Econ 2120: Section 6

The Normal Linear Model

Ashesh Rambachan

Fall 2018

Outline

The Normal Linear Model

- The Normal Distribution
- Set Up – Normal Linear Model
- Canonical Form of the Normal Linear Model
- Least Squares and the Canonical Form
- Confidence Intervals & Ellipses
- F-tests and the Canonical Form

Dominating Least Squares

- The Bias-Variance Tradeoff
- The Decision Problem
- Stein's Unbiased Estimate of Risk (SURE)
- James & Stein (61)
- Prediction
- Preserving our unbiased estimator

Motivation

These lectures on the Normal-Linear Model are my personal favorite part of this course.

There are so many insights packed into a very simple model.

The techniques we'll cover make the proofs incredibly simple and are generally very useful.

Outline

The Normal Linear Model

- The Normal Distribution
- Set Up – Normal Linear Model
- Canonical Form of the Normal Linear Model
- Least Squares and the Canonical Form
- Confidence Intervals & Ellipses
- F-tests and the Canonical Form

Dominating Least Squares

- The Bias-Variance Tradeoff
- The Decision Problem
- Stein's Unbiased Estimate of Risk (SURE)
- James & Stein (61)
- Prediction
- Preserving our unbiased estimator

Outline

The Normal Linear Model

- The Normal Distribution
- Set Up – Normal Linear Model
- Canonical Form of the Normal Linear Model
- Least Squares and the Canonical Form
- Confidence Intervals & Ellipses
- F-tests and the Canonical Form

Dominating Least Squares

- The Bias-Variance Tradeoff
- The Decision Problem
- Stein's Unbiased Estimate of Risk (SURE)
- James & Stein (61)
- Prediction
- Preserving our unbiased estimator

The Normal Distribution

$Z \sim N(0,1)$ with density

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2)$$

with $E[Z] = 0$, $V(Z) = E[Z^2] = 1$. We say Z is a **standard normal distribution**.

We write

$$W \sim N(\mu, \sigma^2)$$

if $W = \mu + \sigma Z$, $Z \sim N(0,1)$.

Joint Normal Distribution

If W is $n \times 1$ random vector with a **joint normal distribution**, we write $W \sim N_n(\mu, \Sigma)$, where $\underset{n\times 1}{\mu}$ is the mean vector and $\underset{n\times n}{\Sigma}$ is the cov. matrix.

Claim 3: If $a_1 \in \mathbb{R}^{m\times m}$, $a_2 \in \mathbb{R}^m$ and $W \sim N_n(\mu, \Sigma)$, then

$$a_1 W + a_2 \sim N(a_1\mu + a_2, a_1\Sigma a_1').$$

Claim 4: If $V \sim N_n(0, I_n)$ and q is $n \times n$ orthogonal matrix, then

$$qV \sim N(0, I_n).$$

Why? We have that $qV \sim N(0, qI_nq') = N(0, I_n)$.

Outline

The Normal Linear Model

- The Normal Distribution
- Set Up – Normal Linear Model
- Canonical Form of the Normal Linear Model
- Least Squares and the Canonical Form
- Confidence Intervals & Ellipses
- F-tests and the Canonical Form

Dominating Least Squares

- The Bias-Variance Tradeoff
- The Decision Problem
- Stein's Unbiased Estimate of Risk (SURE)
- James & Stein (61)
- Prediction
- Preserving our unbiased estimator

Assumptions

Assume **random sampling** of the data

$$(Y_i, Z_i) \overset{i.i.d}{\sim} F, \;\; i = 1, \ldots, n$$

where Y_i is 1×1.

We additionally assume that

$$E[Y_i | X_i = x_i] = x_i' \beta, \text{ and } V(Y_i | X_i = x_i) = \sigma^2.$$

We assume: (1) conditional expectation is linear; (2) homoskedasticity.

Assumptions

Assume that the conditional distribution of Y_i given $Z_i = z_i$ is **normally distributed**:

$$Y_i | X_i = x_i \sim N(x_i', \sigma^2) \text{ for } i = 1, \ldots, n.$$

> We're making **distributional** assumption. This will allow us to make progress in finite sample.

To write the model, we define

$$U_i = Y_i - x_i'\beta, \;\; V_i = U_i / \sigma$$

and so, $V_i | X = x \sim N(0, 1)$ for $i = 1, \ldots, n$.

> Note that we are conditioning on X, not X_i. This holds by the random sampling assumption.

The Normal Linear Model

So, we can write the **normal linear model** as

$$Y = x\beta + \sigma V, \quad V|X = x \sim N(0, I_n) \tag{1}$$

and

$$\underset{n\times 1}{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \ \underset{n\times K}{x} = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix}, \ \underset{n\times 1}{V} = \begin{pmatrix} V_1 \\ \vdots \\ V_n \end{pmatrix}.$$

Note: The lecture notes introduce the matrix $\underset{n\times n}{Z = z}$. Think of this as: We observe z_i and apply some fixed transformations to x_i.

I will just directly condition on $X = x$ to simplify notation.

Outline

- The Normal Linear Model
  - The Normal Distribution
  - Set Up – Normal Linear Model
  - Canonical Form of the Normal Linear Model
  - Least Squares and the Canonical Form
  - Confidence Intervals & Ellipses
  - F-tests and the Canonical Form

- Dominating Least Squares
  - The Bias-Variance Tradeoff
  - The Decision Problem
  - Stein's Unbiased Estimate of Risk (SURE)
  - James & Stein (61)
  - Prediction
  - Preserving our unbiased estimator

The Canonical Form

We have the normal linear model:

$$Y = x\beta + \sigma V, \quad V|X = x \sim N(0, I_n)$$

Goal: Rewrite this model as

$$Y^* = \begin{pmatrix} \mu \\ 0 \end{pmatrix} + \sigma V^*, \quad V^*|X = x \sim N(0, I_n),$$

where μ is a $K \times 1$ vector of means. This is the **canonical form** of the normal linear model.

> Why? It will REALLY simplify a lot of proofs of classic results.
>
> What's happening when we do that? We are "rotating our data" so that only the first k "rotated observations" are used to estimate b – the least squares coefficients.

Linear algebra results I

A $n \times n$ matrix q is **orthogonal** if $q^{-1} = q'$.

If q is orthogonal, so is q'.

If q is orthogonal, then $q'q = qq' = I_n$.

