Solutions

Review Supplement for Math 17

Data set description:
The olive oil data consists of the composition of 8 fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, arachidic, linolenic, eicosenoic) found in the lipid fraction of 572 Italian olive oils, along with variables for the region and area where the olive oil was produced. The composition variables are all quantitative, while region and area are numerically coded categorical variables. There are 3 regions and 9 areas of interest (4 in region 1, 2 in region 2, and 3 in region 3). Units for each fatty acid are not given, but definitely differ (some measured in thousands, some tens, etc.).

Name that Scenario:

For each scenario, determine the most appropriate analysis. On the subsequent pages, there is enough output to complete the correct analysis and check some of the assumptions relating to each. You should be able to jot down a few notes (or calculations in some cases) and your final conclusions with regards to each question. You may assume any assumption you cannot check with the given output is satisfied.

1. Do the three regions have different levels of stearic in their olive oils? ANOVA

2. Is there a relationship between linolenic and linoleic levels in the olive oils? regression

3. Are there equal numbers of olive oils produced in each of the three regions? χ^2 GoF

4. Are higher levels of palmitic found in the olive oils from region 2 compared to region 1 on average?
$\mu_1 - \mu_2$ Hyp. Test

5. Do olive oils tend to have more palmitoleic than arachidic (same units) in their composition?
μ_d Hyp. Test

6. Are there differences between the areas in terms of levels of palmitoleic? ANOVA

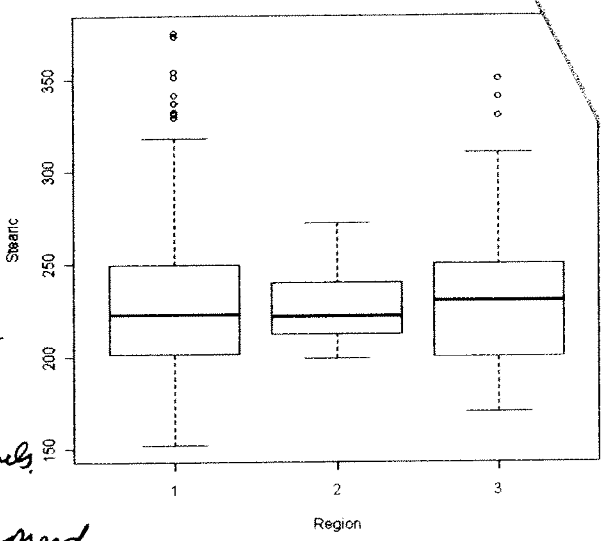
7. Is there a relationship between the levels of palmitic and palmitoleic in the olive oils?
regression

8. Are levels of arachidic greater than 55 on average for olive oils represented by this data set?
μ Hyp Test

9. Do more than 40% of olive oils have "high" eicosenoic levels (high means > 21 units)?
p Hyp Test

10. Does region 3 have more olive oils with "high" oleic levels (high means > 7500 units) than region 1?
$p_1 - p_2$ Hyp Test

1. Do the three regions have different levels of stearic in their olive oils?

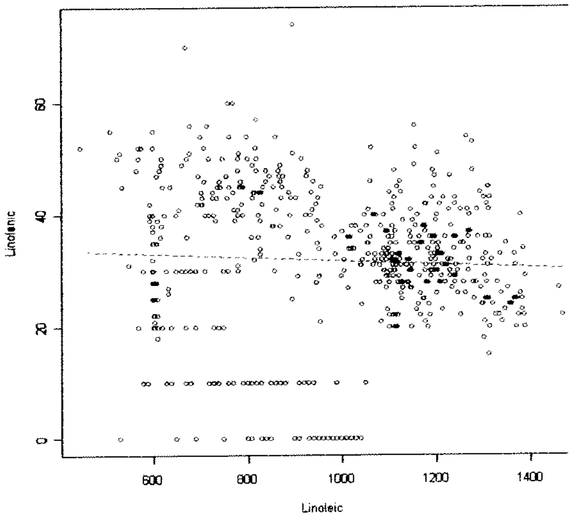


```
summary(AnovaModel.1)
                 Df  Sum Sq  Mean Sq  F value  Pr(>F)
Region            2    1273      637   0.4707  0.6248
Residuals       569  769685     1353
```

There does not seem to be evidence that there is @ least 1 difference among the regions in terms of average stearic levels.

