Multiple Regression

2.1

Ch 2 in Sen & Srivastava

Scalar form

$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$ for $i = 1, \ldots, n$, where

β_j are unknown constants

x_{ij} are observable known constants

$\varepsilon_1 \ldots \varepsilon_n$ are unobservable random variables with

$$E(\varepsilon_i) = 0, \quad Var(\varepsilon_i) = \underset{\text{unknown}}{\sigma^2}, \quad Cov(\varepsilon_i, \varepsilon_j) = 0 \text{ for } i \neq j$$

y_i are observable random variables

Matrix form

WE WILL ALWAYS ASSUME $n > k+1$

$y = X\beta + \varepsilon$, where

β is a $(k+1) \times 1$ vector of unknown constants

X is an $n \times (k+1)$ matrix of observable, known constants

ε is an $n \times 1$ unobservable random vector with

$$E(\varepsilon) = 0, \quad Cov(\varepsilon) = \sigma^2 I, \quad \sigma^2_{1\times 1} \text{ unknown}$$

y is an $n \times 1$ observable random vector

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{y} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & & & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\varepsilon}$$

Put up model

Show $X'X$, $X'y$ Save $X'X$

write first

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$$E(y_i) = E(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) + E(\varepsilon_i)$$

$$= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

Differentiate ~~Least squares estimates~~. Arrive at $(X'X)\beta = X'y$

Least squares estimates b must satisfy

$$(X'X)\,b = X'y \qquad (2.8)$$

If $(X'X)^{-1}$ exists, $b = (X'X)^{-1}X'y \qquad (2.9)$

When does $(X'X)^{-1}$ exist?

Theorem The following are equivalent

(a) The columns of X are linearly independent

(b) $X'X$ is positive definite

(c) $(X'X)^{-1}$ exists

NO The following 3 conditions are equivalent

Theorem $(X'X)^{-1}$ exists iff the cols of X are l.i.

Proof will go

(a) cols l.i. $\Rightarrow$ (b) $X'X$ pd $\Rightarrow$ (c) $(X'X)^{-1}$ exists $\Rightarrow$ (a)

(a) COLS of X L.I.

Means $Xa=0$ implies $a=0$

Seek to show $X'X$ pd. First, it is non-neg def

Because $v'(X'X)v = \underbrace{(Xv)'}_{1\times n}\underbrace{Xv}_{n\times 1} = z'z = \sum_{i=1}^{n} z_i^2 \geq 0$

~~$n\times(k+1)$~~ $(k+1)\times 1$, 1×1

~~X'X pd means...~~ If the only v with $v'X'Xv=0$ is $v=0$, then $X'X$ is pd

$\underbrace{v'X'}_{z'}\underbrace{Xv}_{z} = z'z = 0 \Rightarrow z=0$, i.e. $Xv=0$

~~$Xv=0$~~ $\Rightarrow v = 0$ bec of l.i.

So ~~$X'X$ is pd~~

(b) $X'X$ is pd

~~If X'X is pd then eigenvalues are strictly pos~~

$X'X$ is symmetric, for $(X'X)' = X'X'' = X'X$

So we have spectral decomp $X'X = CDC'$

Bec $X'X$ is pd, all eigenvalues are pos and

$(X'X)^{-1} = CD^{-1}C'$, that is

(c) $(X'X)^{-1}$ exists

Seek to show cols of X l.i. $Xa = 0$, show $a = 0$

$$Xa = \underset{n\times 1}{0} \Rightarrow X'Xa = X'0 = \underset{(k+1)\times 1}{0}$$

$$\Rightarrow (X'X)^{-1}(X'X)a = (X'X)^{-1}0 \Rightarrow a = 0,$$

$\Rightarrow$ (a) Cols of X l.i. $\blacksquare$

Recent HW: $(X'X)^{-1}$ p.d.