Claim 1: Let $a_1, a_2 \in \mathbb{R}^n$ and $q \in \mathbb{R}^{n \times n}$ be orthogonal. Then,

$$\langle qa_1, qa_2 \rangle = \langle a_1, a_2 \rangle.$$

Why? $\langle qa_1, qa_2 \rangle = (qa_1)'(qa_2) = a_1' q' q a_2 = a_1' a_2$.

Orthogonal matrices

Orthonormal matrices can be thought of **rotations**.

Example: In $\mathbb{R}^2$, any orthonormal matrix can be represented by

$$q = \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix}$$

for some angle α. This rotates a vector x by angle α.

This geometric intuition generalizes to higher dimensions.

Orthogonal matrices and rotation

Idea of the Canonical Form: One particular rotation that we focus on will rotate x so that all the information is in the first k elements.

Design matrix x has rank $K < n$ – its columns span a K-dimensional space.

We will change the basis of the n-dimensional space so that we can represent all elements of x using only the first K basis vectors of the transformed space.

Linear algebra results II

Claim 2: Singular Value Decomposition

If x is $n \times K$ with $n \geq K$, then there is an orthogonal matrix $q \in \mathbb{R}^{n \times n}$, diagonal matrix $d = diag\{d_1, \ldots, d_K\} \in \mathbb{R}^{K \times K}$ with $d_1 \geq \ldots d_K \geq 0$ and an orthogonal matrix $s \in \mathbb{R}^{K \times K}$ such that

$$x = q_1 d s',$$

where q_1 is the $n \times K$ matrix formed by the first K columns of q with $\underset{n \times n}{q} = \begin{pmatrix} \underset{n \times K}{q_1} & \underset{n \times n-K}{q_2} \end{pmatrix}$.

Going to the Canonical Form

Assume that $n > K$ and that the columns of x are linearly independent.

That is, if $c \in \mathbb{R}^K$ and $xc = 0$, then $c = 0$.

Why do we need this? Implies that $x'x$ is invertible and so, the matrix d in the SVD of x is non-singular with $d_k > 0$ for $k = 1, \ldots, K$.

Using SVD, we have $x = q_1 ds'$. Plug this into normal linear model:

$$Y = (q_1 ds')\beta + \sigma V.$$

Multiply both sides by $q' = \begin{pmatrix} q_1' \\ q_2' \end{pmatrix}$. So,

$$q'Y = \begin{pmatrix} q_1' \\ q_2' \end{pmatrix} q_1 ds'\beta + \sigma q'V$$

Going to the Canonical Form

$$q'Y = \begin{pmatrix} q_1'q_1 \\ q_2'q_1 \end{pmatrix} ds'\beta + \sigma q'V$$

Because q is orthogonal, we have that $q_1'q_1 = I_K$, $q_2'q_1 = 0_{(n-K)\times K}$. So,

$$q'Y = \begin{pmatrix} I_K \\ 0 \end{pmatrix} \mu + \sigma q'V,$$

where we set $\mu = ds'\beta$. Define

$$Y^* = q'Y, \quad V^* = q'V.$$

Note that $V^*|X = x \sim N(0, I_n)$ and we're done!

The Canonical Form

The canonical form is

$$Y^* = \begin{pmatrix} I_K \\ 0 \end{pmatrix} \mu + \sigma V^*, \quad V^* | X = x \sim N(0, I_n). \tag{2}$$

Note that

$$Y^* | X = x \sim N(\begin{pmatrix} \mu \\ 0 \end{pmatrix}, \sigma^2 I_n).$$

The Canonical Form

We do some additional rewriting. Let

$$Y^*_{(1)} = \begin{pmatrix} Y^*_1 \\ \vdots \\ Y^*_K \end{pmatrix}, \quad V^*_{(1)} = \begin{pmatrix} V^*_1 \\ \vdots \\ V^*_K \end{pmatrix}$$

$$Y^*_{(2)} = \begin{pmatrix} Y^*_{K+1} \\ \vdots \\ Y^*_n \end{pmatrix}, \quad V^*_{(2)} = \begin{pmatrix} V^*_{K+1} \\ \vdots \\ V^*_n \end{pmatrix}.$$

Write the **canonical form** as

$$Y^*_{(1)} = \mu + \sigma V^*_{(1)} \tag{3}$$

$$Y^*_{(2)} = \sigma V^*_{(2)}, \tag{4}$$

where components of $V^*_{(1)}, V^*_{(2)}$ are i.i.d $N(0,1)$ conditional on $X = x$.

Outline

The Normal Linear Model

- The Normal Distribution
- Set Up – Normal Linear Model
- Canonical Form of the Normal Linear Model
- Least Squares and the Canonical Form
- Confidence Intervals & Ellipses
- F-tests and the Canonical Form

Dominating Least Squares

- The Bias-Variance Tradeoff
- The Decision Problem
- Stein's Unbiased Estimate of Risk (SURE)
- James & Stein (61)
- Prediction
- Preserving our unbiased estimator

Least Squares and the Canonical Form

Least squares solves

$$b = \arg\min_{c} \ \|Y - xc\|^2$$

Let's write this in terms of the canonical form.

First, note that

$$\begin{aligned} \|Y - xc\|^2 &= \langle Y - xc, Y - xc \rangle \\ &= \langle q'(Y - xc), q'(Y - xc) \rangle = \|q'(Y - xc)\|^2. \end{aligned}$$

Second, note that

$$x = q_1 ds' \implies \|q'(Y - xc)\| = \|Y^* - \begin{pmatrix} ds' \\ 0 \end{pmatrix} c\|^2.$$

Least Squares and the Canonical Form

So, we have

$$\begin{aligned}\left\| \begin{pmatrix} Y^*_{(1)} \\ Y^*_{(2)} \end{pmatrix} - \begin{pmatrix} ds' \\ 0 \end{pmatrix} c \right\|^2 &= \left\| \begin{pmatrix} Y^*_{(1)} - ds'c \\ Y^*_{(2)} \end{pmatrix} \right\|^2 \\ &= \| Y^*_{(1)} - ds'c \|^2 + \| Y^*_{(2)} \|^2.\end{aligned}$$

The least-squares estimate is simple! It is

$$\begin{aligned} b &= (ds')^{-1} Y^*_{(1)} \\ &= \beta + \sigma s d^{-1} V^*_{(1)}, \end{aligned}$$

where we used $\mu = ds'\beta$. That is, b only depends on the first K elements of Y^*.

