Website of my Linear Model course

http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/

Question

What is Statistics?

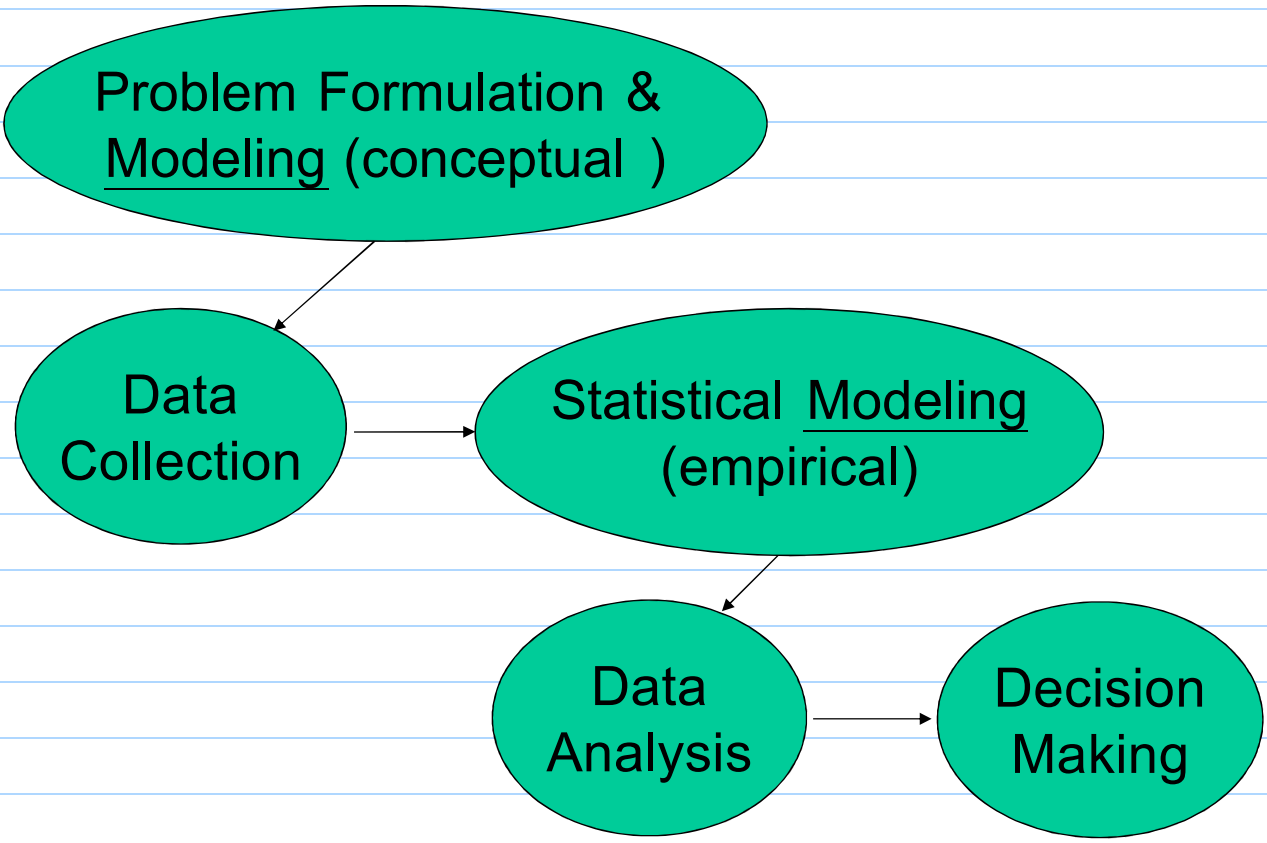
哈利波特	Real Life
占卜學	Statistics
崔老妮	Statisticians
水晶球	Data
未來的資訊	Information

aim of statistics: provide ***insight*** by means of *data*

NTHU STAT 5230, 2024, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

Basic Procedures of Statistics

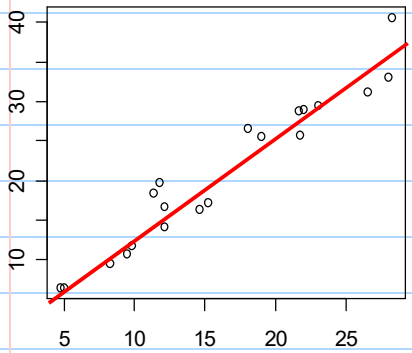
• Statistics divides the study of data into *five* steps:

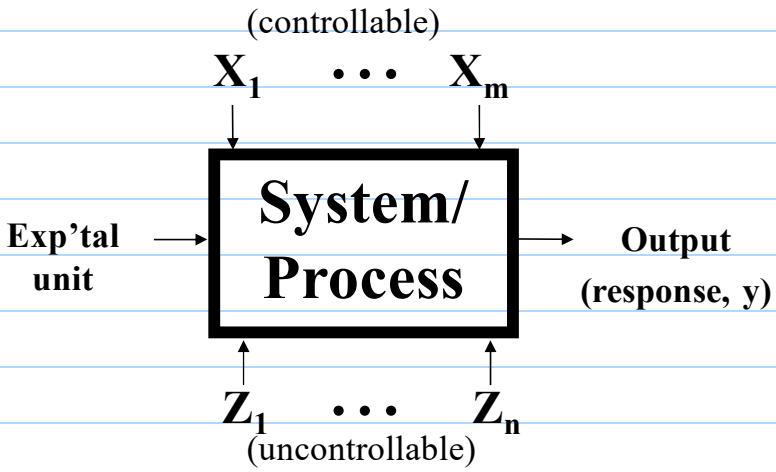


When to use regression analysis (linear model)?

- Regression: a statistical tool for investigating the "linearity relationship" between x and y.
  - ➢ causal relationship: examine the effects of x on y, i.e. how the changes in x *result in* the change in y
  - ➢ Even where no sensible causal relationship exists between x and y, we may wish to relate them by some sore of mathematical equation

Example





NTHU STAT 5230, 2024, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

- data type in regression analysis and some terminologies

  A {*response*, *output*, *dependent*} variable Y is modeled or explained by p effects/functions of m {*predictor*, *input*, *independent*, *regressor*} variables $X_1,\ldots,X_m$

  - ➢ Y: "approximately" continuous
  - ➢ $X_1,\ldots,X_m$: continuous and discrete (quantitative), categorical (qualitative) (**Q**: example?)
  - ➢ p=1, simple regression; p>1, multiple regression
  - ➢ $X_1,\ldots,X_m$

    all quantitative ⇒ multiple regression
    quantitative+qualitative ⇒ analysis of covariance
    all qualitative ⇒ analysis of variance (ANOVA)
  - ➢ more than one Y, multivariate regression

- linear model:

$$\begin{aligned}\mathrm{E}(Y|X_1,\ldots,X_m) &= \textstyle\sum_{i=0}^{p}\beta_i\cdot g_i \\ \mathrm{Var}(Y|X_1,\ldots,X_m) &= \mathrm{Var}(\epsilon)=\sigma^2\end{aligned}$$

$$Y=\underbrace{\sum_{i=0}^{p}\beta_i\cdot g_i(X_1,\ldots,X_m)}_{\text{deterministic component}\ \rightarrow\ \text{mean function}} + \underbrace{\epsilon}_{\substack{\text{error: random component}\ \rightarrow\ \text{variance function}}}$$

  - $X_1,\ldots,X_m$ are regarded as deterministic, i.e., no random phenomenon (when they are random variables, regard the linear model as conditional on $X_1,\ldots,X_m$)
  - $g_0(X_1,\ldots,X_m)$, ..., $g_p(X_1,\ldots,X_m)$: *known* functions of $X_1,\ldots,X_m$, called *effects*
  - (unknown) *parameters* β_0, ..., β_p enter linearly
  - variation due to random error only appears on y-axis

- Rationale: a general model for the relationship between Y and $X_1,\ldots, X_m$, is:
  $Y= f(X_1,\ldots,X_m) + \varepsilon$, where f is **unknown and arbitrary**

Note: # of parameters in f is infinite, usually do not have enough data to estimate f directly (globally), we have to assume that it has some more restricted form

  - local approximation of f may be achievable by a linear model

- **Note:** Because the predictors can be transformed and combined in any way, linear models are actually very flexible.

