

STAT 6338: ADVANCED STATISTICAL METHODS II

Homework 3

March 10, 2015.

1 Problem 18.4

a) The residuals appear to have the same variance.

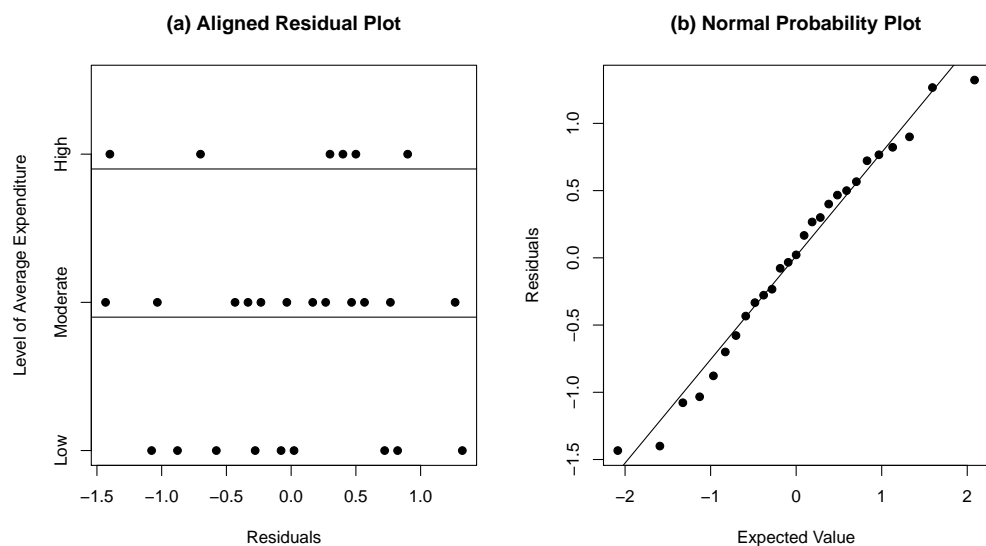


Figure 1: *Aligned residual dot plots by level of expenditure and Normal probability plot*

b) The correlation coefficient between the ordered residuals and their expected values under normality is 0.9919. From the correlation coefficient and the Normal probability plot in Figure 1(b), we can say that the normality assumption appear to be reasonable here.

c) Studentized deleted residuals, t_{ij} :

	j											
i	1	2	3	4	5	6	7	8	9	10	11	12
1	0.956	1.838	-0.101	-1.462	0.029	-0.362	-0.759	1.095	-1.173			
2	-1.982	-0.043	1.720	0.601	-0.428	-0.558	1.001	-0.299	0.213	0.733	-1.374	0.342
3	-0.957	0.677	1.246	-2.039	0.540	0.404						

Hypothesis:

H_0 : There are no outliers

H_a : There is at least one outlier

at $\alpha = 0.01$ familywise error rate.

Decision Rule: If $|t_{ij}| \geq t_{(0.999815, 23)} = 4.168$ reject H_0 , otherwise fail to reject H_0 .

Conclusion: There are no outliers since all $|t_{ij}| \leq 4.168$.

- d) From Figure 2, it doesn't appear that the model can be improved. This is because the residuals do not segregate under home office locations.

