STAT 500 Final Exam

Vu Thi-Hong-Ha; NetID: 851924086

November 24, 2020

1 Question 1

(a) $X = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$

(b) Yes.

(c)

$$\text{Var}\begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \dots \\ \epsilon_{32} \end{pmatrix} = \begin{pmatrix} Var(\epsilon_{11}) & Cov(\epsilon_{11}, \epsilon_{12}) & \dots & Cov(\epsilon_{11}, \epsilon_{32}) \\ Cov(\epsilon_{12}, \epsilon_{11}) & Var(\epsilon_{12}) & \dots & Cov(\epsilon_{12}, \epsilon_{32}) \\ \dots \\ Cov(\epsilon_{32}, \epsilon_{11}) & \dots & \dots & Var(\epsilon_{32}) \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots \\ 0 & \dots & \dots & \sigma^2 \end{pmatrix} = \sigma^2\mathbf{I} \ (\epsilon_{ij} \text{ are i.i.d})$$

Var($\mathbf{Y}$) = Var($\mathbf{X}\beta + \epsilon$) = Var($\epsilon$) because $\mathbf{X}$ is a matrix of constants and β is a vector of unknown parameters.

Hence, the model is a Gauss-Markov model.

(d) $\mathbf{Y} \sim$ N($\mathbf{X}\beta$, $\sigma^2\mathbf{I}$).

(e) rank($\mathbf{X}$) = 3 because there are three linearly independent columns in $\mathbf{X}$.

(f) Let A = $\begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \end{pmatrix}$. Then AX = C.

(i) C = $\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$. Then we have the system of equations:

$a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + a_7 + a_8 = 1$ (1)

$a_1 + a_2 = 1$

$\dots$

$a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + a_7 + a_8 = 0$ (2)

(1) and (2) are conflicting, so the linear function is not estimable.

(ii) $X_7 = X_1 + X_2 + X_3 + X_4 - X_8$. X_7 is a linear combination of the other columns, so β_1 is not estimable.

(iii) $X_8 = X_1 + X_2 + X_3 + X_4 - X_7$. X_8 is a linear combination of the other columns, so β_2 is not estimable.

(iv) C = $\begin{pmatrix} 0 & -1 & 0 & 1 & 0 & 5 \end{pmatrix}$. Then we have the system of equations:

$a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + a_7 + a_8 = 0$ (1)

$a_1 + a_2 = -1$

$a_3 + a_4 + a_5 + a_6 = 0$

$a_7 + a_8 = 1$ $a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + a_7 + a_8 = 5$ (2)

(1) and (2) are conflicting, so the linear function is not estimable.

2 Question 2

(a) For Water Level 3, $X_{2k} = 1$ and $X_{1k} = 0$.

$\mathbf{E}(Y_i) = \mathbf{E}(\beta_0 + \beta_1 * 0 + \beta_2 * 1 + \beta_3 * X_{3k} + \epsilon_k) = \beta_0 + \beta_2 + \beta_3 * X_{3k}$

(b) The estimated value for β_3 is the changes in the conditional mean of $\mathbf{Y}$ for one unit increase in the fertilizer amount, holding all other explanatory variables constant.

(c) By the SAS output, $b_3 = 1.08$ and $S_{b_3} = 0.12396$.

$H_0 : \beta_3 = 0$ against $H_a : \beta_3 \neq 0$

Number of observations $n = 12$, number of explanatory variables $k_{(1)} = 3, \alpha = 0.05$, so $t_{n-(k_{(1)}-1),1-\alpha/2} = t_{8,0.975} =$ 2.306.

Test statistic $t = \dfrac{b_3 - 0}{S_{b_3}} = \dfrac{1.08}{0.12396} = 8.7125 > t_{8,0.975}$

p-value = $P(T > |t|) < 0.001$.

At the significance level of 0.05, we reject the null that $\beta_3 = 0$. We conclude that the amount of fertilizer has a significant impact on the plant heights.

(d)

$\mathbf{x} = \begin{pmatrix} 1 & 1 & 0 & 10 \end{pmatrix}^T$, $\hat{\beta} = \begin{pmatrix} 119.75 & -0.25 & 4 & 1.08 \end{pmatrix}^T$

$\hat{\mu}_{Y|x} = \mathbf{x}^T \hat{\beta} = 130.3$

Based on the given 95% confidence, we have $\hat{\mu}_{Y|x} + t_{8,0.975} S_{\hat{\mu}_{Y|x}} = 130.3 + 2.306 S_{\hat{\mu}_{Y|x}} = 133.16$. Hence, $S_{\hat{\mu}_{Y|x}} = 1.24$.

