STAT 516 sp 2024 exam 02

75 minutes, no calculators or notes allowed

1. Multiple linear regression

Consider fitting on a data set the multiple linear regression model $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$, for $i = 1, \dots, n$, where the ε_i are independent Normal$(0, \sigma^2)$ error terms and the x_{ij} are predictor values.

Suppose the data set has $p = 15$ predictors, but you do not believe all of them are important, so you decide to search for a good model which does not use all 15 predictors.

(a) Suppose you wish to compare all possible models that one can build from the 15 predictors. How many models will you need to fit?

The total number of models is 2^{15}, which is an enormous number! Each predictor is "in" or "out", and there are $\underbrace{2 \times 2 \times 2 \cdots \times 2}_{15}$ sequences of "in" and "out".

(b) Instead of considering all possible models, you decide to start with the model which uses all the predictors and then to remove one predictor at a time according to some criterion. What is the name for such an approach to model selection?

This is called backwards stepwise selection.

(c) Give the name of a criterion for comparing models and explain how to use it.

Akaike's Information Criterion (AIC) is a criterion for comparing models. If you compute it on two models, the model with a smaller AIC is "better".

(d) Explain *why* one would wish to discard some of the 15 predictors. Why not just leave all of 15 of them in the model?

The more predictors in the model, the lower the statistical power to reject $H_0: \beta_j = 0$ for each $j = 1, \dots, p$.

The reduction in power is greater if the predictors are highly correlated with each other.

Therefore it is better not to include "extra" predictors in a model.

Models with fewer predictors are also easier to interpret.

2. One-way ANOVA

A study recorded the tensile strengths of sheet metal specimens sampled from four suppliers. A manufacturer wishes to know whether the mean tensile strength differs across these suppliers.

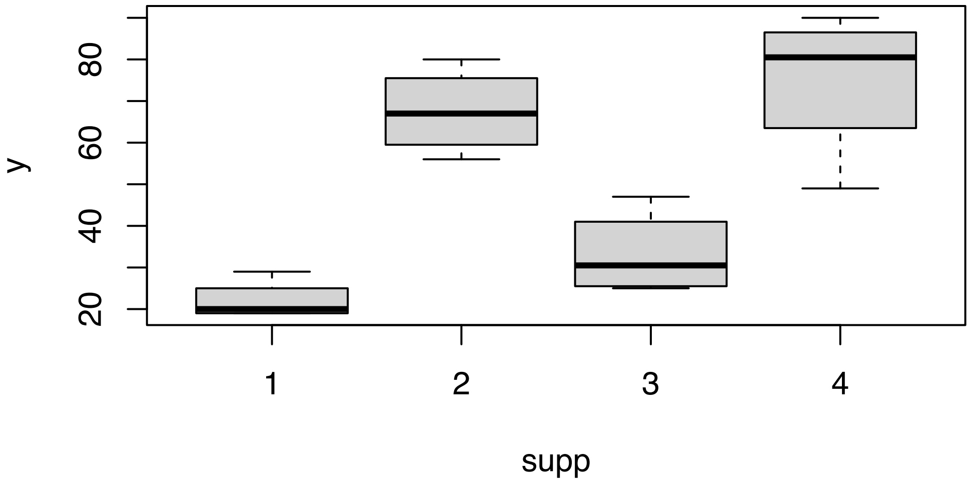
To answer the manufacturer's question, you fit the model

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad i = 1, \dots, a, \quad j = 1, \dots, n,$$

where the ε_{ij} are independent Normal$(0, \sigma^2)$ random variables.

