Projections and Influence

Econometrics II

Douglas G. Steigerwald

UC Santa Barbara

Overview

Reference: B. Hansen Econometrics Chapter 3.10-3.18

- Projection Matrix (Hat Matrix)
- Orthogonal Projection Matrix (Annihilator Matrix)
- How do we estimate σ^2?
- Predicted Values
- Leverage and Influence

Projection Matrix

- projection (hat) matrix

$$\underset{n \times n}{P} = X \left(X^{\mathrm{T}} X \right)^{-1} X^{\mathrm{T}}$$

- why projection?
  - $PX = X \left(X^{\mathrm{T}} X \right)^{-1} X^{\mathrm{T}} X = X$
    - ★ holds for any matrix in the range space of X
- why hat?
  - $Py = X \left(X^{\mathrm{T}} X \right)^{-1} X^{\mathrm{T}} y = X\widehat{\beta} := \widehat{y}$
    - ★ creates fitted values
    - ★ $X = \mathbf{1}$ (n vector of ones) $P = \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}$
    - ★ $Py = \mathbf{1}\overline{y}$ (fitted value is the sample mean)

Projection Matrix Properties

- range space of X consists of matrices formed from columns of X
  - $Z = X\Gamma$ for some matrix Γ

$$PZ = PX\Gamma = X\Gamma = Z$$

- important example, partition $X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$
  - $PX_1 = X_1$
- projection matrix is symmetric

$$P^{\mathrm{T}} = P$$

- projection matrix is idempotent

$$PP = P$$

  - $PX = X$ implies

$$PP = PX\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}} = P$$

Projection Matrix Symmetry (Student Annotation)

Leverage

- i^{th} diagonal element of P

$$h_{ii} = x_i^{\mathrm{T}} \left(X^{\mathrm{T}} X \right)^{-1} x_i$$

  - *leverage* of observation i
  - property 1: $0 \leq h_{ii} \leq 1$
  - property 2: $\sum_{i=1}^{n} h_{ii} = k$

Proof of Property 2

Orthogonal Projection

- orthogonal projection matrix (annihilator matrix)

$$M = I_n - P$$

- why orthogonal projection?
  - $MX = 0$ therefore M and X are orthogonal
- why annihilator matrix?
  - for any matrix Z in the range space of X
    - ★ $MZ = Z - PZ = 0$
  - examples
    - ★ $MX_1 = 0$
    - ★ $MP = 0$
- M creates least squares residuals

$$My = y - Py = y - \widehat{y} = \widehat{u}$$

Properties of Orthogonal Projection

- M satisfies:
  - symmetric $M^{\mathrm{T}} = I_n^{\mathrm{T}} - P^{\mathrm{T}} = M$
  - idempotent $MM = M\left(I_n - P\right) = M$
  - $tr\left(M\right) = n - k$
- special example $X = \mathbf{1}$
  - $M_1 y = \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}\right) = \mathbf{y} - \mathbf{1}\overline{y}$
    - ★ demeaned values
- $\widehat{u} = My = M\left(X\beta + u\right) = Mu$
  - free of dependence on the regression coefficient β

Estimation of the Error Variance (Student Annotation)

An Interesting Fact Regarding the Variance Estimator

consider

$$\begin{aligned}
\widetilde{\sigma}^2 - \widehat{\sigma}^2 &= n^{-1}u^{\mathrm{T}}u - n^{-1}\widehat{u}^{\mathrm{T}}\widehat{u} \\
&= n^{-1}u^{\mathrm{T}}\left(I_n - M\right)u \\
&= n^{-1}u^{\mathrm{T}}Pu \\
&\geq 0
\end{aligned}$$

- the last inequality holds because
  - P is positive semidefinite
  - $u^{\mathrm{T}}Pu$ is a quadratic form
- feasible estimator is numerically smaller than ideal estimator

Analysis of Variance: Orthogonal Decomposition

- orthogonal decomposition

$$y = Py + My := \widehat{y} + \widehat{u}$$

  - orthogonal because $\widehat{y}^{\mathrm{T}}\widehat{u} = y^{\mathrm{T}}PMy = 0$

- it follows that

$$y^{\mathrm{T}}y = \widehat{y}^{\mathrm{T}}\widehat{y} + \widehat{u}^{\mathrm{T}}\widehat{u}$$

  - or

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \widehat{y}_i^2 + \sum_{i-1}^{n} \widehat{u}_i^2$$

