Confidence intervals and regions

estimation — point estimation / interval/region estimation

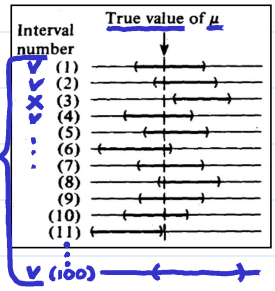
- Q: Why need interval/region estimation? What more information can it provide compared to point estimation?

  e.g., estimate of $\beta = 3.5$, but accept H_0: $\beta = 0$. How to give such result an explanation? Why point estimation cause such confusing?

  ($\hat{\beta}$: a random variable)

- An interval/region estimation provides
  - plausible values for parameter
  - uncertainty in parameter estimator (report $\hat{\beta}$ & $se(\hat{\beta})$ ← both information are contained/combined in C.I.)
  - information about its length and the values it covers may be helpful information related to testing

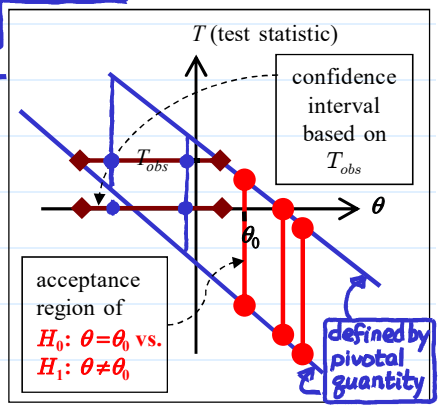
  (can use C.I. or C.R. to do test)



about 95% of these C.I.'s cover μ

A C.I. offers more information than a single test

- meaning of $100(1-\alpha)\%$ confidence interval or region, e.g., 95% confidence interval
- *duality* of interval/region estimation and hypothesis test: For a $100(1-\alpha)\%$ confidence region, any point θ that *lies within* the region represents a null hypothesis that *would not be rejected* at the $100\alpha\%$ significance level while every point θ *outside* represents a null hypothesis that *would be rejected*.



- Model: $Y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I)$; $\hat{\beta}$: OLS estimator $\Rightarrow \hat{\beta} \sim N(\beta, (X^TX)^{-1}\sigma^2)$ (LN p.4-3; $(X^TX)^{-1}X^TY$)
  - Confidence region for $A\beta$, where A is a full rank $d \times p$ matrix and $d \le p$

parameters → $C = (C_1, \cdots, C_d)^T \equiv$ ← (N6) in LN p.4-3

$$A\hat{\beta} \sim N(A\beta, A(X^TX)^{-1}A^T\sigma^2) \Rightarrow [(A\hat{\beta}-A\beta)^T[A(X^TX)^{-1}A^T]^{-1}(A\hat{\beta}-A\beta)]/\sigma^2 \sim \chi^2_d$$

(N1) in LN p.4-3; $d \times d$ matrix of full rank; $-\frac{1}{2}$ and $-\frac{1}{2}$

$$(n-p)\,\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p},$$

RSS/(n-p)

χ^2_d/d / $\chi^2_{n-p}/(n-p)$

and they are **independent**.

$\because \hat{\beta}$ & $\hat{\sigma}^2$ are indep. (LN p.4-4) → pivotal quantity →

$$[(A\hat{\beta}-A\beta)^T[A(X^TX)^{-1}A^T]^{-1}(A\hat{\beta}-A\beta)] / (d\,\hat{\sigma}^2) \sim F_{d,n-p}$$

  - $100(1-\alpha)\%$ confidence region of $A\beta$: collection of $A\beta$'s (or β) that satisfy

general form

$$[(A\hat{\beta}-A\beta)^T[A(X^TX)^{-1}A^T]^{-1}(A\hat{\beta}-A\beta)] / (d\,\hat{\sigma}^2) \le F_{d,n-p}(\alpha)$$

($A\beta \leftarrow C$, C_0; significance level; $H_0: A\beta = C_0$, $H_1: A\beta \neq C_0$)

↕ cf. F-test

The regions are often ellipsoidally shaped (Q: why?)

quadratic form: M: $n \times n$ symmetric matrix, $M = \Gamma^T \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} \Gamma$

$x^TMx = y^T \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} y = \sum_{i=1}^n \lambda_i y_i^2 = \sum_{i=1}^n z_i^2$ ($z = \Gamma x$)

$\lambda_i \ge 0$ if $\forall x: y_i$; M: positive semi-def.

