \sum_i \partial/\partial\theta s_{\overline{\theta}}(z_i)(\hat{\theta} - \theta_0)$$

$$\sqrt{n}(\hat{\theta} - \theta_0) = \left[-\frac{1}{n}\sum_i \partial/\partial\theta s_{\overline{\theta}}(z_i)\right]^{-1} \frac{1}{\sqrt{n}}\sum_i s_{\theta_0}(z_i) \text{ with}$$

$$\left[-\frac{1}{n}\sum_i \partial/\partial\theta s_{\overline{\theta}}(z_i)\right]^{-1} \xrightarrow{p} E\left[\frac{\partial s_{\theta_0}(z)}{\partial\theta}\right]^{-1}, \quad \frac{1}{\sqrt{n}}\sum_i s_{\theta_0}(z_i) \xrightarrow{d} N(0, Var(s_{\theta_0}(z)))$$

$$\text{so } \sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, H^{-1}JH^{-1}) \text{ where } H = E\left[\frac{\partial s_{\theta_0}(z)}{\partial\theta}\right] \text{ and } J = Var(s_{\theta_0}(z)) = E[s_{\theta_0}(z)z_{\theta_0}(z)^T]$$

When correctly specified and under regularity conditions, the Information Matrix Equality ($H = -J$) applies and this asymptotic distribution simplifies to

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, J^{-1})$$

3.5 Cramer-Rao lower bound

Under some regularity conditions, any unbiased estimator $\hat{\theta}$ of θ has variance that is no smaller than

$$Var\left(\frac{\partial \log f_{\theta_0}(z)}{\partial\theta_0}\right)^{-1} = -E\left[\frac{\partial^2 \log f_{\theta_0}(z)}{\partial\theta_0\partial\theta_0^T}\right] \text{ (by the information inequality)}$$
$$= J^{-1} \text{ (as defined above)}$$

3.6 Misspecification and QMLE

The QMLE estimates

$$\theta_0^* = \max_\theta E_{f_0}[\log f_\theta(z)]$$

For which the density $f_{\theta_0^*}(\cdot)$ (in our pre-specified family) is the best approximation to the true density $f_0(\cdot)$, in the sense of minimizing K—L Divergence.

$$D(f_\theta \mid\mid f_0) = E_{f_0}[\log(f_0(z)/f_\theta(z))] = E_{f_0}[\log(f_0(z)] - E_{f_0}[/f_\theta(z))]$$

And when the likelihood is in fact correctly specified, $f_0(z) = f_{\theta_0}(z)$ then $D(f_\theta \mid\mid f_0) = 0$

Method	**True dist.**	**Estimated dist.**	**Estimate**	**K-L Divergence**
MLE	$z \sim f_{\theta_0}$	f_{θ_0}	$\theta_0 = argmax_\theta E_{f_{\theta_0}}[\log f_\theta(z)]$	$D(f_{\theta_0} \mid\mid f_{\theta_0}) = 0$
QMLE	$z \sim f_0$	f_{θ_0}	$\theta_0^* = argmax_\theta E_{f_0}[\log f_\theta(z)]$	$D(f_0 \mid\mid f_{\theta_0^*}) > 0$

Note that the information inequality does not hold under QMLE so we have $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, H^{-1}JH^{-1})$

3.7 Tests

Under our null hypothesis we suppose that some function of our parameters equals zero, $H_0 : a(\theta_0) = 0$ where $a(\theta) : \mathbb{R}^k \to \mathbb{R}^r$.

3.7.1 Test overview

- The three tests are equivalent in large samples
- Wald has a tendency to over-reject in finite samples (size distortion)
- LM has low finite-sample power against some alternative
- Wald and LM are not invariant to reparameterizations; for example, the parameterization $H_0 : \ 1/(\beta_L + \beta_K) - 1 = 0$ could produce a different result in finite samples than $H_0 : \ \beta_L + \beta_K - 1 = 0$, even though it's the same hypothesis

Test	model under null, $\tilde{\theta}$	model under alternative, $\hat{\theta}$
Wald	False	True
Lagrange Multiplier	True	False
Likelihood Ratio	True	True

3.7.2 Wald test

For $\hat{\theta}$ asymptotically normal, and given null, $H_0 : a(\theta_0) = 0$

