- Test: **Q**: how about test H_0: $\beta_1=0$ (or c) in models 1 and 2 when orthogonality exists between $\{x_1, \mathbf{1}\}$ and x_2? will the test results be identical?

  model 1: ω_1: $y=\beta_0+\varepsilon$ vs. Ω_1: $y=\beta_0+\beta_1 x_1+\varepsilon$

  model 2: ω_2 : $y=\beta_0+\beta_2 x_2+\varepsilon$ vs. Ω_2: $y=\beta_0+\beta_1 x_1+\beta_2 x_2+\varepsilon$

  $F=[(RSS_\omega-RSS_\Omega)/(df_\omega-df_\Omega)]/[RSS_\Omega/df_\Omega]\sim F_{1,\,df_\Omega}$

  $RSS_{\omega_1}-RSS_{\Omega_1}=RSS_{\omega_2}-RSS_{\Omega_2}$ (**Q**: why?) but, $RSS_{\Omega_1}\neq RSS_{\Omega_2}$, and $df_{\Omega_1}\neq df_{\Omega_2}$, i.e., $\hat\sigma^2_{\Omega_1}\neq\hat\sigma^2_{\Omega_2}$.

  Q: when will the test results be consistent? ($\hat\sigma^2_{\Omega_1}\approx\hat\sigma^2_{\Omega_2}$) when will be very different?

  Note: although the tests do depend on the presence of x_2, the dependence is usually not as strong as in non-orthogonal cases.

- orthogonality is very unlikely to achieve in observational data (it's a feature of experimental data from a good design. In experimental case, orthogonal design is an important criterion). At best, predictors are almost uncorrelated and "near" orthogonality holds.

- Randomization: In an exp't, suppose that true model is $Y=X\beta+Z\gamma+\varepsilon$, but Z cannot be measured or may not even be suspected $\Rightarrow E(\hat\beta)=\beta+(X^TX)^{-1}X^TZ\gamma$

  Q: what's the best way of controlling X to make X and Z as orthogonal as possible?

❖ **Reading**: Faraway (2005, 1st ed.), 3.6
❖ **Futher reading**: D&S, Appendix 6A

Identifiability

- model: $Y=X\beta+\varepsilon$, where X is an $n\times p$ matrix $\Rightarrow$ OLS estimator $\hat\beta=(X^TX)^{-1}X^TY$

  Q: what if the inverse of X^TX does not exist?

- β (or X) is called *unidentifiable* when X^TX is singular ($\Leftrightarrow$ rank(X)$<p$ $\Leftrightarrow$ dim(Ω)$<p$ $\Leftrightarrow$ at least one column of X is a linear combination of other columns)
  - the normal equation $X^TX\beta=X^TY$ has infinite many solutions. Any $\hat\beta=(X^TX)^-X^TY$, is a solution, but should not be regarded as an estimate of β.
  - $\hat Y$ and $\hat\varepsilon$ are still unique

- **Q**: Why does unidentifiability happen?
  - observational data, some examples:
    - same predictor measured in different scales, and both are in the model
    - $X_1+X_2=X_3$, or $X_1+X_2+X_3=c$, and all three are in the model with intercept
    - X is *supersaturated*: $p>n$, i.e., more effects than observations (**Note**. *saturated* X: when $p=n$ and X^TX is nonsingular $\Rightarrow$ $\hat\beta$ is identifiable, but no degrees of freedom left for estimation of σ because $Y=\hat Y$ and $\hat\varepsilon=\mathbf{0}$ $\Rightarrow$ cannot do testing or C.I.)
    - such problems can be avoided by paying attention.

- experimental data, e.g., two-sample case:

  treatment data: $y_1, \ldots, y_n$ (with mean μ_1) control data: $y_{n+1}, \ldots, y_{m+n}$ (with mean μ_2) Suppose we model the response by an overall mean μ and group effects α_1 and α_2:

  $$y_i=\mu+\alpha_1+\varepsilon_i,\ i=1,\ldots,n; \quad y_i=\mu+\alpha_2+\varepsilon_i,\ i=n+1,\ldots,n+m,$$

  ($\mu_1=\mu+\alpha_1$, $\mu_2=\mu+\alpha_2$)

  $$\begin{pmatrix} y_1 \\ \vdots \\ y_n \\ y_{n+1} \\ \vdots \\ y_{m+n} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \\ \epsilon_{n+1} \\ \vdots \\ \epsilon_{m+n} \end{pmatrix}$$