By the SAS output, $MS_{error} = 5.7625$.

Hence, $S_{\hat{Y}} = \sqrt{MS_{error} + S^2_{\hat{\mu}_{Y|x}}} = \sqrt{5.7625 + 1.24^2} = 2.702$

A 95% prediction interval is $\hat{Y}_i \pm t_{8,0.975} * S_{\hat{Y}} = 130.3 \pm 2.306 * 2.702 = (124.0692; 136.531)$

(e) For Water Level 3, $X_{2k} = 1$ and $X_{1k} = 0$.

$\mathbf{E}(Y_i) = \mathbf{E}(\beta_0 + \beta_1 * 0 + \beta_2 * 1 + \beta_3 * X_{3k} + \beta_4 * 0 * X_{3k} + \beta_5 * 1 * X_{3k} + \epsilon_k) = \beta_0 + \beta_2 + (\beta_3 + \beta_5) * X_{3k}$

(f)

For Water Level 1, the model is $Y_k = \beta_0 + \beta_3 X_{3k} + \epsilon_k$

For Water Level 2, the model is $Y_k = \beta_0 + \beta_1(\beta_3 + \beta_4)X_{3k} + \epsilon_k$

Testing for the significance of the difference between the two models' slopes becomes testing for the significance of β_4.

$H_0 : \beta_4 = 0$ against $H_a : \beta_4 \neq 0$.

Test statistic: $t = \dfrac{-0.66}{0.07303} = -9.04$.

p-value < -0.001.

At the significance level of 0.05, we reject the null that $\beta_4 = 0$. We conclude that there is a significant different in the slope of regression lines between WaterLevel L1 and WaterLevel L2.

(g)

For Water Level 1, the model is $Y_k = \beta_0 + \beta_3 X_{3k} + \epsilon_k$

For Water Level 2, the model is $Y_k = \beta_0 + \beta_1 + (\beta_3 + \beta_4)X_{3k} + \epsilon_k$

For Water Level 2, the model is $Y_k = \beta_0 + \beta_1 + (\beta_3 + \beta_5)X_{3k} + \epsilon_k$

Testing for the significance of the difference between the slopes becomes testing whether $\beta_4 = \beta_5 = 0$.

$H_0 : \beta_4 = \beta_5 = 0$ against H_a : at least β_4 or β_5 is different from 0.

The reduced (null) model is $Y_k = \beta_0 + \beta_1 X_{1k} + \beta_2 X_{2k} + \beta_3 X_{3k} + \epsilon_k$ (number of explanatory variables $k_{(1)}$: 3).

The full model is $Y_k = \beta_0 + \beta_1 X_{1k} + \beta_2 X_{2k} + \beta_3 X_{3k} + \beta_4 X_{1k} X_{3k} + \beta_5 X_{2k} X_{3k} + \epsilon_k$ (number of explanatory variables $k_{(2)}$: 5).

From the SAS output of model (1): $SSE_{r.model} = 5.7625 * (n - k_{(1)} - 1) = 5.7625 * 8 = 46.1$

From the SAS output of model (2): $SSE_{f.model} = 0.33333 * (n - k_{(2)} - 1) = 0.33333 * 6 = 1.99998$

Test statistic $F = \dfrac{(SSE_{r.model} - SSE_{f.model})/2}{MSE_{f.model}} = \dfrac{(46.1 - 1.99998)/2}{0.33333} = 66.15069$.

$F_{2,8,0.95} = 4.45897$ The numerator degree of freedom is 2, and the denominator degree of freedom is 8.

p-value < 0.0001.

At the significance level of 0.05, we reject the null that $\beta_4 = \beta_5 = 0$. We conclude that the difference in the regression slopes of three water levels is significant.

(h)

The assumptions we have to check include:

- Observations Y_k are independent.
- Values of $\mathbf{x}$ are fixed.
- $\mu_{Y|x}$ is a linear function of $\mathbf{x}$.
- Homogeneous error variance: $Var(\epsilon_k) = \sigma^2$.
- Normally distributed errors: ϵ_k i.i.d $N(0, \sigma^2)$.

