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, $Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$, the variation between groups is defined as:

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

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

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

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

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)

Solution: For $y_{ijk} = \mu_{ij} + \varepsilon_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$, we obtain that (omitted the ij subscript for σ's)

$$\boldsymbol{y}_{ij} = \begin{pmatrix} \bar{y}_{ij} \\ \bar{y}_{i\cdot} \\ \bar{y}_{\cdot j} \\ \bar{y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_{ij} \\ \bar{\mu}_{i\cdot} \\ \bar{\mu}_{\cdot j} \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{pmatrix} \right).$$

Let $z_{ij} = \bar{y}_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y} = (1 \quad -1 \quad -1 \quad 1)\boldsymbol{y}_{ij}$, then

$$z_{ij} \sim \mathcal{N}\left(\mu_{ij} - \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot j} - \bar{\mu}, \sigma_z^{(ij)}\right) = \mathcal{N}(\gamma_{ij}, \sigma_z^{(ij)}),$$

With

$$\sigma_z^{(ij)} = (1 \quad -1 \quad -1 \quad 1)\begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{pmatrix}\begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix}$$

$$= (\sigma_{11} + \sigma_{22} + \sigma_{33} + \sigma_{44}) - 2(\sigma_{12} + \sigma_{13} + \sigma_{24} + \sigma_{34}) + 2(\sigma_{14} + \sigma_{23}).$$

So, $E(SSAB) = E\left(\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} z_{ij}^2\right) = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}\left\{\left[E(z_{ij})\right]^2 + \mathrm{Var}(z_{ij})\right\} =$

$\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\gamma_{ij}^2 + \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\sigma_z^{(ij)}$. It remains to show that $\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\sigma_z^{(ij)} = (a-1)(b-1)\sigma^2$.

We have: $(n_{i\cdot} = \sum_{j=1}^{b} n_{ij},\ n_{\cdot j} = \sum_{i=1}^{a} n_{ij},\ n = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij})$

$$\sigma_{11} = \mathrm{Var}(\bar{y}_{ij}) = \mathrm{Var}\left(\frac{\sum_{k=1}^{n_{ij}} y_{ij}}{n_{ij}}\right) = \frac{\sigma^2}{n_{ij}},$$

$$\sigma_{22} = \mathrm{Var}(\bar{y}_{i\cdot}) = \mathrm{Var}\left(\frac{\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ij}}{\sum_{j=1}^{b} n_{ij}}\right) = \frac{\sigma^2}{\sum_{j=1}^{b} n_{ij}} = \frac{\sigma^2}{n_{i\cdot}}, \sigma_{33} = \frac{\sigma^2}{\sum_{i=1}^{a} n_{ij}} = \frac{\sigma^2}{n_{\cdot j}},$$

$$\sigma_{44} = \mathrm{Var}\left(\frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ij}}{\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}}\right) = \frac{\sigma^2}{\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}} = \frac{\sigma^2}{n},$$

$$\sigma_{12} = \mathrm{Cov}(\bar{y}_{ij}, \bar{y}_{i\cdot}) = \mathrm{Cov}\left(\frac{\sum_{k=1}^{n_{ij}} y_{ij}}{n_{ij}}, \frac{\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ij}}{\sum_{j=1}^{b} n_{ij}}\right) = \frac{\sigma^2}{n_{i\cdot}}, \sigma_{13} = \frac{\sigma^2}{n_{\cdot j}},$$

$$\sigma_{14} = \mathrm{Cov}(\bar{y}_{ij}, \bar{y}) = \mathrm{Cov}\left(\frac{\sum_{k=1}^{n_{ij}} y_{ij}}{n_{ij}}, \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ij}}{\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}}\right) = \frac{\sigma^2}{\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}} = \frac{\sigma^2}{n},$$

$$\sigma_{23} = \mathrm{Cov}\left(\frac{\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ij}}{\sum_{j=1}^{b} n_{ij}}, \frac{\sum_{i=1}^{a}\sum_{k=1}^{n_{ij}} y_{ij}}{\sum_{i=1}^{a} n_{ij}}\right) = \frac{n_{ij}\sigma^2}{n_{i\cdot}n_{\cdot j}},$$

$$\sigma_{24} = \mathrm{Cov}\left(\frac{\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ij}}{\sum_{j=1}^{b} n_{ij}}, \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ij}}{\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}}\right) = \frac{\sigma^2}{n}, \sigma_{34} = \frac{\sigma^2}{n}.$$