  (treatment / control) $\Rightarrow X$ (or β) is unidentifiable

$(\mu, \alpha_1, \alpha_2)$	(μ_1, μ_2)
10 5 -1	15 9
5 10 4	15 9
0 15 9	15 9
12 3 -3	15 9

  $\Rightarrow$ *over-parameterized*: some constraint must be imposed on $(\mu, \alpha_1, \alpha_2)$, say (reduce the dimension of parameters from 3 to 2)

  $\mu=0$ or $\alpha_1+\alpha_2=0$

  $\mu=0$: $\alpha_1=\mu_1$: treatment mean, $\alpha_2=\mu_2$: control mean

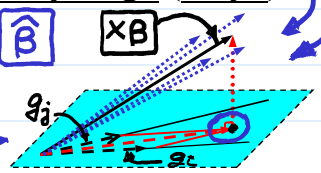
  $\alpha_1+\alpha_2=0$: $\mu=\frac{\mu_1+\mu_2}{2}$: average of 2 means; $\alpha_1=-\alpha_2=\frac{\mu_1-\mu_2}{2}$: (mean difference)/2

- "unidentifiabile" means
  1. insufficient data to estimate the parameters of interest, or
  2. more parameters than are necessary to model the data
- an eigen-decomposition of X^TX will reveal the linear combinations that gave rise to the unidentifiability (check lab)

  ($\exists\, a$ s.t. $a_1x_1+\cdots+a_px_p\approx 0$)

  x_i: ith column of X. Suppose $a_1x_1+\cdots+a_px_p=0$. Let $a=(a_1,\cdots,a_p)$, then $(X^TX)a=0=0\cdot a$. $\Rightarrow a$ is an eigenvector of X^TX, whose corresponding eigenvalue is 0.

  $(X^TX)^{-1}$: $\mathrm{cor}(\hat\beta_i,\hat\beta_j)$ and $\mathrm{cor}(g_i,g_j)$ have opposite signs
- what causes problem is data *close to* "unidentifiable," (i.e., strong collinearity) $\Rightarrow$ model is still identifiable, but standard error of estimates can be very large (why?)
- statistical softwares handle unidentifiability differently. R will automatically fit a reduced model when X is unidentifiable.



graph in LNp.4-13 cf.

❖ **Reading**: Faraway (2005, 1st ed.), 2.9
❖ **Futher reading**: D&S, 4.2, 20.4, Appendix 20A.

Interpreting parameter estimates

LNp.1-1 ← cf. → prediction (Recall. In LNp.5-9)

	X_1, X_2 orthogonal	X_1, X_2 collinear
$\hat\beta_1$	$E(\hat\beta_1)=\beta_1$	$E(\hat\beta_1)=\beta_1+(X_1^TX_1)^{-1}X_1^TX_2\beta_2$
$\hat Y=X_1\hat\beta_1$	$E(\hat Y)=X_1\beta_1$	$E(\hat Y)=X_1\beta_1+[X_1(X_1^TX_1)^{-1}X_1^T](X_2\beta_2)$

- Q: $Y=X\beta+\varepsilon$, what does $\hat\beta$ mean? (定性資訊)

What? How? Why? Some matters needing attention about $\hat\beta$:

- $\hat\beta$ have units [e.g., fuel consumption data, fitted model: true $Y=X_1\beta_1+X_2\beta_2+\varepsilon$, fitted: $Y=X_1\beta_1+\varepsilon$)

  fuel = 154.19 + (– 4.23)Tax + (0.47)Dlic + (–6.14)Income + (18.54)$\log_2$(Miles)]

  its unit = unit of y; its unit = (unit of y)/(unit of x_i)

often more robust than the value of $\hat\beta$

- sign of $\hat\beta$: direction of the relationship between the term and the response
- interpretation of estimated value (see next two slides)
- better to also consider values contained in its confidence interval
  1. what if the C.I. contains 0?
  2. what if (upper bound − lower bound) of the C.I. is very large?
  3. what if (upper bound/lower bound) of the C.I. is very large or almost 1 or 0?
- causality or association
- the parameters β
  - some β_i's have physical interpretation, especially those from a conceptual model [e.g., attach weights x to a spring and measure the extension y]

