What to do today (March 24, 2023)?

Part 3. Important Topics in Statistics (Chp 10-13)

§3.1. Analysis of Variance (ANOVA, Chp 10-11)

§3.2. Introduction to Regression Analysis (Chp 12-13)

§3.2.1 Introduction

§3.2.2 Simple Linear Regression (Chp 12)

§3.2.3 More Advanced Topics (Chp 13)

3.2.3A Multiple linear regression

3.2.3B Regression with transformed variables

3.2.3C Regression with categorical predictors

3.2.3D Discussion

3.2.3E A Comprehensive Example

Some Logistics.

- Homework 9 has been assigned. It's due on Monday March 27.
- The remaining Midterm 2 papers are with Joan Hu.

3.2.3B Regression with transformed variables

Goal. to establish how Y depends on $X_1, \ldots, X_K$

– *always linear relationship?*

Examples ...

- $Y \in (0, \infty)$: $\ln Y = \beta_0 + \beta_1 X + \epsilon$
- Polynominal regression: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$
- Relationship with two predictors (interaction): $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \epsilon$
- more examples in Table 13.1 of the textbook

In general,
$g(Y) = \beta_0 + \beta_1 h_1(X_1) + \ldots + \beta_k h_K(X_K) + \epsilon.$

$$\Longrightarrow Y^* = \beta_0 + \beta_1 X_1^* + \ldots + \beta_K X_K^* + \epsilon$$

with $Y^* = g(Y)$; $X_1^* = h_1(X_1), \ldots, X_K^* = h_K(X_K)$.
$\Longrightarrow$ *A strategy of conducting nonlinear regression analysis by linear regression analysis* (General Linear Regression Analysis)

3.2.3C Regression with categorical predictors

What if an explanatory variable in the regression is qualitative?

- Introducing "dummy" variables to indicate the categories. For example,
  - "predictor"=gender: female vs male. Define

$$X = \begin{cases} 1 & \textit{male} \\ 0 & \textit{female} \end{cases}$$

  - "predictor"=color: red vs yellow vs blue. Define

$$X_1 = \begin{cases} 1 & \textit{red} \\ 0 & \textit{otherwise} \end{cases} \quad X_2 = \begin{cases} 1 & \textit{yellow} \\ 0 & \textit{otherwise} \end{cases}$$

  $\Longrightarrow$ To study "how Y depends on $X_1, \ldots, X_K$?" (*Regression Analysis*)
- "Different Coding"? (parameter interpretation!)

Something further ... ...

For example, a study with response Y and one explanatory varialbe "education" ($\leq$ high school, college, and postgraduate): a One-Factor Study.

(i) One-Factor ANOVA Model: factor "education" with $I = 3$: $\sum_i \alpha_i = 0$, $i = 1, 2, 3$ and $j = 1, \ldots, n_i$

$$Y_{ij} = \mu_i + \epsilon_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

(ii) Using the dummy variables X_1 and X_2:

$$X_1 = \begin{cases} 1 & HighSchool \\ 0 & otherwise \end{cases} \qquad X_2 = \begin{cases} 1 & College \\ 0 & otherwise \end{cases}$$

$\Longrightarrow$ Linear Regression Model: $k = 1, \ldots, n$

$$Y_k = \beta_0 + \beta_1 X_{1k} + \beta_2 X_{2k} + \epsilon_k$$

The two models should be equivalent:

$$\begin{cases} \beta_0 = \mu_3 = \mu + \alpha_3 \\ \beta_1 = \mu_1 - \mu_3 = \alpha_1 - \alpha_3 \\ \beta_2 = \mu_2 - \mu_3 = \alpha_2 - \alpha_3 \end{cases}$$

(iii) An alternative coding:

$$Z_1 = \begin{cases} 1 & \mathit{HighSchool} \\ -1 & \mathit{otherwise} \end{cases} \qquad Z_2 = \begin{cases} 1 & \mathit{College} \\ -1 & \mathit{otherwise} \end{cases}$$

$\Longrightarrow$ Linear Regression Model: $k = 1, \ldots, n$

$$Y_k = \gamma_0 + \gamma_1 Z_{1k} + \gamma_2 Z_{2k} + \epsilon_k$$

Thus,

$$\begin{cases} \mu_1 = \mu + \alpha_1 = \gamma_0 + \gamma_1 \\ \mu_2 = \mu + \alpha_2 = \gamma_0 + \gamma_2 \\ \mu_3 = \mu + \alpha_3 = \gamma_0 - \gamma_1 - \gamma_2 \end{cases}$$

$\Longleftrightarrow$

$$\mu = \gamma_0; \alpha_1 = \gamma_1; \alpha_2 = \gamma_2; \alpha_3 = -\gamma_1 - \gamma_2$$

For another example, to consider how Y is associted with Factors A and B: Factor A with 3 levels, Factor B with 4 levels, n_{ij} can be different.

"Two-Factor Study with Unbalanced Data"!

- $3-1$ dummy variables for Factor A: X_{A1}, X_{A2}; $4-1$ dummy variables for Factor B: X_{B1}, X_{B2}, X_{B3}
- Consider a multiple linear regression model

$$Y = \beta_0 + \beta_1 X_{A1} + \beta_2 X_{A2} + \gamma_1 X_{B1} + \gamma_2 X_{B2} + \gamma_3 X_{B3} + \epsilon$$

  or a regression model with cross-product terms

$$Y = \beta_0 + \beta_1 X_{A1} + \ldots + \gamma_3 X_{B3} + \phi_1 X_{A1} X_{B1} + \ldots + \epsilon$$

Balanced data type is not required in the regression analysis.

3.2.3D Discussion

- What does a regression analysis do?
- Variable selection, model selection in regression analysis
- What if Y is categorical?
  $\Longrightarrow$ Generalized Linear Models. *Categorical Data Analysis*
- What if $E(Y|X)$ is not specified into a linear form, or a parametric form?
- Model checking in (linear) regression analysis
  What if $Y_1, \ldots, Y_n$ are not indpt?
  e.g. they're market prices of a stock recorded on days 1, 2, ..., n? $\Longrightarrow$ Time Series: how to deal with?

What will we study next?

Part 1. Introduction and Review (Chp 1-5)

Part 2. Basic Statistical Inference (Chp 6-9)

Part 3. Important Topics in Statistics (Chp 10-13)

3.1A One-Factor Analysis of Variance (Chp 10)

3.1B Multi-Factor ANOVA (Chp 11)

3.2A Simple Linear Regression Analysis (Chp 12)

3.2B More on Regression (Chp 13)

Part 4. Further Topics (Selected from Chp 14-16)

4.1 Distribution-Free Procedures (Chp15.1, 15.2)

4.2 Quality Control Methods (Chp16.1, 16.2, 16.3)