Might be concerned region 2 has smaller spread.

2. Is there a relationship between linolenic and linoleic levels in the olive oils?



```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     34.896354   2.256080  15.468   <2e-16
Linoleic        -0.003068   0.002234  -1.374     0.17

Residual standard error: 12.96 on 570 degrees of
freedom
Multiple R-squared: 0.003299
F-statistic: 1.887 on 1 and 570 DF,p-value: 0.1701
```

There is no evidence that the slope is non-zero.

⇒ No relationship between linoleic and linolenic levels.

3. Are there equal numbers of olive oils produced in each of the three regions?

Region 1 has 323 oils, Region 2 has 98 and Region 3 has 151 based on this sample of olive oils.

χ^2 GoF $H_0: p_1 = p_2 = p_3 = 1/3$ $n = 572$

$EC = \frac{1}{3}(572) = 190.6\bar{6}$ All EC ≥ 5. ✓ $df = 3-1 = 2$

$$\chi^2 = \frac{(323-190.6\bar{6})^2}{190.6\bar{6}} + \frac{(98-190.6\bar{6})^2}{190.6\bar{6}} + \frac{(151-190.6\bar{6})^2}{190.6\bar{6}} =$$

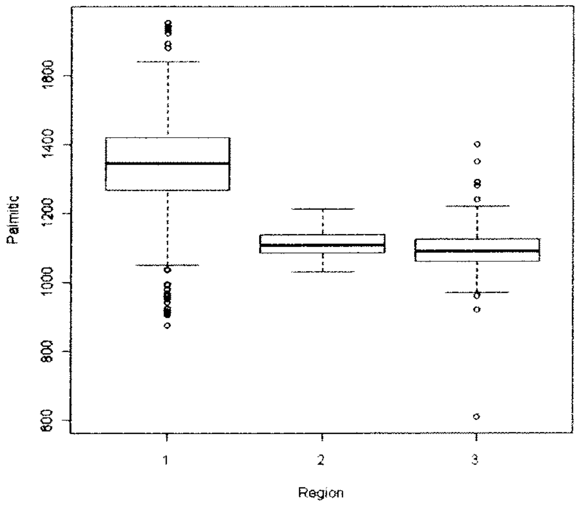
$$= 91.8592 + 45.03239 + 8.248946$$

$$= 145.1405$$

Reject H_0

We have evidence the 3 regions do not have equal proportions (#s) of olive oils.

4. Are higher levels of palmitic found in the olive oils from region 2 compared to region 1 on average?



```
Welch Two Sample t-test

data:  Dataset$Palmitic[1:323] (Region 1) -
Dataset$Palmitic[324:421] (Region 2)
t = 23.4097, df = 414.402, p-value = 1
alternative hypothesis: true difference in
means is less than 0
95 percent confidence interval:
     -Inf 236.4999
sample estimates:
mean of x mean of y
 1332.288  1111.347
```

Definitely not. There is no evidence that region 2 has higher levels of palmitic than region 1, on average.

Extra: What would your test stat, p-value, and conclusion have been if the question was whether region 1 had higher levels of palmitic than region 2 on average?

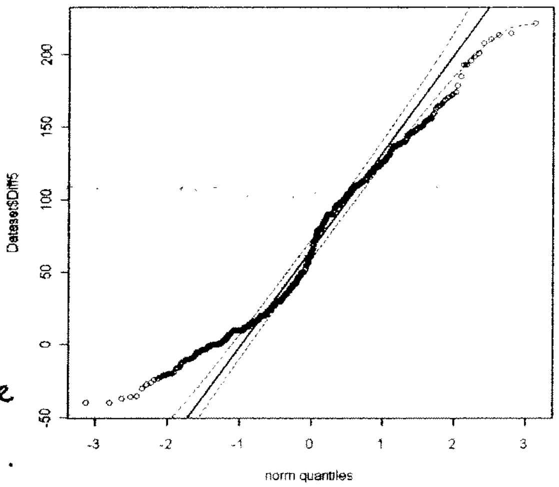
$t = -23.4097$ if swap order   p-value ≈ 0

Very strong evidence that region 1 has higher average palmitic levels than region 2.

