Linear Regression

Kosuke Imai

Princeton University

POL572 Quantitative Analysis II
Spring 2016

Simple Linear Regression

Simple Linear Regression Model

- The linear regression with a single variable:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

  where $\mathbb{E}(\epsilon_i) = 0$

- Data
  - Y_i: outcome (response, dependent) variable
  - X_i: predictor (explanatory, independent) variable, covariate
- Parameters
  - α: intercept
  - β: slope
- ϵ_i: error term, disturbance, ~~residual~~
  - $\mathbb{E}(\epsilon_i) = 0$ is not really an assumption because we have α

- Why simple regression? It's simple and gives lots of intuition

Causal Interpretation

- Association: you can always regress Y_i on X_i and vice versa
- The previous model written in terms of potential outcomes:

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

where $\mathbb{E}(\epsilon_i) = 0$

  - X_i is the treatment variable
  - $\alpha = \mathbb{E}(Y_i(0))$
  - $\beta = Y_i(1) - Y_i(0)$ for all $i \Longleftrightarrow$ Constant additive unit causal effect

- A more general model:

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

where $\mathbb{E}(\epsilon_i(1)) = \mathbb{E}(\epsilon_i(0)) = 0$

  - Relax the assumption of $\epsilon_i = \epsilon_i(1) = \epsilon_i(0)$
  - $Y_i(1) - Y_i(0) = \{\alpha + \beta + \epsilon_i(1)\} - \{\alpha + \epsilon_i(0)\} = \beta + \epsilon_i(1) - \epsilon_i(0)$
  - ATE: $\beta = \mathbb{E}(Y_i(1) - Y_i(0))$
  - $\alpha = \mathbb{E}(Y_i(0))$ as before

Assumptions

- Exogeneity: $\mathbb{E}(\epsilon_i \mid X) = \mathbb{E}(\epsilon_i) = 0$ where $X = (X_1, X_2, \ldots, X_n)$
  1. Orthogonality: $\mathbb{E}(\epsilon_i X_j) = 0$
  2. Zero correlation: $\mathrm{Cov}(\epsilon_i, X_j) = 0$
- Homoskedasticity: $\mathbb{V}(\epsilon_i \mid X) = \mathbb{V}(\epsilon_i) = \sigma^2$

- Classical randomized experiments:
  1. Randomization of treatment: $(Y_i(1), Y_i(0)) \perp\!\!\!\perp X_i$ for all i
     - $\mathbb{E}(Y_i(x) \mid X_i) = \mathbb{E}(Y_i(x)) \Longleftrightarrow \mathbb{E}(\epsilon_i(x) \mid X_i) = \mathbb{E}(\epsilon_i(x)) = 0$
     - $\mathbb{E}(Y_i(x)) = \mathbb{E}(Y_i \mid X_i = x) = \alpha + \beta x$
  2. Random sampling of units:
     - $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp \{Y_j(1), Y_j(0)\}$ for any $i \neq j$
     - $\{\epsilon_i(1), \epsilon_i(0)\} \perp\!\!\!\perp \{\epsilon_j(1), \epsilon_j(0)\}$ for any $i \neq j$
  3. Variance of potential outcomes:
     - $\mathbb{V}(\epsilon_i(x)) = \mathbb{V}(\epsilon_i(x) \mid X_i) = \mathbb{V}(Y_i(x) \mid X_i) = \mathbb{V}(Y_i(x)) = \sigma_x$ for $x = 0, 1$

Least Squares Estimation

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

$$(\hat{\alpha}, \hat{\beta}) = \underset{(\alpha,\beta)}{\text{argmin}} \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

which yields

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \xrightarrow{p} \frac{\text{Cov}(X_i, Y_i)}{\mathbb{V}(X_i)} = \rho_{XY}\sqrt{\frac{\mathbb{V}(Y_i)}{\mathbb{V}(X_i)}}$$

Unbiasedness of Least Squares Estimator

- When X_i is binary, $\hat{\beta} =$ Difference-in-Means estimator
- So, $\hat{\beta}$ is unbiased from the design-based perspective too

- Model-based estimation error:

$$\hat{\beta} - \beta \;=\; \frac{\sum_{i=1}^{n}(X_i - \overline{X})\epsilon_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

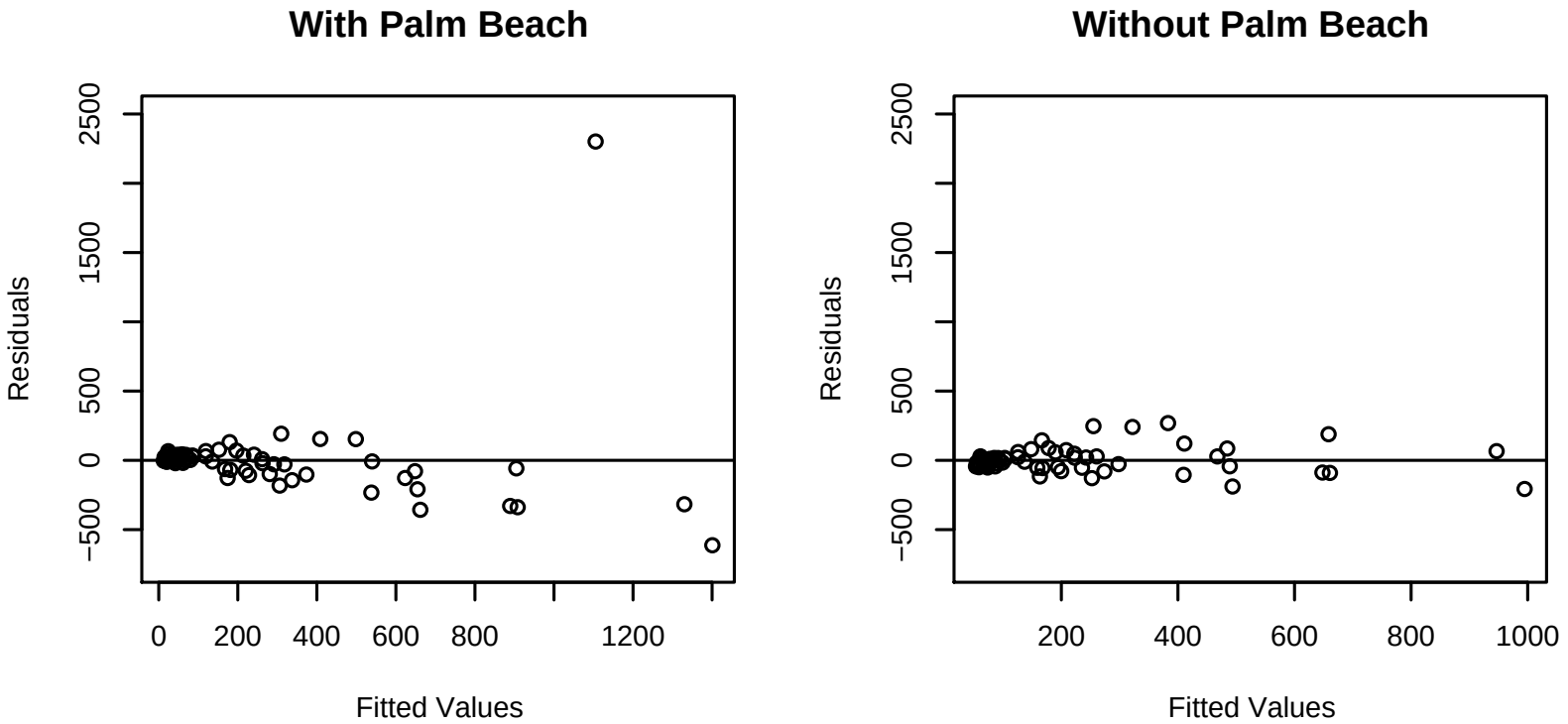
- Thus, the exogeneity assumption implies,

$$\mathbb{E}(\hat{\beta}) - \beta \;=\; \mathbb{E}\{\mathbb{E}(\hat{\beta} - \beta \mid X)\} \;=\; 0$$

- Similarly, $\hat{\alpha} - \alpha = \overline{\epsilon} - (\hat{\beta} - \beta)\overline{X}$
- Thus, $\mathbb{E}(\hat{\alpha}) - \alpha = 0$

Residuals

- Estimated error term
- Zero mean: $\sum_{i=1}^{n} \hat{\epsilon}_i = 0$
- Orthogonality: $\langle \hat{\epsilon}, X \rangle = \hat{\epsilon}^\top X = \sum_{i=1}^{n} \hat{\epsilon}_i X_i = 0$
- Sample correlation between $\hat{\epsilon}_i$ and X_i is exactly 0
- Does not imply $\mathbb{E}(\langle \epsilon, X \rangle) = 0$ or $\mathrm{Cor}(\epsilon_i, X_i)$
- Residual plot ($Y =$ 2000 Buchanan votes, $X =$ 1996 Perot votes):