- Examples:
  - confidence region for β, i.e, $A = I_{p\times p}$

the overall F-test ↔ $A = \begin{bmatrix} 0 & 1 & 0 & \\ 0 & 0 & 1 & \cdots \\ \vdots & & \ddots & \\ 0 & 0 & 0 & \cdots\ 1 \end{bmatrix}$, θ_0

$$(\hat{\beta}-\beta)^TX^TX(\hat{\beta}-\beta) \le (p\,\hat{\sigma}^2)\,F_{p,n-p}(\alpha)$$

(Q: What's the confidence region for all effects?)

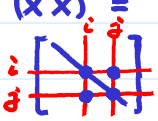
  - confidence region of β_i, β_j, i.e, $A = \begin{pmatrix} 0, \cdots, 0, \overset{i\text{th}}{1}, 0, \cdots, 0, 0, 0, \cdots, 0 \\ 0, \cdots, 0, 0, 0, \cdots, 0, \overset{j\text{th}}{1}, 0, \cdots, 0 \end{pmatrix}$

test a pair of effects

$\begin{bmatrix} \hat{\beta}_i \\ \hat{\beta}_j \end{bmatrix}$ = $\begin{bmatrix} \beta_i \\ \beta_j \end{bmatrix}$ $\begin{bmatrix} (X^TX)^{-1}_{ii} & (X^TX)^{-1}_{ij} \\ (X^TX)^{-1}_{ji} & (X^TX)^{-1}_{jj} \end{bmatrix}$

$$[(A\hat{\beta}-A\beta)^T[A(X^TX)^{-1}A^T]^{-1}(A\hat{\beta}-A\beta)] \le (2\,\hat{\sigma}^2)F_{2,n-p}(\alpha)$$

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

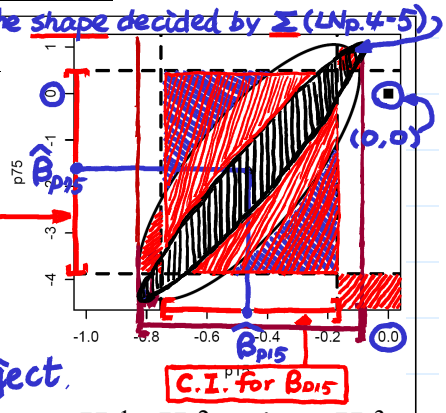


$$\mathbf{Z} \equiv \begin{pmatrix} \hat{\beta}_i \\ \hat{\beta}_j \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \beta_i \\ \beta_j \end{pmatrix}, \sigma^2 \begin{pmatrix} (\mathbf{X}^T\mathbf{X})^{-1}_{ii} & (\mathbf{X}^T\mathbf{X})^{-1}_{ij} \\ (\mathbf{X}^T\mathbf{X})^{-1}_{ji} & (\mathbf{X}^T\mathbf{X})^{-1}_{jj} \end{pmatrix} \right) \equiv \mathrm{N}(\mu, \sigma^2\Sigma) \Leftarrow \Sigma = A(X^TX)^{-1}A^T$$

contour plot in LNp.4-5 (cf.)

confidence region of β_i and β_j: $\{\mu \mid (\mathbf{Z}-\mu)^T\Sigma^{-1}(\mathbf{Z}-\mu) \le c\}$ for some c

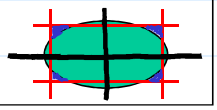
- example: confidence region and intervals of β_{p15} and β_{p75}



  - **Q1**: why the straight lines not tangential to the ellipse?