NTHU STAT 5230, 2024, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

Matrix representation

- Given the data matrix,

Y	X_1	X_2	...	X_m
y_1	x_{11}	x_{12}	...	x_{1m}
y_2	x_{21}	x_{22}	...	x_{2m}
...		...		
y_n	x_{n1}	x_{n2}	...	x_{nm}

a row: one group of observations

a column: one variable (response or predictor)

- We may write a linear model as follows (*functional form*): for $i= 1, 2, \ldots, n$,

$$y_i = \beta_0 + \beta_1\, g_1(x_{i1},\ldots,x_{im}) + \beta_2\, g_2(x_{i1},\ldots,x_{im}) + \ldots + \beta_{p-1}\, g_{p-1}(x_{i1},\ldots,x_{im}) + \varepsilon_i,$$

Y	**1**	$\boldsymbol{g}_1$	$\boldsymbol{g}_2$	...	$\boldsymbol{g}_{p-1}$
y_1	1	g_{11}	g_{12}	...	g_{1p-1}
y_2	1	g_{21}	g_{22}	...	g_{2p-1}
...			...		
y_n	1	g_{n1}	g_{n2}	...	g_{np-1}

a row: one group of observations

a column: response or effect

where $g_{ij}= g_j(x_{i1},\ldots,x_{im})$

  - the expression is (i) ugly notation (ii) conceptually awkward
  - matrix/vector notation is more elegant

Y	=	$\beta_0 \times 1$	+	$\beta_1 \times g_1$	+	$\beta_2 \times g_2$	+	...	+	$\beta_{p-1} \times g_{p-1}$	+	ε
y_1	=	1	+	g_{11}	+	g_{12}	+	...	+	g_{1p-1}	+	ε_1
y_2	=	1	+	g_{21}	+	g_{22}	+	...	+	g_{2p-1}	+	ε_2
...	=					...					+	...
y_n	=	1	+	g_{n1}	+	g_{n2}	+	...	+	g_{np-1}	+	ε_n

a row: one group of observations

a column: response or effect

- Matrix form of the linear model:

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

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & g_{11} & \cdots & g_{1p-1} \\ 1 & g_{21} & \cdots & g_{2p-1} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & g_{n1} & \cdots & g_{np-1} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_{p-1} \end{bmatrix}, \varepsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdots \\ \epsilon_n \end{bmatrix}.$$

and $E(\varepsilon) = 0$ and $var(\varepsilon) = \sigma^2 I$ (**Note**: the assumption that errors are normally distributed is not required at the estimation stage)

NTHU STAT 5230, 2024, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

Estimating β

- (ordinary) least square estimator
  - assume ε are (i) uncorrelated (ii) equal variance ($Var(\varepsilon) = \sigma^2 I$)
  - define the best $\hat{\beta}$ as that minimizes *sum of squared error*: $\varepsilon^T\varepsilon = \sum_{i=1}^n \epsilon_i^2$

$$\varepsilon^T\varepsilon = (Y - X\beta)^T(Y - X\beta) = Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta \qquad (*)$$

$\Rightarrow$ a second-order polynomial of β

  - One method of finding the minimizer is to differentiate (*) w.r.t. β and set the derivatives equal to zero

$$\Rightarrow \frac{\partial}{\partial\beta}\epsilon^T\epsilon = -2X^TY + 2X^TX\beta = 0$$

  - By calculus, $\hat{\beta}$ is the solution of

$$X^TX\beta = X^TY \quad \Leftarrow \text{ called } \textit{normal equation}$$

  - assume X^TX is non-singular,

$$\hat{\beta} = (X^TX)^{-1}X^TY \quad \Rightarrow \quad X\hat{\beta} = X(X^TX)^{-1}X^TY \equiv HY$$

  - *predicted values*: $\hat{Y} = X\hat{\beta} = HY$
  - *residuals*: $\hat{\varepsilon} = Y - X\hat{\beta} = Y - \hat{Y} = (I-H)Y$
  - *residual sum of squares* (*RSS*): $\hat{\varepsilon}^T\hat{\varepsilon} = [Y^T(I-H)^T][(I-H)Y] = Y^T(I-H)Y$

- ➢ OLS $\hat{\boldsymbol{\beta}}$ results from orthogonal projection, makes sense geometrically
- ➢ (FYI) if $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I)$, $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator
- ➢ Gauss-Markov thm states $\hat{\boldsymbol{\beta}}$ is BLUE ("Best" Linear Unbiased Estimator)

• mean and covariance matrix of OLS estimator $\hat{\boldsymbol{\beta}}$

$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ is a $p\times 1$ vector of random variables, so

- ➢ mean: $E(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{Y}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$ (i.e., *unbiased*)
- ➢ $Cov(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2 I\, \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2$ ($\Rightarrow$ irrelevant to $\boldsymbol{Y}$ and $\boldsymbol{\beta}$ $\Rightarrow$ **Note**: if we can control $\boldsymbol{X}$, can decide the var-cov matrix *before* observing $\boldsymbol{Y}$)

  Since $\hat{\boldsymbol{\beta}}$ is a random vectors, $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2$ is a variance-covariance matrix.
- ➢ $se(\hat{\beta}_i) = \sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{ii}}\,\hat{\sigma}$
- ➢ how to calculate the correlation between $\hat{\beta}_i$ and $\hat{\beta}_j$?

Estimating σ^2

- estimate σ^2 by $\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}/(n-p) = RSS/(n-p) \Rightarrow$ an unbiased estimator
- $\hat{\sigma}^2$ has the minimum variance among all quadratic unbiased estimators of σ^2
- $\hat{\sigma} = \sqrt{RSS/(n-p)}$
- (FYI) if $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I)$, the maximum likelihood estimator of σ^2 is $\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}/n = RSS/n$

NTHU STAT 5230, 2024, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

goodness-of-fit: how well does the model fit the data?

- R^2, coefficient of determination or **percentage of variance explained**

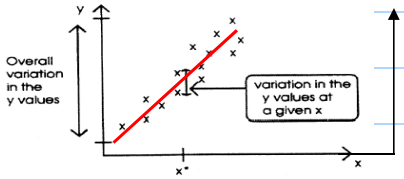
$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2} = \frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2} = \left(\frac{\sum(y_i-\bar{y})(\hat{y}_i-\bar{y})}{\sqrt{\sum(y_i-\bar{y})^2\sum(\hat{y}_i-\bar{y})^2}}\right)^2$$

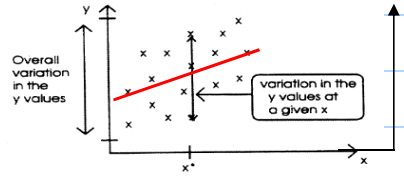
RSS is calculated from model with all independent variables,

TSS from model without any independent variables

Interpretation of R^2: "proportion of total variation in $\boldsymbol{y}$ that can be explained by the independent variables"

- ➢ R=correlation between $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$; for simple regression, R=correlation between $\boldsymbol{x}$ and $\boldsymbol{y}$ (from the geometry viewpoint, ...)
- ➢ $0 \le R^2 \le 1$, values closer to 1 indicate better fits. (what if $n \approx p$?)
- ➢ What is a good value of R^2? It depends.





