# Business Analytics (107-1)

# Assignment 3 – Solutions

**1. (DADM, P12.60)**

(a)

```
th <- read.table("toothpaste.txt", header=T, sep='\t')
th.0 <- lm(Sales ~ StoreLocation + StoreType + Display, data=th)
summary(th.0)
```

```
Call:
lm(formula = Sales ~ StoreLocation + StoreType + Display)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     35.556      2.613  13.608  0.00536 **
StoreLocationS   6.667      2.419   2.756  0.11031
StoreLocationU  17.000      2.419   7.028  0.01965 *
StoreTypeDI      8.333      2.419   3.445  0.07492 .
StoreTypeGR    -10.667      2.419  -4.409  0.04778 *
DisplayB        14.000      2.419   5.787  0.02858 *
DisplayC         7.667      2.419   3.169  0.08679 .
---
Residual standard error: 2.963 on 2 degrees of freedom
Multiple R-squared: 0.9865,    Adjusted R-squared: 0.9459
F-statistic: 24.29 on 6 and 2 DF,  p-value: 0.04006
```

Relevel the predictors:

```
StoreLocation <- relevel(StoreLocation, ref="U")
StoreType <- relevel(StoreType, ref="GR")

th.1 <- lm(Sales ~ StoreLocation + StoreType + Display)
summary(th.1)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     41.889      2.613  16.032  0.00387 **
StoreLocationR -17.000      2.419  -7.028  0.01965 *
StoreLocationS -10.333      2.419  -4.272  0.05067 .
StoreTypeDE     10.667      2.419   4.409  0.04778 *
StoreTypeDI     19.000      2.419   7.854  0.01583 *
```

```
DisplayB        14.000     2.419   5.787  0.02858 *
DisplayC         7.667     2.419   3.169  0.08679 .
```

The regression fit is quite good, and all dummies are significant, at least at the 0.08 level. Be aware, however, that it's risky to use this many explanatory variables when there are only 9 observations. It would be better if we had more data.

(b)
Based on the regression coefficients, the best location is urban, the best type is discount, and the best display is B.

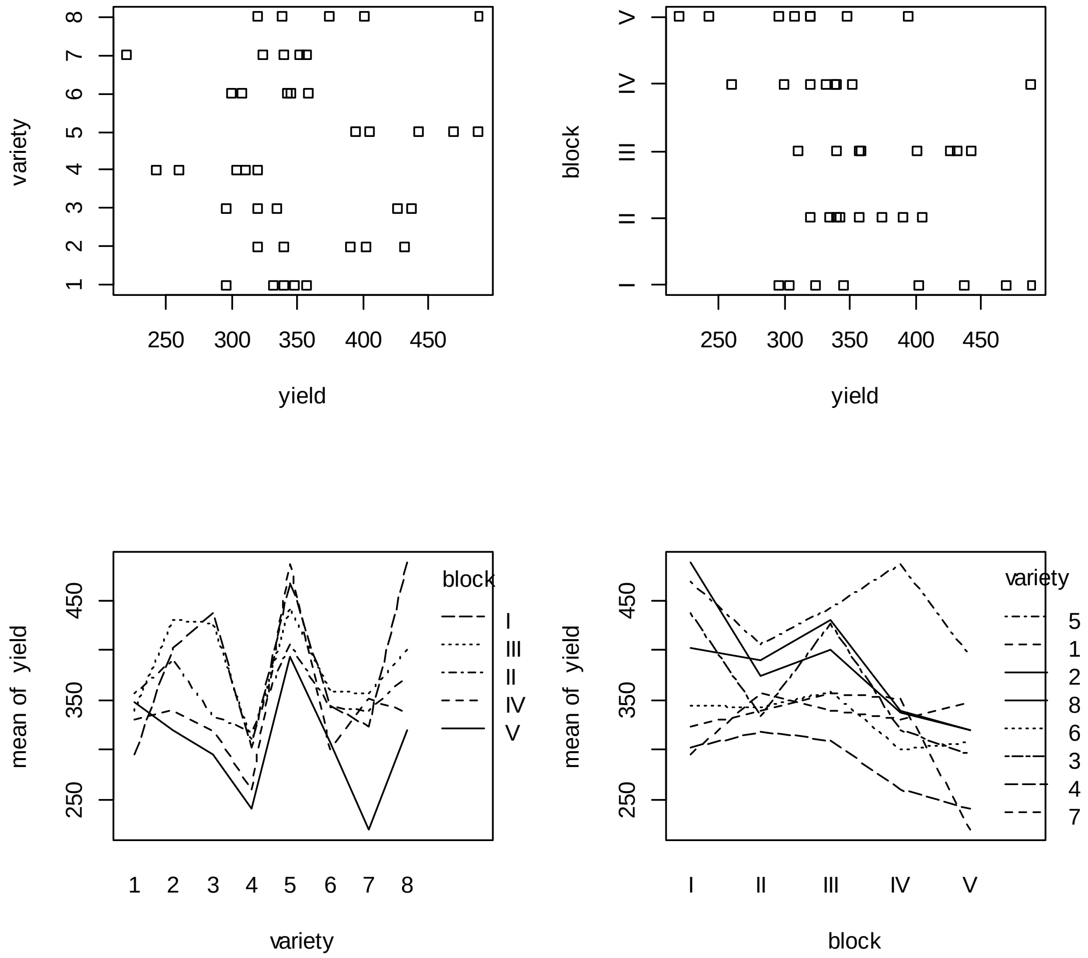
(c)
```
s.hat <- predict(th.1,
newdata=data.frame(StoreLocation="U",StoreType="DI",Display="B"))

summary(th.1)$sigma
1-pnorm(80, s.hat, summary(th.1)$sigma)
```

The probability that 80 or more toothpaste will be sold during a week is 0.0423.

