What to do today (March 21, 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)

Some Logistics.

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

3.2.2D Residual analysis

(model checking)

- *(raw) Residual:* $e_i = Y_i - \hat{Y}_i$ ($observed - fitted$) with $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = \bar{Y} + \hat{\beta}_1 (X_i - \bar{X})$

$$E(e_i) = 0; \quad V(e_i) = \sigma^2 [1 - \frac{1}{n} - \frac{(X_i - \bar{X})^2}{S_{XX}}]$$

  and $e_i \sim normal$

- *Standardized residual:*

$$e_i^* = e_i / s\sqrt{1 - \frac{1}{n} - \frac{(X_i - \bar{X})^2}{S_{XX}}}$$

  with $s^2 = \hat{\sigma}^2 = MSS_e$. $\Longrightarrow e_i^* \sim N(0,1)$ roughly

- *Pearson residual:* $\tilde{e}_i = e_i / s$
  when $n \gg 1$ and $S_{XX} \gg (x_i - \bar{x})^2$, $\tilde{e}_i \approx e_i^*$.

Commonly-used diagnostic plots:

- x_i vs e_i (or e_i^* or $\tilde{e}_i$)
- $\hat{y}_i$ vs e_i (or e_i^* or $\tilde{e}_i$)
- y_i vs $\hat{y}_i$
- a normal probability plot of e_i (Z-percentile vs e_i):
    - rank $e_i, i = 1, \dots, n$, denoted by

$$e_{(1)}, \dots, e_{(n)}$$

    - scatter plot:

$$(Z_{1/n-1/2n}, e_{(1)}), (Z_{2/n-1/2n}, e_{(2)}), \dots, (Z_{1-1/2n}, e_{(n)})$$

Example 7.1 (cont'd) Consider a regression analysis under model $Y = \beta_0 + \beta_1 X + \epsilon$, assuming $\epsilon \sim N(0, \sigma^2)$:

- (i) ANOVA

Source of Variation	df	SS	MSS	F-value
regression	1	398030.2	MSS_{reg}	$F_{obs} = \frac{MSS_{reg}}{MSS_e} = 294.74$
error	$n-2=12$	16205.45	MSS_e	$>> F_{.95}(1,12) = 3.17$
total	$n-1=13$	SS_T		

$(SS_T = SS_{reg} + SS_e)$

- (ii) Test on $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 > 0$

$$T = \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} \sim t(12) \text{ under } H_0$$

$T_{obs} = 1.711/.0997 = 17.16 >> t_{.95}(12) = 1.78.$

Example 7.1 (cont'd) Consider a regression analysis under model $Y = \beta_0 + \beta_1 X + \epsilon$, assuming $\epsilon \sim N(0, \sigma^2)$:

- (iiia) Estimation for $E(Y|X = x^*)$ with $x^* = 225$.

$$\hat{Y}|x^* = \hat{\mu}_{Y|x^*} = \hat{\beta}_0 + \hat{\beta}_1 x^* = 339.43$$

$1 - \alpha$ CI:

$$\hat{Y}|x^* \pm t_{\alpha/2}(n-2)\sqrt{MSS_e[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}]} \Longrightarrow (317.90, 360.95)$$

- (iiib) Prediction for Y with $X = x^* = 225$.

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

$1 - \alpha$ prediction interval (PI)

$$\hat{Y}|x^* \pm t_{\alpha/2}(n-2)\sqrt{MSS_e[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}]} \Longrightarrow (256.51, 422.34)$$

§3.2.3 More on Regression (Chp 13)

3.2.3A Multiple linear regression

- **Modeling**

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

$E(\epsilon) = 0$, $V(\epsilon) = \sigma^2$.

With the data from n indpt units: $\{(Y_i, X_{1i}, \ldots, X_{Ki}) : i = 1, \ldots, n\}$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_K X_{Ki} + \epsilon_i$$

ϵ_i indpt with $E(\epsilon_i) = 0$ and $V(\epsilon_i) = \sigma^2$.

Alternative presentation: $\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\epsilon}$ with $E(\underline{\epsilon}) = \underline{0}$ and $V(\underline{\epsilon}) = diag(\sigma^2, \ldots, \sigma^2)$.

Inferences with the Multiple Linear Regression Model

(a) Estm of $\beta_0, \ldots, \beta_K$ by **LSE:** $\hat{\beta}_0, \ldots, \hat{\beta}_K$

$$L(\hat{\beta}_0, \ldots, \hat{\beta}_K) = \min_{\underline{\beta}} L(\beta_0, \ldots, \beta_K)$$

with $L(\beta_0, \ldots, \beta_K) = \sum_{i=1}^{n} \left(Y_i - [\beta_0 + \beta_1 X_{1i} + \ldots + \beta_K X_{Ki}]\right)^2$.

$$\underline{\hat{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underline{Y}$$

Properties of LSE:

(i) $\underline{\hat{\beta}}$ is the best unbiased linear estm

(ii) If $\epsilon_i \sim N(0, \sigma^2)$ iid, $\underline{\hat{\beta}}$ is the MLE and $\underline{\hat{\beta}} \sim N(\underline{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$

(b) Estimation of σ^2: unbiased $\hat{\sigma}^2 = MSS_e$

$$SS_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

with $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_K X_{Ki}$. Let $MSS_e = SS_e/[n-(K+1)]$.