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V) \implies \sqrt{n}(a(\hat{\theta}) - a(\theta_0)) \xrightarrow{d} N\left(0, \frac{\partial a(\theta_0)}{\partial \theta} V \frac{\partial a(\theta_0)}{\partial \theta}^T\right) \text{ by the delta method}$$

$$H_0 : \sqrt{n} a(\hat{\theta}) \xrightarrow{d} N\left(0, \frac{\partial a(\theta_0)}{\partial \theta} V \frac{\partial a(\theta_0)}{\partial \theta}^T\right)$$

$$H_0 : \sqrt{n}\left(\frac{\partial a(\theta_0)}{\partial \theta} V \frac{\partial a(\theta_0)}{\partial \theta}^T\right)^{-\frac{1}{2}} a(\hat{\theta}) \xrightarrow{d} N(0, I)$$

$$H_0 : n\left(\left(\frac{\partial a(\theta_0)}{\partial \theta} V \frac{\partial a(\theta_0)}{\partial \theta}^T\right)^{-\frac{1}{2}} a(\hat{\theta})\right)^T \left(\left(\frac{\partial a(\theta_0)}{\partial \theta} V \frac{\partial a(\theta_0)}{\partial \theta}^T\right)^{-\frac{1}{2}} a(\hat{\theta})\right) \xrightarrow{d} \chi^2_r \stackrel{d}{=} \sum_{j=1}^{r} Z_j^2$$

$$H_0 : n * a(\hat{\theta})^T \left(\frac{\partial a(\theta_0)}{\partial \theta} V \frac{\partial a(\theta_0)}{\partial \theta}^T\right)^{-1} a(\hat{\theta}) \xrightarrow{d} \chi^2_r$$

Rejecting the null hypothesis when

$$W = n * a(\hat{\theta})^T \left(\frac{\partial a(\theta_0)}{\partial \theta} V \frac{\partial a(\theta_0)}{\partial \theta}^T\right)^{-1} a(\hat{\theta}) > Q_{1-\alpha(\chi^2_r)}$$

We must assume that $\frac{\partial a(\theta_0)}{\partial \theta}$ has full row rank, r. I.e., the number of hypothesis, r does not exceed the number of parameters, k, and the null hypothesis are not redundant or mutually inconsistent.

3.7.3 Likelihood ratio test

Here we look at the difference in log likelihoods between an unrestricted estimator and a restricted estimator follows a Chi squared distribution

$$\hat{\theta} = argmax_{\theta \in \Theta} \hat{Q}_n(\theta) \text{ the unrestricted estimator}$$

$$\tilde{\theta} = argmax_{\theta \in \Theta} \hat{Q}_n(\theta) \; s.t. \; a(\theta) = 0 \text{ the restricted estimator}$$

$$H_0 : 2(\log L_n(\hat{\theta}) - \log L_n(\tilde{\theta})) \xrightarrow{d} \chi^2_r$$

Rejecting the null hypothesis when

$$LR = 2(\log L_n(\hat{\theta}) - \log L_n(\tilde{\theta})) > Q_{1-\alpha(\chi^2_r)}$$

3.7.4 Lagrange multiplier (score) test

The motivation of this test is to see what the gradient of the log likelihood is under the restricted parameters. Under the null hypothesis, we assume this gradient is close to zero (the maximizer). For the same restricted and unrestricted estimators defined above

$$H_0 : \frac{1}{n} \frac{\partial \log L_n(\tilde{\theta})}{\partial \theta}^T \hat{J}^{-1} \frac{\partial \log L_n(\tilde{\theta})}{\partial \theta} \xrightarrow{d} \chi^2_r \text{ where } \hat{J} \text{ is an efficient estimator for the Fischer Information Matrix}$$

Rejecting the null hypothesis when

$$LM = \frac{1}{n} \frac{\partial \log L_n(\tilde{\theta})}{\partial \theta}^T \hat{J}^{-1} \frac{\partial \log L_n(\tilde{\theta})}{\partial \theta} > Q_{1-\alpha(\chi^2_r)}$$

4 Generalized method of moments (GMM)

The generalized methods of moments estimand is a vector function, $g(z, \theta)$, such that the moment, $E[g(z, \theta)]$ identifies θ_0:

$$E[g(z,\theta)] = 0 \iff \theta = \theta_0, \text{ equivalently we have}$$
$$\theta_0 = argmin_\theta E[g(z,\theta)]^T W E[g(z,\theta)] \text{ for any } W \in S_{++}$$

By analogy principle, we get the generalized method of moments estimator

$$\hat{\theta} = argmin_\theta \left(\frac{1}{n}\sum_{i=1}^{n} g(z_i,\theta)\right)^T \hat{W} \left(\frac{1}{n}\sum_{i=1}^{n} g(z_i,\theta)\right) \text{ for any } W \in S_{++}$$

Note, this is an extremum estimator, hence we get with it all the properties of consistency and asymptotic distribution!