> This is what we meant when describing the rotation as putting "all the information" on the first K elements.

Least Squares and the Canonical Form

So from

$$b = \beta + \sigma s d^{-1} V^*_{(1)},$$

we immediately have the following result.

Result #1:

$$b | X = x \sim N(\beta, \sigma^2 s d^{-2} s'),$$

where $x = q_1 d s' \implies x'x = s d^2 s'$ and $(x'x)^{-1} = s d^{-2} s'$.

Sum of Squared Residuals

Using this expression for b, we can write SSR:

$$\begin{aligned} SSR &= \|Y - xb\|^2 \\ &= \|Y^*_{(1)} - ds'b\|^2 + \|Y^*_{(2)}\|^2 = \|Y^*_{(2)}\|^2, \end{aligned}$$

where

$$\begin{aligned} \|Y^*_{(2)}\|^2 &= \sigma^2 \|V^*_{(2)}\|^2 \\ &= \sigma^2 \sum_{i=K+1}^{n} (V^*_i)^2, \end{aligned}$$

where each V^*_i is i.i.d. $N(0,1)$ conditional on $X = x$.

Sum of Squared Residuals

So from,

$$SSR = \sigma^2 \sum_{i=K+1}^{n} (V_i^*)^2,$$

we immediately have the following result.

<u>**Result #2**</u>:

$$SSR|X = x \sim \sigma^2 \cdot \text{Chi}^2(n-K),$$

and

$$SSR \perp\!\!\!\perp b|X = x.$$

Moreover, we immediately have an unbiased estimator of σ^2:

$$\hat{\sigma}^2 = \frac{SSR}{n-K}, \;\; E[\hat{\sigma}^2|X = x] = \sigma^2.$$

using that the expectation of a Chi-squared r.v. is its degrees of freedom.

Outline

The Normal Linear Model

- The Normal Distribution
- Set Up – Normal Linear Model
- Canonical Form of the Normal Linear Model
- Least Squares and the Canonical Form
- Confidence Intervals & Ellipses
- F-tests and the Canonical Form

Dominating Least Squares

- The Bias-Variance Tradeoff
- The Decision Problem
- Stein's Unbiased Estimate of Risk (SURE)
- James & Stein (61)
- Prediction
- Preserving our unbiased estimator

Confidence Intervals

As in Note 4, we construct CI's for a linear combo of coefficients

$$l'\beta = \sum_{j=1}^{K} l_j \beta_j.$$

The standard error is

$$SE = [\hat{\sigma}^2 l'(x'x)^{-1} l]^{1/2},$$

where $\hat{\sigma}^2 = \frac{SSR}{n-K}$
We consider our usual t-statistic

$$t = \frac{l'(b-\beta)}{SE}.$$

Confidence Intervals

Definition: If WS with $W \sim N(0,1)$, $S \sim Chi^2(m)$, then

$$\frac{W}{(S/m)^{1/2}} \sim t(m).$$

Claim 5:

$$\frac{l'(b-\beta)}{SE}|X = x \sim t(n-K).$$

Why? We re-write by dividing by σ^2/σ^2 and substituting in for $\hat{\sigma}^2$. We get

$$\frac{l'(b-\beta)}{[\sigma^2 l'(x'x)^{-1}l]^{1/2}}/[\frac{SSR}{\sigma^2(n-K)}]|X = x \sim \frac{N(0,1)}{[\frac{Chi^2(n-K)}{n-K}]^{1/2}} \sim t(n-K).$$

So, we can construct confidence intervals by just substituting in the critical values for a $t(n-K)$ distribution.

> These will be exact, finite-sample confidence intervals – NOT asymptotic CIs.

Confidence Intervals

Let c denote the 97.5 quantile of a $t(n-K)$ distribution. Then, our 95% confidence interval is

$$P(l'b - c \cdot SE \leq l'\beta \leq l'b + c \cdot SE | X = x) = 0.95.$$

Since the probability does not depend on x, it holds unconditionally as well. So,

$$P(l'b - c \cdot SE \leq l'\beta \leq l'b + c \cdot SE) = 0.95.$$

Confidence Ellipses

Now suppose we wish to obtain confidence regions for multiple linear combinations of coefficients.

L is an $r \times K$ matrix. So $L\beta$ is $r \times 1$. We want to construct a region for $L\beta$.

From the conditional distribution for b, we have that

$$Lb|X = x \sim N(L\beta, \sigma^2 L(x'x)^{-1}L').$$

Define our estimator of the covariance matrix as

$$\hat{Cov}(Lb) = \hat{\sigma}^2 L(x'x)^{-1}L', \quad \hat{\sigma}^2 = \frac{SSR}{n-K}.$$

Confidence Ellipses: Useful Probability Results

Claim 6: If $W \sim N_r(\mu, \Sigma)$, then

$$(W - \mu)'\Sigma^{-1}(W - \mu) \sim Chi^2(r).$$

Definition: If $S_1 \perp\!\!\!\perp S_2$ with $S_1 \sim Chi^2(r)$ and $S_2 \sim Chi^2(m)$, then

$$\frac{S_1/r}{S_2/m} \sim F(r, m).$$

The ratio follows an F-distribution with degrees of freedom r, m.

Confidence Ellipses: Usual F-stat

Claim 7: Conditional on $X = x$,

$$(Lb - L\beta)'[C\hat{o}v(Lb)]^{-1}(Lb - L\beta)/r \sim F(r, n-K).$$

Why? Again divide by σ^2/σ^2. We have

$$\frac{(Lb - L\beta)'[\sigma^2 L(x'x)^{-1}L']^{-1}(Lb - L\beta)/r}{SSR/[\sigma^2(n-K)]}|X = x \sim \frac{Chi^2(r)/r}{Chi^2(n-K)/(n-K)}$$
$$\sim F(r, n-K).$$

Outline

The Normal Linear Model

- The Normal Distribution
- Set Up – Normal Linear Model
- Canonical Form of the Normal Linear Model
- Least Squares and the Canonical Form
- Confidence Intervals & Ellipses
- F-tests and the Canonical Form

Dominating Least Squares

- The Bias-Variance Tradeoff
- The Decision Problem
- Stein's Unbiased Estimate of Risk (SURE)
- James & Stein (61)
- Prediction
- Preserving our unbiased estimator

Canonical Form of the F-Test

We just covered the F-test... Why are we covering it again?