Through the diagnosis plots, we can check:

- Homogeneous variance. As we can see from the Residual vs Predicted Value, the points seem to be spreading out when the predicted values increase, creating a megaphone shape. This indicates that the variances may not be homogeneous. However, we only have 12 observations, so it is hard to conclude.
- We can check the normality of the errors by looking at the q-q plot. By this q-q plot, the data seems to be light tailed, and does not follow normal distribution. However, we only have 12 observations, so it is hard to conclude.
- From the RStudent vs Predict Value and RStudent vs Leverage plots, there does not seem to have any outlier as all the data points have residuals and leverage within thresholds.

(i)

The diagnosis plots cannot help us check the following assumptions:
- Observations Y_k are independent.
- Values of $\mathbf{x}$ are fixed.
- $\mu_{Y|x}$ is a linear function of $\mathbf{x}$.

Discussion:
- It is safe to assume that the observations Y_k are independent because in the experiment designs, there is no indication of the samples being related (in terms of time, space, cluster) with each other. The treatments are also assigned randomly.
- By the definition of the variables and how the data is collected, we can assume the variables are fixed.
- In order to check if there is a linear relationship, we can use scatter plot or residual plot of $\mathbf{Y}$ against each variable. If the linear relationship assumption is violated, we can transform the values of $\mathbf{Y}$ or the variables to have linear relationship.

(j)

Model 1:

$adjR^2 = 1 - \frac{MS_{error}}{SS_{total}/(n-1)} = 1 - \frac{5.7625}{529/(12-1)} = 0.8802$

$C_p = \frac{SS_{error}}{\hat{sigma}^2} - (n - 2(k_{(1)} + 1)) = k_{(1)} + 1$

$AIC = 12log(\frac{5.7625 * 8}{12}) + 8 = 15.014$

Model 2:

$adjR^2 = 1 - \frac{MS_{error}}{SS_{total}/(n-1)} = 1 - \frac{0.33333}{529/(12-1)} = 0.9931$

$C_p = \frac{SS_{error}}{\hat{sigma}^2} - (n - 2(k_{(1)} + 1)) = k_{(2)} + 1$

$AIC = 12log(\frac{0.33333 * 6}{12}) + 12 = 2.6621$

Comparing Model 1 and Model 2, we can see Model 2 have better $adjR^2$ and AIC while C_p of both models are good. Therefore, I would choose model 2.

(k)

Number of observations $n = 24$.

Number of distinct variables $r = 3 * 4 = 12$ (3 levels of water, 4 levels of fertilizer).

The numerator df is $r - 2 = 10$, and the denominator df is $n - r = 24 - 12 = 12$.

3 Question 3

(a)

- Treatment: fed with granules or not.

- Factors: colors of granules (level = 3), sizes of granules (level = 3), replications (level = 8).

- Experimental units: birds.

- Observation units: unit concentration of blood enzyme.

- Response variable: birds' amount of increase in concentration of blood enzyme.

- Replication: yes.

- Randomization: yes: Birds were chosen randomly. Birds can move freely in the pen. Food and granules were scatter randomly. Treatments where assigned randomly to pens.

- Blocking: no.

- This is a complete factorial design.

(b)

To answer the researchers' questions, consider the model:

$Y_{ijkl} = \mu + \alpha_i + \tau_j + \beta_k + (\alpha\tau)_{ij} + (\alpha\beta)_{ik} + (\tau\beta)_{jk} + (\alpha\tau\beta)_{ijk} + \epsilon_{ijkl}$

To use this model, we assume: Independence, Homogeneous Variance and Normality of the residuals. We will have to check these assumptions.

The questions will correspond to checking the following:

(i) the main effects of color and size on increase in the measured concentration of the blood enzyme:

+ main effect of colors: $\bar{\mu_{1..}} = \bar{\mu_{2..}} = \bar{\mu_{3..}}$

+ main effect of sizes: $\bar{\mu_{.1.}} = \bar{\mu_{.2.}} = \bar{\mu_{.3.}}$

(ii) whether the effects of color depend on size or not:

$(\bar{\mu_{11.}} - \bar{\mu_{21.}} - \bar{\mu_{31.}}) - (\bar{\mu_{12.}} - \bar{\mu_{22.}} - \bar{\mu_{32.}}) - (\bar{\mu_{13.}} - \bar{\mu_{23.}} - \bar{\mu_{33.}})$

(iii) whether the effects of blue and natural depend on size or not:

$(\bar{\mu_{11.}} - \bar{\mu_{21.}} - \bar{\mu_{31.}}) - (\bar{\mu_{13.}} - \bar{\mu_{23.}} - \bar{\mu_{33.}})$

