

Linear Regression

Kosuke Imai

Princeton University

POL572 Quantitative Analysis II
Spring 2012

The Model and Interpretation

- The model: $Y_i = \alpha + \beta T_i + \epsilon_i$ where $\mathbb{E}(\epsilon_i) = 0$
- Potential outcomes:

$$Y_i(T_i) = \alpha + \beta T_i + \epsilon_i$$

- **Constant additive unit causal effect:** $Y_i(1) - Y_i(0) = \beta$ for all i
- $\alpha = \mathbb{E}(Y_i(0))$
- A general model:

$$Y_i(T_i) = \alpha + \beta T_i + \epsilon_i(T_i) \quad \text{where} \quad \mathbb{E}(\epsilon_i(t)) = 0$$

- $Y_i(1) - Y_i(0) = \beta + \epsilon_i(1) - \epsilon_i(0)$
- $\beta = \mathbb{E}(Y_i(1) - Y_i(0))$
- $\alpha = \mathbb{E}(Y_i(0))$ as before

Assumptions and Interpretation

- **Exogeneity:** $\mathbb{E}(\epsilon_i | T) = \mathbb{E}(\epsilon_i) = 0$
 - ① Orthogonality: $\mathbb{E}(\epsilon_i T_j) = 0$
 - ② Zero correlation: $\text{Cov}(\epsilon_i, T_j) = 0$
- **Homoskedasticity:** $\mathbb{V}(\epsilon_i | T) = \mathbb{V}(\epsilon_i) = \sigma^2$
- **Randomization:**
 - $(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i$ for all i
 - $\mathbb{E}(Y_i(t) | T_i) = \mathbb{E}(Y_i(t)) \iff \mathbb{E}(\epsilon_i | T_i) = \mathbb{E}(\epsilon_i)$
 - $\mathbb{E}(Y_i(t)) = \mathbb{E}(Y_i | T_i = t) = \alpha + \beta t$
- **Random sampling:**
 - $(Y_i(1), Y_i(0)) \perp\!\!\!\perp (Y_j(1), Y_j(0))$ for any $i \neq j$
 - $\epsilon_i \perp\!\!\!\perp \epsilon_j$
- **Equal variance:**
 - $\mathbb{V}(Y_i(t) | T_i) = \mathbb{V}(Y_i(t)) = \sigma^2$

Least Squares Estimation

- Model parameters: (α, β)
- Estimates: $(\hat{\alpha}, \hat{\beta})$
- Predicted (fitted) value: $\hat{Y}_i = \hat{\alpha} + \hat{\beta}T_i$
- Residual: $\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta}T_i$
- Minimize the **sum of squared residuals**:

$$SSR = \sum_{i=1}^n \hat{\epsilon}_i^2$$

which yields

$$\hat{\alpha} = \bar{Y}_n - \hat{\beta}\bar{T}_n \quad \text{and} \quad \hat{\beta} = \frac{\sum_{i=1}^n (Y_i - \bar{Y}_n)(T_i - \bar{T}_n)}{\sum_{i=1}^n (T_i - \bar{T}_n)^2}$$

Unbiasedness of Least Squares Estimator

- When T_i is binary, $\hat{\beta} =$ Difference-in-Means estimator!
- So, $\hat{\beta}$ is unbiased from the design-based perspective
- Model-based estimation error:

$$\hat{\beta} - \beta = \frac{\sum_{i=1}^n (T_i - \bar{T}_n) \epsilon_i}{\sum_{i=1}^n (T_i - \bar{T}_n)^2}$$

- Thus, the exogeneity assumption implies,

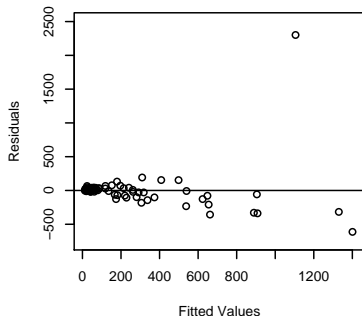
$$\mathbb{E}(\hat{\beta}) - \beta = \mathbb{E}\{\mathbb{E}(\hat{\beta} - \beta \mid \mathbf{T})\} = 0$$

- Similarly, $\hat{\alpha} - \alpha = \bar{\epsilon}_n - (\hat{\beta} - \beta)\bar{T}_n$
- Thus, $\mathbb{E}(\hat{\alpha}) - \alpha = 0$

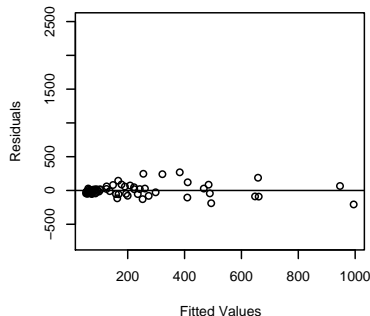
Residuals

- Estimated error term
- Zero mean: $\sum_{i=1}^n \hat{\epsilon}_i = 0$
- Orthogonality: $\hat{\epsilon} \cdot T = \sum_{i=1}^n \hat{\epsilon}_i T_i = 0$
- (Sample and population) correlation between $\hat{\epsilon}_i$ and T_i is 0
- Does not imply $\mathbb{E}(\epsilon \cdot T) = 0$ or $\text{Cor}(\epsilon_i, T_i)$
- Residual plot ($Y = 2000$ Buchanan votes, $T = 1996$ Perot votes):

With Palm Beach



Without Palm Beach



The Coefficient of Determination

- How much variation in Y does the model explain?
- **Total Sum of Squares:**

$$TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

- The coefficient of determination or R^2 :

$$R^2 \equiv \frac{TSS - SSR}{TSS}$$

- $0 \leq R^2 \leq 1$
- $R^2 = 0$ when $\hat{\beta} = 0$
- $R^2 = 1$ when $Y_i = \hat{Y}_i$ for all i
- Example: 0.85 (without PB) vs. 0.51 (with PB)