Here is some R output:

```
tensile <- data.frame( y = c(19,80,47,90,21,71,26,49,19,63,25,83,29,56,35,78),
                       supp = as.factor(c(1,2,3,4,1,2,3,4,1,2,3,4,1,2,3,4)))
boxplot(y ~ supp, data = tensile)
```

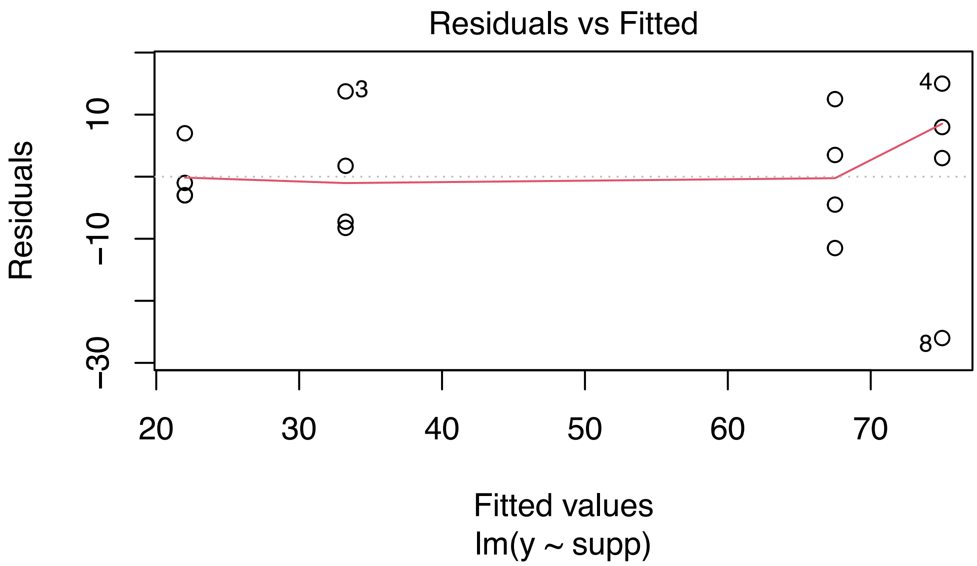


```
lm_out <- lm(y ~ supp, data = tensile)
lm_out
```

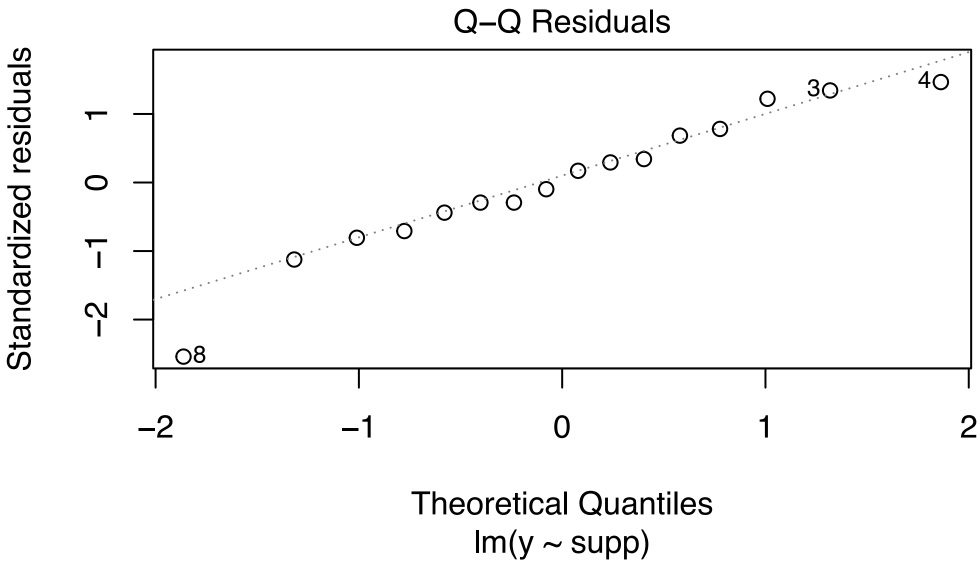
```
Call:
lm(formula = y ~ supp, data = tensile)

Coefficients:
(Intercept)        supp2        supp3        supp4
      22.00        45.50        11.25        53.00
```

```
# summary(lm_out)
# anova(lm_out)
plot(lm_out, which = 1)
```



```
plot(lm_out, which = 2)
```



(a) Give a and n for these data.

a = 4

n = 4

(b) Give each of the treatment group means $\bar{Y}_{i.}$ for $i = 1, 2, 3, 4$ using the estimated model coefficients (you do not need a calculator to do this).