- If $r < k$, the model is not identified (no unique solution)
- If $r = k$, the model is just identified and the choice of $\hat{W}$ is inconsequential
- If $r > k$, the model is overidentified. In this case, the choice of $\hat{W}$ affects the estimator

4.1 Asymptotic normality

Using Taylor expansions and the mean value theorem we can write the scaled difference between the estimator and the truth with respect to an influence function:

$$\sqrt{n}(\hat{\theta} - \theta) = -(\hat{G}^T \hat{W} \overline{G})^{-1} \hat{G}^T \hat{W} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(z_i, \theta_0), \text{ where } \hat{G} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial g(z_i,\hat{\theta})}{\partial \theta}^T, \overline{G} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial g(z_i,\overline{\theta})}{\partial \theta}^T$$

Since we can write it in this way, when we also assume the conditions in the consistency theorem, we have asymptotic normality

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, (G^T W G)^{-1} G^T W \Omega W G (G^T W G)^{-1}), \text{ by Slutsky's Lemma where } \Omega = E[g(z,\theta_0) g(z,\theta_0)^T]$$

Note when GMM is just-identified

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, G^{-1} W^{-1} G^{-T} G^T W \Omega W G G^{-1} W^{-1} G^{-T}), \text{ since we can now distribute the inverse}$$
$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, G^{-1} \Omega G^{-T})$$

Note when GMM is over-identified, we can minimize the variance of our estimator by choosing $W = c\Omega^{-1}$

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, (G^T (c\Omega^{-1}) G)^{-1} G^T (c\Omega^{-1}) \Omega (c\Omega^{-1}) G (G^T (c\Omega^{-1}) G)^{-1}) = N(0, (G^T \Omega^{-1} G)^{-1})$$

4.2 OLS as a special case

OLS is a special case of GMM. With assumptions $y = x^T\beta + \epsilon$, $E[\epsilon \mid x] = 0$, we have

$$E[g(z,\theta)] = 0 \iff E[x\epsilon] = 0 \iff E[x(y - x^T\beta)] = 0$$

Which is simply the first order condition that we solve in OLS! We note that MLE is a special case of GMM too.

4.3 MLE as a special case

MLE is a special case of GMM. With $g(z, \theta) = s_\theta(z) = \partial/\partial\theta \log f_\theta(z)$ we note that

$$\hat{\theta}_{MLE} = argmin_\theta E[\log f_\theta(z)] \implies \partial/\partial\theta \log f_\theta(z) = s_\theta(z) = 0$$

4.4 Example set-up

Set up for simple heteroskedastic linear regression

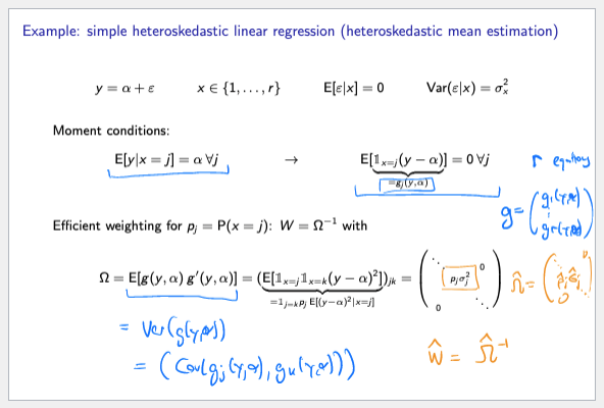

Figure 1: Example of simple heteroskedastic linear regression

5 Inconsistency in OLS

OLS consistency:

- $\epsilon_i \sim N(0, \sigma^2)$: Normal errors (or mean zero, same variance)
- $\epsilon_i \perp x_i$: Fully independent errors
- $E[\epsilon_i \mid x_i] = 0$: Conditional mean independent errors
- $E[\epsilon_i x_i] = 0$: Uncorrelated errors

