Lecture Notes for Theory of Linear Models

(chapter 7)

- Estimation in Linear Models
- BLUE
- Distribution of SSR, SSE, $\hat{\beta}$
- R^2 and R_a^2
- Overfitting and Underfitting

Longhai Li
Department of Mathematics and Statistics
University of Saskatchewan

Assumptions in Linear Models

Suppose that on a random sample of n units (patients, animals, trees, etc.) we observe a response variable Y and explanatory variables $X_1, \ldots, X_k$.

Our data are then $(y_i, x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$, or, in vector/matrix form $\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_k$ where $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$; or $\mathbf{y}, \mathbf{X}$ where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$.

Either by design or by conditioning on their observed values, $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are regarded as vectors of known constants.

The linear model in its classical form makes the following assumptions:

A1. (additive error) $\mathbf{y} = \boldsymbol{\mu} + \mathbf{e}$ where $\mathbf{e} = (e_1, \ldots, e_n)^T$ is an unobserved random vector with $\mathrm{E}(\mathbf{e}) = \mathbf{0}$. This implies that $\boldsymbol{\mu} = \mathrm{E}(\mathbf{y})$ is the unknown mean of $\mathbf{y}$.

A2. (linearity) $\boldsymbol{\mu} = \beta_1\mathbf{x}_1 + \cdots + \beta_k\mathbf{x}_k = \mathbf{X}\boldsymbol{\beta}$ where $\beta_1, \ldots, \beta_k$ are unknown parameters. This assumption says that $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu} \in \mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_k) = C(\mathbf{X})$ lies in the column space of $\mathbf{X}$; i.e., it is a linear combination of explanatory vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ with coefficients the unknown parameters in $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$.

- Linear in $\beta_1, \ldots, \beta_k$ not in the x's.

A3. (independence) $e_1, \ldots, e_n$ are independent random variables (and therefore so are $y_1, \ldots, y_n$).

A4. (homoscedasticity) $e_1, \ldots, e_n$ all have the same variance σ^2; that is, $\mathrm{var}(e_1) = \cdots = \mathrm{var}(e_n) = \sigma^2$ which implies $\mathrm{var}(y_1) = \cdots = \mathrm{var}(y_n) = \sigma^2$.

A5. (normality) $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}, \quad i = 1, \ldots, n,$$

or

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}}_{=\mathbf{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{=\boldsymbol{\beta}} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

or

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$.

or $y = \beta_0 \hat{j}_n + \beta_1 x_1 + \cdots + \beta_k x_k + e$

Taken together, all five assumptions can be stated more succinctly as $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

or $y \sim N_n(\mu_y, \sigma^2 I_n)$

$\mu_y = X\beta \in C(X)$

$y \sim N_n(X\beta, \sigma^2 I_n)$

A Note:

In addition, this effect depends upon what other explanatory variables are present in the model. For example, β_0 and β_1 in the model

$$\mathbf{y} = \beta_0 \mathbf{j}_n + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \mathbf{e}$$

will typically be different than β_0^* and β_1^* in the model

$$\mathbf{y} = \beta_0^* \mathbf{j}_n + \beta_1^* \mathbf{x}_1 + \mathbf{e}^*.$$

We will first consider the case that rank(X) = k+1

Example 7.3.1a. We use the data in Table 7.1 to illustrate computation of $\hat{\boldsymbol{\beta}}$ using (7.6).

$$\mathbf{y} = \begin{pmatrix} 2 \\ 3 \\ 2 \\ 7 \\ 6 \\ 8 \\ 10 \\ 7 \\ 8 \\ 12 \\ 11 \\ 14 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 0 & 2 \\ 1 & 2 & 6 \\ 1 & 2 & 7 \\ 1 & 2 & 5 \\ 1 & 4 & 9 \\ 1 & 4 & 8 \\ 1 & 4 & 7 \\ 1 & 6 & 10 \\ 1 & 6 & 11 \\ 1 & 6 & 9 \\ 1 & 8 & 15 \\ 1 & 8 & 13 \end{pmatrix}, \quad \mathbf{X}'\mathbf{X} = \begin{pmatrix} 12 & 52 & 102 \\ 52 & 395 & 536 \\ 102 & 536 & 1004 \end{pmatrix},$$

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 90 \\ 482 \\ 872 \end{pmatrix}, \quad (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} .97476 & .24290 & -.22871 \\ .24290 & .16207 & -.11120 \\ -.22871 & -.11120 & .08360 \end{pmatrix},$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} 5.3754 \\ 3.0118 \\ -1.2855 \end{pmatrix}.$$

Theorem 7.3b. If $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, then $\hat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}$.

$E(Ay) = A \cdot E(y)$

PROOF

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}) \quad \text{[by (3.38)]} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned} \tag{7.13}$$

□

Theorem 7.3c. If $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, the covariance matrix for $\hat{\boldsymbol{\beta}}$ is given by $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Cov = Var

$Var(Ay) = A \cdot Var(y) \cdot A'$

PROOF

$$\begin{aligned} \text{cov}(\hat{\boldsymbol{\beta}}) &= \text{cov}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{cov}(\mathbf{y})[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \quad \text{[by (3.44)]} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \tag{7.14}$$

$V(\hat{\beta}) = \sigma^2 \cdot (X'X)^{-1}$

Note: no assumption of normality.

$\hat{A} = (X'X)^{-1} X'$

Best Linear Unbiased Estimator

BLUE

Theorem 7.3d (Gauss–Markov Theorem). If $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, the least-squares estimators $\hat{\beta}_j, j = 0, 1, \ldots, k$, have minimum variance among all linear unbiased estimators.

$V(\hat{\beta}_j) \le V(\tilde{\beta}_j)$

PROOF. We consider a linear estimator $\mathbf{Ay}$ of $\boldsymbol{\beta}$ and seek the matrix $\mathbf{A}$ for which $\mathbf{Ay}$ is a minimum variance unbiased estimator of $\boldsymbol{\beta}$. In order for $\mathbf{Ay}$ to be an unbiased estimator of $\boldsymbol{\beta}$, we must have $E(\mathbf{Ay}) = \boldsymbol{\beta}$. Using the assumption $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, this can be expressed as

$\tilde{\beta} = Ay$

$$E(\mathbf{Ay}) = \mathbf{A}E(\mathbf{y}) = \mathbf{AX}\boldsymbol{\beta} = \boldsymbol{\beta},$$

$\text{cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}$

which gives the unbiasedness condition

$$\mathbf{AX} = \mathbf{I}_{k+1}$$

since the relationship $\mathbf{AX}\boldsymbol{\beta} = \boldsymbol{\beta}$ must hold for any possible value of $\boldsymbol{\beta}$ [see (2.44)].

The covariance matrix for the estimator $\mathbf{Ay}$ is given by

$\text{cov}(\tilde{\beta}) = \text{cov}(\mathbf{Ay}) = \mathbf{A}(\sigma^2\mathbf{I})\mathbf{A}' = \sigma^2\mathbf{AA}'.$ ($\mathbf{I}_n$)

$\hat{\beta} = [(X'X)^{-1}X']y = \hat{A}y$

The variances of the $\hat{\beta}_j$'s are on the diagonal of $\sigma^2\mathbf{AA}'$, and we therefore need to choose $\mathbf{A}$ (subject to $\mathbf{AX} = \mathbf{I}$) so that the diagonal elements of $\mathbf{AA}'$ are minimized. To relate $\mathbf{Ay}$ to $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, we add and subtract $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ to obtain

$AA' = (A - \hat{A} + \hat{A})(A - \hat{A} + \hat{A})'$

$$\mathbf{AA}' = [\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']'.$$

Expanding this in terms of $\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, we obtain four terms, two of which vanish because of the restriction $\mathbf{AX} = \mathbf{I}$. The result is

$$\mathbf{AA}' = [\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' + (\mathbf{X}'\mathbf{X})^{-1}. \tag{7.17}$$

