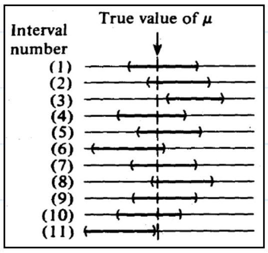
Confidence intervals and regions

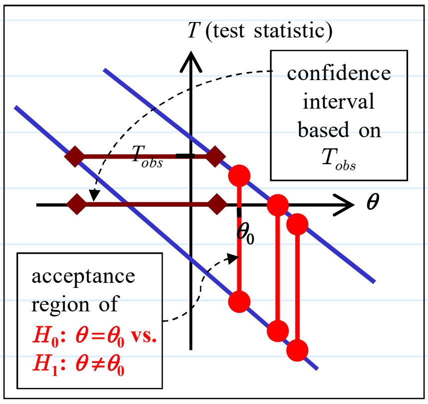
- **Q**: Why need interval/region estimation? What more information can it provide compared to point estimation?

  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?

- An interval/region estimation provides
  - plausible values for parameter
  - uncertainty in parameter estimator



  - information about its length and the values it covers may be helpful
  - information related to testing
- 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: $Y = X\beta + \varepsilon$, $\varepsilon \sim N(\mathbf{0}, \sigma^2 I)$; $\hat{\beta}$: OLS estimator $\Rightarrow \hat{\beta} \sim N(\beta, (X^TX)^{-1}\sigma^2)$
  - Confidence region for $A\beta$, where A is a full rank $d\times p$ matrix and $d \leq p$

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

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

and they are **independent**.

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

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

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

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

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

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

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

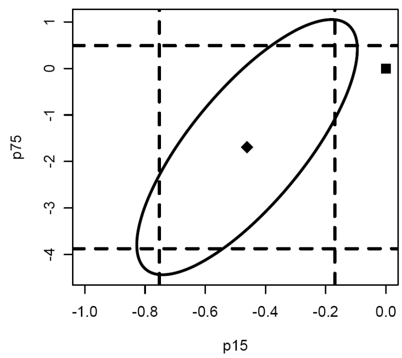
  - confidence region of β_i, β_j, i.e, $A = \begin{pmatrix} 0,\cdots,0,1,0,\cdots,0,0,0,\cdots,0 \\ 0,\cdots,0,0,0,\cdots,0,1,0,\cdots,0 \end{pmatrix}$

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

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

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

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

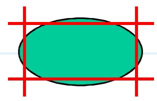


  - **Q1**: why the straight lines not tangential to the ellipse?

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

  - **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?
  - **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)
  - **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?
  - **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?
  - **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?



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

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

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

$$\Rightarrow (\hat\beta_i - \beta_i)/\left(\hat\sigma\sqrt{(X^TX)^{-1}_{ii}}\right) \sim t_{n-p} \qquad \Rightarrow \text{C.I.: } \hat\beta_i \pm t_{n-p}^{(\alpha/2)} \times \left(\hat\sigma\sqrt{(X^TX)^{-1}_{ii}}\right).$$

- confidence interval for prediction of mean response at x_0

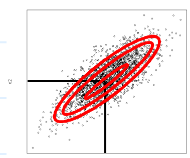
$$x_0^T\hat{\boldsymbol\beta} - x_0^T\boldsymbol\beta \sim N(\mathbf{0}, (x_0^T(X^TX)^{-1}x_0)\sigma^2) \Rightarrow (x_0^T\hat{\boldsymbol\beta} - x_0^T\boldsymbol\beta)/\left(\hat\sigma\sqrt{x_0^T(X^TX)^{-1}x_0}\right) \sim t_{n-p}$$