Note: Code to run this previous test has to be altered from Rcmdr defaults due to data structure.

5. Do olive oils tend to have more palmitoleic than arachidic (same units) in their composition?

```
Paired t-test

data:  Dataset$Palmitoleic - Dataset$Arachidic
t = 29.4791, df = 571, p-value < 2.2e-16
alternative hypothesis: true difference in means
is greater than 0
```

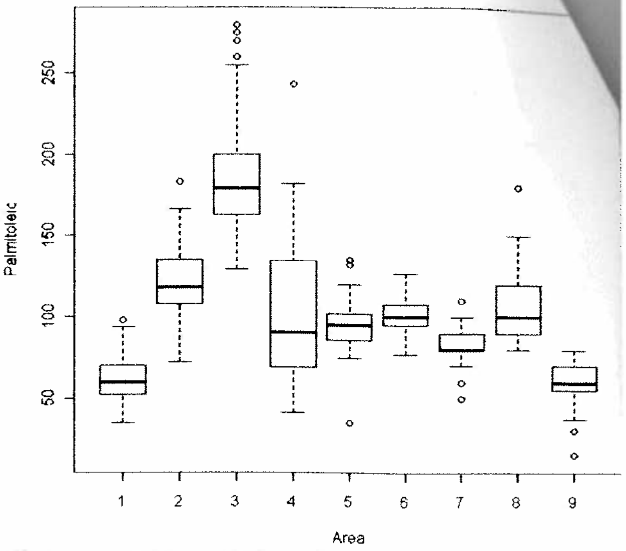


QQ Plot of differences

Yes. There is strong evidence that olive oils tend to have more palmetoleic than arachidic, on avg.

(Note some differences are −, but most are +)

6. Are there differences between the areas in terms of levels of palmitoleic?



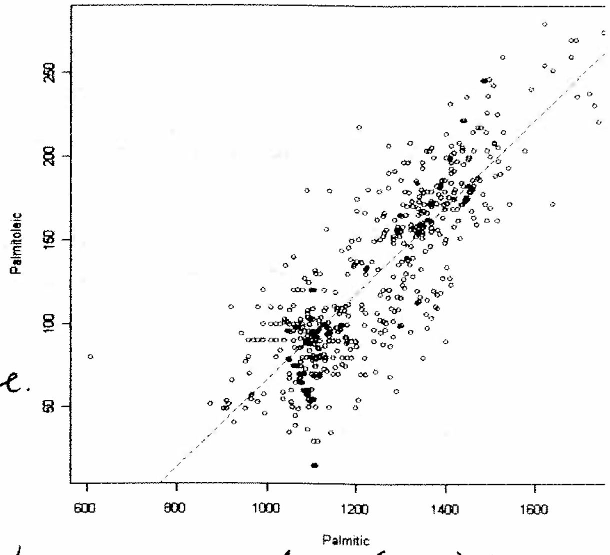
```
summary(AnovaModel.1)
           Df  Sum Sq   Mean Sq   F value      Pr(>F)
Area        8 1224057    153007    246.53   < 2.2e-16
Res       563  349424       621
```

It looks like the areas do have diff levels, however, since we do not have equal spread btw the groups, we cannot use ANOVA.

Note: You should have run into a major assumption concern here.

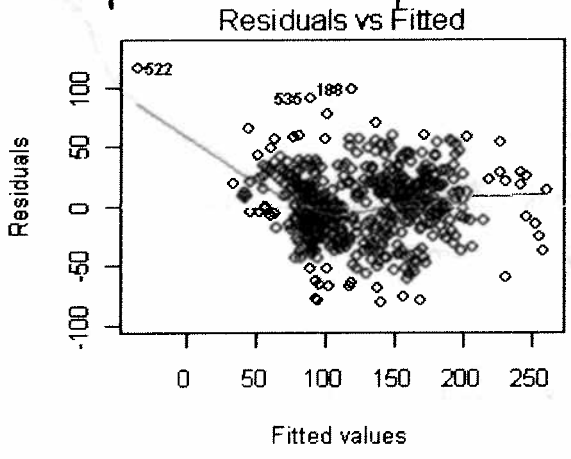
7. Is there a relationship between the levels of palmitic and palmitoleic in the olive oils?