The matrix $[\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']'$ on the right side of (7.17) is positive semidefinite (see Theorem 2.6d), and, by Theorem 2.6a (ii), the diagonal elements are greater than or equal to zero. These diagonal elements can be made equal to zero by choosing $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. (This value of $\mathbf{A}$ also satisfies the unbiasedness condition $\mathbf{AX} = \mathbf{I}$.) The resulting minimum variance estimator of $\boldsymbol{\beta}$ is

$$\mathbf{Ay} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

which is equal to the least–squares estimator $\hat{\boldsymbol{\beta}}$. □

proof of $(A - \hat{A})\hat{A}' = 0$

$\hat{A} = (X'X)^{-1} X'$, $\hat{\beta}_{LS} = \hat{A} y$

$$(A - \hat{A})\,\hat{A}'$$

$$= A \cdot \hat{A}' - \hat{A} \cdot \hat{A}'$$

$$= A \cdot \hat{A}' - (X'X)^{-1}$$

$$= A \cdot (X \cdot (X'X)^{-1}) - (X'X)^{-1}$$

$$= \underbrace{I_{k+1}}\,(X'X)^{-1} - (X'X)^{-1} = 0$$

$AX = I$

Let $\hat{A} = (X'X)^{-1}X'$, i.e. $\hat{\beta}_{LS} = \hat{A} \cdot y$

$$A A' = (A - \hat{A}) \cdot (A - \hat{A})' + (X'X)^{-1} + 0$$

$\forall \alpha \in \mathbb{R}^{k+1}$

$$\frac{1}{\sigma^2} V(\alpha' \tilde{\beta}) = V(\alpha' A y)/\sigma^2 = \alpha' A A' \alpha$$

$$= \alpha'(A - \hat{A})(A - \hat{A})'\alpha \geq 0 + \alpha'(X'X)^{-1}\alpha$$

$$\geq \alpha'(X'X)^{-1}\alpha = \frac{1}{\sigma^2} \mathrm{Var}(\alpha' \hat{\beta})$$

Let $\alpha_j = (0, \cdots, 0, 1, 0, 0, \cdots 0)$ (1 in the j^{th} position)

$$V(\tilde{\beta}_j) = \alpha_j' A A' \alpha_j \cdot \sigma^2$$

$$\geq \alpha_j' (X'X)^{-1} \alpha_j \cdot \sigma^2$$

$$= V(\hat{\beta}_j)$$

Notes on Gauss-Markov Thm:

1) The remarkable feature of the Gauss–Markov theorem is its distributional generality. The result holds for any distribution of $\mathbf{y}$; normality is not required. The only assumptions used in the proof are $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$. If these assumptions do not hold, $\hat{\boldsymbol{\beta}}$ may be biased or each $\hat{\beta}_j$ may have a larger variance than that of some other estimator.

2) **Corollary 1.** If $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, the best linear unbiased estimator of $\mathbf{a}'\boldsymbol{\beta}$ is $\mathbf{a}'\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the least–squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

3) **Theorem 7.3e.** If $\mathbf{x} = (1, x_1, \ldots, x_k)'$ and $\mathbf{z} = (1, c_1x_1, \ldots, c_kx_k)'$, then $\hat{y} = \hat{\boldsymbol{\beta}}'\mathbf{x} = \hat{\boldsymbol{\beta}}_z'\mathbf{z}$, where $\hat{\boldsymbol{\beta}}_z$ is the least squares estimator from the regression of y on $\mathbf{z}$.

PROOF. From (2.29), we can rewrite $\mathbf{z}$ as $\mathbf{z} = \mathbf{D}\mathbf{x}$, where $\mathbf{D} = \text{diag}(1, c_1, c_2, \ldots, c_k)$. Then, the $\mathbf{X}$ matrix is transformed to $\mathbf{Z} = \mathbf{X}\mathbf{D}$ [see (2.28)]. We substitute $\mathbf{Z} = \mathbf{X}\mathbf{D}$ in the least-squares estimator $\hat{\boldsymbol{\beta}}_z = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$ to obtain

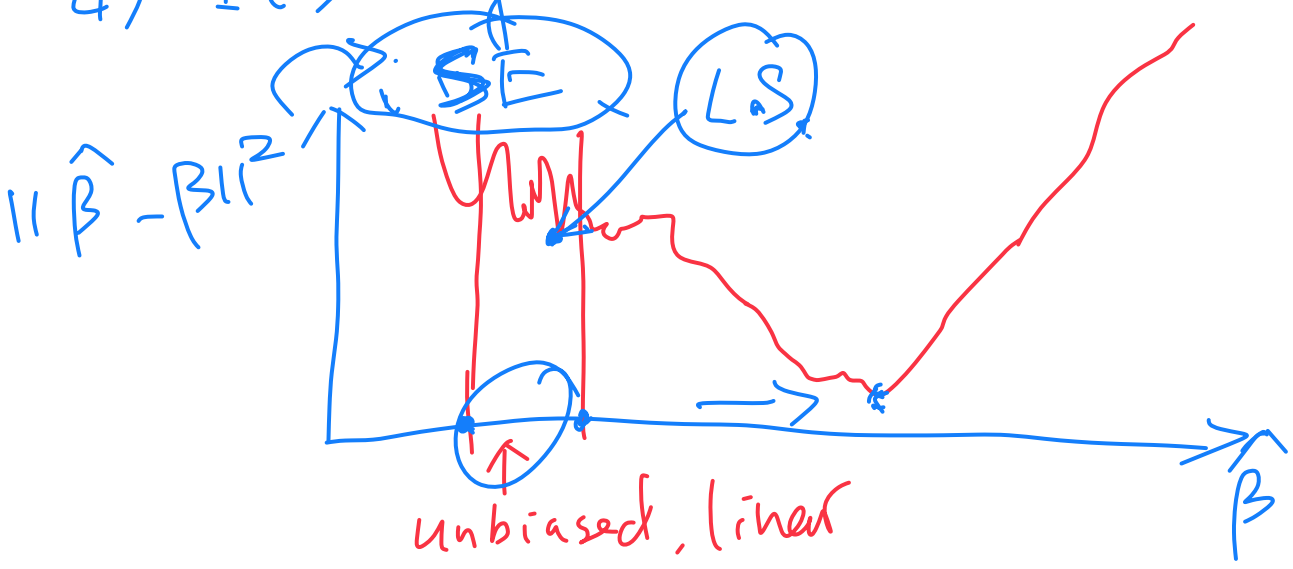
$$\begin{aligned}\hat{\boldsymbol{\beta}}_z &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = [(\mathbf{X}\mathbf{D})'(\mathbf{X}\mathbf{D})]^{-1}(\mathbf{X}\mathbf{D})'\mathbf{y} \\ &= \mathbf{D}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \qquad \text{[by (2.49)]} \\ &= \mathbf{D}^{-1}\hat{\boldsymbol{\beta}}, \qquad (7.18)\end{aligned}$$

where $\hat{\boldsymbol{\beta}}$ is the usual estimator for y regressed on the x's. Then

$$\hat{\boldsymbol{\beta}}_z'\mathbf{z} = (\mathbf{D}^{-1}\hat{\boldsymbol{\beta}})'\mathbf{D}\mathbf{x} = \hat{\boldsymbol{\beta}}'\mathbf{x}.$$

□

4) It's only the best in a subset



Estimator of σ^2

We estimate σ^2 by a corresponding average from the sample

$$s^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2, \tag{7.22}$$

where n is the sample size and k is the number of x's. Note that, by the corollary to Theorem 7.3d, $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$ is the BLUE of $\mathbf{x}_i'\boldsymbol{\beta}$.

Using (7.7), we can write (7.22) as

$$s^2 = \frac{1}{n-k-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{7.23}$$

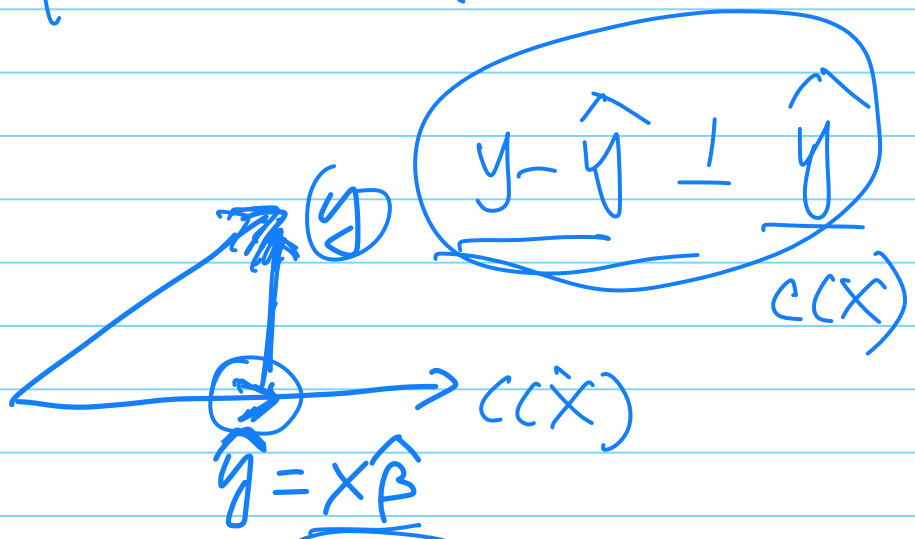
$$= \frac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}}{n-k-1} = \frac{\text{SSE}}{n-k-1}, \tag{7.24}$$

where $\text{SSE} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$. With the denominator $n-k-1$, s^2 is an unbiased estimator of σ^2, as shown below.