$$\begin{aligned} 1-\alpha &= P(\{(\beta_{p15},\beta_{p75}) \in \text{C. Region}\}) \\ &= P(\{(\beta_{p15},\beta_{p75}) \in \text{C.I.}_{p15} \times \mathbb{R}\}) \\ &= P(\{(\beta_{p15},\beta_{p75}) \in \mathbb{R} \times \text{C.I.}_{p75}\}) \end{aligned}$$

  - **Q2**: what can you say, based on the plot, about the results of testing H_0^1:β_{p15}=0, H_0^2: β_{p75}=0, and H_0^3: β_{p15}=β_{p75}=0? Ans. H_0^1: reject, H_0^2: accept, H_0^3: reject
  - **Q3**: where will be the point (0,0) located if the data accept H_0^1, H_0^2, reject H_0^3? how to explain the result if (0,0) falls in other regions? (exercise) (LNp.4-13)
  - **Q4**: what is the correlation between $\hat{\beta}_{p15}$ and $\hat{\beta}_{p75}$? how will the shape of ellipse change when the correlation becomes larger or smaller? (LNp.4-5) (LNp.3-8) $\frac{(X^TX)^{-1}_{ij}}{\sqrt{(X^TX)^{-1}_{ii}(X^TX)^{-1}_{jj}}}$
  - **Q5**: can you see why the situation in Q3 will happen more frequently when the correlation between $\hat{\beta}_{p15}$ and $\hat{\beta}_{p75}$ gets larger? (cf.) LNp.4-13. graph for $\hat{\beta}_3, \hat{\beta}_4$ with strong collinearity
  - **Q6**: if $\hat{\beta}_{p15}$ and $\hat{\beta}_{p75}$ are uncorrelated, what would be the shape of the confidence region? why situation in Q3 less possible to occur? (LNp.4-5)



confidence interval for β_i, i.e, A=(0,...,0,1,0,...,0) (1 at ith)

test just one effect

$$(\hat{\beta}_i-\beta_i)^2/(X^TX)^{-1}_{ii} \le \hat{\sigma}^2 F_{1,n-p}(\alpha) \Rightarrow |(\hat{\beta}_i-\beta_i)/\left(\hat{\sigma}\sqrt{(X^TX)^{-1}_{ii}}\right)| \le t_{n-p}(\alpha/2)$$

equivalent; take square root on both sides

alternative method:

① $\hat{\beta}_i \sim N(\beta_i, \sigma^2(X^TX)^{-1}_{ii})$, ② $(n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$, and ③ they are independent

pivotal quantity $\Rightarrow \left|(\hat{\beta}_i-\beta_i)/\left(\hat{\sigma}\sqrt{(X^TX)^{-1}_{ii}}\right)\right| \sim t_{n-p}$ (LNp.5-1) $\Rightarrow$ C.I.: $\hat{\beta}_i \pm t_{n-p}(\alpha/2) \times \left(\hat{\sigma}\sqrt{(X^TX)^{-1}_{ii}}\right)$

$N(0,1) \sim \frac{\hat{\beta}_i-\beta_i}{\sigma\sqrt{(X^TX)^{-1}_{ii}}} \Big/ \sqrt{\frac{(n-p)\hat{\sigma}^2/\sigma^2}{n-p}} \sim \sqrt{\frac{\chi^2_{n-p}}{n-p}}$

center of C.I. | critical value | se($\hat{\beta}_i$)

- confidence interval for prediction of mean response at x_0 (LNp.4-4) → $x_0^T\beta \Rightarrow A = x_0^T$

LNp.4-4: $x_0^T\hat{\beta} - x_0^T\beta \sim N(\mathbf{0}, (x_0^T(X^TX)^{-1}x_0)\sigma^2) \Rightarrow \left|(x_0^T\hat{\beta}-x_0^T\beta)/\left(\hat{\sigma}\sqrt{x_0^T(X^TX)^{-1}x_0}\right)\right| \sim t_{n-p}(\alpha/2)$