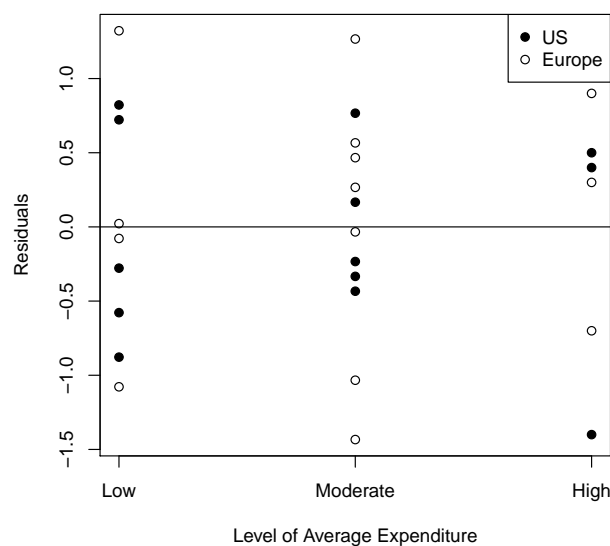
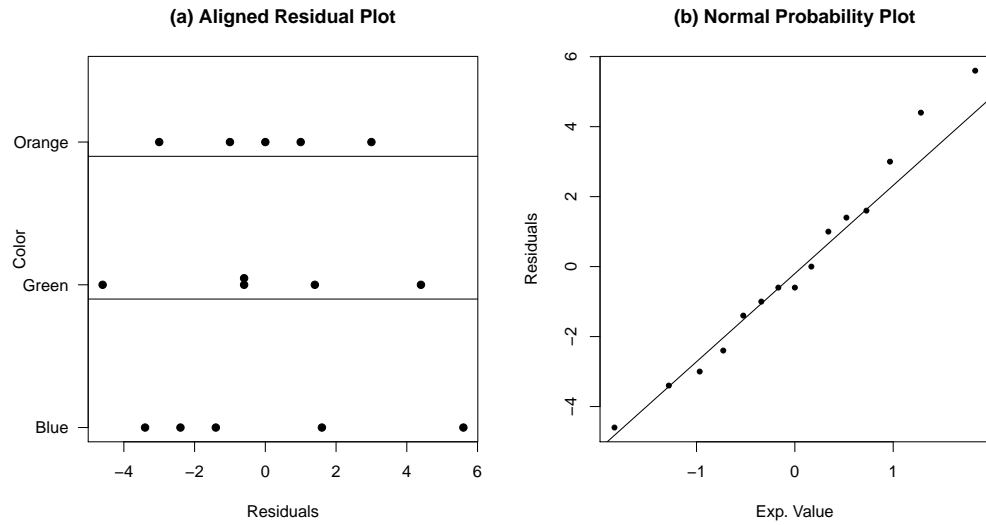
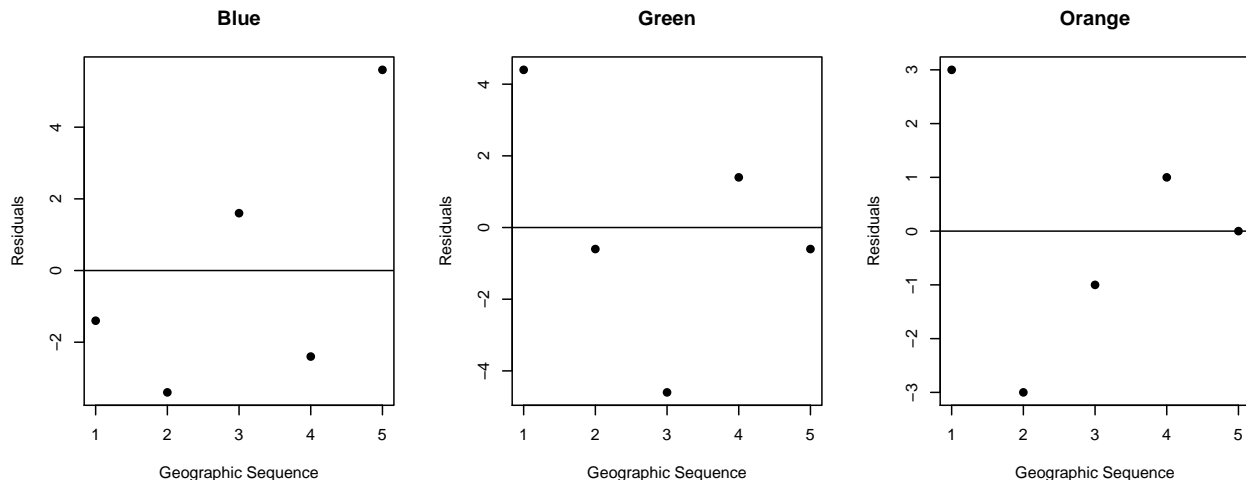


Figure 2: *Aligned residual dot plots by level of expenditure with with home office identified*

2 Problem 18.5

- From Figure 3(a), the constancy of variance assumption for errors is violated.
- The coefficient of correlation between the ordered residuals and their expected value under normality is 0.9908. From Figure 3(b), the normality assumption appears to be reasonable.
- From Figure 4, the residuals are uncorrelated.
- Studentized deleted residuals, t_{ij} :

Figure 3: *Aligned residual dot plots by color and Normal probability plot*Figure 4: *Residual sequence plots by color*

Color	j				
	1	2	3	4	5
Blue (1)	-0.486	-1.249	0.557	-0.852	2.363
Green (2)	1.699	-0.207	-1.798	0.486	-0.207
Orange (3)	1.085	-1.085	-0.346	0.346	0.000

Hypothesis:

H_0 : There are no outliers.

H_a : There is at least an outlier.

at $\alpha = 0.025$ familywise error rate.

Decision Rule: If $|t_{ij}| \geq t_{(0.99917, 11)} = 4.1319$ reject H_0 , otherwise fail to reject H_0 .