$\bar{Y}_{1.} = 22.00$ $\qquad$ $\bar{Y}_{3.} = 22.00 + 11.25 = 33.25$

$\bar{Y}_{2.} = 22.00 + 45.50 = 67.50$ $\qquad$ $\bar{Y}_{4.} = 22.00 + 53.00 = 75.00$

(c) Each value listed below appears in the ANOVA table for these data.

15 7978.2 19.044 139.65 12 1675.7 7.401×10^{-5} 2659.40 3 9653.9 Largest, so SS_{Tot}

Put each value in the right place (you do not need a calculator to do this):

Source	Df	SS	MS	F value	p-value
Supplier	3	7978.2	2659.4	19.044	7.401×10^{-5}
Error	12	1675.7	139.65		
Total	15	9653.9			

(d) State whether you think the model assumptions are satisfied by these data. Write a couple of sentences. If you do not think the assumptions are satisfied, give some advice about what to do.

There is some indication in the residuals vs fitted values plot that the variance is not constant across the treatment groups.

One could try log-transforming the response values and fitting the model again.

(e) Write down the null hypotheses for which the F value in the ANOVA table serves as a test statistic. ALSO state whether you reject the null hypothesis with these data.

This is the null hypothesis

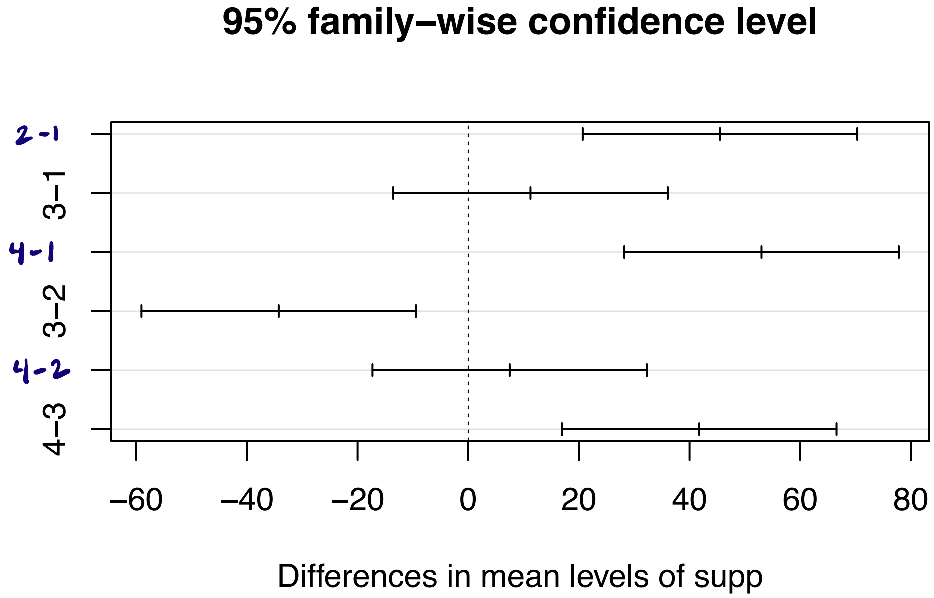
$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4,$$

where $\mu_1, \mu_2, \mu_3,$ and μ_4 are the treatment group means.

Since the p-value is very small (7.401×10^{-5}), we would reject H_0 and conclude that not all the treatment means are the same.

(f) Assuming the assumptions are satisfied, write (three or four sentences) an interpretation of the output of the plot below. What can you tell the manufacturer about the differences in mean tensile strength between the four manufacturers? Does a ranking of the suppliers emerge? Can you relate this picture to the boxplots shown earlier in this question? Address such questions in your answer.

```r
Tukey_out <- TukeyHSD(aov(y~supp,data=tensile))
plot(Tukey_out)
```



From Tukey's pairwise comparisons of the supplier means, we see that suppliers 4 and 2 each have means greater than suppliers 1 and 3, but we do not have evidence to say that 4 and 2 differ or that 1 and 3 differ.

This confirms what we might expect from the side-by-side boxplots, which seem to separate the suppliers into these two groups.