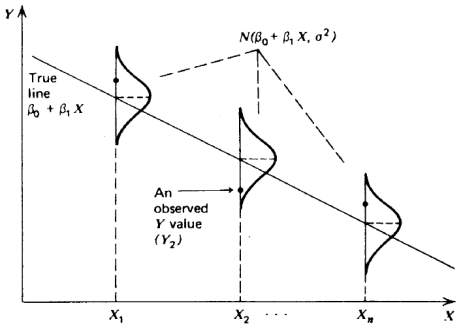
• alternative measure for goodness of fit: $\hat{\sigma}$

- ➢ it's related to standard error of estimates of $\boldsymbol{\beta}$ and prediction
- ➢ it's measured in the unit of the response (cf., R^2 is free of unit)

Normality assumption

- **Note**: up till now, **haven't assumed** any distributional form for $\boldsymbol{\varepsilon}$. If we want to perform any hypothesis tests or make any confidence intervals, we will need to do this. The usual assumption is:

$$\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$



  - ➢ model: $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$

$$\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I})$$

    - **Q**: what does the model describe? e.g.,

$$y_x = \beta_0 + \beta_1 x + \varepsilon_x,\ \varepsilon_x\text{'s} \sim \text{i.i.d. } N(0, \sigma^2)$$

$$\Rightarrow E(y_x) = \beta_0 + \beta_1 x$$

$$\Rightarrow\ y_x\text{'s are independent and } y_x \sim N(\beta_0 + \beta_1 x, \sigma^2) \text{ at } x = x_i,\ i = 1, \ldots, n.$$

- Some properties of linear models when $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$:
  - ➢ distribution of $\boldsymbol{Y}$ $[= \boldsymbol{X\beta} + \boldsymbol{\varepsilon}] \sim N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I})$
  - ➢ distribution of $\hat{\boldsymbol{\beta}}$ $[= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] \sim N(\boldsymbol{\beta}, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2)$
  - ➢ distribution of $\hat{\boldsymbol{\varepsilon}}$ $[= (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}] \sim N(\boldsymbol{0}, (\boldsymbol{I}-\boldsymbol{H})\sigma^2)$, which has a singular covariance matrix $\boldsymbol{I}-\boldsymbol{H}$ with rank $n-p$ (Note: $\dim(\hat{\boldsymbol{\varepsilon}}) = n-p$)

NTHU STAT 5230, 2024, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

  - ➢ distribution of RSS $[= (n-p)\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}] \sim \sigma^2\chi^2_{n-p}$
  - ➢ distribution of $\hat{\boldsymbol{Y}}$ $[= \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{HY}] \sim N(\boldsymbol{X\beta}, \boldsymbol{H}\sigma^2)$, which has a singular covariance matrix with rank p (Note: $\dim(\hat{\boldsymbol{Y}}) = p$)
  - ➢ $\hat{\boldsymbol{\beta}}$ is independent of $\hat{\sigma}^2$ (Note: $\text{cov}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}, (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}) = \boldsymbol{0}$)
  - ➢ $\hat{\boldsymbol{Y}}$ is independent of $\hat{\boldsymbol{\varepsilon}}$ (Note: $\text{cov}(\boldsymbol{HY}, (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}) = \boldsymbol{0}$)
  - ➢ distribution of prediction for a new set of predictors, $\boldsymbol{x}_0 = (g_1(x_{10}, \ldots, x_{m0}), \ldots, g_p(x_{10}, \ldots, x_{m0}))^T$

$$\text{model: } y = \sum_{j=1}^{p} \beta_j \cdot g_j(x_1, \ldots, x_m) + \epsilon$$

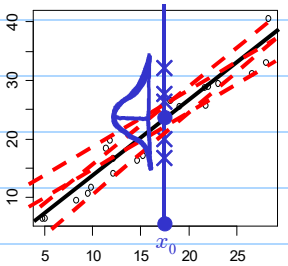
cf. fitted model: $\hat{\beta}_j$, $\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}$

    - mean response v.s. future observation (**Q**: what different?)
      - Example: average yield when $\boldsymbol{x} = \boldsymbol{x}_0$? and tomorrow's yield when $\boldsymbol{x} = \boldsymbol{x}_0$?
      - same predicted value $\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}$, but different distributions
    - distribution of prediction error for mean response at $\boldsymbol{x}_0$

$$\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} - \boldsymbol{x}_0^T\boldsymbol{\beta} \sim N(\boldsymbol{0}, (\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0)\sigma^2)$$

    - distribution of prediction error for future observations at $\boldsymbol{x}_0$

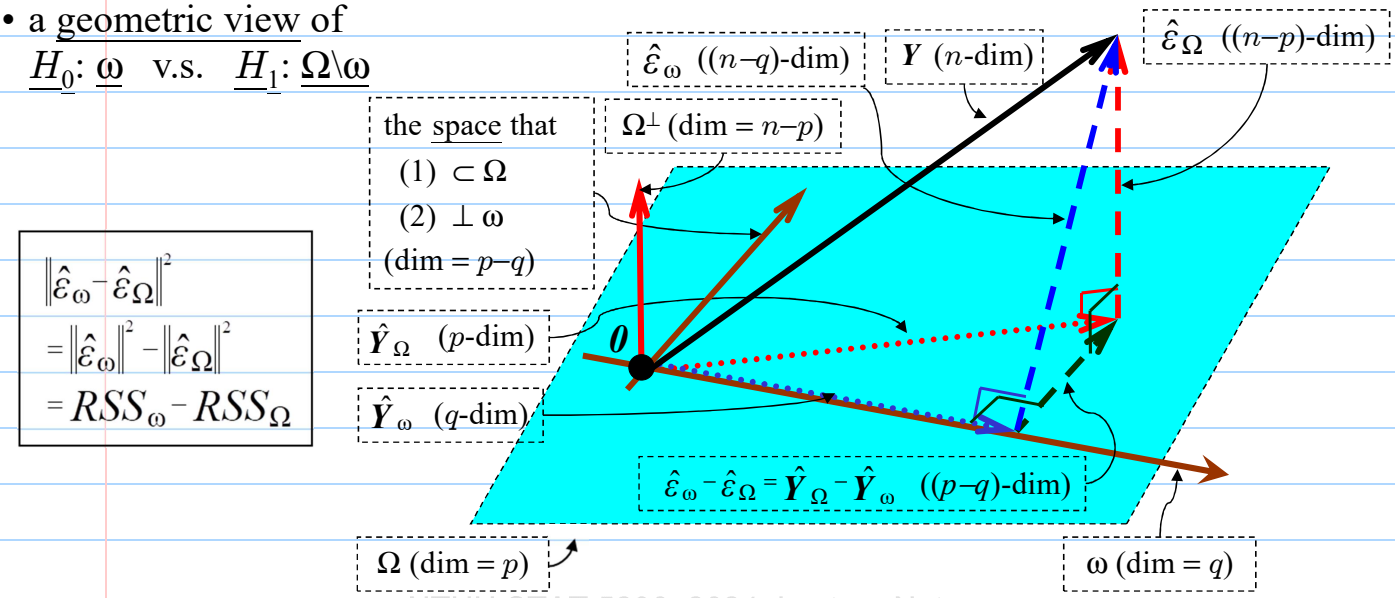
$$\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} - (\boldsymbol{x}_0^T\boldsymbol{\beta} + \varepsilon) \sim N(\boldsymbol{0}, (\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0 + 1)\sigma^2)$$



hypothesis testings (for β)

- formulation of hypothesis testing from the view of **comparing models** (**model spaces**)
  - ➢ a model space ≡ the space spanned by columns of some X (model matrix)
  - ➢ consider a large model space, Ω, and a smaller model space, ω, where $\omega \subset \Omega$ (i.e., ω represents a subset/subspace of Ω). Suppose dimension (# of parameters) of Ω is p and $\dim(\omega)=q$, where $p>q$.
  - ➢ to answer "which of the model spaces is more adequate" in statistical language ⇒ perform the test H_0: ω ($A\beta=c$) v.s. H_1: Ω\ω

- a geometric view of H_0: ω v.s. H_1: Ω\ω



NTHU STAT 5230, 2024, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

- ➢ suppose dimension (# of parameters) of Ω is p and $\dim(\omega)=q$.
  Under the null H_0: ω,

$$(RSS_\omega - RSS_\Omega)/\sigma^2 \sim \chi^2_{p-q},$$

$$RSS_\Omega/\sigma^2 \sim \chi^2_{n-p}$$

and they are *independent*.

