Lecture Notes for Theory of Linear Models

Statistical Inference for Linear Models (Ch8)

Longhai Li
Department of Mathematics and Statistics
University of Saskatchewan

Testing Reduced Model vs Full Model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} = (\mathbf{X}_1, \mathbf{X}_2)\begin{pmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{pmatrix} + \mathbf{e}$$
$$= \underbrace{\mathbf{X}_1}_{n\times(k+1-h)}\boldsymbol{\beta}_1 + \underbrace{\mathbf{X}_2}_{n\times h}\boldsymbol{\beta}_2 + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \qquad \text{(FM)}$$

where we are interested in the hypothesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$.

Under $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ the model becomes

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1^* + \mathbf{e}^*, \quad \mathbf{e}^* \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \qquad \text{(RM)}$$

$u = E(y)$

The problem is to test

$$H_0 : \boldsymbol{\mu} \in C(\mathbf{X}_1) \quad \text{(RM)} \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \notin C(\mathbf{X}_1) \quad \text{(Full model)}$$

under the *maintained hypothesis* that $\boldsymbol{\mu} \in C(\mathbf{X}) = C([\mathbf{X}_1, \mathbf{X}_2])$ (FM).

Example:

H_0:

$$y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \varepsilon$$

H_1:

$$y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \dots + \beta_k x^k + \varepsilon$$

$k > 2$

3

Note that under RM, $\boldsymbol{\mu} \in C(\mathbf{X}_1) \subset C(\mathbf{X}) = C([\mathbf{X}_1, \mathbf{X}_2])$. Therefore, if RM is true, then FM must be true as well. So, if RM is true, then the least squares estimates of the mean $\boldsymbol{\mu}$: $\mathbf{P}_{C(\mathbf{X}_1)}\mathbf{y}$ and $\mathbf{P}_{C(\mathbf{X})}\mathbf{y}$ are estimates of the same thing.

This suggests that the difference between the two estimates

$$\mathbf{P}_{C(\mathbf{X})}\mathbf{y} - \mathbf{P}_{C(\mathbf{X}_1)}\mathbf{y} = (\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y} = \hat{y} - \hat{y}_0$$

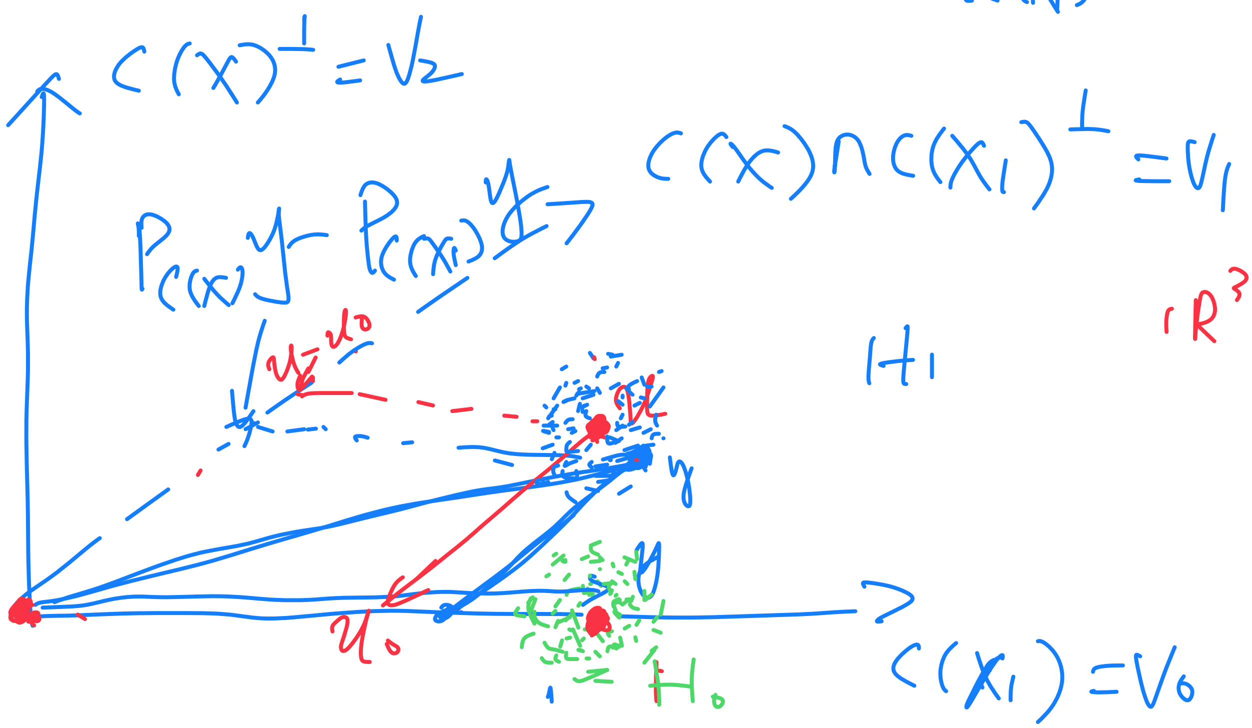
should be small under $H_0 : \boldsymbol{\mu} \in C(\mathbf{X}_1)$.

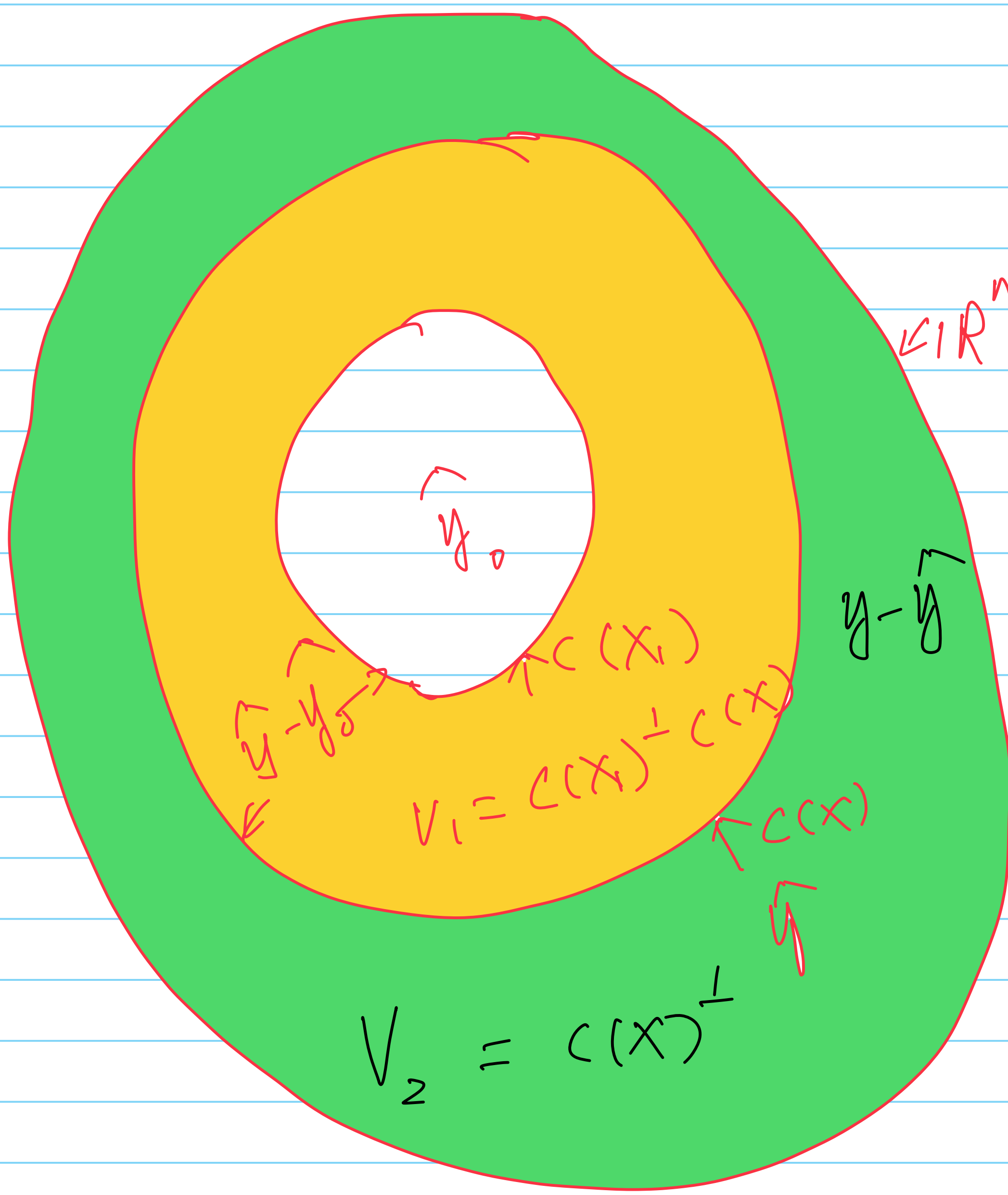
- Note that $\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)}$ is the projection matrix onto $C(\mathbf{X}_1)^{\perp} \cap C(\mathbf{X})$, the orthogonal complement of $C(\mathbf{X}_1)$ with respect to $C(\mathbf{X})$, and $C(\mathbf{X}_1) \oplus [C(\mathbf{X}_1)^{\perp} \cap C(\mathbf{X})] = C(\mathbf{X})$.

So, under H_0, $(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}$ should be "small". A measure of the "smallness" of this vector is its squared length:

$$SSH = \|(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}\|^2 = \mathbf{y}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}.$$

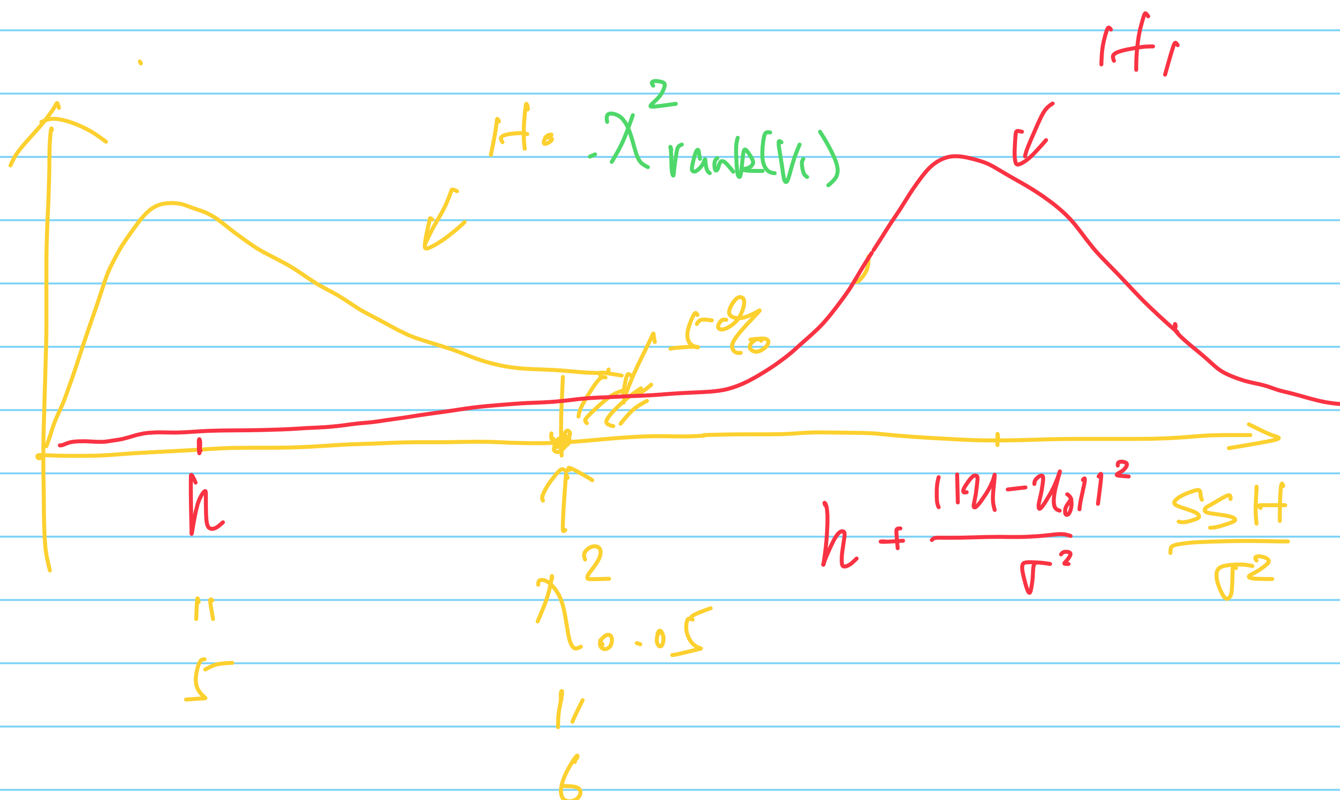
$\mu_0 = P_{C(X_1)} \cdot \mu$





$\mathbb{R}^n$
$\hat{y}_0$
$C(X_1)$
$\hat{y}-\hat{y}_0$
$C(X)^{\perp} C(X)$
$V_1 = C(X_1)^{\perp} C(X)$
$C(X)$
$\hat{y}$
$y-\hat{y}$
$V_2 = C(X)^{\perp}$

