

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 $Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$, for $i = 1, \dots, n$, where the ε_i are independent $\text{Normal}(0, \sigma^2)$ 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?

- (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?

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

- (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?

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.

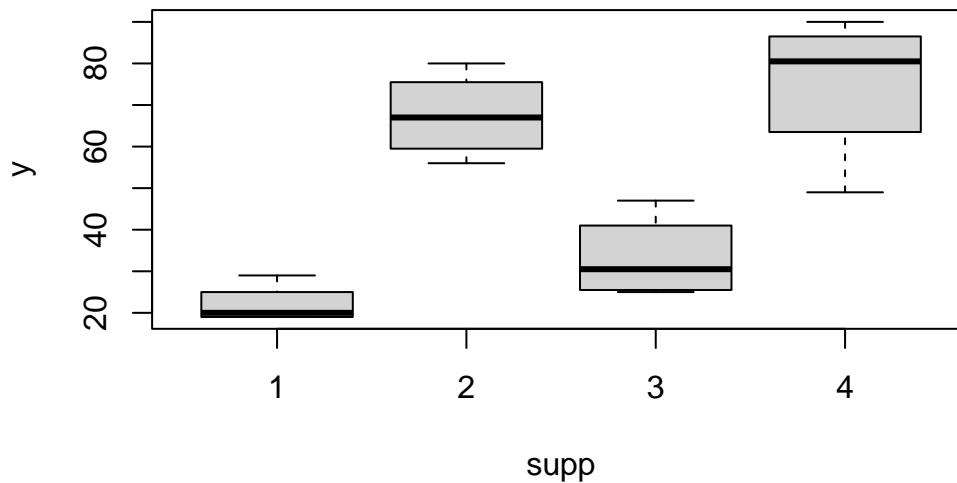
To answer the manufacturer's question, you fit the model