So, we have $$F = \frac{(RSS_\omega - RSS_\Omega)/(p-q)}{RSS_\Omega/(n-p)} \sim F_{p-q,n-p}.$$

Therefore, reject if $F > F_{p-q,n-p}^{(\alpha)}$ (usually check if p-value $< \alpha$)

- ➢ **General form**: because the degree of freedom of residuals in a model is the number of observations minus the number of parameters (in β), this test statistics can be written as:

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

where $df_\omega = \dim(\omega^\perp) = n-q$ and $df_\Omega = \dim(\Omega^\perp) = n-p$.

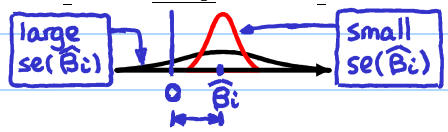
- ➢ The test is widely used in regression and ANOVA. The beauty of this approach is you only need to know the general form.
- ➢ This test is the likelihood-ratio test.

- Example 1: test of all predictors
  - Q: are any of the predictors g_i's useful in predicting the response?
    - Ω: $y = \beta_0 + \beta_1 g_1 + \cdots + \beta_{p-1} g_{p-1} + \epsilon$, $\dim(\Omega) = p$, $df_\Omega = n-p$
    - ω: $y = \beta_0 + \epsilon$, $\dim(\omega) = 1$, $df_\omega = n-1$
    - H_0: $\beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ H_1: at least one of $\beta_1, \cdots, \beta_{p-1}$ is not zero
    - RSS_Ω: $\hat{\epsilon}_\Omega^T \hat{\epsilon}_\Omega = \sum_{i=1}^n (y_i - \hat{y}_{i,\Omega})^2$ RSS_ω: $(Y-\bar{Y}\mathbb{1})^T(Y-\bar{Y}\mathbb{1}) = \sum_{i=1}^n (y_i - \bar{y})^2$
    - (the overall F) $F = \frac{(RSS_\omega - RSS_\Omega)/(df_\omega - df_\Omega)}{RSS_\Omega/df_\Omega} = \frac{[\Sigma(y_i-\bar{y})^2 - \Sigma(y_i-\hat{y}_i)^2]/(p-1)}{\Sigma(y_i-\hat{y}_i)^2/n-p}$

they are functionally related. cf. $R^2 = 1 - \frac{RSS_\Omega}{RSS_\omega} = 1 - \frac{1}{(1 + \frac{p-1}{n-p} F)}$

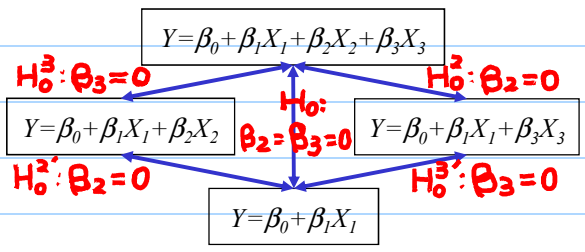
- Example 2: testing just one predictor
  - Q: Can one particular predictor, say $g_i(x)$, be dropped from the model?
    - Ω: $y = \beta_0 + \cdots + \beta_i g_i + \cdots + \beta_{p-1} g_{p-1} + \epsilon$, $\dim(\Omega) = p$, $df_\Omega = n-p$
    - ω: $y = \beta_0 + \cdots + \cancel{\beta_i g_i} + \cdots + \beta_{p-1} g_{p-1} + \epsilon$, $\dim(\omega) = p-1$, $df_\omega = n-p+1$
    - H_0: $\beta_i = 0$ $(\beta_j \in \mathbb{R}, \text{for } j \neq i)$ H_1: $\beta_i \neq 0$ $(\beta_j \in \mathbb{R}, \text{for } j \neq i)$
    - $F = [(RSS_\omega - RSS_\Omega)/(df_\omega - df_\Omega)]/(RSS_\Omega/df_\Omega) \sim F_{df_\omega - df_\Omega, df_\Omega}$ (1 / $\hat{\sigma}^2_\Omega$ / 1)
  - alternative method t-test: $t_i = \hat{\beta}_i / \text{se}(\hat{\beta}_i) \sim t_{n-p}$ [Note. $t_i^2 \sim F_{1,n-p}$, and $t_i^2 = F$]

$t_i^2 = \left(\frac{\hat{\beta}_{i,\Omega}}{\sqrt{(X_\Omega^T X_\Omega)^{-1}_{ii}}\,\hat{\sigma}_\Omega}\right)^2 = F \Leftarrow RSS_\omega - RSS_\Omega = \left(\frac{\hat{\beta}_{i,\Omega}}{\sqrt{(X_\Omega^T X_\Omega)^{-1}_{ii}}}\right)^2$ (exercise) (1/σ, 1/σ)



NTHU STAT 5230, 2024, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

- (sequential) ANOVA (A: 3 levels; B: 4 levels)
  - anova($y \sim 1 + A + B + A:B$)
    1) test ω_1:model 1 ($y \sim 1$) against Ω_1:model 2 ($y \sim 1 + A$) [$df_\omega - df_\Omega = 2$]
    2) test ω_2:model 2 ($y \sim 1 + A$) against Ω_2:model 4 ($y \sim 1 + A + B$) [$df_\omega - df_\Omega = 3$]
    3) test ω_3:model 4 ($y \sim 1 + A + B$) against Ω_3:model 5 ($y \sim 1 + A + B + A:B$) [$df_\omega - df_\Omega = 6$]
    - $$F = \frac{(RSS_\omega - RSS_\Omega)/(df_\omega - df_\Omega)}{RSS_{\text{model 5}}/df_{\text{model 5}}} \sim F_{df_\omega - df_\Omega, df_{\text{model 5}}}$$
    - invariant to the choice of dummy variables if they generate same ω and Ω
  - ANOVA could have different results when the order of effect sequence is changed, e.g., anova($y \sim 1 + B + A + A:B$):
    α) test ω_1:model 1 ($y \sim 1$) against Ω_1:model 3 ($y \sim 1 + B$) [$df_\omega - df_\Omega = 3$]
    β) test ω_2:model 3 ($y \sim 1 + B$) against Ω_2:model 4 ($y \sim 1 + B + A$) [$df_\omega - df_\Omega = 2$]
    χ) test ω_3:model 4 ($y \sim 1 + B + A$) against Ω_3:model 5 ($y \sim 1 + B + A + A:B$) [$df_\omega - df_\Omega = 6$]



Confidence intervals and regions

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

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

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

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

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

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

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

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

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

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

NTHU STAT 5230, 2024, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

  - 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

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

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

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

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

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

  - a general form for confidence interval:

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

Interpreting parameter estimates

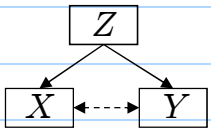
- Q: $Y = X\beta + \varepsilon$, what does $\hat{\beta}$ mean?