More notation

$$\underset{n\times 1}{\hat{y}} = Xb$$

predicted y

$\hat{y}_i = b_0 + b_1 x_{i1} + \dots + b_k x_{ik}$

point on the least-squares plane

$$e = y - \hat{y}$$

residual

vert distr $e_i = y_i - \hat{y}_i$

Very useful result

Thm 2.1 from the text (the better of 2 parts)

$$\underset{(k+1)\times n}{X'}\ \underset{n\times 1}{e} = \underset{(k+1)\times 1}{0}$$

Geometrically interesting, later

Proof $X'e = X'(y - \hat{y}) = X'y - X'\hat{y}$

$= X'y - X'Xb = X'y - \underbrace{X'X(X'X)^{-1}}X'y$

$= X'y - X'y = 0$ ∎

And of course $e'X = 0'$ $1\times(k+1)$

Found $b = (X'X)^{-1}X'y$ by minimizing

$S = \Sigma(\qquad) = (y - X\beta)'(y - X\beta)$

Does it really yield a minimum?

Eigenvalues of Hessian $\left[\frac{\partial^2 S}{\partial\beta_i\,\partial\beta_j}\right]$ all pos?

No thanks!

Basic trick (yielding a formula ~~good~~ useful for distribution theory later)

$$S = (y - X\beta)'(y - X\beta)$$

$$= (y - \underbrace{Xb}_{\hat{y}} + Xb - X\beta)'(y - \underbrace{Xb}_{\hat{y}} + Xb - X\beta)$$

This would be cheating if we did not know about b

$$= (e + X(b-\beta))'(e + X(b-\beta))$$

$$= (\underbrace{e'}_{1\times n} + \underbrace{(b-\beta)'}_{1\times n}X')(e + X(b-\beta))$$

$$= e'e + \cancel{e'X(b-\beta) + (b-\beta)'X'e} \quad 1\times 1$$
$$+ (b-\beta)'X'X(b-\beta)$$

$$= e'e + (b-\beta)'X'X(b-\beta)$$

$\parallel$

$\Sigma(y_i - \hat{y}_i)^2$

≥ 0

and free of β

pos def

The second term is strictly positive unless $\beta = b$

<u>Unique MINIMUM</u>

Don't write this

So $b = (X'X)^{-1}X'y$ <u>minimizes</u>

$$S = (y - X\beta)'(y - X\beta) \quad \text{at } \beta = b$$

$$\hat{y} = Xb = X(X'X)^{-1}X'y = Hy$$

The "Hat Matrix" H puts a hat on y

- Symmetric
- Idempotent

Show

Also note

$$e = y - \hat{y} = y - Hy = Iy - Hy$$
$$= (I - H)y = My \quad \text{(book's notation)}$$

$= M\varepsilon$ a good exercise

- Symmetric
- Idempotent

Projections (not in text)

$$V = \{ v \in \mathbb{R}^n : v = Xa,\ a \in \mathbb{R}^{k+1} \}$$

The space <u>spanned</u> by the cols of X

All linear combos of the cols of X

Some important vectors are in V

- $E(y) = X\beta$, β is an a ($E(X\beta + \varepsilon)$)
- $\hat{y} = Xb$ b is an a
- How about y? Is $y \in V$?

VERY UNLIKELY   write it b's

Suppose $y = Xa$, some a

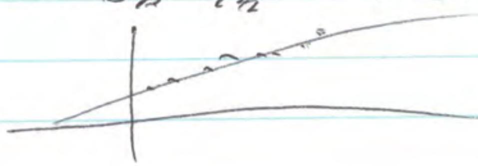
$$\Rightarrow X'Xa = X'y \Rightarrow (X'X)^{-1}X'Xa = (X'X)^{-1}X'y = b$$

$\Rightarrow a = b$, so that

$$y = Xb, \text{ exactly}$$

In scalar form,

$$y_i = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik} \quad \text{exactly,}$$