Theorem 7.3f. If s^2 is defined by (7.22), (7.23), or (7.24) and if $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, then

$$E(s^2) = \sigma^2. \tag{7.25}$$

Various expressions of $\|\hat{y}\|^2$



$$\langle y, \hat{y} \rangle = \langle \hat{y}, \hat{y} \rangle$$

$$\langle y-\hat{y}, \hat{y} \rangle = \langle y, \hat{y} \rangle - \langle \hat{y}, \hat{y} \rangle = 0$$

$$\|\hat{y}\|^2 = \langle \hat{y}, \hat{y} \rangle = \langle y, \hat{y} \rangle$$

$$= \hat{\beta}' X' y$$

$$= \hat{\beta}' X' X \hat{\beta}$$

$$= \|X\hat{\beta}\|^2$$

Directly, $\hat{y}' y = \hat{y}' (\hat{y} + y - \hat{y})$

$$= \hat{y}' \hat{y} + 0$$

Recall:

$E(y'Ay) = tr(A\cdot\Sigma) + \mu' A \mu$

or $E(\|Py\|^2) = rank(P)\cdot\sigma^2 + \|P\mu\|^2$ when $Var(y)=\sigma^2 I$

proof of Thm 7.3 f:

$H = X(X'X)^{-1}X'$ (proj matrix onto $C(X)$)

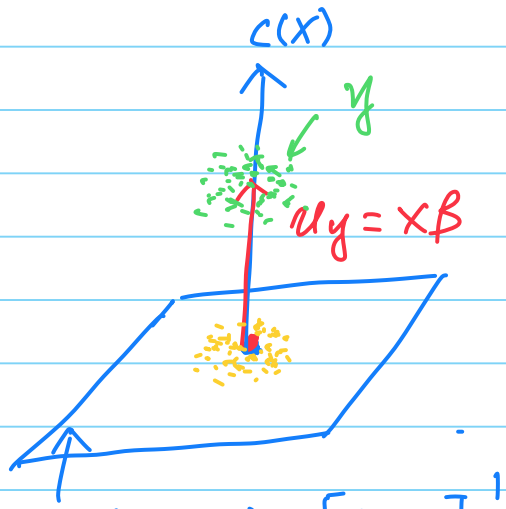
$\hat{y} = H\cdot y = X\hat{\beta}$

$y - \hat{y} = (I_n - H)y$

$SSE = \|y-\hat{y}\|^2$

$= y'\cdot(I_n - H)y$

$= \|(I_n - H)y\|^2$



$C(I-H) = [C(X)]^{\perp}$

$E(SSE)$

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

$+ \mu_y'(I_n - H)\mu_y$ (0)

$= (n-k+1)\sigma^2 + 0.$

$\mu_y = X\beta$

$(I_n - H)\mu_y = 0$

proof of $(I-H)\mu_y = 0$, where $\mu_y = X\beta$

$\mu_y = X\beta \in C(X)$

$(I - H)\mu_y = 0$

$\mu_y - H\mu_y$

$= X\beta - X(X'X)^{-1}X'X\beta$ (I)

$= X\beta - X\beta = 0$

so $\|(I-H)\mu_y\|^2 = 0$

Variance-Cov. matrix of $\hat{\beta}$

Corollary 1. An unbiased estimator of $\text{cov}(\hat{\boldsymbol{\beta}})$ in (7.14) is given by

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = s^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{7.27}$$

□

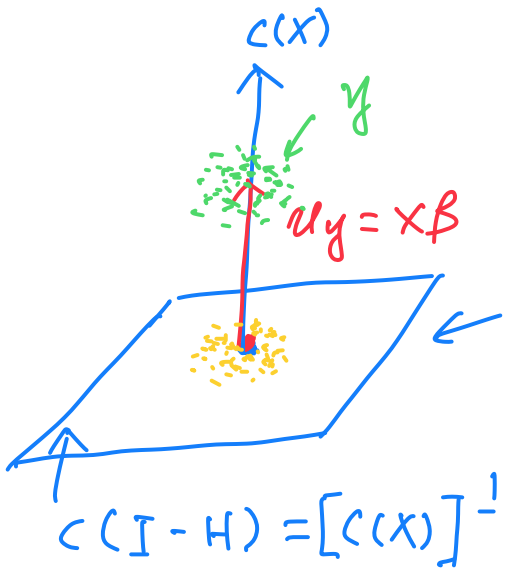
$\text{Cov}(\hat{\beta}) = \sigma^2 \cdot (X'X)^{-1}$

Distributions of $\hat{\beta}$ and s^2

$y \sim N(X\beta, \sigma^2 I)$

Theorem: Under assumptions A1–A5 of the classical linear model,

i. $\hat{\boldsymbol{\beta}} \sim N_{k+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$,
ii. $(n-k-1)s^2/\sigma^2 \sim \chi^2(n-k-1)$, and
iii. $\hat{\boldsymbol{\beta}}$ and s^2 are independent.



$\text{rank}(I-H)$
$= \text{tr}(I-H) = n-k-1$

$\mathcal{U}_y = X\beta$

$(I_n - H)\mathcal{U}_y = 0$

Proof:

(i)

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

$$y \sim N(X\beta, \sigma^2 I)$$

$$E(\hat{\beta}) = \beta$$

$$V(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

$$\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})$$

(ii)

$$SSE = y'(I - H)y$$

$$= y' P_{C(X)^\perp} y$$

$$\frac{SSE}{\sigma^2} \sim \chi^2\left(\underbrace{\text{rank}(I-H)}_{n-k-1}, \frac{1}{2} \cdot \frac{\|(I-H)X\|^2}{\sigma^2}\right)$$

$$(I - H)u = (I - H)X\beta = 0$$

(iii)

$\underline{H} = P_{C(X)} = X(X'X)^{-1}X'$

$\left[(X'X)^{-1}X'\right] X \cdot (X'X)^{-1}X'y$

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

That is, $\hat{\beta} = (X'X)^{-1}X' \cdot \hat{y}$

$\hat{y} = Hy$ and $(\underline{I} - H)y$ are indep

because $H \cdot (\underline{I} - H) = 0$

Maximum Likelihood Estimator

Theorem 7.6a. If $\mathbf{y}$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where $\mathbf{X}$ is $n \times (k+1)$ of rank $k+1 < n$, the maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2 are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{7.48}$$

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \quad = \frac{SSE}{n} \tag{7.49}$$

PROOF. We sketch the proof. For the remaining steps, see Problem 7.21. The likelihood function (joint density of $y_1, y_2, \ldots, y_n$) is given by the multivariate normal density (4.9)

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2) = f(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) &= \frac{1}{(2\pi)^{n/2}|\sigma^2\mathbf{I}|^{1/2}} e^{-(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/2\sigma^2}. \end{aligned} \tag{7.50}$$

[Since the y_i's are independent, $L(\boldsymbol{\beta}, \sigma^2)$ can also be obtained as $\prod_{i=1}^n f(y_i; \mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$.] Then $\ln L(\boldsymbol{\beta}, \sigma^2)$ becomes

$$\ln L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{7.51}$$

Taking the partial derivatives of $\ln L(\boldsymbol{\beta}, \sigma^2)$ with respect to $\boldsymbol{\beta}$ and σ^2 and setting the results equal to zero will produce (7.48) and (7.49). To verify that $\hat{\boldsymbol{\beta}}$ maximizes (7.50) or (7.51), see (7.10). □

Note: $s^2 = \dfrac{(y - X\hat{\beta})'(y - X\beta)}{n-k-1}$

$E(\hat{\sigma}^2) = \dfrac{(n-k+1)\sigma^2}{n} < \sigma^2$

For $k=1$, $MSE(\hat{\sigma}^2) < MSE(s^2)$

Linear Models in Centered Form

The regression model can be written

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i \\ &= \alpha + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + \cdots + \beta_k(x_{ik} - \bar{x}_k) + e_i, \end{aligned}$$

for $i = 1, \ldots, n$, where

$$\alpha = \beta_0 + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2 + \cdots + \beta_k \bar{x}_k, \qquad (\heartsuit)$$

and where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$.

In matrix form, the equivalence between the original model and centered model that we've written above becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} = (\mathbf{j}_n, \mathbf{X}_c)\begin{pmatrix} \alpha \\ \boldsymbol{\beta}_1 \end{pmatrix} + \mathbf{e},$$

where $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_k)^T$, and

$$\mathbf{X}_c = \underbrace{(\mathbf{I} - \frac{1}{n}\mathbf{J}_{n,n})}_{=\mathbf{P}_{\mathcal{L}(\mathbf{j}_n)^\perp}} \mathbf{X}_1 = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nk} - \bar{x}_k \end{pmatrix},$$

and $\mathbf{X}_1$ is the matrix consisting of all but the first columns of $\mathbf{X}$, the original model matrix.