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

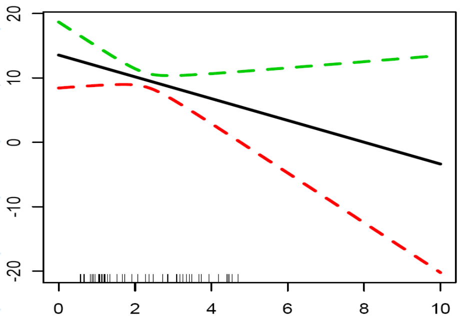
  - **Q**: for a given dataset and α, the length of the C.I. is related to x_0 only. What x_0 will cause a wider C.I.? **Ans**: x_0 that is away from the "center" of X
  - interpolation and extrapolation
    - interpolation: x_0 lie "within the range" of X
    - extrapolation: x_0 lie "outside the range" 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: $y=\beta_0+\beta_1 x+\varepsilon$)



  - **Q1**: why the confidence intervals get wider when we move away from the range of data?
  - **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?
  - **Q4**: does the plot represent a *simultaneous* confidence band for **all** prediction of mean response?

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

$$x_0^T\hat{\boldsymbol{\beta}} - (x_0^T\boldsymbol{\beta} + \varepsilon) \sim N(\mathbf{0}, (x_0^T(X^TX)^{-1}x_0 + 1)\sigma^2)$$

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

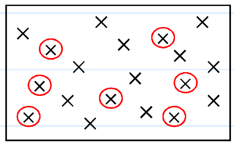
➢ a general form for confidence interval:

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

❖ **Reading**: Faraway (2005, 1st ed.), 3.4, 3.5
❖ **Futher reading**: D&S, 5.3, 5.4, 5.5

Sampling model

population



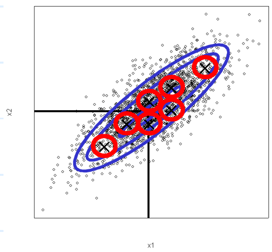
- experimental data vs. observational data

  It depends on whether we have *control* over predictors

  examples: yield of crop.

  experimental: fertilizer, …,

  observational: exposure, weather, …



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

Ans: experimental data: causation,

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

➢ **Q**: Is this model description, $Y=X\boldsymbol{\beta}+\varepsilon$, $\varepsilon \sim N(\mathbf{0}, \sigma^2 I)$, 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

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

example:

- 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?)

- the data should be generated using a "(simple) *random* sample" of the population so that they can be representative
- conditional distribution of multivariate normal: If

$$\mathbb{Z} = \begin{bmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \end{bmatrix} \sim \boldsymbol{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right),$$

then

$$\boldsymbol{Z}_1 \big| \boldsymbol{Z}_2 = \boldsymbol{z}_2 \sim \boldsymbol{N}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{z}_2 - \boldsymbol{\mu}_2),\ \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)$$

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

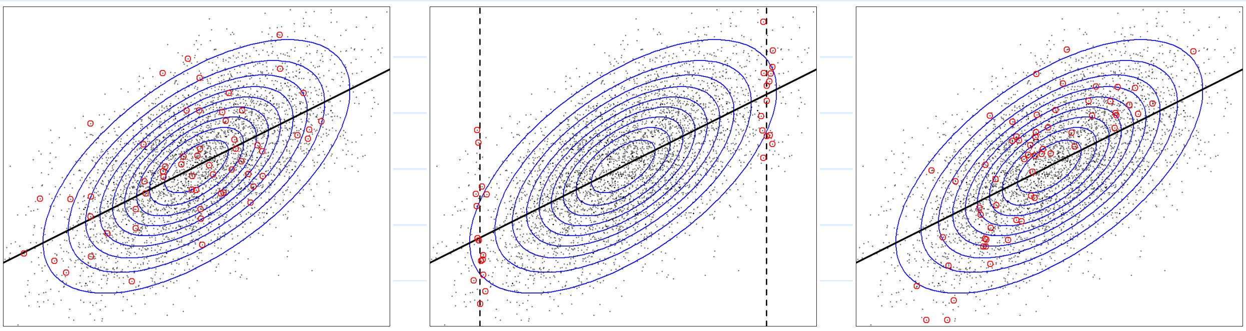
$$\begin{bmatrix} Y_i \\ \boldsymbol{X}_i \end{bmatrix} \sim \boldsymbol{N}\left(\begin{bmatrix} \mu_Y \\ \boldsymbol{\mu}_X \end{bmatrix}, \begin{bmatrix} \sigma_Y^2 & \boldsymbol{\Sigma}_{XY}^T \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{bmatrix}\right) \Rightarrow y_i | \boldsymbol{X}_i = \boldsymbol{x}_i \sim N((\mu_Y - \boldsymbol{\beta}^T\boldsymbol{\mu}_X) + \boldsymbol{\beta}^T\boldsymbol{x}_i, \sigma^2).$$