OLS is always consistent for something, often times called a pseudo-value. The question is whether that something is what you want in the model. How to assess consistency:

1. Propose the estimation procedure: what we do with the data.
2. Make assumptions about the true data generating process. This is from our theory and NOT from the data.
3. Decide on the meaning of a parameter in the context of the assumed true data generating process
4. Ask if the estimator is consistent for the parameter under the assumed true data generating process

5.1 Omitted variable bias

Omitted variable bias biases our estimator (no longer consistent for true estimator). For

- True model: $y = \beta_0 + \beta_i X + \beta_o U + \epsilon$
- Estimation: $y = \beta_0 + \beta_i X$

We have

$$\begin{aligned}\hat{\beta}_i \longrightarrow & Cov(X, y)Var(X)^{-1} = Cov(X, \beta_0 + \beta_i X + \beta_o U + \epsilon)Var(X)^{-1} \\ & = \beta_i + \beta_o Cov(X, U)Var(X)^{-1}\end{aligned}$$

In summary, use long regression if and only if $\beta_o \neq 0$. See slide 115+ for details.

- When $Cov(X, U) \neq 0$, only the long regression is consistent
- When $Cov(X, U) = 0$, the long regression is more efficient. Intuitively, accounting for x reduces the error variance.
- When $\beta_o = 0$, but $Cov(X, U) \neq 0$, the short regression is more efficient. Intuitively, a parsimonious model with less parameters is more precisely estimated.

Nonlinearity is a form of omitted variable bias:

- True model: $y = \beta_0 x^2 + 0x$
- Estimation: $y = \beta x$

5.2 Measurement error bias

See p123 of lecture notes. Instead of measuring x, we can only measure $w = x + u$ where $u \perp x, e$. A similar model is a *proxy* model where we assume $x = w + v$ with $v \perp w, e$. Under proxy assumptions, OLS is consistent (but people mostly don't make this assumption)

5.3 Sample selection

Motivated by wanting to study weekly wage. If a person doesn't work, then we can't calucate this estimate (since we can't divide by zero). f we want to interpret the coefficient on education as the return to education for the average person in the population (regardless of whether they are currently working or not), then such a conditional regression might suffer from sample selection bias. To account for such bias, we need to jointly model the wage equation and the labor force participation decision.

This requires a two-stage approach, or a more elaborate likelihood function that also includes the propensity to work.

Problem set-up:

$$\begin{aligned} \ln W &= X_w'\beta_w + \epsilon_w \text{ log weekly wage} \\ U &= X_u'\gamma_u + \epsilon_u = X_w'\gamma_w + Z'\gamma_z + \epsilon_u \text{ utility from working} \\ P(U > 0 \mid X_u) &= P(\epsilon_u > X_u'\gamma_u \mid X_u) = 1 - \Phi(-X_u'\gamma_u) \\ (\epsilon_w, \epsilon_u) &\sim N(0, \Sigma_\epsilon) \end{aligned}$$

See p133 of class notes for walkthrough of the Heckman two-stage least squares and the Mills Ratio.

- Typically the second stage $\hat{\beta}_w$ and $\hat{\theta}$ will be consistent (as if the true α is known) as long as the first stage $\hat{\alpha}$ is consistent. Here $\theta = \sigma_{uv}/\sigma_u$
- Furthermore, the second stage $\hat{\beta}_w$ and $\hat{\theta}$ will be $\sqrt{n}$ consistent and asymptotically normal (as if the true α is known) as long as the first stage $\hat{\alpha}$ is consistent and asymptotically normal
- However, both nonrobust AND robust standard errors are in general incorrect
  - Solution 1: derive an analytic correction formula (MLE)
  - Solution 2: Use bootstrap; valid for i.i.d. sample.
  - Solution 3: rely on software

See p141 of class notes for walkthrough of the MLE approach. This considers two cases. Case 1, not working: $P(\text{not work} \mid X_u) = \Phi(-X_u'\gamma_u)$. Case 2, log wage and working: $P(\ln W, \text{work}|X_u) = f(\ln W \mid X_u)P(\text{work} \mid X_u)$

Comparison between two-step and MLE:

- Two step estimators are numerically easier to implement, and are closely related to identification arguments
- Statistical inference requires additional work to account for the sampling noise from the first stage, which MLE achieves
- Two step estimators tend to be less statistically efficient than maximum likelihood, unless additional steps are designed to improve efficiency.