All pts exactly on the LS line

DISCUSS

So assume $y \notin \mathcal{V}$, because otherwise it's a miracle

What's the point p in $\mathcal{V}$ that is CLOSEST to y? Euclidean distance is

$$\sqrt{(y_1-p_1)^2+(y_2-p_2)^2+\cdots+(y_n-p_n)^2}$$

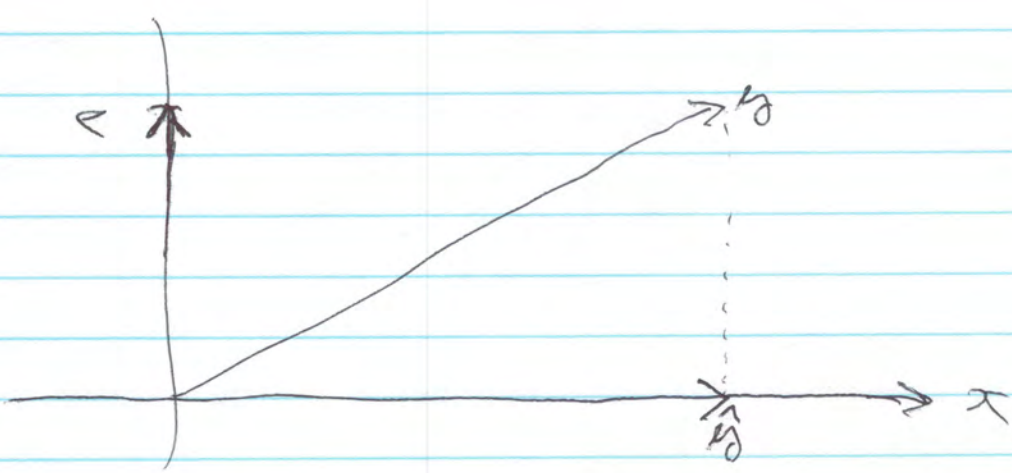
$p = Xa$, some $a \in \mathbb{R}^{k+1}$ to find it,

Minimize $(y-p)'(y-p) = (y-Xa)'(y-Xa)$

<u>Hey</u>, we've already done this, calling a by the name β

Answer is $p = Xb = \hat{y}$

So $\hat{y}$ is the point in $\mathcal{V}$ that is closest to y



$e = y - \hat{y}$

$\hat{y} + e = y$

$\hat{y}$ is the projection (shadow) of y onto $\mathcal{V}$

$\hat{y} = Hy$, H is the projection operator

For any $y \in \mathbb{R}^n$, Hy is the pt in $\mathcal{V}$ closest to y.

8.5

If so, should have

- If $p \in \mathcal{V}$, $Hp = p$
  - Includes $\hat{y} = Xb$
  - " $E(y) = X\beta$
  - Includes any ~~vector~~ col of X
- $\hat{y} \perp e$
- $p \perp e \quad \forall p \in \mathcal{V}$
- One meaning of $X'e = 0$

What are we estimating?

(More detail later)

Human Resources example

X_1 = University GPA ~~@ subject~~

X_2 = Interview score

X_3 = Test score

Y_i = % salary increase after 1 yr

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

$\beta_1\ \beta_2\ \beta_3$ are links between IV & DV

β_0 is for curve fitting — no good interpretation

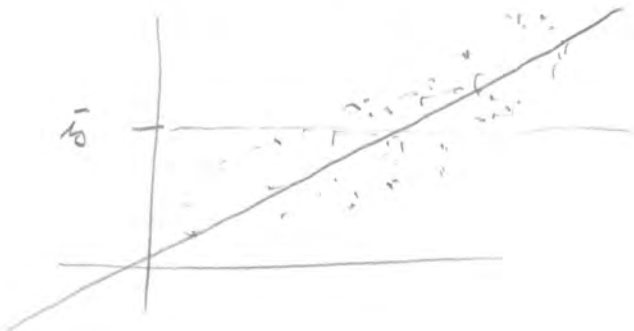
Question

Holding Interview & Test Scores constant, how much does GPA matter?