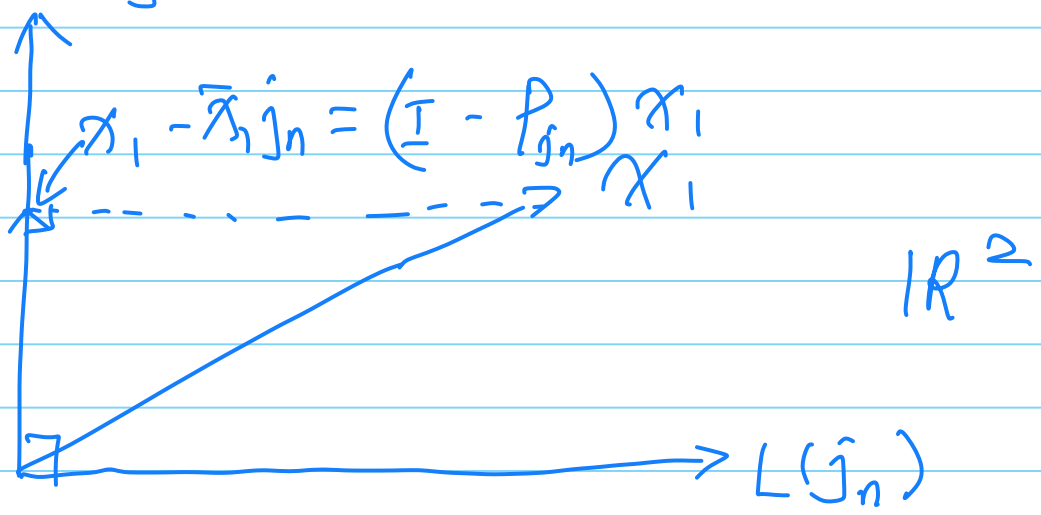
$$\begin{pmatrix} \hat{\alpha} \\ \hat{\boldsymbol{\beta}}_1 \end{pmatrix} = [(\mathbf{j}_n, \mathbf{X}_c)^T(\mathbf{j}_n, \mathbf{X}_c)]^{-1}(\mathbf{j}_n, \mathbf{X}_c)^T\mathbf{y} = \begin{pmatrix} n & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_c^T\mathbf{X}_c \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{j}_n^T \\ \mathbf{X}_c^T \end{pmatrix}\mathbf{y}$$
$$= \begin{pmatrix} n^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{X}_c^T\mathbf{X}_c)^{-1} \end{pmatrix}\begin{pmatrix} n\bar{y} \\ \mathbf{X}_c^T\mathbf{y} \end{pmatrix} = \begin{pmatrix} \bar{y} \\ (\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{y} \end{pmatrix},$$

$\hat{\alpha} = \bar{y}$, and

$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{y}.$

$\hat{\beta}_0 = \hat{\alpha} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x} - \cdots - \hat{\beta}_k\bar{x}_k = \bar{y} - \hat{\boldsymbol{\beta}}_1'\bar{\mathbf{x}}$

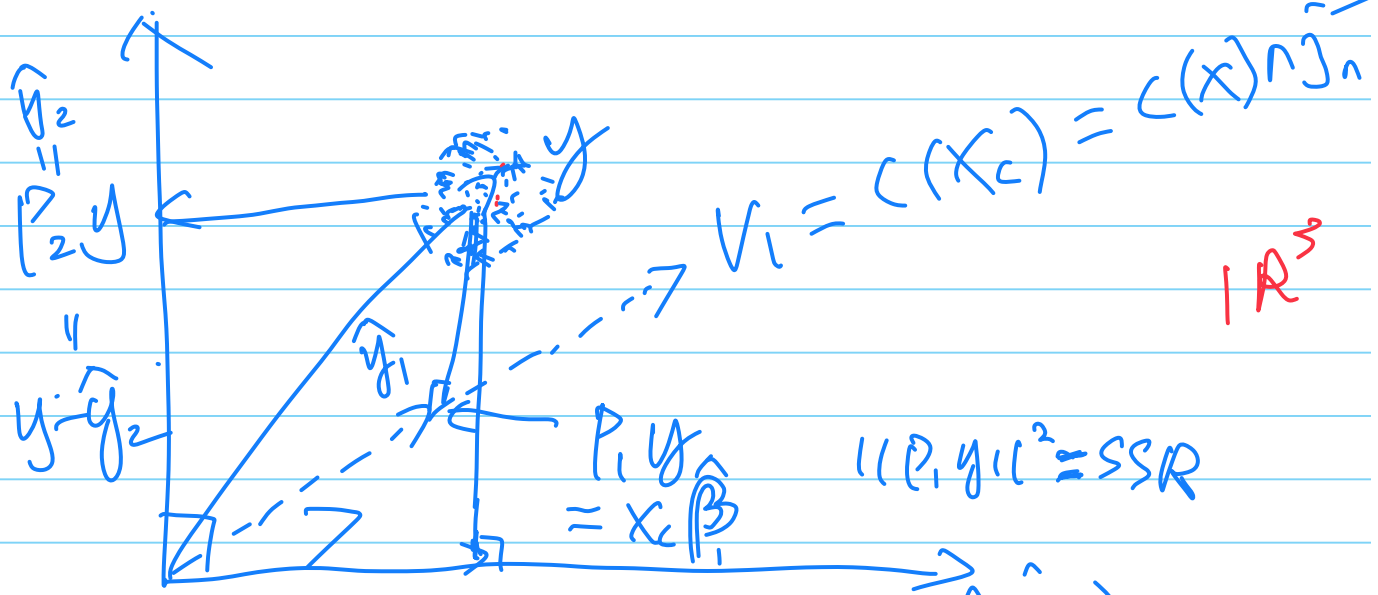
$L(\hat{j}_n)^{\perp} \cdot L(x_c) = c(x_{1,c})$



$$c(x_c) = c(x) \cap [L(\hat{j}_n)]^{\perp}$$
$$= c(P_{c(x)} - P_{L(\hat{j}_n)})$$
$$= c(P_{L(\hat{j}_n)^{\perp}} \cdot x)$$
$$= c(x - P_{L(\hat{j}_n)} x)$$

$$c(x) = c(x_c) \oplus L(\hat{j}_n)$$

$V_2 = C(X)^{\perp} = C(I - H)$



$H = X \cdot (X'X)^{-1} X'$

$C(X) = C([J_n, X_c])$

$P_1 y = X_c \cdot \hat{\beta}_1 = X_c (X_c' X_c)^{-1} X_c' y$

$SSR = \|P_1 y\|^2 = \hat{\beta}_1' X_c' X_c \hat{\beta}_1$

Notations of S_{xx} & S_{xy}

$$y_c = y - \bar{y}\,\hat{j}_n = (I - P_{\hat{j}_n})\,y$$

$$x_c' y_c = x_c' \cdot (y - \bar{y}\,\hat{j}_n)$$

$$= x_c' y - \bar{y}\,(x_c' \hat{j}_n)$$

$$= x_c' y$$

$$\frac{x_c' y}{n-1} = \widehat{Cov}(x, y) \equiv S_{xy}$$

$$\frac{x_c' x_c}{n-1} = \widehat{Cov}(x, x) \equiv S_{xx}$$

Example 7.3.1a. We use the data in Table 7.1 to illustrate computation of $\hat{\boldsymbol{\beta}}$ using (7.6).

$$\mathbf{y} = \begin{pmatrix} 2 \\ 3 \\ 2 \\ 7 \\ 6 \\ 8 \\ 10 \\ 7 \\ 8 \\ 12 \\ 11 \\ 14 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 0 & 2 \\ 1 & 2 & 6 \\ 1 & 2 & 7 \\ 1 & 2 & 5 \\ 1 & 4 & 9 \\ 1 & 4 & 8 \\ 1 & 4 & 7 \\ 1 & 6 & 10 \\ 1 & 6 & 11 \\ 1 & 6 & 9 \\ 1 & 8 & 15 \\ 1 & 8 & 13 \end{pmatrix}, \quad \mathbf{X}'\mathbf{X} = \begin{pmatrix} 12 & 52 & 102 \\ 52 & 395 & 536 \\ 102 & 536 & 1004 \end{pmatrix},$$

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 90 \\ 482 \\ 872 \end{pmatrix}, \quad (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} .97476 & .24290 & -.22871 \\ .24290 & .16207 & -.11120 \\ -.22871 & -.11120 & .08360 \end{pmatrix},$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} 5.3754 \\ 3.0118 \\ -1.2855 \end{pmatrix}.$$

$\hat{y} = \bar{y}$

$x_{1,c} = (0 - \bar{x}_1, 2 - \bar{x}_1, \ldots, 8 - \bar{x}_1)'$

$x_{2,c} = (2 - \bar{x}_2, 6 - \bar{x}_2, \ldots, 13 - \bar{x}_2)'$

$X_c = (x_{1,c}, x_{2,c})$: 12×2

$\hat{\beta}_1 = (X_c' X_c)^{-1} X_c' y$

Example 7.5. For the data in Table 7.1, we calculate $\hat{\boldsymbol{\beta}}_1$ and $\hat{\beta}_0$ using (7.46) and (7.47).