5.4 Instrumental variables

See p150 of class notes for instrumental variables to address simultenaity (modeling supply and demand on prices)

For $y_i = x_i'\beta + \epsilon_i$

- OLS is inconsistent when $Cov(x_i, \epsilon_i) \neq 0$, meaning that at least one element of x_i is correlated with ϵ_i
- A vector of instruments z_i is available so that $Cov(z_i, \epsilon_i) = E[z_i\epsilon_i] = 0$
- z_i includes other predictors excluded from x_i, as well as other predictors in x_i (i.e., z_i and x_i share common components w_i)
- Require $dim(z_i) \geq dim(x_i)$
- A good instrument z_i is one that is independent of ϵ_i and correlated with x_i. Larger covariance $Cov(z_i, z_i)$ leads to smaller variance $Var(\hat{\beta}_{IV})$.

When $dim(z_i) = dim(x_i)$, $E[z_i\epsilon_i] = E[z_i(y_i - x_i'\beta)] = 0 \implies E[z_iy_i] = E[z_ix_i']\beta \implies \beta = E[z_ix_i']^{-1}E[z_iy_i]$, and

$$\begin{aligned} \sqrt{n}(\hat{\beta}_{IV} - \beta) &= \left(\frac{1}{n}\sum_i^n z_ix_i'\right)^{-1} \frac{1}{n}\sum_i^n z_i\epsilon_i \\ &\xrightarrow{d} N(0, E[z_ix_i']^{-1}E[z_iz_i\epsilon_i^2]E[z_ix_i']^{-1}) \end{aligned}$$

When $dim(z_i) > dim(x_i)$, we still expect in the populationthat $E[z_i(y_i - x_i'\beta)] = E[z_i\epsilon_i] = 0$, so we can choose to minimize $\frac{1}{n}\sum_i^n z_i(y_i - x_i'\hat{\beta})$

$$\hat{\beta} = argmin_b \left(\frac{1}{n}\sum_i^n z_i(y_i - x_i'b)\right)^T W_n \left(\frac{1}{n}\sum_i^n z_i(y_i - x_i'b)\right)$$
$$\hat{\beta}_{2SLS} = argmin_b \left(\frac{1}{n}\sum_i^n z_i(y_i - x_i'b)\right)^T \left(\frac{1}{n}\sum_i^n z_i z_i'\right)^{-1} \left(\frac{1}{n}\sum_i^n z_i(y_i - x_i'b)\right)$$

With robust asymptotic distribution

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} N(0, (Exz'(Ezz')^{-1}Ezx')^{-1}Exz'(Ezz')^{-1}E\epsilon^2 zz'(Ezz')^{-1}Ezx'(Exz'(Ezz')^{-1}Ezx')^{-1})$$
$$\xrightarrow{d} N(0, (GWG')^{-1}GW\Omega WG'(GWG')^{-1})$$

where $G = Exz'$ is Jacobian of moments wrt parameters

$W = (Ezz')^{-1}$ is weight matrix

$\Omega = E\epsilon^2 zz'$ is var matrix of moments

See p174 of class notes for 3SLS or "2-step GMM"

- If the linear model is correct, then calculating standard errors with 2SLS and 3SLS should be similar
- Large difference raises concern of model misspecification, which can be tested statistically. If the linear model is correct, there is no first order asymptotic benefit in iterating; If the linear model is misspecified, iteration might not converge and exposes the possibility of misspecification.

5.4.1 J-test

When using "optimal weighting matrix" $W = \Omega - 1$ is that

$$\left(\frac{1}{n}\sum_i^n z_i(y_i - x_i'\beta)\right)^T W_n \left(\frac{1}{n}\sum_i^n z_i(y_i - x_i'\beta)\right) \sim \chi^2_d$$

Under H_0: all instruments are correct where d = [n. instruments] - [n. regressors]

- This forms the basis of a J-test, or over-identification test. It is used to test the joint validity of all the instruments.
- he estimator is more efficient if the J-test does not reject.

6 Nonlinear models

Typical steps in analyzing a nonlinear model:

- Show estimator consistency
- Derive estimator asymptotic distribution
- Provide a consistent estimate of the asymptotic distribution
- Conduct Monte Carlo simulation to evaluate the estimator finite sample properties
- Use the Delta method to conduct inference on the function(al) of the parameter.

