Econ271: Econometrics II, linear regression

Erich Trieschman

2023 Winter quarter class notes

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: $Var(\tilde{\beta}) - Var(\hat{\beta}) \preceq 0 \Longrightarrow Var(\tilde{\beta}) - Var(\hat{\beta}) \in S_{++}$
- Note $\tilde{\beta} = Wy \Longrightarrow WX = I$ since $E[\tilde{\beta} \mid X] = \beta \Longrightarrow WX\beta = \beta$
- 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 $S \in S_{++}$

2.5.2 Cramer-Rao lower bound

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

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

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

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

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

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

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

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

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 Misspecification and QMLE

3.6 Tests

4 Generalized method of moments (GMM)

5 Bayesian regression

6 Machine learning