Distribution of $\frac{SSH}{\sigma^2}$



6

$$
\begin{aligned}
\mathrm{E}[\mathbf{y}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}] &= \sigma^2 \dim[C(\mathbf{X}_1)^\perp \cap C(\mathbf{X})] + \boldsymbol{\mu}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\boldsymbol{\mu} \\
&= \sigma^2 h + [(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\boldsymbol{\mu}]^T[(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\boldsymbol{\mu}] \\
&= \sigma^2 h + (\mathbf{P}_{C(\mathbf{X})}\boldsymbol{\mu} - \mathbf{P}_{C(\mathbf{X}_1)}\boldsymbol{\mu})^T(\mathbf{P}_{C(\mathbf{X})}\boldsymbol{\mu} - \mathbf{P}_{C(\mathbf{X}_1)}\boldsymbol{\mu})
\end{aligned}
$$

$\|(P_{C(X)} - P_{C(X_1)})\mu\|^2$

$\|$

$\mu - \mu_0$

Under H_0, $\boldsymbol{\mu} \in C(\mathbf{X}_1)$ and $\boldsymbol{\mu} \in C(\mathbf{X})$, so

$$(\mathbf{P}_{C(\mathbf{X})}\boldsymbol{\mu} - \mathbf{P}_{C(\mathbf{X}_1)}\boldsymbol{\mu}) = \boldsymbol{\mu} - \boldsymbol{\mu} = \mathbf{0}.$$

Under H_1,

$$\mathbf{P}_{C(\mathbf{X})}\boldsymbol{\mu} = \boldsymbol{\mu}, \quad \text{but} \quad \mathbf{P}_{C(\mathbf{X}_1)}\boldsymbol{\mu} \neq \boldsymbol{\mu}.$$

I.e., letting $\boldsymbol{\mu}_0$ denote $p(\boldsymbol{\mu}|C(\mathbf{X}_1))$,

$$\mathrm{E}[\mathbf{y}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}] = \begin{cases} \sigma^2 h, & \text{under } H_0; \\ \sigma^2 h + \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|^2, & \text{under } H_1. \end{cases}$$

$\mu - \mu_0 = (X_1\beta_1 + X_2\beta_2) - P_{C(X_1)} \cdot \mu = X_2\beta_2$

if $X_1 \perp X_2$

Therefore, if σ^2 is known

$$\frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2}{\sigma^2 h} = \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2/h}{\sigma^2} \begin{cases} \approx 1, & \text{under } H_0 \\ > 1, & \text{under } H_1 \end{cases}$$

is an appropriate test statistic for testing H_0.

Typically, σ^2 will not be known, so it must be estimated. The appropriate estimator is $s^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2/(n-k-1)$, the mean squared error from FM, the model which is valid under H_0 and under H_1. Our test statistic then becomes

$$F = \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2/h}{s^2} = \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2/h}{\|\mathbf{y} - \hat{\mathbf{y}}\|^2/(n-k-1)} \begin{cases} \approx 1, & \text{under } H_0 \\ > 1, & \text{under } H_1. \end{cases}$$

MSE

σ^2

$h = \text{rank}(P_{C(X)} - P_{C(X_1)})$

Theorem: Suppose $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ where $\mathbf{X}$ is $n \times (k+1)$ of full rank where $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$, and $\mathbf{X}_2$ is $n \times h$. Let $\hat{\mathbf{y}} = p(\mathbf{y}|C(\mathbf{X})) = \mathbf{P}_{C(\mathbf{X})}\mathbf{y}$, $\hat{\mathbf{y}}_0 = p(\mathbf{y}|C(\mathbf{X}_1)) = \mathbf{P}_{C(\mathbf{X}_1)}\mathbf{y}$, and $\boldsymbol{\mu}_0 = p(\boldsymbol{\mu}|C(\mathbf{X}_1)) = \mathbf{P}_{C(\mathbf{X}_1)}\boldsymbol{\mu}$. Then

(i) $\frac{1}{\sigma^2}\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \frac{1}{\sigma^2}\mathbf{y}^T(\mathbf{I} - \mathbf{P}_{C(\mathbf{X})})\mathbf{y} \sim \chi^2(n-k-1)$;

(ii) $\frac{1}{\sigma^2}\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2 = \frac{1}{\sigma^2}\mathbf{y}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y} \sim \chi^2(h, \lambda_1)$, where

$$\lambda_1 = \frac{1}{2\sigma^2}\|(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\boldsymbol{\mu}\|^2 = \frac{1}{2\sigma^2}\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|^2;$$

and

$\mu_0 = P_{C(X_1)} \cdot \mu$

(iii) $\frac{1}{\sigma^2}\|\mathbf{y} - \hat{\mathbf{y}}\|^2$ and $\frac{1}{\sigma^2}\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2$ are independent.

$SSH = \|\hat{y} - \hat{y}_0\|^2$, $SSE = \|y - \hat{y}\|^2$

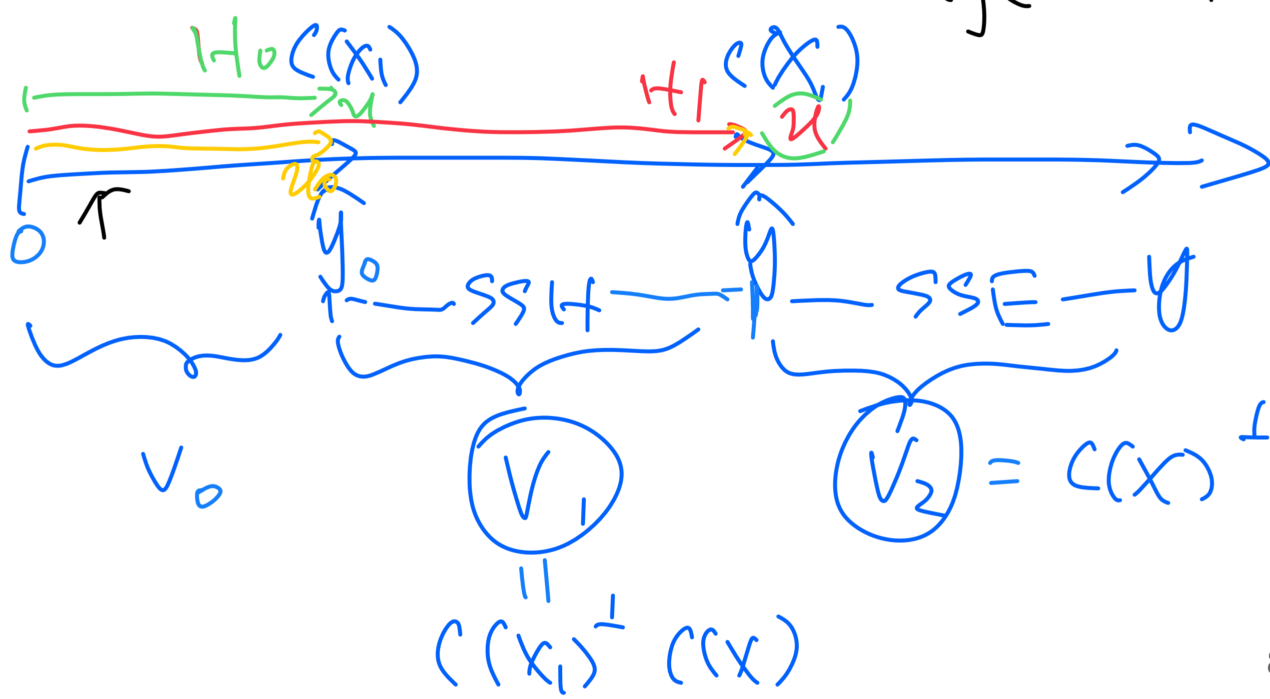
Theorem: Under the conditions of the previous theorem,

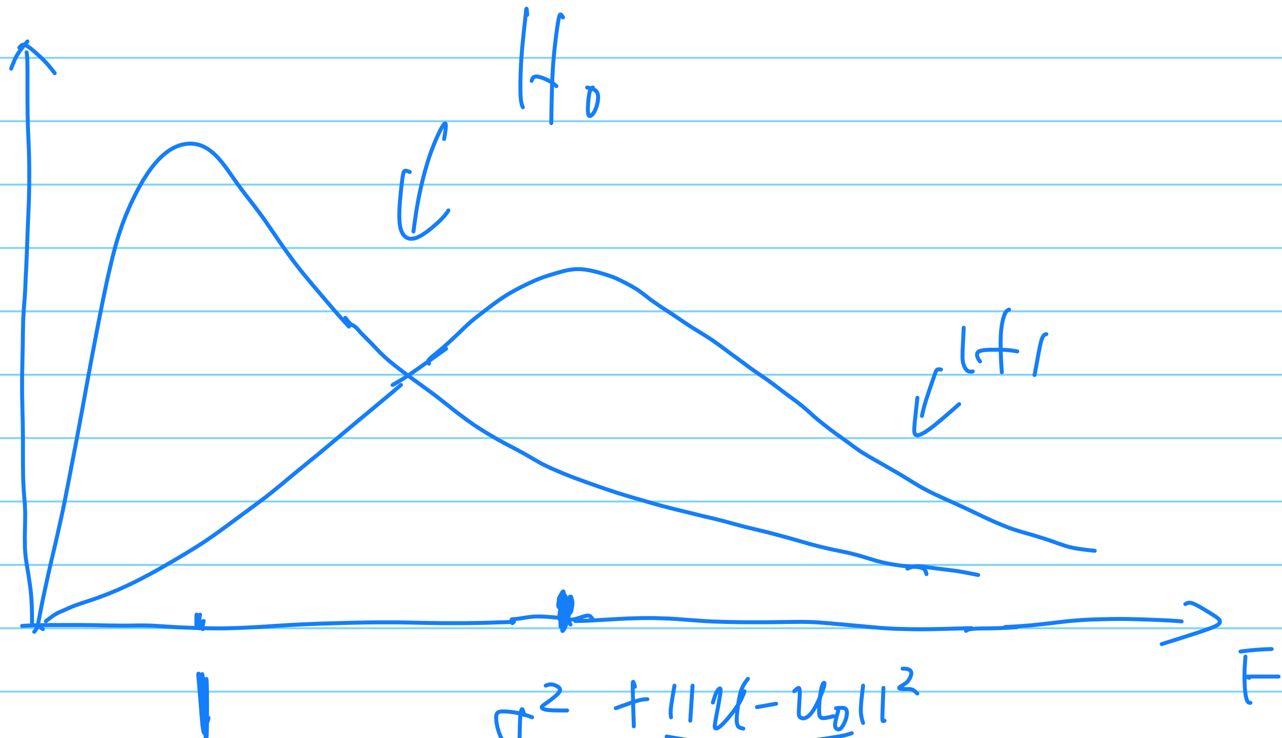
$$F = \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2/h}{s^2} = \frac{\mathbf{y}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}/h}{\mathbf{y}^T(\mathbf{I} - \mathbf{P}_{C(\mathbf{X})})\mathbf{y}/(n-k-1)} = \frac{SSH/h}{SSE/df_e}$$

$$\sim \begin{cases} F(h, n-k-1), & \text{under } H_0\text{; and} \\ F(h, n-k-1, \lambda_1), & \text{under } H_1, \end{cases}$$

where λ_1 is as given in the previous theorem.