Therefore

$$\sigma_z^{(ij)} = \left(\frac{\sigma^2}{n_{ij}} + \frac{\sigma^2}{n_{i\cdot}} + \frac{\sigma^2}{n_{\cdot j}} + \frac{\sigma^2}{n}\right) - 2\left(\frac{\sigma^2}{n_{i\cdot}} + \frac{\sigma^2}{n_{\cdot j}} + \frac{\sigma^2}{n} + \frac{\sigma^2}{n}\right) + 2\left(\frac{\sigma^2}{n} + \frac{n_{ij}\sigma^2}{n_{i\cdot}n_{\cdot j}}\right)$$

$$= \frac{\sigma^2}{n_{ij}} - \frac{\sigma^2}{n_{i\cdot}} - \frac{\sigma^2}{n_{\cdot j}} - \frac{\sigma^2}{n} + \frac{2n_{ij}\sigma^2}{n_{i\cdot}n_{\cdot j}}$$

$$\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\sigma_z^{(ij)} = \sigma^2 \sum_{i=1}^{a}\sum_{j=1}^{b}\left(1 - \frac{n_{ij}}{n_{i\cdot}} - \frac{n_{ij}}{n_{\cdot j}} - \frac{n_{ij}}{n} + \frac{2n_{ij}^2}{n_{i\cdot}n_{\cdot j}}\right)$$

$$= \sigma^2\left(ab - a - b - 1 - \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{2n_{ij}^2}{n_{i\cdot}n_{\cdot j}}\right)$$

When n_{ij}'s are all equal $\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{2n_{ij}^2}{n_{i\cdot}n_{\cdot j}} = 2 \Rightarrow \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\sigma_z^{(ij)} = (a-1)(b-1)\sigma^2.$

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

Solution: We have that

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

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

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

Let n_{ij} all equal to n_0, then:

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{y}_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y})(\bar{y}_{i\cdot} - \bar{y}) = n_0\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y})(\bar{y}_{i\cdot} - \bar{y})$$

$$= n_0\sum_{i=1}^{a}\left[(\bar{y}_{i\cdot} - \bar{y})\sum_{j=1}^{b}(\bar{y}_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y})\right]$$

$$= n_0\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y})(b\bar{y}_{i\cdot} - b\bar{y}_{i\cdot} - b\bar{y} + b\bar{y}) = 0.$$

Similarly

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{y}_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y})(\bar{y}_{\cdot j} - \bar{y}) = n_0\sum_{j=1}^{b}(\bar{y}_{\cdot j} - \bar{y})(a\bar{y}_{\cdot j} - a\bar{y}_{\cdot j} - a\bar{y} + a\bar{y}) = 0.$$

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{y}_{i\cdot} - \bar{y})(\bar{y}_{i\cdot} - \bar{y}) = n_0\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y})(b\bar{y} - b\bar{y}) = 0.$$

Therefore, $SSM = SSA + SSB + SSAB$.

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)

Solution:

$$\bar{y}_A = \frac{40+47+38+42+45+46}{6} = 43$$
$$\bar{y}_B = \frac{26+34+30+28+32+33}{6} = 30.5$$
$$\bar{y}_C = \frac{39+40+48+50+49+32}{6} = 43$$
$$s_A^2 = \frac{(40-43)^2+(47-43)^2+\cdots+(46-43)^2}{5} = 12.8$$
$$s_B^2 = \frac{(26-30.5)^2+(34-30.5)^2+\cdots+(33-30.5)^2}{5} = 9.5$$
$$s_C^2 = \frac{(39-43)^2+(40-43)^2+\cdots+(32-43)^2}{5} = 51.2$$
$$\bar{y} = \frac{\bar{y}_A+\bar{y}_B+\bar{y}_C}{3} = 38.33$$
$$SSB = \sum n_i(\bar{y}_i-\bar{y})^2 = 6[(43-38.33)^2+(30.5-38.33)^2+(43-38.33)^2] = 625$$
$$SSW = \sum(n_i-1)s_i^2 = 5s_A^2+5s_B^2+5s_C^2 = 367.5$$

Therefore, the ANOVA table is:

ANOVA Table					
Source	DF	Sum of Squares	Mean Square	F value	P-value
Between	2	625	312.5	12.76	0.0006
Within	15	367.5	24.5		
Total	17	992.5			

As P-value < 0.05, $H_0: \mu_A = \mu_B = \mu_C$ is rejected at significance level $\alpha = 0.05$, indicating that the lifespans of the auto parts produced by the three factories are not all equal.