Analysis of Variance Formula

- subtracting $\overline{y}$ from both sides of the decomposition

$$y - 1\overline{y} = (\widehat{y} - 1\overline{y}) + \widehat{u}$$

- orthogonal decomposition when X contains a constant: $1^{\mathrm{T}}\widehat{u} = 0$
- $(y - 1\overline{y})^{\mathrm{T}} (y - 1\overline{y}) = (\widehat{y} - 1\overline{y})^{\mathrm{T}} (\widehat{y} - 1\overline{y}) + \widehat{u}^{\mathrm{T}}\widehat{u}$
- analysis of variance formula for LS regression

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 + \sum_{i-1}^{n} \widehat{u}_i^2$$

- coefficient of determination (algebraic measure of fit, we have better measures that require statistical derivation)

$$R^2 = \frac{\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = 1 - \frac{\sum_{i-1}^{n} \widehat{u}_i^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

Regression Components

- partition $\quad X = [X_1 \quad X_2]$
- OLS regression of y on X yields
  - $y = X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2 + \widehat{u}$
- algebraic expressions for $\widehat{\beta}_1$ and $\widehat{\beta}_2$ identical to algebra for population coefficients

$$
\begin{aligned}
\widehat{\beta}_1 &= \left(X_1^{\mathrm{T}} M_2 X_1\right)^{-1} \left(X_1^{\mathrm{T}} M_2 y\right) \\
\widehat{\beta}_2 &= \left(X_2^{\mathrm{T}} M_1 X_2\right)^{-1} \left(X_2^{\mathrm{T}} M_1 y\right)
\end{aligned}
$$

  - $M_1 = I_n - X_1 \left(X_1^{\mathrm{T}} X_1\right)^{-1} X_1^{\mathrm{T}}$
  - $M_2 = I_n - X_2 \left(X_2^{\mathrm{T}} X_2\right)^{-1} X_2^{\mathrm{T}}$
  - $\widehat{\beta}_1$ - projection onto M_2 removes component correlated with X_2
    - in essence, "holding X_2 constant"

Matrix Algebra Derivation

Residual Regression

First recognized by Frisch and Waugh (1933)

- because $M_1 = M_1 M_1$

$$\begin{aligned}\widehat{\beta}_2 &= \left(X_2^{\mathrm{T}} M_1 M_1 X_2\right)^{-1}\left(X_2^{\mathrm{T}} M_1 M_1 y\right)\\ &= \left(\widetilde{X}_2^{\mathrm{T}}\widetilde{X}_2\right)^{-1}\left(\widetilde{X}_2^{\mathrm{T}}\overline{u}_1\right)\end{aligned}$$

  - $\widetilde{X}_2 = M_1 X_2 \qquad \overline{u}_1 = M_1 y$
  - proves the following theorem

Theorem (Frisch-Waugh-Lovell). In the linear model $y = X_1\beta_1 + X_2\beta_2 + u$ the OLS estimator of β_2 and the OLS residuals $\widehat{u}$ may be equivalently computed by either the OLS regression or via the following algorithm:

1. *Regress y on X_1, obtain residuals $\overline{u}_1$;*
2. *Regress X_2 on X_1, obtain residuals $\widetilde{X}_2$;*
3. *Regress $\overline{u}_1$ on $\widetilde{X}_2$, obtain OLSE $\widehat{\beta}_2$ and residuals $\widehat{u}$.*

Residual Regression Continued

- the estimated coefficient $\widehat{\beta}_2$ numerically equals the regression of y on the covariates X_2 after the covariates X_1 have been linearly projected out
- important example (deviations from means): $X_1 = 1 \quad X_2$ the observed covariates
  - $M_1 = I_n - 1\left(1^{\mathrm{T}}1\right)^{-1} 1^{\mathrm{T}}$
  - $\widetilde{X}_2 = X_2 - \overline{X}_2 \qquad \overline{u}_1 = y - \overline{y}$

$$\widehat{\beta}_2 = \left(\sum_{i=1}^{n} (x_{2i} - \overline{x}_2)(x_{2i} - \overline{x}_2)^{\mathrm{T}}\right)^{-1} \left(\sum_{i=1}^{n} (x_{2i} - \overline{x}_2)(y_i - \overline{y})\right)$$

- Ragnar Frisch:
  - co-winner (with Jan Tinbergen) of 1^{st} Nobel prize in Economics in 1969
  - formalized consumer, producer, and business cycle theory