$df_e = n - k - 1$





$$\frac{\sigma^2 + \frac{\|u-u_0\|^2}{n}}{\sigma^2}$$

$$\shortparallel$$

$$1 + \frac{\|u-u_0\|^2}{n \cdot \sigma^2}$$

$u_0 = P_{C(X_1)} u$, $\quad u = X_1\beta_1 + X_2\beta_2$

when $X_1 \perp X_2$, $\quad u - u_0 = X_2\beta_2$

Expression of F with SSE and SSR

It is worth noting that the numerator of this F test can be obtained as the difference in the SSE's under FM and RM divided by the difference in the dfE (degrees of freedom for error) for the two models. This is so because the Pythagorean Theorem yields

$$\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}_0\|^2 - \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \mathrm{SSE(RM)} - \mathrm{SSE(FM)}.$$

The difference in the dfE's is $(n - h - k - 1) - (n - k - 1) = h$. Therefore,

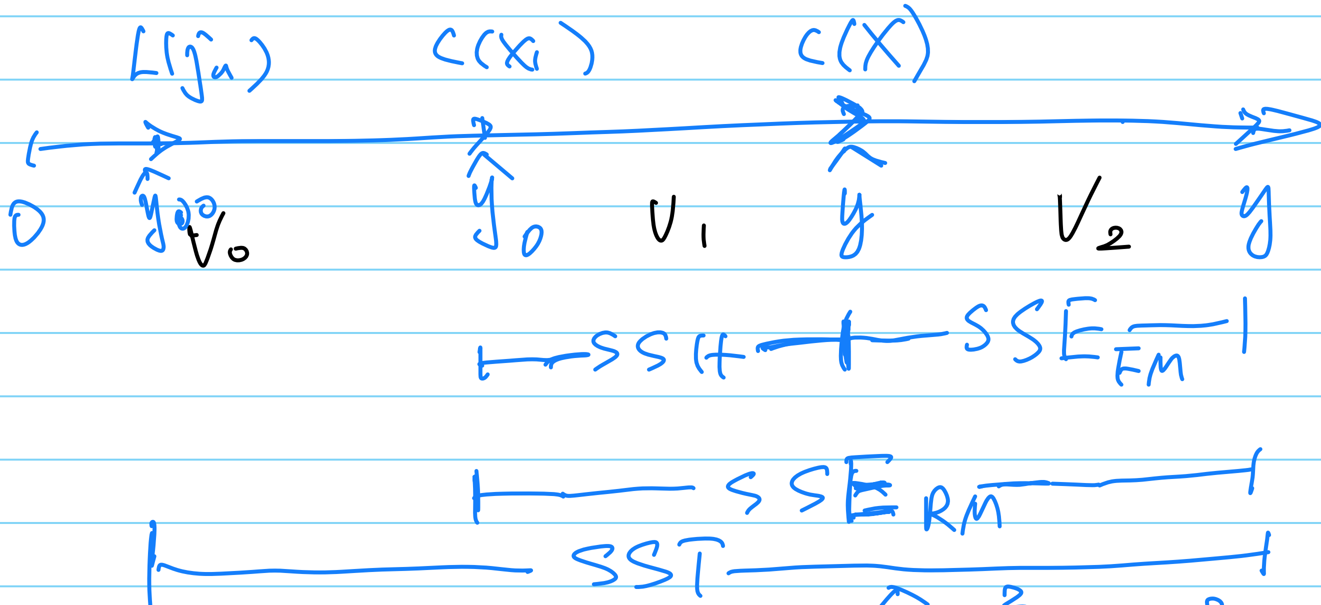
$$F = \frac{[\mathrm{SSE(RM)} - \mathrm{SSE(FM)}]/[\mathrm{dfE(RM)} - \mathrm{dfE(FM)}]}{\mathrm{SSE(FM)}/\mathrm{dfE(FM)}}.$$

In addition, because SSE $=$ SST $-$ SSR,

$$\begin{aligned}\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2 &= \mathrm{SSE(RM)} - \mathrm{SSE(FM)} \\ &= \mathrm{SST} - \mathrm{SSR(RM)} - [\mathrm{SST} - \mathrm{SSR(FM)}] \\ &= \mathrm{SSR(FM)} - \mathrm{SSR(RM)} \equiv \mathrm{SS}(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)\end{aligned}$$

which we denote as $\mathrm{SS}(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)$, and which is known as the *"extra" regression sum of squares due to $\boldsymbol{\beta}_2$ after accounting for $\boldsymbol{\beta}_1$.*

SSH in terms of SSE



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

$$SSE_{FM} = \|y - \hat{y}\|^2 = \|y\|^2 - \|\hat{y}\|^2$$

$$SSE_{RM} - SSE_{FM}$$

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

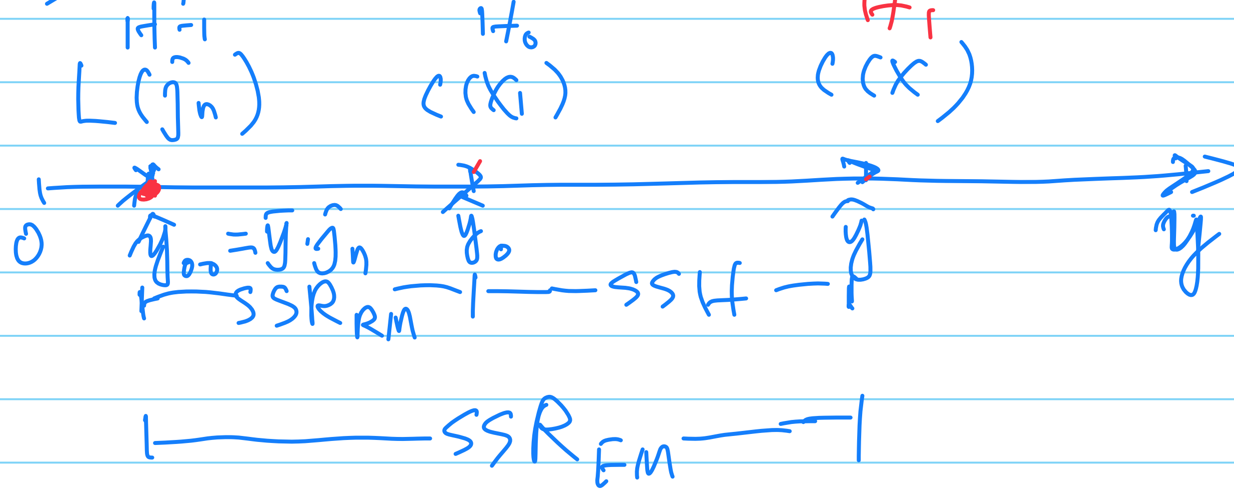
$$SSE_{RM} = SSH + SSE.$$

$$(SST = SSR + SSE)$$

$\uparrow$

$SSE \mid H_0: \mu \in L(j_n)$

SSH in terms of SSR



$$\hat{y}_{00} = P_{\hat{j}_n} y = \bar{y} \cdot \hat{j}_n$$

$$SSR_{RM} = \| \hat{y}_0 - \hat{y}_{00} \|^2$$

$$SSR_{FM} = \| \hat{y} - \hat{y}_{00} \|^2$$

$$SSH = SSR_{FM} - SSR_{RM}$$

ANOVA table

$L(j_n)$ $\quad C(X_1)$ $\quad C(X)$

$0 \quad \hat{y}_{00} \quad \hat{y}_0 \quad \hat{y} \quad y$

$\vdash SSH \dashv \vdash SSE_{FM} \dashv$

$\vdash SSR_{RM} \dashv \vdash SSE_{RM} \dashv$

$\vdash SST \dashv$

$L(j_n)$ $\quad L(j_n)^{\perp} C(X_1)$ $\quad C(X_1)^{\perp} C(X)$ $\quad C(X)^{\perp}$

Source	df	SS	MS
X_1	h_1	SSR_{RM}	SSR_{RM}/h_1
$SS(\beta_2 \mid \beta_1)$	h	SSH	SSH/h
error	$n-k-1$	SSE	$SSE/(n-k-1)$
sum	$n-1$	SST	

$h_1 = \text{rank}(X_1)$, $h = \text{rank}(X_2)$, $k = h_1 + h$

13

Overall Regression Test

14

An important special case of the test of $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ that we have just developed is when we partition $\boldsymbol{\beta}$ so that $\boldsymbol{\beta}_1$ contains just the intercept and when $\boldsymbol{\beta}_2$ contains all of the regression coefficients. That is, if we write the model as

$$\begin{aligned}\mathbf{y} &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e} \\ &= \beta_0\mathbf{j}_n + \underbrace{\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}}_{=\mathbf{X}_2} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}}_{=\boldsymbol{\beta}_2} + \mathbf{e}\end{aligned}$$

then our hypothesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ is equivalent to

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0,$$

The test of this hypothesis is called the **overall regression test** and occurs as a special case of the test of $\boldsymbol{\beta}_2 = \mathbf{0}$ that we've developed. Under H_0,

$$\hat{\mathbf{y}}_0 = p(\mathbf{y}|C(\mathbf{X}_1)) = p(\mathbf{y}|\mathcal{L}(\mathbf{j}_n)) = \bar{y}\,\mathbf{j}_n$$

and $h = k$, so the numerator of our F-test statistic becomes

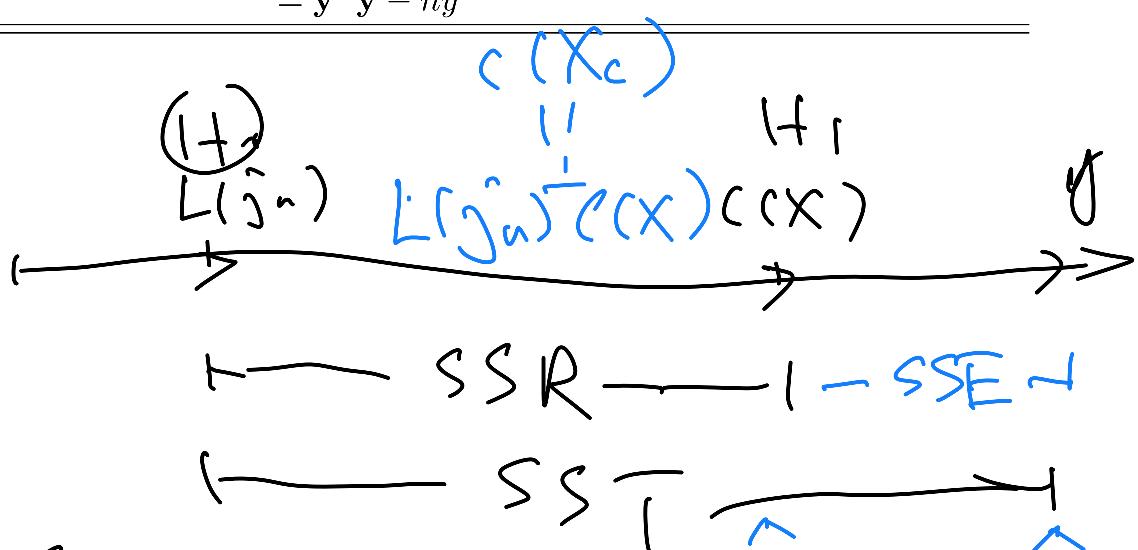
$$\begin{aligned}\frac{1}{k}\mathbf{y}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{\mathcal{L}(\mathbf{j}_n)})\mathbf{y} &= \frac{1}{k}(\mathbf{y}^T\mathbf{P}_{C(\mathbf{X})}\mathbf{y} - \mathbf{y}^T\mathbf{P}_{\mathcal{L}(\mathbf{j}_n)}\mathbf{y}) \\ &= \frac{1}{k}\{(\mathbf{P}_{C(\mathbf{X})}\mathbf{y})^T\mathbf{y} - \mathbf{y}^T\mathbf{P}^T_{\mathcal{L}(\mathbf{j}_n)}\underbrace{\mathbf{P}_{\mathcal{L}(\mathbf{j}_n)}\mathbf{y}}_{=\bar{y}\mathbf{j}_n}\} \\ &= \frac{1}{k}(\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y} - n\bar{y}^2) = SSR/k \equiv MSR\end{aligned}$$

Thus, the test statistic of overall regression is given by

$$F = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE}$$
$$\sim \begin{cases} F(k, n-k-1), & \text{under } H_0 : \beta_1 = \cdots = \beta_k = 0 \\ F(k, n-k-1, \frac{1}{2\sigma^2}\boldsymbol{\beta}_2^T \mathbf{X}_2^T \mathbf{P}_{\mathcal{L}(\mathbf{j}_n)^\perp} \mathbf{X}_2 \boldsymbol{\beta}_2), & \text{otherwise.} \end{cases}$$

The ANOVA table for this test is given below. This ANOVA table is typically part of the output of regression software (e.g., PROC REG in SAS).