We will formulate the F-test in the canonical form of the normal linear model.

> This will allow us easily formulate the F-statistic as the difference of the R^2 in a "restricted" and "unrestricted" regressions.
>
> This will also serve as a launching point for our discussion of dominating least squares.

The Problem: We have a bunch of regressors and are only interested in testing the joint significance of some subset.

> E.g. We include a set of controls in our regression and are only interested in testing whether the coefficient on the treatment of interest is significant.
>
> E.g. Interested in only testing the coefficient on education in a regression of log wages on education with some additional controls.

Canonical Form of the F-test – Set up

Consider the normal linear model

$$Y = x\beta + \sigma V, \;\; V|x \sim N(0, I_n),$$

where x is $n \times K$. Partition the covariates into two sets

$$Y = x_1\beta_1 + x_2\beta_2 + \sigma V, \;\; V|x \sim N(0, I_n), \tag{5}$$

where x_1 is $n \times K_1$ and x_2 is $n \times K_2$.

Goal: Test $H_0 : \beta_2 = 0$.

We begin by re-writing the normal linear model into a canonical form that partitions μ into subvectors μ_1, μ_2 and then we can simply test $\mu_2 = 0$.

Canonical Form of the F-test – Set up

First: Residualize x_2.

Let $\tilde{x}_2$ be the residual from projecting x_2 onto x_1. So, we have

$$\tilde{x}_2 = x_2 - x_1 t, \; x_1' \tilde{x}_2 = 0.$$

Since $x_2 = \tilde{x}_2 + x_1 t$, we plug this into the normal linear model to get

$$x\beta = x_1\alpha + \tilde{x}_2\beta, \; \alpha = \beta_1 + t\beta_2.$$

Canonical Form of the F-test – Set up

Second: Construct SVDs

Let $rank(x_1) = h \leq K_1$. The SVD of x_1 is

$$x_1 = q_1 d_1 s_1',$$

where q_1 is $N \times h$ with $q_1' q_1 = q_1 q_1' = I_h$, d_1 is $h \times h$ diagonal matrix with positive elements on the diagonal, s_1 is $K_1 \times h$ with $s_1' s_1 = s_1 s_1' = I_h$.

Let $rank(\tilde{x}_2) = r \leq K_2$. The SVD of $\tilde{x}_2$ is

$$\tilde{x}_2 = q_2 d_2 s_2',$$

where q_2 is $N \times r$ with $q_2' q_2 = q_2 q_2' = I_r$, d_2 is $h \times h$ diagonal matrix with positive elements on the diagonal, s_2 is $K_2 \times r$ with $s_2' s_2 = s_2 s_2' = I_r$.

Canonical Form of the F-test – Set up

Since $\tilde{x}_2$ is orthogonal to x_1, we have that

$$0 = \tilde{x}_2' x_1 = s_2 d_2 q_2' q_1 d_1 s_1'.$$

Pre-multiplying by $d_2^{-1} s_2'$ and post-multiplying by $s_1 d_1^{-1}$, we get that

$$0 = q_2' q_1.$$

So q_2 is orthogonal to q_1.

Consider the linear subspace spanned by the columns of $\begin{pmatrix} q_1 & q_2 \end{pmatrix}$. This has dimension $h + r \leq K$ and its orthogonal complement $\begin{pmatrix} q_1 & q_2 \end{pmatrix}^{\perp}$ has dimension $n - h - r$.

> Define an $n \times (n - h - r)$ matrix q_3 whose columns form an orthonormal basis of $\begin{pmatrix} q_1 & q_2 \end{pmatrix}^{\perp}$. So

$$q_3' q_1 = 0, \quad q_3' q_2 = 0, \quad q_3' q_3 = I_{n-h-r}.$$

Canonical Form of the F-test – Set up

Third: Form $q = \begin{pmatrix} q_1 & q_2 & q_3 \end{pmatrix}$ and proceed to the canonical form of the normal linear model.

We have

$$
\begin{aligned}
Y &= x_1\alpha + \tilde{x}_2\beta_2 + \sigma V \\
q'Y &= \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix} \left(q_1 d_1 s_1'\alpha + q_2 d_2 s_2'\beta_2 \right) + \sigma q'V \\
Y^* &= \begin{pmatrix} I_h \\ 0 \\ 0 \end{pmatrix} d_1 s_1'\alpha + \begin{pmatrix} 0 \\ I_r \\ 0 \end{pmatrix} d_2 s_2'\beta_2 + \sigma V^* \\
Y^* &= \begin{pmatrix} \mu_1 \\ \mu_2 \\ 0 \end{pmatrix} + \sigma V^*, \quad V^*|X = x \sim N(0, I_n)
\end{aligned}
$$

where $Y^* = q'Y$, $V^* = q'V$, $\mu_1 = d_1 s_1'\alpha$, $\mu_2 = d_2 s_2'\beta_2$.

Canonical Form of the F-test – Set up

We're done! We have

$$\begin{aligned} Y^*_{(1)} &= \mu_1 + \sigma V_{(1)}, \\ Y^*_{(2)} &= \mu_2 + \sigma V_{(2)}, \\ Y^*_{(3)} &= \sigma V_{(3)}, \end{aligned}$$

where $Y^*_{(1)}$ is the first h elements, $Y^*_{(2)}$ is the next r elements, and $Y^*_{(3}$ is remaining $n - h - r$ elements of Y^*. Moreover, conditional on $X = x$, the elements of $V^*_{(1)}, V^*_{(2)}, V^*_{(3)}$ are all i.i.d. $N(0, 1)$.

Canonical Form of the F-test – The F-Stat

We will now construct the F-stat in terms of SSR of restricted (just using x_1) and unrestricted (using both x_1, x_2) regressions.

Define

$$SSR_r = \min_{c \in \mathbb{R}^h} \|Y - x_1 c\|^2.$$

Note that

$$\begin{aligned} \|Y - x_1 c\|^2 = \|q'(Y - x_1 c)\| = \|Y^* - \begin{pmatrix} I_h \\ 0 \\ 0 \end{pmatrix} d_1 s_1' c\|^2 \\ = \|Y^*_{(1)} - d_1 s_1' c\|^2 + \|Y_{(2)}\|^2 + \|Y_{(3)}\|^2. \end{aligned}$$

So,

$$SSR_r = \|Y_{(2)}\|^2 + \|Y_{(3)}\|^2.$$

Canonical Form of the F-test – The F-Stat

Similarly, define

$$SSR_u = \min_{c_1 \in \mathbb{R}^h, c_2 \in \mathbb{R}^r} \|Y - x_1 c_1 - \tilde{x}_2 c_2\|^2.$$

As before, we can show

$$SSR_u = \|Y_{(3)}\|^2,$$

with least-squares estimates

$$\hat{\alpha} = s_1 d_1^{-1} Y^*_{(1)}, \quad b_2 = s_2 d_2^{-1} Y^*_{(2)}$$

and we can recover b_1 as $b_1 = \hat{\alpha} - t b_2$, where $t = s_1 d_1^{-1} q_1' x_2$ (linear regresion algebra).