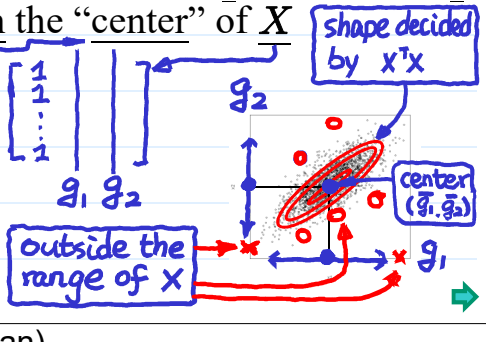
($x_0^T\hat{\beta}$: estimate)

$\Rightarrow$ C.I.: $x_0^T\hat{\beta} \pm t_{n-p}(\alpha/2) \times \left(\hat{\sigma}\sqrt{x_0^T(X^TX)^{-1}x_0}\right)$

center of C.I. | estimate of $x_0^T\beta$ | se($x_0^T\hat{\beta}$)

- a quadratic form
- Mahalanobis distance (future lecture)

  - **Q**: for a given dataset and α, the length of the C.I. is related to x_0 only. What x_0 will cause a wider C.I.? **Ans**: x_0 that is away from the "center" of X
  - interpolation and extrapolation (check LNp.3-7, projection view of centering)
    - 內插 interpolation: x_0 lie "within the range" of X
    - 外插 extrapolation: x_0 lie "outside the range" of X

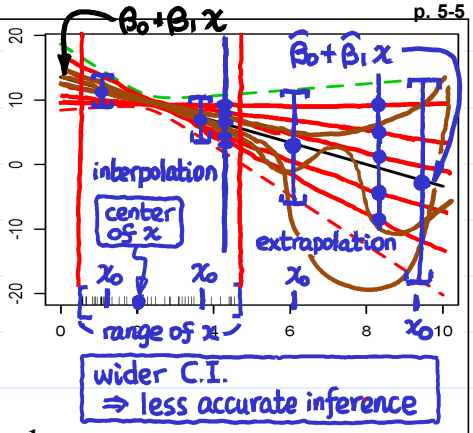


(**Q**: fitted model still hold outside the range?)

✓ quantitative x_0 and qualitative x_0

- Example: 95% *pointwise* confidence band for prediction of mean responses (model: $y=\beta_0+\beta_1 x+\varepsilon$)

pointwise confidence band



- Q1: why the confidence intervals get wider when we move away from the range of data?
  - Ans: wider C.I. ⇐ ε cause it
  - Ans: fitted model may approximate the true model very badly or wrongly & you don't have any information about it.
- Q2: what's the danger of extrapolation?
- Q3: does the widening reflect the possibility that the mean structure of the model may change outside the range? → No. Because the confidence band is constructed under the assumption that the fitted model still hold outside the range of x
- Q4: does the plot represent a *simultaneous* confidence band for **all** prediction of mean response? → No. S.C.B will be wider than P.C.B.

➢ C.I. for prediction of future observation at x_0

y_{x_0} (r.v.) $x_0^T\hat\beta - (x_0^T\beta+\varepsilon) \sim N(\mathbf{0}, (x_0^T(X^TX)^{-1}x_0+1)\sigma^2)$ (estimate; LNp.4-4)

$\Rightarrow$ C.I.: $x_0^T\hat\beta \pm t_{n-p}^{(\alpha/2)} \times \left(\hat\sigma\sqrt{1+x_0^T(X^TX)^{-1}x_0}\right)$

cf. $x_0^T\beta$ (parameter); estimate of y_{x_0}; not reduced when n↑; $se(x_0^T\hat\beta-\varepsilon)$

➢ a general form for confidence interval: ← check the C.I. for β_i's, prediction, ...

estimate ± **(critical value)** × **(standard error of estimate)**

C.I.: a combination of $\hat\theta$ & $se(\hat\theta)$.