(2) Apply Levene's test with $z_{ij} = |y_{ij} - \bar{y}_i|$ 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)

Solution: For the Levene's test, compute the dispersion variable $z_{ij} = |y_{ij} - \bar{y}_i|$:

Factory	A	B	C
z_{ij}	3 4 5 1 2 3	4.5 3.5 0.5 2.5 1.5 2.5	4 3 5 7 6 11

Then perform ANOVA on z_{ij}. The ANOVA table based on z_{ij} following the same computation steps in (1) is

ANOVA Table					
Source	DF	Sum of Squares	Mean Square	F value	P-value
Between	2	43	21.5	5.38	0.0174
Within	15	60	4		
Total	17	103			

As P-value < 0.05, $H_0: \sigma_A^2 = \sigma_B^2 = \sigma_C^2$ is rejected at significance level $\alpha = 0.05$, indicating that the equal variance assumption is violated.

(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)

Solution:

Step I: Rank the observations ignoring group membership:

y_{ij}	26	28	30	32	32	33	34	38	39	40	40	42	45	46	47	48	49	50
r_{ij}	1	2	3	4.5	4.5	6	7	8	9	10.5	10.5	12	13	14	15	16	17	18
Group	B	B	B	B	C	B	B	A	C	A	C	A	A	A	A	C	C	C

Then we can obtain:

$$\bar{r}_A = \frac{8 + 10.5 + 12 + 13 + 14 + 15}{6} = 12.0833$$

$$\bar{r}_B = \frac{1 + 2 + 3 + 4.5 + 6 + 7}{6} = 3.9167$$

$$\bar{r}_C = \frac{4.5 + 9 + 10.5 + 16 + 17 + 18}{6} = 12.5$$

$$\bar{r} = \frac{\bar{r}_A + \bar{r}_B + \bar{r}_C}{3} = 9.5$$

So that the test statistic of the Kruskal-Wallis test is

$$KW = \frac{(n-1)\sum n_i(\bar{r}_i - \bar{r})^2}{\sum\sum(r_{ij} - \bar{r})^2} = 9.8830.$$

AS $KW \sim_{approx} \chi_2^2$ under H_0, the P-value is computed to be $0.0071 < 0.05$. The equality of group medians is rejected at $\alpha = 0.05$.

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)

Solution: The result based on the one-way ANOVA model to test the equality of mean number of glassware produced by the four workers is given below:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	20.9206349	6.9735450	0.84	0.4841
Error	26	215.7460317	8.2979243		
Corrected Total	29	236.6666667			

The P-value= $0.4841 > 0.1$, indicating that the data do not provide enough evidence that the mean numbers of glassware produced by the four workers are different.

(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)

Solution: The result based on the one-way ANOVA model to test the equality of mean number of glassware produced with the three tools is given below:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	75.4303030	37.7151515	6.32	0.0056
Error	27	161.2363636	5.9717172		
Corrected Total	29	236.6666667			

The P-value= $0.0056 < 0.1$, indicating that the data provide sufficient evidence that the mean numbers of glassware produced with the three tools are different.

(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)

Solution: The simultaneous confidence intervals using the Tukey-Kramer method is

Comparisons significant at the 0.1 level are indicated by ***.				
tool Comparison	Difference Between Means	Simultaneous 90% Confidence Limits		
A - C	1.018	-1.270	3.306	
A - B	3.867	1.461	6.273	***
C - A	-1.018	-3.306	1.270	
C - B	2.848	0.495	5.202	***
B - A	-3.867	-6.273	-1.461	***
B - C	-2.848	-5.202	-0.495	***

The confidence interval of $\mu_A - \mu_B$ is $[1.461, 6.273]$, indicates a significant difference between tool A and tool B.

The confidence interval of $\mu_A - \mu_C$ is $[-1.270, 3.306]$, indicates no significant difference between tool A and tool C.

The confidence interval of $\mu_B - \mu_C$ is $[-5.202, -0.495]$, indicates a significant

difference between tool B and tool C.