  Some matters needing attention about $\hat{\beta}$:

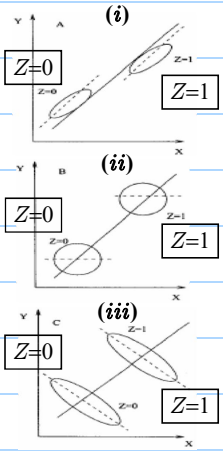
  - $\hat{\beta}$ have units [e.g., fuel consumption data, fitted model:
    fuel = 154.19 + (– 4.23)Tax + (0.47)Dlic + (–6.14)Income + (18.54)$\log_2$(Miles)]
  - sign of $\hat{\beta}$: direction of the relationship between the term and the response
  - interpretation of estimated value (see next two 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

NTHU STAT 5230, 2024, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

- 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 $E(Y)=\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_1 X_2=\beta_0+(\beta_1+\beta_3 X_2)X_1+\beta_2 X_2$
- 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. 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 $\hat{\boldsymbol{\beta}}$:

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
- directly interpretable and success may be measured in future
- dangers of extrapolation, be cautious when $\boldsymbol{x}_0$ is outside the range of $\boldsymbol{X}$

NTHU STAT 5230, 2024, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

Mean structure

- idea: data are generated from an underlying system, which is assumed to have the form: $y = f(x_1, \ldots, x_m) + \varepsilon$, where f is ***unknown***.
- regression *approximates* the mean structure f by a linear combination of (known) *base functions* $g_i(x_1, \ldots, x_m)$'s, $i=1, \ldots, p$, i.e.,

$$f \longleftarrow \sum_{i=1}^{p} \beta_i \cdot g_i(x_1, \ldots, x_m)$$

➢ when the structure of f is simple and almost linear, it can be approximated by a simple structure with fewer terms, e.g.,

$$E(y) = f \approxeq \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

- **Q**: nature is simple?
- **Q**: are there sufficient data to support/fit a complex model?

➢ when f is complex and non-linear ⇒ need more terms to get a good approximation

- more parameters, need more degrees of freedom, i.e., more data
- e.g., 2 levels, only linear effects; 3 levels, linear and quadratic effects
- **Q**: what other complex models?

- base functions for quantitative and qualitative predictors x_i's are defined in different ways

Polynomial regression

- one predictor case: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_d x^d + \varepsilon$
- two predictors x_1, x_2 case:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon \quad (d=2, 2^{\text{nd}}\text{-order model})$$

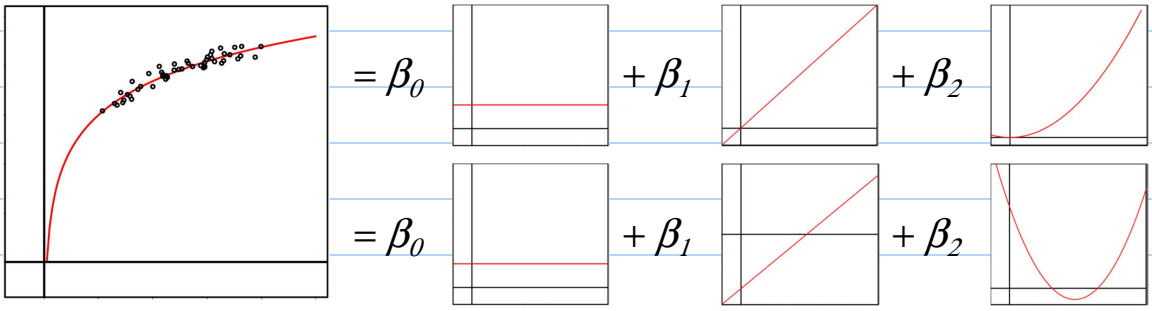
  - the cross-product term $x_1 x_2$ can be interpreted as an "interaction" effect, e.g.,

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2, \text{ where } x_1, x_2 \in \{-1, 1\}$$
$$x_1 = +1 \Longrightarrow E(y) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) x_2$$
$$x_1 = -1 \Longrightarrow E(y) = (\beta_0 - \beta_1) + (\beta_2 - \beta_3) x_2$$

  - models for more predictors can be similarly extended

$$y = \beta_0 + \sum_{i=1}^{m} \beta_{1,i} x_i + \sum_{i=1}^{m} \beta_{2,i} x_i^2 + \sum_{1 \le i < j \le m} \beta_{3,ij} x_i x_j + \epsilon$$

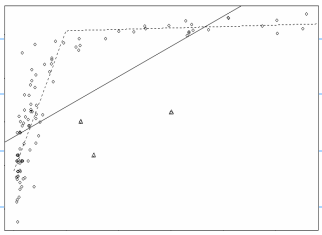
- orthogonal polynomials
  - polynomial terms can cause numerical instability (especially when d large) and collinearity
  - example: 2^{nd}-order model



NTHU STAT 5230, 2024, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

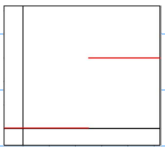
broken stick (line) regression (segmented regression)

- Recall. polynomial regression: suitable for smooth mean structure, but cannot capture local abrupt change (example?)

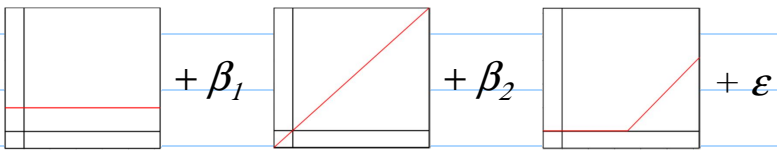


- suppose the break occurs at the *known* value c, define the base function (where c is called a *knot*):

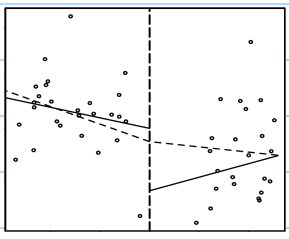
$$d_c(x) = \begin{cases} 1, & \text{if } x > c, \\ 0, & \text{if } x \le c. \end{cases}$$



- model: $y = \beta_0 + \beta_1 x + \beta_2 (x - c) d_c(x) + \varepsilon$



$$E(y) = \begin{cases} \beta_0 + \beta_1 x, & \text{if } x \le c, \\ (\beta_0 - \beta_2 c) + (\beta_1 + \beta_2) x, & \text{if } x > c, \end{cases}$$



  - the two lines meet at $c \Rightarrow$ continuous fit
  - notice only 3 parameters in the model $\Rightarrow$ one degree of freedom is saved because of the continuity restriction

dummy variable (indicator variable, coding)

- categorical (qualitative) predictors
  - nominal v.s. ordinal
  - examples: male/female, treatment/control, eye colors, blocks, ...
  - qualitative in nature:
    - values are symbols, no quantitative meaning
    - no value exist between categories
  - **Q**: what properties can we explore for qualitative predictor?

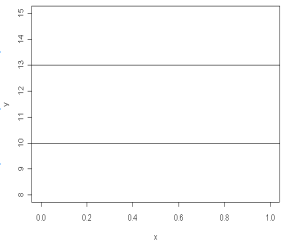
    category $i \to y_{ij}$, $\mu_i = E(y_{ij}) \Rightarrow$ can only study *difference* between μ_i's (cf., quantitative predictor)
  - **Q**: how to fit these predictors into the format of linear regression model $\boldsymbol{Y} = \boldsymbol{X\beta}+\boldsymbol{\varepsilon}$? $\Rightarrow$ Ans: dummy variables
- one dichotomous predictor: two categories
  - for a dichotomous predictor C with two categories c_1 and c_2, define a dummy variable d:
$$d(C) = \begin{cases} 0, & \text{if } C = c_1, \\ 1, & \text{if } C = c_2. \end{cases}$$
  - for a data set with response y, one quantitative predictor x, and one qualitative predictor C (dummy variable d), possible models are:

    model 1: $y = \beta_0 + \beta_1 d + \varepsilon$, model 2: $y = \beta_0 + \beta_1 x + \varepsilon$,
    model 3: $y = \beta_0 + \beta_1 d + \beta_2 x + \varepsilon$, model 4: $y = \beta_0 + \beta_1 x + \beta_2 xd + \varepsilon$,
    model 5: $y = \beta_0 + \beta_1 d + \beta_2 x + \beta_3 xd + \varepsilon$

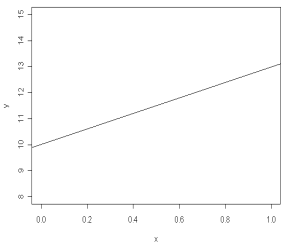
NTHU STAT 5230, 2024, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