Source of Variation	Sum of Squares	df	Mean Squares	F
Regression	$SSR = \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{y} - n\bar{y}^2$	k	$\frac{SSR}{k}$	$\frac{MSR}{MSE}$
Error	$SSE = \mathbf{y}^T(\mathbf{I} - \mathbf{P}_{C(\mathbf{X})})\mathbf{y}$	$n-k-1$	$\frac{SSE}{n-k-1}$	
Total (Corr.)	$SST = \mathbf{y}^T\mathbf{y} - n\bar{y}^2$			



$SSR_{RM} = 0$ $(\hat{y}_0 = \hat{y}_0)$

$SSR_{FM} = SSR$

16

$\underline{F \text{ test in terms of } R^2:}$

The F test statistics we have just developed can be written in terms of R^2, the coefficient of determination. This relationship is given by the following theorem.

Theorem: The F statistic for testing $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ in the full rank model $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$ (top of p. 138) can be written in terms of R^2 as

$$F = \frac{(R^2_{FM} - R^2_{RM})/h}{(1 - R^2_{FM})/(n-k-1)},$$

where R^2_{FM} corresponds to the full model $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$, and R^2_{RM} corresponds to the reduced model $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1^* + \mathbf{e}^*$.

$$\frac{SSH}{SST}$$

$$\frac{SSH}{SST} = \frac{SSR_{FM}}{SST} - \frac{SSR_{RM}}{SST}$$

$$= R^2_{FM} - R^2_{RM}$$

$$\frac{SSE_{FM}}{SST}$$

$$\frac{SSE_{FM}}{SST} = 1 - R^2_{FM}$$

$$\frac{SSH}{SSE} = \frac{R^2_{FM} - R^2_{RM}}{1 - R^2_{FM}}$$

Corollary: The F statistic for overall regression (for testing $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$) in the full rank model, $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + e_i$, $i = 1, \ldots, n$, $e_1, \ldots, e_n \overset{iid}{\sim} N(0, \sigma^2)$ can be written in terms of R^2, the coefficient of determination from this model as follows:

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}.$$

$$F = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE}$$

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

This is a non-decreasing transformation of R^2. $P(R^2 > R^{2\,(obs)}) = P(F > F^{obs})$

No need to think about the dist. of R^2.

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

$$F = \frac{SSR}{SSE} \cdot \frac{k}{(n-k-1)}$$

General Test

Motivation:

$$y = X_1 \beta_1 + X_2 \beta_2 + e$$
$$= (X_1, X_2) \cdot \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + e$$

$$H_0: \beta_2 = 0 \quad \text{vs} \quad H_1: \beta_2 \neq 0$$

$$(0, I_h) \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = 0$$

$$\beta_2 \in \mathbb{R}^h, \quad C = (0, I_h)$$

The hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$ is called the general linear hypothesis. Her
$\mathbf{C}$ is a $q \times (k+1)$ matrix of (known) coefficients with rank$(\mathbf{C}) = q$. W
will consider the slightly simpler case $H : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ (i.e., $\mathbf{t} = \mathbf{0}$) first.

Most of the questions that are typically asked about the coefficients of
linear model can be formulated as hypotheses that can be written in th
form $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, for some $\mathbf{C}$. For example, the hypothesis $H_0 : \boldsymbol{\beta}_2 =$
in the model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

can be written as

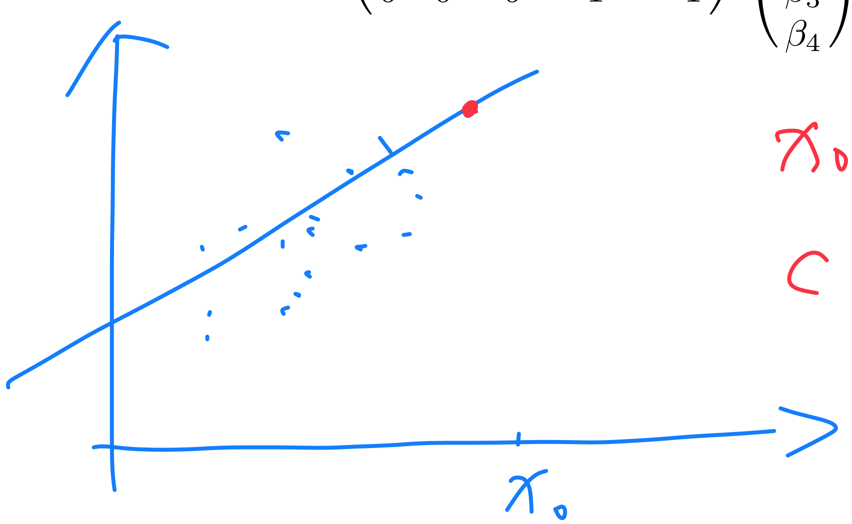
$$H_0 : \mathbf{C}\boldsymbol{\beta} = (\underbrace{\mathbf{0}}_{h\times(k+1-h)}, \mathbf{I}_h)\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix} = \boldsymbol{\beta}_2 = \mathbf{0}.$$

The test of overall regression can be written as

$$H_0 : \mathbf{C}\boldsymbol{\beta} = (\underbrace{\mathbf{0}}_{k\times 1}, \mathbf{I}_k)\left(\begin{array}{c}\beta_0\\ \begin{pmatrix}\beta_1\\ \vdots\\ \beta_k\end{pmatrix}\end{array}\right) = \begin{pmatrix}\beta_1\\ \vdots\\ \beta_k\end{pmatrix} = \mathbf{0}.$$

Hypotheses encompassed by $H: \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ are not limitted to ones in whic
certain regression coefficients are set equal to zero. Another example tha
can be handled is the hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k$. For exampl
suppose $k = 4$, then this hypothesis can be written as

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \begin{pmatrix}0 & 1 & -1 & 0 & 0\\ 0 & 0 & 1 & -1 & 0\\ 0 & 0 & 0 & 1 & -1\end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix} = \begin{pmatrix}\beta_1 - \beta_2\\ \beta_2 - \beta_3\\ \beta_3 - \beta_4\end{pmatrix} = \mathbf{0}.$$



$x_0'\beta$

$C = x_0'$

21

The test statistic for $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ is based on comparing $\mathbf{C}\hat{\boldsymbol{\beta}}$ to its null value $\mathbf{0}$, using a squared statistical distance (quadratic form) of the form

$$
\begin{aligned}
Q &= \{\mathbf{C}\hat{\boldsymbol{\beta}} - \underbrace{\mathrm{E}_0(\mathbf{C}\hat{\boldsymbol{\beta}})}_{=\mathbf{0}}\}^T \{\widehat{\mathrm{var}}_0(\mathbf{C}\hat{\boldsymbol{\beta}})\}^{-1} \{\mathbf{C}\hat{\boldsymbol{\beta}} - \mathrm{E}_0(\mathbf{C}\hat{\boldsymbol{\beta}})\} \\
&= (\mathbf{C}\hat{\boldsymbol{\beta}})^T \{\widehat{\mathrm{var}}_0(\mathbf{C}\hat{\boldsymbol{\beta}})\}^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}}).
\end{aligned}
$$

- Here, the 0 subscript is there to indicate that the expected value and variance are computed under H_0.

Recall that $\hat{\boldsymbol{\beta}} \sim N_{k+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$. Therefore,

$$
\mathbf{C}\hat{\boldsymbol{\beta}} \sim N_q(\mathbf{C}\boldsymbol{\beta}, \sigma^2 \mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T).
$$

Theorem: If $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ where $\mathbf{X}$ is $n \times (k+1)$ of full rank and $\mathbf{C}$ is $q \times (k+1)$ of rank $q \leq k+1$, then

(i) $\mathbf{C}\hat{\boldsymbol{\beta}} \sim N_q[\mathbf{C}\boldsymbol{\beta}, \sigma^2\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]$;
(ii) $(\mathbf{C}\hat{\boldsymbol{\beta}})^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}/\sigma^2 \sim \chi^2(q, \lambda)$, where

$$\lambda = (\mathbf{C}\boldsymbol{\beta})^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}\mathbf{C}\boldsymbol{\beta}/(2\sigma^2);$$

(iii) $\mathrm{SSE}/\sigma^2 \sim \chi^2(n-k-1)$; and
(iv) $(\mathbf{C}\hat{\boldsymbol{\beta}})^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}$ and SSE are independent.

$\hat{y}$ $\qquad$ $y - \hat{y}$

Theorem: If $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ where $\mathbf{X}$ is $n \times (k+1)$ of full rank and $\mathbf{C}$ is $q \times (k+1)$ of rank $q \leq k+1$, then

SSH

$$\begin{aligned} F &= \frac{(\mathbf{C}\hat{\boldsymbol{\beta}})^T\{\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\}^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}/q}{\mathrm{SSE}/(n-k-1)} \\ &= \frac{\mathrm{SSH}/q}{\mathrm{SSE}/(n-k-1)} \\ &\sim \begin{cases} F(q, n-k-1), & \text{if } H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0} \text{ is true;} \\ F(q, n-k-1, \lambda), & \text{if } H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0} \text{ is false,} \end{cases} \end{aligned}$$

where λ is as in the previous theorem.

Recall:

$X \sim N_p(\mu, \Sigma)$

$$(X-\mu_0)' \Sigma^{-1} (X-\mu_0) \sim \chi^2(p, \lambda)$$

where $\lambda = \frac{1}{2}(\mu-\mu_0)'\Sigma^{-1}(\mu-\mu_0)$

Pf: 1) $X-\mu_0 \sim N_p(\mu-\mu_0, \Sigma)$ ✓

$$A = \Sigma^{-1}, \quad A\Sigma = I_p \Leftarrow$$

2) $\Sigma^{-\frac{1}{2}}(X-\mu_0) \sim N_p(\mu-\mu_0, I_p)$

$$(X-\mu_0)'\Sigma^{-1}(X-\mu_0)$$

$$= \quad \| \Sigma^{-\frac{1}{2}}(X-\mu_0) \|^2$$

$$\sim \chi^2\left(p, \frac{1}{2} \| \Sigma^{-\frac{1}{2}}(\mu-\mu_0)\|^2\right)$$

$$C\hat{\beta} \sim N_q(C\beta,\ C(X'X)^{-1}C')$$

$$H_0: C\beta = 0$$

$$C\hat{\beta} \sim N_q(0, \sigma^2 C(X'X)^{-1}C')$$

$$\frac{SSH}{\sigma^2} = (C\hat{\beta})'[C(X'X)^{-1}C']^{-1}(C\hat{\beta})/\sigma^2$$

$$= \hat{\beta}'C'[C(X'X)^{-1}C']^{-1}C\hat{\beta}/\sigma^2$$

$$\sim \chi^2_q$$

$$H_1: C\beta \neq 0$$

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

$$\frac{SSH}{\sigma^2} = (C\hat{\beta})'A(C\hat{\beta})$$

where $A = [C(X'X)^{-1}C']^{-1}\frac{1}{\sigma^2}$

$\frac{SSH}{\sigma^2} \sim \chi^2(q, \lambda_1)$, where

$$\lambda_1 = (C\beta)'A\,C\beta/2$$

25

$$\frac{SSE}{\sigma^2} \sim \chi^2_{n-k-1}$$

SSE indep SSH

Under H_0

$$F = \frac{\dfrac{(C\hat{\beta})'\left[C(X'X)^{-1}C'\right]^{-1}(\hat{\beta} / \cancel{\sigma^2}}{q}}{\dfrac{SSE / \cancel{\sigma^2}}{n-k-1}}$$

$$\sim F_{q,\, n-k-1}$$

The general linear hypothesis is a test of nested models

Theorem: The F test for the general linear hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ is a full-and-reduced-model test.