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

Solution: The result based on the two-way ANOVA model is given below

Source	DF	Type III SS	Mean Square	F Value	Pr > F
worker	3	38.52257123	12.84085708	3.19	0.0487
tool	2	89.05192878	44.52596439	11.05	0.0007
worker*tool	6	53.92077518	8.98679586	2.23	0.0874

The P-value of the interaction effect between workers and tools is $0.0874 < 0.1$, so that the data provide sufficient evidence to indicate an interaction effect between workers and tools.

(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)
Solution: Output the residuals from the two-way ANOVA model using PROC GLM, then test normality of the residuals using PROC UNIVARIATE, the result is given below

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.97525	Pr < W	0.6901
Kolmogorov-Smirnov	D	0.1	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.036512	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.235878	Pr > A-Sq	>0.2500

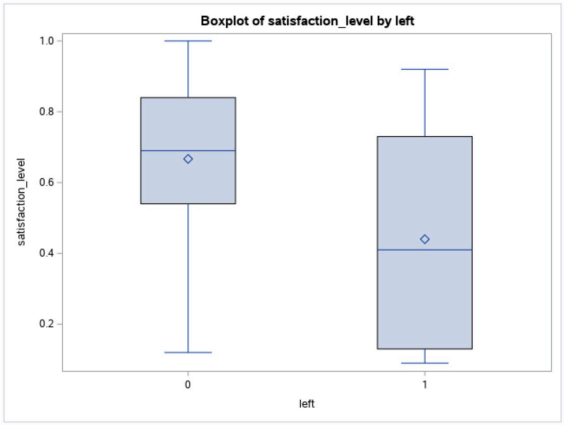
As the sample size is small, we could look at the Shapiro-Wilk test, the P-value is $0.6901 > 0.1$, therefore the data do not provide enough evidence to indicate departure from a normal distribution for the residuals.

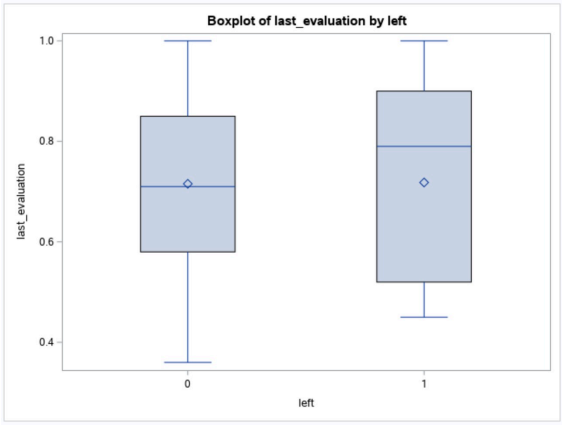
4.4 A big company wants to understand why its employees are leaving the company. The data "HRdata.csv" for 14999 employees are provided by the HR department including satisfaction level, latest evaluation (yearly), number of projects worked on, average monthly hours, time spent in the company (in years), 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).

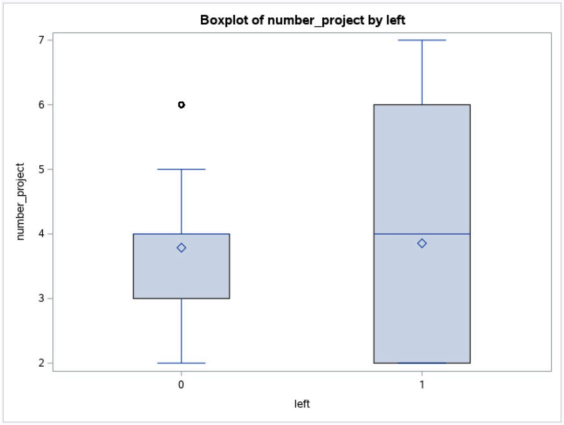
(1) Use graphical methods to explore the possible differences in the explanatory variables between employees who left and who didn't leave the company. Interpret your findings. (10 points)

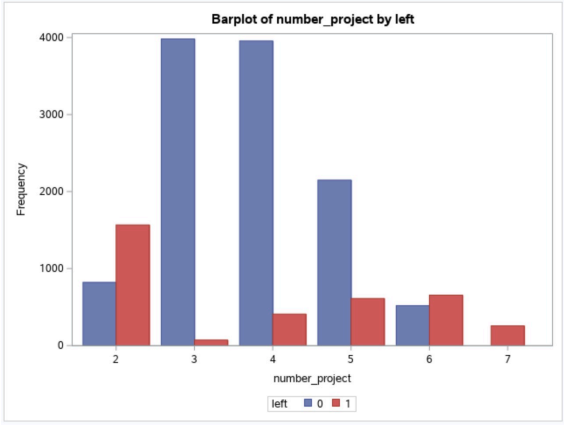
Solution: Treat satisfaction level, last evaluation, number of projects worked on, average monthly hours, time spent in the company as numerical variables, obtain the boxplots of