The Coefficient of Determination

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

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

- 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 $Y_i = \widehat{Y}_i$ for all i
- Example: 0.85 (without Palm Beach) vs. 0.51 (with Palm Beach)

Model-Based Variance and Its Estimator

- The homoskedasticity assumption implies

$$\mathbb{V}(\hat{\beta} \mid X) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

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

$$\widehat{\mathbb{V}(\hat{\beta} \mid X)} = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \quad \text{where} \quad \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\epsilon}_i^2$$

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

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

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

$$\mathbb{V}(\hat{\beta}) = \mathbb{E}\{\mathbb{V}(\hat{\beta} \mid X)\} = \mathbb{E}[\mathbb{E}\{\widehat{\mathbb{V}(\hat{\beta} \mid X)} \mid X\}] = \mathbb{E}\{\widehat{\mathbb{V}(\hat{\beta} \mid X)}\}$$

Model-Based Prediction

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

$$\widehat{Y(x)} \;=\; \mathbb{E}(Y \mid X_i = x) \;=\; \hat{\alpha} + \hat{\beta}x$$

- Predicted value for $X_i = x$:

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

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

$$\begin{aligned}
\mathbb{V}(Y(x) \mid X) &= \mathbb{V}(\hat{\alpha} \mid X) + \mathbb{V}(\hat{\beta} \mid X) \cdot x^2 + 2 \cdot x \cdot \mathrm{Cov}(\hat{\alpha}, \hat{\beta} \mid X) + \sigma^2 \\
&= \sigma^2 \left(\frac{\sum_{i=1}^n (X_i - x)^2}{n \sum_{i=1}^n (X_i - \overline{X})^2} + 1 \right)
\end{aligned}$$

where $\mathbb{V}(\hat{\alpha} \mid X) = \frac{\sigma^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \overline{X})^2}$ and $\mathrm{Cov}(\hat{\alpha}, \hat{\beta} \mid X) = -\frac{\sigma^2 \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2}$

Model-Based Finite Sample Inference

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

$$\hat{\beta} - \beta \sim \mathcal{N}\left(0, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)$$

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

$$\frac{\hat{\beta} - \beta}{\text{s.e.}} = \underbrace{\frac{\hat{\beta} - \beta}{\sigma / \sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}}}_{\sim \mathcal{N}(0,1)} \Bigg/ \sqrt{\underbrace{\frac{(n-2)\hat{\sigma}^2}{\sigma^2}}_{\sim \chi^2_{n-2}} \times \frac{1}{n-2}} \sim t_{n-2}$$

Key Distributions for Linear Regression and Beyond

1. Chi-square distribution with ν degrees of freedom: $X \sim \chi^2_\nu$
   - Construction I: $X = \sum_{i=1}^{\nu} Y_i^2$ where $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$
   - Construction II: $X = Y^\top \Sigma^{-1} Y$ where $Y \sim \mathcal{N}_\nu(0, \Sigma)$
   - Mean and variance: $\mathbb{E}(X) = \nu$ and $\mathbb{V}(X) = 2\nu$
   - $(X_1 + X_2) \sim \chi^2_{\nu_1+\nu_2}$ if $X_1 \sim \chi^2_{\nu_1}$ and $X_2 \sim \chi^2_{\nu_2}$, which are independent
2. Student's t distribution with ν degrees of freedom: $X \sim t_\nu$
   - Construction: $X = Y/\sqrt{Z/\nu}$ if $Y \sim \mathcal{N}(0,1)$ and $Z \sim \chi^2_\nu$
   - Mean and variance: $\mathbb{E}(X) = 0$ and $\mathbb{V}(X) = \nu/(\nu-2)$
   - Cauchy distribution ($\nu = 1$): $\mathbb{E}(X) = \mathbb{V}(X) = \infty$
   - Normal distribution ($\nu = \infty$): $\mathbb{E}(X) = 0$ and $\mathbb{V}(X) = 1$
3. F distribution with ν_1 and ν_2 degrees of freedom: $X \sim F(\nu_1, \nu_2)$
   - Construction I: $X = \frac{Y_1/\nu_1}{Y_2/\nu_2}$ if $Y_1 \sim \chi^2_{\nu_1}$ and $Y_2 \sim \chi^2_{\nu_2}$, which are independent
   - Construction II ($\nu_1 = 1$): $X = Y^2$ where $Y \sim t_{\nu_2}$
   - Construction III: $X = 1/Y$ where $Y \sim F(\nu_2, \nu_1)$
   - Chi-square distribution: $X \xrightarrow{d} \chi^2_{\nu_1}$ as $\nu_2 \to \infty$

Model-Based Asymptotic Inference

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

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

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:

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

- 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 $\mathbb{V}(\epsilon_i \mid X) = \sigma^2(X_i) \neq \sigma^2$
- Heteroskedasticity consistent robust variance estimator (more later):

$$\widehat{\mathbb{V}((\hat{\alpha}, \hat{\beta}) \mid X)} = \left(\sum_{i=1}^{n} \widetilde{X}_i \widetilde{X}_i^\top \right)^{-1} \left(\sum_{i=1}^{n} \hat{\epsilon}_i^2 \widetilde{X}_i \widetilde{X}_i^\top \right) \left(\sum_{i=1}^{n} \widetilde{X}_i \widetilde{X}_i^\top \right)^{-1}$$

  where in this case $\widetilde{X}_i = (1, X_i)$ is a column vector of length 2
- Model-based justification: asymptotically valid in the presence of heteroskedastic errors
- Design-based evaluation:

$$\text{Finite Sample Bias} = \mathbb{E}(\widehat{\mathbb{V}(\hat{\beta} \mid X)}) - \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0} \right) = -\left(\frac{\sigma_1^2}{n_1^2} + \frac{\sigma_0^2}{n_0^2} \right)$$

- Bias vanishes asymptotically

Cluster Randomized Experiments

- Units: $i = 1, 2, \ldots, n_j$
- Clusters of units: $j = 1, 2, \ldots, m$
- 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:

$$\begin{aligned} \text{SATE} &\equiv \frac{1}{\sum_{j=1}^{m} n_j} \sum_{j=1}^{m} \sum_{i=1}^{n_j} (Y_{ij}(1) - Y_{ij}(0)) \\ \text{PATE} &\equiv \mathbb{E}(Y_{ij}(1) - Y_{ij}(0)) \end{aligned}$$

- Random sampling of clusters and units

Design-Based Inference

- For simplicity, assume the following:
  1. equal cluster size, i.e., $n_j = n$ for all j
  2. we observe all units for a selected cluster (no sampling of units)

- The difference-in-means estimator:

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

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

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

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

Intracluster Correlation Coefficient

- Comparison with the standard variance:

$$\mathbb{V}(\hat{\tau}) = \frac{\sigma_1^2}{m_1 n} + \frac{\sigma_0^2}{m_0 n}$$

- Correlation of potential outcomes across units within a cluster

$$\begin{aligned}
\mathbb{V}(\overline{Y_j(t)}) &= \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} Y_{ij}(t)\right) \\
&= \frac{1}{n^2}\left\{\sum_{i=1}^{n}\mathbb{V}(Y_{ij}(t)) + \sum_{i\neq i'}\sum_{i'=1}^{n}\mathrm{Cov}(Y_{ij}(t), Y_{i'j}(t))\right\} \\
&= \frac{\sigma_t^2}{n}\{1 + (n-1)\rho_t\} \overset{\text{typically}}{\geq} \frac{\sigma_t^2}{n}
\end{aligned}$$

Cluster Standard Error

- Cluster robust variance estimator:

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

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

- Design-based evaluation:

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

- Bias vanishes asymptotically as $m \to \infty$ 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: $\mathbb{E}(Y_i(t) \mid X_i = x)$ is continuous in x
- Contrast this with the "as-if random" assumption within a window

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid c_0 \leq X_i \leq c_1$$

- Estimand: $\mathbb{E}(Y_i(1) - Y_i(0) \mid X_i = c)$
- Regression modeling:

$$\begin{aligned}
\mathbb{E}(Y_i(1) \mid X_i = c) &= \lim_{x \downarrow c} \mathbb{E}(Y_i(1) \mid X_i = x) = \lim_{x \downarrow c} \mathbb{E}(Y_i \mid X_i = x) \\
\mathbb{E}(Y_i(0) \mid X_i = c) &= \lim_{x \uparrow c} \mathbb{E}(Y_i(0) \mid X_i = x) = \lim_{x \uparrow c} \mathbb{E}(Y_i \mid X_i = x)
\end{aligned}$$