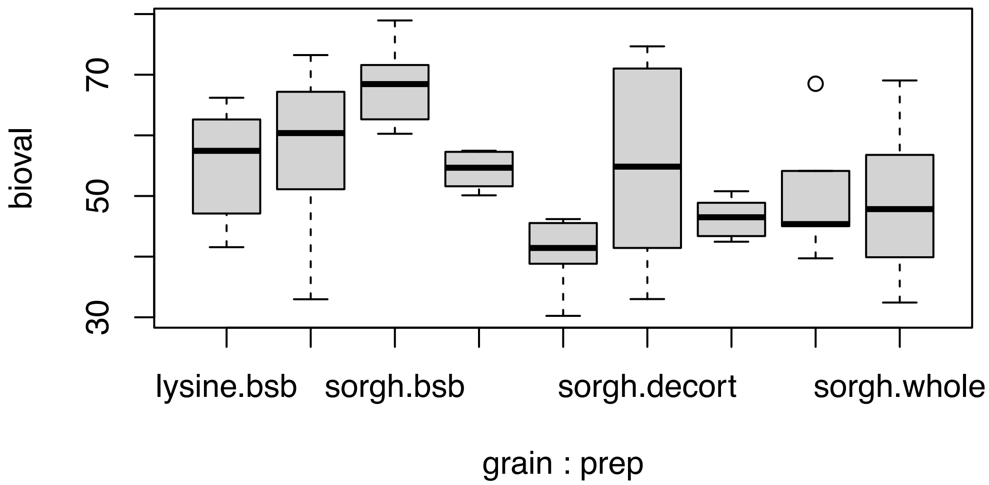
3. Two-way factorial design

Fifty-four rats were randomly assigned to receive one of nine diets such that six rats were assigned to each diet. All combinations of three grain types (sorghum, high-lysine sorghum, millet) and three preparations (whole; decorticated; decorticated, boiled, and soaked) comprised the nine diets. The response for each rat is a biological measurement taken after the rat was fed the diet for some amount of time.

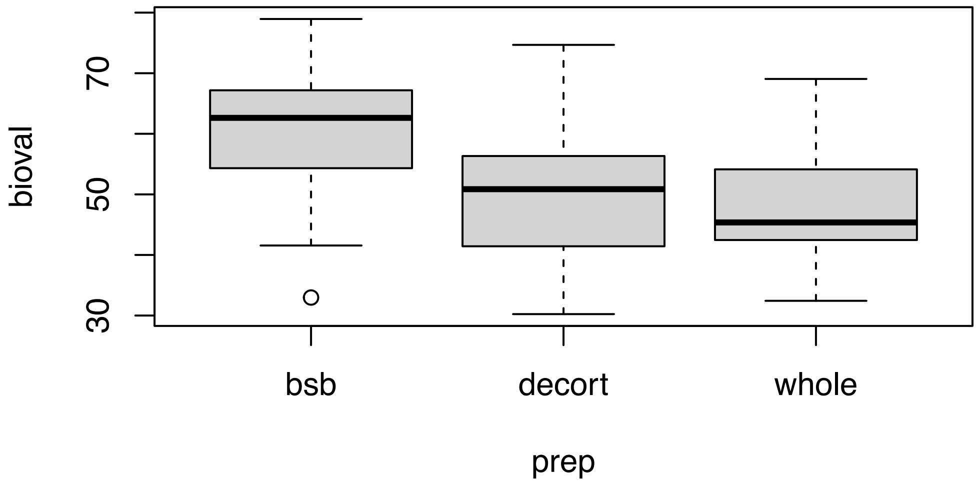
```
head(diet,n=12)
```

```
   grain   prep bioval
1  sorgh  whole  40.61
2  sorgh  whole  56.78
3  sorgh  whole  69.05
4  sorgh  whole  39.90
5  sorgh  whole  55.06
6  sorgh  whole  32.43
7  sorgh decort  74.68
8  sorgh decort  56.33
9  sorgh decort  71.02
10 sorgh decort  53.35
11 sorgh decort  41.43
12 sorgh decort  33.00
```

