STAT 481 -- Midterm II

Exam Time: 1:00 - 1:50 PM, March 18, 2016

Name: _____

UIN: _____

— Solution —

Score Table

Problems	Score	
1	30	
2	30	
3	40	
Total		

1. **[30pt]** The tensile strength of a paper product is related to the amount of hardwood in the pulp. Ten samples are produced in the pilot study. We will fit a linear regression model relating strength to percent hardwood as follows:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1,...,10,.$$

(1). It is known that SSE=38.8, SSR= 1262. Construct the ANOVA table.

SS TO = SSR + SSE = 165

Source	DF	SS	MS	F
Regression	1	126.2	126.2	26
Error	8	38.8	4.85	
Total	9	165		

(2). State the hypotheses. What conclusion can you draw for the test for significance of regression?

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

$F = 26 > F_{0.05}(1, 8) = 5.318$

Reject H_0, i.e. there exists linear relationship between the percent hard-wood and strength.

(3). Derive the distribution of the estimate $\hat{\beta}_1$ first, and show that the sampling distribution of $\hat{\beta}_1$ is a t distribution. Given that

$$\hat{\beta}_1 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{Sxx} y_i = \sum_{i=1}^{n} c_i y_i, \text{ where } Sxx = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, independent responses.

$\sum_i c_i = \sum_i (x_i - \bar{x}) / S_{xx}$

$\sum_i c_i x_i = \sum_i c_i (x_i - \bar{x}) = 1$

$\sum_i c_i^2 = \frac{1}{S_{xx}}$

$$E\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i \cdot E y_i = \sum_i c_i (\beta_0 + \beta_1 x_i)$$

$$= \beta_0 \cdot \sum_i c_i + \beta_1 \left(\sum_i c_i x_i\right) = \beta_1$$

$$Var\left(\sum_i c_i y_i\right) = \sum_i c_i^2 \cdot Var(y_i) = \sum_i c_i^2 (\sigma^2) = \frac{\sigma^2}{S_{xx}}$$

$$\therefore \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right), \text{ i.e. } \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / S_{xx}}} \sim N(0,1)$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSE}{S_{xx}}}} = \frac{(\hat{\beta}_1 - \beta_1)/\sqrt{\sigma^2/S_{xx}}}{\sqrt{\frac{SSE}{n-2}/\sigma^2}} \sim t(n-2) \quad \text{as} \begin{cases} \hat{\beta}_1 \perp\!\!\!\perp SSE \\ \frac{SSE}{\sigma^2} \sim \chi^2(n-2) \\ \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1) \end{cases}$$

(4). (Graduate Only) Show that $E(SSR) = \sigma^2 + \beta_1^2 S_{xx}$.

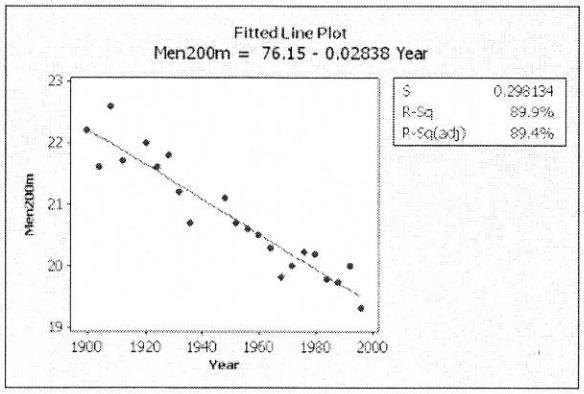
$$SSR = \sum_i (\hat{y}_i - \bar{y})^2 \quad \left| \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \ \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \right.$$

$$SSR = \sum_i \hat{\beta}_1^2 \cdot (x_i - \bar{x})^2 = \hat{\beta}_1^2 \cdot \sum_i (x_i - \bar{x})^2$$

$$E(\hat{\beta}_1^2) = \frac{\sigma^2}{S_{xx}} + \beta_1^2 \quad \text{From (3)}$$

$$E(SSR) = \sigma^2 + \beta_1^2 S_{xx}$$

2. **[30pt]** The data set (mens200m.txt) contains the winning times (in seconds) of the 22 men's 200 meter olympic sprints held between 1900 and 1996. (Notice that the Olympics were not held during the World War I and II years.) Is there a linear relationship between year and the winning times? Below include a scatter plot and partial computer output by fitting a simple linear regression model.



Predictor	Coef	SE Coef	T	P
Constant	76.153	4.152	18.34	0.000
Year	-0.0284	0.00213	-13.33	0.000

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	15.796	15.796	177.7	0.000
Residual Error	20	1.778	0.089		
Total	21	17.574			

(1). To answer the research question about the linear relationship, let's conduct the *F*-test. What is the null and the alternative hypotheses for the study? What decision will you draw based on the ANOVA table above? ↑ what is the statistic value?

Linear relationship $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

$F = \frac{MSR}{MSE} = 177.7$, P-value (given) < 0.01

There exists strong linear relationship between the year and the winning time.

(2). Based on the information given above, find the fitted regression line with intercept and slope estimates. Calculate the coefficient of the determination and explain it.

$\hat{\beta}_0 = 76.153$ $\hat{\beta}_1 = -0.028\%$

Fitted line: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$, $i = 1, \ldots, n$

$R^2 = \frac{SSR}{SSTO} = \frac{15.796}{17.574} \approx 90\%$

95%

(3). Construct the confidence interval for the slope β_1.