Model-Based Variance and Its Estimator

- The homoskedasticity assumption implies

$$\mathbb{V}(\hat{\beta} | T) = \frac{\sigma^2}{\sum_{i=1}^n (T_i - \bar{T}_n)^2}$$

- Standard model-based (conditional) variance estimator for $\hat{\beta}$:

$$\widehat{\mathbb{V}}(\hat{\beta} | T) = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (T_i - \bar{T}_n)^2} \quad \text{where} \quad \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\epsilon}_i^2$$

- (Conditionally) Unbiased: $\mathbb{E}(\hat{\sigma}^2 | T) = \sigma^2$ implies

$$\mathbb{E}(\widehat{\mathbb{V}}(\hat{\beta} | T) | T) = \mathbb{V}(\hat{\beta} | T)$$

- (Unconditionally) Unbiased: $\mathbb{V}(\mathbb{E}(\hat{\beta} | T)) = 0$ implies

$$\mathbb{V}(\hat{\beta}) = \mathbb{E}(\mathbb{V}(\hat{\beta} | T)) = \mathbb{E}(\mathbb{E}(\widehat{\mathbb{V}}(\hat{\beta} | T) | T)) = \mathbb{E}(\widehat{\mathbb{V}}(\hat{\beta} | T))$$

Model-Based Prediction

- (Estimated) **Expected value** for $T_i = t$:

$$\widehat{Y}(t) = \mathbb{E}(Y \mid T_i = t) = \hat{\alpha} + \hat{\beta}t$$

- **Predicted value** for $T_i = t$:

$$Y(t) = \widehat{Y}(t) + \epsilon_i$$

- Variance (point estimate is still $\widehat{Y}(t)$):

$$\begin{aligned}\mathbb{V}(Y(t) \mid T) &= \mathbb{V}(\hat{\alpha} \mid T) + \mathbb{V}(\hat{\beta} \mid T)t^2 + 2t\text{Cov}(\hat{\alpha}, \hat{\beta} \mid T) + \sigma^2 \\ &= \sigma^2 \left(\frac{\sum_{i=1}^n (T_i - t)^2}{n \sum_{i=1}^n (T_i - \bar{T})^2} + 1 \right)\end{aligned}$$

$$\text{where } \mathbb{V}(\hat{\alpha} \mid T) = \frac{\sigma^2 \sum_{i=1}^n T_i^2}{n \sum_{i=1}^n (T_i - \bar{T})^2} \text{ and } \text{Cov}(\hat{\alpha}, \hat{\beta} \mid T) = -\frac{\sigma^2 \bar{T}}{\sum_{i=1}^n (T_i - \bar{T})^2}$$

Model-Based Finite Sample Inference

- Standard error: $s.e. = \sqrt{\widehat{\text{V}}(\hat{\beta} | T)}$
- Sampling distribution:

$$\hat{\beta} - \beta \sim \mathcal{N}\left(0, \frac{\sigma^2}{\sum_{i=1}^n (T_i - \bar{T}_n)^2}\right)$$

- Inference under $\epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$:

$$\frac{\hat{\beta} - \beta}{s.e.} = \frac{\hat{\beta} - \beta}{\underbrace{\sigma / \sqrt{\sum_{i=1}^n (T_i - \bar{T}_n)^2}}_{\sim \mathcal{N}(0,1)}} / \sqrt{\underbrace{\frac{(n-2)\hat{\sigma}^2}{\sigma^2}}_{\sim \chi_{n-2}^2} \frac{1}{n-2}} \sim t_{n-2}$$

Model-Based Asymptotic Inference

- Consistency: $\hat{\beta} \xrightarrow{p} \beta$
- Asymptotic distribution and inference:

$$\begin{aligned}\sqrt{n}(\hat{\beta} - \beta) &= \underbrace{\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n (T_i - \mathbb{E}(T_i)) \epsilon_i + (\mathbb{E}(T_i) - \bar{T}_n) \frac{1}{n} \sum_{i=1}^n \epsilon_i \right)}_{\xrightarrow{d} \mathcal{N}(0, \sigma^2 \mathbb{V}(T_i))} \\ &\quad \times \underbrace{\left(\frac{1}{n} \sum_{i=1}^n (T_i - \bar{T}_n)^2 \right)^{-1}}_{\xrightarrow{p} \mathbb{V}(T_i)^{-1}} \\ &\xrightarrow{d} \mathcal{N} \left(0, \frac{\sigma^2}{\mathbb{V}(T_i)} \right). \\ \frac{\hat{\beta} - \beta}{\text{s.e.}} &\xrightarrow{d} \mathcal{N}(0, 1).\end{aligned}$$

Bias of Model-Based Variance

- The design-based perspective: use Neyman's exact variance
- What is the bias of the model-based variance estimator?
- Finite sample bias:

$$\begin{aligned}\text{Bias} &= \mathbb{E} \left(\frac{\hat{\sigma}^2}{\sum_{i=1}^n (T_i - \bar{T}_n)^2} \right) - \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0} \right) \\ &= \frac{(n_1 - n_0)(n - 1)}{n_1 n_0 (n - 2)} (\sigma_1^2 - \sigma_0^2)\end{aligned}$$

- Bias is zero when $n_1 = n_0$ or $\sigma_1^2 = \sigma_0^2$
- In general, bias can be negative or positive and does not asymptotically vanish

Robust Standard Error

- Suppose $\text{Var}(\epsilon_i | T) = \sigma^2(T_i) \neq \sigma^2$
- **Heteroskedasticity consistent robust variance estimator** (more later):

$$\text{Var}(\widehat{(\hat{\alpha}, \hat{\beta})} | T) = \left(\sum_{i=1}^n x_i x_i^\top \right)^{-1} \left(\sum_{i=1}^n \hat{\epsilon}_i^2 x_i x_i^\top \right) \left(\sum_{i=1}^n x_i x_i^\top \right)^{-1}$$

where in this case $x_i = (1, T_i)$ is a column vector of length 2

- Model-based justification: asymptotically valid in the presence of heteroskedastic errors
- Design-based evaluation:

$$\text{Finite Sample Bias} = - \left(\frac{\sigma_1^2}{n_1^2} + \frac{\sigma_0^2}{n_0^2} \right)$$