    $\Rightarrow$ unfortunately, such cases are rare

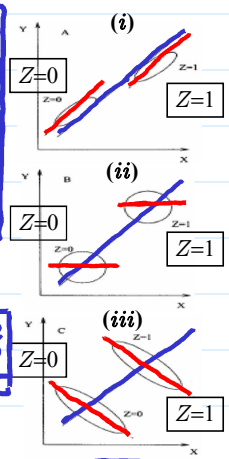
    check LNp.1-2
  - usually, β_i's do not have such physical interpretation

    $\Rightarrow$ in the case, the model $Y=X\beta+\varepsilon$ is only an *empirical model*, i.e., a convenience for representing a complex reality within the range of X $\Rightarrow$ the real meaning of a particular β_i is not obvious, interpretation is difficult

local approximation

Some interpretations of parameter estimates

- a naive interpretation:
  "A unit increase in X_i will *cause* an average change of $\hat{\beta}_i$ in Y" ⇐ causality statement
  [e.g., Y: annual income, and X: years of education]
  - Q: what if there exist lurking variables?
    [e.g., X: shoe size, Y: reading abilities, Z: age of child]
    ⇒ causal conclusion is doubtful
  - Q: what if the roles of predictor and response are mistakenly switched?
    [e.g., Y: fire damage, and X: numbers of firefighters called out]
  - Q: what if some important effects are not included in model?
    - X fixed. $E(\hat{\beta}_1) = \beta_1 + (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2\beta_2$
    - X random. true model: $E(Y \mid \mathbf{X}_1, \mathbf{X}_2) = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2$,
      fitted model: $E(Y \mid \mathbf{X}_1) = \mathbf{X}_1\beta_1$
      $E(Y \mid \mathbf{X}_1) \doteq \mathbf{X}_1\beta_1 + E(\mathbf{X}_2 \mid \mathbf{X}_1)\beta_2$
      $Var(Y \mid \mathbf{X}_1) \doteq \sigma^2 + \beta_2^T\, Var(\mathbf{X}_2 \mid \mathbf{X}_1)\, \beta_2$
  - even though we have all important variables in the model and no lurking variables, there still are problems, e.g.:
    $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon = \beta_0 + (\beta_1 - \beta_2) X_1 + \beta_2 (X_1 + X_2) + \varepsilon$
  - in a properly designed experiment, the naive interpretation is more reasonable (because of its use of orthogonal designs and randomization); but for observational data, it's often questionable.



- an alternative interpretation
  " A unit increase in X_i with all the other (specified) terms *held constant* will be associated with an average change of $\hat{\beta}_i$ in Y"
  - Q: can other terms be held constant? e.g.
    - X_1 and X_2 are highly correlated
    - consider the model $E(Y)=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_1X_2=\beta_0+(\beta_1+\beta_3X_2)X_1+\beta_2X_2$
  - it requires the specification of the other terms/effects
    Q: what will happen in the analysis when strong collinearity exists between effects?
    ⇒ estimates and tests of β_i's may significantly change according to *what other effects are included*. It makes the interpretation almost impossible (check lab). In some cases, the problem can be removed by redefining the terms into new linear combinations that may be easier to interpret.
- an interpretation from prediction viewpoint
  regarding the parameters and their estimates as fictional quantities, and concentrating on prediction enable a rather cautious interpretation of $\hat{\beta}$:
  given $(g_{1,0},\ldots,g_{i,0},\ldots,g_{p-1,0}) \to \hat{y}_0$, observe $(g_{1,0},\ldots,g_{i,0}+1,\ldots,g_{p-1,0}) \to \hat{y}_0 + \hat{\beta}_i$
  - prediction is more stable than parameter estimation (check lab)
  - directly interpretable and success may be measured in future
  - dangers of extrapolation, be cautious when x_0 is outside the range of X