(iv) which treatment group(s) results in the smallest increase in the measured concentration of the blood enzyme

$\bar{\mu_{11.}} = \bar{\mu_{21.}} = \bar{\mu_{31.}} = \bar{\mu_{12.}} = \bar{\mu_{22.}} = \bar{\mu_{32.}} = \bar{\mu_{13.}} = \bar{\mu_{23.}} = \bar{\mu_{33.}}$

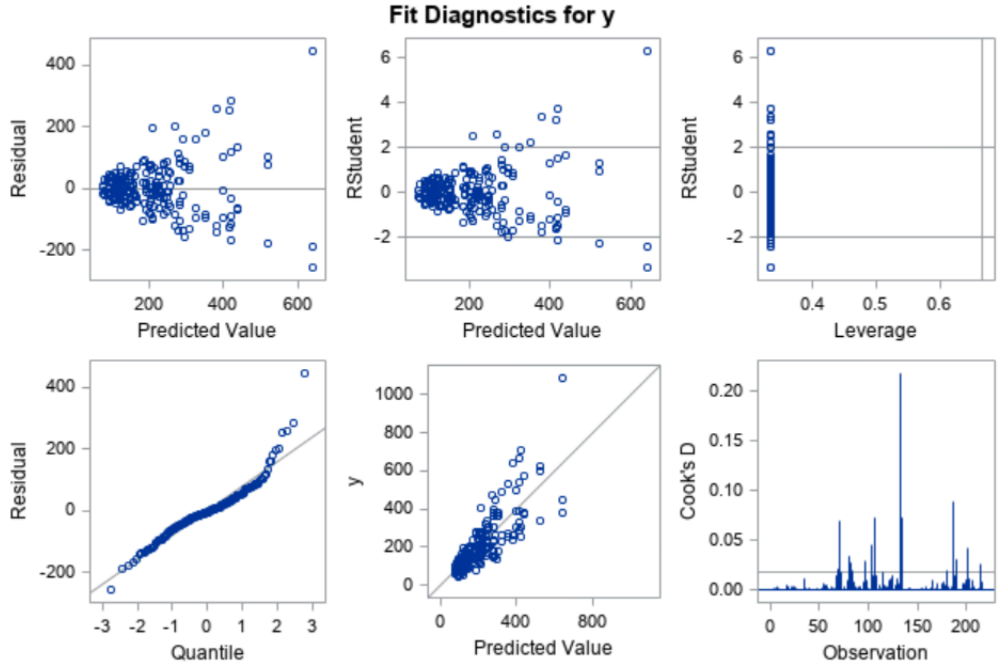
(v) whether the replication effect should be considered when designing future experiments

$(\alpha\beta)_{ik} = (\tau\beta)_{jk} = (\alpha\tau\beta)_{ijk} = 0$

If none of the interactions are significant, test $\beta_k = 0$

(c) **Check model assumptions:** - From the experimental designs, we can assume the residuals are independent.

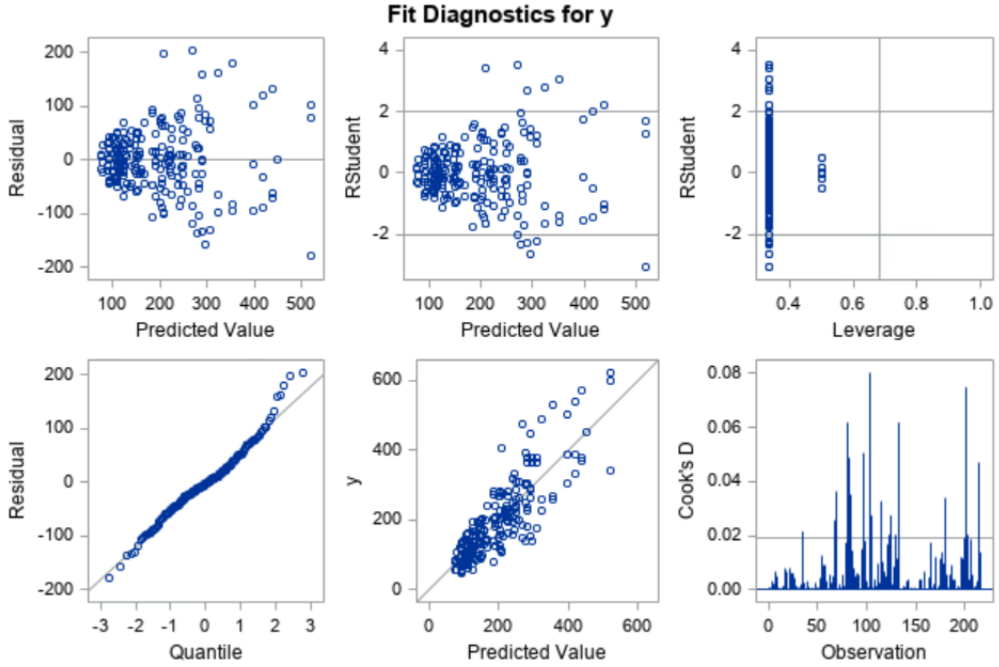
- To check homogeneous variances and normality assumption, we look at the diagnosis plots:



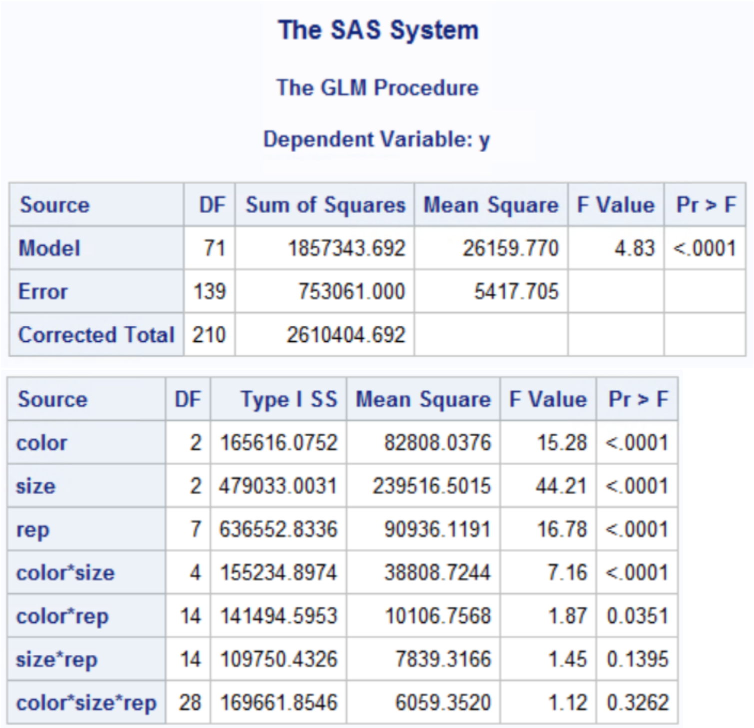
From the Residual vs Predicted value, we can see there are potential outliers that make the plot megaphone shape. From the RStudent vs Predicted value, the outliers look more obvious. Therefore, for downstream analysis, I will exclude the data points that have absolute RStudent value greater than 3.

From the q-q plot, the data looks quite good except for some outliers.

I will exclude the 5 outliers whose $|RStudent| > 3$ and check the assumptions again:



The diagnosis plots look much better, so we can do further testing.



The SAS System

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	71	1857343.692	26159.770	4.83	<.0001
Error	139	753061.000	5417.705		
Corrected Total	210	2610404.692			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
color	2	165616.0752	82808.0376	15.28	<.0001
size	2	479033.0031	239516.5015	44.21	<.0001
rep	7	636552.8336	90936.1191	16.78	<.0001
color*size	4	155234.8974	38808.7244	7.16	<.0001
color*rep	14	141494.5953	10106.7568	1.87	0.0351
size*rep	14	109750.4326	7839.3166	1.45	0.1395
color*size*rep	28	169661.8546	6059.3520	1.12	0.3262

(i) the main effects of color and size on increase in the measured concentration of the blood enzyme:

+ main effect of colors:

$H_0 : \bar{\mu_{1..}} = \bar{\mu_{2..}} = \bar{\mu_{3..}}$

H_a : at least one mean is different from zero.

Test statistic: $F = MS_1/MS_{error} = 82808.0376/5417.705 = 15.28471$

$F_{a-1;ab(n-1);1-\alpha} = F_{2,139,0.95} = 3.06123$

At the significane level of 0.05, we reject the null.

+ main effect of sizes:

$H_0 : \bar{\mu_{.1.}} = \bar{\mu_{.2.}} = \bar{\mu_{.3.}}$

H_a : at least one mean is different from zero.

Test statistic: $F = MS_2/MS_{error} = 239516.5015/5417.705 = 44.20996$

$F_{a-1;ab(n-1);1-\alpha} = F_{2,139,0.95} = 3.06123$

At the significane level of 0.05, we reject the null.