C.I. for β_1: $\hat{\beta}_1 \pm t_{0.025}(n-2)\, se(\hat{\beta}_1)$ $\quad n-2=20$, $t_{0.025}(20)=2.086$

$(-0.0284) \pm (2.086) * (0.00213)$

$(-0.24, -0.033) \neq 0$

(4). Is there a negative association between year and the wining time? State your hypothesis for this research objective. Choose appropriate test and draw your conclusion. $\alpha=5\%$

$H_0: \beta_1=0$ vs $H_1: \beta_1<0$, left-sided test

$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{-0.0284}{0.00213} = -13.3 < -t_{0.05}(20) = -1.725$

Reject $H_0: \beta_1=0$, there exists strong negative association between year and the winning time.

3. **[40 pt]** Suppose the National Transportation Safety Board (NTSB) wants to examine the safety of compact cars, midsize cars, and full-size cars. It collects a sample of three for each of the treatments (cars types). Using the hypothetical data provided below, test whether the mean pressure applied to the driver's head during a crash test is equal for each types of car. Use $\alpha=5\%$.

Car Type	Pressure on head	Group mean ($\bar{y}_{i\bullet}$)	Sample variance (S_i^2)
Compact	644, 655, 702	667	949
Midsize	468, 427, 524	473	2374
Full-size	484, 456, 401	447	1783

Write down the model you will use for the analysis

(1). Calculate SSTR and SSE. $\bar{y}_{..} = (667+473+447)/3 = 529$

$SSTR = \sum_i n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = 3 \times \{[667-529]^2 + (473-529)^2 + (447-529)^2\} = 86712$

$SSE = \sum_i (n_i-1) S_i^2 = 2 \times [949 + 2374 + 1783] = 10212$

$SSTO = SSE + SSR = 97302$

$Y_{ij} = \mu_i + \varepsilon_{ij}$, $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$, μ_i: group mean are the parameters of interest

σ^2: variance of error, nuisance parameter

or $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, $\sum_{i=1}^{k} n_i \tau_i = 0$ $(\sum_{i=1}^{3} \tau_i = 0)$

(2). State the null and alternative hypotheses for this study. Calculate the appropriate test statistic based on the calculation in (1) and draw your decision accordingly.

Let μ_i be the mean pressure of the i-th group, $i=1,2,3$ $(K=3)$

$H_0: \mu_1=\mu_2=\mu_3$ vs $H_1:$ at least one μ_i is different from others

$$F=\frac{MSTR}{MSE}=\frac{SSTR/(K-1)}{SSE/(N-K)}=\frac{8671.2/2}{1021.2/6}=25.5, \quad (MSE=170.2)$$

$$F_0=25.5 > F_{0.01}(2,6)=10.925 \qquad p\text{-value} < 0.01$$

(3). Construct the Tukey's simultaneous confidence interval and decide if there is difference among the mean pressures in the crash test for three types of cars.

Tukey's SCI for $\mu_i-\mu_j$: $(\bar{Y}_i-\bar{Y}_j) \pm q_\alpha(k,N-k)\sqrt{\frac{MSE}{n}}$ $(n_1=n_2=n_3=3)$

$(\mu_1-\mu_2)$: 19.4 ± 103.38 $(+,+)$ $q_{0.05}(3,6)=4.34$

$(\mu_2-\mu_3)$: 26 ± 103.38 $(-,+)$ $\sqrt{\frac{MSE}{n}}=23.82$

$(\mu_1-\mu_3)$: 220 ± 103.38 $(+,+)$ $ME=103.38$

Mean pressure in Compact group is significantly from the other two groups. But the mean pressures in mid-size and full-size groups show no significant difference.

Grouping: Compact 1
Mid size 2
Full size 2

(4). Show that $E(MSE) = \sigma^2$, where σ^2 is the constant response variance.

$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2 = \sum_{i=1}^{k} (n_i - 1) S_i^2$, where $S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2$ is the sample variance of i-th group

$\because Y_{ij} \overset{iid}{\sim} N(\mu_i, \sigma^2)$, then

$$\frac{(n_i - 1) S_i^2}{\sigma^2} \sim \chi^2(n_i - 1), \quad i = 1, \ldots, k$$

In addition, all data are independent, then $\sum_{i=1}^{k} \frac{(n_i - 1) S_i^2}{\sigma^2} \sim \chi^2\left(\sum_{i=1}^{k} (n_i - 1)\right) = \chi^2(N - k)$

i.e. $\frac{SSE}{\sigma^2} \sim \chi^2(N - k) \Rightarrow E\left[\frac{SSE}{\sigma^2}\right] = N - k$, $E[SSE] = \sigma^2 (N - k)$

$\therefore E[MSE] = E\left[\frac{SSE}{N - k}\right] = \sigma^2$

(5). Comment on the difference between the application of the confidence interval (C.I.) and the simultaneous confidence interval (S.C.I.) for pairwise comparisons between the treatment levels.

(1) C.I. is an interval estimate for a single parameter, pairwise difference of a particular pair (of the levels)

SCI. is for comparing multiple pairwise difference among possibly all treatment levels

(2) SCI can group the trt levels based on the significance differences of those trt level differences.

(3) C.I. has shorter length than SCI because SCI is a joint event of several C.I.s. eg. Bonferroni's SCI, to achieve the overall confidence level $(1-\alpha)$, each individual CI needs to have a confidence level $(1-\alpha^*)$, and $\alpha^* = \frac{\alpha}{m}$, m is the # of pairs for comparison.