**STAT 481** **Midterm II – Review** **Spring 2016**

1. A simple linear regression model $Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$ is used to fit n=12 data points.
   a) It was found that SSTO=240, SSR=160. Construct the ANOVA Table.
   b) Calculate R-square and interpret it.
   c) Use two different tests to test the hypotheses $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ , do you reach the same decisions?
   d) Given that the least square estimator $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, derive its distribution. What is the standard error of $\hat{\beta}_1$.
   e) What is the necessary model assumption for a regression model?
   f) For a simple data analysis, what model checking procedures are needed before you run the simple linear regression analysis?

2. In the computer science department of a large university, many students change their major after the first year. A detailed study[1] of the 256 students enrolled as first-year computer science majors in one year was undertaken to help understand this phenomenon. Students were classified on the basis of their status at the beginning of their second year, and several variables measured at the time of their entrance to the university were obtained. Here are summary statistics for the SAT mathematics scores:

| Second-year major | $n$ | $\bar{x}$ | $s$ |
|---|---|---|---|
| Computer science | 103 | 619 | 86 |
| Engineering and other sciences | 31 | 629 | 67 |
| Other | 122 | 575 | 83 |

   a) Write down the mean model and the required assumption for the analysis.
   b) Show that the sample group mean $\bar{Y}_{i\bullet}$ is the least square estimator for the group mean $\mu_i, i = 1,2,3..$
   c) Write down effect model and model restrictions.
   d) Compute SSTO and SSTR and construct ANOVA table.
   e) Given that F(0.01, 2, 253) = 4.7, what is the p-value for the test for the difference of the SAT scores among three groups of students? What's your conclusion?
   f) Prove that E(MSE)=$\sigma^2$. What is E(MSTR) in general?
   g) What is the sampling distribution for the sample mean of the SAT scores in the first group (CS) $\bar{Y}_{1\bullet}$? Construct its 100(1-α)% confidence interval for the group mean $\mu_1$.

h) Calculate the 95% confidence interval for the difference between the SAT scores of computer science and engineering science students.

i) Use Tukey's 95% simultaneous confidence interval to decide if there is difference among the groups, and further group the majors. Use q(0.05, 3, 253)=3.32.

**Brief Keys:**

1. a) ANOVA

| Source | SS | DF | MS | F |
|---|---|---|---|---|
| Regression | 160 | 1 | 160 | 20 |
| Error | 80 | 10 | 8 | |
| Total | 240 | 11 | | |

b) Coefficent of determination: $R^2$=66.7%. By introducing two predictors, the variation of the response is reduced by 66.67%. Or 66.67% of the variation of the response can be explained by the linear model.

c) Use F test or two-sided t test, $F_o$=20, $t_o$= 4.47. $F_{0.05}(1,10)$=4.96, $t_{0.05}(10)$=2.23. Both tests lead to the same decisions, reject $H_0$,

$$d)\ \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{S_{xx}} = \sum_{i=1}^{n} c_i y_i, \text{ where } \sum_{i=1}^{n} c_i = 0, \sum_{i=1}^{n} c_i x_i = 1, \sum_{i=1}^{n} c_i^2 = S_{xx}^{-1},$$

$$\because y_i \sim ind.\ N(\beta_0+\beta_1 x_i, \sigma^2) \Rightarrow \hat{\beta}_1 \sim N\left(\sum_{i=1}^{n} c_i(\beta_0+\beta_1 x)_i, \sum_{i=1}^{n} c_i^2\sigma^2,\right) = N\left(\beta_1, S_{xx}^{-1}\sigma^2\right)$$

$$se(\hat{\beta}_1) = \sqrt{S_{xx}^{-1}\cdot\hat{\sigma}^2} = \sqrt{S_{xx}^{-1}\cdot MSE}$$

e) Independent and idendically distributed errors : $\varepsilon_i \sim N(0,\sigma^2)$,

f) Checking for linearity between the predictor and the response (scatter plot), Normality of the error distribution (QQ-plot, nonparametric test), constant variance (residual plot), dependence among responses (autocorrelation test), and outliers/influential points.

2. a) Mean Model: $Y_{ij} = \mu_i + \varepsilon_{ij},\ i=1,2,3,\ j=1,2...,n_i,\ n_1=103, n_2=31, n_3=122$, where errors are $\varepsilon_{ij}$ i.i.d. $\sim N(0,\sigma^2)$. Variable Y is the response, SAT score. Mean $\mu_i$ is the average score of each of the three majors, C.S., Eng., other major (factor level).

b) Objective function : $Q(\mu_1,\mu_2,\mu_3) = \sum_{i=1}^{3}\sum_{j=1}^{n_i}(Y_{ij}-\mu_i)^2$.

Estimating equation : $\frac{\partial Q}{\partial \mu_i} = (-2)\sum_{j=1}^{n_i}(Y_{ij}-\mu_i) = 0,$

Then the least square estimator (solution) : $\hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} = \overline{Y}_{i\bullet}$

c) Effect Model: $Y_{ij} = \mu + \tau_i + \varepsilon_{ij},\ i=1,2,3,\ j=1,2...,n_i,\ n_1=103, n_2=31, n_3=122$

where $\varepsilon_{ij}$ i.i.d. $\sim N(0,\sigma^2)$ and $\sum n_i\tau_i = 0$.

d) SSTO=1862003, SSTR=139372.

$d$) Overall Mean : $\bar{x}_{\bullet\bullet} = \frac{1}{N}\sum_i n_i \bar{x}_{i\bullet} = 599.24$

$$SSTR = \sum_{i=1}^{3} n_i \left(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet}\right)^2 = 139357,\ SSE = \sum_{i=1}^{3}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_{i\bullet}\right)^2 = \sum_{i=1}^{3}\sum_{j=1}^{n_i}\left(n_i - 1\right)s_i^2 = 1722631$$

| Source | SS | DF | MS | F |
|---|---|---|---|---|
| Treatment | 139357 | 2 | 69678.5 | 10.23 |
| Error | 1722631 | 253 | 6808.8 | |
| Total | 1861988 | 255 | | |

e) P-value<0.01, reject $H_0$, i.e. three groups of students do have different SAT math scores.

f) $E\left(MSE\right) = \sigma^2, E\left(MSTR\right) = \sigma^2 + \frac{1}{k-1}\sum_{i=1}^{k} n_i \tau_i^2, k = 3$

g) Reference Distribution of $\bar{Y}_{1\bullet}$ : $\bar{Y}_{1\bullet} \sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right), n_1 = 103.$

Sampling distribution of $\bar{Y}_{1\bullet}$ : $\frac{\bar{Y}_{1\bullet} - \mu_1}{\sqrt{\frac{MSE}{n_1}}} \sim t\left(N-k\right).\ N = 256, k = 3$

Confidence interval for $\mu_1$ : $\bar{Y}_{1\bullet} \pm t_{\frac{\alpha}{2}}\left(N-k\right)\sqrt{\frac{MSE}{n_1}}$

h) $-10 \pm 33.42$

i) Grp 1vs Grp 2: $-10 \pm 40$; Grp 2vs Grp 3: $54 \pm 39$; Grp 1vs Grp 3: $44 \pm 26$. It shows that the SAT score of group 3 is significantly from those of group 1 and 2, meanwhile no strong evidence show that group 1 and group 2 are different.