6.1 Model based and robust standard errors in nonlinear models

There are two general principle for deriving asymptotic variance:

1. Robust sandwich variance that does not take into account
2. Model based asymptotic variance that takes into account correct model specification.

The difference in robust versus model based std errors lies in how to compute $E[]$: robust standard errors replace $E[] \leftarrow \frac{1}{n}\sum_{i=1}^n()$. Model based standard errors calculate $E_{\hat{\theta}}$ by integrating $f(y \mid x, \hat{\theta})$ first.

Robust: The sandwich form $\hat{H}^{-1}\Omega\hat{H}^{-1}$ for the asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta_*)$ is the general form of the Huber-White robust variance estimate. It is robust under general misspecification that essentially only requires the existence of θ_*

Model-based: If we believe that the conditional model is correctly specified and $\theta_0 = \theta_*$. They are not robust against model misspecification.

6.2 The Delta Method for functions of parameters

6.2.1 Analytic delta method

For $\sqrt{n}(\hat{\theta} - \theta_0) \longrightarrow N(0, \Sigma)$

$$\sqrt{n}(\rho(\hat{\theta}) - \rho(\theta_0)) = \frac{\partial \rho(\theta)'}{\partial \theta_0} N(0, \Sigma) = N\left(0, \frac{\partial \rho(\theta)'}{\partial \theta_0} \Sigma \frac{\partial \rho(\theta)}{\partial \theta_0}\right)$$

See p102 of class notes for analytic derivation of asymptotic distribution of partial effects for probit and logit models.

6.2.2 Asymptotic delta method

Analytic delta method can be costly to compute or not possible. One alternative is to bootstrap, but that requires recomputing $\hat{\theta}$ repeatedly, which can be expensive. The asymptotic delta method is an alternative:

1. Draw $\theta_i \sim N(0, \frac{1}{n}\hat{\Sigma})$ for $i \in \{1, \ldots, r\}$
2. Recompute $\rho(\theta_i)$
3. Use the resulting histogram to form confidence intervals (by calculating percentiles)

7 Binary choice models

- For $y \in \{0, 1\}$, $E[y \mid x] = 1 * p(y = 1 \mid x) + 0 * p(y = 0 \mid x) = p(y = 1 \mid x)$
- Since $x'\beta$ isn't bounded between 0 and 1, it's not the best model to use to model probability. but for $F : \mathbb{R} \to [0, 1]$, we can model $p(y = 1 \mid x) = F(x'\beta)$. CDFs can make great link functions

7.1 Deviance

Likelihood is related to the probability of your data given parameters. You want to make it as big as possible. Deviance refers to a notion of distance between data and the fit of the model. You want to make it as small as possible.

$$\text{Deviance} = -2\log[\text{likelihood}] + C \text{ for logistic regression}$$

$$\text{Deviance} = \sum_i^n \hat{e}^2$$

7.2 Partial effects

Unlike the linear OLS model, now the partial effect is a function of x.

$$\frac{\partial E[y \mid x]}{\partial x_j} = \frac{\partial F(x'\beta_0)}{\partial x_j} = f(x'\beta_0)\beta_j$$

- Average partial effects: $\frac{1}{n}\sum_i^n \frac{\partial}{\partial x_j} F(x_i'\hat{\beta}) = \frac{1}{n}\sum_i^n f(x_i'\hat{\beta})$
- Partial effects at the mean: $\frac{\partial E[y|\overline{x}]}{\partial x_j} = \frac{\partial F(\overline{x}'\hat{\beta})}{\partial x_j} = f(\overline{x}'\hat{\beta})$

8 Time series

- Transformation methods include detrending, deseasonizing, differencing, log-differencing (approximating growth rate), etc.
- Identical and independently distributed (iid) in a crossectional setting maps to stationary and weekly dependent in a timeseries setting
- Week stationarity: mean and covariance are finite and don't depend on t. Strong stationarity requires the distribution of any series isn't dependent on t
- Weekly dependent: observations far apart are virtually independent