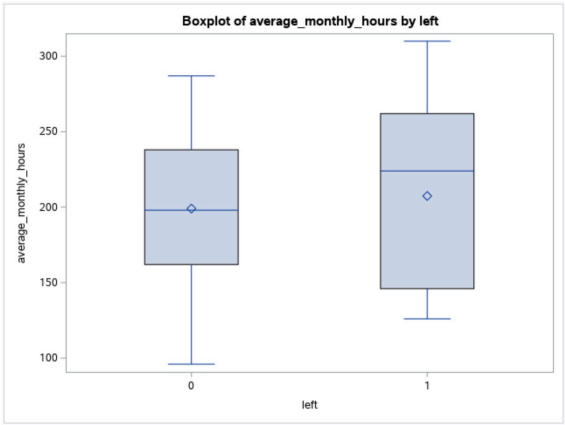
these variables by whether the employee left the company:

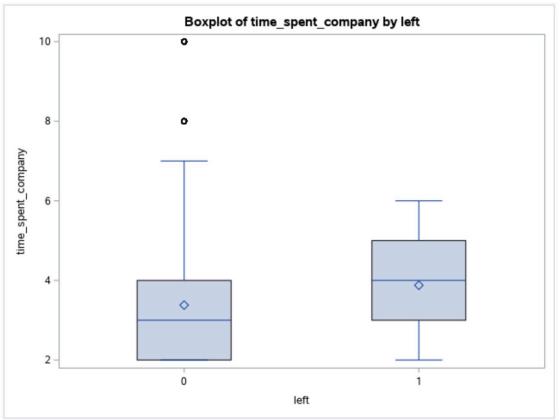












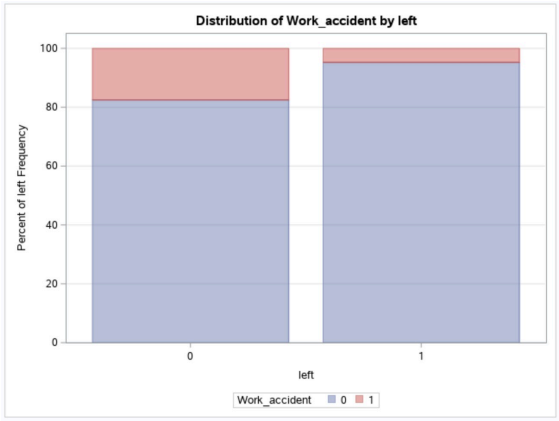
The plots show that:

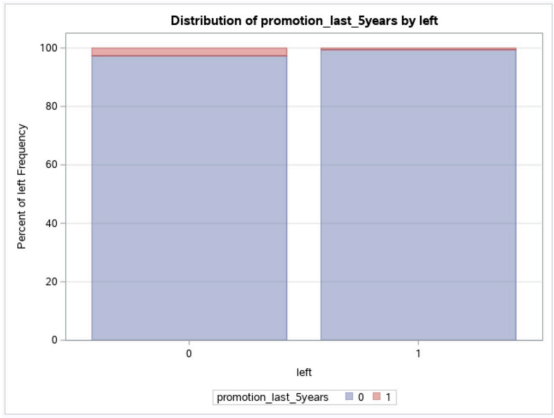
- The employees who left the company seem to have lower average satisfaction level compared to those who stayed. This is not surprising, as employees who are less satisfied with the company tend to leave.
- The employees who left the company seem to have higher average last evaluation compared to those who stayed. It seems surprising at the first glance, but actually it is reasonable. As employees with higher evaluation are more outstanding, they may attract

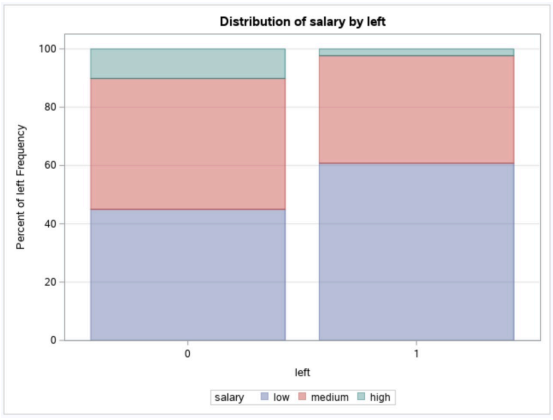
more headhunters and are easier to get positions in other companies.