Under H_0,

$$
\begin{aligned}
\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad &\text{and} \qquad \mathbf{C}\boldsymbol{\beta} = \mathbf{0} \\
\Rightarrow \quad \mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} &= \mathbf{0} \\
\Rightarrow \quad \mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\mu} &= \mathbf{0} \\
\Rightarrow \quad \mathbf{T}^T\boldsymbol{\mu} &= \mathbf{0} \quad \text{where} \quad \mathbf{T} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T.
\end{aligned}
$$

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

$\beta = X^- \mu$

$C\beta = C \cdot X^- X\beta$

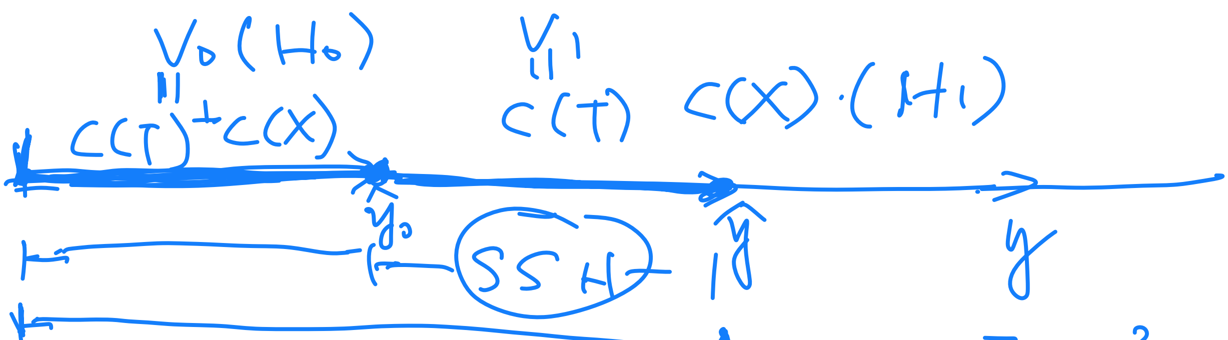
$= (C X^-)^T$

That is, under H_0, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \in C(\mathbf{X}) = V$ and $\boldsymbol{\mu} \perp C(\mathbf{T})$, or

$$\boldsymbol{\mu} \in [C(\mathbf{T})^\perp \cap C(\mathbf{X})] = V_0$$

where $V_0 = C(\mathbf{T})^\perp \cap C(\mathbf{X})$ is the orthogonal complement of $C(\mathbf{T})$ with respect to $C(\mathbf{X})$.

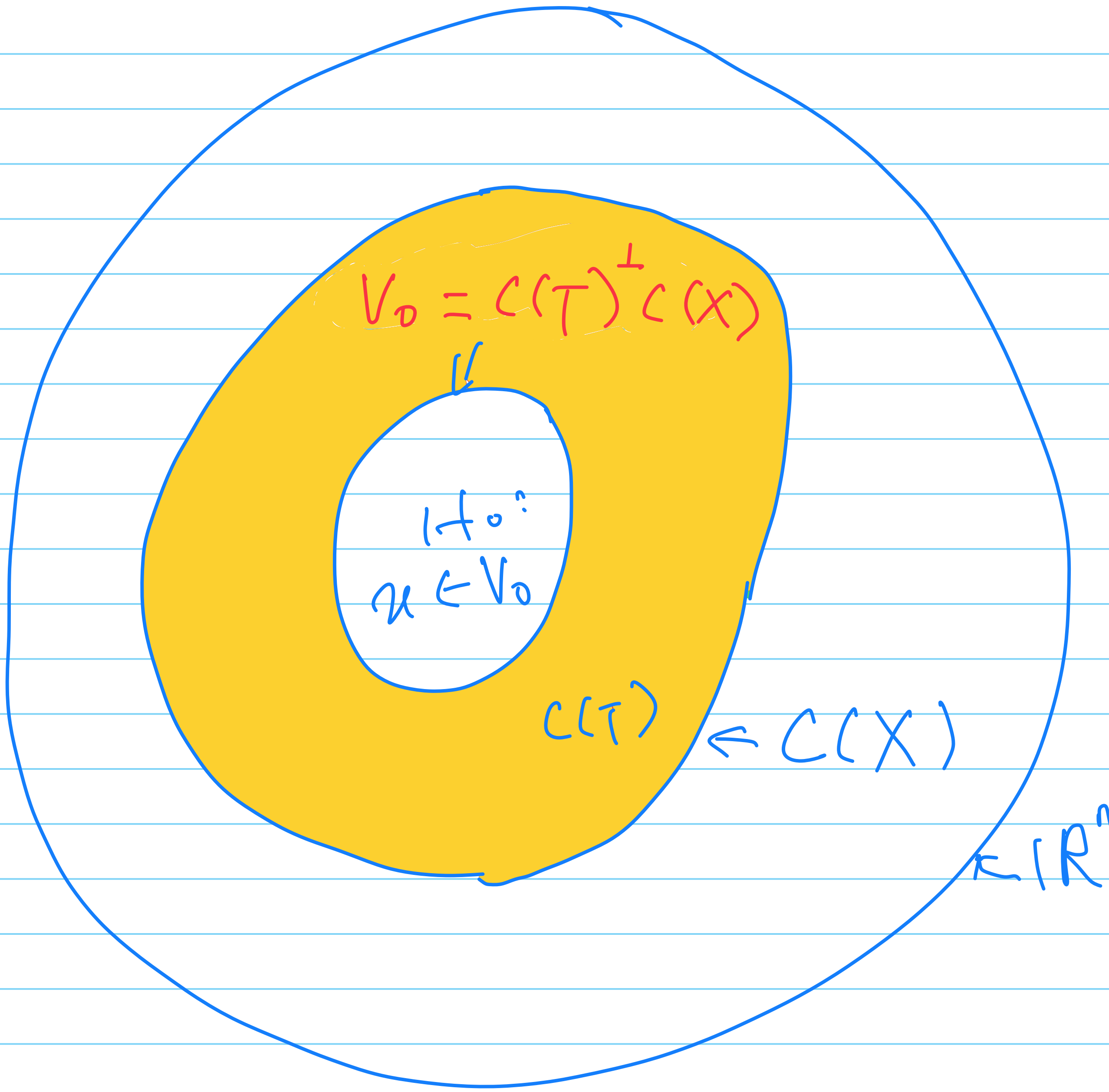
Thus, under $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, $\boldsymbol{\mu} \in V_0 \subset V = C(\mathbf{X})$, and under $H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{0}$, $\boldsymbol{\mu} \in V$ but $\boldsymbol{\mu} \notin V_0$. That is, these hypotheses correspond to nested models. It just remains to establish that the F test for these nested models is the F test for the general linear hypothesis



$SSH = \| [P_{C(X)} - (P_{C(X)} - P_{C(T)})] y \|^2$

$= \| P_{C(T)} y \|^2$

28



$$C(T) = V_0^\perp \cdot C(X)$$

$$T = X(X'X)^{-1}C'$$

$$\mathrm{rank}(T) = \mathrm{rank}(T'T)$$

$$= \mathrm{rank}\left(C(X'X)^{-1}X' \cdot X(X'X)^{-1}C'\right)$$

$$= \mathrm{rank}\left(C(X'X)^{-1}C'\right)$$

$$= \mathrm{rank}\left(C(X'X)^{-\frac{1}{2}}\right)$$

$$= \mathrm{rank}(C) = q$$

$T = X(X'X)^{-1}C'$

Thus the full vs. reduced model F statistic becomes

$$F = \frac{\mathbf{y}^T[\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{V_0}]\mathbf{y}/q}{\text{SSE}/(n-k-1)} = \frac{\mathbf{y}^T[\mathbf{P}_{C(\mathbf{X})} - (\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{T})})]\mathbf{y}/q}{\text{SSE}/(n-k-1)}$$
$$= \frac{\mathbf{y}^T\mathbf{P}_{C(\mathbf{T})}\mathbf{y}/q}{\text{SSE}/(n-k-1)}$$

where

$$\begin{aligned}
\mathbf{y}^T\mathbf{P}_{C(\mathbf{T})}\mathbf{y} &= \mathbf{y}^T\mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y} \\
&= \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\{\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\}^{-1}\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= \underbrace{\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}}_{=\hat{\boldsymbol\beta}^T}\mathbf{C}^T\{\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\}^{-1}\mathbf{C}\underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}_{=\hat{\boldsymbol\beta}} \\
&= \hat{\boldsymbol\beta}^T\mathbf{C}^T\{\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\}^{-1}\mathbf{C}\hat{\boldsymbol\beta}
\end{aligned}$$

$P_{V_0} = P_{C(X)} - P_{C(T)}$

$V_0 = C(T)^{\perp} \cdot C(X)$

$P_{C(X)} - P_{V_0} = P_{C(T)}$

Example

In an effort to obtain maximum yield in a chemical reaction, the values of the following variables were chosen by the experimenter:

$$x_1 = \text{temperature } (^\circ\text{C})$$
$$x_2 = \text{concentration of a reagent } (\%)$$
$$x_3 = \text{time of reaction (hours)}$$

Two different response variables were observed:

$$y_1 = \text{percent of unchanged starting material}$$
$$y_2 = \text{percent converted to the desired product}$$

TABLE 7.4 Chemical Reaction Data

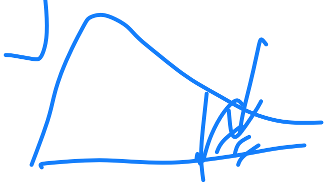
y_1	y_2	x_1	x_2	x_3
41.5	45.9	162	23	3
33.8	53.3	162	23	8
27.7	57.5	162	30	5
21.7	58.8	162	30	8
19.9	60.6	172	25	5
15.0	58.0	172	25	8
12.2	58.6	172	30	5
4.3	52.4	172	30	8
19.3	56.9	167	27.5	6.5
6.4	55.4	177	27.5	6.5
37.6	46.9	157	27.5	6.5
18.0	57.3	167	32.5	6.5
26.3	55.0	167	22.5	6.5
9.9	58.9	167	27.5	9.5
25.0	50.3	167	27.5	3.5
14.1	61.1	177	20	6.5
15.2	62.9	177	20	6.5
15.9	60.0	160	34	7.5
19.6	60.6	160	34	7.5

Example 8.4.1b. Consider the dependent variable y_1 in the chemical reaction data in Table 7.4. For the model $y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, we test $H_0 : 2\beta_1 = 2\beta_2 = \beta_3$ using (8.27) in Theorem 8.4b. To express H_0 in the form $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, the matrix $\mathbf{C}$ becomes

$$C\beta = \begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 2 & -1 \end{pmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

and we obtain

0.101



$$\mathbf{C}\hat{\boldsymbol{\beta}} = \begin{pmatrix} -.1214 \\ -.6118 \end{pmatrix},$$

2.6776

$$\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}' = \begin{pmatrix} .003366 & -.006943 \\ -.006943 & .044974 \end{pmatrix},$$

$$F = \frac{\begin{pmatrix} -.1214 \\ -.6118 \end{pmatrix}' \begin{pmatrix} .003366 & -.006943 \\ -.006943 & .044974 \end{pmatrix}^{-1} \begin{pmatrix} -.1214 \\ -.6118 \end{pmatrix} / 2}{5.3449}$$

↑ MSE (s^2)

$$= \frac{28.62301/2}{5.3449} = 2.6776,$$

which has $p = .101$. □

H_0: $y_1 = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \frac{1}{2}\beta_1 x_3 + e$

$= \beta_0 + \beta_1 (x_1 + x_2 + \frac{1}{2} x_3) + e$

$k = 1$

H_1 $k = 3$, $F = \dfrac{(SSE_0 - SSE_1)/2}{SSE_1/(n-3)}$

Example: One-way ANOVA

$$y_{ij} \sim N(\mu_i, \sigma^2)$$

$$i = 1, \dots, k, \quad j = 1, \dots, n_i$$

We want to test:

$H_0: \mu_1 = \cdots = \mu_k$ ($k-1$ equations)