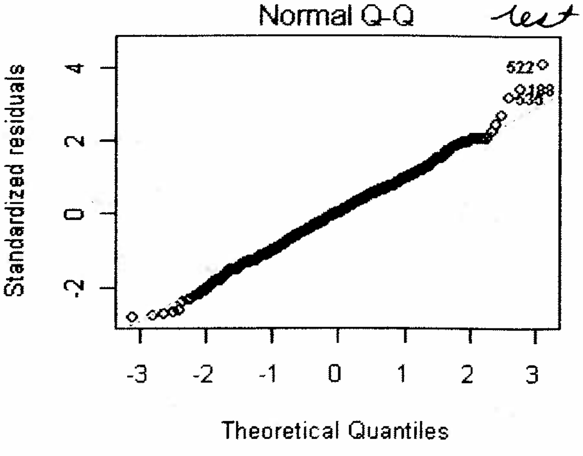
```
Coefficients:
           Estimate    Std. Error t value Pr(>|t|)
(Inter)  -1.944e+02  8.907e+00  -21.82    <2e-16
Palmitic  2.602e-01  7.164e-03   36.32    <2e-16

Res standard error: 28.86 on 570 df
Multiple R-squared: 0.6982
F-statistic:  1319 on 1 and 570 DF, p-value: <
2.2e-16
```



There is evidence of a non-zero slope. So, yes, there is a linear relationship btw palm. & palmitoleic.

All regression assumptions look good. We see an outlier (522) but rest are good





8. Are levels of arachidic greater than 55 on average for olive oils represented by this data set?

```
data:  Dataset$Arachidic
t = 63.0724, df = 571, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 56.28868 59.90712
sample estimates:
mean of x
  58.0979
```

↗ adjust to be one-sided

$> \Rightarrow$ still small

$\text{p-value} < 1.1 \times 10^{-16}$

There is evidence that the average arachidic level is > 55.

9. Do more than 40% of olive oils have "high" eicosenoic levels (high means > 21 units)?

4 decimals ↓

246 of the 572 olive oils in this sample have "high" eicosenoic levels

$H_0: p = .4 \qquad H_A: p > .40 \qquad \hat{p} = \frac{246}{572} = .4301$

$np_0, n(1-p_0) \geq 10$?

228.8, 343.2 ✓

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.4301 - .4}{\sqrt{\frac{.4(.6)}{572}}} = \frac{.0301}{.0205} = 1.47$$

Assuming $\alpha = .05$, we cannot reject H_0.

$\text{p-value} = P(z > 1.47) = .0708$

We do not have evidence that more than 40% of oils have high eicosenoic levels.

10. Does region 3 have more olive oils with "high" oleic levels (high means > 7500 units) than region 1?

42 of the 323 olive oils from region 1 have high oleic levels while 146 of the 151 olive oils from region 3 have high oleic oils.

p_1 = prop from region 1

p_2 = " " region 3

$n_1\hat{p}_1 = 42$

$n_2\hat{p}_2 = 146$

$n_1(1-\hat{p}_1) = 281$

$n_2(1-\hat{p}_2) = 5$

ACK!

but using $\hat{p}_c$ ← smaller is ≥ 10. So ok.

$H_0: p_1 = p_2$

$H_A: p_1 < p_2 \qquad p_1 - p_2 < 0$

$\hat{p}_1 = \frac{42}{323} \qquad \hat{p}_2 = \frac{146}{151}$

$$\hat{p} = \frac{42 + 146}{323 + 151} = \frac{188}{474} = .3966$$

$$z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} = \frac{.1300 - .9669}{\sqrt{\frac{.3966(.6034)}{323} + \frac{.3966(.6034)}{151}}}$$

$$= \frac{-.8369}{.0482} = -17.36$$

$P(z < -17.36) \approx 0$

off chart

There is evidence that region 3 has a higher proportion of olive oils with high oleic levels than region 1.