- Bias vanishes asymptotically

Cluster Randomized Experiments

- Units: $i = 1, 2, \dots, n_j$
- Clusters of units: $j = 1, 2, \dots, m$
- Treatment at cluster level: $T_j \in \{0, 1\}$
- Outcome: $Y_{ij} = Y_{ij}(T_j)$
- Random assignment: $(Y_{ij}(1), Y_{ij}(0)) \perp\!\!\!\perp T_j$
- **No interference** between units of different clusters
- Possible interference between units of the same cluster
- Estimands at unit level:

$$\text{SATE} \equiv \frac{1}{\sum_{j=1}^m n_j} \sum_{j=1}^m \sum_{i=1}^{n_j} (Y_{ij}(1) - Y_{ij}(0))$$

$$\text{PATE} \equiv \mathbb{E}(Y_{ij}(1) - Y_{ij}(0))$$

- Random sampling of clusters and units

Design-Based Inference

- For simplicity, assume equal cluster size, i.e., $n_j = n$ for all j
- The difference-in-means estimator:

$$\hat{\tau} \equiv \frac{1}{m_1} \sum_{j=1}^m T_j \bar{Y}_j - \frac{1}{m_0} \sum_{j=1}^m (1 - T_j) \bar{Y}_j$$

where $\bar{Y}_j \equiv \sum_{i=1}^{n_j} Y_{ij} / n_j$

- Easy to show $\mathbb{E}(\hat{\tau} \mid \mathcal{O}) = \text{SATE}$ and thus $\mathbb{E}(\hat{\tau}) = \text{PATE}$
- Exact population variance:

$$\text{Var}(\hat{\tau}) = \frac{\text{Var}(\overline{Y_j(1)})}{m_1} + \frac{\text{Var}(\overline{Y_j(0)})}{m_0}$$

- **Intracluster correlation coefficient** ρ_t :

$$\text{Var}(\overline{Y_j(t)}) = \frac{\sigma_t^2}{n} \{1 + (n-1)\rho_t\} \leq \sigma_t^2$$

Cluster Standard Error

- **Cluster robust variance estimator:**

$$\text{Var}(\widehat{(\hat{\alpha}, \hat{\beta})} \mid T) = \left(\sum_{j=1}^m \mathbf{X}_j^\top \mathbf{X}_j \right)^{-1} \left(\sum_{j=1}^m \mathbf{X}_j^\top \hat{\epsilon}_j \hat{\epsilon}_j^\top \mathbf{X}_j \right) \left(\sum_{j=1}^m \mathbf{X}_j^\top \mathbf{X}_j \right)^{-1}$$

where in this case $\mathbf{X}_j = [1 \ T_j]$ is an $n_j \times 2$ matrix and $\hat{\epsilon}_j = (\hat{\epsilon}_{1j}, \dots, \hat{\epsilon}_{n_j j})$ is a column vector of length n_j

- Design-based evaluation (assume $n_j = n$ for all j):

$$\text{Finite Sample Bias} = - \left(\frac{\mathbb{V}(\overline{Y_j(1)})}{m_1^2} + \frac{\mathbb{V}(\overline{Y_j(0)})}{m_0^2} \right)$$

- Bias vanishes asymptotically as $m \rightarrow \infty$ with n fixed
- **Implication:** cluster standard errors by the unit of treatment assignment

Regression Discontinuity Design

- Idea: Find an arbitrary cutpoint c which determines the treatment assignment such that $T_i = \mathbf{1}\{X_i \geq c\}$
- Assumption: $\mathbb{E}(Y_i(t) | X_i = x)$ is continuous in x
- Estimand: $\mathbb{E}(Y_i(1) - Y_i(0) | X_i = c)$
- Regression modeling:

$$\mathbb{E}(Y_i(1) | X_i = c) = \lim_{x \downarrow c} \mathbb{E}(Y_i(1) | X_i = x) = \lim_{x \downarrow c} \mathbb{E}(Y_i | X_i = x)$$

$$\mathbb{E}(Y_i(0) | X_i = c) = \lim_{x \uparrow c} \mathbb{E}(Y_i(0) | X_i = x) = \lim_{x \uparrow c} \mathbb{E}(Y_i | X_i = x)$$

- Advantage: internal validity
- Disadvantage: external validity
- Make sure nothing else is going on at $X_i = c$

Close Elections as RD Design (Lee)

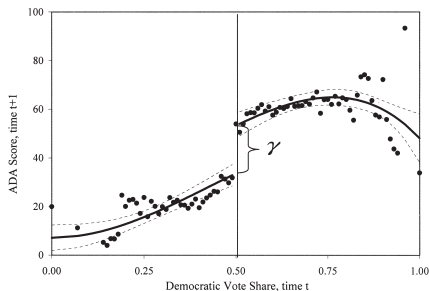


FIGURE I
Total Effect of Initial Win on Future ADA Scores: γ

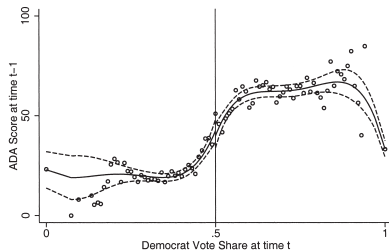


FIGURE V
Specification Test: Similarity of Historical Voting Patterns between Bare Democrat and Republican Districts

- **Placebo test** for natural experiments
- What is a good placebo?
 - 1 expected not to have any effect
 - 2 closely related to outcome of interest

The Model and Interpretation

- Scalar representation: $Y_i = \beta_1 + \beta_2 X_{i2} + \dots + \beta_K X_{iK} + \epsilon_i$
- Vector representation: $Y_i = X_i^\top \beta + \epsilon_i$
- Matrix representation: $Y = X\beta + \epsilon$
- The first column of X is a vector of 1's
- X_i^\top is the i th row of X