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

Close Elections as RD Design (Lee)

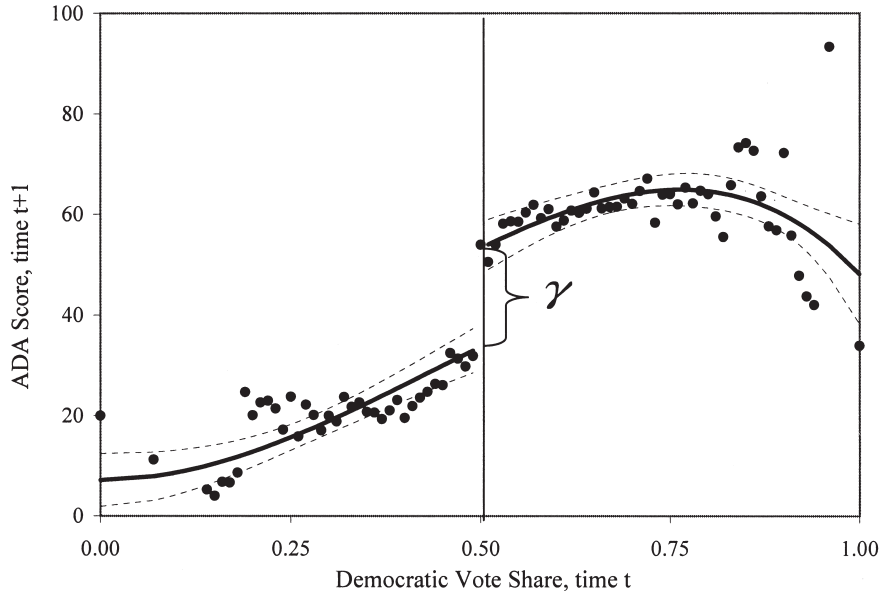


FIGURE I
Total Effect of Initial Win on Future ADA Scores: γ

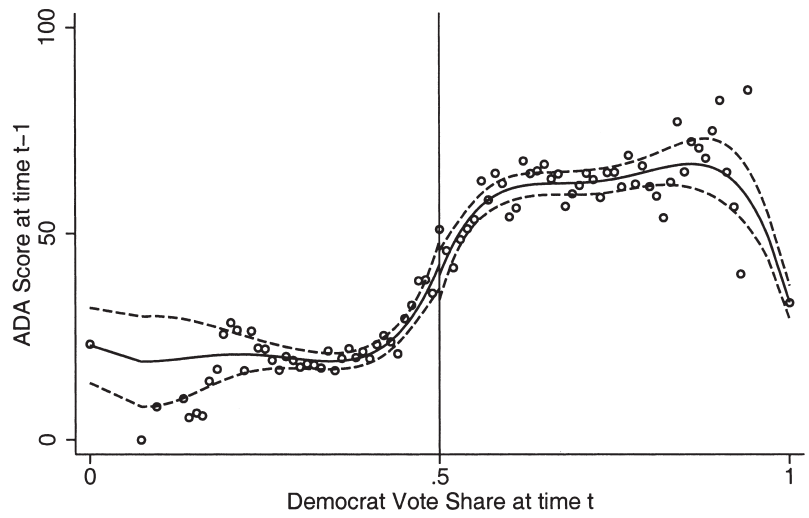


FIGURE V
Specification Test: Similarity of Historical Voting Patterns between Bare Democrat and Republican Districts

- Placebo test for natural experiments
- What is a good placebo?
  1. expected not to have any effect
  2. closely related to outcome of interest
- Close election controversy: de la Cuesta and Imai (2016). "Misunderstandings about the Regression Discontinuity Design in the Study of Close Elections" *Annual Review of Political Science*

Analysis Methods under the RD Design

- As-if randomization $\rightsquigarrow$ Difference-in-means within a window
- Continuity $\rightsquigarrow$ linear regression within a window
- We want to choose a window in a principled manner
- We want to relax the functional form assumption
- Higher-order polynomial regression $\rightsquigarrow$ not robust
- Local linear regression: better behavior at the boundary than other nonparametric regressions

$$f(x) \;=\; \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \{Y_i - \alpha - (X_i - x)\beta\}^2 K_h\left(\frac{X_i - x}{h}\right)$$

- Weighted regression with a kernel of one's choice:
  - uniform kernel: $k(u) = \frac{1}{2}\mathbf{1}\{|u| < 1\}$
  - triangular kernel: $k(u) = (1 - |u|)\mathbf{1}\{|u| < 1\}$
- Choice of bandwidth parameter h: cross-validation, analytical minimization of MSE

Multiple Regression from Simple Regressions

- Consider simple linear regression without an intercept:

$$\hat{\beta} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} = \frac{\langle X, Y \rangle}{\langle X, X \rangle}$$

  where $\langle \cdot, \cdot \rangle$ represents the inner product of two vectors

- Another predictor Z that is orthogonal to X, i.e., $\langle X, Z \rangle = 0$
- Then, regressing Y on X and Z separately give you the same coefficients as regressing Y on X and Z together:

$$\begin{aligned} \text{SSR} &= \langle Y - \beta X - \gamma Z, Y - \beta X - \gamma Z \rangle \\ &= \langle Y - \beta X, Y - \beta X \rangle + \langle Y - \gamma Z, Y - \gamma Z \rangle - \langle Y, Y \rangle \end{aligned}$$

- Orthogonal predictors have no impact on each other in multiple regression (experimental designs)

Orthogonalization

- How to orthogonalize Z with respect to X? Regress Z on X and obtain residuals!
- An alternative way to estimate β in $Y = \alpha\mathbf{1} + \beta X + \epsilon$
  1. Orthogonalize X with respect to $\mathbf{1}$:

$$\langle X - \overline{X}\mathbf{1}, \mathbf{1} \rangle = 0$$

  2. Now, we can get $\hat{\beta}$ from a simple regression:

$$\hat{\beta} = \frac{\langle X - \overline{X}\mathbf{1}, Y \rangle}{\langle X - \overline{X}\mathbf{1}, X - \overline{X}\mathbf{1} \rangle}$$

- $\hat{\beta}$ represents additional contribution of X on Y after X has been adjusted for $\mathbf{1}$
- Generalization of this idea: Gram-Schmidt procedure