❖ **Reading**: Faraway (2005, 1st ed.), 3.4, 3.5

❖ **Futher reading**: D&S, 5.3, 5.4, 5.5

Sampling model

no plan with a sampling plan

- experimental data vs. observational data

graph in LNp.2-1

It depends on whether we have *control* over predictors

examples: yield of crop ← response Y

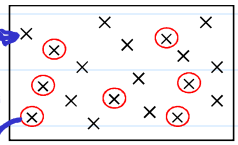
experimental: fertilizer, …, $(X_1,\cdots,X_m)$

observational: exposure, weather, …

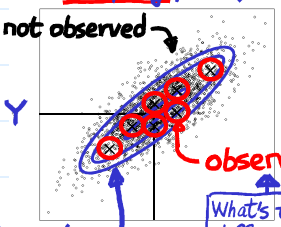
no plan; operators (sampling plan)

an object or a unit → $(Y, X_1, \cdots, X_m)$ r.v.

population



observational data (no plan or with a sampling plan)



What's the difference btw no plan & with a sampling plan?

➢ Q: What difference between inferences based on experimental data and observational data?

Ans: experimental data: causation, 因果 cf. 關聯

e.g. smoking & lung cancer

e.g., father/son height

observational data: often only association (**Note**. lurking variable)

➢ Q: Is this model description, $Y=X\beta+\varepsilon$, $\varepsilon\sim N(0, \sigma^2 I)$, appropriate for observational data? **Note** that

(1) observational X are random variables

(2) in LM, X are treated as fixed values, i.e., no distribution assigned for X

① X→Y ② Y→X ③ X ↔ Y ④ random disturbance

LNp.5-12

➢ Q: difference between "X is random" and "X measured with (random) error"

example: X is random (why?), but measured accurately | measurement error model (future lecture)

$x^* = x + \delta$ (x^*: observed value (random); x: true value (fixed or random); δ: random error)

survey sampling

for some data sets, we can regard the data as a *sample* drawn from a *population*. In the case, we want to say something about the unknown population values using estimated values that are obtained from the sampled data. (example?)

→ better chance to achieve pattern in sample ≈ pattern in population ⇒ good sample (representative, LNp.1-3)

the data should be generated using a "(simple) *random* sample" of the population so that they can be representative → each object/unit in the population has the same (known) probability to be included in the sample

- conditional distribution of multivariate normal: If

Z in LNp.4-3 ↔ cf. p+q variables

$$\underset{(p+q)\times 1}{\mathbb{Z}} = \begin{bmatrix} \underset{p\times 1}{Z_1} \\ \underset{q\times 1}{Z_2} \end{bmatrix} \sim N\left(\begin{bmatrix} \underset{p\times 1}{\mu_1} \\ \underset{q\times 1}{\mu_2} \end{bmatrix}, \begin{bmatrix} \underset{p\times p}{\Sigma_{11}} & \underset{p\times q}{\Sigma_{12}} \\ \underset{q\times p}{\Sigma_{21}} & \underset{q\times q}{\Sigma_{22}} \end{bmatrix} \right), \quad \Sigma_{12}=\Sigma_{21}^T$$

check LNp.3-12

then

$$Z_1 \mid Z_2 = z_2 \sim N\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(z_2-\mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \right)$$

Schur complement

regression line ← cf. TSS in R^2 ← cf. → cf. RSS in R^2

Recall simple regression (LNp.3-7)

$$\frac{y-\mu_Y}{\sigma_Y} = \rho\left(\frac{x-\mu_X}{\sigma_X}\right) \quad \text{cf. regression effect (LNp.2-7)}$$

$$\Rightarrow y = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X) = \mathrm{cov}(x,y)\times(\sigma_x^2)^{-1}$$