- **Q**: how to interpret β_i's in models 1~5?
  - model 1: $y = \beta_0 + \beta_1 d + \epsilon$

$$\begin{array}{lll} C = c_1: & \mu_1 = E(y|d=0) = \beta_0 & \Rightarrow \quad \beta_0 = \mu_1 \\ C = c_2: & \mu_2 = E(y|d=1) = \beta_0 + \beta_1 & \Rightarrow \quad \beta_1 = \mu_2 - \mu_1 \end{array}$$



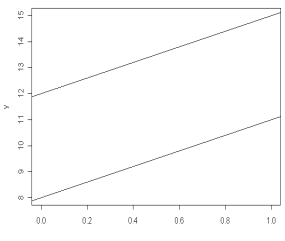
  - model 2: $y = \beta_0 + \beta_1 x + \epsilon$



  - model 3: $y = \beta_0 + \beta_1 d + \beta_2 x + \epsilon$

$$\begin{array}{ll} C = c_1: & \mu_{1,x} = E(y|d=0,x) = \beta_0 + \beta_2 x \\ C = c_2: & \mu_{2,x} = E(y|d=1,x) = (\beta_0 + \beta_1) + \beta_2 x \end{array}$$

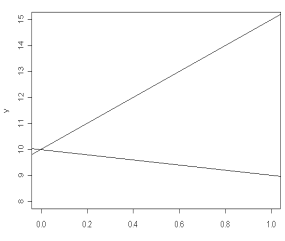
$$\begin{array}{lll} & \beta_0 = \mu_{1,0} & \text{(intercept in } c_1 \text{ group)} \\ \Rightarrow & \beta_1 = \mu_{2,x} - \mu_{1,x} & \text{(difference of intercepts)} \\ & \beta_2 = \text{slope} & \text{(same slope in two categories)} \end{array}$$



  - model 4: $y = \beta_0 + \beta_1 x + \beta_2 (d \cdot x) + \epsilon$

$$\begin{array}{ll} C = c_1: & \mu_{1,x} = E(y|d=0,x) = \beta_0 + \beta_1 x \\ C = c_2: & \mu_{2,x} = E(y|d=1,x) = \beta_0 + (\beta_1 + \beta_2) x \end{array}$$

$$\begin{array}{lll} & \beta_0 = \mu_{1,0} = \mu_{2,0} & \text{(same intercept in two categories)} \\ \Rightarrow & \beta_1 = \text{slope of category } c_1 & \\ & \beta_2 = \text{difference in slopes} & \end{array}$$



- model 5: $y = \beta_0 + \beta_1 d + \beta_2 x + \beta_3 (d \cdot x) + \epsilon$

$C = c_1$: $\mu_{1,x} = E(y|d=0, x) = \beta_0 + \beta_2 x$

$C = c_2$: $\mu_{2,x} = E(y|d=1, x) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) x$

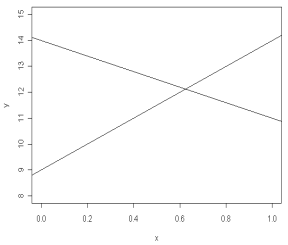
$\Rightarrow$

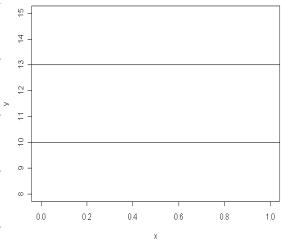
β_0 = $\mu_{1,0}$ (intercept of category c_1) ← reference

β_2 = slope of category c_1 ← reference

β_1 = difference in intercepts

β_3 = difference in slopes





➢ alternative coding of dummy variable (better orthogonality)

$$d(C) = \begin{cases} -1, & \text{if } C = c_1, \\ 1, & \text{if } C = c_2. \end{cases}$$

Q: how to interpret β_i's in models 1~5 under this coding?

- model 1: $y = \beta_0 + \beta_1 d + \epsilon$

$C = c_1$: $\mu_1 = E(y|d=-1) = \beta_0 - \beta_1$

$C = c_2$: $\mu_2 = E(y|d=1) = \beta_0 + \beta_1$

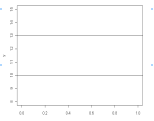
$\Rightarrow$ $\beta_0 = (\mu_1 + \mu_2)/2$, $\beta_1 = (\mu_2 - \mu_1)/2$

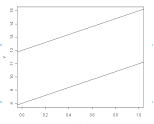
➢ analysis strategy: start from the full model (model 5) if there are enough degrees of freedom, and then test if some terms can be eliminated

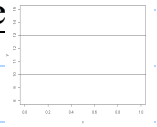
➢ identical methodology applies for more than 2 categories and more quantitative predictors

NTHU STAT 5230, 2024, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

➢ *ANalysis of COVAriance*: testing model 3 (Ω) against model 2 (ω) (more than 2 categories and more quantitative predictors is allowed). The quantitative predictor is called *covariate* and is expected to have the same effect in all categories. The difference between categories is assumed to be an additive effect.







• one polytomous predictor: more than two categories

➢ for k categories, $k-1$ dummy variables are needed to depict the difference between categories (one parameter is used to represent constant term)

➢ various coding of dummy variables: 4 categories c_1, c_2, c_3, c_4 example

treatment coding

	d_1	d_2	d_3
c_1	0	0	0
c_2	1	0	0
c_3	0	1	0
c_4	0	0	1

Helmert coding

	d_1	d_2	d_3
c_1	−1	−1	−1
c_2	1	−1	−1
c_3	0	2	−1
c_4	0	0	3

sum coding

	d_1	d_2	d_3
c_1	−1	−1	−1
c_2	1	0	0
c_3	0	1	0
c_4	0	0	1

➢ consider the model: $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \epsilon$

- properties of treatment coding:

$C = c_1$: $\mu_1 = E(y|d_1=0, d_2=0, d_3=0) = \beta_0$

$C = c_2$: $\mu_2 = E(y|d_1=1, d_2=0, d_3=0) = \beta_0 + \beta_1$

$C = c_3$: $\mu_3 = E(y|d_1=0, d_2=1, d_3=0) = \beta_0 + \beta_2$

$C = c_4$: $\mu_4 = E(y|d_1=0, d_2=0, d_3=1) = \beta_0 + \beta_3$

$\Rightarrow$ $\beta_0 = \mu_1$, $\beta_1 = \mu_2 - \mu_1$, $\beta_2 = \mu_3 - \mu_1$, $\beta_3 = \mu_4 - \mu_1$

- treats c_1 as a reference
- it is convenient if a "standard" categories exists
- d_1, d_2, and d_3 are mutually orthogonal, but not orthogonal to constant term

- properties of Helmert coding: $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \epsilon$

$$
\begin{aligned}
C = c_1: &\quad \mu_1 = E(y|d_1 = -1, d_2 = -1, d_3 = -1) = \beta_0 - \beta_1 - \beta_2 - \beta_3 \\
C = c_2: &\quad \mu_2 = E(y|d_1 = 1, d_2 = -1, d_3 = -1) = \beta_0 + \beta_1 - \beta_2 - \beta_3 \\
C = c_3: &\quad \mu_3 = E(y|d_1 = 0, d_2 = 2, d_3 = -1) = \beta_0 + 2\beta_2 - \beta_3 \\
C = c_4: &\quad \mu_4 = E(y|d_1 = 0, d_2 = 0, d_3 = 3) = \beta_0 + 3\beta_3
\end{aligned}
$$

$$
\Rightarrow \quad
\begin{aligned}
\beta_0 &= \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4} \equiv \bar{\mu} \\
\beta_1 &= \frac{\mu_2 - \mu_1}{2} \\
\beta_2 &= \frac{\mu_3 - ((\mu_1 + \mu_2)/2)}{3} \\
\beta_3 &= \frac{\mu_4 - ((\mu_1 + \mu_2 + \mu_3)/3)}{4}
\end{aligned}
$$

- constant term, d_1, d_2, and d_3 are orthogonal when there are equal # of observations in each categories
- hard to interpret parameters
- may suitable for ordinal qualitative predictor