Proof by A Picture

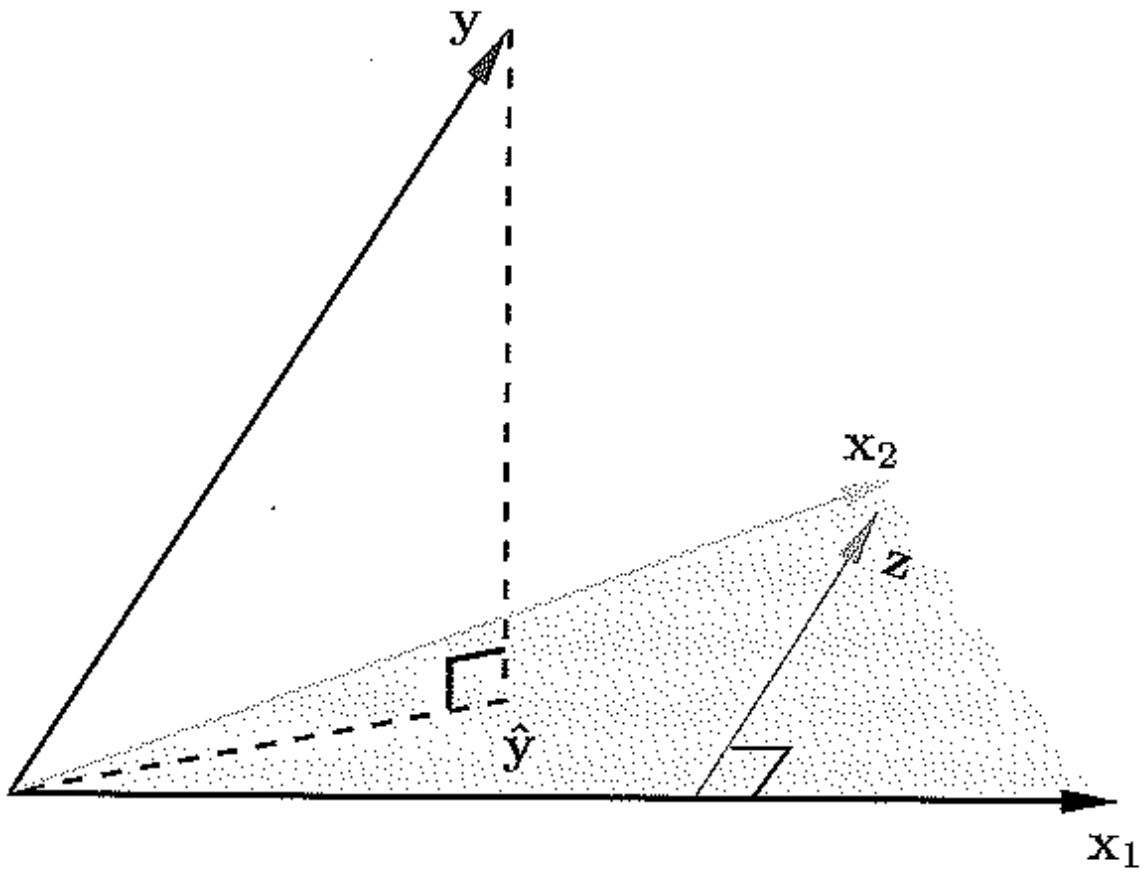


FIGURE 3.4. *Least squares regression by orthogonalization of the inputs. The vector* $\mathbf{x}_2$ *is regressed on the vector* $\mathbf{x}_1$*, leaving the residual vector* $\mathbf{z}$*. The regression of* $\mathbf{y}$ *on* $\mathbf{z}$ *gives the multiple regression coefficient of* $\mathbf{x}_2$*. Adding together the projections of* $\mathbf{y}$ *on each of* $\mathbf{x}_1$ *and* $\mathbf{z}$ *gives the least squares fit* $\hat{\mathbf{y}}$*.*

Multiple Linear Regression

Multiple Linear Regression Model

1. Scalar representation:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_K X_{iK} + \epsilon_i$$

where typically $X_{i1} = 1$ is the intercept

2. Vector representation:

$$Y_i = X_i^\top \beta + \epsilon_i \quad \text{where} \quad X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{nK} \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix}$$

3. Matrix representation:

$$Y = \mathbf{X}\beta + \epsilon \quad \text{where} \quad \mathbf{X} = \begin{pmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_n^\top \end{pmatrix} \quad \text{and} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Causal Interpretation

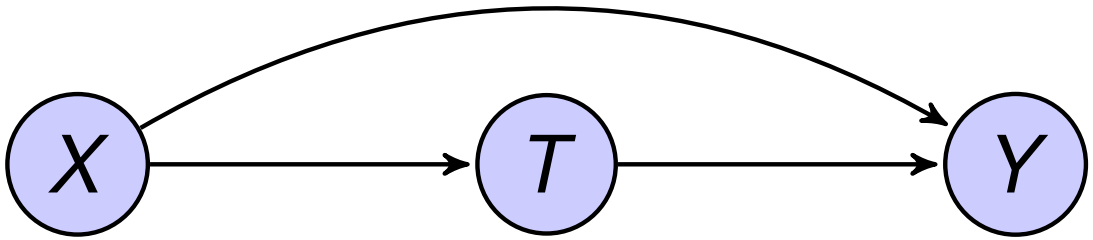
- Potential outcomes:

$$Y_i(T_i) = \alpha + \beta T_i + X_i^\top \gamma + \epsilon_i$$

  - T_i is a treatment variable
  - X_i is a vector of *pre-treatment* confounders
  - Constant unit treatment effect: $\beta = Y_i(1) - Y_i(0)$
- A general model with PATE $\beta = \mathbb{E}(Y_i(1) - Y_i(0))$:

$$Y_i(T_i) = \alpha + \beta T_i + X_i^\top \gamma + \epsilon_i(T_i)$$

- Post-treatment bias: X_i shouldn't include post-treatment variables



- γ does NOT have a causal interpretation

Assumptions and Interpretation

- Exogeneity: $\mathbb{E}(\epsilon \mid \mathbf{X}) = \mathbb{E}(\epsilon) = 0$
  1. Conditional Expectation Function (CEF): $\mathbb{E}(Y \mid \mathbf{X}) = \mathbf{X}\beta$
  2. Orthogonality: $\mathbb{E}(\epsilon_i \cdot X_{jk}) = 0$ for any i, j, k
  3. Zero correlation: $\mathrm{Cov}(\epsilon_i, X_{jk}) = 0$ for any i, j, k
- Homoskedasticity: $\mathbb{V}(\epsilon \mid \mathbf{X}) = \mathbb{V}(\epsilon) = \sigma^2 \mathbf{I}_n$
  1. spherical error variance
  2. $\mathbb{V}(\epsilon_i \mid \mathbf{X}) = \sigma^2$ and $\mathrm{Cov}(\epsilon_i, \epsilon_j \mid \mathbf{X}) = 0$ for any $i \neq j$

- Ignorability (unconfoundedness): $(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i \mid X_i$
  - $\mathbb{E}(Y_i(t) \mid T_i, X_i) = \mathbb{E}(Y_i(t) \mid X_i)$
  - $\mathbb{E}(Y_i(t) \mid X_i) = \mathbb{E}(Y_i \mid T_i = t, X_i) = \alpha + \beta t + X_i^\top \gamma$
  - $\mathbb{E}(\epsilon_i \mid T_i, X_i) = \mathbb{E}(\epsilon_i \mid X_i) \stackrel{?}{=} \mathbb{E}(\epsilon_i) = 0$
- Random sampling: $(Y_i(1), Y_i(0), T_i, X_i)$ is i.i.d.
- Equal variance: $\mathbb{V}(Y_i(t) \mid T_i, X_i) = \mathbb{V}(Y_i(t)) = \sigma^2$

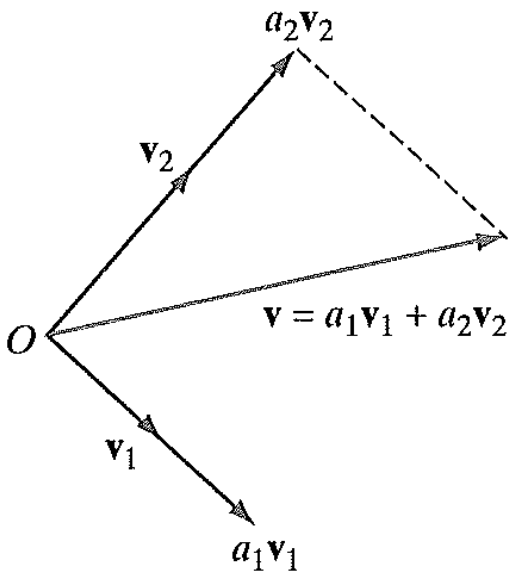
Rank and Linear Independence

- Assumption (No multicollinearity): $\mathbf{X}$ is full column rank

$$\text{rank}(\mathbf{X}) \;=\; K \leq n$$

- Column (row) rank of a matrix = maximal # of linearly independent columns (rows)
- Linear independence: $\mathbf{X}c = 0$ iff c is a column vector of 0s

$$c_1 X_1 + c_2 X_2 + \cdots + c_K X_K \;=\; 0 \quad \Longleftrightarrow \quad c_1 = c_2 = \cdots = c_K = 0$$



- Geometric interpretation of matrix rank:

$$\text{rank}(\mathbf{X}) \;=\; \dim(\mathcal{S}(\mathbf{X}))$$

where $\mathcal{S}(\mathbf{X}) = \{\mathbf{X}c : c \in \mathbf{R}^K\}$ is the column space of $\mathbf{X}$

Properties of Matrix Rank

- Property 1: column rank of $\mathbf{X}$ = row rank of $\mathbf{X}$
- Proof:
  1. Let column rank be $L \leq K$. There exist basis vectors b such that $X_k = c_{1k}b_1 + c_{2k}b_2 + \cdots + c_{Lk}b_L$ for each column k of $\mathbf{X}$.
  2. We can write $\mathbf{X} = \mathbf{BC}$. Since every row of $\mathbf{X}$ is a linear combination of rows of $\mathbf{C}$, row rank of $\mathbf{X}$ is no greater than row rank of $\mathbf{C}$, or equivalently, column rank of $\mathbf{X}$, i.e., L.
  3. Applying the same argument to $\mathbf{X}^\top$ to show row rank is no greater than column rank.