TSS − RSS $= (\rho\sigma_X\sigma_Y)^2/\sigma_X^2 = \rho^2\times\sigma_Y^2$

- an alternative view of regression: data (y_i, x_i), $i=1,\ldots,n$, are randomly sampled from a multivariate Normal population,

vector of p r.v.'s

$$\begin{bmatrix} \underset{1\times 1}{Y_i} \\ \underset{p\times 1}{X_i} \end{bmatrix} \overset{\text{iid}}{\sim} N\left(\begin{bmatrix} \underset{1\times 1}{\mu_Y} \\ \underset{p\times 1}{\mu_X} \end{bmatrix}, \begin{bmatrix} \underset{1\times 1}{\sigma_Y^2} & \underset{1\times p}{\Sigma_{XY}^T} \\ \underset{p\times 1}{\Sigma_{XY}} & \underset{p\times p}{\Sigma_{XX}} \end{bmatrix} \right) \Rightarrow y_i \mid X_i=x_i \sim N((\mu_Y-\beta^T\mu_X)+\beta^T x_i, \sigma^2).$$

required? cf. $y_x \sim N(X^*\beta^*, \sigma^2)$ fixed intercept $\beta_1 X_1 + \cdots + \beta_p X_p$ constant variance

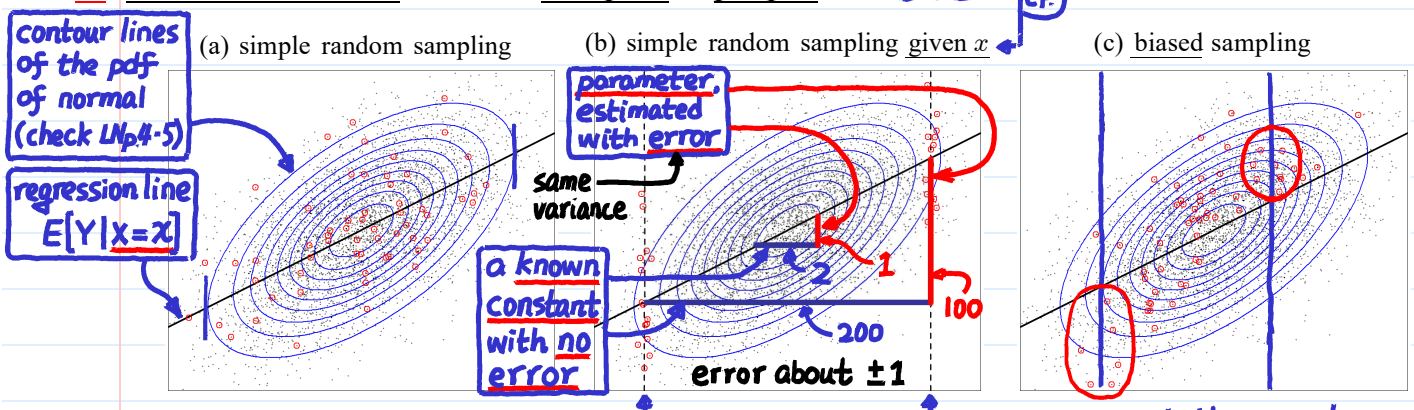
It is a linear model with $\beta=\Sigma_{XX}^{-1}\Sigma_{XY}$, $\sigma^2 = \sigma_Y^2 - \Sigma_{XY}^T\Sigma_{XX}^{-1}\Sigma_{XY} \equiv \sigma_Y^2(1-r^2)$.

$\Sigma_{XY}=0 \Rightarrow \beta=0$ ← $\beta^T = \Sigma_{XY}^T\Sigma_{XX}^{-1}$ ← R^2 estimate → $r^2 = 1-(\sigma^2/\sigma_Y^2) = \sigma_Y^2(0/\sigma_Y^2)$ ← population multiple correlation

When we are interested in the "transformed" parameters, regression can be applied.

- Q: what information in these samples is proper?

DOE ← cf.



	(a)	(b)	(c)
Joint distribution (X, Y)	✓	×	×
Conditional distribution $Y\|X$	✓	✓✓	×
Marginal distribution X	✓	×	✓ or ×

a better way to collect information about $E[Y|X=x]$ than simple random sampling (a) (check graphs in LNp.3-18)

- Q: what will happen if the sample is not random?