- **Q**: how to make a stronger case for causality (be associated with → cause)?
  - include all relevant variables/effects ⇒ however, even though you try hard to do so, the possibility of an unsuspected lurking variable will always exists
    (important variable)
  - Recall Level of evidence (LN p.1-12) ⇒ fit a variety of models and see if a similar effect is observed, i.e., whether the estimates of β_i similar no matter what the fitted models are?
    Note. when orthogonality exists ⇒ identical $\hat{\beta}_i$
  - e.g. smoking → lung cancer ⇒ use non-statistical knowledge of the physical nature of the relationship ⇒ conceptual model is more persuasive than empirical model
  - ⇒ multiple studies under different conditions can help confirm a relationship.
  - in a few cases, one can infer causality from an observational study.
    biased sample? [e.g., Dahl and Moretti (2003): parents of a single girl are 5% more likely to divorce than parents of a single boy. This observational study functions like an experimental design because the sex of a child is a purely random matter.]
    response: divorce, predictor: sex
    X_1: sex, X_2: other possible variables
    caused by some lurking variables or unobserved important predictors?
    X_1, X_2: independent ⇒ orthogonality ⇒ randomization ⇒ remove the possibility of lurking variables
  - even if these steps are accomplished, one can never be 100% sure of the causality relationship purely based on a statistical analysis. For example, consider the history of the study of the link between smoking and lung cancer ⇒ it takes decades of studies to go from association to causality
    FYI. a new field in statistics: causal inference

❖ **Reading**: Faraway (2005, 1st ed.), 3.6, 3.7

What can go wrong? many many things ...

$$Y = X\beta + \varepsilon,\quad \varepsilon \sim N(0, \sigma^2 I)$$

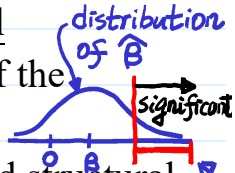
- source and quality of the data (**Q**: how was the data collected?)
  Note. the information contained in the data might be limited or wrong from the beginning.
  - data may not be a random sample of the population. Situations such as biased sample, a sample of convenience, or sample=population → check graphs in LN p.5-8
  - important predictors may not have been observed (**Q**: how may you find out?)
  - observational data often make causal conclusions problematic, reason: lack of orthogonality, collinearity, lurking variables, …
    e.g. • too low R^2 • too large $\hat{\sigma}^2$ • check with expertise
  - the range of X and qualitative nature of some predictors may limit effective predictions, it's unsafe to extrapolate too much
  - Key: data collected should be *representative* of the *population of interest*
    the objective of data collection → DOE, survey
- error component [we hope $\varepsilon \sim N(0, \sigma^2 I)$] ← 隨機
  - ε may have unequal variance ← WLS
  - ε may be correlated ← GLS
  - ε may not be normally distributed
    e.g., ε ~ heavy-tailed distribution like Cauchy → outlier → robust regression
    detection
    Note. $\hat{\beta} = (X^TX)^{-1}X^TY$
    need it in testing & C.R., not estimation
    - this is less serious when sample size is large. Notice that even if ε is not normal, $\hat{\beta}$ might tend to normality due to CLT. With large sample size, normality of data is not much of a problem
      cf. normality of $\hat{\beta}$
      Inference based only on normality of $\hat{\beta}$ might still be valid.
    - for small sample sizes, bootstrap method offers a solution
      treat empirical cdf of $\hat{\varepsilon}$ as the true cdf of ε

- structural component [$E(Y)=X\beta$] ← 規律
  - errors in X → measurement error model
  - serious collinearity in X → • principal component regression • ridge regression • shrinkage method (like LASSO, ⋯)
  - some inferences strongly rely on the choice of full model, $X\beta$ (example?) ← Lab5-4

  Q: where does the full model come from? ← conceptual model

  confidence decrease
  1. physical theory may suggest a model --- wonderful but relatively uncommon
  2. experience from past data --- may help suggesting a reliable model (cf.)
  3. no prior experience --- explore current data to find an empirical model
     - confidence in inference will depend on confidence in the model
     - an empirical model can be regarded as a *local approximation* of the underlying true system on some "safe" range of X

  truth inflation



- many statistical theory rests on the assumption that the model (error and structural components) is correct. In practice, the best one can hope for is often "empirical model≈underlying system". [Box: "*all models are wrong but some are useful*"]
- publication and experimenter bias / confirmation bias — $H_0: \beta_i = 0$ is true
  - significant level, say 5% ⇒ keep studying, sooner or later, one will come up with a significant result (about 5% chance) even if one really does not exist.

  5%↑ Problem: significant results get published but not insignificant results
  - experimenter bias ⇒ many ways of analyzing data, experimenters may be tempted to pick the one that gives them the results they want/*expect* → 不想要則視而不見

  看不見的大猩猩

❖ **Reading**: Faraway (2005, 1st ed.), 3.8