Canonical Form of the F-Test – The F-stat

Consider the following test statistic for the null, $\beta_2 = 0$:

$$F = \frac{(SSR_r - SSR_u)/r}{SSR_u/(n-h-r)}.$$

Claim: Under $H_0 : \beta_2 = 0$,

$$F|X = x \sim F_{r,n-h-r}.$$

See notes for simple argument.

Outline

The Normal Linear Model

- The Normal Distribution
- Set Up – Normal Linear Model
- Canonical Form of the Normal Linear Model
- Least Squares and the Canonical Form
- Confidence Intervals & Ellipses
- F-tests and the Canonical Form

Dominating Least Squares

- The Bias-Variance Tradeoff
- The Decision Problem
- Stein's Unbiased Estimate of Risk (SURE)
- James & Stein (61)
- Prediction
- Preserving our unbiased estimator

Dominating Least Squares

Machine learning all the rage today... Techniques provide high quality predictions in high-dimensional settings.

> One key idea: Shrinkage – shrink coefficients towards zero. This produces more biased but lower variance predictions. Provided this tradeoff between bias and variance is "tuned" correctly, this can greatly improves prediction accuracy

Old result: James-Stein (61). We can show that least-squares is dominated by a shrinkage estimator in the normal linear model.

> Provides a closed form expression for the optimal level of shrinkage – where the shrinkage factor will depend on the data.

Outline

The Normal Linear Model

- The Normal Distribution
- Set Up – Normal Linear Model
- Canonical Form of the Normal Linear Model
- Least Squares and the Canonical Form
- Confidence Intervals & Ellipses
- F-tests and the Canonical Form

Dominating Least Squares

- The Bias-Variance Tradeoff
- The Decision Problem
- Stein's Unbiased Estimate of Risk (SURE)
- James & Stein (61)
- Prediction
- Preserving our unbiased estimator

The Value of Shrinkage

Consider the canonical form

$$\begin{aligned} Y^*_{(1)} &= \mu + \sigma V^*_{(1)}, \\ Y^*_{(2)} &= \sigma V^*_{(2)}. \end{aligned}$$

Goal: Construct an accurate estimator $\hat{\mu}$ of μ, where accurate means low MSE

$$E[\sum_{k=1}^{K} (\hat{\mu}_k - \mu_k)^2.$$

Shrinking our estimator towards zero will produce biased estimator but may lead to a large decrease in the variance of the estimator. This could lead to a net large decrease in the MSE.

> E.g. $\hat{\mu}_{LS} = Y^*_{(1)}$, and a shrinkage estimator could be $\hat{\mu}_{shrinkage} = c \cdot \hat{\mu}_{LS}$ for $0 \leq c \leq 1$.

The Bias-Variance Tradeoff

Recall:

$$V(\hat{\mu}) = E[(\hat{\mu} - E[\hat{\mu}])(\hat{\mu} - E[\hat{\mu}])'] = E[\hat{\mu}\hat{\mu}'] - E[\hat{\mu}]E[\hat{\mu}]' = V.$$

Bias is $E[\hat{\mu} - \mu] = b$.

Consider the MSE objective. With some algebra, we can write

$$\begin{aligned} E[(\hat{\mu} - \mu)'(\hat{\mu} - \mu)] &= E[(\hat{\mu} - E[\hat{\mu}] + E[\hat{\mu}] - \mu)'(\hat{\mu} - E[\hat{\mu}] + E[\hat{\mu}] - \mu)] \\ &= E[(\hat{\mu} - E[\hat{\mu})'(\hat{\mu} - E[\hat{\mu})] \\ &+ 2E[(\hat{\mu} - E[\hat{\mu}])'(E[\hat{\mu}] - \mu)] \\ &+ E[(E[\hat{\mu}] - \mu)'(E[\hat{\mu}] - \mu)] \\ &= E[(\hat{\mu} - E[\hat{\mu})'(\hat{\mu} - E[\hat{\mu})] + E[(E[\hat{\mu}] - \mu)'(E[\hat{\mu}] - \mu)] \\ &= E[(\hat{\mu} - E[\hat{\mu})'(\hat{\mu} - E[\hat{\mu})] + b'b. \end{aligned}$$

The Bias-Variance Tradeoff

We'll now rewrite the first-term, $E[(\hat{\mu}-E[\hat{\mu})'(\hat{\mu}-E[\hat{\mu})]$. It is a scalar and so, we can write

$$\begin{aligned} E[(\hat{\mu}-E[\hat{\mu})'(\hat{\mu}-E[\hat{\mu})] &= E[Trace((\hat{\mu}-E[\hat{\mu})'(\hat{\mu}-E[\hat{\mu}))] \\ &= E[Trace((\hat{\mu}-E[\hat{\mu})(\hat{\mu}-E[\hat{\mu})')] \\ &= Trace(E[(\hat{\mu}-E[\hat{\mu})(\hat{\mu}-E[\hat{\mu})']) \\ &= Trace(V), \end{aligned}$$

where we used that $Trace(AB) = Trace(BA)$ and that we can exchange the trace with the expectation because both are linear.

So, we have that our MSE is

$$MSE = E[(\hat{\mu}-\mu)'(\hat{\mu}-\mu)] = Trace(V) + b'b,$$

where the first term depends on the variance of our estimator and the second depends on the bias.

The Bias-Variance Tradeoff

This is a general decomposition for MSE. For scalar case,

$$MSE = E[(\hat{\mu}) - \mu)^2] = V(\hat{\mu}) + Bias(\hat{\mu})^2.$$

Now lets compare the least-squares estimator and the shrinkage estimator.