Prediction Errors

- $\widehat{u}_i$ constructed from full sample, including y_i
  - not a prediction error
  - proper prediction should exclude y_i
- leave-one-out estimator excludes y_i

$$\begin{aligned}
\widehat{\beta}_{(-i)} &= \left(\frac{1}{n-1}\sum_{j\neq i} x_j x_j^{\mathrm{T}}\right)^{-1}\left(\frac{1}{n-1}\sum_{j\neq i} x_j y_j\right) \\
&= \left(X_{(-i)}^{\mathrm{T}} X_{(-i)}\right)^{-1}\left(X_{(-i)}^{\mathrm{T}} y_{(-i)}\right) \qquad \text{note } X_{(-i)} \text{ excludes row } i
\end{aligned}$$

- leave-one-out predicted value $\widetilde{y}_i = x_i^{\mathrm{T}}\widehat{\beta}_{(-i)}$
- prediction error (residual) $\widetilde{u}_i = y_i - \widetilde{y}_i$
- sample mean squared prediction error $n^{-1}\sum_{i=1}^{n}\widetilde{u}_i^2$

Prediction Error Construction

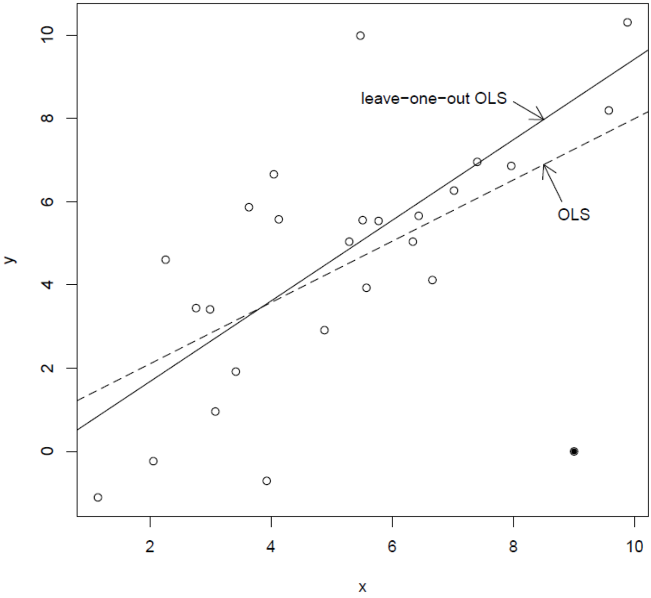
- convenient expression: $\widehat{\beta}_{(-i)} = \widehat{\beta} - (1 - h_{ii})^{-1} \left(X^{\mathrm{T}} X\right)^{-1} x_i \widehat{u}_i$
  - recall, leverage value (scalar) $h_{ii} = x_i^{\mathrm{T}} \left(X^{\mathrm{T}} X\right)^{-1} x_i$
- resulting simplified expression for prediction error, $\widetilde{u}_i =$
  - $= y_i - x_i^{\mathrm{T}} \widehat{\beta}_{(-i)}$
  - $= y_i - x_i^{\mathrm{T}} \widehat{\beta} + (1 - h_{ii})^{-1} x_i^{\mathrm{T}} \left(X^{\mathrm{T}} X\right)^{-1} x_i \widehat{u}_i$
  - $= \left(1 + (1 - h_{ii})^{-1} h_{ii}\right) \widehat{u}_i$
  - $= (1 - h_{ii})^{-1} \widehat{u}_i$
- for $M^* \overset{def}{=} diag \left\{ (1 - h_{11})^{-1}, \ldots, (1 - h_{nn})^{-1} \right\}$

$$\widetilde{u} = M^* \widehat{u}$$

  - computation of $\widetilde{u}$ does not require n estimations

Influential Observations

- influential if omission of observation induces a substantial change in the estimate
- example: consider the following figure with data generated as
  - $x_i \sim U[1, 10] \quad y_i \sim \mathcal{N}(x_i, 4)$
  - outlier $x_{26} = 9 \quad y_{26} = 0$
  - note: must examine joint behavior to detect outlier
    - ★ neither x_{26} nor y_{26} are unusual relative to their marginal distributions