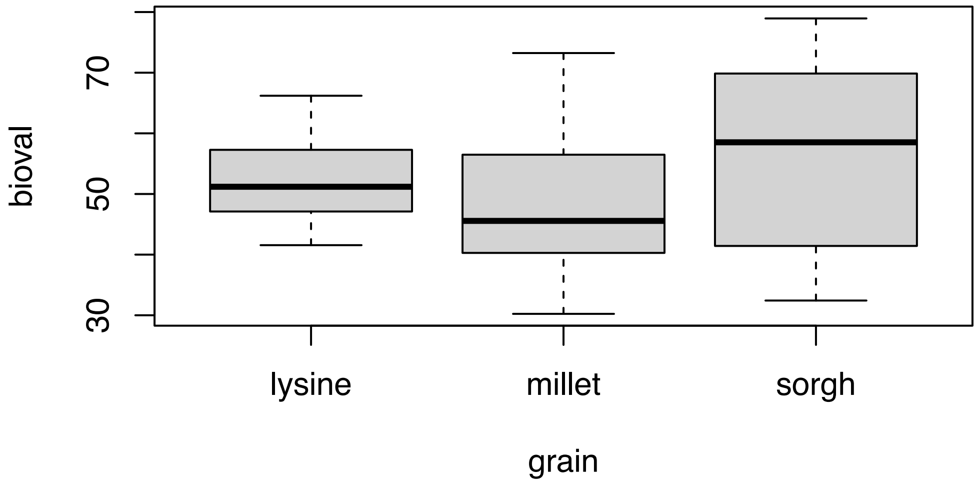
```
boxplot(bioval ~ grain + prep, data = diet)
```



```r
boxplot(bioval ~ prep, data = diet)
```



```r
boxplot(bioval ~ grain, data = diet)
```



Consider modeling the data with the two-way treatment effects model

$$Y_{ijk} + \mu + \tau_i + \gamma_j + (\tau\gamma)_{ij} + \varepsilon_{ijk}, \quad i = 1, \dots, a, \quad j = 1, \dots, b, \quad k = 1, \dots, n_{ij},$$

where the ε_{ijk} are independent Normal$(0, \sigma^2)$ random variables.

(a) Let the grain type be factor A and the preparation type be factor B. Give a and b as well as n_{ij} for all i, j.

$a = 3$

$b = 3$

$n_{ij} = n = 6$ for all i, j.

(b) Use the estimated coefficients (printed below) from the two-way treatment effects model to ~~compute~~ give an expression for the mean of the responses in the group of rats fed the diet at the factor level combination sorghum × whole.

$$\bar{Y}_{\text{sorghum} \times \text{whole}} = 55.397 + 12.997 - 8.982 - 10.440$$

```
lm_out <- lm(bioval ~ grain + prep + grain:prep, data = diet)
lm_out
```

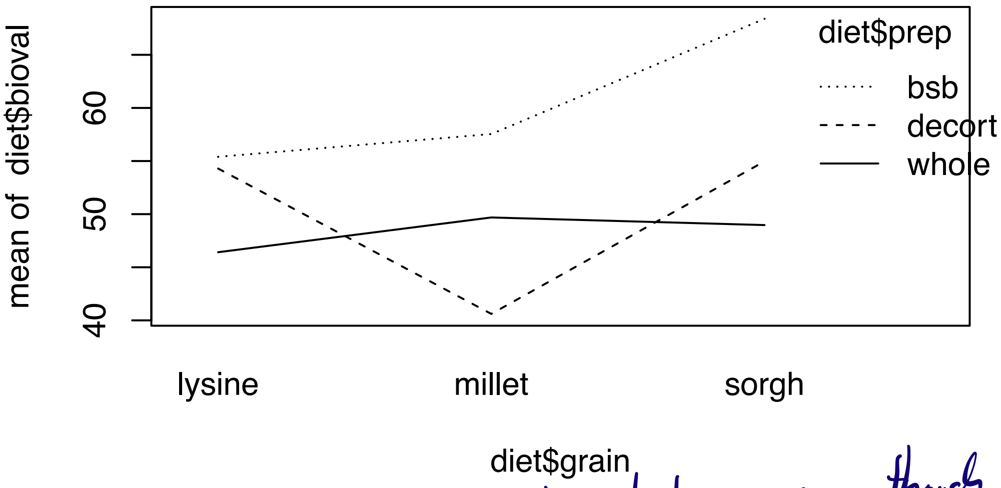
```
Call:
lm(formula = bioval ~ grain + prep + grain:prep, data = diet)

Coefficients:
          (Intercept)             grainmillet              grainsorgh  
               55.397                   2.145                  12.997  
           prepdecort               prepwhole  grainmillet:prepdecort  
               -1.102                  -8.982                 -15.833  
grainsorgh:prepdecort   grainmillet:prepwhole    grainsorgh:prepwhole  
              -12.323                   1.127                 -10.440  
```