Least-squares is unbiased and $V(\hat{\mu}_{LS}) = V(Y^*_{(1)}) = \sigma^2 I_K$. So,

$$E[(\hat{\mu}_{LS} - \mu)'(\hat{\mu}_{LS} - \mu)] = \sigma^2 K.$$

For the shrinkage estimator, we have $V(\hat{\mu}_{shrinkage}) = c^2\sigma^2 I_K$ and $E[\hat{\mu}_{shrinkage} - \mu] = (c-1)\mu$. So,

$$E[(\hat{\mu}_{shrinkage} - \mu)'(\hat{\mu}_{shrinkage} - \mu)] = c^2\sigma^2 K + (1-c)^2\mu'\mu.$$

Shrinkage in High-Dimensional Problems

Note that the MSE of the least-squares estimator grows linearly with dimension K.

For high-dimensional (large K) problems, this produces poor predictions. In jargon, the least-squares estimator will be too high variance and will "overfit."

For the right choice of the shrinkage parameter, we may be able to produce an estimator with a lower MSE than least-squares.

If we minimize the MSE of the shrinkage estimator with respect to c, we see that

$$c^* = \frac{\mu'\mu}{\sigma^2 K + \mu'\mu} < 1$$

and so, we could in fact always improve the MSE of our estimator with shrinkage. But this shrinkage rate depends on unknown population parameters. Is there a **feasible** shrinkage rate that dominates least squares?

Outline

The Normal Linear Model

- The Normal Distribution
- Set Up – Normal Linear Model
- Canonical Form of the Normal Linear Model
- Least Squares and the Canonical Form
- Confidence Intervals & Ellipses
- F-tests and the Canonical Form

Dominating Least Squares

- The Bias-Variance Tradeoff
- The Decision Problem
- Stein's Unbiased Estimate of Risk (SURE)
- James & Stein (61)
- Prediction
- Preserving our unbiased estimator

The Decision Problem

Consider the canonical form of the F-test

$$\begin{aligned} Y^*_{(1)} &= \mu_1 + \sigma V^*_{(1)}, \\ Y^*_{(2)} &= \mu_2 + \sigma V^*_{(2)} \\ Y^*_{(3)} &= \sigma V^*_{(3)}. \end{aligned}$$

The parameter space is

$$\Theta = \mathbb{R}^h \times \mathbb{R}^r \times \mathbb{R}_+ = \{(\mu_1, \mu_2, \sigma)\}.$$

The problem: We want to estimate μ_2 (e.g. we want to prediction $Y^*_{(2)}$). We specify a **loss function**:

$$L(\theta, a) = \|a - \mu_2\|^2,$$

where a is an action (estimate of μ_2).

The Decision Rule

A **decision rule** d specifies an action as a function of the data, $\hat{m}u_2 = d(Y^*)$. The **risk function** R gives the expected loss of d under the likelihood, P_θ

$$R(\theta, d) = E_\theta[L(\theta, d(Y^*))].$$

We want to choose decision rules that will minimize our risk/expected loss.

Shrinkage Estimator

The least-squares estimator for μ_2 is simply

$$\hat{\mu}_{2,LS} = Y^*_{(2)}.$$

Consider the shrinkage estimator

$$\hat{\mu}_{2,c} = c \cdot \hat{\mu}_{2,LS} = c \cdot Y^*_{(2)}$$

for $0 \leq c \leq 1$. What is the risk of this? We have that

$$\begin{aligned} R(\theta, \hat{\mu}_{2,c}) &= E_\theta \| c Y^*_{(2)} - \mu_2 \|^2 \\ &= E_\theta \| c(\mu_2 + \sigma V^*_{(2)}) - \mu_2 \| \\ &= (1-c)^2 \|\mu_2\|^2 + c^2 r \sigma^2. \end{aligned}$$

We'd want to choose c to minimize this. However, this is infeasible because θ is unknown. What do we do?

Outline

The Normal Linear Model

- The Normal Distribution
- Set Up – Normal Linear Model
- Canonical Form of the Normal Linear Model
- Least Squares and the Canonical Form
- Confidence Intervals & Ellipses
- F-tests and the Canonical Form

Dominating Least Squares

- The Bias-Variance Tradeoff
- The Decision Problem
- Stein's Unbiased Estimate of Risk (SURE)
- James & Stein (61)
- Prediction
- Preserving our unbiased estimator

Stein's Unbiased Estimate of Risk (SURE)

We construct a statistic whose expectation under P_θ is $R(\theta, \hat{\mu}_{2,c})$. We then minimize this unbiased estimate of the risk with respect to c.

> This statistic is known as **Stein's Unbiased Estimate of Risk** or SURE.

How do we get there? First, consider

$$E_\theta \| Y^*_{(2)} \|^2 = \| \mu_2 + \sigma V^*_{(2)} \|^2 = \| \mu_2 \|^2 + r\sigma^2.$$

We have an unbiased estimator of σ^2, $\hat{\sigma}^2 = \frac{SSR}{n-K}$, where $K = h + r$. So,

$$\| Y^*_{(2)} \|^2 - r\hat{\sigma}^2$$

is an unbiased estimator of $\|\mu_2\|$.

Stein's Unbiased Estimate of Risk (SURE)

Our unbiased estimate of risk is simply a linear combination of the unbiased estimates of $\|\mu_2\|^2$ and σ^2. It is

$$\begin{aligned} Q_c &= (1-c)^2[\|Y^*_{(2)}\|^2 - r\hat{\sigma}^2] + c^2 r\hat{\sigma}^2 \\ &= (1-c)^2[\|Y^*_{(2)}\|^2 - r\frac{SSR}{n-K}] + c^2 r\frac{SSR}{n-K}, \end{aligned}$$

where $K = h + r$.

The **SURE estimator** is the shrinkage estimator that chooses c to minimize SURE. That is,

$$\hat{c}_{SURE} = \arg\min_{0\leq c\leq 1} Q_c.$$

If the minimizer is negative, we set $c = 0$.

SURE Estimator

Taking the first-order condition, we immediately see that

$$\hat{c}_{SURE} = \left(1 - \frac{SSR}{\|Y^*_{(2)}\|}\frac{r}{n-h-r}\right)_+.$$

But, we just had that

$$F = \frac{(SSR_r - SSR_u)/r}{SSR_u/n-h-r} = \frac{\|Y^*_{(2)}\|}{SSR}\frac{n-h-r}{r}.$$

So,

$$\hat{c}_{SURE} = \left(1 - \frac{1}{F}\right)_+$$

$$\hat{\mu}_{2,s} = \left(1 - \frac{1}{F}\right)_+ Y^*_{(2)}.$$

This is the **SURE Estimator**.

