Partial Specification for Linear Regression and Robust Standard Errors

Last time:

- Parametric confidence regions (e.g. Wald's test) can give wrong answer
- Fix: Bootstrap
  - non-parametric method that generates more robust uncertainty estimates

- <u>This time:</u>
  - derive algebraic equations that work even when model is false, as long as certain "orthogonality conditions" hold

linear regression
↳ Gaussian errors + same σ^2 $\forall$ data points
→ relax this
"partial specification"

<u>Setup</u> $(x_1, y_1), \ldots, (x_n, y_n) \sim P$

$$Y_i \mid X_i \sim N(\beta^T X_i, \sigma^2)$$

$Y_i = \beta^T X_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$

Estimate β:

$$\hat{\beta} = \underset{\hat{\beta}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}^T x_i)^2$$

$$(X^TX)^{-1}X^Ty = \left(\tfrac{1}{n}X^TX\right)^{-1}\left(\tfrac{1}{n}X^Ty\right)$$

$$\hat{\beta} = \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right)^{-1}}_{\text{"}\mathbb{E}[xx^T]\text{"}} \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right)}_{\text{"}\mathbb{E}[xy]\text{"}}$$

$$S = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$$

$$= S^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i(\beta^T x_i + \varepsilon_i)\right)$$

$$= S^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\beta + x_i\varepsilon_i\right)$$

$$= S^{-1}\left(S\beta + \frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i\right)$$

$$= \beta + S^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i\right)$$

$$\Rightarrow \hat{\beta} - \beta = S^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i\right)$$

Gaussian

$$\mathbb{E}[\hat{\beta} - \beta \mid x_1, \dots, x_n] = 0 \quad \text{b/c } \mathbb{E}[\varepsilon_i \mid x_1, \dots, x_n] = 0 \ \forall i$$

$$\begin{aligned}
\mathrm{Cov}[\hat{\beta} - \beta \mid x_1, \dots, x_n] &= S^{-1}\, \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i\right) \cdot S^{-1} \\
&= \frac{1}{n^2} S^{-1} \sum_{i=1}^{n} \mathrm{Cov}(x_i \varepsilon_i)\, S^{-1} \\
&= \frac{1}{n^2} S^{-1} \cdot \Big(\sum_{i=1}^{n} \sigma^2 \underbrace{x_i x_i^T}_{n \cdot S}\Big) \cdot S^{-1} \\
&= \frac{\sigma^2}{n} \cdot S^{-1} \cdot S \cdot S^{-1} = \frac{\sigma^2}{n} \cdot S^{-1}
\end{aligned}$$

Upshot: $\hat{\beta} - \beta$ is Gaussian w/ cov. $\frac{\sigma^2}{n} \cdot S^{-1}$

(at least assuming ε_i is actually Gaussian)

$\Rightarrow$ standard error for β_i:

$$\sigma \sqrt{(S^{-1})_{ii} / n}$$

(non-robust) standard error get by default in most software packages

Assumptions: $\mathbb{E}[\varepsilon_i \mid x_i] = 0$

$\mathbb{E}[\varepsilon_i^2 \mid x_i] = \sigma^2$ for all i

Robust standard errors: $\varepsilon_i = y_i - \beta^T x_i$

$$\hat{\beta} - \beta = S^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i\right)$$

$$\hat{\beta} - \beta = S^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i\right)$$

Assume (x_i, y_i) are independent across i

$$\mathbb{E}\left[\hat{\beta} - \beta \mid x_1, \dots, x_n\right]$$

$$= S^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i \,\mathbb{E}[\varepsilon_i \mid x_i]\right)$$

$= 0$ assuming $\mathbb{E}[\varepsilon_i \mid x_i] = 0 \;\; \forall i$

Just need $\sum_{i=1}^{n} x_i \,\mathbb{E}[\varepsilon_i \mid x_i] = 0$ $\rightarrow 0$ if $\mathbb{E}[x\varepsilon] = 0$

$\rightarrow$ "$\mathbb{E}[x \cdot \varepsilon] = 0$"

"signal uncorrelated w/ noise"

Moral. Linear regression "works" as long as signal X and noise Z are uncorrelated (orthogonal).

Variance of $\hat{\beta}-\beta$.