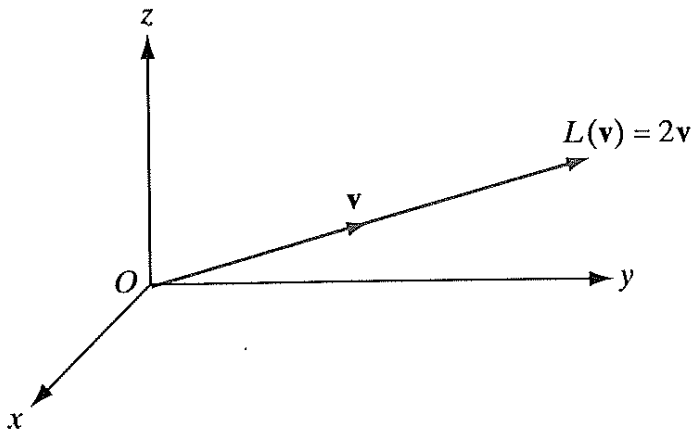
- Property 2: $\text{rank}(\mathbf{AB}) \leq \min(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B}))$
- Proof is similar to the previous one

- Property 3: If $\mathbf{A}$ and $\mathbf{C}$ are non-singular, $\text{rank}(\mathbf{ABC}) = \text{rank}(\mathbf{B})$
- Proof:
  1. $\text{rank}(\mathbf{ABC}) \leq \text{rank}(\mathbf{AB}) \leq \text{rank}(\mathbf{B})$
  2. $\text{rank}(\mathbf{B}) = \text{rank}(\mathbf{A}^{-1}(\mathbf{ABC})\mathbf{C}^{-1}) \leq \text{rank}(\mathbf{A}^{-1}(\mathbf{ABC})) \leq \text{rank}(\mathbf{ABC})$

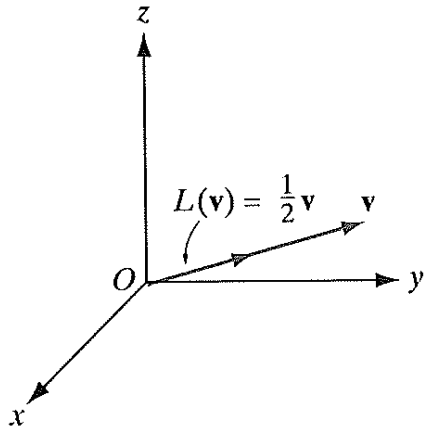
Matrix and Linear Transformation

- A matrix **A** can be interpreted as a linear transformation operator:

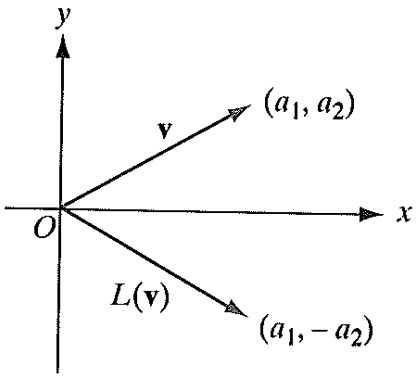
$$\text{transformed vector} = \mathbf{A}v$$

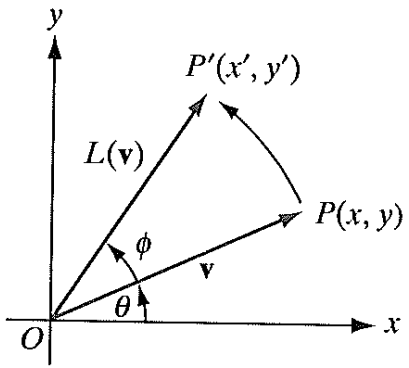


(a) Dilation: $r > 1$.



(b) Contraction: $0 < r < 1$.



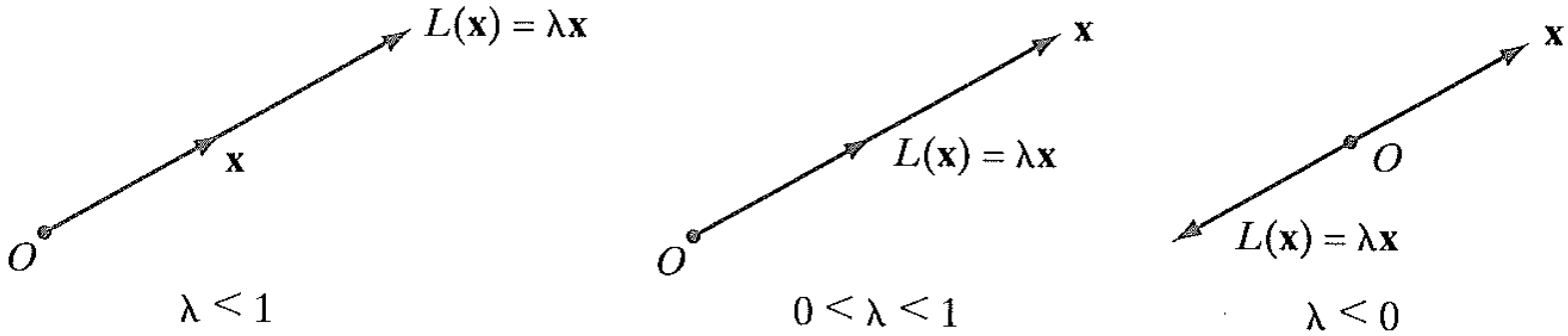


(b) Rotation.

Eigenvalues and Eigenvectors

- Definition: $\mathbf{A}v = \lambda v$
  1. a square matrix $\mathbf{A}$
  2. a non-zero column vector v: an eigenvector of $\mathbf{A}$
  3. a scalar λ: an eigenvalue of $\mathbf{A}$ associated with v

- Interpretation: direction of v is not affected by linear transform $\mathbf{A}$

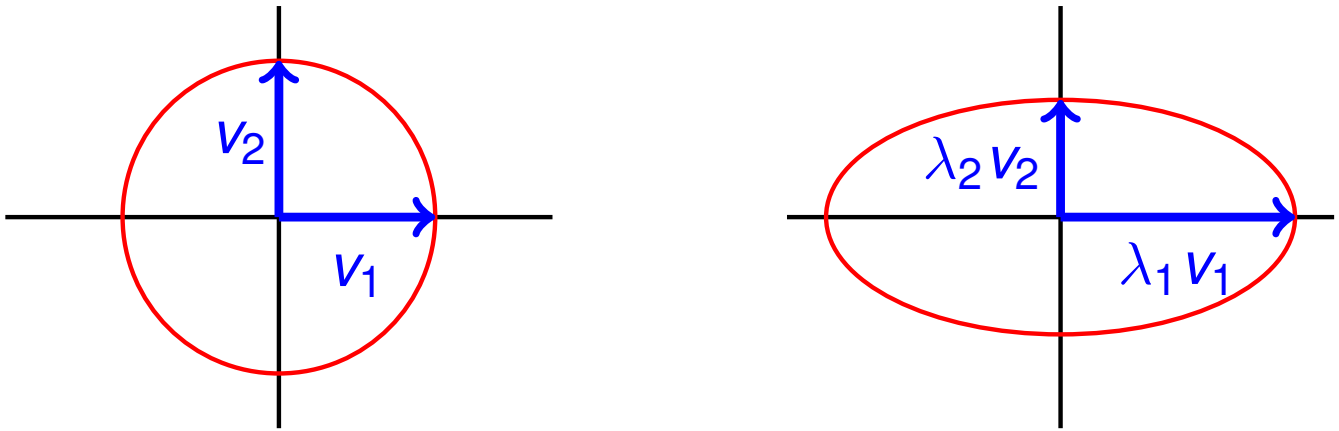


Spectoral Theorem

If $\mathbf{A}$ is a symmetric $n \times n$ matrix, then eigenvalues $\lambda_1, \ldots, \lambda_n$ and their corresponding orthonormal eigenvectors $v_1, \ldots, v_n$ exist. Moreover, the following spectoral decomposition (eigen decomposition) applies,

$$\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^\top \quad \text{where} \quad \mathbf{D} = \text{diag}(\lambda_1, \ldots, \lambda_n) \text{ and } \mathbf{V} = [v_1 \ \cdots v_n]$$

- Orthonormality implies $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_n$ or $\mathbf{V}^{-1} = \mathbf{V}^\top$. Thus, $\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}$
- Geometric interpretation: $\mathbf{A}\mathbf{V} = \mathbf{V}\mathbf{D} = [\lambda_1 v_1 \ \cdots \ \lambda_n v_n]$



- Corollaries: $\text{trace}(\mathbf{A}) = \lambda_1 + \cdots + \lambda_n$ and $\det(\mathbf{A}) = \lambda_1 \times \cdots \times \lambda_n$