$$E(y_i) = (\beta_0 + \beta_2 x_2 + \beta_3 x_3) + \beta_1 x_1$$

A measure of good model fit

For simple regression $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \le \sum_{i=1}^{n}(y_i-\bar{y})^2$



so $\dfrac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2} \in (0,1)$ is a measure of bad fit

and $R^2 = 1 - \dfrac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}$ is a measure of good fit

Why squared presently

To generalize to multiple regression, need

Decomposition of Sums of squares

Suppose the model has an intercept (not all do)

Thm (Cor 2.2 in the text)

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

$$\underset{\text{Total variation to Explain}}{SSTO} = \underset{\text{unexplained}}{SSE} + \underset{\text{Explained}}{SSR}$$

Proof Because the model has an intercept,

$X'e = 0$ implies $\sum_{i=1}^{n} e_i = 0$

$$SSTO = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$= \sum_{i=1}^{n}(e_i + \hat{y}_i - \bar{y})^2$$

$$= \sum_{i=1}^{n} e_i^2 + 2\sum_{i=1}^{n} e_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$= SSE + SSR + 2\Big(\sum_{i=1}^{n} e_i \hat{y}_i - \bar{y}\underbrace{\sum_{i=1}^{n} e_i}_{0}\Big)$$

$$= SSE + SSR + 2\hat{y}'e$$

$$= SSE + SSR + 2(Xb)'e$$

$$= SSE + SSR + 2b'\underbrace{X'e}_{0} = SSE + SSR \quad \blacksquare$$

$$R^2 = \frac{SSR}{SSTO} = \frac{\text{Explained}}{\text{Total}} = \frac{SSTO - SSE}{SSTO}$$

$$= 1 - \frac{SSE}{SSTO} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Same expression given for simple regression

<u>Thm</u> For simple regression, $R^2 = r^2$

We'll need formulas

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2 \Sigma(y-\bar{y})^2}} \qquad b_0 = \bar{y} - b_1\bar{x}, \quad b_1 = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

$$R^2 = \frac{SSR}{SSTO}$$

<u>Proof</u>

For simple regression

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}(b_0 + b_1 x_i - \bar{y})^2$$

$$= \sum_{i=1}^{n}(\cancel{\bar{y}} - b_1\bar{x} + b_1 x_i - \cancel{\bar{y}})^2$$

$$= \sum_{i=1}^{n}\left(b_1(x_i - \bar{x})\right)^2 = b_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \left(\frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x-\bar{x})^2}\right)^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{\left(\Sigma(x-\bar{x})(y-\bar{y})\right)^2}{\sum_{i=1}^{n}(x-\bar{x})^2}$$

And

$$R^2 = \frac{SSR}{SST} = \frac{\left(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right)^2}{\Sigma(x-\bar{x})^2\ \Sigma(y-\bar{y})^2}$$

$$= \left(\qquad \qquad \right)^2 = r^2$$

10 13

Expected value & variance - covariance

$y = X\beta + \varepsilon$, ~~X & β Constants~~,

- X & β constants
- $E(\varepsilon) = 0$
- $\operatorname{cov}(\varepsilon) = \sigma^2 I_n$

unknown constants

First $E(y)$
$\operatorname{cov}(y)$

* β are important Estimate with

$\underset{(k+1)\times 1}{b} = (X'X)^{-1}X'y$ A random vector

$$E(b) = E\left((X'X)^{-1}X'y\right) = (X'X)^{-1}X'E(y)$$
$$= (X'X)^{-1}X'X\beta = \beta$$

Say b is an UNBIASED ESTIMATOR of β, meaning $E(b) = \beta \quad \forall \beta \in \mathbb{R}^{k+1}$

$\operatorname{cov}(b)$ Use $\operatorname{cov}(Aw) = A\Sigma A'$

$$\operatorname{cov}\left(\underbrace{(X'X)^{-1}X'}_{A}y\right) = (X'X)^{-1}X'\operatorname{cov}(y)\left((X'X)^{-1}X'\right)'$$
$$= (X'X)^{-1}X'\sigma^2 I X(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}$$