$$\hat{\boldsymbol{\beta}}_1 = \mathbf{S}_{xx}^{-1}\mathbf{s}_{yx} = \begin{pmatrix} 6.4242 & 8.5455 \\ 8.5455 & 12.4545 \end{pmatrix}^{-1} \begin{pmatrix} 8.3636 \\ 9.7273 \end{pmatrix}$$

$$= \begin{pmatrix} 3.0118 \\ -1.2855 \end{pmatrix},$$

$$\hat{\beta}_0 = \bar{y} - \mathbf{s}_{yx}'\mathbf{S}_{xx}^{-1}\bar{\mathbf{x}}$$

$$= 7.5000 - (3.0118, \ -1.2855)\begin{pmatrix} 4.3333 \\ 8.5000 \end{pmatrix}$$

$$= 7.500 - 2.1246 = 5.3754.$$

These values are the same as those obtained in Example 7.3.1a. □

Sum Squares in Centered

$$SST = \|y - \bar{y}\mathbf{j}_n\|^2 = \|(I - P_{\mathbf{j}_n})y\|^2$$

$$\begin{aligned}\text{SSE} &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = (\mathbf{y} - \mathbf{P}_{C(\mathbf{X})}\mathbf{y})^T(\mathbf{y} - \mathbf{P}_{C(\mathbf{X})}\mathbf{y}) \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{P}_{C(\mathbf{X})}\mathbf{y} - \mathbf{y}^T\mathbf{P}_{C(\mathbf{X})}\mathbf{y} + \mathbf{y}^T\mathbf{P}_{C(\mathbf{X})}\mathbf{y} \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{P}_{C(\mathbf{X})}\mathbf{y} = \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y}.\end{aligned}$$

$\hat{y}_0 = P_{\mathbf{j}_n} y$

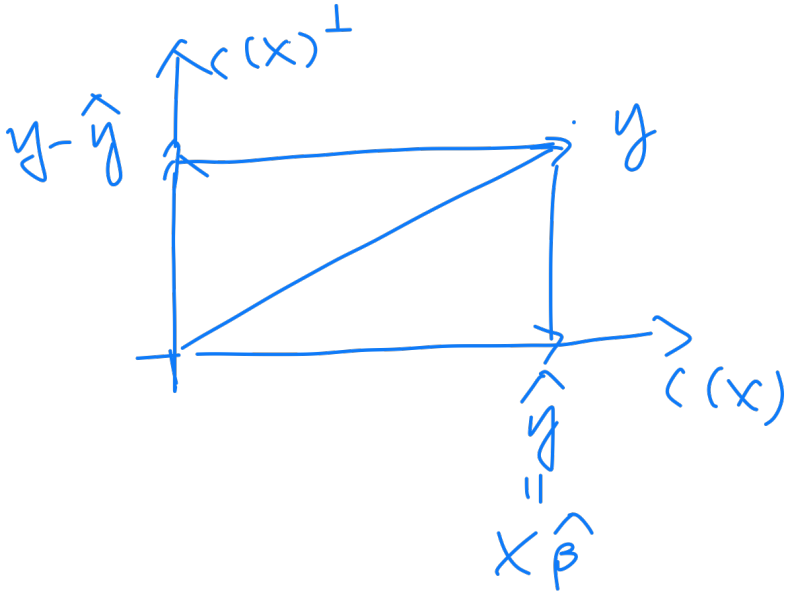
$= \bar{y}\mathbf{j}_n$

$\uparrow \|\hat{y}\|^2$

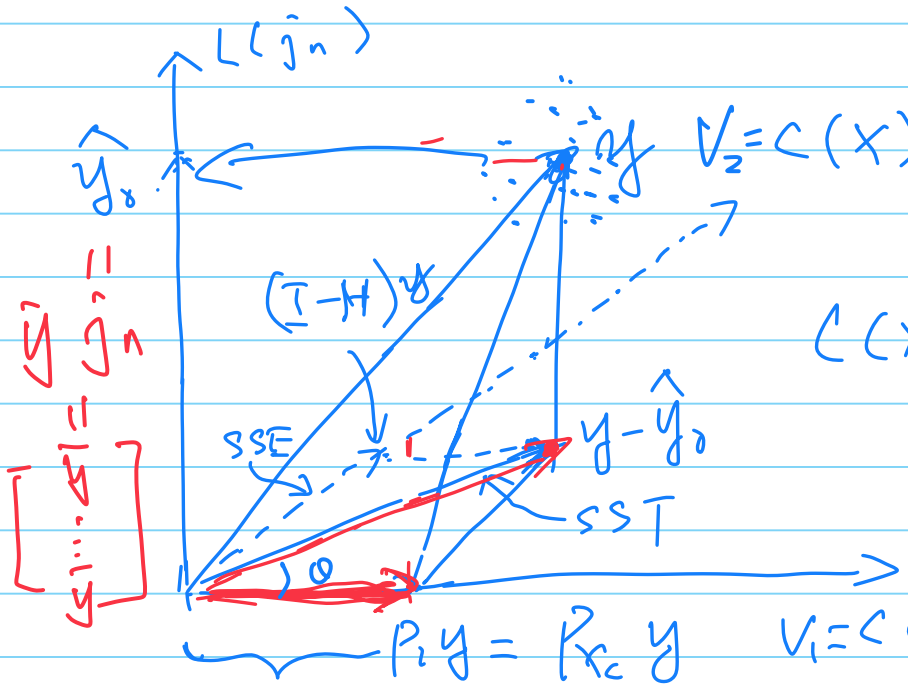
$$\begin{aligned}\text{SSE} &= \mathbf{y}^T\mathbf{y} - (\hat{\alpha}, \hat{\boldsymbol{\beta}}_1^T)\begin{pmatrix}\mathbf{j}_n^T \\ \mathbf{X}_c^T\end{pmatrix}\mathbf{y} \\ &= \mathbf{y}^T\mathbf{y} - \bar{y}\mathbf{j}_n^T\mathbf{y} - \hat{\boldsymbol{\beta}}_1^T\mathbf{X}_c^T\mathbf{y} \\ &= (\mathbf{y} - \bar{y}\mathbf{j}_n)^T\mathbf{y} - \hat{\boldsymbol{\beta}}_1^T\mathbf{X}_c^T\mathbf{y} \\ &= (\mathbf{y} - \bar{y}\mathbf{j}_n)^T(\mathbf{y} - \bar{y}\mathbf{j}_n) - \hat{\boldsymbol{\beta}}_1^T\mathbf{X}_c^T\mathbf{y} \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - \hat{\boldsymbol{\beta}}_1^T\mathbf{X}_c^T\mathbf{y}\end{aligned}$$

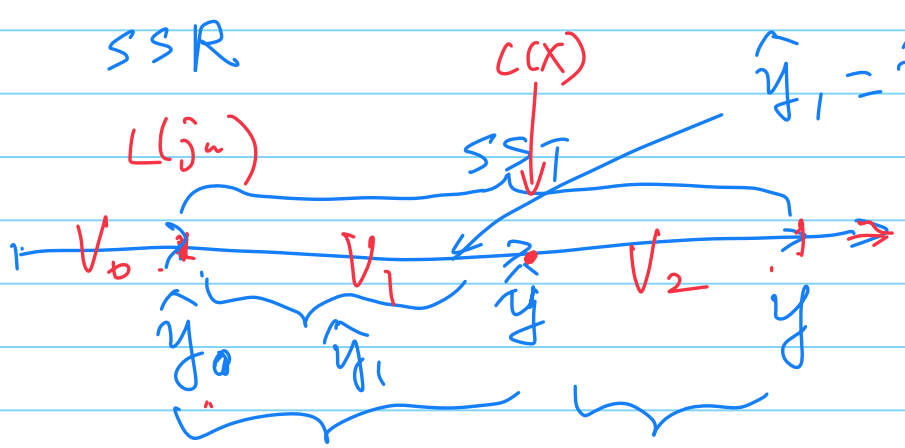
$$= \Sigma(y_i - \bar{y})^2 - y' X_c (X_c' X_c)^{-1} X_c' y$$