(c) Fill in the missing Df values in the ANOVA table below.

	Source	Df	SS	MS	F value	p value
3-1	A	2	SS_A	MS_A	2.9334	0.06346
3-1	B	2	SS_B	MS_B	7.3265	0.00176
(3-1)(3-1)	AB	4	SS_{AB}	MS_{AB}	1.8531	0.13533
3·3(6-1)	Error	45	SS_{Error}	105.89		
3·3·6-1	Total	53	SS_{Tot}			

(d) In light of the results in the ANOVA table give a careful interpretation of the plot below (more than one sentence).



Though the plot makes it look as though there is an interaction between the two factors, the p-value for testing for an interaction effect was quite large — 0.135. Therefore the crossing of the lines in the interaction plot is likely due to random noise in the data than to a true interaction.

(e) The p-values 0.06346 and 0.00176 appear in the ANOVA table above. Carefully write down the null hypotheses to which these two p-values correspond.

$H_0: \bar{\mu}_{1\cdot} = \bar{\mu}_{2\cdot} = \bar{\mu}_{\cdot 3}$, where $\bar{\mu}_{1\cdot}, \bar{\mu}_{2\cdot}$, and $\bar{\mu}_{3\cdot}$ are the marginal means for the grain factor.

and $H_1: \bar{\mu}_{\cdot 1} = \bar{\mu}_{\cdot 2} = \bar{\mu}_{\cdot 3}$, where $\bar{\mu}_{\cdot 1}, \bar{\mu}_{\cdot 2}$, and $\bar{\mu}_{\cdot 3}$ are the marginal means for the preparation factor.

(f) What can we conclude on the basis of the p-value which is equal to 0.00176? What does it mean in the terms of the study?

We can conclude that the way in which the grain is prepared has a significant effect on the response across all types of grain.

(g) Explain in detail what the following code is doing. Give also a careful interpretation of the printed output. Write a few sentences.

```
a <- 3
n <- 6

y.1. <- mean(diet$bioval[diet$prep == "whole"])
y.2. <- mean(diet$bioval[diet$prep == "decort"])
y.3. <- mean(diet$bioval[diet$prep == "bsb"])

me <- 2.29 * sqrt(105.89) * sqrt( 2 / (a*n))
CIs <- rbind(c(y.2. - y.1. - me,y.2. - y.1. + me),
             c(y.3. - y.1. - me,y.3. - y.1. + me))
rownames(CIs) <- c("decort - whole","bsb - whole")
colnames(CIs) <- c("lower","upper")
CIs
```

```
                   lower     upper
decort - whole -6.256030  9.453808
bsb - whole     4.231192 19.941030
```

The marginal means of the preparation factor are being compared to the "whole" level as to a baseline level using Dunnett's method.

The "bsb" level has a significantly greater marginal mean than the "whole" level, while there is no significant difference between the "decort" and "whole" levels.

4. Cell-means model for the two-way factorial design

Let $Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$, $i = 1, 2$, $j = 1, 2, 3$, $k = 1, \dots, n_{ij}$, where the ε_{ijk} are independent Normal$(0, \sigma^2)$ random variables. Let i index the levels of one factor and j index the levels of another factor in a two-way factorial experiment. Moreover, suppose the cell means are

$$\mu_{11} = 18, \quad \mu_{12} = 20, \quad \mu_{13} = 21, \quad \mu_{21} = 14, \quad \mu_{22} = 16, \quad \mu_{23} = 17.$$

(a) Compute the marginal means $\bar{\mu}_{i.}$ for $i = 1, 2$ and $\bar{\mu}_{.j}$ for $j = 1, 2, 3$.