4 /14

$\hat{y} = Xb = Hy, \quad H = X(X'X)^{-1}X'$

- $E(\hat{y}) = X\beta$
- $\mathrm{cov}(\hat{y}) =$

$e = y - \hat{y} = (I - H)y$

- $E(e) = 0$
- $\mathrm{cov}(\hat{y}) =$

An unbiased estimator of σ^2

$y = X\beta + \varepsilon$, $Cov(\varepsilon) = E(\varepsilon\varepsilon') = \sigma^2 I_n$

Base somehow on $SSE = e'e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Recall $e = (I-H)y \overset{\text{surprising fact}}{=} (I-H)\varepsilon$

$$= (I-H)(X\beta+\varepsilon) = X\beta - \underbrace{HX\beta}_{X\beta \text{ projection}} + (I-H)\varepsilon$$

And another surprising fact about H

$$tr(H) = tr(X(X'X)^{-1}X') = tr(X'X(X'X)^{-1})$$
$$= k+1$$

Now $E(e'e) \overset{\text{TRICK}}{=} E(tr(e'e))$

$$= E(tr(ee')) = tr\,E(ee')$$

$$= tr\,E\{(I-H)\varepsilon((I-H)\varepsilon)'\}$$

$$= tr\,E\{(I-H)\varepsilon\varepsilon'(I-H)'\}$$

$$= tr\,(I-H)E\{\varepsilon\varepsilon'\}(I-H)$$

$$= tr((I-H)\sigma^2 I(I-H)) = \sigma^2 tr(I-H)$$

Estimating Linear Combinations of β values

16

y = 1st year university GPA

x_1 = HS GPA

x_2 = Grade 12 English

x_3 = # Math courses

x_4 = # Extra-Cirricular activities

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Estimate $a'\beta$, for example

$$\beta_4 = (0\ 0\ 0\ 0\ 1)\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix}$$

For fixed ACADEMIC background, what's the connection of extra-cirriculars to success in university

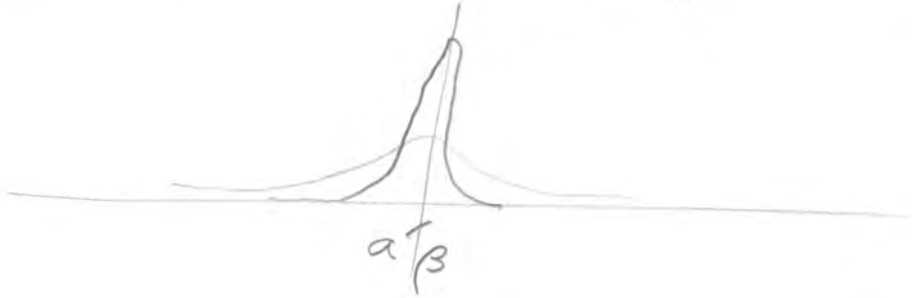
Or, what's the expected 1st year GPA for a student with HSGPA of 80, 70 in HS Engl, took 5 math courses, 2 extra-c ?

$$(1\ \ 80\ \ 70\ \ 5\ \ 2)\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix}$$

Natural estimate of $a'\beta$ is $a'b$

It's unbiased $E(a'b) = a'E(b) = a'\beta$

Small variance in an unbiased estimator is good. It's the variance of the SAMPLING DISTRIBUTION



Note that the natural estimator $a'b$ is a linear combination of the y values:

$$a'b = \underbrace{a'(X'X)^{-1}X'}_{c_0' \;\leftarrow\; 1\times n} y = c_0' y$$

Save

Let $L = c'y = c_1 y_1 + c_2 y_2 + \cdots + c_n y_n$, another linear combination of the ys with