$$= SST - \|P_{X_c} y\|^2$$



$$u_y = u \cdot \hat{j}_n$$





$$SSR = \| P_{X_c} y \|^2 = X_c (X_c' X_c)^{-1} X_c' y$$

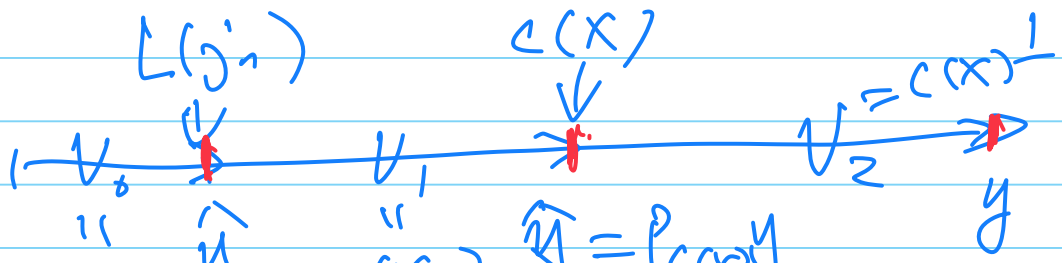
$$SSE = \| P_{C(X)^\perp} y \|^2 = \| (I-H) y \|^2$$

$$SST = SSR + SSE = \| y - \bar{y} j_n \|^2$$

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

$$H = X (X'X)^{-1} X' = P_X$$

$L(j_n) \subseteq C(X)$



$P_{X_c} y = \hat{y} - \hat{y}_0$

$\vdash$ SSR $\dashv\vdash$ SSE $\dashv$

$\vdash$ SST $\dashv$

$SST = SSR + SSE$

$C(X) = C(j_n, X_c)$

$= L(j_n) \oplus C(X_c)$

$j_n \perp X_c$

$SSR = \|\hat{y} - \hat{y}_0\|^2 = \|\hat{y}\|^2 - \|\hat{y}_0\|^2$

$SSE = \|y - \hat{y}\|^2 = \|y\|^2 - \|\hat{y}\|^2$

$SST = \|y - \bar{y} j_n\|^2 = \|y\|^2 - n \cdot \bar{y}^2$

R^2, the Estimated Coefficient of Determination

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \hat{\boldsymbol{\beta}}_1^T \mathbf{X}_c^T \mathbf{y} + \text{SSE}$$

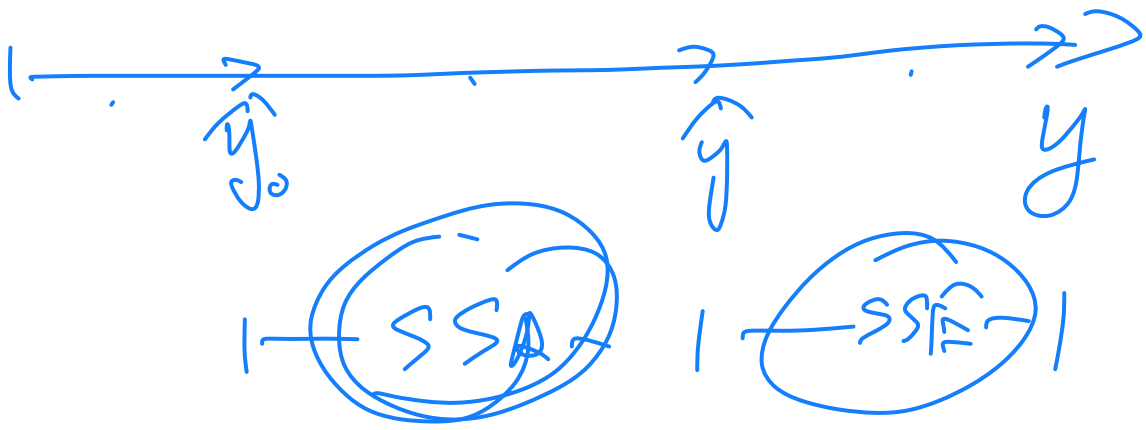
or $\text{SST} = \text{SSR} + \text{SSE}$

SST = SSR + SSE

$$\text{SSR} = \hat{\boldsymbol{\beta}}_1^T \mathbf{X}_c^T \mathbf{y} = \hat{\boldsymbol{\beta}}_1^T \mathbf{X}_c^T \mathbf{X}_c (\mathbf{X}_c^T \mathbf{X}_c)^{-1} \mathbf{X}_c^T \mathbf{y} = (\mathbf{X}_c \hat{\boldsymbol{\beta}}_1)^T (\mathbf{X}_c \hat{\boldsymbol{\beta}}_1).$$

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\hat{\boldsymbol{\beta}}_1^T \mathbf{X}_c^T \mathbf{X}_c \hat{\boldsymbol{\beta}}_1}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{y} - n\bar{y}^2}{\mathbf{y}^T \mathbf{y} - n\bar{y}^2}.$$

$= 1 - \frac{SSE}{SST}$



$1 - \frac{SSR}{SST}$ $\quad$ $1 - \frac{SSE}{SST} = 1$

$0 \le \frac{SSR}{SST} \le 1$

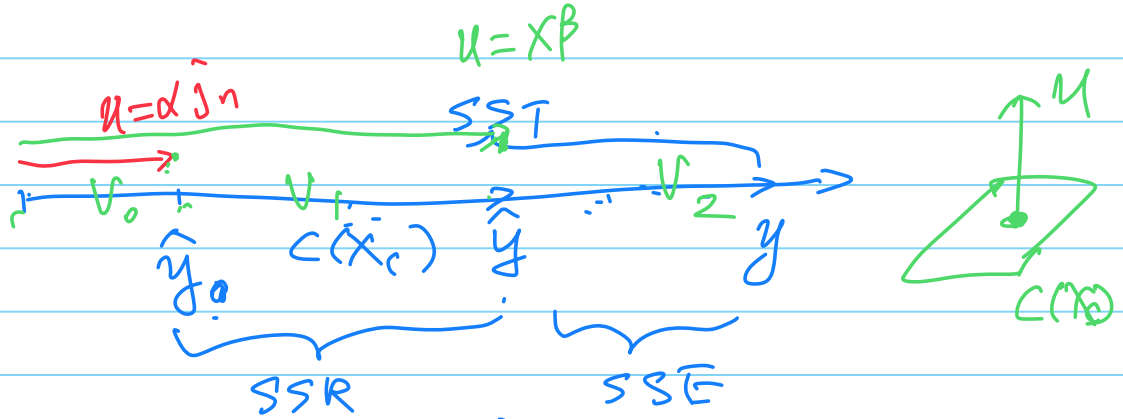
Adjusted R^2

$E(y) = \mu, \quad V(y) = \sigma^2 I$

P is a proj matrix with $\text{rank}(P) = r$. then

$E(\|Py\|^2) = E(y'Py) = \text{tr}(P \cdot \sigma^2 I) + \mu' P \mu$

$= \text{tr}(P) \cdot \sigma^2 + \|P\mu\|^2 = \text{rank}(P) \cdot \sigma^2 + \|P\mu\|^2$



$SSR = \|P_{X_c} y\|^2$

$SSE = \|P_{C(X)^\perp} y\|^2 = \|(I - P_X) y\|^2$

$\mu = E(y) = X\beta = \alpha j_n + X_c \beta_1$

$E(SSR) = \text{rank}(P_{X_c})\sigma^2 + \|P_{X_c}\mu\|^2$ $\quad j_n \perp X_c$

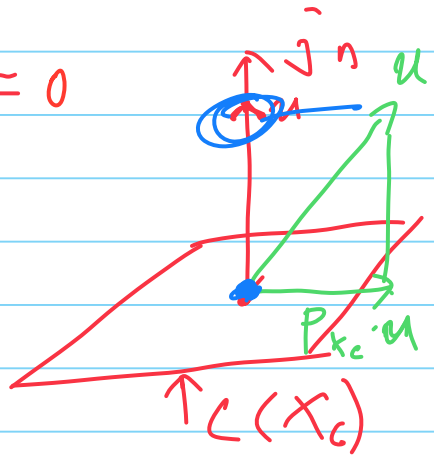
$= k\sigma^2 + \|X_c\beta_1\|^2$

$E(SSE) = (n-k-1)\sigma^2 + \|P_{C(X)^\perp} X\beta\|^2$

$= (n-k-1)\sigma^2$

$E(R^2)$ when $\beta_1 = \cdots = \beta_k = 0$

$[\beta_1 = 0]$



$u \in C(\vec{j}_n) \perp C(X_c)$

$P_{X_c} u = X_c \beta_1 = 0$

$E(SSR) = \text{rank}(X_c)\sigma^2 = k \cdot \sigma^2$

$E(SSE) = \text{rank}(I - P_X)\sigma^2 + \|(I - P_X)u\|^2$

$= (n-k-1)\sigma^2 + 0$

since $u = X\beta \in C(X) \perp C(I - P_X)$.

Also, SSR indep SSE.