- The boxplot of the number of projects worked on does not provide much information so we also look at the barplot of it. The barplot indicates that employees who worked on 2, 6, or 7 projects seem to have a higher probability of leaving the company.
- The employees who left the company seem to have higher average monthly hours compared to those who stayed.
- The employees who left the company seem to have higher average time spent in the company compared to those who stayed.

On the other hand, the work accident, promotion within the past 5 years, and salary level are treated as categorical variables, the distributions of these variables by whether the employee left the company are:







The plots show that:

- The employees who left the company seem to have lower proportion of work accident compared to those who stayed.
- The percentages of employees who got promotion in the past 5 years are low, but the percentage seems to be higher among those who stayed in the company.
- The employees who left the company seem to have lower proportion of medium or high salary level compared to those who left.

(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)

Solution: Fit the logistic regression model with PROC LOGISTIC and make sure that the reference levels of the categorical variables are correctly specified, the odds ratio estimates are given below:

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
satisfaction_level	0.016	0.013	0.019
last_evaluation	2.068	1.545	2.767
number_project	0.730	0.700	0.761
average_monthly_hour	1.004	1.003	1.005
time_spent_company	1.299	1.260	1.338
Work_accident 1 vs 0	0.215	0.181	0.256
promotion_last_5year 1 vs 0	0.227	0.138	0.375
salary high vs low	0.135	0.105	0.173
salary medium vs low	0.586	0.536	0.641

The odds rate estimates indicate that:

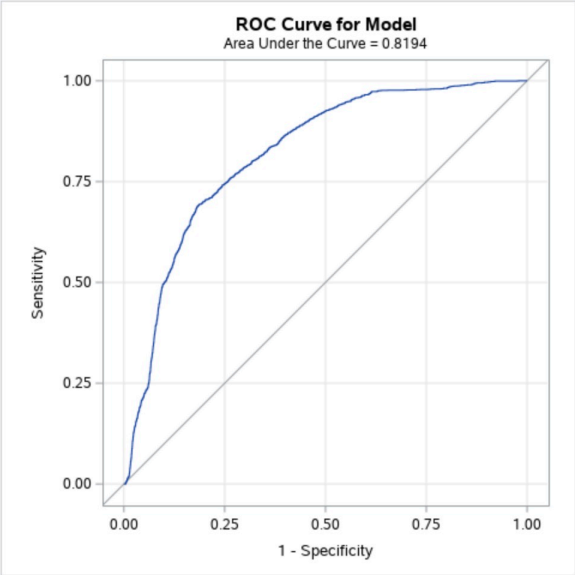
- With other variables fixed, an employee with 1-unit higher satisfaction level has on average 0.016 times the odds of leaving the company.
- With other variables fixed, an employee with 1-unit higher last evaluation has on average 2.068 times the odds of leaving the company.
- With other variables fixed, an employee with 1 more project experience has on average 0.730 times the odds of leaving the company.
- With other variables fixed, an employee with 1 more monthly working hour has on average 1.004 times the odds of leaving the company.
- With other variables fixed, an employee with 1 more year spent in the company has on average 1.299 times the odds of leaving the company.
- With other variables fixed, an employee with work accident has on average 0.215 times the odds of leaving the company compared to one without work accident.
- With other variables fixed, an employee with promotion in the past 5 years has on average 0.227 times the odds of leaving the company compared to one without promotion.
- With other variables fixed, an employee with high salary has on average 0.135 times

the odds of leaving the company compared to one with low salary level.

- With other variables fixed, an employee with medium salary level has on average 0.586 times the odds of leaving the company compared to one with low salary level.

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

Solution: The ROC curve is given below:



The AUC= 0.8194, which means that if a positive sample (an employee who left the company) and a negative sample (an employee who stayed) are randomly chosen, then the probability that the positive sample has a higher predicted value than the negative sample from the logistic regression model is 0.8194.