- The model: $Y_i = \alpha + \beta T_i + X_i^\top \gamma + \epsilon_i$
- Potential outcomes: $Y_i(T_i) = \alpha + \beta T_i + X_i^\top \gamma + \epsilon_i$
- X_i is a vector of *pre-treatment* covariate
- **Post-treatment bias**: X_i shouldn't include post-treatment variables
- $\beta = Y_i(1) - Y_i(0)$
- A general model: $Y_i(T_i) = \alpha + \beta T_i + X_i^\top \gamma + \epsilon_i(T_i)$
- $\beta = \mathbb{E}(Y_i(1) - Y_i(0))$

Assumptions and Interpretation

- **Exogeneity:** $\mathbb{E}(\epsilon | X) = \mathbb{E}(\epsilon) = 0$
 - ① Conditional Expectation Function (CEF): $\mathbb{E}(Y | X) = X\beta$
 - ② Orthogonality: $\mathbb{E}(\epsilon_i X_{jk}) = 0$ for any i, j, k
 - ③ Zero correlation: $\text{Cov}(\epsilon_i, X_{jk}) = 0$
- **Homoskedasticity:** $\mathbb{V}(\epsilon | X) = \mathbb{V}(\epsilon) = \sigma^2 I_n$
 - ① spherical error variance
 - ② $\mathbb{V}(\epsilon_i | X) = \sigma^2$ and $\text{Cov}(\epsilon_i, \epsilon_j | X) = 0$
- **Ignorability:** $(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i | X_i$
 - $\mathbb{E}(Y_i(t) | T_i, X_i) = \mathbb{E}(Y_i(t) | X_i)$
 - $\mathbb{E}(Y_i(t) | X_i) = \mathbb{E}(Y_i | T_i = t, X_i) = \alpha + \beta t + X_i^\top \gamma$
 - $\mathbb{E}(\epsilon_i | T_i, X_i) = \mathbb{E}(\epsilon_i | X_i) \stackrel{?}{=} \mathbb{E}(\epsilon_i) = 0$
- **Random sampling:** $(Y_i(1), Y_i(0), T_i, X_i)$ is i.i.d.
- **Equal variance:** $\text{Var}(Y_i(t) | T_i, X_i) = \text{Var}(Y_i(t)) = \sigma^2$

Rank Condition

- Assumption: Rank of X is K (i.e., full column rank)
- Rank of a matrix = # of linearly independent columns (rows)
- Linear independence: $Xc = 0$ iff c is a column vector of 0s
- **No multicollinearity**: no perfect linear column dependence
- $n \geq K$

- A square matrix has full rank iff it is non-singular
- $Xc = b$ has a unique solution $c = X^{-1}b$
- X is of full column rank iff $X^T X$ is non-singular
- Suppose X has full rank: assume $X^T Xc = 0$ for a non-zero c
 $\implies c^T X^T Xc = 0 \implies \|Xc\|^2 = 0 \implies Xc = 0$, contradiction
- Suppose $X^T X$ has full rank: assume $Xc = 0$ for a non-zero c
 $\implies X^T Xc = 0$, contradiction

Least Squares Estimation

- Model parameters: β
- Estimates: $\hat{\beta}$
- Predicted (fitted) value: $\hat{Y} = X\hat{\beta}$
- Residual: $\hat{\epsilon} = Y - \hat{Y} = Y - X\hat{\beta}$
- Minimize the **sum of squared residuals**:

$$SSR = \|\hat{\epsilon}\|^2$$

which yields

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Geometry of Least Squares

- 1 Using $\hat{\beta} = \beta + (X^T X)^{-1} X^T \epsilon$, show the orthogonality $X^T \hat{\epsilon} = X^T (Y - X\hat{\beta}) = 0$. i.e., $x_k \cdot \hat{\epsilon} = 0$ for any column vector x_k of X
- 2 Using the orthogonality, show $SSR = \|Y - X\tilde{\beta}\|^2 = \|\hat{\epsilon} + X(\hat{\beta} - \tilde{\beta})\|^2 = \|\hat{\epsilon}\|^2 + \|X(\hat{\beta} - \tilde{\beta})\|^2$
- 3 SSR is minimized when $\tilde{\beta} = \hat{\beta}$

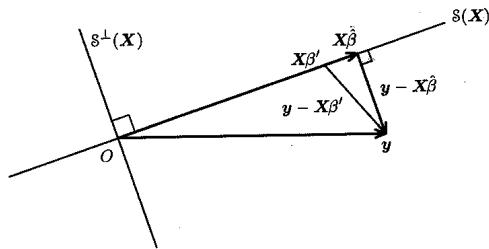


Figure 1.2 The projection of y onto $S(X)$

- $S(X)$: subspace of \mathfrak{R}^n spanned by columns of X
- $\hat{Y} = X\hat{\beta}$: projection of Y onto $S(X)$
- $\hat{\epsilon}$ is orthogonal to all columns of X and thus to $S(X)$

Derivation with Calculus

- 1 Vector calculus: Let a, b be column vectors and A be a matrix
 - 1 $\frac{\partial a^\top b}{\partial b} = a$
 - 2 $\frac{\partial Ab}{\partial b} = A^\top$
 - 3 $\frac{\partial b^\top Ab}{\partial b} = 2Ab$ when A is symmetric
- 2 $SSR = \|Y - X\hat{\beta}\|^2 = Y^\top Y - 2Y^\top X\hat{\beta} + \hat{\beta}^\top X^\top X\hat{\beta}$
- 3 First order condition:
 $\frac{\partial SSR}{\partial \hat{\beta}} = 0 \implies (X^\top X)\hat{\beta} = X^\top Y$ (normal equation)
- 4 Second order condition:
 $\frac{\partial^2 SSR}{\partial \hat{\beta} \partial \hat{\beta}^\top} = X^\top X \geq 0$ in a matrix sense
- 5 A square matrix A is **positive semi-definite** (non-negative definite) if A is symmetric and $c^\top A c \geq 0$ for any column vector c
 $X^\top X$ is symmetric and $c^\top X^\top X c = \|Xc\|^2 \geq 0$