$E(L) = a'\beta$ for every $\beta \in \mathbb{R}^{k+1}$

That is, its an unbiased estimator of $a'\beta$

If we can find L, unbiased, with

$Var(L) < Var(a'b)$, we should use that instead

A serious L. Let

$$\hat{\beta}_w = \underbrace{(X'WX)^{-1}X'W}_{C'} y \text{ , where } W \text{ is an } n \times n$$

matrix of rank at least $k+1$ so $(X'WX)^{-1}$ exists.

$$E(\hat{\beta}_w) = (X'WX)^{-1}X'W\,E(y)$$
$$= \underbrace{(X'WX)^{-1}X'WX}_{I}\beta = \beta, \text{ so}$$

let $L = a'\hat{\beta}_w$, $E(L) = a'\beta$

The matrix W could be specially chosen to suit a. Should we try to find a W with $Var(a'\hat{\beta}_w) < Var(a'b)$?

The Gauss-Markov Theorem says don't bother

Thm Let $L = c'y$ be an unbiased estimator of $a'\beta$. $Var(a'b) \le Var(L)$, with equality only if $c = c_0$

$$(a'b = \underbrace{a'(X'X)^{-1}X'}_{c_0'} y$$

Proof of GM

$L = c'y$ unbiased means $E(c'y) = a'\beta$ for all $\beta \in \mathbb{R}^{k+1}$, even special crazy βs.

$$\underbrace{a'}_{1\times(k+1)}\beta = E(c'y) = c'E(y) = \underbrace{c'X}_{1\times(k+1)}\beta = r'\beta$$

β could have a 1 in the first position, so $a_1 = r_1$
β " " " " " " second " " $a_2 = r_2$

$\vdots$

$a_{k+1} = r_{k+1}$

Thus $\boxed{a' = c'X \iff a = X'c}$ save

Now, the variance of a scalar RV is the variance-covariance matrix of a 1×1 random vector, so

$$Var(L) - Var(a'b) = Var(c'y) - Var(a'b)$$

$$= cov(c'y) - cov(a'b) = c'cov(y)c - a'cov(b)a$$

$$= c'\sigma^2 I_n c - a'\sigma^2 (X'X)^{-1} a$$

$$= \sigma^2 (c'Ic - c'X(X'X)^{-1}X'c)$$

$$= \sigma^2 (c'Ic - c'Hc)$$

$$= \sigma^2 c'(I-H)c$$

$$= \sigma^2 c'(I-H)'(I-H)c$$

$$= \sigma^2 \left((I-H)c\right)' \underset{n\times n}{(I-H)} \underset{n\times 1}{c} = \sigma^2 z'z \geq 0$$

Wow, so variance of L can't be less

If $z'z = 0 \implies z = (I-H)c = 0$

$$\implies c = Hc = X(X'X)^{-1}\underbrace{X'c}_{a} = c_0$$

when $a'b = \underbrace{a'(X'X)^{-1}X'}_{c_0'}y = c_0'y$

So the variance of L can be as small as $var(a'b)$ only if $L = a'b$

$\blacksquare$

Another nice way

$$L = c'y = (Hc)'y = c'Hy$$

$$= c'Hy = c'X\underbrace{(X'X)^{-1}X'y}_{b} = c'Xb = a'b$$

Recap and one more amazing thing

21

Want to estimate $a'\beta$. Use

$$a'b = \underbrace{a'(X'X)^{-1}X'}_{c_0'}y = c_0'y,$$

$$c_0 = X(X'X)^{-1}a$$

Letting $L = c'y$ with $E(L) = a'\beta \ \forall \beta$,

$E(L) = a'\beta = E(c'y) = c'X\beta$, so that

$$\boxed{a' = c'X \iff a = X'c}$$

Proved $Var(L) > Var(a'b)$ with equality only if

$L = a'b$ & $c = c_0$

Now look at the PROJECTION (onto V)

$$Hc = X(X'X)^{-1}\underbrace{X'c}_{a} = X(X'X)^{-1}a = c_0$$

The projection of <u>every</u> vector of coefficients c giving an unbiased estimate is c_0.

There are ∞ many, & it's the closest point in V to all of them.