$$\begin{array}{|c|c|c|} \hline \mu_{11} & \mu_{12} & \mu_{13} \\ \hline \mu_{21} & \mu_{22} & \mu_{23} \\ \hline \end{array} = \begin{array}{ccc|c} 18 & 20 & 21 & 59/3 \\ 14 & 16 & 17 & 47/3 \\ \hline 32/2 & 36/2 & 38/2 & \end{array}$$

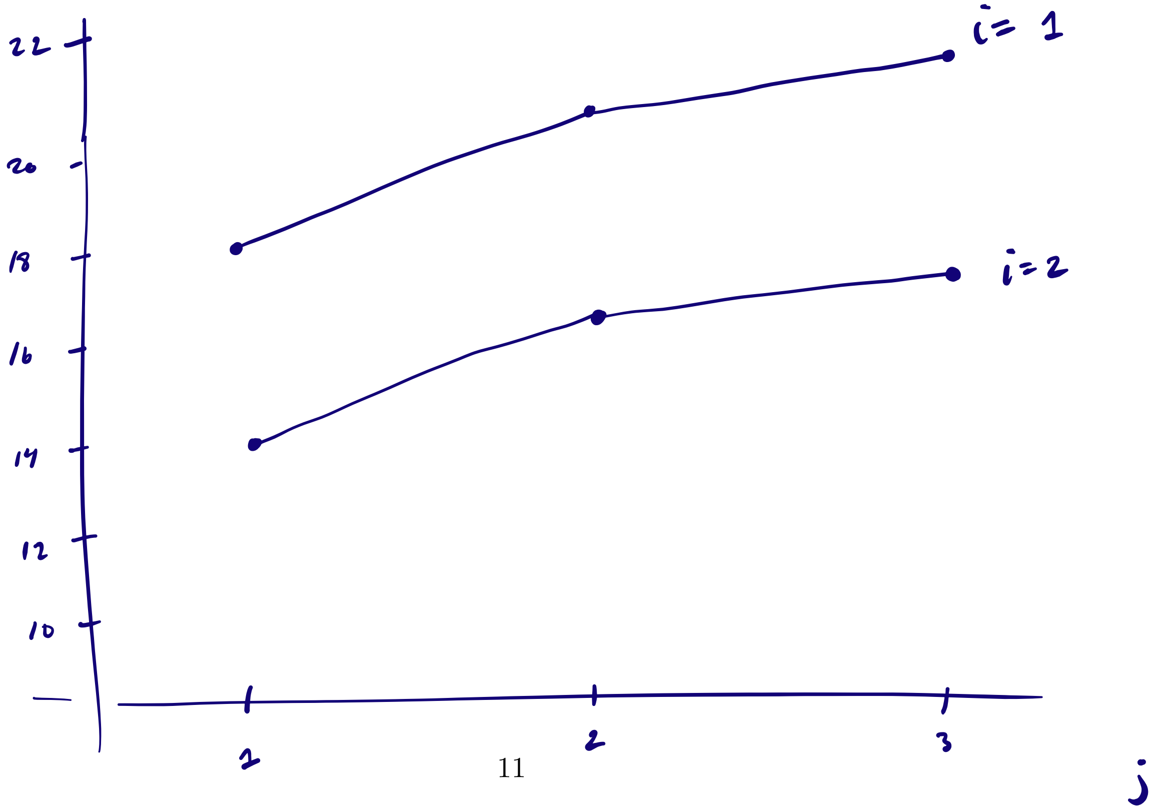
$\bar{\mu}_{1.} = 19\frac{1}{3}$ $\quad \bar{\mu}_{2.} = 15\frac{2}{3}$

$\bar{\mu}_{.1} = 16$, $\bar{\mu}_{.2} = 18$, $\bar{\mu}_{.3} = 19$.

(b) Is there interaction between the two factors? Explain your answer.

No, because $18 - 14 = 20 - 16 = 21 - 17$, so the effect of factor 1 is the same at the three levels of factor 2.

(c) Carefully draw an interaction plot with the level $j = 1, 2, 3$ along the horizontal axis.



We see that there is no interaction.