Outline

The Normal Linear Model

- The Normal Distribution
- Set Up – Normal Linear Model
- Canonical Form of the Normal Linear Model
- Least Squares and the Canonical Form
- Confidence Intervals & Ellipses
- F-tests and the Canonical Form

Dominating Least Squares

- The Bias-Variance Tradeoff
- The Decision Problem
- Stein's Unbiased Estimate of Risk (SURE)
- James & Stein (61)
- Prediction
- Preserving our unbiased estimator

James & Stein (61)

Theorem: James & Stein (61)

If $r \geq 3$ and if c satisfies

$$0 < c < \frac{2(r-2)}{n-h-r+2},$$

then

$$\hat{\mu}_{2,shrinkage} = (1 - c\frac{\| Y^*_{(3)} \|^2}{\| Y^*_{(2)} \|^2})\hat{\mu}_{2,LS}$$

dominates the least-squares estimator

$$R(\theta, \hat{\mu}_{2,shrinkage}) < R(\theta, \hat{\mu}_{2,LS}) \; \forall \theta \in \Theta.$$

$R(\theta, \hat{\mu}_{2,shrinkage})$ is minimized at $c^* = \frac{r-2}{n-h-r+2}$.

James & Stein (61)

Theorem: James & Stein (61) – continued.

The positive-part estimator

$$\hat{\mu}^{+}_{shrinkage} = (1 - c\frac{\|Y^*_{(3)}\|^2}{\|Y^*_{(2)}\|^2})_+ \hat{\mu}_{2,LS}$$

dominates $\hat{\mu}_{LS}$ under the same conditions and dominates $\hat{\mu}_{shrinkage}$.

Apply James & Stein (61) to our results

Our shrinkage estimator $\hat{\mu}_{2,s} = (1 - \frac{r}{n-h-r} \frac{\|Y^*_{(3)}\|^2}{\|Y^*_{(2)}\|^2})_+ \hat{\mu}_{LS}$ set $c = \frac{r}{n-h-r}$. So, it dominates least squares if

$$\frac{r}{n-h-r} < \frac{2(r-2)}{n-h-r+2}.$$

A sufficient condition for this $r \geq 5, n-h-r > 10$.

Outline

The Normal Linear Model

- The Normal Distribution
- Set Up – Normal Linear Model
- Canonical Form of the Normal Linear Model
- Least Squares and the Canonical Form
- Confidence Intervals & Ellipses
- F-tests and the Canonical Form

Dominating Least Squares

- The Bias-Variance Tradeoff
- The Decision Problem
- Stein's Unbiased Estimate of Risk (SURE)
- James & Stein (61)
- Prediction
- Preserving our unbiased estimator

Prediction and Shrinkage

We will now restate the results in the context of prediction.

> This may be a more intuitive presentation but it is the same fundamental idea. By shrinking our predictions, we are introducing some bias in exchange for a large reduction in variance.

The Problem: We observe $\underset{n\times 1}{Y}$, which follows the normal linear model

$$Y|X = x \sim N(x\beta, \sigma^2 I_n).$$

We wish to predict the value of an independent draw $\underset{n\times 1}{\tilde{Y}}$ from the same distribution

$$\tilde{Y}|X = x \sim N(x\beta, \sigma^2 I_n),$$

where $\theta = (\beta, \sigma) \in \mathbb{R}^K \times \mathbb{R}_+$. Assume $Y \perp\!\!\!\perp \tilde{Y}|X = x$.

Prediction: The Decision Problem

The action a is a point in $\mathbb{R}^n$ – it is a prediction for $\tilde{Y}$.

The loss function is

$$\begin{aligned} L(\theta, a) &= E_\theta[\sum_{j=1}^{n}(\tilde{Y}_j - a_j)^2] \\ &= E_\theta[\|\tilde{Y} - a\|^2] \\ &= E_\theta[\|x\beta + \sigma\tilde{V} - a\|^2] \\ &= \|x\beta - a\|^2 + n\sigma^2. \end{aligned}$$

The decision rule specifies an action as a function of the observation Y. The risk function is

$$R(\theta, d) = E_\theta[L(\theta, d(Y))].$$

Shrinkage and Prediction

As before, we partition the regressors into two groups

$$x\beta = x_1\beta_1 + x_2\beta_2 = x_1\alpha + \tilde{x}_2\beta_2,$$

where $\tilde{x}_2 = x_2 - x_1 t, x_1'\tilde{x}_2 = 0, \alpha = \beta_1 + t\beta_2$. We will consider a prediction function that just shrinks the coefficients on x_2 (i.e. $\tilde{x}_2$).

The least-squares decision rule is

$$d_{LS}(Y) = xb = x_1\hat{\alpha} + \tilde{x}_2 b_2,$$

with $(x_1'x_1)\hat{\alpha} = x_1'Y$, $(\tilde{x}_2'\tilde{x}_2)b_2 = \tilde{x}_2'Y$.

Shrinkage and Prediction

We shrink the least squares estimate b_2 with

$$\hat{\beta}_{2,S} = (1 - \frac{1}{F})_+ b_2$$

and consider the decision rule

$$d_S(Y) = x_1\hat{\alpha} + \tilde{x}_2\hat{\beta}_{2,S}.$$

Note that

$$d_S(Y) = wd_{LS}(Y) + (1-w)x_1\hat{\alpha} \quad w = (1 - \frac{1}{F})_+.$$

This will be useful later.

Comparing the Risk Functions

We will translate back into the canonical form to simplify the risk functions. First, for the shrinkage prediction, we have

$$\begin{aligned} R((\beta,\sigma),d_S) &= E_\theta[\|q'[x_1(\hat{\alpha}-\alpha)+\tilde{x}_2(\hat{\beta}_{2,S}-\beta_2)]\|]+n\sigma^2 \\ &= E_\theta[\|Y^*_{(1)}-\mu_1\|^2]+E_\theta[\|\hat{\mu}_{2,S}-\mu_2\|^2]+n\sigma^2 \end{aligned}$$

Next, for least-squares

$$\begin{aligned} R((\beta,\sigma),d_{LS}) &= E_\theta[\|Y^*_{(1)}-\mu_1\|^2]+E_\theta[\|Y^*_{(2)2,S}-\mu_2\|^2]+n\sigma^2 \\ &= (n+h+r)\sigma^2. \end{aligned}$$

So, $R(\theta,d_S) < R(\theta,d_{LS})$ for all θ if and only if

$$E_\theta[\|\hat{\mu}_{2,S}-\mu_2\|^2] < r\sigma^2.$$

A sufficient condition (as before) is $r \geq 5$, $n-h-r > 10$.