(X_i, Y_i) indpt. across i

$Z_i = Y_i - \beta^T X_i$

$\Rightarrow Z_i$ indpt. across i

$$\hat{\beta}-\beta = S^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Z_i\right)$$

$$\mathrm{Cov}(\hat{\beta}-\beta \mid X_1,\dots,X_n)$$

$$= \frac{1}{n^2} S^{-1}\left(\sum_{i,j=1}^{n} \mathrm{Cov}(X_i Z_i, X_j Z_j \mid X_1,\dots,X_n)\right) S^{-1}$$

$$= \frac{1}{n^2} S^{-1}\left(\sum_{i,j=1}^{n} X_i X_j^T \mathrm{Cov}(Z_i, Z_j \mid X_i, X_j)\right) S^{-1}$$

$$= \frac{1}{n^2} S^{-1}\left(\sum_{i=1}^{n} X_i X_i^T \mathrm{Var}[Z_i \mid X_i]\right) S^{-1}$$

Define $\Omega = \frac{1}{n}\sum_{i=1}^{n} X_i \mathrm{Var}[Z_i \mid X_i] X_i^T$

"$\Omega = E[XZX^T]$" (recall TV-robust linear regression)

$$= \boxed{\frac{1}{n} S^{-1} \Omega S^{-1}} \longrightarrow \text{before: } \frac{\sigma^2}{n} S^{-1}$$

Interpretation: $\mathrm{Var}[\varepsilon_i \mid x_i] \preceq \sigma^2$

$$\Rightarrow \Omega \preceq \sigma^2 \cdot S$$

$$\Rightarrow \frac{1}{n} S^{-1} \Omega S^{-1} \preceq \frac{\sigma^2}{n} S^{-1}$$

Can also estimate Ω:

$$\mathrm{Var}[\varepsilon_i \mid x_i] = \mathbb{E}[\varepsilon_i^2 \mid x_i]$$

$$= \mathbb{E}[(y_i - \beta^T x_i)^2 \mid x_i]$$

$$\approx \underbrace{(y_i - \hat{\beta}^T x_i)}_{\triangleq u_i}{}^2 = u_i^2$$

$\sum u_i^2 \leq \sum \varepsilon_i^2$

$$\hat{\Omega}_n = \frac{1}{n} \sum_{i=1}^{n} x_i u_i^2 x_i^T$$

robust standard estimate using $\frac{1}{n} S^{-1} \hat{\Omega}_n S^{-1}$

robust std error for β_i: $\sqrt{\frac{1}{n} (S^{-1} \hat{\Omega}_n S^{-1})_{ii}}$.

"heteroskedasticity-consistent standard error"

Comments: u_i downward-biased b/c $\hat{\beta}$ can overfit

- Simplest correction: $\frac{n}{n-d}\hat{\Omega}_n$
- Fancier correction: using jackknife

$$u_i' = u_i / (1 - k_i) \approx \frac{n}{n-d} u_i$$

$$k_i = \frac{1}{n} x_i^T S^{-1} x_i$$

$$\frac{1}{n}\sum_{i=1}^n k_i = \frac{1}{n^2}\sum_{i=1}^n x_i^T S^{-1} x_i$$

$$= \left\langle \frac{1}{n^2}\sum_{i=1}^n x_i x_i^T, S^{-1} \right\rangle$$

$$= \frac{1}{n}\langle S, S^{-1} \rangle = tr(I) = \frac{d}{n}$$

2nd mom

$$\Omega_n' = \frac{1}{n}\sum_{i=1}^n x_i (u_i')^2 x_i^T$$

"\zeta" ζ ζ \ x_i

1st mom

$$\zeta = \frac{1}{n}\sum_{i=1}^n x_i u_i'$$

$$\frac{1}{n} S^{-1}\left(\Omega_n' - \zeta\zeta^T\right) S^{-1}$$

<u>OOD error:</u> $(x_1, y_1), \dots, (x_n, y_n) \sim p$

$$\bar{x}_1, \dots, \bar{x}_m \sim \bar{p}$$

Goal. Estimate error on $\bar{p}$.