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

where the ε_{ij} are independent $\text{Normal}(0, \sigma^2)$ 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
```

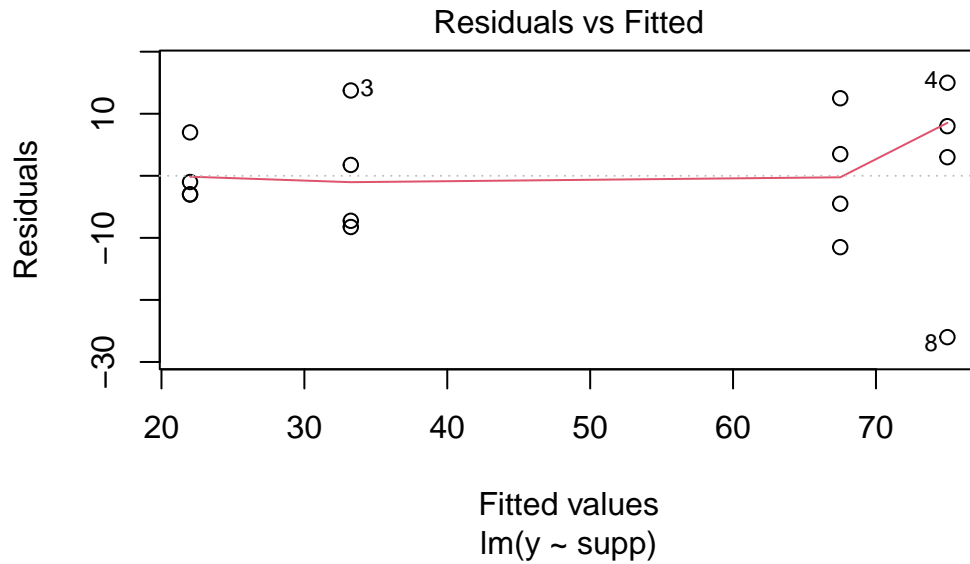
Call:

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

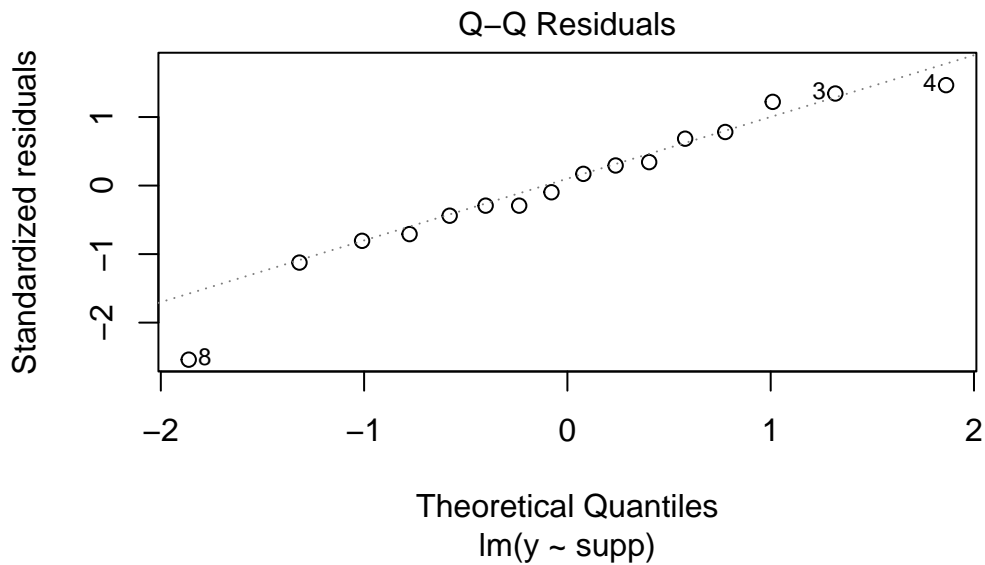
Coefficients:

(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.

(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).

(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

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					
Error					
Total					

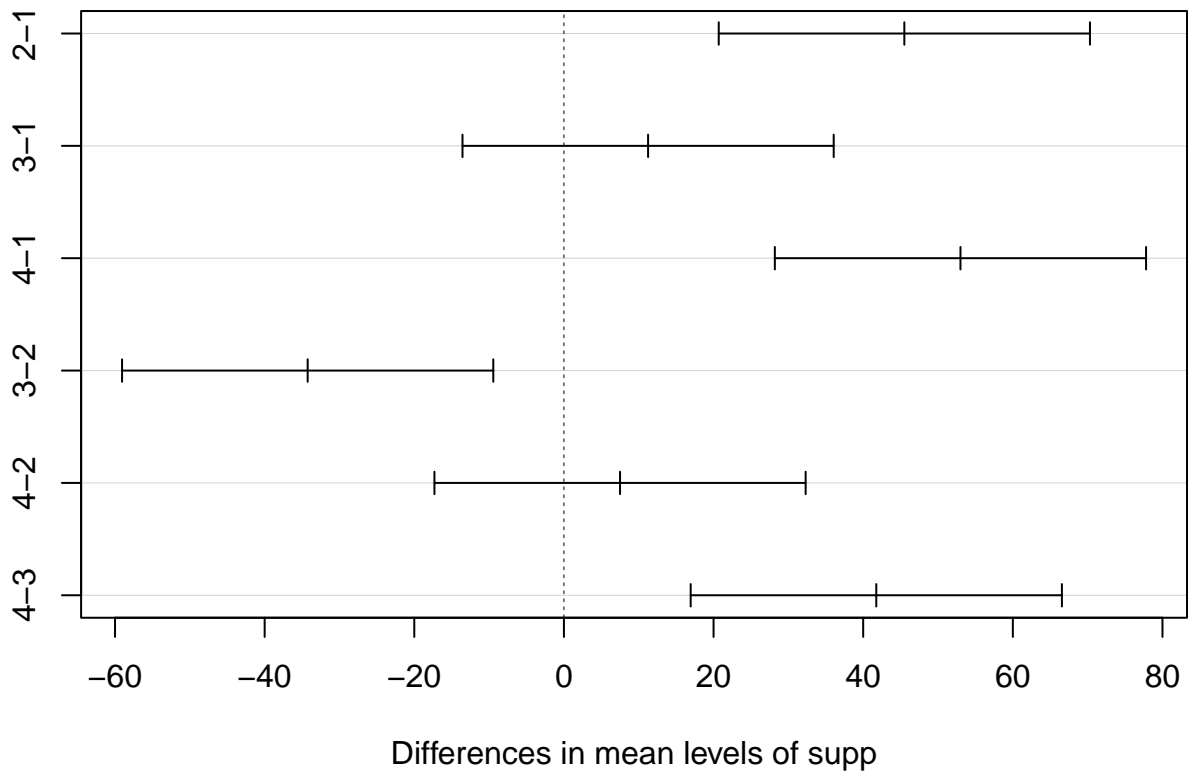
(d) State whether you think the model assumptions are satisfied by these data. Write a couple of sentences. If you do not think the assumptions are satisfied, give some advice about what to do.

(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.

- (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.