Conclusion: There are no outliers since all $|t_{ij}| \leq 4.1319$.

3 Problem 18.11

Hypothesis:

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ (The residuals have constant variance)

$H_a : \text{at least one } \sigma_i^2, i = 1, 2, 3, \text{ is different.}$

at 5% significance level.

Decision Rule: If $F^* > F_{(0.95, 2, 24)} = 3.4028$, reject H_0 , otherwise fail to reject H_0

Conclusion: Brown-Forsythe test statistic $F^* = 0.0245 < 3.4028$. At 5% significance level, we fail to reject H_0 and conclude that the residuals have constant variance. The $p - \text{value} = 0.9758$.

4 Problem 18.23

a) Hypothesis:

$H_0 : \mu_1 = \mu_2 = \mu_3$ (Mean infection risk is the same for all regions)

$H_a : \text{at least one } \mu_i, i = 1, 2, 3, \text{ is different.}$

at 5% significance level.

ANOVA Table

Source	df	SS	MS	F-value	p-value
Treatment	2	941.63	470.81	16.25	3.452×10^{-5}
Residuals	24	695.38	28.97		
Total	26	1637.01			

Decision: If $F_R^* > F_{(0.95, 2, 24)} = 3.4028$ reject H_0 , otherwise fail to reject H_a .

Conclusion: Since $F_R^* = 16.25 > 3.4028$, we reject the null hypothesis. Hence there is enough evidence to conclude that, at 5% level of significance, at least one of the factor level means differ.

b) $p - \text{value} = 3.452 \times 10^{-5}$

c) The conclusion is the same as the one in Problem 16.7e.

d) The non-parametric test is not needed since the normality and constancy of variance assumptions were shown to be satisfied in Problems 18.4 and 18.11 respectively.

e) $100(1 - \alpha)\%$ Bonferroni confidence interval for pairwise comparison based on ranks is given by

$$\bar{R}_i. - \bar{R}_{i'}. \pm B \sqrt{\frac{n_T(n_T + 1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_{i'}} \right)}$$

where

$$B = z_{\left(1 - \frac{\alpha}{2g}\right)} \quad \text{and} \quad g = \frac{r(r-1)}{2}$$

Level(<i>i</i>)	Low (1)	Moderate (2)	High (3)
n_i	9	12	6
\bar{R}_i	6.50	15.50	22.25

For a 90% family confidence coefficient, the Bonferroni confidence intervals for all pairwise comparisons between three levels are:

Pairwise Comparison	Difference $\bar{R}_i - \bar{R}_{i'}$	90% Bonferroni Confidence Interval	
		Lower	Upper
$\bar{R}_2 - \bar{R}_1$	9.00	1.5518	16.4482
$\bar{R}_3 - \bar{R}_1$	15.75	6.8477	24.6523
$\bar{R}_3 - \bar{R}_2$	6.75	-1.6954	15.1954

Moderate level is not significantly different from High level. However, Low level is significantly different from the other two levels.

5 Problem 19.4

a) The factor A level means are:

$$\mu_{1.} = \frac{34 + 23 + 36}{3} = 31, \quad \mu_{2.} = \frac{40 + 29 + 42}{3} = 37$$

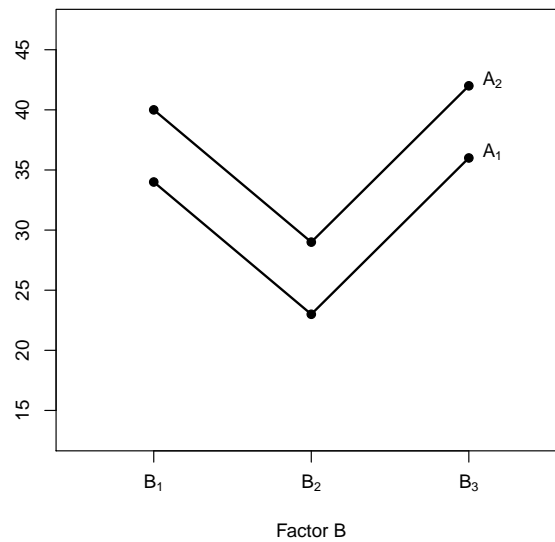
b)

$$\mu_{..} = \frac{34 + 23 + 36 + 40 + 29 + 42}{6} = 34$$

The main effects of factor A are:

$$\alpha_1 = \mu_{1.} - \mu_{..} = 31 - 34 = -3, \quad \alpha_2 = \mu_{2.} - \mu_{..} = 37 - 34 = 3$$

- c) No, this doesn't imply A and B interact. These are just difference between the mean responses between two levels in factor B . For interaction to be present, we need to have, at least, the difference between the mean responses for two levels of factor B to be different for the levels of factor A .
- d) From Figure 5, we can see that the treatment mean curves for the different level of factor A are parallel. Therefore, there is no interaction between the two factors.