8.1 CLT under stationarity

$$Var(S_n) = \frac{1}{n}\left(n\Sigma + \sum_{l=1}^{n}(n-l)\Gamma(l)\right) = \Sigma + \sum_{l=1}^{n}\left(1-\frac{l}{n}\right)(\Gamma(l)+\Gamma^T(l)) = \sum_{l=-n}^{n}\left(1-\frac{l}{n}\right)\Gamma(l)$$

$$\text{Where } \Sigma = Eu_tu_t' = \Gamma(0),\ \ \Gamma(l) = Eu_tu_{t-l},\ \ \Gamma(l) = \Gamma^T(-l)$$

$$Var(S_n) \longrightarrow \sum_{l=-\infty}^{\infty}\Gamma(l) := \Omega$$

Estimated by Newey-West Heteroskedasticity and Autocorrelation Consistent (HAC) estimator. Newey-West HAC generalizes Huber-White Heteroscedastic consistent robust standard errors to allow for serial correlation.

$$\hat{\Omega} := \sum_{l=-M}^{M}\left(1-\frac{|l|}{M+1}\right)\hat{\Gamma}(l)$$

$$\text{Where } \hat{\Gamma}(l) := \frac{1}{n}\sum_{t=l+1}^{n}u_tu_{t-l}'$$

- The Newey-West Heteroscedasticity and Autocorrelation-Consistent Covariance Matrix extends the CLT to parameter estimates.
- If the true process is AR(p) with white noise or i.i.d. et, then we only need the conventional nonrobust standard errors.
- If we don't think et is white noise or i.i.d, but only believe that $E[y_t|y_{t-1},\ldots,y_{t-p}] = \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p}$ then we can use Huber-White HC robust std errors. (assumes correct specification, or "dynamic completeness")
- Without dynamic completeness, need to use Newey-West HAC std errors.

8.2 Moving average model and autoregressive model

Any stationary y_t can be represented as an MA(∞) series

$$\begin{aligned} y_t &= \mu + \Sigma_{j=1}^{\infty}\theta_j e_{t-j}, \text{ where } e_t \text{ is white noise} \\ &= \mu + \Sigma_{j=1}^{\infty}\theta_j L^j e_t, \text{ where } L^i e_t = e_{t-i} \\ &= \mu + \theta(L)e_t, \text{ where } \theta(L) = \theta_0 + \theta_1 L^1 + \ldots \end{aligned}$$

Under most conditions, y_t can also be represented as an AR(∞) seriers

$$\begin{aligned} y_t &= \mu_a + \Sigma_{j=1}^{\infty}a_t y_{t-j} + e_t = \mu_a + \Sigma_{j=1}^{\infty}a_j L^j y_t + e_t, \text{ where } L^i y_t = y_{t-i} \\ &= \mu_a + a(L)y_t + e_t, \text{ where } \theta(L) = \theta_0 + \theta_1 L^1 + \ldots \end{aligned}$$

These are related through a simple mapping

- $a(l) = \theta(L)^{-1}$
- $u_a = a(1)\mu$
- Solving for $\theta(l), a(l)$ can be done with recursive formulas. See p224 of lecture notes for inversion approach.

8.2.1 Moments of MA(q) and MAM(1)

Moment	**MA(q)**	**MA(1)**
$E[y_t]$	μ	μ
$Var(y_t)$	$\left(\sum_{j=0}^{q}\theta_j^2\right)\sigma^2$	$(1+\theta^2)\sigma^2$
$\gamma(k), k \le q$	$\left(\sum_{j=0}^{q-k}\theta_{j+k}\theta_j\right)\sigma^2$	$\theta\sigma^2$
$\gamma(k), k > q$	0	0
$\rho(k), k \le q$	$\gamma(k)/Var(y_t)$	$\gamma(k)/Var(y_t)$
$\rho(k), k > q$	0	0

8.3 Partial autocorrelation function

- Partial autocorrelation function (pacf) is related to but differs from autocorrelation function.
- pacf(k) is defined as coefficient on y_{t-k} of a population regression of y_t on $y_{t-1}, \ldots, y_{t-k}$
- Only for k = 1, pacf(1) = acf(1). When k ¿ 1, pacf(1) $\neq$ acf(1)
- ACF and PACF functions are collectively used to differentiate between MA, AR, and ARMA processes.
  - AR(p): geometric decay in acf(k), complete drop off in pacf(k) after p
  - MA(p): complete drop off in acf(k) after p, geometric decay in pacf(k)