NTHU STAT 5230, 2024, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

- properties of sum coding: $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \epsilon$

$$
\begin{aligned}
C = c_1: &\quad \mu_1 = E(y|d_1 = -1, d_2 = -1, d_3 = -1) = \beta_0 - \beta_1 - \beta_2 - \beta_3 \\
C = c_2: &\quad \mu_2 = E(y|d_1 = 1, d_2 = 0, d_3 = 0) = \beta_0 + \beta_1 \\
C = c_3: &\quad \mu_3 = E(y|d_1 = 0, d_2 = 1, d_3 = 0) = \beta_0 + \beta_2 \\
C = c_4: &\quad \mu_4 = E(y|d_1 = 0, d_2 = 0, d_3 = 1) = \beta_0 + \beta_3
\end{aligned}
$$

$$
\Rightarrow \quad
\begin{aligned}
\beta_0 &= \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4} \equiv \bar{\mu} \\
\beta_1 &= \mu_2 - \bar{\mu} \\
\beta_2 &= \mu_3 - \bar{\mu} \\
\beta_3 &= \mu_4 - \bar{\mu}
\end{aligned}
$$

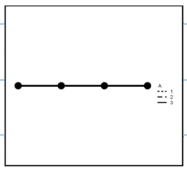
- β_0 represent overall mean
- compare each category with the overall mean
- lesser orthogonal

➢ Note: the choice of coding does not affect the R^2, $\hat{\sigma}$ and overall F-test (to test H_0: $\beta_1=\beta_2=\beta_3= 0$, the three codings have same ω and Ω)

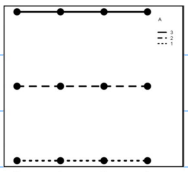
➢ the overall F-test is one-way ANOVA (ANalysis Of VAriance)

➢ **Q**: how to work with quantitative predictors? ⇒ identical methodology as in 2 categories case. **Q**: how to interpret parameters in the case?

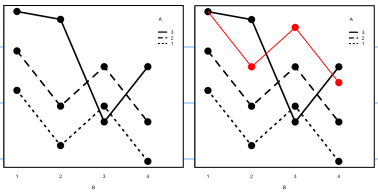
- two qualitative predictors (say, A: 3 categories a_1, a_2, a_3; B: 4 categories, b_1, b_2, b_3, b_4)
  - number of different category combinations = 3×4 =12,
    denote their means as μ_{ij}, i=1, 2, 3 and j=1, 2, 3, 4, i.e.,
    $$y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \quad k = 1, 2, \ldots, n_{ij},$$
    n_{ij} = number of observations in category $A=a_i$ and $B=b_j$

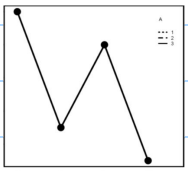


  - Q: how to depict the difference between μ_{ij}'s?
    consider the following linear models:
    - model 1: $E(y_{ijk}) = \beta_0$
    - model 2: $E(y_{ijk}) = \beta_0 + \beta_1 d_1^A + \beta_2 d_2^A$
    - model 3: $E(y_{ijk}) = \beta_0 + \beta_1 d_1^B + \beta_2 d_2^B + \beta_3 d_3^B$
    - model 4: $E(y_{ijk}) = \beta_0 + \beta_1 d_1^A + \beta_2 d_2^A + \beta_3 d_1^B + \beta_4 d_2^B + \beta_5 d_3^B$

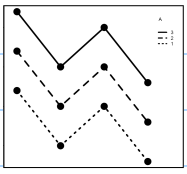


    Q: how to perform interaction coding? what is interaction?





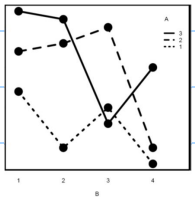
interaction plot: replace μ_{ij}'s by cell means
$\bar{y}_{ij} = \sum_{k=1}^{n_{ij}} y_{ijk}/n_{ij}$



    - model 5:

2-factor interaction

$$E(y_{ijk}) = \beta_0 + \beta_1 d_1^A + \beta_2 d_2^A + \beta_3 d_1^B + \beta_4 d_2^B + \beta_5 d_3^B + \sum_{i=1}^{2}\sum_{j=1}^{3} \beta_{ij} d_{ij}$$



# of parameters: 1 + 2 + 3 + 6 = 12

NTHU STAT 5230, 2024, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

- identical methodology applies for more qualitative (3-factor interaction, 4-factor interaction, …) and quantitative predictors (similar modeling to what in LNp.1-26~27)

Transformation

- transformation of response
  - Box-Cox transformation family: $t_\lambda(y) = \begin{cases} (y^\lambda - 1)/\lambda, & \text{if } \lambda \neq 0, \\ \log(y), & \text{if } \lambda = 0. \end{cases}$
    - $t_\lambda(y)$ is continuous in λ: for fixed $y>0$,
      $$\lim_{\lambda\to 0} t_\lambda(y) = \lim_{\lambda\to 0} (y^\lambda - 1)/\lambda = \lim_{\lambda\to 0} (y^\lambda \log(y))/1 = \log(y)$$
    - λ=1 ⇒ no transformation, λ=0 ⇒ log, $\lambda \neq 0$ or 1 ⇒ power transformation
    - model: $t_\lambda(y) = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$
      - parameters: λ, $\boldsymbol{\beta}$, σ
      - can write down likelihood for estimation and testing of λ
      - choice of transformation becomes a estimation/test problem
    - the log-likelihood of λ is
      $$L(\lambda) = \underbrace{(-n/2)\log(RSS_\lambda/n)}_{\text{goodness of fit}} + \underbrace{(\lambda - 1)\Sigma \log(y_i)}_{\text{adjustment}},$$
      where RSS_λ = residual sum of square when using $t_\lambda(y)$ as response, i.e.,
      $$RSS_\lambda = [t_\lambda(y)]^T (I - H) t_\lambda(y)$$

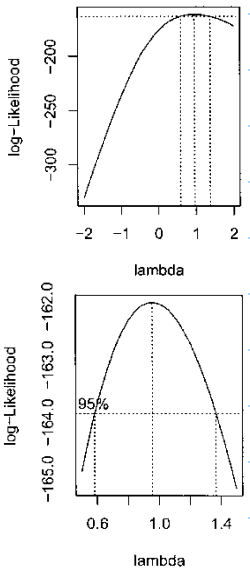
- estimation of λ: choose λ to fit data well using *maximum likelihood*.
  - can compute $L(\lambda)$ for various values of λ and compute $\hat{\lambda}$ exactly to maximize $L(\lambda)$
  - but usually $\hat{\lambda}$ is not a nice round number, e.g., $\hat{\lambda} = -0.17$. It would be hard to explain what this new response means.
  - to avoid this, maximize $L(\lambda)$ over a grid of values, such as $\{-2, -1, -1/2, 0, 1/2, 1, 2\}$. This helps with interpretation.
  - for $\hat{\lambda}$ outside $[-2, 2]$, pay more attention on whether such transformation is required
  - Q: why not just minimize RSS_λ to estimate λ?
- test of λ: is the transformation really necessary?
  - we can answer the question form a C.I. for λ
  - *likelihood ratio test* (H_0: $\lambda = \lambda_0$ vs. H_A: $\lambda \neq \lambda_0$):

$$-2[L(\lambda_0) - L(\hat{\lambda})] \sim \chi_1^2 \text{ under } H_0$$

  - a $100(1-\alpha)\%$ C.I. for λ can be formed by:

$$\{\lambda \mid L(\lambda) > L(\hat{\lambda}) - (1/2)\chi_1^2(1-\alpha)\}$$

  - is $\lambda=1$ in the C.I.? if so, may as well stay with no transformation.
  - if rounding $\hat{\lambda}$, check that rounded value is in the C.I.