(ii) whether the effects of color depend on size or not:

$(\bar{\mu_{11.}} - \bar{\mu_{21.}} - \bar{\mu_{31.}}) - (\bar{\mu_{12.}} - \bar{\mu_{22.}} - \bar{\mu_{32.}}) - (\bar{\mu_{13.}} - \bar{\mu_{23.}} - \bar{\mu_{33.}})$

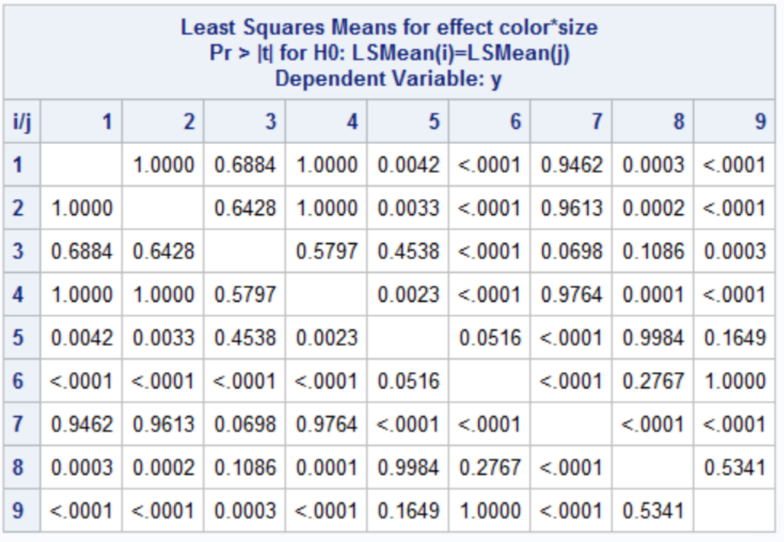
F value = 7.16, pvalue < 0.001. We reject the null at the significance level of 0.05.

(iii) whether the effects of blue and natural depend on size or not:

$H_0 : (\bar{\mu_{11.}} - \bar{\mu_{21.}} - \bar{\mu_{31.}}) - (\bar{\mu_{13.}} - \bar{\mu_{23.}} - \bar{\mu_{33.}}) = 0$

F value = 22.84, pvalue < 0.001. We reject the null at the significance level of 0.05.

(iv) $\bar{\mu_{11.}} = \bar{\mu_{21.}} = \bar{\mu_{31.}} = \bar{\mu_{12.}} = \bar{\mu_{22.}} = \bar{\mu_{32.}} = \bar{\mu_{13.}} = \bar{\mu_{23.}} = \bar{\mu_{33.}}$



Least Squares Means for effect color*size
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: y

i/j	1	2	3	4	5	6	7	8	9
1		1.0000	0.6884	1.0000	0.0042	<.0001	0.9462	0.0003	<.0001
2	1.0000		0.6428	1.0000	0.0033	<.0001	0.9613	0.0002	<.0001
3	0.6884	0.6428		0.5797	0.4538	<.0001	0.0698	0.1086	0.0003
4	1.0000	1.0000	0.5797		0.0023	<.0001	0.9764	0.0001	<.0001
5	0.0042	0.0033	0.4538	0.0023		0.0516	<.0001	0.9984	0.1649
6	<.0001	<.0001	<.0001	<.0001	0.0516		<.0001	0.2767	1.0000
7	0.9462	0.9613	0.0698	0.9764	<.0001	<.0001		<.0001	<.0001
8	0.0003	0.0002	0.1086	0.0001	0.9984	0.2767	<.0001		0.5341
9	<.0001	<.0001	0.0003	<.0001	0.1649	1.0000	<.0001	0.5341	

(I did not have enough time for this question, but I wanted to test if there is a significant difference between each treatment, and then carry out one tail test for the difference between the pairs of treatments, and get the smallest one.)

(v) From the ANOVA table, we can see the interaction between color and replication has a F value of 1.87 and pvalue of 0.0351. At the significant level of 0.05, we reject the null that the interaction effect is non-zero. Then, the replication factor has some effects on the response. (d) Conclusion:

(i) Color has a significant effect on the increase of blood enzyme.

Size has a significant effect on the increase of blood enzyme.

(ii) Effects of color depend of sizes.

(iii) The difference in effects between blue and natural color depends on size.

(iv)

(v) We should include replication factor in the future.