Outline

The Normal Linear Model

- The Normal Distribution
- Set Up – Normal Linear Model
- Canonical Form of the Normal Linear Model
- Least Squares and the Canonical Form
- Confidence Intervals & Ellipses
- F-tests and the Canonical Form

Dominating Least Squares

- The Bias-Variance Tradeoff
- The Decision Problem
- Stein's Unbiased Estimate of Risk (SURE)
- James & Stein (61)
- Prediction
- Preserving our unbiased estimator

Preserving our unbiased estimator

Suppose we have

$$Y = x_1\beta_1 + x_2\beta_2 + \sigma V, \;\; V|X = x \sim N(0, I_n).$$

We are interested in the coefficients on x_1, β_1.

> E.g. x_1 contains our treatment of interest and x_2 contains a large set of additional controls.

We wish to apply our shrinkage estimator to β_2 but then, if we used that to recover an estimate of β_1, it will be biased. That is,

$$\hat{\beta}_{1,S} = \hat{\alpha} - t\hat{\beta}_{2,S}$$

is biased for β_1 in general. How can we preserve an unbiased estimator for β_1 but still maintain the dominance result for the prediction MSE?

Preserving our unbiased estimator – Set Up

The notation here is confusing so pay attention.

Suppose we have some set of covariates p and we want an unbiased estimator for the coefficients on p. That is, we are after an unbiased estimator of β_1 in

$$Y = p\beta_1 + x_2\beta_2 + \sigma V.$$

In other words, we wish to preserve the least-squares estimate b_1, which satisfies

$$(p - Proj(p|x_2))'[Y - (p - Proj(p|x_2)b_1] = 0 \tag{6}$$

by residual regression.

Preserving our unbiased estimator – Set Up

Construct the fitted values from the least-squares projection of p on x_2, where

$$x_2'(p - \hat{p}) = 0, \quad \hat{p}'(p - \hat{p}) = 0.$$

Now define

$$x_1 = \begin{pmatrix} p & \hat{p} \end{pmatrix}$$

and so,

$$x = \begin{pmatrix} x_1 & x_2 \end{pmatrix}.$$

The least squares projection of Y on x has fitted value

$$xb = pb_{11} + \hat{p}b_{12} + x_2 b_2$$

and we want to preserve the coefficient vector b_{11} and thereby obtain an unbiased estimator of $\beta_{1,1}$.

Preserving an unbiased estimator of $b_{1,1}$

Consider the shrinkage estimator. It produces predictions

$$\begin{aligned} d_S(Y) &= x_1\hat{\beta}_{1,S} + x_2\hat{\beta}_{2,S} \\ &= x_1\hat{\alpha} + \tilde{x}_2\hat{\beta}_{2,S}. \end{aligned}$$

Recall that

$$d_S(Y) = w d_{LS}(Y) + (1-w)x_1\hat{\alpha}, \quad w = (1-\frac{1}{F})_+.$$

So, we have that

$$d_S(Y) = w(pb_{11} + \hat{p}b_{12} + x_2 b_2) + (1-w)(p\hat{\alpha}_{11} + \hat{p}\hat{\alpha}_{12})$$

Preserving an unbiased estimator of $b_{1,1}$

The least squares fit of Y on x_1 gives

$$x_1\hat{\alpha} = p\hat{\alpha}_{11} + \hat{p}\hat{\alpha}_{12}.$$

We can use residual regression to obtain $\hat{\alpha}_{11}$.

> Let $\tilde{p}$ be the residual from a least-squares fit of p on $\hat{p}$. Because $\hat{p}'(p - \hat{p}) = 0$, we have that the coefficients of the least-squares fit of p on $\hat{p}$ are all one and so,
>
> $$\tilde{p} = p - \hat{p}.$$
>
> $\hat{\alpha}_{11}$ then solves
>
> $$\begin{aligned} \tilde{p}'[Y - \tilde{p}\hat{\alpha}_{11}] &= 0 \\ (p - \hat{p})'[Y - (p - \hat{p})\hat{\alpha}_{11}] &= 0. \end{aligned}$$

Preserving an unbiased estimator of $b_{1,1}$

We again use residual regression to represent $b_{1,1}$, where $b_{1,1}$ is the coefficient vector on p in the least squares fit of Y on $\begin{pmatrix} p & \hat{p} & x_2 \end{pmatrix}$.

So, we residualize p on $\begin{pmatrix} \hat{p} & x_2 \end{pmatrix}$. Since $x_2'(p - \hat{p}) = 0$ and $\hat{p}'(p - \hat{p}) = 0$, it folows that the fitted value is $\hat{p}$ and the residuals are $p - \hat{p}$.

Then, $b_{1,1}$ satisfies

$$(p - \hat{p})'[Y - (p - \hat{p})b_{11}] = 0.$$

$b_{1,1}$ and $\hat{\alpha}_1$ satisfy the same orthogonality condition and so, we conclude that

$$\hat{\alpha}_{11} = b_{11}.$$

This is also the same orthogonality condition as in Equation 6. So,

$$\hat{\alpha}_{11} = b_{11} = b_1.$$

Preserving an unbiased estimator of $b_{1,1}$

So, we then have that

$$\begin{aligned} d_S(Y) &= p\hat{\beta}_{11,s} + \hat{}\hat{\beta}_{12,s} + x_2\hat{\beta}_{2,s} \\ &= w(pb_{11} + \hat{p}b_{12} + x_2b_2) + (1-w)(p\hat{\alpha}_{11} + \hat{p}\hat{\alpha}_{12}) \\ &= pb_{11} + w(\hat{p}b_{12} + x_2b_2) + (1-w)\hat{p}\hat{\alpha}_{12}) \end{aligned}$$

and conclude immediately that

$$\hat{\beta}_{11,s} = b_{11}, \quad E[\hat{\beta}_{11,s}] = E[b_{11}] = \beta_{11}.$$

Applications

The lecture note illustrates how to apply this result to our fixed effects model for panel data.

Next time: Work through Chamberlain (2016) – applies these results to estimation of neighborhood effects.