$H_1: \exists\, i, i'$, s.t. $\mu_i \neq \mu_{i'}$

1) Contruct the full linear model

$$Y = X\beta + \varepsilon$$

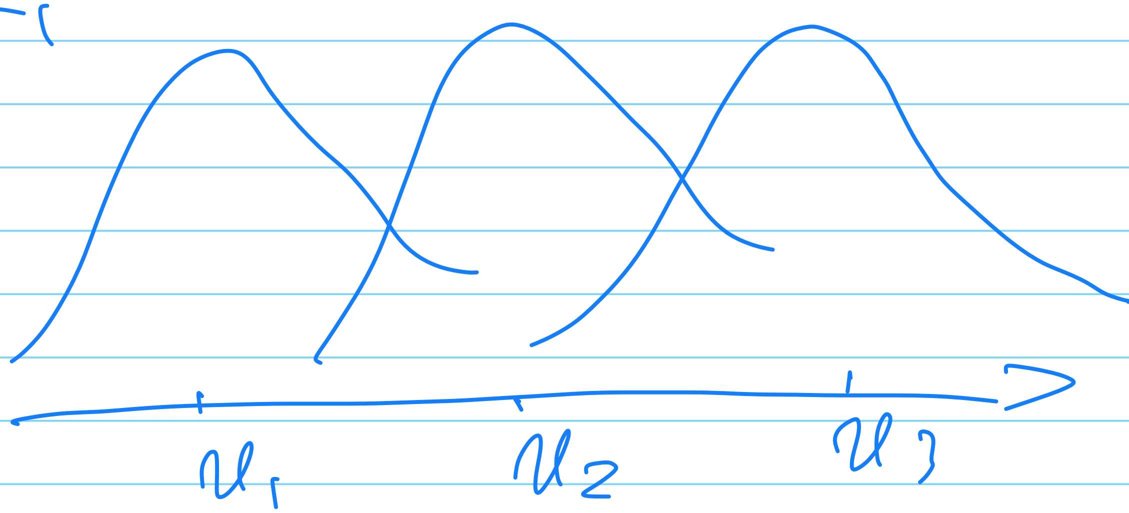
$$\begin{bmatrix} y_{11} \\ \vdots \\ \vdots \\ y_{kn_k} \end{bmatrix} = \begin{array}{c} \begin{bmatrix} 1 & & 0 \\ \vdots & & 0 \\ 1 & & 0 \\ 0 & \cdots & \vdots \\ 0 & & 1 \\ \vdots & & \vdots \\ 0 & & 1 \end{bmatrix} \\ \uparrow \ \ \uparrow \qquad \uparrow \\ x_1 \ \ x_2 \qquad x_k \end{array} \begin{array}{c} \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_k \end{bmatrix} \\ \uparrow \\ \beta \end{array} + \varepsilon$$

Remark: No intercept

$$\varepsilon \sim N_n(0, \sigma^2 I_n)$$

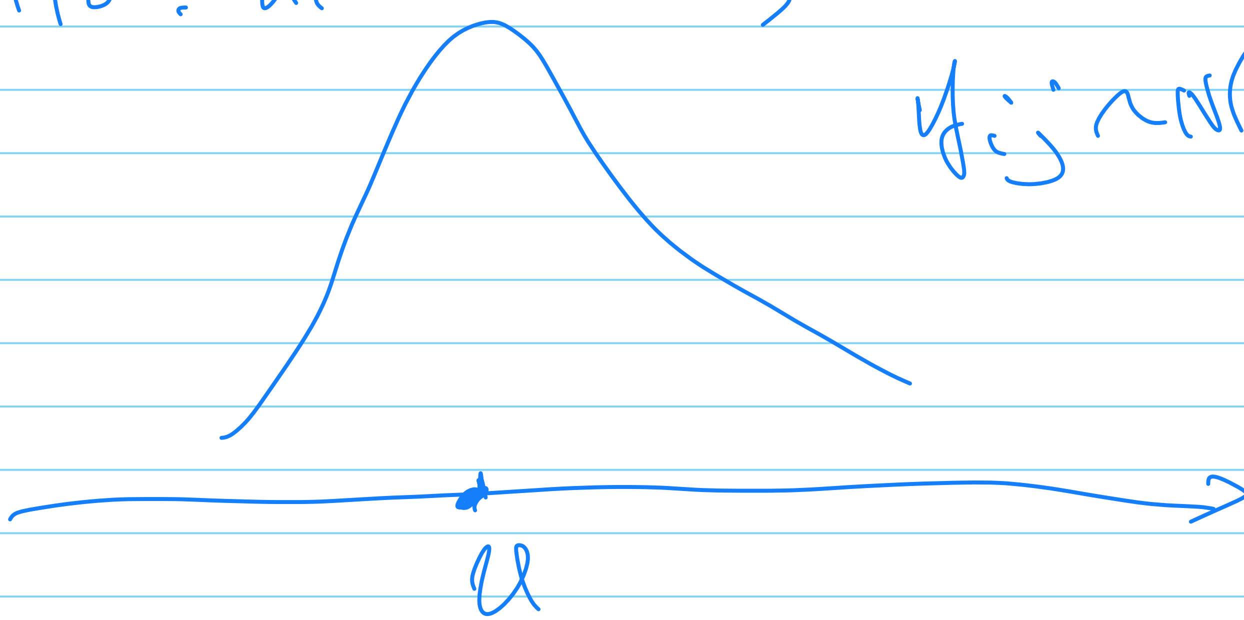
$$n = n_1 + \cdots + n_k$$

H_1



$H_0 : \mu_1 = \mu_2 = \mu_3$

$y_{ij} \sim N(\mu, \sigma^2)$



2) The reduced model:

$$y = X\beta + \varepsilon = \mathbf{1}_n \cdot \mu + \varepsilon$$

$$n = n_1 + \cdots + n_k$$

3.1) Find $\hat{y}_0$ & $\hat{y}_1$ for the reduced & full models respectively

3.2) Find SSR_0, SSE_0, SSR_1, SSE_1 of the reduced & full model.

4) Find the statistic F

$$F = \frac{(SSE_0 - SSE_1)/(k-1)}{SSE_1/(n-k)}$$

5) Derive the F-statistic with the formula for the general test applied to the full model.

$$H_0: C\beta = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} = 0$$

36

6) Show that the two test statistic (given in 4) & 5) are the same

The case $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$ where $\mathbf{t} \neq \mathbf{0}$:

Extension to this case is straightforward. The only requirement is that the system of equations $\mathbf{C}\boldsymbol{\beta} = \mathbf{t}$ be consistent, which is ensured by $\mathbf{C}$ having full row rank q.

Then the F test statistic for $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$ is given by

$$F = \frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{t})^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{t})/q}{\text{SSE}/(n-k-1)} \sim \begin{cases} F(q, n-k-1), & \text{under } H_0 \\ F(q, n-k-1, \lambda), & \text{otherwise,} \end{cases}$$

where $\lambda = (\mathbf{C}\boldsymbol{\beta} - \mathbf{t})^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}(\mathbf{C}\boldsymbol{\beta} - \mathbf{t})/(2\sigma^2)$.

Some Specific Tests

Tests on β_j and on $\mathbf{a}^T\boldsymbol{\beta}$:

Tests of $H_0 : \beta_j = 0$ or $H_0 : \mathbf{a}^T\boldsymbol{\beta} = 0$ occur as special cases of the tests we have already considered. To test $H_0 : \mathbf{a}^T\boldsymbol{\beta} = 0$, we use $\mathbf{a}^T$ in place of $\mathbf{C}$ in our test of the general linear hypothesis $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$. In this case $q = 1$ and the test statistic becomes

$$F = \frac{(\mathbf{a}^T\hat{\boldsymbol{\beta}})^T[\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}]^{-1}\mathbf{a}^T\hat{\boldsymbol{\beta}}}{\mathrm{SSE}/(n-k-1)} = \frac{(\mathbf{a}^T\hat{\boldsymbol{\beta}})^2}{s^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}$$
$$\sim F(1, n-k-1) \quad \text{under } H_0 : \mathbf{a}^T\boldsymbol{\beta} = 0.$$

Note that since $t^2(\nu) = F(1,\nu)$, an equivalent test of $H_0 : \mathbf{a}^T\boldsymbol{\beta} = 0$ is given by the t-test with test statistic

$V(AX) = A\,V(X)\cdot A'$

$$t = \frac{\mathbf{a}^T\hat{\boldsymbol{\beta}}}{s\sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}} \sim t(n-k-1) \quad \text{under } H_0.$$

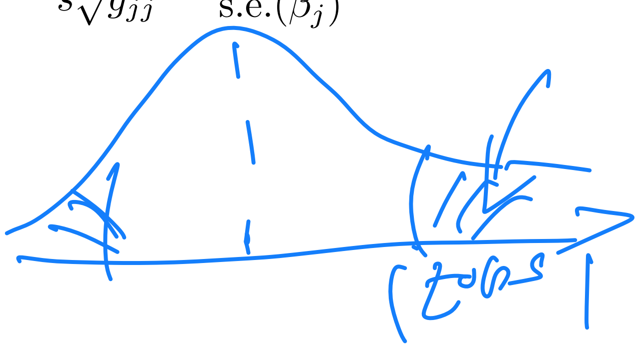
$$a = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow j\text{th}$$

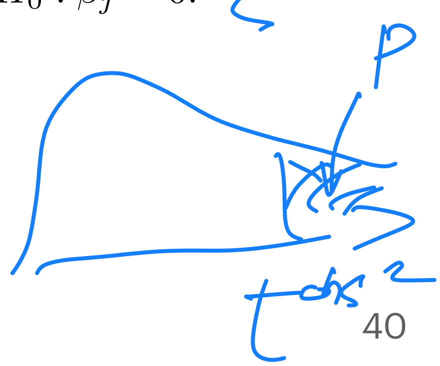
The test statistic for this hypothesis simplifies from that given above to yield

$$F = \frac{\hat{\beta}_j^2}{s^2 g_{jj}} \sim F(1, n-k-1) \quad \text{under } H_0 : \beta_j = 0,$$

where g_{jj} is the j^{th} diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$. Equivalently, we could use the t test statistic

$$t = \frac{\hat{\beta}_j}{s\sqrt{g_{jj}}} = \frac{\hat{\beta}_j}{\text{s.e.}(\hat{\beta}_j)} \sim t(n-k-1) \quad \text{under } H_0 : \beta_j = 0.$$





$\hat{\beta}_j$	$se(\hat{\beta}_j)$	p-value

Confidence and Prediction Intervals

Confidence Region for β:

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

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/(k+1)}{s^2} \sim F(k+1, n-k-1)$$

From this distributional result, we can make the probability statement,

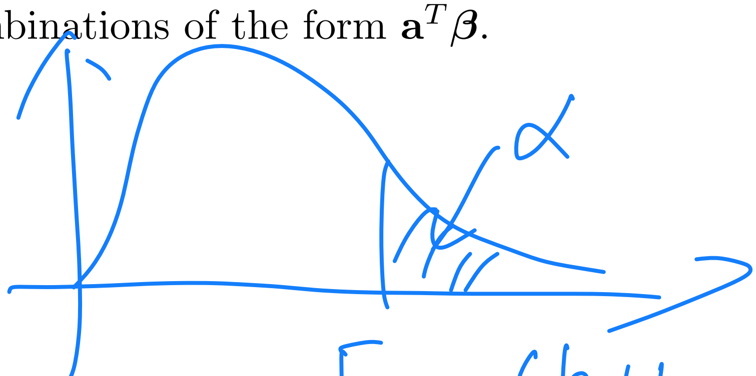
$$\Pr\left\{ \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{s^2 (k+1)} \leq F_{1-\alpha}(k+1, n-k-1) \right\} = 1 - \alpha.$$

Therefore, the set of all vectors $\boldsymbol{\beta}$ that satisfy

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq (k+1) s^2 F_{1-\alpha}(k+1, n-k-1)$$