```
Tukey_out <- TukeyHSD(aov(y~supp,data=tensile))  
plot(Tukey_out,cex = .5)
```

95% family-wise confidence level



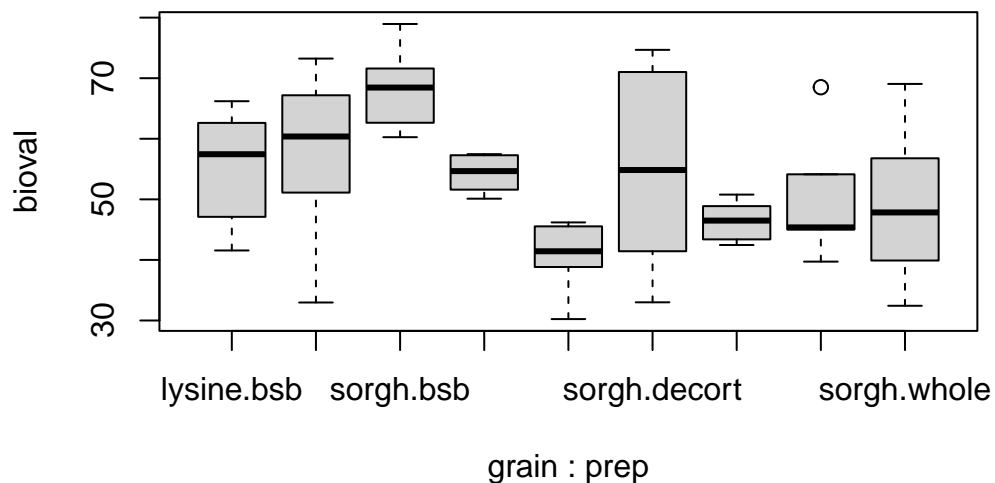
3. Two-way factorial design

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.

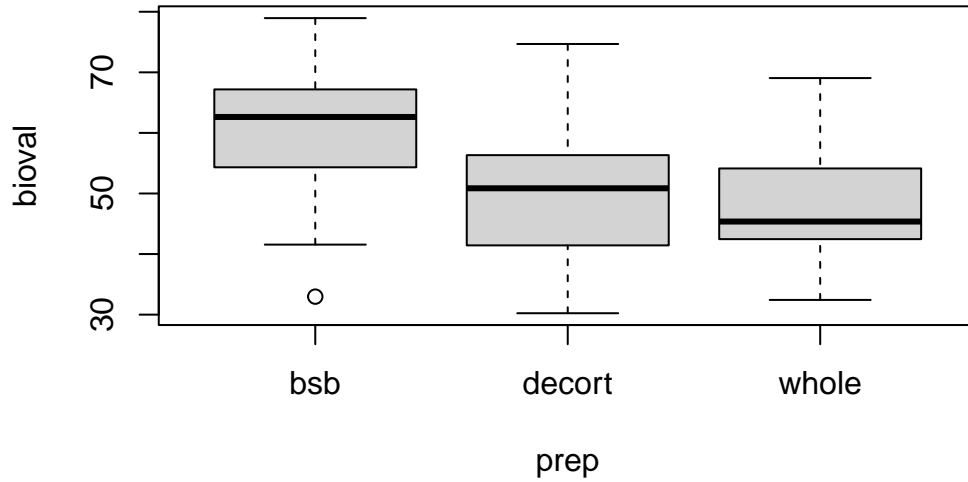
```
head(diet,n=12)
```

```
   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
```

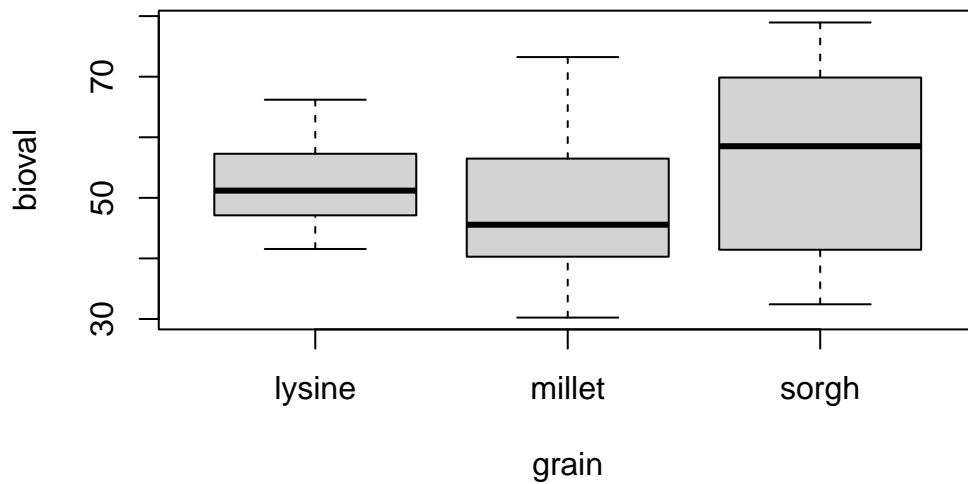
```
boxplot(bioval ~ grain + prep, data = diet)
```



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



```
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 $\text{Normal}(0, \sigma^2)$ 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 .
- (b) Use the estimated coefficients (printed below) from the two-way treatment effects model to write an expression giving the mean of the responses in the group of rats fed the diet at the factor level combination sorghum \times whole (you do not have to evaluate your expression).

```
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
A		SS_A	MS_A	2.9334	0.06346
B		SS_B	MS_B	7.3265	0.00176
AB		SS_{AB}	MS_{AB}	1.8531	0.13533
Error		SS_{Error}	105.89		
Total		SS_{Tot}			

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



- (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.

- (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?

(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
```

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 $\text{Normal}(0, \sigma^2)$ 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 $\bar{\mu}_{i\cdot}$ for $i = 1, 2$ and $\bar{\mu}_{\cdot j}$ for $j = 1, 2, 3$.
- (b) Is there interaction between the two factors? Explain your answer.
- (c) Carefully draw an interaction plot with the level $j = 1, 2, 3$ along the horizontal axis.