Figure 5: *Treatment means plot*

6 Problem 19.6

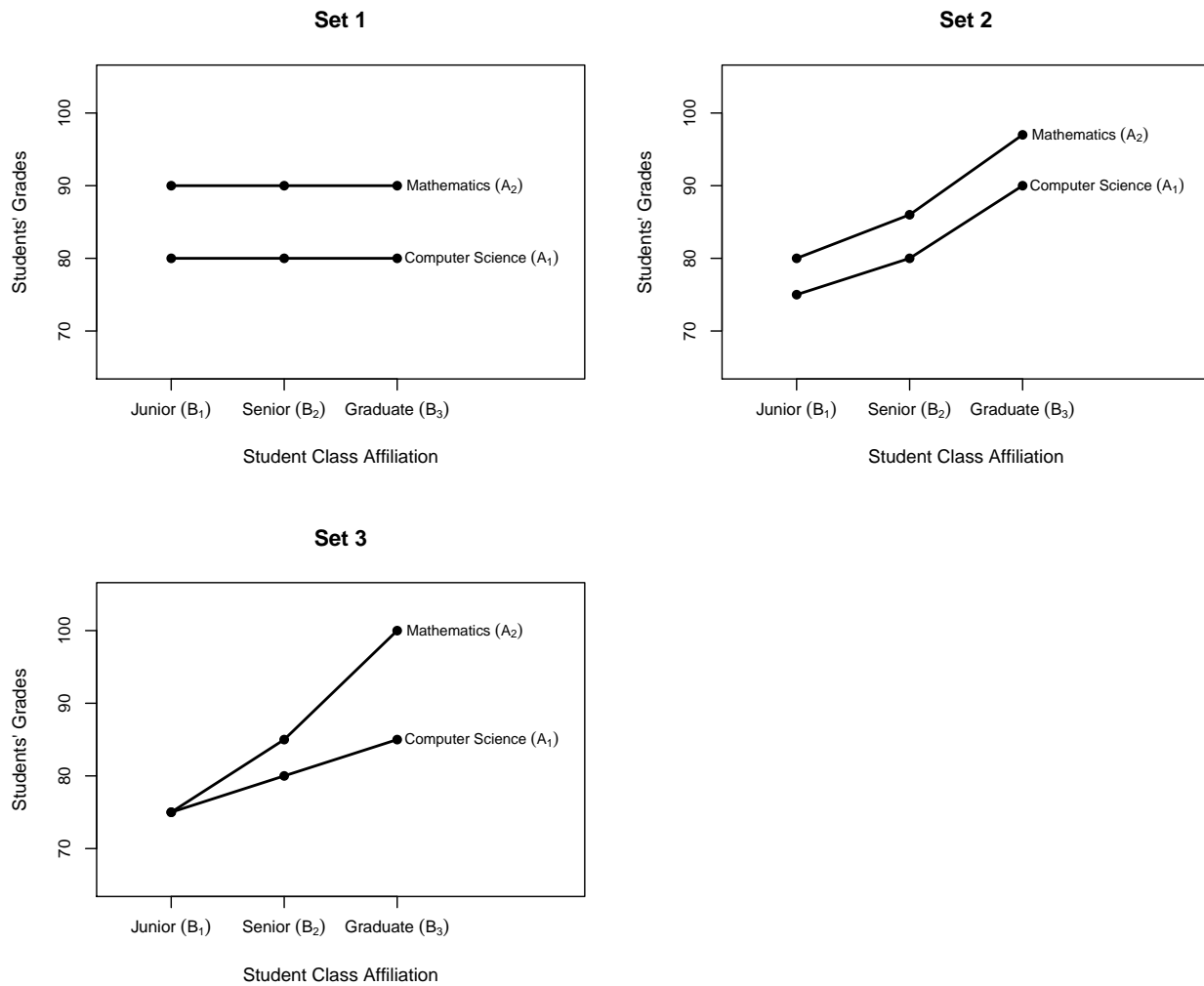
- i) In Set 1 (fig 6), the zero slope of each curve indicates that *student's class affiliation* has no effect on *student's grades*. The difference in the heights of the two curves show the *student's major* effect. There are no interactions in this set.
- ii) In Set 2 (fig 6), *student's class affiliation*, as well as *student's major*, affects *student's grades*. The difference in the height of the two curves reflects the *student's affiliation* difference and the departure from horizontal for each of the curves reflects *student's major* effect. The two curves are almost parallel. The interaction effects are so small that they are considered unimportant interactions.
- iii) In Set 3 (fig 6), the two factors interact. They indicate that there are relatively large *student's grade* for *graduates*. The interaction would therefore be considered as important interactions.