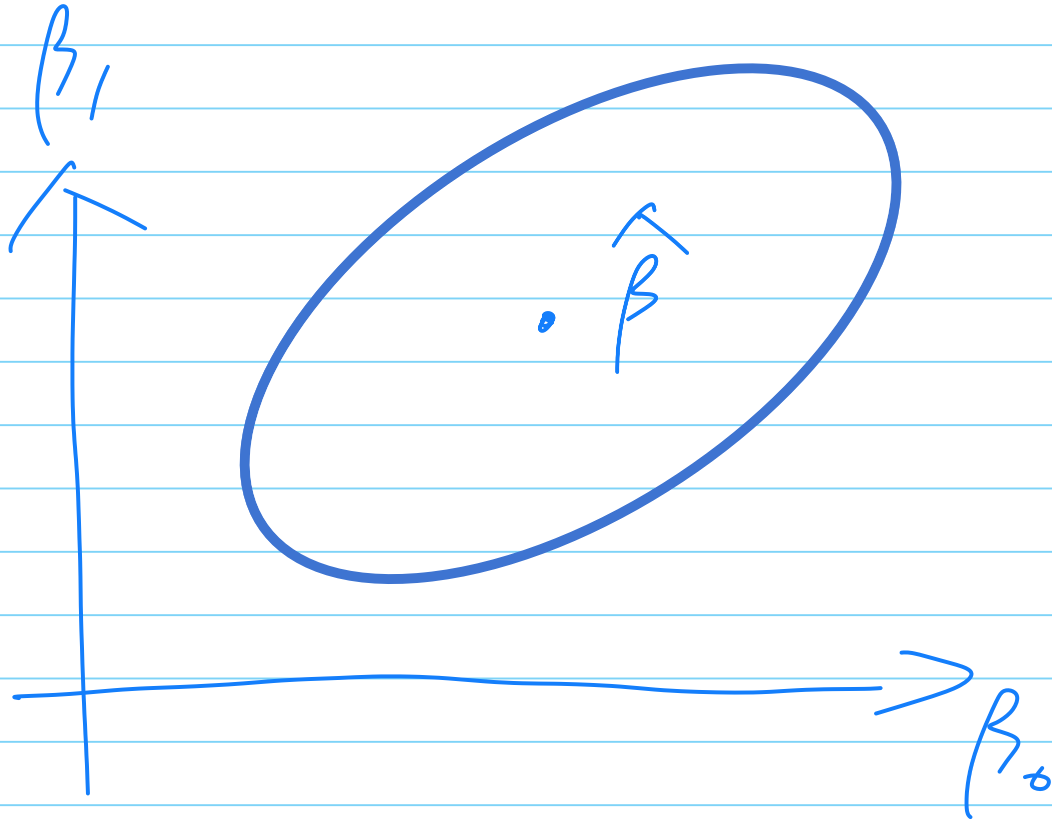
forms a $100(1-\alpha)\%$ confidence region for $\boldsymbol{\beta}$.

- Such a region is an ellipse, and is only easy to draw and make easy interpretation of for $k = 1$ (e.g., simple linear regression).
- If one can't plot the region and then plot a point to see whether its in or out of the region (i.e., for $k > 1$) then this region isn't any more informative than the test of $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$. To decide whether $\boldsymbol{\beta}_0$ is in the region, we essentially have to perform the test!
- More useful are confidence intervals for the individual β_j's and for linear combinations of the form $\mathbf{a}^T \boldsymbol{\beta}$.



$F_{1-\alpha}(k+1, n-k-1)$

43


β_1
$\hat{\beta}$
β_0

Confidence Interval for $\mathbf{a}^T\boldsymbol{\beta}$:

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

$$\frac{(\mathbf{a}^T\hat{\boldsymbol{\beta}} - \mathbf{a}^T\boldsymbol{\beta})^2}{s^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}} \sim F(1, n-k-1)$$

which implies

$$\frac{(\mathbf{a}^T\hat{\boldsymbol{\beta}} - \mathbf{a}^T\boldsymbol{\beta})}{s\sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}} \sim t(n-k-1).$$

From this distributional result, we can make the probability statement,

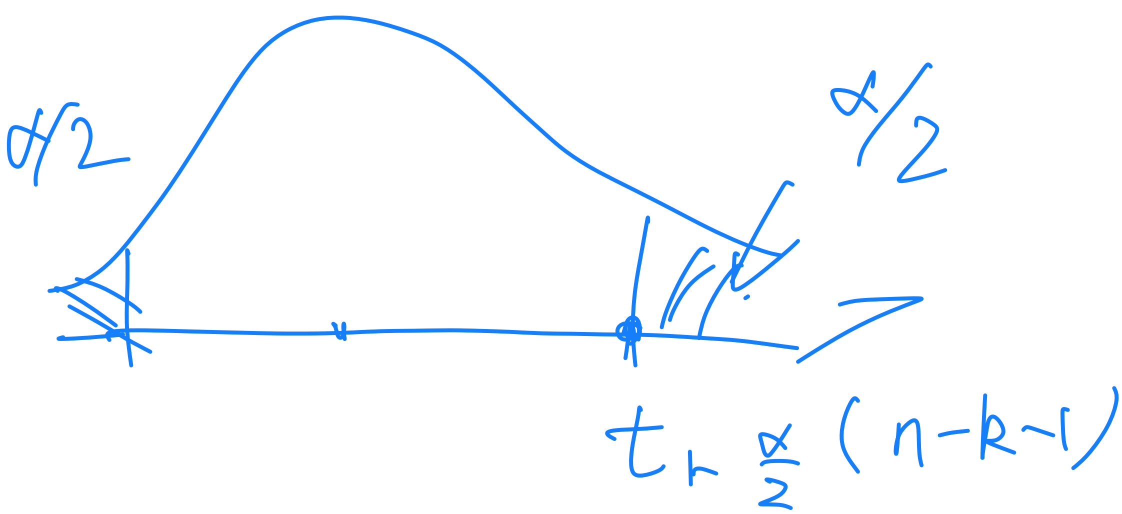
$$\Pr\left\{\underbrace{t_{\alpha/2}(n-k-1)}_{-t_{1-\alpha/2}(n-k-1)} \le \frac{(\mathbf{a}^T\hat{\boldsymbol{\beta}} - \mathbf{a}^T\boldsymbol{\beta})}{s\sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}} \le t_{1-\alpha/2}(n-k-1)\right\} = 1-\alpha.$$

Rearranging this inequality so that $\mathbf{a}^T\boldsymbol{\beta}$ falls in the middle, we get

$$\Pr\Big\{\mathbf{a}^T\hat{\boldsymbol{\beta}} - t_{1-\alpha/2}(n-k-1)s\sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}} \le \mathbf{a}^T\boldsymbol{\beta}$$
$$\le \mathbf{a}^T\hat{\boldsymbol{\beta}} + t_{1-\alpha/2}(n-k-1)s\sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}\Big\} = 1-\alpha.$$

Therefore, a $100(1-\alpha)\%$ CI for $\mathbf{a}^T\boldsymbol{\beta}$ is given by

$$\mathbf{a}^T\hat{\boldsymbol{\beta}} \pm t_{1-\alpha/2}(n-k-1)s\sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}.$$



45

Confidence Interval for β_j:

A special case of this interval occurs when $\mathbf{a} = (0, \ldots, 0, 1, 0, \ldots, 0)^T$, where the 1 is in the $j+1$th position. In this case $\mathbf{a}^T\boldsymbol{\beta} = \beta_j$, $\mathbf{a}^T\hat{\boldsymbol{\beta}} = \hat{\beta}_j$, and $\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a} = \{(\mathbf{X}^T\mathbf{X})^{-1}\}_{jj} \equiv g_{jj}$. The confidence interval for β_j is then given by

$$\hat{\beta}_j \pm t_{1-\alpha/2}(n-k-1)s\sqrt{g_{jj}}.$$

46

Confidence Interval for $\mathrm{E}(y)$:

Let $\mathbf{x}_0 = (1, x_{01}, x_{02}, \ldots, x_{0k})^T$ denote a particular choice of the vector of explanatory variables $\mathbf{x} = (1, x_1, x_2, \ldots, x_k)^T$ and let y_0 denote the corresponding response.

We assume that the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ applies to $(y_0, \mathbf{x}_0)$ as well. This may be because $(y_0, \mathbf{x}_0)$ were in the original sample to which the model was fit (i.e., $\mathbf{x}_0^T$ is a row of $\mathbf{X}$), or because we believe that $(y_0, \mathbf{x}_0)$ will behave similarly to the data $(\mathbf{y}, \mathbf{X})$ in the sample. Then

$$y_0 = \mathbf{x}_0^T\boldsymbol{\beta} + e_0, \quad e_0 \sim N(0, \sigma^2)$$

where $\boldsymbol{\beta}$ and σ^2 are the same parameters in the fitted model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$.

Suppose we wish to find a CI for

$a = x_0$

$$\mathrm{E}(y_0) = \mathbf{x}_0^T\boldsymbol{\beta}.$$

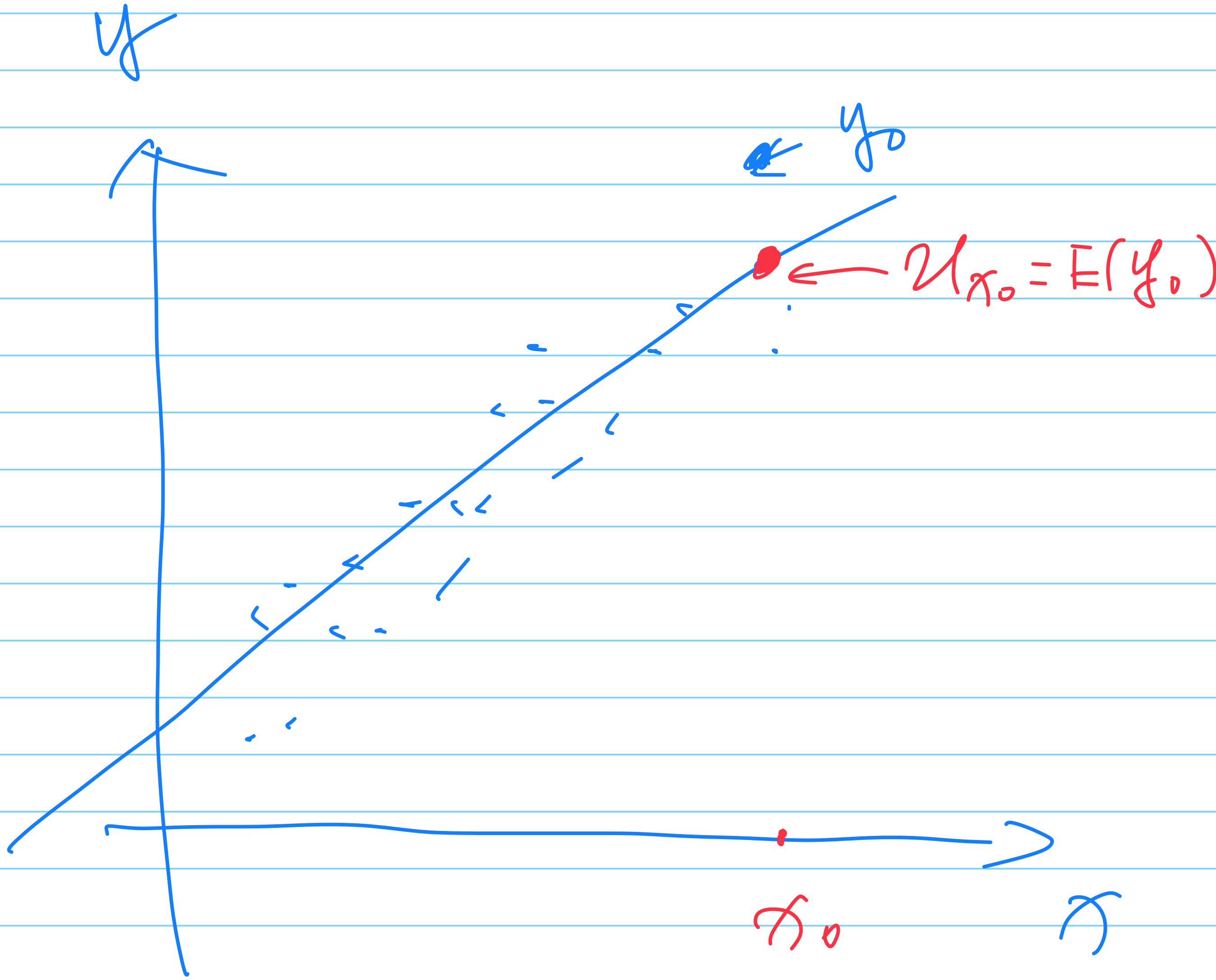
This quantity is of the form $\mathbf{a}^T\boldsymbol{\beta}$ where $\mathbf{a} = \mathbf{x}_0$, so the BLUE of $\mathrm{E}(y_0)$ is $\mathbf{x}_0^T\hat{\boldsymbol{\beta}}$ and a $100(1-\alpha)\%$ CI for $\mathrm{E}(y_0)$ is given by

$$\mathbf{x}_0^T\hat{\boldsymbol{\beta}} \pm t_{1-\alpha/2}(n-k-1)s\sqrt{\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}.$$

- This confidence interval holds for a particular value $\mathbf{x}_0^T\boldsymbol{\beta}$. Sometimes, it is of interest to form simultaneous confidence intervals around each and every point $\mathbf{x}_0^T\boldsymbol{\beta}$ for all $\mathbf{x}_0$ in the range of $\mathbf{x}$. That is, we sometimes desire a simultaneous confidence band for the entire regression line (or plane, for $k > 1$). The confidence interval given above, if plotted for each value of $\mathbf{x}_0$, does not give such a simultaneous band; instead it gives a "point-wise" band. For discussion of simultaneous intervals see §8.6.7 of our text.