(c) Associated ANOVA with $\epsilon_i \sim N(0, \sigma^2)$ iid

$$SS_T = \sum_{i=1}^{n}(Y_i - \bar{Y})^2, \quad SS_e = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad SS_{reg} = SS_T - SS_e$$

$H_0 : \beta_1 = \ldots = \beta_K = 0$ vs H_1 : *otherwise*. [*Test for Model Utility*]

Source of Variation	df	SS	F-value
regression	K	SS_{reg}	$F = \frac{MSS_{reg}}{MSS_e}$
error	$n-(K+1)$	SS_e	$\sim F(K, n-(K+1))$ under H_0
total	$n-1$	SS_T	

– *Coef of determination.* $R^2 = SS_{reg}/SS_T = 1 - SS_e/SS_T$, multiple correlation coef R
– *Adjusted* R^2. $\bar{R}^2 = 1 - MSS_e/MSS_T$, to adjust for num of explanatory variables in the model.

(d) Interval estimation

(i) CI of β_k with $1-\alpha$ level: $\hat{\beta}_k \pm t_{\alpha/2}(n-(K+1))s_{\hat{\beta}_k}$

(ii) CI of $\mu_{Y|X_1^*,\ldots,X_K^*} = \beta_0 + \beta_1 X_1^* + \ldots + \beta_K X_K^*$ with $1-\alpha$ level $\hat{\mu}_{Y|X_1^*,\ldots,X_K^*} \pm t_{\alpha/2}(n-(K+1))s_{\hat{\mu}_{\underline{X}^*}}$

Here estm $\hat{Y}_{\underline{X}^*} = \hat{\mu}_{\underline{X}^*}$; $V(\hat{\mu}_{\underline{X}^*}) = \sigma^2 \underline{x}^{*'}(\mathbf{X}'\mathbf{X})^{-1}\underline{x}^*$

(iii) PI of Y at $\underline{x}^{*'} = (x_1^*, \ldots, x_K^*)$ with $1-\alpha$ level

$$\hat{\mu}_{Y|X_1^*,\ldots,X_K^*} \pm t_{\alpha/2}(n-(K+1))\sqrt{s^2 + s^2_{\hat{\mu}_{\underline{X}^*}}}$$

Anything really new in multiple linear regression?
$\Longrightarrow$ (e) Variable selection: to be studied soon ... ...

Example 7.2 (textbook p583)

- **Study.** reported by an article in *J of the Amer Ceramic Soc* (2012):
  Y = glass microhardness resulted from various compositions;
  $X_1 = N$; $X_2 = F$
- Data: $n = 18$ indpt obs

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i, \quad i = 1, \ldots, n$$

assuming $\epsilon_i \sim N(0, \sigma^2)$.

- the fitted model: $Y = 6.23 + 0.0618X_1 - 0.0387X_2$
- $R^2 = 98.4\%$
- 95% CI for β_1: (.0573,.0663); 95% CI for β_2: (-.0626, -.0148)
- 95% CI for average hardness when $X_1 = 20, X_2 = 1$: (7.35, 7.50)
- 95% PI for an observed hardness when $X_1 = 20, X_2 = 1$: (7.20, 7.65)

(e) Variable selection

To select from $X_1, \ldots, X_K$ the "important" (significant) explanatory variables for Y

- *Procedures* when $K \gg 1$, the exhaustive search hard to implement
  - (i) Forward Selection: starting from the most important explanatory variable and gradually adding to the list the important ones
  - (ii) Backward Elimination: starting from the full list and gradually eliminating the non-important variables
  - (iii) Forward-Backward/Backward-Forward Selection:

Assessing "importance": various criteria

(i) R^2-value, Adjusted R^2-value;

(ii) Akaike's Information Criterion (AIC):
$-2 \ln L + 2p$ ($L =$ maximum likelihood, closely related to Kullback-Leibler information) *the smaller, the better*;

(iii) Bayesian Information Criterion (BIC):
$-2 \ln L + p \ln n$ *the smaller, the better*.

(iv) Hypothesis testings

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)**
  - *3.2.3A Multiple linear regression*
  - **3.2.3B Regression with transformed variables**
  - **3.2.3C Categorical predictors**
  - **3.2.3D Discussion**

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