What to do today (March 17, 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 8 has been assigned. It's due on Monday March 20.
- Marked Midterm 2 papers will be distributed at next week's tutorial.

Example 7.1

- **Study.** reported in "An Experimental Correlation of Oxides of Nitrogen Emissions from Power Boilers Based on Field Data", *J. of Engr for Power*, July 1973: 165-70.

obs	x_i	y_i	$\hat{y}_i$	e_i
1	100	150		
2	125	140		
3	125	180		
4	150	210		
5	150	190		
6	200	320		
7	200	280		
8	250	400		
9	250	430		
10	300	440		
11	300	390		
12	350	600		
13	400	610		
14	400	670		
total	3300	5010	...	...
squares				
total	913750	2207100	...	...
	$\sum_i x_i y_i = 1413500$			

Consider a regression analysis under model $Y = \beta_0 + \beta_1 X + \epsilon$:

obs	x_i	y_i	$\hat{y}_i$	e_i
1	100	150		
2	125	140		
...	...	...		
14	400	670		
total	3300	5010	...	...
sample mean	**235.71**	**357.86**	...	...
squares				
total	913750	2207100	...	...
sum squares	**135892.9**	**414235.7**	...	...

$\sum_i x_i y_i = 1413500$

$S_{XY} = 232571.4$

$\hat{\beta}_1 = S_{XY}/S_{XX} = 1.711$; $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -45.55$

Consider a regression analysis under model $Y = \beta_0 + \beta_1 X + \epsilon$:

obs	x_i	y_i	$\hat{y}_i$	e_i
1	100	150	**125.59**	**24.41**
2	125	140	**168.38**	**-28.38**
...	...	...		
14	400	670	**639.02**	**30.98**
total	3300	5010	**5010**	**0**
sample mean	235.71	357.86	**357.86**	**0**
squares				
total	913750	2207100	**2190895**	**16205.4**
sum squares	135892.9	414235.7	**397666.3**	**16205.4**

$\sum_i x_i y_i = 1413500$

$S_{XY} = 232571.4$

$\hat{\beta}_1 = S_{XY}/S_{XX} = 1.711$; $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -45.55$

Estimate σ^2: $\hat{\sigma}^2 = 16205.4/(n-2) = 1350.45$

3.2.2C Additional inferences

Consider the **Simple Linear Regression Model** with the additional normality assumption:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

for $i = 1, \ldots, n$, ϵ_i's are iid $N(0, \sigma^2)$.
$\Longrightarrow$ $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$ and indpt with each other.

- *The MLE and the LSE of β_0, β_1 are the same:*

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_i (X_i - \bar{X}) Y_i}{S_{XX}} \sim N(\beta_1, \frac{\sigma^2}{S_{XX}}) \\ \hat{\beta}_0 &\sim N\big(\beta_0, \sigma^2 [\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}]\big) \end{aligned}$$

The LSE/MLE are the best unbiased estimators of β_0, β_1.

- *Inferences concerning* β_1.

  $1-\alpha$ *CI*:

  $$\hat{\beta}_1 \pm t_{\alpha/2}(n-2)s_{\hat{\beta}_1}$$

  standard error of $\hat{\beta}$: $s_{\hat{\beta}_1} = \sqrt{\frac{MSS_e}{S_{xx}}}$

  $H_0 : \beta_1 = \beta_{10}$ *vs* H_1 : *otherwise*

  $$T = \frac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}} \sim t(n-2)$$

  under H_0.

What does $\beta_1 = 0, < 0,$ *or* > 0 *mean?*

- *Inference concerning* $E(Y|X = x^*) = \beta_0 + \beta_1 x^*$ (average/mean of a future Y, $\mu_{Y|x^*}$)

$$\begin{aligned} \hat{\mu}_{Y|x^*} &= \hat{\beta}_0 + \hat{\beta}_1 x^* = \hat{Y}|x^* \\ &\sim N(\mu_{Y|x^*}, \sigma^2[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}]) \end{aligned}$$

$1 - \alpha$ CI:

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

- *Prediction for* Y *at* $X = x^*$ *(a future* Y *value)*

$$Y = \beta_0 + \beta_1 x^* + \epsilon \quad \Rightarrow \hat{\beta}_0 + \hat{\beta}_1 x^* + \hat{\epsilon}$$

$\hat{\epsilon}$ unavailable but mean 0 and variance σ^2
$\Longrightarrow 1 - \alpha$ prediction interval (PI)

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

- ANOVA in regression analysis: to answer "is the regression model good enough?"

Source of Variation	df	SS	MSS	F-value
regression	1	SS_{reg}	MSS_{reg}	$F = \frac{MSS_{reg}}{MSS_e}$
error	$n-2$	SS_e	MSS_e	
total	$n-1$	SS_T		

$SS_T = SS_{reg} + SS_e$:
$SS_T = S_{YY} = \sum_i (Y_i - \bar{Y})^2$; $SS_e = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$;
$SS_{reg} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_1^2 S_{XX}$.

$$F = \frac{MSS_{reg}}{MSS_e} = \frac{\hat{\beta}_1^2}{MSS_e / S_{XX}} \sim F(1, n-2)$$

under $H_0 : \beta_1 = 0$

Remarks.

- *Coefficient of determination*:

$$R = \frac{SS_{reg}}{SS_T} = \hat{\beta}_1^2 \frac{S_{XX}}{S_{YY}}$$

  a measure of the portion of observed Y variation that can be explained/captured by the simple linear regression model

  In practice, check R-value in regression analysis. **Coefficient of determination in Example 7.1:** $R = SS_R/SS_T = 0.96$

- *Relationship with sample correlation coef* (corr: a measure of X and Y's linear relationship):

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} = \hat{\beta}_1 \sqrt{\frac{S_{XX}}{S_{YY}}}$$

$\Longrightarrow r^2 = R$-value in the SLRM.

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 $e_i, i = 1, \ldots, n$, denoted by

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

  - scatter plot:

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

For example, ...

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)

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

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)