Assuming model is well-specified (on $\bar{p}$):

$$\mathbb{E}\left[(\bar{y}_i - \hat{\beta}^T \bar{x}_i)^2\right] = \sigma^2 + \boxed{\langle \hat{\beta} - \beta, \bar{x}_i \rangle^2}$$

$$\frac{1}{m} \sum_{j=1}^{m} \langle \hat{\beta} - \beta, \bar{x}_j \rangle^2$$

$$= (\hat{\beta} - \beta)^T \underbrace{\left(\frac{1}{m} \sum_{j=1}^{m} \bar{x}_j \bar{x}_j^T\right)}_{= \bar{S}} (\hat{\beta} - \beta)$$

$$= (\hat{\beta} - \beta)^T \bar{S} (\hat{\beta} - \beta) = \langle \bar{S}, (\hat{\beta} - \beta)(\hat{\beta} - \beta)^T \rangle$$

expected error (over $z_1, \ldots, z_n$)

$$= \langle \bar{S}, \mathbb{E}_z[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] \rangle$$

assuming Gaussian model is correct on both train + test

$$\langle \bar{S}, \frac{\sigma^2}{n} S^{-1} \rangle = \frac{\sigma^2}{n} \mathrm{tr}(\bar{S} S^{-1}).$$

Interpretation.

$\bar{\Sigma} = \Sigma \Rightarrow \frac{\sigma^2}{n} tr(I) = \sigma^2 \cdot \frac{d}{n}$.

Σ low-rank but $\bar{\Sigma}$ full-rank ; $\frac{\sigma^2}{n} tr(\bar{\Sigma}\Sigma^{-1}) = \infty$

Σ full-rank $\quad \Sigma = I \quad \frac{\sigma^2}{n} \cdot tr(\bar{\Sigma}\Sigma^{-1})$

$\bar{\Sigma}$ low-rank $\quad \bar{\Sigma} = \frac{d}{k} P_k \quad = \frac{\sigma^2}{n} \cdot tr\left(\frac{d}{k} P_k\right)$

$= \sigma^2 \cdot \frac{d}{n}$

$\Sigma = \bar{\Sigma} = \frac{d}{k} P_k \Rightarrow \sigma^2 \cdot \frac{k}{n}$

So far, assume model is well-specified on train + test.

What happens if not?

$\bar{Y}_i = \beta^T \bar{X}_i + \bar{Z}_i$

Error on $\bar{X}_1, \dots, \bar{X}_m$:

$$\left(\beta^T \bar{X}_i + \bar{Z}_i - \hat{\beta}^T \bar{X}_i \right)^2$$

$$= \left((\beta - \hat{\beta})^T \bar{X}_i + \bar{Z}_i \right)^2$$

$$\frac{1}{m} \sum_{j=1}^{m} \left((\beta - \hat{\beta})^T \bar{X}_j + \bar{Z}_j \right)^2 \qquad \sigma^2$$

$$= \underbrace{\frac{1}{m} \sum_{j=1}^{m} \left((\beta - \hat{\beta})^T \bar{X}_i \right)^2}_{\frac{\sigma^2}{n} \langle \bar{S}, S^{-1} \rangle} + \frac{1}{m} \sum_{j=1}^{m} \bar{Z}_j^2$$

$$+ \underbrace{\frac{2}{m} \sum_{j=1}^{m} \langle \beta - \hat{\beta}, \bar{X}_j \rangle \cdot \bar{Z}_j}_{-2 \langle \hat{\beta} - \beta, \frac{1}{m} \sum_{j=1}^{m} \bar{X}_j \mathbb{E}[\bar{Z}_j | \bar{X}_j] \rangle}$$

$$\frac{\sigma^2}{n} \langle \bar{S}, S^{-1} \rangle$$

$$\Downarrow$$

$$\frac{1}{n} \langle \bar{S}, S^{-1} \Omega S^{-1} \rangle$$

$$-2 \left\langle \hat{\beta} - \beta, \underbrace{\frac{1}{m} \sum_{j=1}^{m} \bar{X}_j \mathbb{E}[\bar{Z}_j | \bar{X}_j]}_{\bar{b}} \right\rangle$$

$$\mathbb{E}[\hat\beta - \beta \mid x_1, \dots, x_n] = 0$$

$$\Downarrow$$

$$\mathbb{E}[\hat\beta - \beta] = S^{-1}\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} x_i \,\mathbb{E}[z_i \mid x]\right)}_{b}$$

$$= -2\langle S^{-1} b, \bar{b} \rangle$$

$$= \boxed{-2\,\bar{b}^T S^{-1} b}$$

Take β such that $b=0$ on train dist^n.

↳ Then β is minimizer of signed error on P

$\bar{z}_j = \bar{Y}_j - \beta^T \bar{X}_j$ ← If linear model is correct for P but not $\bar{P}$, $\bar{z}_j$ might be quite large.

Partial specification:

- Robust standard errors for OLS
- Clustering

- Econometrics: "generalized method of moments"

2014 Nobel Prize in Economics

Lars Hansen

"Uncertainty Inside and Outside of Economic Models"