- The confidence interval given above is for $\mathrm{E}(y_0)$, **not** for y_0 itself. $\mathrm{E}(y_0)$ is a parameter, y_0 is a random variable. Therefore, we can't *estimate* y_0 or form a confidence interval for it. However, we can predict its value, and an interval around that prediction that quantifies the uncertainty associated with that prediction is called a prediction interval.

C.I. for E(y)



$$\mu_{x_0} = E(y_0) = \underline{x_0'\beta}$$

a unknown value.
(fixed)

$$E(y_0 - (x_0'\hat{\beta})) = x_0'\beta - x_0'\beta = 0$$

Prediction Interval for an Unobserved y-value:

For an unobserved value y_0 with known explanatory vector $\mathbf{x}_0$ assumed to follow our linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, we predict y_0 by

$$\hat{y}_0 = \mathbf{x}_0^T \hat{\boldsymbol{\beta}}.$$

To form a CI for the estimator $\mathbf{x}_0^T \hat{\boldsymbol{\beta}}$ of $\mathrm{E}(y_0)$ we examine the variance of the error of estimation:

$$\mathrm{var}\{\mathrm{E}(y_0) - \mathbf{x}_0^T \hat{\boldsymbol{\beta}}\} = \mathrm{var}(\mathbf{x}_0^T \hat{\boldsymbol{\beta}}).$$

In contrast, to form a PI for the predictor $\mathbf{x}_0^T \hat{\boldsymbol{\beta}}$ of y_0, we examine the variance of the error of prediction:

$$\begin{aligned}
\mathrm{var}(y_0 - \mathbf{x}_0^T \hat{\boldsymbol{\beta}}) &= \mathrm{var}(y_0) + \mathrm{var}(\mathbf{x}_0^T \hat{\boldsymbol{\beta}}) - \underbrace{2\,\mathrm{cov}(y_0, \mathbf{x}_0^T \hat{\boldsymbol{\beta}})}_{0} \\
&= \mathrm{var}(\mathbf{x}_0^T \boldsymbol{\beta} + e_0) + \mathrm{var}(\mathbf{x}_0^T \hat{\boldsymbol{\beta}}) \\
&= \mathrm{var}(e_0) + \mathrm{var}(\mathbf{x}_0^T \hat{\boldsymbol{\beta}}) = \sigma^2 + \sigma^2 \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0.
\end{aligned}$$

r.v. y_0 r.v $y_1, \ldots, y_n$

$\downarrow 0$ as $n \uparrow +\infty$

It's not hard to show that

$$\frac{y_0 - \hat{y}_0}{s\sqrt{1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}} \sim t(n - k - 1),$$

therefore

$$\mathrm{Pr}\left\{ -t_{1-\alpha/2}(n-k-1) \le \frac{y_0 - \hat{y}_0}{s\sqrt{1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}} \le t_{1-\alpha/2}(n-k-1) \right\} = 1 - \alpha.$$

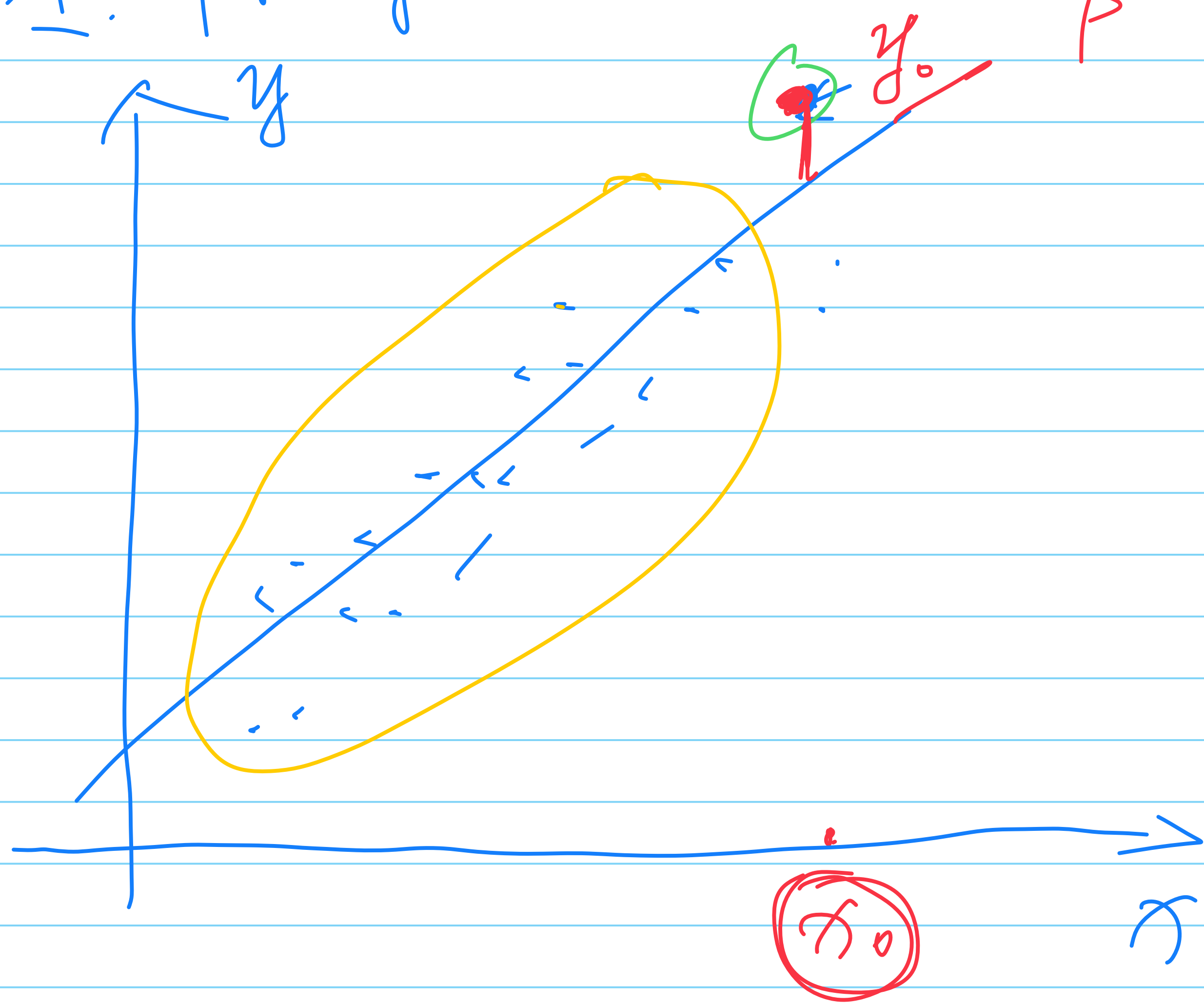
Rearranging,

$$\begin{aligned}
\mathrm{Pr}\Big\{ &\hat{y}_0 - t_{1-\alpha/2}(n-k-1) s \sqrt{1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0} \le y_0 \\
&\le \hat{y}_0 + t_{1-\alpha/2}(n-k-1) s \sqrt{1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0} \Big\} = 1 - \alpha.
\end{aligned}$$

Therefore, a $100(1-\alpha)\%$ prediction interval for y_0 is given by

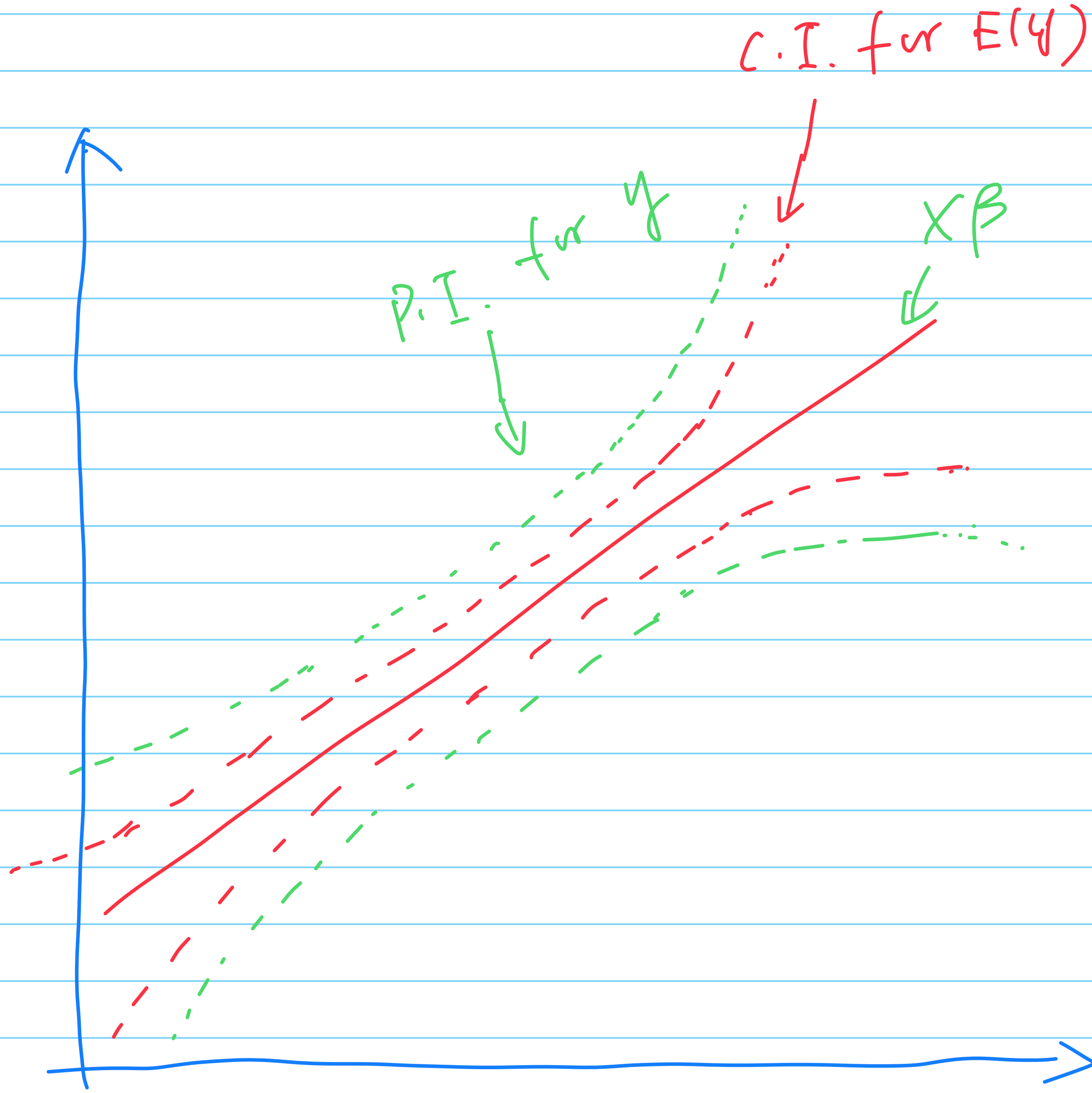
49

P.I. for y



$$\mu_{x_0} = E(y_0) = \underline{x_0' \beta}$$

a unknown value (fixed)


C.I. for E(y)
P.I. for y
Xβ