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

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

- **Q**: what information in these samples is proper?

(a) simple random sampling (b) simple random sampling given x (c) biased sampling



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

- **Q**: what will happen if the sample is not random?
  (i) biased sample
  (ii) sample of convenience
  (iii) sample = population
  these nonrandom samples can cause problems in the inference (e.g., R^2, LNp.3-18)

❖ **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?)

  [also, test H_0: $\beta_1=0$ (or c) not identical]

  an *exception*: when x_1 and x_2 are orthogonal

fitted model=model 1: $Y=X_1\beta_1+\varepsilon$

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

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

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?)

$$X^T X = \begin{pmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{pmatrix} = \begin{pmatrix} X_1^T X_1 & \boxed{0} \\ 0 & X_2^T X_2 \end{pmatrix}$$

$$\Rightarrow \quad (X^T X)^{-1} = \begin{pmatrix} (X_1^T X_1)^{-1} & \boxed{0} \\ 0 & (X_2^T X_2)^{-1} \end{pmatrix}$$

- Estimation: $\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T Y$, $\hat{\beta}_2 = (X_2^T X_2)^{-1} X_2^T Y$, 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).

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

- 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?

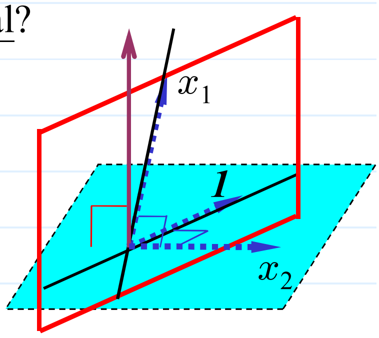
  model 1: ω_1: $y=\beta_0+\varepsilon$ vs. Ω_1: $y=\beta_0+\beta_1 x_1+\varepsilon$

  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?)

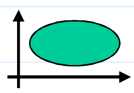
  but, $RSS_{\Omega 1} \neq RSS_{\Omega 2}$, and $df_{\Omega 1} \neq df_{\Omega 2}$, i.e., $\hat{\sigma}^2_{\Omega_1} \neq \hat{\sigma}^2_{\Omega_2}$.



  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 $\Rightarrow$ $E(\hat{\beta})=\beta+(X^T X)^{-1} X^T Z\gamma$

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

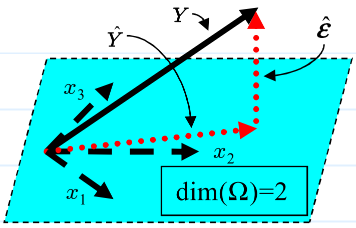
❖ **Reading**: Faraway (2005, 1st ed.), 3.6

❖ **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 $\hat{\varepsilon}=0$ $\Rightarrow$ cannot do testing or C.I.)
    - such problems can be avoided by paying attention.