7 Problem 19.7

- a) $\sigma = 1.4$, $n = 10$, $a = 2$, $b = 3$, $\alpha_1 = -3$, $\alpha_2 = 3$.

$$E\{MSE\} = \sigma^2 = 1.4^2 = 1.96, \quad E\{MSA\} = \sigma^2 + nb \frac{\sum \alpha_i^2}{a-1} = 1.4^2 + (10)(3) \frac{(-3)^2 + (3)^2}{2-1} = 541.96$$

- b) $E\{MSA\}$ is substantially larger than $E\{MSE\}$. There is factor A main effects.

Figure 6: *Treatment means plot for each set of μ_{ij}*

8 Problem 19.20

a) Fitted values for ANOVA model:

$$\begin{array}{lll} \bar{Y}_{11\cdot} = 222.00 & \bar{Y}_{12\cdot} = 106.50 & \bar{Y}_{13\cdot} = 60.50 \\ \bar{Y}_{21\cdot} = 62.25 & \bar{Y}_{22\cdot} = 44.75 & \bar{Y}_{23\cdot} = 38.75 \end{array}$$

b) e_{ijk} :

	$j = 1$	$j = 2$	$j = 3$
$i = 1$	18.00	3.50	-4.50
	-16.00	11.50	-0.50
	-5.00	-3.50	7.50
	3.00	-11.50	-2.50
$i = 2$	8.75	2.25	-1.75
	-9.25	7.25	-5.75
	5.75	-13.75	1.25
	-5.25	4.25	6.25

c) From Figure 7(a), the residuals seems to have non-constant (increasing) variance.

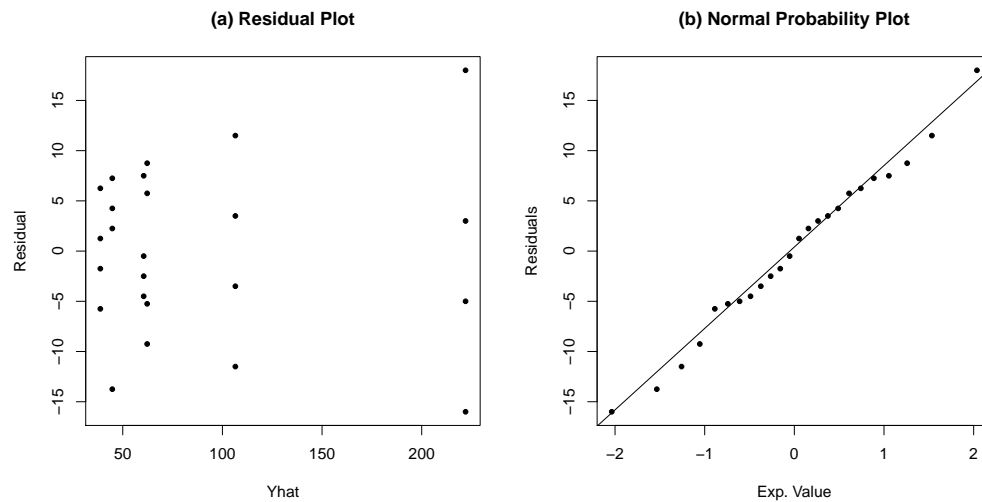


Figure 7: Residual dot plots for the treatment (a) and normal probability plot of the residuals (b)

d) The coefficient of correlation between the ordered residuals and their expected values under normality is 0.9942. Thus, from Figure 7(b) and the correlation coefficient, the normality assumption is reasonable here.

9 Problem 19.21

a) From Figure 8, factor effects are present. Programmers with *small systems only* experience had higher prediction error than their counterparts with *small and large systems* experience. Also, programmers with more years of experience tend to have less prediction errors on the average.

b) ANOVA Table:

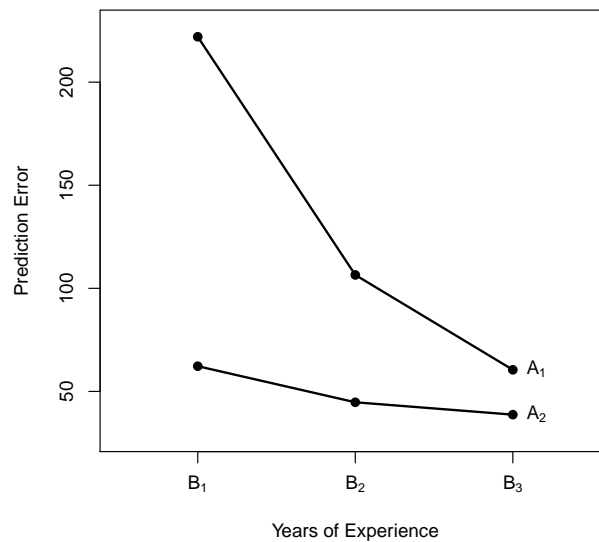


Figure 8: *Treatment means plot for prediction errors from programmers with different type and years of experience*

Source	df	SS	MS	<i>F</i> value	<i>p</i> -value
Treatments	5	96024.38	19204.88	222.98	1.55×10^{-15}
A (Type)	1	39447.04	39447.04	458.02	2.98×10^{-14}
B (Years)	2	36412.00	18206.00	211.39	3.16×10^{-13}
AB Interactions	2	20165.33	10083.67	117.07	4.82×10^{-11}
Residuals	18	1550.25	86.13		
Total	23	97574.63			

No source accounts for most of the total variability. This is because, the sums of squares accounted for by factor A , factor B , and their interaction term AB are not much different.

c) Hypothesis:

H_0 : all $(\alpha\beta)_{ij} = 0$ (The two factors don't interact)

H_a : not all $(\alpha\beta)_{ij} = 0$ (The two factors interact)

at $\alpha = 1\%$ significance level.

Test Statistics: $F^* = 117.07$

Decision: If $F^* > F_{0.99,2,18} = 6.01$ reject H_0 , otherwise fail to reject H_0 .

Conclusion: Since $F^* > 6.01$, we reject H_0 and conclude that there is interaction between the type of experience and years of experience of a programmer, at 1% significance level. $p\text{-value} = 4.82 \times 10^{-11}$

d) Hypothesis:

H_0 : $\alpha_1 = \alpha_2 = 0$ (There is no main effect for type of experience)

$H_a : \alpha_1 \neq 0$ or $\alpha_2 \neq 0$ (There is main effect for type of experience)

at $\alpha = 1\%$ significance level.

Test Statistics: $F^* = 458.02$

Decision: If $F^* > F_{0.99,1,18} = 8.29$ reject H_0 , otherwise fail to reject H_0 .

Conclusion: Since $F^* > 8.29$, we reject H_0 and conclude that there is main effect for type of experience of a programmer, at 1% significance level. $p\text{-value} = 2.98 \times 10^{-14}$

Hypothesis:

$H_0 : \beta_1 = \beta_2 = 0$ (There is no main effect for years of experience)

$H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$ (There is main effect for years of experience)

at $\alpha = 1\%$ significance level.

Test Statistics: $F^* = 211.39$

Decision: If $F^* > F_{0.99,2,18} = 6.01$ reject H_0 , otherwise fail to reject H_0 .

Conclusion: Since $F^* > 6.01$, we reject H_0 and conclude that there is main effect for years of experience of a programmer, at 1% significance level. $p\text{-value} = 3.16 \times 10^{-13}$

e) $\alpha \leq 1 - (0.99)^3 = 0.0297$.

f) The results in (c) and (d) confirms the analysis on the graph in part (a).

10 Problem 23.6

a) ANOVA model:

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

Equivalent regression model:

$$Y_{ijk} = \mu_{..} + \alpha_1 X_{ijk1} + \alpha_2 X_{ijk2} + \beta_1 X_{ijk3} + (\alpha\beta)_{11} X_{ijk1} X_{ijk3} + (\alpha\beta)_{21} X_{ijk2} X_{ijk3} + \varepsilon_{ijk}$$

$$X_{ijk1} = \begin{cases} 1, & \text{if case from level 1 for factor A (Young)} \\ -1, & \text{if case from level 3 for factor A (Elderly)} \\ 0, & \text{otherwise} \end{cases}$$

$$X_{ijk2} = \begin{cases} 1, & \text{if case from level 2 for factor A (Middle)} \\ -1, & \text{if case from level 3 for factor A (Elderly)} \\ 0, & \text{otherwise} \end{cases}$$

$$X_{ijk3} = \begin{cases} 1, & \text{if case from level 1 for factor B (Male)} \\ -1, & \text{if case from level 2 for factor B (Female)} \end{cases}$$

b)

$$\beta = \begin{bmatrix} \mu_{..} & \alpha_1 & \alpha_2 & \beta_1 & (\alpha\beta)_{11} & (\alpha\beta)_{21} \end{bmatrix}'$$