Sigularity

1. If $\mathbf{A}$ is a symmetric $n \times n$ matrix, $\text{rank}(\mathbf{A})$ equals the number of non-zero eigenvalues of $\mathbf{A}$
   - Proof: $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{V}\mathbf{D}\mathbf{V}^{-1}) = \text{rank}(\mathbf{D})$

2. A symmetric $n \times n$ matrix has full rank iff it is non-singular
   - Proof: non-singular $\Leftrightarrow \det(\mathbf{A}) \neq 0 \Leftrightarrow$ all eigenvalues are non-zero
   - $\mathbf{A}c = b$ has a unique solution $c = \mathbf{X}^{-1}b$ (K unknowns and K equations)

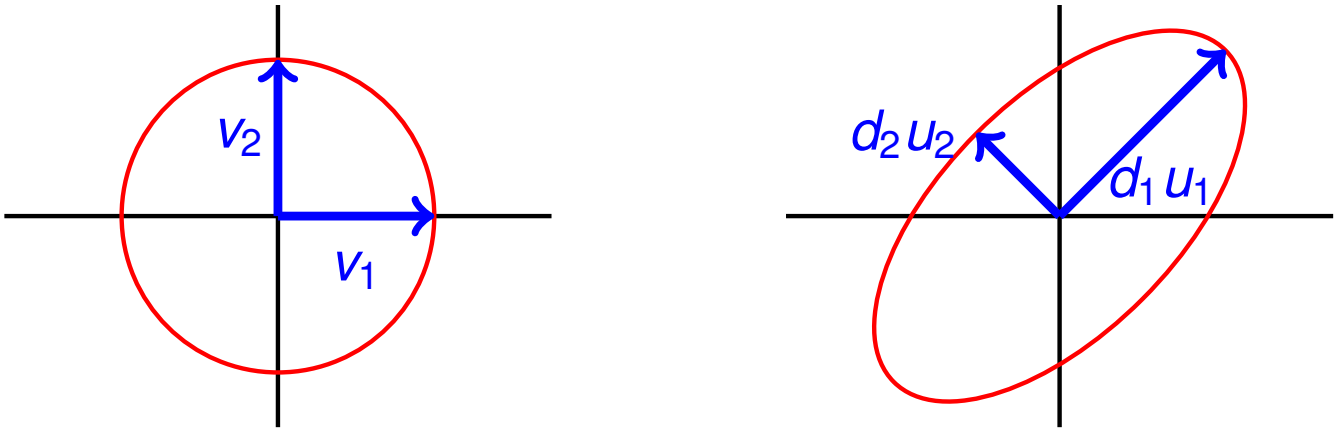
3. $\mathbf{X}$ is of full column rank iff $\mathbf{X}^\top\mathbf{X}$ is non-singular
   - Suppose $\mathbf{X}$ has full rank: assume $\mathbf{X}^\top\mathbf{X}c = 0$ for a non-zero c $\Longrightarrow c^\top\mathbf{X}^\top\mathbf{X}c = 0 \Longrightarrow \|\mathbf{X}c\|^2 = 0 \Longrightarrow \mathbf{X}c = 0$, contradiction
   - Suppose $\mathbf{X}^\top\mathbf{X}$ has full rank: assume $\mathbf{X}c = 0$ for a non-zero c $\Longrightarrow \mathbf{X}^\top\mathbf{X}c = 0$, contradiction

Singular Value Decomposition (SVD)

- For any $m \times n$ matrix $\mathbf{A}$,

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top \quad \text{where} \quad \mathbf{D} = \text{diag}(d_1, \ldots, d_n)$$

  - singular values of $\mathbf{A}$: $d_1 \geq \cdots \geq d_n \geq 0$
  - orthogonal matrix $\mathbf{U}$ whose columns span the column space of $\mathbf{A}$
  - orthogonal matrix $\mathbf{V}$ whose columns span the row space of $\mathbf{A}$
  - Geometric interpretation: $\mathbf{AV} = \mathbf{UD}$



- Relationship to spectoral decomposition:

$$\mathbf{A}^\top\mathbf{A} = (\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top(\mathbf{U}\mathbf{D}\mathbf{V}^\top) = \mathbf{V}\mathbf{D}^2\mathbf{V}^\top$$

Columns of $\mathbf{V}$ are eigenvectors and $\text{diag}(\mathbf{D}^2)$ are eigenvalues

Principal Components

- An $n \times K$ matrix of predictors: $\mathbf{X}$
- "Centered" predictors: $\widetilde{\mathbf{X}}$ where $\widetilde{X}_k = X_k - \text{mean}(X_k)$
- Sample covariance matrix: $\frac{1}{n}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}$
- Singular value decomposition of $\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}$:

$$\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} = \mathbf{V}\mathbf{D}^2\mathbf{V}^\top$$

- Columns of $\mathbf{V}$ are principal components direction of $\widetilde{\mathbf{X}}$
- $z_k = \widetilde{\mathbf{X}}v_k = u_k d_k$ is called the kth principal component
- Variances of principal components:

$$\mathbb{V}(z_1) = \frac{d_1^2}{n} \geq \mathbb{V}(z_2) = \frac{d_2^2}{n} \geq \cdots \geq \mathbb{V}(z_K) = \frac{d_K^2}{n}$$

- Frequently used to summarize multiple measurements

Least Squares Estimation and Prediction

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

$$\text{SSR} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \langle \hat{\epsilon}, \hat{\epsilon} \rangle = \|\hat{\epsilon}\|^2$$

which yields

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\ \text{SSR} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y$$

- Computation via SVD: $\mathbf{X}^\top \mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^\top$ and $(\mathbf{X}^\top \mathbf{X})^{-1} = \mathbf{V}\mathbf{D}^{-2}\mathbf{V}^\top$

Geometry of Least Squares

1. Using $\hat{\beta} = \beta + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \epsilon$, show the orthogonality

$$\mathbf{X}^\top \hat{\epsilon} = \mathbf{X}^\top (Y - \mathbf{X}\hat{\beta}) = 0$$

   that is, any column vector of $\mathbf{X}$ is orthogonal to $\hat{\epsilon}$
2. Using the orthogonality, show

$$\text{SSR} = \|Y - \mathbf{X}\tilde{\beta}\|^2 = \|\hat{\epsilon} + \mathbf{X}(\hat{\beta} - \tilde{\beta})\|^2 = \|\hat{\epsilon}\|^2 + \|\mathbf{X}(\hat{\beta} - \tilde{\beta})\|^2$$

3. SSR is minimized when $\tilde{\beta} = \hat{\beta}$

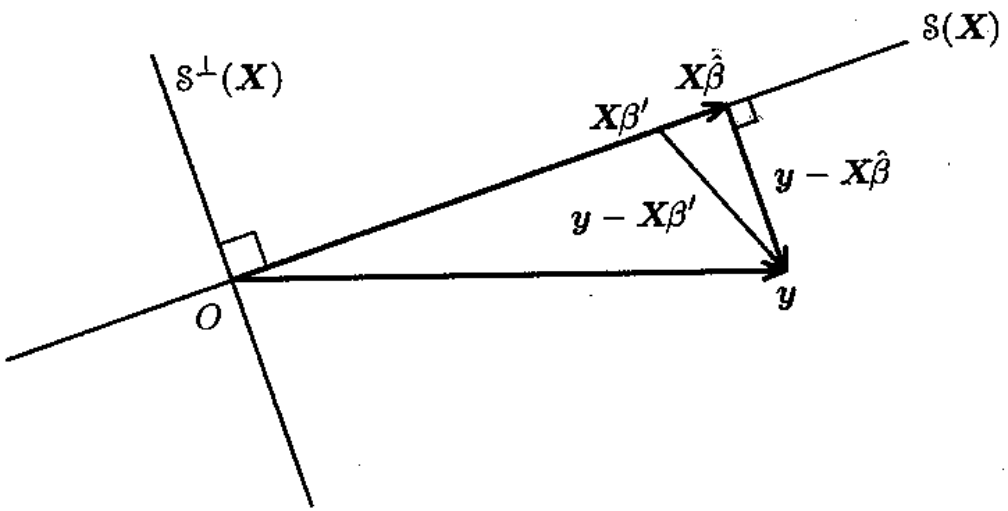


Figure 1.2 The projection of y onto $\mathcal{S}(\boldsymbol{X})$

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

Derivation with Calculus

1. Vector calculus: Let a, b be column vectors and $\mathbf{A}$ be a matrix
   1. $\frac{\partial a^\top b}{\partial b} = \frac{\partial}{\partial b}(a_1 b_1 + a_2 b_2 + \cdots + a_K b_K) = a$
   2. $\frac{\partial \mathbf{A} b}{\partial b} = \mathbf{A}^\top$
   3. $\frac{\partial b^\top \mathbf{A} b}{\partial b} = 2\mathbf{A}b$ when $\mathbf{A}$ is symmetric
2. $\text{SSR} = \|Y - \mathbf{X}\hat{\beta}\|^2 = Y^\top Y - 2Y^\top \mathbf{X}\hat{\beta} + \hat{\beta}^\top \mathbf{X}^\top \mathbf{X}\hat{\beta}$
3. First order condition:
   $\frac{\partial \text{SSR}}{\partial \hat{\beta}} = 0 \Longrightarrow (\mathbf{X}^\top \mathbf{X})\hat{\beta} = \mathbf{X}^\top Y$ (normal equation)
4. Second order condition:
   $\frac{\partial^2 \text{SSR}}{\partial \hat{\beta} \partial \hat{\beta}^\top} = \mathbf{X}^\top \mathbf{X} \geq 0$ in a matrix sense
5. Theorem: A square matrix $\mathbf{A}$ is positive semi-definite if $\mathbf{A}$ is symmetric and $c^\top \mathbf{A} c \geq 0$ for any column vector c