Unbiasedness of Least Squares Estimator

- Recall $\hat{\beta} = \beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \epsilon$
- Conditional unbiasedness: $\mathbb{E}(\hat{\beta} \mid \mathbf{X}) = \beta$
- Unconditional unbiasedness: $\mathbb{E}(\hat{\beta}) = \beta$
- Conditional variance: $\mathbb{V}(\hat{\beta} \mid \mathbf{X}) = \mathbb{V}(\hat{\beta} - \beta \mid \mathbf{X}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$
- Unconditional variance: $\mathbb{V}(\hat{\beta}) = \mathbb{E}\{\mathbb{V}(\hat{\beta} - \beta \mid \mathbf{X})\} = \sigma^2 \mathbb{E}(\mathbf{X}^\top \mathbf{X})^{-1}$

Gauss-Markov Theorem

- Under exogeneity and homoskedasticity assumptions, $\hat{\beta}$ is the **Best Linear Unbiased Estimator**
- A linear estimator: $\tilde{\beta} = AY = \hat{\beta} + BY$ where $B = A - (X^T X)^{-1} X^T$
- $\tilde{\beta} = \{(X^T X)^{-1} X^T + B\} Y = \beta + \{(X^T X)^{-1} X^T + B\} \epsilon + BX\beta$
- $\mathbb{E}(\tilde{\beta} | X) = \mathbb{E}(\hat{\beta} | X) = \beta$ and exogeneity imply $BX = 0$

$$\begin{aligned}\mathbb{V}(\tilde{\beta} | X) &= \{(X^T X)^{-1} X^T + B\} \text{Var}(\epsilon | X) \{(X^T X)^{-1} X^T + B\}^T \\ &= \sigma^2 \{(X^T X)^{-1} + BB^T\} \\ &\geq \sigma^2 (X^T X)^{-1} \\ &= \mathbb{V}(\hat{\beta} | X)\end{aligned}$$

- Also, $\mathbb{V}(\tilde{\beta}) \geq \mathbb{V}(\hat{\beta})$ so long as $\mathbb{E}(\tilde{\beta} | X) = \beta$ holds
- Don't take it too seriously!: bias, nonlinearity, heteroskedasticity

More Geometry

- **Orthogonal projection matrix** or “Hat” matrix:

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = P_X Y$$

- P_X projects Y onto $\mathcal{S}(X)$ and thus $P_X X = X$
- $M_X \equiv I_n - P_X$ projects Y onto $\mathcal{S}^\perp(X)$: $M_X X = 0$
- **Orthogonal decomposition**: $Y = P_X Y + M_X Y$

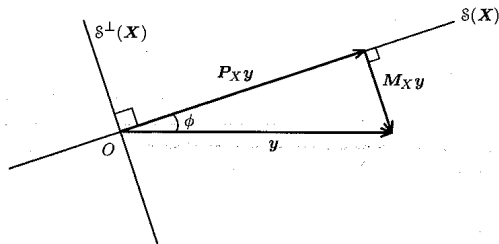


Figure 1.3 The orthogonal decomposition of y

- Symmetric: $P_X = P_X^T$ and $M_X = M_X^T$
- Idempotent: $P_X P_X = P_X$ and $M_X M_X = M_X$
- Annihilator:
 $M_X P_X = P_X M_X = 0$

Residuals

- $\hat{\epsilon} = Y - X\hat{\beta} = M_X Y$
- Orthogonality:
 - 1 $\hat{Y} \cdot \hat{\epsilon} = (P_X Y) \cdot M_X Y = 0$
 - 2 $x_k \cdot \hat{\epsilon} = (P_X x_k) \cdot M_X Y = 0$ for any column k
 - 3 More generally, $x \cdot \hat{\epsilon} = 0$ for any $x \in \mathcal{S}(X)$
- Zero mean: $\sum_{i=1}^n \hat{\epsilon}_i = 0$ since $x_1 = (1, \dots, 1) \in \mathcal{S}(X)$
- (Sample and population) correlation between $\hat{\epsilon}_i$ and x_{ik} is 0 for any column k
- Does not imply $\mathbb{E}(\epsilon | X) = 0$, $\mathbb{E}(x_k \cdot \epsilon) = 0$ or $\text{Cor}(\epsilon_i, x_{ik})$

The Coefficient of Determination

- The (*centered*) coefficient of determination:

$$R^2 \equiv 1 - \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\text{explained variance}}{\text{original variance}}$$

- Recall $Y = X\hat{\beta} + \hat{\epsilon}$, $\hat{Y} \cdot \hat{\epsilon} = 0$, and $\bar{Y}\mathbf{1} \cdot \hat{\epsilon} = 0$
- **Pythagoras' Theorem:**

$$\|Y - \bar{Y}\mathbf{1}\|^2 = \|X\hat{\beta} + \hat{\epsilon} - \bar{Y}\mathbf{1}\|^2 = \|X\hat{\beta} - \bar{Y}\mathbf{1}\|^2 + \|\hat{\epsilon}\|^2$$

- Note $\bar{Y} = \overline{X\hat{\beta}}$
- Thus, $\text{var}(Y_i) = \text{var}(X_i\hat{\beta}) + \text{var}(\hat{\epsilon}_i)$

Variance Estimation Under Homoskedasticity

- Under homoskedasticity, standard variance estimator is:

$$\widehat{\mathbb{V}(\hat{\beta} | \mathbf{X})} = \hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1} \quad \text{where} \quad \hat{\sigma}^2 = \frac{\|\hat{\epsilon}\|^2}{n - K}$$

- (Conditionally) Unbiased: $\mathbb{E}(\hat{\sigma}^2 | \mathbf{X}) = \sigma^2$ implies

$$\mathbb{E}(\widehat{\mathbb{V}(\hat{\beta} | \mathbf{X})} | \mathbf{X}) = \mathbb{V}(\hat{\beta} | \mathbf{X})$$