leave−one−out OLS
OLS
y
x
0
2
4
6
8
10
2
4
6
8
10

Calculation of Influence

- for coefficients of interest, calculate for each i
  - $\widehat{\beta} - \widehat{\beta}_{(-i)} = \left(X^{\mathrm{T}} X\right)^{-1} x_i \widetilde{u}_i \qquad \widetilde{u}_i = (1 - h_{ii})\,\widehat{u}_i$
    - DFBETA - post estimation diagnostic in STATA
    - Is there a meaningful change? (no magic threshold)
    - hard to recommend other proposed diagnostics (DFITS, Cook's Distance, Welsch Distance) - not based on statistical theory
- for general assessment, study predicted value
  - $\mathit{Influence} = \max_{1 \le i \le n} \left| \widehat{y}_i - \widetilde{y}_i \right|$
  - $\widehat{y}_i - \widetilde{y}_i = x_i^{\mathrm{T}} \widehat{\beta} - x_i^{\mathrm{T}} \widehat{\beta}_{(-i)} = h_{ii} \widetilde{u}_i$
  - observation i is influential for the predicted value if h_{ii} and $\left|\widetilde{u}_i\right|$ are large
    - h_{ii} large - x_i is far from its sample mean, leverage point
    - leverage points are not necessarily influential

What to do with Influential Observations?

- due to data entry error, delete, termed "cleaning the data"
  - e.g. individual who is employed but has $0 earnings
  - requires judgement, therefore proper empirical practice
    - keep: source data in original form, revised data after cleaning, record describing the cleaning process
- not due to data entry error
  - do nothing, or alter the specification to properly model the influential observation
  - delete the observation - reduces the integrity of the results (viewed skeptically)

Influential Observation Example

- log wage regression for single Asian males
- $n = 268 \quad \mathit{Influence} = 0.29$
  - most influential observation, when included, changes a fitted value of log wage by 0.29, or the wage by 29%!
- for this observation $h_{ii} = 0.33$ (recall, h_{ii} positive and sum to 1)
  - 1/3 of the leverage for the entire sample is contained in this observation
  - individual is 65 years old, 8 years of education, thus 51 years of (potential) experience
  - next highest level of experience is 41 years
- essentially estimating the conditional mean of *experience=51* with only 1 observation
  - solution, estimate over a smaller range of experience, restrict sample to $\mathit{experience} \leq 45$
  - $\widehat{\log \mathit{wage}} = 0.144 \mathit{ed} + 0.043 \mathit{exp} - 0.095 \mathit{exp}^2/100 + 0.531$
  - coefficient on exp and $\exp^2$ increase slightly and $\mathit{Influence} = 0.11$
  - more robust estimate of conditional mean for most levels of experience
- Which to report? A matter of judgement

Normal Regression Model

- linear regression model with u_i independent of x_i with a normal distribution
  - $u_i | x_i \sim \mathcal{N}\left(0, \sigma^2\right)$ which implies $y_i | x_i \sim \mathcal{N}\left(x_i^{\mathrm{T}}\beta, \sigma^2\right)$
- log-likelihood function

$$\begin{aligned} \log L\left(\beta, \sigma^2\right) &= \sum_{i=1}^{n} \log \frac{1}{\left(2\pi\sigma^2\right)^{\frac{1}{2}}} \exp\left(\frac{-1}{2\sigma^2}\left(y_i - x_i^{\mathrm{T}}\beta\right)\right)^2 \\ &= -\frac{n}{2}\log\left(2\pi\right) - \frac{n}{2}\log\left(\sigma^2\right) + \frac{1}{2\sigma^2} SSE_n\left(\beta\right) \end{aligned}$$

  - β enters only through $SSE_n(\beta)$ thus $\widehat{\beta}_{mle} = \widehat{\beta}_{ols}$
  - $\widehat{\sigma}^2_{mle}$ - maximize $\log L\left(\widehat{\beta}_{mle}, \sigma^2\right)$
    - ★ FOC

$$\frac{\partial}{\partial \sigma^2} \log L\left(\widehat{\beta}_{mle}, \widehat{\sigma}^2\right) = -\frac{n}{2}\frac{1}{\widehat{\sigma}^2} + \frac{SSE_n\left(\widehat{\beta}_{mle}\right)}{2\left(\widehat{\sigma}^2\right)^2} = 0$$

    - ★ $\widehat{\sigma}^2_{mle} = \frac{1}{n}\sum \widehat{u}_i^2$

Normal (Gaussian) Regression Model 2

- the sample value of the log-likelihood

$$\log L\left(\widehat{\beta}_{mle}, \widehat{\sigma}^2_{mle}\right) = -\frac{n}{2}\left(\log\left(2\pi\right) + 1\right) - \frac{n}{2}\log\left(\widehat{\sigma}^2_{mle}\right)$$