   $\mathbf{X}^\top \mathbf{X}$ is symmetric and $c^\top \mathbf{X}^\top \mathbf{X} c = \|\mathbf{X}c\|^2 \geq 0 \Longrightarrow \mathbf{X}^\top \mathbf{X}$ is positive semi-definite

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 **B**est **L**inear **U**nbiased **E**stimator
- A linear estimator: $\tilde{\beta} = \mathbf{A}Y = \hat{\beta} + \mathbf{B}Y$ where $\mathbf{B} = \mathbf{A} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$
- $\tilde{\beta} = \{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{B}\}Y = \beta + \{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{B}\}\epsilon + \mathbf{B}\mathbf{X}\beta$
- $\mathbb{E}(\tilde{\beta} \mid \mathbf{X}) = \mathbb{E}(\hat{\beta} \mid \mathbf{X}) = \beta$ and exogeneity imply $\mathbf{B}\mathbf{X} = \mathbf{0}$

$$
\begin{aligned}
\mathbb{V}(\tilde{\beta} \mid \mathbf{X}) &= \{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{B}\}\mathbb{V}(\epsilon \mid \mathbf{X})\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{B}\}^\top \\
&= \sigma^2\{(\mathbf{X}^\top\mathbf{X})^{-1} + \mathbf{B}\mathbf{B}^\top\} \\
&\geq \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= \mathbb{V}(\hat{\beta} \mid \mathbf{X})
\end{aligned}
$$

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

More Geometry

- Orthogonal projection matrix or "Hat" matrix:

$$\widehat{Y} = \mathbf{X}\hat{\beta} = \underbrace{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top}_{\mathbf{P_X}} Y = \mathbf{P_X} Y$$

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

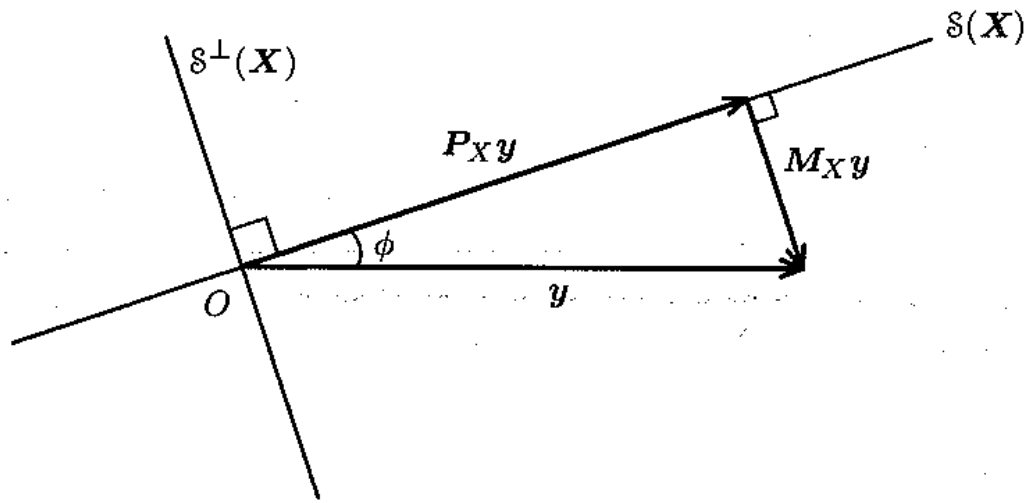


Figure 1.3 The orthogonal decomposition of y

- Symmetric: $\mathbf{P_X} = \mathbf{P_X}^\top$ and $\mathbf{M_X} = \mathbf{M_X}^\top$
- Idempotent: $\mathbf{P_X}\mathbf{P_X} = \mathbf{P_X}$ and $\mathbf{M_X}\mathbf{M_X} = \mathbf{M_X}$
- Annihilator: $\mathbf{M_X}\mathbf{P_X} = \mathbf{P_X}\mathbf{M_X} = \mathbf{0}$

Residuals

- Projection of Y onto $\mathcal{S}^{\top}(\mathbf{X})$:

$$\hat{\epsilon} \;=\; Y - \mathbf{X}\hat{\beta} = \mathbf{M}_{\mathbf{X}} Y$$

- Orthogonality:
  1. $\left\langle \widehat{Y},\ \hat{\epsilon} \right\rangle = \langle \mathbf{P}_{\mathbf{X}} Y,\ \mathbf{M}_{\mathbf{X}} Y \rangle = 0$
  2. $\langle X_k,\ \hat{\epsilon} \rangle = \langle \mathbf{P}_{\mathbf{X}} X_k,\ \mathbf{M}_{\mathbf{X}} Y \rangle = 0$ for any column X_k of $\mathbf{X}$
  3. More generally, $x \cdot \hat{\epsilon} = 0$ for any $x \in \mathcal{S}(\mathbf{X})$
- Zero mean: $\sum_{i=1}^{n} \hat{\epsilon}_i = 0$ since $x_1 = (1, \ldots, 1) \in \mathcal{S}(\mathbf{X})$
- (Sample and population) correlation between $\hat{\epsilon}_i$ and x_{ik} is 0 for any column k
- Does not imply $\mathbb{E}(\epsilon \mid \mathbf{X}) = 0$, $\mathbb{E}(x_k \cdot \epsilon) = 0$ or $\mathrm{Cor}(\epsilon_i, x_{ik})$

The Coefficient of Determination

- The (*centered*) coefficient of determination:

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

- Recall $Y = \mathbf{X}\hat{\beta} + \hat{\epsilon}$, $\left\langle \widehat{Y}, \hat{\epsilon} \right\rangle = 0$, and $\left\langle \overline{Y}\mathbf{1}, \hat{\epsilon} \right\rangle = 0$
- Pythagoras' Theorem:

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

- Note $\overline{Y} = \overline{\mathbf{X}\hat{\beta}}$
- Thus, $\mathbb{V}(Y_i) = \mathbb{V}(X_i\hat{\beta}) + \mathbb{V}(\hat{\epsilon}_i)$ (holds even in sample)

Variance Estimation Under Homoskedasticity

- Under homoskedasticity, standard variance estimator is:

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

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

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

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

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

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

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

Model-Based Prediction

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

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

- Predicted value for $X_i = x$:

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

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

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

Causal Inference with Interaction Terms

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

$$\beta + x^\top \delta$$

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

$$(x^* - x)^\top \delta$$

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

Model-Based Finite Sample Inference

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

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

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

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

Derivation (see Proposition 1.3 of Hayashi)