- (Unconditionally) Unbiased: $\mathbb{V}(\mathbb{E}(\hat{\beta} | \mathbf{X})) = 0$ implies

$$\mathbb{V}(\hat{\beta}) = \mathbb{E}(\mathbb{V}(\hat{\beta} | \mathbf{X})) = \mathbb{E}(\widehat{\mathbb{V}(\hat{\beta} | \mathbf{X})} | \mathbf{X}) = \mathbb{E}(\widehat{\mathbb{V}(\hat{\beta} | \mathbf{X})})$$

Proof that $\hat{\sigma}^2$ is Unbiased (Freedman, Theorem 4)

- Recall $\hat{\epsilon} = M_X Y = M_X(X\beta + \epsilon) = M_X \epsilon$
- Then $\|\hat{\epsilon}\|^2 = \|M_X \epsilon\|^2 = \epsilon^\top M_X \epsilon = \text{trace}(\epsilon^\top M_X \epsilon)$
- $\epsilon^\top M_X \epsilon = \sum_{i=1}^m \sum_{j=1}^m \epsilon_i \epsilon_j m_{ij}$ where m_{ij} is the (i, j) element of M_X
- Homoskedasticity: $\mathbb{E}(\epsilon_i \epsilon_j | X) = 0$ for $i \neq j$ and $\mathbb{E}(\epsilon_i^2 | X) = \sigma^2$
- $\mathbb{E}(\|\hat{\epsilon}\|^2 | X) = \sigma^2 \text{trace}(M_X)$
- Recall the following properties of trace operator:
 - 1 $\text{trace}(A + B) = \text{trace}(A) + \text{trace}(B)$
 - 2 $\text{trace}(AB) = \text{trace}(BA)$ where A is $m \times n$ and B is $n \times m$
- $\text{trace}(M_X) = \text{trace}(I_n) - \text{trace}\{X(X^\top X)^{-1}X^\top\} = n - \text{trace}(I_K) = n - K$

Model-Based Prediction

- (Estimated) **Expected value** for $X_i = x$:

$$\widehat{Y(x)} = \mathbb{E}(Y_i | \widehat{X_i} = x) = x^\top \widehat{\beta}$$

- **Predicted value** for $X_i = x$:

$$Y(x) = \widehat{Y(x)} + \epsilon_i$$

- Variance (point estimate is still $\widehat{Y(x)}$):

$$\begin{aligned} \mathbb{V}(Y(x) | X) &= x^\top \mathbb{V}(\widehat{\beta} | X) x + \sigma^2 \\ &= \sigma^2 \left\{ x^\top (X^\top X)^{-1} x + 1 \right\} \end{aligned}$$

Causal Inference with Interaction Terms

- A Model: $Y_i = \alpha + \beta T_i + \mathbf{X}_{1i}^\top \gamma_1 + \mathbf{X}_{2i}^\top \gamma_2 + T_i \mathbf{X}_{1i}^\top \delta + \epsilon_i$
- Average causal effect depends on *pre-treatment* covariates \mathbf{X}_{1i}
- Average causal effect when $\mathbf{X}_{1i} = \mathbf{x}$:

$$\beta + \mathbf{x}^\top \delta$$

- Variance: $\mathbb{V}(\hat{\beta} \mid T, \mathbf{X}) + \mathbf{x}^\top \mathbb{V}(\hat{\delta} \mid T, \mathbf{X}) \mathbf{x} + 2\mathbf{x}^\top \text{Cov}(\hat{\beta}, \hat{\delta} \mid T, \mathbf{X})$
- Difference in the average causal effects between the case with $\mathbf{X}_{1i} = \mathbf{x}^*$ and the case with $\mathbf{X}_{1i} = \mathbf{x}$:

$$(\mathbf{x}^* - \mathbf{x})^\top \delta$$

- Variance: $(\mathbf{x}^* - \mathbf{x})^\top \mathbb{V}(\hat{\delta} \mid T, \mathbf{X})(\mathbf{x}^* - \mathbf{x})$

Model-Based Finite Sample Inference

- Standard error for $\hat{\beta}_k$: $\text{s.e.} = \sqrt{\widehat{\text{V}}(\hat{\beta} | \mathbf{X})_{kk}}$
- Sampling distribution:

$$\hat{\beta}_k - \beta_k \sim \mathcal{N}\left(0, \sigma^2(\mathbf{X}^\top \mathbf{X})_{kk}^{-1}\right)$$

- Inference under $\epsilon_j \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$:

$$\frac{\hat{\beta}_k - \beta_k}{\text{s.e.}_k} = \frac{\hat{\beta}_k - \beta_k}{\underbrace{\sigma \sqrt{(\mathbf{X}^\top \mathbf{X})_{kk}^{-1}}}_{\sim \mathcal{N}(0,1)}} / \sqrt{\underbrace{\frac{(n-K)\hat{\sigma}^2}{\sigma^2}}_{\sim \chi_{n-K}^2} \frac{1}{n-K}} \sim t_{n-K}$$

Derivation (see Proposition 1.3 of Hayashi)

- 1 Note that if $x \sim \mathcal{N}(0, I_K)$, then $x^\top A x \sim \chi_\nu^2$ where $\nu = \text{rank}(A)$
- 2 If X is an idempotent matrix, then $\text{rank}(X) = \text{trace}(X)$
- 3 Recall $\hat{\epsilon} = M_X \epsilon$
- 4 Given X ,
$$(n - K)\hat{\sigma}^2/\sigma^2 = \|M_X \epsilon/\sigma\|^2 \sim \chi_{\text{rank}(M_X)}^2 = \chi_{\text{trace}(M_X)}^2 = \chi_{n-K}^2$$
- 5 Recall $\hat{\beta} = \beta + (X^\top X)^{-1} X^\top \epsilon$
- 6 $\text{Cov}(\hat{\beta}, \hat{\epsilon} | X) = \text{Cov}(\hat{\beta} - \beta, M_X \epsilon | X) = (X^\top X)^{-1} X^\top \mathbb{E}(\epsilon \epsilon^\top | X) M_X = 0$
- 7 **Theorem:** If $X \sim \mathcal{N}(\mu, \Sigma)$, then $AX \sim \mathcal{N}(A\mu, A\Sigma A^\top)$
- 8 Thus, $(\hat{\beta}, \hat{\epsilon})$ have a multivariate Normal distribution given X
- 9 $\hat{\beta}$ and $\hat{\epsilon}$ are independent given X