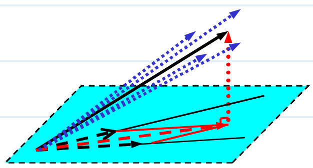
  - experimental data, e.g., two-sample case:

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

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

    $$\begin{pmatrix} y_1 \\ \cdots \\ y_n \\ y_{n+1} \\ \cdots \\ y_{m+n} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ & \cdots & \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ & \cdots & \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \cdots \\ \epsilon_n \\ \epsilon_{n+1} \\ \cdots \\ \epsilon_{m+n} \end{pmatrix} \Rightarrow X \text{ (or } \beta\text{) is unidentifiable}$$

    $\Rightarrow$ *over-parameterized*: some constraint must be imposed on $(\mu, \alpha_1, \alpha_2)$, say $\mu=0$ or $\alpha_1+\alpha_2=0$
- "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)
- 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.



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

Interpreting parameter estimates

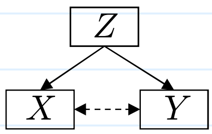
- **Q**: $Y = X\beta + \varepsilon$, what does $\hat{\beta}$ mean?

  Some matters needing attention about $\hat{\beta}$:
  - $\hat{\beta}$ have units [e.g., fuel consumption data, fitted model:
    fuel = 154.19 + (– 4.23)Tax + (0.47)Dlic + (–6.14)Income + (18.54)$\log_2$(Miles)]
  - 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
  - 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
    - 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

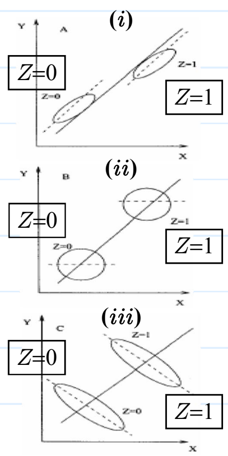
- Some interpretations of parameter estimates
  - a naive interpretation:
    "A unit increase in X_i will *cause* an average change of $\hat{\beta}_i$ in Y" $\Leftarrow$ causality statement
    [e.g., Y: annual income, and X: years of education]
  - **Q**: what if there exist lurking variables?
    [e.g., X: shoe size, Y: reading abilities, Z: age of child]
    $\Rightarrow$ 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: $E(Y \mid \mathbf{X}_1) = \mathbf{X}_1\beta_1$

      $$E(Y \mid \mathbf{X}_1) = \mathbf{X}_1\beta_1 + E(\mathbf{X}_2 \mid \mathbf{X}_1)\beta_2$$
      $$Var(Y \mid \mathbf{X}_1) = \sigma^2 + \beta_2^T\, Var(\mathbf{X}_2 \mid \mathbf{X}_1)\, \beta_2$$



  - 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?

  $\Rightarrow$ 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 $(g_{1,0},\ldots,g_{i,0},\ldots,g_{p-1,0}) \rightarrow \hat{y}_0$, observe $(g_{1,0},\ldots,g_{i,0}+1,\ldots,g_{p-1,0}) \rightarrow \hat{y}_0+\hat{\beta}_i$

- 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 $\rightarrow$ cause)?
  - include all relevant variables/effects $\Rightarrow$ however, even though you try hard to do so, the possibility of an unsuspected lurking variables will always exists
  - 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?
  - use non-statistical knowledge of the physical nature of the relationship $\Rightarrow$ 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.

    [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.]

  - 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 $\Rightarrow$ it takes decades of studies to go from association to causality

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

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

What can go wrong? many many things ...

- source and quality of the data (Q: how was the data collected?)
  - data may not be a random sample of the population. Situations such as biased sample, a sample of convenience, or sample=population
  - 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, …
  - 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*
- error component [we hope $\varepsilon \sim N(\mathbf{0}, \sigma^2 I)$]
  - ε may have unequal variance
  - ε may be correlated
  - ε may not be normally distributed
    - 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
    - for small sample sizes, bootstrap method offers a solution

- structural component [$E(Y)=X\beta$]
  - errors in X
  - serious collinearity in X
  - some inferences strongly rely on the choice of full model, $X\beta$ (example?)

    Q: where does the full model come from?

    1. physical theory may suggest a model --- wonderful but relatively uncommon
    2. experience from past data --- may help suggesting a reliable model
    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
- 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
  - 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. 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*

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