NTHU STAT 5230, 2024, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

Generalized Least Square (GLS)

- model: $\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$, $E(\boldsymbol{\varepsilon})=\boldsymbol{0}$ and $\text{var}(\boldsymbol{\varepsilon})=\sigma^2\boldsymbol{I} \Rightarrow \boldsymbol{\varepsilon}$: uncorrelated and constant variance
- Consider the case $\text{var}(\boldsymbol{\varepsilon})=\sigma^2\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ ($\neq\boldsymbol{I}$) is *known* but σ^2 is unknown, i.e., we know the *correlation* and *relative variance* between the errors but we don't know the absolute scale
- Because $\boldsymbol{\Sigma}_{n\times n}$ is symmetric and positive definite, we can write $\boldsymbol{\Sigma}=\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}}$, where $\boldsymbol{S}$ is an $n\times n$ nonsigular matrix (by Cholesky or spectral decompositions)

$$\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon} \Rightarrow \boldsymbol{S}^{-1}\boldsymbol{Y}=\boldsymbol{S}^{-1}\boldsymbol{X\beta}+\boldsymbol{S}^{-1}\boldsymbol{\varepsilon} \Rightarrow \boldsymbol{Y'}=\boldsymbol{X'\beta}+\boldsymbol{\varepsilon'}, \text{ where}$$

$$\boldsymbol{Y'}=\boldsymbol{S}^{-1}\boldsymbol{Y},\ \boldsymbol{X'}=\boldsymbol{S}^{-1}\boldsymbol{X},\ \boldsymbol{\varepsilon'}=\boldsymbol{S}^{-1}\boldsymbol{\varepsilon}, \text{ and}$$

$$E(\boldsymbol{\varepsilon'})=\boldsymbol{0} \text{ and } \text{var}(\boldsymbol{\varepsilon'})=\text{var}(\boldsymbol{S}^{-1}\boldsymbol{\varepsilon})=\boldsymbol{S}^{-1}\text{var}(\boldsymbol{\varepsilon})\boldsymbol{S}^{-\mathrm{T}}=\boldsymbol{S}^{-1}\sigma^2\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}^{-\mathrm{T}}=\sigma^2\boldsymbol{I}$$

$\Rightarrow$ For $\boldsymbol{Y'}$ and $\boldsymbol{X'}$, the assumption in ordinary least square is satisfied

- GLS: find $\boldsymbol{\beta}$ that minimize

$$\boldsymbol{\varepsilon'}^T\boldsymbol{\varepsilon'}=(\boldsymbol{Y'}-\boldsymbol{X'\beta})^T(\boldsymbol{Y'}-\boldsymbol{X'\beta})=(\boldsymbol{Y}-\boldsymbol{X\beta})^T\boldsymbol{S}^{-\mathrm{T}}\boldsymbol{S}^{-1}(\boldsymbol{Y}-\boldsymbol{X\beta})=(\boldsymbol{Y}-\boldsymbol{X\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{X\beta})$$

$$\Rightarrow \hat{\boldsymbol{\beta}}=(\boldsymbol{X'}^T\boldsymbol{X'})^{-1}\boldsymbol{X'}^T\boldsymbol{Y'}=(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$$

$$\Rightarrow \text{var}(\hat{\boldsymbol{\beta}})=\sigma^2(\boldsymbol{X'}^T\boldsymbol{X'})^{-1}=\sigma^2(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$$

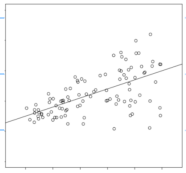
GLS is like OLS regressing $\boldsymbol{Y'}=\boldsymbol{S}^{-1}\boldsymbol{Y}$ on $\boldsymbol{X'}=\boldsymbol{S}^{-1}\boldsymbol{X}$

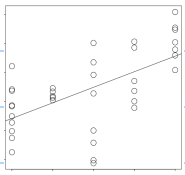
- The practical problem is that $\boldsymbol{\Sigma}$ may not be known. It's usually necessary to make some assumptions and examine the residuals to estimate $\boldsymbol{\Sigma}$ (check IRWLS)

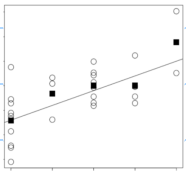
Weighted Least Square (WLS)

- Sometimes, the errors are uncorrelated, but have unequal variance where the form of the inequality is known ($\Rightarrow \boldsymbol{\Sigma}$ is diagonal, it's a special case of GLS), example:
  - ➢ error variance proportional to a function of predictors [e.g., $\text{var}(\varepsilon_i)=x_i^2\sigma^2$]
  - ➢ data with replicates, which show a pattern of unequal variance [e.g., $\text{var}(\varepsilon_i) \approx$ sample variance of observations with same x_i]
  - ➢ the observed y_i's are actually averages of several observations. [e.g., suppose y_i is the average of n_i observations, $\text{var}(\varepsilon_i)=\sigma^2/n_i$]
- $\boldsymbol{\varepsilon}$: uncorrelated, but not constant variance $\Rightarrow \boldsymbol{\Sigma}$ is diagonal. Write

$$\Sigma = \begin{pmatrix} 1/w_1 & 0 & \cdots & 0 \\ 0 & 1/w_2 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & 1/w_n \end{pmatrix} \Rightarrow \Sigma^{-1} = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & w_n \end{pmatrix}$$







where w_i's ($\propto 1/\text{var}(\varepsilon_i)$) are called *weights*.

low weight $\Leftrightarrow$ high variance; high weight $\Leftrightarrow$ low variance

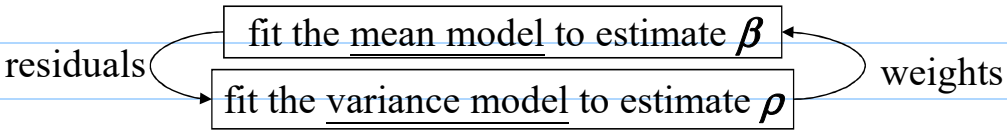
- $\boldsymbol{S}=\text{diag}(1/\sqrt{w_1}, ..., 1/\sqrt{w_n})$, then $\boldsymbol{\Sigma}=\boldsymbol{S}\boldsymbol{S}^{\text{T}}$

$\Rightarrow$ OLS regress $\boldsymbol{S}^{-1}\boldsymbol{Y}$ (i.e., $\sqrt{w_i}\,y_i$) on $\boldsymbol{S}^{-1}\boldsymbol{X}$ ($\sqrt{w_i}\,x_i$) (Note. the column of ones, i.e., intercept needs to be replaced with $\sqrt{w_i}$)

$\Rightarrow$ convenient for regression package without a weighted options

NTHU STAT 5230, 2024, Lecture Notes
made by S.-W. Cheng (NTHU, Taiwan)

- Q: Why observations with smaller variance should be multiplied by heavier weight? intuitive interpretation?
- iteratively re-weighted least squares (IRWLS): In all the previous examples, weights (or $\boldsymbol{\Sigma}$ in GLS) are assumed known. Q: what if $\text{var}(\varepsilon_i)$ is not completely known, what weights should we use? Q: where can you find the information of weights?
  - ➢ model the mean response for $\boldsymbol{Y}$, $\text{E}(\boldsymbol{Y})=\boldsymbol{X}\boldsymbol{\beta}$
  - ➢ model the variance in $\boldsymbol{Y}$, $\text{var}(\boldsymbol{Y})=f(\boldsymbol{X}, \boldsymbol{\rho})$, where $\boldsymbol{\rho}$ are parameters for the variance model



  - ➢ Example: $\text{var}(\varepsilon_i)=\rho_0+\rho_1 x_{i1}$
    1. start with $w_i=1$
    2. use weighted least square to estimate $\boldsymbol{\beta}$
    3. use the residuals to estimate ρ_0 and ρ_1, perhaps by regressing residuals^2 on x_1
    4. re-compute the weights and go to 2. Continue until convergence

    Problems: converge? how is the inference about $\boldsymbol{\beta}$ affected? d.f.=? ...etc
  - ➢ alternative approach: jointly estimate the mean and variance parameters using likelihood based method (in R, use `gls()` function in the `nlme` library)