Testing Linear Null Hypothesis

- Null hypothesis $H_0 : A\beta = a$ where A is of full row rank
- Any linear restriction: $b_1\beta_1 + b_2\beta_2 + \cdots + b_K\beta_K = c$
- Any number of linearly independent such restrictions
- F -statistic:

$$\begin{aligned} F &\equiv \frac{(A\hat{\beta} - a)^\top \{A(X^\top X)^{-1}A^\top\}^{-1}(A\hat{\beta} - a)}{\hat{\sigma}^2 \text{rank}(A)} \\ &= \frac{\overbrace{(A\hat{\beta} - a)^\top \{\sigma^2 A(X^\top X)^{-1}A^\top\}^{-1}(A\hat{\beta} - a)}^{\sim \chi^2_{\text{rank}(A)}}}{\underbrace{\|\hat{\epsilon}\|^2 / \sigma^2}_{\sim \chi^2_{n-K}}} \times \frac{n-K}{\text{rank}(A)} \\ &\sim F_{\text{rank}(A), n-K} \end{aligned}$$

- If F is larger than the critical value, reject the null

Influential Observations and Leverage Points

- Leverage for unit i in $\mathcal{S}(X)$:

$$p_i \equiv X_i^\top (X^\top X)^{-1} X_i = \text{the } i\text{th diagonal element of } P_X = \|P_X v(i)\|^2$$

where $v(i)$ is a vector such that $v(i)_i = 1$ and $v(i)_{i'} = 0$ with $i \neq i'$

- Interpretation: Projecting $v(i)$ on $\mathcal{S}(X)$

- $0 \leq p_i \leq \|v(i)\|^2 = 1$

- $\bar{p} \equiv \sum_{i=1}^n p_i/n = \text{trace}(P_X)/n = K/n$

- How much one observation can alter the estimate?

- OLS estimate without the i th observation:

$$\hat{\beta}_{(i)} = (X_{(i)}^\top X_{(i)})^{-1} X_{(i)}^\top Y_{(i)} = \hat{\beta} - (X^\top X)^{-1} X_i^\top \frac{\hat{\epsilon}_i}{1 - p_i}$$

- Influential points: (1) high leverage, (2) outlier, and (3) both

- Cook's distance: $D_i \equiv (\hat{\beta}_{(i)} - \hat{\beta})^\top X^\top X (\hat{\beta}_{(i)} - \hat{\beta}) / (\hat{\sigma}^2 K)$

Model-Based Asymptotic Inference

- An alternative expression: $\hat{\beta} = (\sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top)^{-1} (\sum_{i=1}^n \mathbf{X}_i Y_i)$
- Consistency: only $\mathbb{E}(\mathbf{X}_i \epsilon_i) = \mathbf{0}$ is required

$$\hat{\beta} - \beta = \left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \epsilon_i \right) \xrightarrow{P} \mathbf{0}$$

- Asymptotic distribution and inference:

$$\sqrt{n}(\hat{\beta} - \beta) = \underbrace{\left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top \right)^{-1}}_{\xrightarrow{P} \{\mathbb{E}(\mathbf{X}_i \mathbf{X}_i^\top)\}^{-1}} \times \underbrace{\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \epsilon_i \right)}_{\xrightarrow{D} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{E}(\mathbf{X}_i \mathbf{X}_i^\top))}$$

$$\xrightarrow{D} \mathcal{N}(\mathbf{0}, \sigma^2 \{\mathbb{E}(\mathbf{X}_i \mathbf{X}_i^\top)\}^{-1})$$

$$\frac{\hat{\beta}_k - \beta_k}{\text{s.e.}_k} \xrightarrow{D} \mathcal{N}(0, 1)$$

Robust Standard Errors

- **Heteroskedasticity:** $\mathbb{V}(\epsilon_i | \mathbf{X}) \neq \sigma^2$
- $\mathbb{V}(\hat{\beta} | \mathbf{X}) = \mathbb{V}(\hat{\beta} - \beta | \mathbf{X}) = (\mathbf{X}^\top \mathbf{X})^{-1} \{ \mathbf{X}^\top \mathbb{E}(\epsilon \epsilon^\top | \mathbf{X}) \mathbf{X} \} (\mathbf{X}^\top \mathbf{X})^{-1}$
- How do we estimate $\mathbb{E}(\epsilon \epsilon^\top | \mathbf{X}) = \mathbb{V}(\epsilon | \mathbf{X})$?
- **Sandwich estimator:** meat = $\mathbf{X}^\top \widehat{\mathbb{V}(\epsilon | \mathbf{X})} \mathbf{X}$, bread = $(\mathbf{X}^\top \mathbf{X})^{-1}$
- asymptotically consistent
- i.i.d.: $\hat{\sigma}^2 I_n$
- independence: $\text{diag}(\hat{\epsilon}_i^2)$, $n \text{diag}(\hat{\epsilon}_i^2)/(n - K)$, etc.
- clustering:
$$\begin{pmatrix} \hat{\epsilon}_1 \hat{\epsilon}_1^\top & 0 & \cdots & 0 \\ 0 & \hat{\epsilon}_2 \hat{\epsilon}_2^\top & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\epsilon}_G \hat{\epsilon}_G^\top \end{pmatrix} \text{ or}$$
$$\text{meat} = \sum_{g=1}^G \mathbf{X}_g^\top \hat{\epsilon}_g \hat{\epsilon}_g^\top \mathbf{X}_g$$
- autocorrelation, panel-correction, etc.
- **WARNING:** only fixes asymptotic standard error but not bias!