1. Note that if $x \sim \mathcal{N}(0, \mathbf{I}_K)$, then $x^\top \mathbf{A} x \sim \chi^2_\nu$ where $\nu = \text{rank}(\mathbf{A})$
2. If $\mathbf{A}$ is a projection matrix, then $\text{rank}(\mathbf{A}) = \text{trace}(\mathbf{A})$
   - Proof: Let $\mathbf{A}v = \lambda v$. Then, $\mathbf{A}v = \mathbf{A}^2 v = \lambda \mathbf{A} v = \lambda^2 v \Longrightarrow \lambda = 1$. Thus, $\text{trace}(\mathbf{A})$ equals the # of non-zero eigenvalues, i.e., $\text{rank}(\mathbf{A})$
3. Recall $\hat{\epsilon} = \mathbf{M}_\mathbf{X} \epsilon$
4. Given $\mathbf{X}$, $(n-K)\hat{\sigma}^2/\sigma^2 = \|\mathbf{M}_\mathbf{X}\epsilon/\sigma\|^2 \sim \chi^2_{\text{rank}(\mathbf{M}_\mathbf{X})} = \chi^2_{\text{trace}(\mathbf{M}_\mathbf{X})} = \chi^2_{n-K}$
5. Recall $\hat{\beta} = \beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \epsilon$
6. $\text{Cov}(\hat{\beta}, \hat{\epsilon} \mid \mathbf{X}) = \text{Cov}(\hat{\beta} - \beta, \mathbf{M}_\mathbf{X}\epsilon \mid \mathbf{X}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}(\epsilon\epsilon^\top \mid \mathbf{X}) \mathbf{M}_\mathbf{X} = 0$
7. Theorem: If $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$, then $\mathbf{A}\mathbf{X} \sim \mathcal{N}(\mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}^\top)$
8. Thus, $(\hat{\beta}, \hat{\epsilon})$ have a multivariate Normal distribution given $\mathbf{X}$
9. $\hat{\beta}$ and $\hat{\epsilon}$ are independent given $\mathbf{X}$

Testing Linear Null Hypothesis

- Null hypothesis $H_0 : \mathbf{A}\beta = b$ where $\mathbf{A}$ is of full row rank
- Any linear restriction: $a_1\beta_1 + a_2\beta_2 + \cdots + a_K\beta_K = b$
- Any number of linearly independent such restrictions
- *F*-statistic:

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

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

Influential Observations and Leverage Points

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

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

  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 $\mathcal{S}(\mathbf{X})$
- $0 \leq p_i \leq \|v(i)\|^2 = 1$
- $\bar{p} \equiv \sum_{i=1}^n p_i/n = \text{trace}(\mathbf{P_X})/n = K/n$

- How much one observation can alter the estimate?
- OLS estimate without the ith observation:

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

- Influential points: (1) high leverage, (2) outlier, and (3) both
- Cook's distance: $D_i \equiv (\hat{\beta}_{(i)} - \hat{\beta})^\top \mathbf{X}^\top \mathbf{X} (\hat{\beta}_{(i)} - \hat{\beta}) / (\hat{\sigma}^2 K)$

Model-Based Asymptotic Inference

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

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

- Asymptotic distribution and inference:

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

$$\xrightarrow{d} \mathcal{N}\left(0,\, \sigma^2 \{\mathbb{E}(X_i X_i^\top)\}^{-1}\right)$$

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

Robust Standard Errors

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

Asymptotic Tests

- Null hypothesis $H_0 : \mathbf{A}\beta = b$ where $\mathbf{A}$ is of full row rank
- Any linear restriction: $a_1\beta_1 + a_2\beta_2 + \cdots + a_K\beta_K = b$
- Any number of linearly independent such restrictions
- Wald statistic:

$$
\begin{aligned}
W &\equiv (\mathbf{A}\hat{\beta} - b)^\top \{\mathbf{A}\widehat{\mathbb{V}(\hat{\beta})}\mathbf{A}^\top\}^{-1}(\mathbf{A}\hat{\beta} - b) \\
&= \underbrace{\sqrt{n}(\mathbf{A}\hat{\beta} - b)^\top}_{\xrightarrow{d}\mathcal{N}(0,\, n\mathbf{A}\mathbb{V}(\hat{\beta})\mathbf{A}^\top)} \underbrace{\{n\mathbf{A}\widehat{\mathbb{V}(\hat{\beta})}\mathbf{A}^\top\}^{-1}}_{\xrightarrow{p}\{n\mathbf{A}\mathbb{V}(\hat{\beta})\mathbf{A}^\top\}^{-1}} \underbrace{\sqrt{n}(\mathbf{A}\hat{\beta} - b)}_{\xrightarrow{d}\mathcal{N}(0,\, n\mathbf{A}\mathbb{V}(\hat{\beta})\mathbf{A}^\top)} \\
&\xrightarrow{d} \chi^2_{\text{rank}(\mathbf{A})}
\end{aligned}
$$

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

Generalized Least Squares (GLS)

- Known heteroskedasticity: $\mathbb{V}(\epsilon \mid \mathbf{X}) = \sigma^2\Omega$ where Ω is a positive definite matrix
- OLS estimator is no longer **BLUE**
- GLS estimator: $\hat{\beta}_{GLS} = (\mathbf{X}^\top\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\Omega^{-1}Y$
- $\hat{\beta}_{GLS}$ is **BLUE**
- Consider the transformed regression: $Y^* = \mathbf{X}^*\beta + \epsilon^*$ where $Y^* = \Omega^{-1/2}Y$, $\mathbf{X}^* = \Omega^{-1/2}\mathbf{X}$, and $\epsilon^* = \Omega^{-1/2}\epsilon$
- Cholesky decomposition: $\Omega = \Omega^{1/2}\Omega^{1/2^\top}$ where $\Omega^{1/2}$ is a lower triangular matrix with strictly positive diagonal elements
- Variance: $\mathbb{V}(\hat{\beta}_{GLS} \mid \mathbf{X}) = \sigma^2(\mathbf{X}^\top\Omega^{-1}\mathbf{X})^{-1}$ 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. $\hat{\beta}_{FGLS} = (\mathbf{X}^\top\widehat{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\widehat{\Omega}^{-1}Y$
  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
- $w_i =$ Sampling weights $= 1/\Pr(\text{being selected into the sample})$
- 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 U: $Y = \mathbf{X}\beta + \underbrace{U\gamma + \epsilon^*}_{\epsilon}$ with $\mathbb{E}(\epsilon^* \mid \mathbf{X}, U) = 0$
- Decomposition: $U = \mathbf{P}_{\mathbf{X}} U + \hat{\eta} = \mathbf{X}\hat{\delta} + \hat{\eta}$ where $\mathbf{X}^\top \hat{\eta} = 0$
- Then,

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

  with $\mathbb{E}(\gamma\hat{\eta}_i + \epsilon_i^*) = 0$ and $\mathbb{E}\{X_i(\gamma\hat{\eta}_i + \epsilon_i^*)\} = 0$
- $\hat{\beta}_k \xrightarrow{p} \beta_k + \gamma\delta_k$ where $\hat{\delta} \xrightarrow{p} \delta$
- If $\delta_{k'} = 0$ for all $k' \neq k$, then

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

Measurement Error

- Three types of ME: classical, nondifferential, differential

- ME in Y_i: $Y_i = Y_i^* + e_i$
- Omitted variable problem: $Y_i = X_i^\top \beta + e_i + \epsilon_i$

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

Model Assessment and Selection

- Need a criteria to assess model fitting
- R^2: risk of overfitting
- Adjusted R^2 (coefficient of determination):

$$R^2_{\text{adj}} = 1 - \frac{\widehat{\mathbb{V}(\epsilon_i)}}{\widehat{\mathbb{V}(Y_i)}} = 1 - \frac{\frac{1}{n-K}\sum_{i=1}^n \hat{\epsilon}_i^2}{\frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2}$$

$$= R^2 - \underbrace{(1 - R^2)\frac{K-1}{n-K}}_{\text{model complexity penalty}}$$

- Alternative: in-sample average expected error for new observations:

$$\overline{\text{Error}} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\{(Y_i^{\text{new}} - X_i^\top \hat{\beta})^2 \mid \mathbf{X}, Y\}$$

where the expectation is taken over Y_i^{new} given all the data $(\mathbf{X}, Y)$

The C_p Statistic

- Compare $\overline{\text{Error}}$ with the average SSR under the homoskedasticity assumption:

$$\begin{aligned}\mathbb{E}\left(\overline{\text{Error}} - \frac{1}{n}\text{SSR} \,\middle|\, \mathbf{X}\right) &= \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\{(Y_i - X_i^\top\beta)(X_i\hat{\beta} - X_i^\top\beta) \mid \mathbf{X}\} \\ &= \frac{2}{n}\sum_{i=1}^{n}\text{Cov}(Y_i, \widehat{Y}_i \mid \mathbf{X})\end{aligned}$$

  where the expectation is taken with respect to Y given $\mathbf{X}$
- One can show:

$$\frac{2}{n}\sum_{i=1}^{n}\text{Cov}(Y_i, \widehat{Y}_i \mid \mathbf{X}) = \frac{2}{n}\text{trace}\{\text{Cov}(Y, \mathbf{P_X}Y \mid \mathbf{X})\} = \frac{2\sigma^2 K}{n}$$

- The C_p statistic is defined:

$$\widehat{\overline{\text{Error}}} = \frac{1 + K/n}{n - K}\text{SSR}$$

Key Points

- 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