  - this value, or the negative of this value, is reported as a measure of fit
- no surprise that $\widehat{\beta}_{mle} = \widehat{\beta}_{ols}$ - most loss functions have an ML equivalent
- Carl Friedrich Gauss (1777-1855) mathematician
  - proposed normal regression model, derived the OLSE as the MLE
  - claims to have discovered this in 1795 at the age of eighteen
  - not published until 1809
  - interest in the result reinforced by Laplace's simultaneous discovery of the CLT, which provided justification for viewing random disturbances as approximately normal

Proof of Projection Matrix Property 2

$$
\begin{aligned}
tr\,(P) &= tr\left(X\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}\right) \\
&= tr\left(\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}X\right) \\
&= tr\,(I_k) \\
&= k
\end{aligned}
$$

Return to Leverage

Derivation of Matrix Components

$$\widehat{Q}_{xx} = \begin{bmatrix} \widehat{Q}_{11} & \widehat{Q}_{12} \\ \widehat{Q}_{21} & \widehat{Q}_{22} \end{bmatrix} = \begin{bmatrix} n^{-1} X_1^{\mathrm{T}} X_1 & n^{-1} X_1^{\mathrm{T}} X_2 \\ n^{-1} X_2^{\mathrm{T}} X_1 & n^{-1} X_2^{\mathrm{T}} X_2 \end{bmatrix}$$

$$\widehat{Q}_{xy} = \begin{bmatrix} \widehat{Q}_{1y} \\ \widehat{Q}_{2y} \end{bmatrix} = \begin{bmatrix} n^{-1} X_1^{\mathrm{T}} y \\ n^{-1} X_2^{\mathrm{T}} y \end{bmatrix}$$

- partitioned matrix inversion formula yields

$$\widehat{Q}_{xx}^{-1} \overset{def}{=} \begin{bmatrix} \widehat{Q}^{11} & \widehat{Q}^{12} \\ \widehat{Q}^{21} & \widehat{Q}^{22} \end{bmatrix} = \begin{bmatrix} \widehat{Q}_{11\cdot 2}^{-1} & -\widehat{Q}_{11\cdot 2}^{-1} \widehat{Q}_{12} \widehat{Q}_{22}^{-1} \\ -\widehat{Q}_{22\cdot 1}^{-1} \widehat{Q}_{21} \widehat{Q}_{11}^{-1} & \widehat{Q}_{22\cdot 1}^{-1} \end{bmatrix}$$

$$\begin{aligned} \widehat{\beta}_1 &= \widehat{Q}_{11\cdot 2}^{-1} \left(\frac{1}{n} X_1^{\mathrm{T}} y - \frac{1}{n} X_1^{\mathrm{T}} X_2 \left(\frac{1}{n} X_2^{\mathrm{T}} X_2 \right)^{-1} \frac{1}{n} X_2^{\mathrm{T}} y \right) \\ &= \widehat{Q}_{11\cdot 2}^{-1} \left(\frac{1}{n} X_1^{\mathrm{T}} M_2 y \right) \end{aligned}$$

Derivation Continued

$$\widehat{Q}_{11\cdot 2} = \widehat{Q}_{11} - \widehat{Q}_{12}\widehat{Q}_{22}^{-1}\widehat{Q}_{21}$$
$$\widehat{Q}_{22\cdot 1} = \widehat{Q}_{22} - \widehat{Q}_{21}\widehat{Q}_{11}^{-1}\widehat{Q}_{12}$$

$$\begin{aligned}
\widehat{Q}_{11\cdot 2} &= \widehat{Q}_{11} - \widehat{Q}_{12}\widehat{Q}_{22}^{-1}\widehat{Q}_{21} \\
&= \frac{1}{n}X_1^{\mathrm{T}}X_1 - \frac{1}{n}X_1^{\mathrm{T}}X_2\left(\frac{1}{n}X_2^{\mathrm{T}}X_2\right)^{-1}\frac{1}{n}X_2^{\mathrm{T}}X_1 \\
&= \frac{1}{n}X_1^{\mathrm{T}}M_2X_1
\end{aligned}$$

- therefore

$$\begin{aligned}
\widehat{\beta}_1 &= \widehat{Q}_{11\cdot 2}^{-1}\left(\frac{1}{n}X_1^{\mathrm{T}}M_2y\right) \\
&= \left(X_1^{\mathrm{T}}M_2X_1\right)^{-1}X_1^{\mathrm{T}}M_2y
\end{aligned}$$

Return to Regression Components