A	B	Freq	X_{ijk1}	X_{ijk2}	X_{ijk3}	$X_{ijk1}X_{ijk3}$	$X_{ijk2}X_{ijk3}$
1	1	6	1	1	0	1	0
1	2	6	1	1	0	-1	-1
2	1	5	1	0	1	1	0
2	2	6	1	0	1	-1	0
3	1	6	1	-1	-1	1	-1
3	2	5	1	-1	-1	-1	1

c) $X\beta$:

A	B	$X_{ij}\beta$
1	1	$\mu_{..} + \alpha_1 + \beta_1 + (\alpha\beta)_{11}$
1	2	$\mu_{..} + \alpha_1 - \beta_1 - (\alpha\beta)_{11} = \mu_{..} + \alpha_1 + \beta_2 + (\alpha\beta)_{12}$
2	1	$\mu_{..} + \alpha_2 + \beta_1 + (\alpha\beta)_{12}$
2	2	$\mu_{..} + \alpha_2 - \beta_1 - (\alpha\beta)_{12} = \mu_{..} + \alpha_2 + \beta_2 + (\alpha\beta)_{22}$
3	1	$\mu_{..} - \alpha_1 - \alpha_2 + \beta_1 - (\alpha\beta)_{11} - (\alpha\beta)_{12} = \mu_{..} + \alpha_3 + \beta_1 + (\alpha\beta)_{31}$
3	2	$\mu_{..} - \alpha_1 - \alpha_2 - \beta_1 + (\alpha\beta)_{11} + (\alpha\beta)_{12} = \mu_{..} + \alpha_3 + \beta_2 + (\alpha\beta)_{32}$

d) The reduced regression model for testing for interaction effects:

$$Y_{ijk} = \mu_{..} + \alpha_1 X_{ijk1} + \alpha_2 X_{ijk2} + \beta_1 X_{ijk3} + \varepsilon_{ijk}$$

e) The fitted **full model** is

$$\hat{Y} = 23.5667 - 2.0667X_1 + 4.1667X_2 + 0.3667X_3 - 0.2X_1X_3 - 0.3X_2X_3$$

and $SSE(F) = 71.333$ with $df_F = 28$.The fitted **reduced model** is

$$\hat{Y} = 23.5909 - 2.0909X_1 + 4.1691X_2 + 0.3602X_3$$

and $SSE(R) = 75.521$ with $df_R = 30$.Hypothesis: $H_0 : (\alpha\beta)_{11} = (\alpha\beta)_{21} = 0$ (There is no interaction effect) $H_a : (\alpha\beta)_{11} \neq 0$ or $(\alpha\beta)_{21} \neq 0$ (There is interaction effect)

at $\alpha = 5\%$ significance level.

Test Statistic:

$$F^* = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F} = \frac{(75.521 - 71.333)/(30 - 28)}{71.333/28} = 0.8219$$

Decision Rule: If $F^* > F_{(0.95, 2, 28)} = 3.3404$, we reject H_0 , otherwise we fail to reject H_0 .

Conclusion: Since $F^* < 3.3404$, we fail to reject H_0 and conclude that, at 5% significance level, there is no interaction effect between age and gender. The p -value = 0.4499.

f) Fitted reduced model to test **age effect**:

$$\hat{Y} = 23.5 + 0.1768X_3 - 0.0101X_1X_3 - 0.4949X_2X_3$$

and $SSE(R) = 359.94$ with $df_R = 30$.

Hypothesis:

$H_0 : \alpha_1 = \alpha_2 = 0$ (There is no age main effect)

$H_a : \alpha_1 \neq 0$ or $\alpha_2 \neq 0$ (There is age main effect)

at $\alpha = 5\%$ significance level.

Test Statistic:

$$F^* = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F} = \frac{(359.94 - 71.333)/(30 - 28)}{71.333/28} = 56.6428$$

Decision Rule: If $F^* > F_{(0.95, 2, 28)} = 3.3404$, we reject H_0 , otherwise we fail to reject H_0 .

Conclusion: Since $F^* > 3.3404$, we reject H_0 and conclude that, at 5% significance level, there is age main effect on the cash offer. The p -value = 1.4416×10^{-10} .

Fitted reduced model to test **gender effect**:

$$\hat{Y} = 23.5667 - 2.0667X_1 + 4.1323X_2 - 0.1771X_1X_3 - 0.3115X_2X_3$$

and $SSE(R) = 75.871$ with $df_R = 29$.

Hypothesis:

$H_0 : \beta_1 = 0$ (There is no gender main effect)

$H_a : \beta_1 \neq 0$ (There is gender main effect)

at $\alpha = 5\%$ significance level.