Asymptotic Tests

- Null hypothesis $H_0 : A\beta = a$ where A is of full row rank
- Any linear restriction: $b_1\beta_1 + b_2\beta_2 + \dots + b_K\beta_K = c$
- Any number of linearly independent such restrictions
- **Wald statistic:**

$$\begin{aligned} W &\equiv (A\hat{\beta} - a)^\top \{A\widehat{V}(\hat{\beta})A^\top\}^{-1} (A\hat{\beta} - a) \\ &= \underbrace{\sqrt{n}(A\hat{\beta} - a)^\top}_{\xrightarrow{D} \mathcal{N}(0, nAV(\hat{\beta})A^\top)} \underbrace{\{nA\widehat{V}(\hat{\beta})A^\top\}^{-1}}_{\xrightarrow{P} \{nAV(\hat{\beta})A^\top\}^{-1}} \underbrace{\sqrt{n}(A\hat{\beta} - a)}_{\xrightarrow{D} \mathcal{N}(0, nAV(\hat{\beta})A^\top)} \\ &\xrightarrow{D} \chi_{\text{rank}(A)}^2 \end{aligned}$$

- If W is larger than the critical value, reject the null

Generalized Least Squares (GLS)

- Known heteroskedasticity: $\mathbb{V}(\epsilon | X) = \sigma^2 \Omega$ where Ω is a positive definite matrix
- OLS estimator is no longer BLUE
- GLS estimator: $\hat{\beta}_{GLS} = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} Y$
- $\hat{\beta}_{GLS}$ is BLUE
- Consider the transformed regression: $Y^* = X^* \beta + \epsilon^*$ where $Y^* = \Omega^{-1/2} Y$, $X^* = \Omega^{-1/2} X$, and $\epsilon^* = \Omega^{-1/2} \epsilon$
- Cholesky decomposition: $\Omega = \Omega^{1/2} \Omega^{1/2 T}$ where $\Omega^{1/2}$ is a lower triangular matrix with strictly positive diagonal elements
- Variance: $\mathbb{V}(\hat{\beta}_{GLS} | X) = \sigma^2 (X^T \Omega^{-1} X)^{-1}$ with $\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^n (\hat{\epsilon}_i^*)^2$
- Feasible GLS (FGLS):
 - 1 Estimate the unknown Ω
 - 2 $\hat{\beta}_{FGLS} = (X^T \hat{\Omega}^{-1} X)^{-1} X^T \hat{\Omega}^{-1} Y$
 - 3 Iterate until convergence

Weighted Least Squares

- $\Omega = \text{diag}(1/w_i)$ with $\Omega^{-1/2} = \text{diag}(\sqrt{w_i})$
- Independence across units, but different variance for each unit
- Never know variance in practice
- $w_i = \text{Sampling weights} = 1 / \text{Pr}(\text{being selected into the sample})$
- Know a priori, independent sampling of units
- OLS estimator in the finite population:

$$\beta_P = \left(\sum_{i=1}^N X_i X_i^\top \right)^{-1} \sum_{i=1}^N X_i Y_i$$

- WLS estimator is consistent for β_P :

$$\hat{\beta}_{WLS} = \left(\sum_{i=1}^N \mathbf{1}\{i \in S\} w_i X_i X_i^\top \right)^{-1} \sum_{i=1}^N \mathbf{1}\{i \in S\} w_i X_i Y_i$$

Omitted Variables Bias

- Omitted variables, unmeasured confounders
- Scalar confounder: $Y = X\beta + \underbrace{U\gamma + \epsilon^*}_{\epsilon}$ with $\mathbb{E}(\epsilon^* | X, U) = 0$
- Projection: $U = P_X U + \hat{\eta} = X\hat{\delta} + \hat{\eta}$ where $\mathbb{E}(\hat{\eta}) = \mathbb{E}(\hat{\eta} \cdot x_k) = 0$
- Then,

$$Y = X(\beta + \gamma\hat{\delta}) + \gamma\hat{\eta} + \epsilon^*$$

with $\mathbb{E}(\gamma\hat{\eta}_i + \epsilon_i^*) = \mathbb{E}(X_{ik}(\gamma\hat{\eta}_i + \epsilon_i^*)) = 0$

- $\hat{\beta}_k \xrightarrow{P} \beta_k + \gamma\delta_k$ where $\hat{\delta} \xrightarrow{P} \delta$
- If $\delta_{k'} = 0$ for all $k' \neq k$, then

$$\hat{\beta}_k \xrightarrow{P} \beta_k + \gamma \frac{\text{Cov}(U_i, X_{ik})}{\mathbb{V}(X_{ik})} = \beta_k + \gamma \text{Cor}(U_i, X_{ik}) \sqrt{\frac{\mathbb{V}(U_i)}{\mathbb{V}(X_{ik})}}$$

Measurement Error

- Three types of ME: classical, nondifferential, differential
- ME in Y_i : $Y_i = Y_i^* + e_i$
- Omitted variable problem: $Y_i = \mathbf{X}_i^\top \beta + e_i + \epsilon_i$
- ME in x_k : $X_{ik} = X_{ik}^* + e_i$
- Classical ME assumption: $\text{Cov}(e_i, X_{ik}^*) = \text{Cov}(e_i, X_{ik'}) = 0$ for all $k' \neq k$, but $\text{Cov}(e_i, X_{ik}) \neq 0$
- Omitted variable problem: $Y_i = \mathbf{X}_i^\top \beta - \beta_k e_i + \epsilon_i$
- **Attenuation bias**: $\hat{\beta}_k \xrightarrow{P} \beta_k \left(\frac{\text{V}(X_{ik}^*)}{\text{V}(X_{ik}^*) + \text{V}(e_i)} \right) \leq \beta_k$
- ME in x_k yields an inconsistent estimate of $\beta_{k'}$ unless $\text{Cov}(e_i, X_{ik'}) = 0$

Concluding Remarks about Linear Regression

- For experimental data, no need to run regression!
- Use covariate adjustment before randomization of treatment (e.g., matched-pair design, randomized block design) with design-based estimator
- Robust and cluster standard errors can be justified from the design-based point of view

- In observational studies, regression adjustment is common
- Many results depend on linearity and exogeneity
- Non/semi-parametric regression
- Preprocess the data with matching methods to make parametric inference robust