$E(SST) = (n-1)\sigma^2 = \text{rank}(\vec{j}_n^{\perp}) \cdot \sigma^2$

Simply, $E(\|Py\|^2) = \text{Dim}(P)\sigma^2$

$$R^2 = \frac{SSR}{SSR+SSE}$$

$\xrightarrow{E}$ $(n-k-1)\sigma^2$

$$\frac{1}{R^2} = 1 + \frac{SSE}{SSR}$$

note $SSE \xrightarrow{E} k\sigma^2$ indep SSR

$$E\left(\frac{1}{R^2}\right) = 1 + \frac{n-k-1}{k}$$

$k \leftarrow \dim\left(\hat{J}_n^{-\frac{1}{2}}\right)$

$$= \frac{n-1}{k}$$

$n-1 \leftarrow \dim(X_c)$

$$E(R^2) \simeq \frac{k}{n-1}$$ as $k \uparrow$

We expect $E(R^2) = 0$ when

$\beta_1 = \beta_2 = \cdots = \beta_k = 0$

Indeed SSE always decreases as $k \uparrow$.

Let $df_e = n - k - 1 = \text{rank}(C(X)^{\perp})$

$df_{reg} = k = \text{rank}(X_c)$

$df_e + df_{reg} = n - 1$

Adjusted R^2:

$$R^2_a = 1 - \frac{SSE / df_e}{SST / (df_e + df_{reg})}$$

(SSE / df_e ← MSE; $SST / (df_e + df_{reg})$ ← MST)

$$= 1 - \frac{SSE}{SST} \cdot \frac{n-1}{n-k-1}$$

$$E(R_a^2) \approx 0$$

when $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

$$E\left(\frac{SSE}{SST}\right) \approx \frac{n-k-1}{n-1}$$

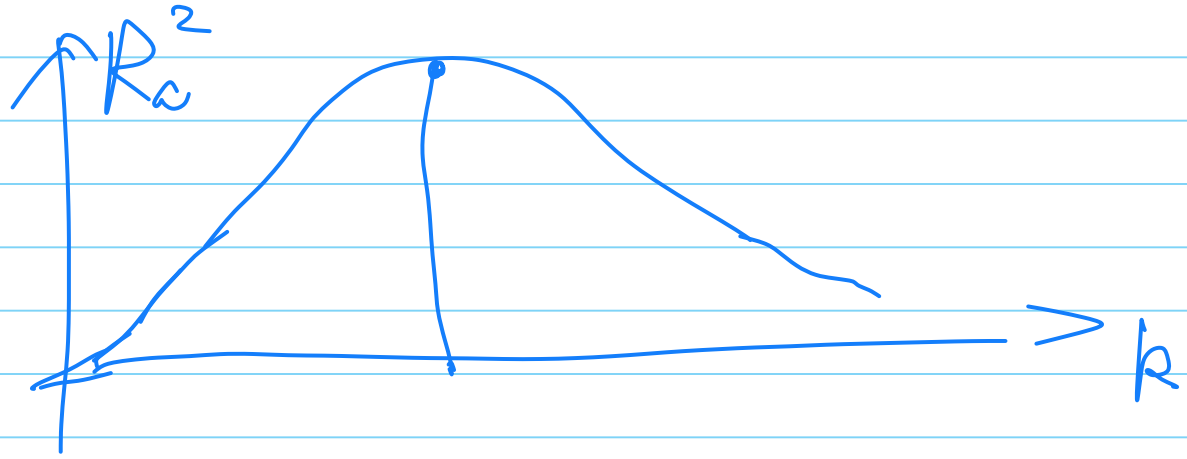
$$E(R_a^2) \approx 1 - 1 = 0$$

This is desired as

$$\beta_1 = \beta_2 = \cdots = \beta_k = 0$$

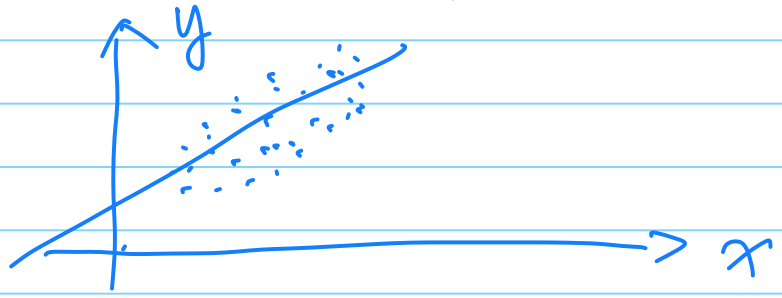
Large R^2 but low R_a^2 is a sign of overfitting

R_a^2 is a criterion for selecting k:



R_a^2 as an estimate of $\rho^2 = 1 - \frac{E(V(y|x))}{V(y)}$

$$V(y) = E(V(y|x)) + V(E(y|x))$$



suppose $y = x\beta + e$, $Cov(x, e) = 0$

$SSE/(n-k-1) = \widehat{Var}(e)$

$SST/(n-1) = \widehat{Var}(y)$ [x is a r.v. too]

$R_a^2 = 1 - \frac{\widehat{Var}(e)}{\widehat{Var}(y)}$ ($\sim \frac{Var(e)}{Var(y)}$)

$= 1 - \frac{MSE}{MST}$.

Underfitting and Overfitting

Suppose that the true model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\mathbf{e}$ where we return to the spherical errors case: $\text{var}(\mathbf{e}) = \sigma^2\mathbf{I}$. We want to consider what happens when we omit some explanatory variable is $\mathbf{X}$ and when we include too many x's. So, let's partition our model as

$$\begin{aligned}\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} &= (\mathbf{X}_1, \mathbf{X}_2)\begin{pmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{pmatrix} + \mathbf{e} \\ &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}. \qquad (\dagger)\end{aligned}$$

- If we leave out $\mathbf{X}_2\boldsymbol{\beta}_2$ when it should be included (when $\boldsymbol{\beta}_2 \neq \mathbf{0}$) then we are **underfitting**.
- If we include $\mathbf{X}_2\boldsymbol{\beta}_2$ when it doesn't belong in the true model (when $\boldsymbol{\beta}_2 = \mathbf{0}$) then we are **overfitting**.
- We will consider the effects of both overfitting and underfitting on the bias and variance of $\hat{\boldsymbol{\beta}}$. The book also consider effects on predicted values and on the MSE s^2.

Underfitting:

Suppose model (†) holds, but we fit the model

reduced

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1^* + \mathbf{e}^*, \quad \text{var}(\mathbf{e}^*) = \sigma^2\mathbf{I}. \qquad (\clubsuit)$$

The following theorem gives the bias and var-cov matrix of $\hat{\boldsymbol{\beta}}_1^*$ the OLS estimator from ♣.

Theorem: If we fit model ♣ when model (†) is the true model, then the mean and var-cov matrix of the OLS estimator $\hat{\boldsymbol{\beta}}_1^* = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y}$ are as follows:

(i) $\text{E}(\hat{\boldsymbol{\beta}}_1^*) = \boldsymbol{\beta}_1 + \mathbf{A}\boldsymbol{\beta}_2$, where $\mathbf{A} = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2$.

(ii) $\text{var}(\hat{\boldsymbol{\beta}}_1^*) = \sigma^2(\mathbf{X}_1^T\mathbf{X}_1)^{-1}$.

Proof:

(i)

$$\begin{aligned}\text{E}(\hat{\boldsymbol{\beta}}_1^*) &= \text{E}[(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y}] = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\text{E}(\mathbf{y}) \\ &= (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2) \\ &= \boldsymbol{\beta}_1 + \mathbf{A}\boldsymbol{\beta}_2. \qquad \neq \beta_1\end{aligned}$$

(ii)

$$\begin{aligned}\text{var}(\hat{\boldsymbol{\beta}}_1^*) &= \text{var}[(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y}] \\ &= (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T(\sigma^2\mathbf{I})\mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1} \\ &= \sigma^2(\mathbf{X}_1^T\mathbf{X}_1)^{-1}.\end{aligned}$$

■

- This result says that when underfitting, $\hat{\boldsymbol{\beta}}_1^*$ is biased by an amount that depends upon both the omitted and included explanatory variables.

Corollary If $\mathbf{X}_1^T\mathbf{X}_2 = \mathbf{0}$, i.e.. if the columns of $\mathbf{X}_1$ are orthogonal to the columns of $\mathbf{X}_2$, then $\hat{\boldsymbol{\beta}}_1^*$ is unbiased.

Remarks:
1) Orthogonal design
2) randomization

overfitting

True model: (Reduced)

$$y = X_1 \beta_1^* + \epsilon$$

Fitted Model: (Full)

$$y = X_1 \beta_1 + X_2 \beta_2 + \epsilon$$

Note that the fitted model is the true model for the $y \mid X_1, X_2$ with true parameter

$$\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \beta_1^* \\ 0 \end{bmatrix}.$$

$$E\left(\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}\right) = \begin{pmatrix} \beta_1^* \\ 0 \end{pmatrix} \Rightarrow E(\hat{\beta}_1) = \beta_1^*$$

fitting

$\hat{\beta}_1$ (from the Full model) is unbiased.

