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

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

2.2 Estimation

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

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

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

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: $X^T X$ 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 $\tilde{\beta} = \hat{\beta} + W(I - X(X^T X)^{-1} X^T)y$
- Note $Cov(\hat{\beta}, W(I - X(X^T X)^{-1} X^T)y) = 0$
- 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$\beta \quad := argmin_b E[(y - x^T b)^2]$ 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 $E[y \mid x]$

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

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:

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

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.

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

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 $\theta_0 = argmaxQ_0(\theta)$. 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

$$\begin{aligned}
\sum_i s_{\hat{\theta}}(z_i) =& 0 \text{ where } s_\theta = \partial/\partial\theta \log f_\theta(z)\\
s_{\hat{\theta}}(z_i) &\approxeq s_{\theta_0}(z_i) + \partial/\partial\theta s_{\theta_0}(z_i)(\hat{\theta} - \theta_0)\\
s_{\hat{\theta}}(z_i) &= s_{\theta_0}(z_i) + \partial/\partial\theta s_{\overline{\theta}}(z_i)(\hat{\theta} - \theta_0) \text{ by mean-value theorem for } \left\|\overline{\theta} - \theta_0\right\|_x \leq \left\|\hat{\theta} - \theta_0\right\|_x\\
0 = \sum_i s_{\hat{\theta}}(z_i) &= \sum_i s_{\theta_0}(z_i) + \sum_i \partial/\partial\theta s_{\overline{\theta}}(z_i)(\hat{\theta} - \theta_0)\\
\sqrt{n}(\hat{\theta} - \theta_0) = \left[-\frac{1}{n}\sum_i \partial/\partial\theta s_{\overline{\theta}}(z_i)\right]^{-1} & \frac{1}{\sqrt{n}} \sum_i s_{\theta_0}(z_i) \text{ with}\\
\left[-\frac{1}{n}\sum_i \partial/\partial\theta s_{\overline{\theta}}(z_i)\right]^{-1} & \xrightarrow{p} E\left[\frac{\partial s_{\theta_0}(z)}{\partial\theta}\right]^{-1},\ \ \frac{1}{\sqrt{n}}\sum_i s_{\theta_0}(z_i) \xrightarrow{d} N(0, Var(s_{\theta_0}(z)))\\
\text{so } \sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, H^{-1}JH^{-1}) & \text{ where } H = E\left[\frac{\partial s_{\theta_0}(z)}{\partial\theta}\right] \text{ and } J = Var(s_{\theta_0}(z)) = E[s_{\theta_0}(z) z_{\theta_0}(z)^T]
\end{aligned}$$

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

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

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

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

- 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 $H_0 : \quad 1/(\beta_L + \beta_K) - 1 = 0$ could produce a different result in finite samples than $H_0 : \quad \beta_L + \beta_K - 1 = 0$, even though it's the same hypothesis

3.7.2 Wald test

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

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

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

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

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

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

Rejecting the null hypothesis when

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

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

$$\hat{\theta} = argmax_{\theta \in \Theta} \hat{Q}_n(\theta) \text{ the unrestricted estimator}$$
$$\tilde{\theta} = argmax_{\theta \in \Theta} \hat{Q}_n(\theta) \; s.t. \; a(\theta) = 0 \text{ the restricted estimator}$$
$$H_0 : 2(\log L_n(\hat{\theta}) - \log L_n(\tilde{\theta})) \xrightarrow{d} \chi_r^2$$

Rejecting the null hypothesis when

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

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

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

Rejecting the null hypothesis when

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

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.

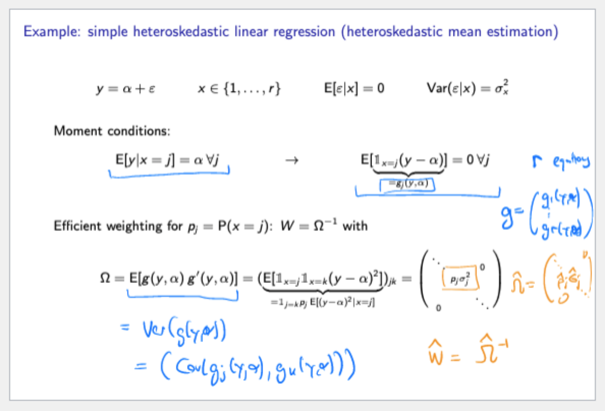
4.3 MLE as a special case

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

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

4.4 Example set-up

Set up for simple heteroskedastic linear regression



Example: simple heteroskedastic linear regression (heteroskedastic mean estimation)

$$y = \alpha + \varepsilon \qquad x \in \{1, \dots, r\} \qquad \mathrm{E}[\varepsilon|x] = 0 \qquad \mathrm{Var}(\varepsilon|x) = \sigma_x^2$$

Moment conditions:

$$\underbrace{\mathrm{E}[y|x=j] = \alpha \,\forall j}_{} \quad \rightarrow \quad \underbrace{\mathrm{E}[\underbrace{\mathbb{1}_{x=j}(y-\alpha)}_{\boxed{=g_j(y,\alpha)}}] = 0 \,\forall j}_{} \quad r \text{ equations}$$

$$g = \begin{pmatrix} g_1(y,\alpha) \\ \vdots \\ g_r(y,\alpha) \end{pmatrix}$$

Efficient weighting for $p_j = \mathrm{P}(x=j)$: $W = \Omega^{-1}$ with

$$\Omega = \underbrace{\mathrm{E}[g(y,\alpha)\, g'(y,\alpha)]}_{} = (\underbrace{\mathrm{E}[\mathbb{1}_{x=j}\mathbb{1}_{x=k}(y-\alpha)^2]}_{=\mathbb{1}_{j=k}p_j\,\mathrm{E}[(y-\alpha)^2|x=j]})_{jk} = \begin{pmatrix} \ddots & & 0 \\ & \boxed{p_j\sigma_j^2} & \\ 0 & & \ddots \end{pmatrix} \quad \hat\Omega = \begin{pmatrix} \ddots & & 0 \\ & \hat p_j \hat\sigma_j^2 & \\ 0 & & \ddots \end{pmatrix}$$

$$= \mathrm{Var}(g(y,\alpha))$$

$$= (\mathrm{Cov}(g_j(y,\alpha), g_k(y,\alpha)))$$

$$\hat W = \hat\Omega^{-1}$$

Figure 1: Example of simple heteroskedastic linear regression