Test Statistic:

$$F^* = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F} = \frac{(75.871 - 71.333)/(29 - 28)}{71.333/28} = 1.7813$$

Decision Rule: If $F^* > F_{(0.95,1,28)} = 4.196$, we reject H_0 , otherwise we fail to reject H_0 .

Conclusion: Since $F^* < 4.196$, we fail to reject H_0 and conclude that, at 5% significance level, there is no gender main effect on the cash offer. The p -value = 0.187.

$$g) \hat{\alpha}_1 = -2.0667, \hat{\alpha}_2 = 4.1667, s^2\{\hat{\alpha}_1\} = 0.1462, s^2\{\hat{\alpha}_2\} = 0.1534, s\{\hat{\alpha}_1, \hat{\alpha}_2\} = -0.0734$$

$$\hat{D}_1 = \hat{\alpha}_1 - \hat{\alpha}_2 = -6.2334$$

$$\hat{D}_2 = \hat{\alpha}_1 - \hat{\alpha}_3 = \hat{\alpha}_1 - (-\hat{\alpha}_1 - \hat{\alpha}_2) = 2\hat{\alpha}_1 + \hat{\alpha}_2 = 0.0333$$

$$\hat{D}_3 = \hat{\alpha}_2 - \hat{\alpha}_3 = \hat{\alpha}_2 - (-\hat{\alpha}_1 - \hat{\alpha}_2) = \hat{\alpha}_1 + 2\hat{\alpha}_2 = 6.2667$$

$$s^2\{\hat{D}_1\} = s^2\{\hat{\alpha}_1\} + s^2\{\hat{\alpha}_2\} - 2s\{\hat{\alpha}_1, \hat{\alpha}_2\} = 0.1462 + 0.1534 - 2(-0.0734) = 0.4463$$

$$s^2\{\hat{D}_2\} = 4s^2\{\hat{\alpha}_1\} + s^2\{\hat{\alpha}_2\} + 4s\{\hat{\alpha}_1, \hat{\alpha}_2\} = 4(0.1462) + 0.1534 + 4(-0.0734) = 0.4448$$

$$s^2\{\hat{D}_3\} = s^2\{\hat{\alpha}_1\} + 4s^2\{\hat{\alpha}_2\} + 4s\{\hat{\alpha}_1, \hat{\alpha}_2\} = 0.1462 + 4(0.1534) + 4(-0.0734) = 0.4461$$

$$\text{Bonferroni: } B = t_{(1-\alpha/2g; n_T-ab)} = t_{(0.9833; 28)} = 2.2383$$

$$\text{Scheffé: } S = \sqrt{(a-1)F_{(1-\alpha; a-1, n_T-ab)}} = \sqrt{2F_{(0.9; 2, 28)}} = \sqrt{2(2.5028)} = 2.2373$$

$$\text{Tukey } T = \frac{1}{\sqrt{2}}q_{(1-\alpha; a, n_T-ab)} = \frac{1}{\sqrt{2}}q_{(0.9; 3, 28)} = \frac{1}{\sqrt{2}}(3.0257) = 2.1395$$

90% family confidence intervals:

$$\hat{D}_1 \pm Ts\{\hat{D}_1\} \implies -6.2334 \pm 2.1395\sqrt{0.4463} \implies -7.6628 \leq D_1 \leq -4.8040$$

$$\hat{D}_2 \pm Ts\{\hat{D}_2\} \implies 0.0333 \pm 2.1395\sqrt{0.4448} \implies -1.3936 \leq D_2 \leq 1.4602$$

$$\hat{D}_3 \pm Ts\{\hat{D}_3\} \implies 6.2667 \pm 2.1395\sqrt{0.4461} \implies 4.8060 \leq D_3 \leq 7.7274$$

h)

$$\hat{L} = 0.3\bar{Y}_{12.} + 0.6\bar{Y}_{22.} + 0.1\bar{Y}_{32.} = 0.3(21.3333) + 0.6(27.6667) + 0.1(20.6) = 25.06$$

$$s^2\{\hat{L}\} = MSE \sum_{i=1}^3 \frac{c_i^2}{n_{i2}} = \frac{71.333}{28} \left(\frac{0.3^2}{6} + \frac{0.6^2}{6} + \frac{0.1^2}{5} \right) = 0.1962; \quad s\{\hat{L}\} = 0.4429$$

$$t_{(0.975, 28)} = 2.0484$$

95% confidence interval for mean cash offer for this population is

$$\hat{L} \pm t_{(0.975, 28)}s\{\hat{L}\} \implies 25.06 \pm 2.0484(0.4429) \implies 24.1528 \leq L \leq 25.9673$$