Note that in the above theorem the var-cov matrix of $\hat{\boldsymbol{\beta}}_1^*$, $\sigma^2(\mathbf{X}_1^T\mathbf{X}_1)^{-1}$ is not the same as the var-cov matrix of $\hat{\boldsymbol{\beta}}_1$, the corresponding portion of the OLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ from the full model. How these var-cov matrices differ is established in the following theorem:

n.n.d not negative definite

Theorem: Let $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ from the full model (†) be partitioned as

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}$$

and let $\hat{\boldsymbol{\beta}}_1^* = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y}$ be the estimator from the reduced model ♣. Then

$$\text{var}(\hat{\boldsymbol{\beta}}_1) - \text{var}(\hat{\boldsymbol{\beta}}_1^*) = \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^T$$

a n.n.d. matrix. Here, $\mathbf{A} = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2$ and $\mathbf{B} = \mathbf{X}_2^T\mathbf{X}_2 - \mathbf{X}_2^T\mathbf{X}_1\mathbf{A}$.

- Thus $\text{var}(\hat{\beta}_j) \geq \text{var}(\hat{\beta}_j^*)$, meaning that underfitting results in smaller variances of the $\hat{\beta}_j$'s and overfitting results in larger variances of the $\hat{\beta}_j$'s.

Proof: Partitioning $\mathbf{X}^T\mathbf{X}$ to conform to the partitioning of $\mathbf{X}$ and $\boldsymbol{\beta}$, we have

$$\begin{aligned} \text{var}(\hat{\boldsymbol{\beta}}) = \text{var}\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2\begin{pmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_2 \\ \mathbf{X}_2^T\mathbf{X}_1 & \mathbf{X}_2^T\mathbf{X}_2 \end{pmatrix}^{-1} \\ &= \sigma^2\begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{pmatrix}^{-1} = \sigma^2\begin{pmatrix} \mathbf{H}^{11} & \mathbf{H}^{12} \\ \mathbf{H}^{21} & \mathbf{H}^{22} \end{pmatrix}, \end{aligned}$$

$\begin{pmatrix} X_1' \\ X_2' \end{pmatrix} \cdot (X_1, X_2)$

where $\mathbf{H}_{ij} = \mathbf{X}_i^T\mathbf{X}_j$ and $\mathbf{H}^{ij}$ is the corresponding block of the inverse matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ (see p. 54).

So, $\text{var}(\hat{\boldsymbol{\beta}}_1) = \sigma^2\mathbf{H}^{11}$. Using the formulas for inverses of partitioned matrices,

$$\mathbf{H}^{11} = \mathbf{H}_{11}^{-1} + \mathbf{H}_{11}^{-1}\mathbf{H}_{12}\mathbf{B}^{-1}\mathbf{H}_{21}\mathbf{H}_{11}^{-1},$$

where

$$\mathbf{B} = \mathbf{H}_{22} - \mathbf{H}_{21}\mathbf{H}_{11}^{-1}\mathbf{H}_{12}.$$

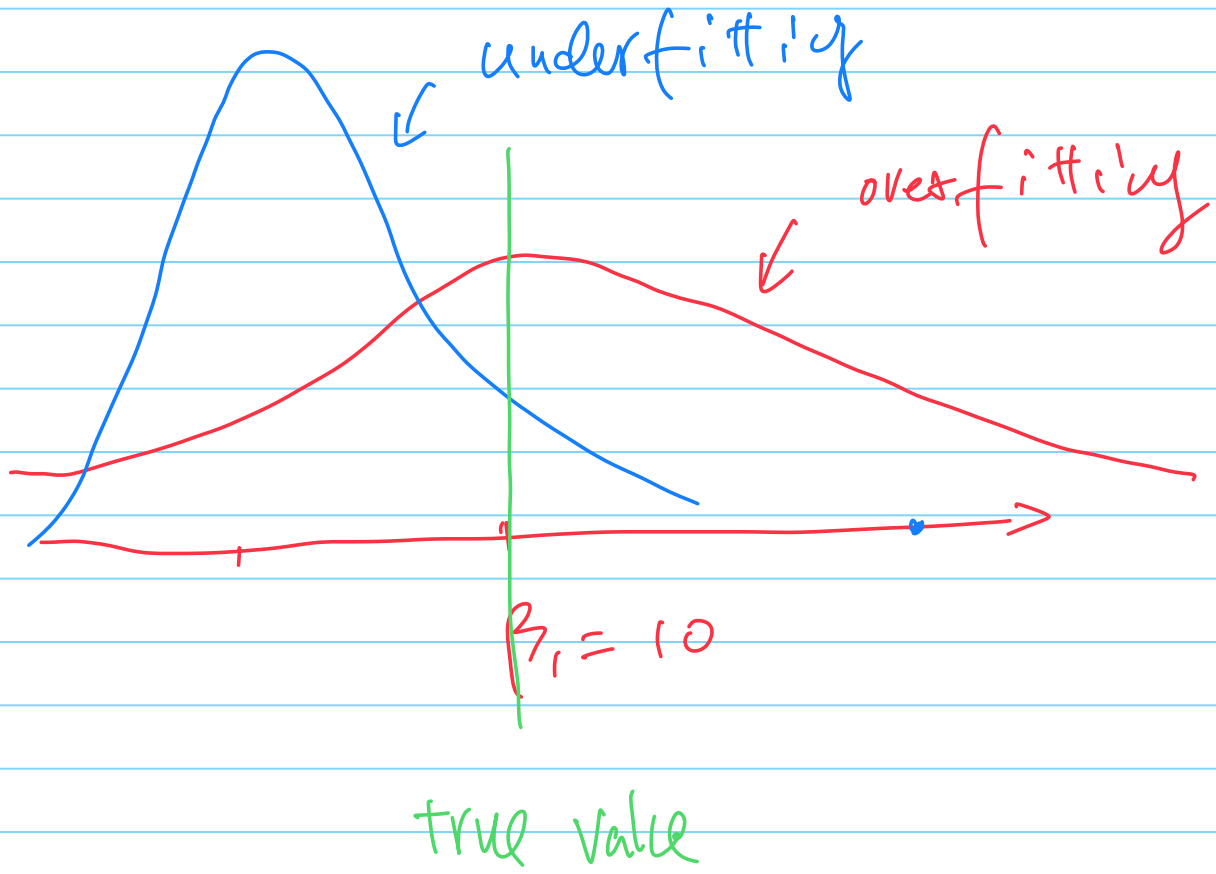
B is n.n.d. b/c X'X is n.n.d.

In the previous theorem, we showed that $\text{var}(\hat{\boldsymbol{\beta}}_1^*) = \sigma^2(\mathbf{X}_1^T\mathbf{X}_1)^{-1} = \sigma^2\mathbf{H}_{11}^{-1}$. Hence,

$$\begin{aligned} \text{var}(\hat{\boldsymbol{\beta}}_1) - \text{var}(\hat{\boldsymbol{\beta}}_1^*) &= \sigma^2(\mathbf{H}^{11} - \mathbf{H}_{11}^{-1}) \\ &= \sigma^2(\mathbf{H}_{11}^{-1} + \mathbf{H}_{11}^{-1}\mathbf{H}_{12}\mathbf{B}^{-1}\mathbf{H}_{21}\mathbf{H}_{11}^{-1} - \mathbf{H}_{11}^{-1}) \\ &= \sigma^2(\mathbf{H}_{11}^{-1}\mathbf{H}_{12}\mathbf{B}^{-1}\mathbf{H}_{21}\mathbf{H}_{11}^{-1}) \\ &= \sigma^2[(\mathbf{X}_1^T\mathbf{X}_1)^{-1}(\mathbf{X}_1^T\mathbf{X}_2)\mathbf{B}^{-1}(\mathbf{X}_2^T\mathbf{X}_1)(\mathbf{X}_1^T\mathbf{X}_1)^{-1}] \\ &= \sigma^2\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^T. \end{aligned}$$

$V(\alpha_1'\hat{\beta}_1) - \text{Var}(\alpha'\hat{\beta}_1^*) = \sigma^2 \cdot \alpha' A B^{-1} A' \alpha$

where $\alpha_1' = (\alpha', 0, \cdots, 0)$ for X_2


underfitting
overfitting
$\beta_1 = 10$
true value

To summarize, we've seen that underfitting reduces the variances of regression parameter estimators, but introduces bias. On the other hand, overfitting produces unbiased estimators with increased variances. Thus it is the task of a regression model builder to find an optimum set of explanatory variables to balance between a biased model and one with large variances.

Occam's razor, Ockham's razor, Ocham's razor (Latin: *novacula Occami*), also known as the principle of parsimony or the law of parsimony (Latin: *lex parsimoniae*), is the problem-solving principle that "entities should not be multiplied beyond necessity".