MA409: Statistical Data Analysis (SAS)

Assignment 4 (Apr 30 – May 21)

Note: Please work on 4.2 by hand calculation (p-values can be obtained with any software) and 4.3-4.4 by SAS procedures.

4.1 Under two-way ANOVA model, $Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij}$, the variation between groups is defined as:

$$SSM = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}\left(y_{ijk} - \bar{y}\right)^2.$$

The variation between groups due to factor A, B, and interaction of A and B are defined as:

$$SSA = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{y}_{i\cdot} - \bar{y})^2, SSB = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}\left(\bar{y}_{\cdot j} - \bar{y}\right)^2,$$

$$SSAB = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}\left(\bar{y}_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}\right)^2.$$

Show that

(1) $E(SSAB) = (a-1)(b-1)\sigma^2 + \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\gamma_{ij}^2$. (5 points)

(2) $SSM = SSA + SSB + SSAB$ when the n_{ij}'s are all equal. (10 points)

4.2 There are three factories (factory A, B, and C) producing the same type of auto parts. The lifespan of six random samples from each of the three factories are given below:

	1	2	3	4	5	6
Factory A	40	47	38	42	45	46
Factory B	26	34	30	28	32	33
Factory C	39	40	48	50	49	32

Let $(\bar{y}_A, s_A^2, \mu_A, \sigma_A^2)$, $(\bar{y}_B, s_B^2, \mu_B, \sigma_B^2)$, $(\bar{y}_C, s_C^2, \mu_C, \sigma_C^2)$ denote the sample mean, sample variance, population mean, population variance of the lifespan of the auto parts produced by the factory A, B, and C, respectively.

(1) Compute $\bar{y}_A$, $\bar{y}_B$, $\bar{y}_C$, s_A^2, s_B^2, s_C^2, and the overall mean $\bar{y}$, variation between groups SSB, variation within groups SSW. Construct the ANOVA table based on the data. Test $H_0: \mu_A = \mu_B = \mu_C$ vs. $H_1: \mu_A, \mu_B, \mu_C$ are not all equal at significance level $\alpha = 0.05$. (10 points)

(2) Apply Levene's test with $z_{ij} = \left|y_{ij} - \bar{y}\right|$ to test equality of group variances at $\alpha = 0.05$:

$$H_0: \sigma_A^2 = \sigma_B^2 = \sigma_C^2 \text{ vs. } H_1: \sigma_A^2, \sigma_B^2, \sigma_C^2 \text{ are not all equal.}$$

Please provide the detailed steps of the test. (10 points)

(3) Apply the Kruskal-Wallis test to test the equality of group medians at $\alpha = 0.05$. Please provide the detailed steps of the test. (10 points)

4.3 Four workers from a glass manufacturer produces glassware with three different tools. Random samples on the number of glassware produced per day by a worker with a specific tool are given below

	Worker 1	**Worker 2**	**Worker 3**	**Worker4**
Tool A	14 10 12	11 11 10	13 19	10 12
Tool B	9 7	10 8 9	7 8 11	6
Tool C	5 11 8	13 14 11	12 13	14 10 12

(1) Do the data provide sufficient evidence to indicate differences among the mean number of glassware produced by the four workers? Use $\alpha = 0.1$. (5 points)

(2) Do the data provide sufficient evidence to indicate differences among the mean number of glassware produced with the three tools? Use $\alpha = 0.1$. (5 points)

(3) Let μ_A, μ_B, μ_C be the mean number of glassware produced with tool A, B, and C, respectively. Compute the simultaneous confidence interval of $\mu_A - \mu_B$, $\mu_A - \mu_C$, $\mu_B - \mu_C$ using the Tukey-Kramer method. Do the results indicate any difference between any pair of tools? Use $\alpha = 0.1$. (5 points)

(4) Do the data provide sufficient evidence to indicate an interaction effect between workers and tools? Use $\alpha = 0.1$. (5 points)

(5) For the two-way ANOVA model with the interaction effect between workers and tools, test whether the residuals follow a normal distribution. Use $\alpha = 0.1$ (5 points)

4.4 A big company wants to understand why its employees are leaving the company. The data "HRdata.csv" for 14999 employees are provided by the HR department including satisfaction level, latest evaluation (yearly), number of project worked on, average monthly hours, time spent in the company (in years), work accident (within the past 2 years, 0 – no, 1 - yes), promotion within the past 5 years (0 – no, 1 - yes), and salary level (low, medium, high). The response variable of interest is whether the employee left the company (0 – no, 1 – yes).

(1) Use graphical methods to explore the possible differences in the explanatory variables between employees who left and who didn't leave the company. Interpret your findings. (10 points)

(2) Fit the following logistic regression model on the probability that an employee would leave the company:

$$\begin{aligned}\text{logit}(p) = \beta_0 &+ \beta_1 \text{satisfaction\_level} + \beta_2 \text{last\_evaluation} + \beta_3 \text{number\_project} \\ &+ \beta_4 \text{average\_monthly\_hours} + \beta_5 \text{time\_spent\_company} \\ &+ \beta_6 \text{I}(\text{work\_accident} = 1) + \beta_7 \text{I}(\text{promotion\_last\_5years} = 1) \\ &+ \beta_8 \text{I}(\text{salary} = \text{medium}) + \beta_9 \text{I}(\text{salary} = \text{high}).\end{aligned}$$

Obtain the odds ratio estimates and interpret the odds ratios. (15 points)

(3) Plot the ROC curve of the model in (2), obtain the AUC, and interpret the AUC. (5 points)