(i) biased sample → sample is not representative

census survey (ii) sample of convenience → may or may not be a good sample ← large population ⊃ small population ← sampling

(iii) sample = population → (1) estimation is exact (2) testing is questionable. ← permutation test

these nonrandom samples can cause problems in the inference (e.g., R^2, LNp.3-18)

sampling probability of each object/unit could be unknown

❖ **Reading**: Faraway (2005, 1st ed.), 3.8, nonrandom samples
Weisberg (2005), *Applied Linear Regression*, 3rd Ed., 4.2, 4.3

Orthogonality

- Q: consider the two models:
  model 1: $y = \beta_0 + \beta_1 x_1 + \varepsilon$,
  model 2: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
  In general, $\hat{\beta}_1$, in the two models are not identical (Q: why?) → check the graph in LN p.4-13 3-13
  [also, test H_0: $\beta_1 = 0$ (or c) not identical]
  an *exception*: when x_1 and x_2 are orthogonal

• span$(X_1) \perp$ span(X_2) $\Leftrightarrow$ orthogonality
• span$(X_1) \supseteq$ span(X_2) $\Leftrightarrow$ aliasing (in DOE)

fitted model=model 1: $Y = X_1\beta_1 + \varepsilon$
true model=model 2: $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$

- $E(\hat{\beta}_1) = \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2$ (II? 0)
- $E(X_1\hat{\beta}_1) = X_1\beta_1$ (II? 0)

$\hat{Y}$: Hat matrix of X_1 $+ X_1(X_1^T X_1)^{-1} X_1^T (X_2\beta_2)$ LN p.3-11

Note. If fitted model=model 2
- $E(\hat{\beta}_1) = \beta_1$

- $Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$, where $\beta = [\beta_1\ \beta_2]^T$ and $X = [X_1\ X_2]$ with the property
  $X_1^T X_2 = 0 \Rightarrow X_1$ and X_2 are orthogonal (generalization?)

$X = [X_1\ X_2 \cdots X_k]$

$\hat{\beta} = \begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix} = (X^TX)^{-1}X^TY = \begin{bmatrix}\square & 0\\ 0 & \square\end{bmatrix}\begin{bmatrix}X_1^TY\\ X_2^TY\end{bmatrix} = \begin{bmatrix}(X_1^TX_1)^{-1}X_1^TY\\ (X_2^TX_2)^{-1}X_2^TY\end{bmatrix}$

$$X^TX = \begin{pmatrix} X_1^TX_1 & X_1^TX_2 \\ X_2^TX_1 & X_2^TX_2 \end{pmatrix} = \begin{pmatrix} X_1^TX_1 & 0 \\ 0 & X_2^TX_2 \end{pmatrix}$$

$$\Rightarrow (X^TX)^{-1} = \begin{pmatrix} (X_1^TX_1)^{-1} & 0 \\ 0 & (X_2^TX_2)^{-1} \end{pmatrix}$$

$X^TX = \begin{bmatrix} X_1^TX_1 & 0 & 0 \\ 0 & X_2^TX_2 & \\ 0 & & \ddots \\ & & & X_k^TX_k \end{bmatrix}$

By (N4) in LN p.4-3

- Estimation: $\hat{\beta}_1 = (X_1^TX_1)^{-1}X_1^TY$, $\hat{\beta}_2 = (X_2^TX_2)^{-1}X_2^TY$, and $\hat{\beta}_1, \hat{\beta}_2$ independent
  $\Rightarrow$ note that $\hat{\beta}_1$ will be the same regardless of whether X_2 is in the model or not (and vise versa).

fitted model: $Y = X_1\beta_1 + \varepsilon \Rightarrow \hat{\beta}_1 = (X_1^TX_1)^{-1}X_1^TY$

$X_i \perp X_j \Leftrightarrow X_i^TX_j = 0 \Leftrightarrow (X^TX)_{ij} = 0$
But. $X_1^TX_2 \neq 0 \not\Rightarrow (X^TX)^{-1}_{ij} = 0$
$\Rightarrow \hat{\beta}_i, \hat{\beta}_j$ may not be indep.

- Q: what if only two predictors, say some x_i in X_1 and some x_j in X_2, are orthogonal?