(d)
A glance at the pattern in columns B, C, and D of the Data sheet shows that the displays are "scrambled" for each group of three weeks. So there is no relationship between the predictors, hence no problem with multicollinearity.

**2.**
(a)
Randomized Block Design

(b)
```
oatvar <- read.table("oatvar.txt", header=T, sep="\t")
attach(oatvar)
xtabs(yield ~ variety + block)
par(mfrow=c(2,2))
stripchart(yield ~ variety,xlab="yield",ylab="variety")
stripchart(yield ~ block,xlab="yield",ylab="block")
interaction.plot(variety,block,yield)
interaction.plot(block,variety,yield)
```




From the plots above, interaction effects between the variety of oats and the growing area block need to be taken into account.

(c)

```
oatvar$variety <- as.factor(oatvar$variety)
ot <- lm(yield ~ block+variety, oatvar)
summary(ot); anova(ot)
```

*Analysis of Variance Table*

*Response: yield*

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | |
|---|---|---|---|---|---|---|
| block | 4 | 33396 | 8349 | 6.2449 | 0.001008 | ** |
| variety | 7 | 77524 | 11075 | 8.2839 | 1.804e-05 | *** |
| Residuals | 28 | 37433 | 1337 | | | |

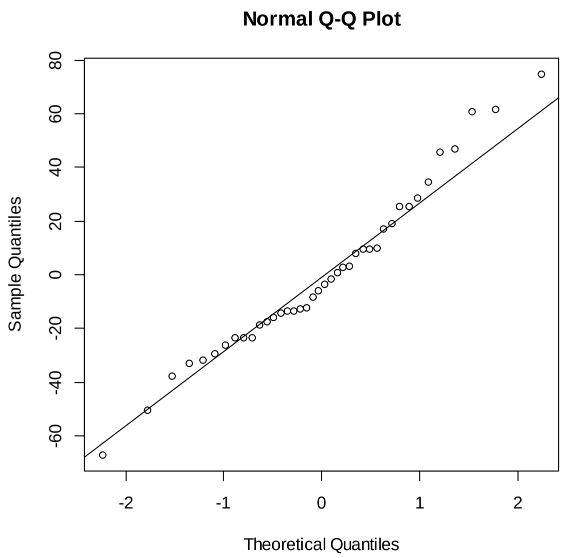
H0: There is no difference in population mean yield of oats based on varieties

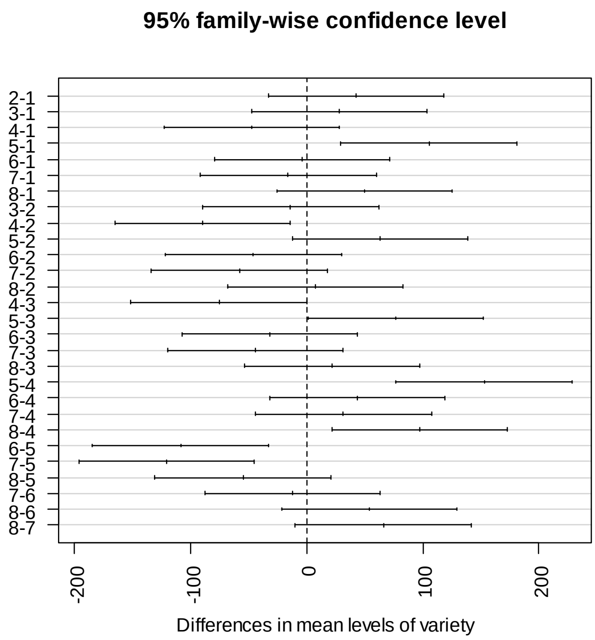
P-value is `1.804e-05` < 0.05. The data suggests to reject H0.
We conclude that yield of oats is affected by different varieties. Further details are provided by model summary `summary(ot)`.

(d)

```
plot(fitted(ot),residuals(ot),xlab="Fitted",ylab="Residuals")
abline(h=0)
qqnorm(residuals(ot))
qqline(residuals(ot))
```

By and large, the QQ plot looks fine.




(e)

```
othsd <- TukeyHSD(aov(yield ~ block+variety, oatvar), "variety")
par(mfrow=c(1,1))
plot(othsd, las=2)
```

**3.**

(a)

Latin Square Design

(b)

$$y_{ijl} = \mu + \alpha_i + \beta_j + \tau_l + \varepsilon_{ijl} \quad i, j, l = 1, \ldots, 4$$

(c)

```
fabric <- read.table("fabric.txt", header=T, sep="\t")
fabric
ab <- lm(result ~ area+factor(run)+factor(position), fabric)
drop1(ab, test="F")
```

```
Single term deletions
Model:
result ~ area + factor(run) + factor(position)
                 Df Sum of Sq     RSS     AIC     F value     Pr(F)
<none>                          16.000  20.000
area              3    40.000    56.000  34.044      5   0.0451975 *
factor(run)       3    24.000    40.000  28.661      3   0.1169598
factor(position)  3   216.000   232.000  56.786     27   0.0006987 ***
```

```
anova(ab)
```

```
                 Df  Sum Sq   Mean Sq   F value    Pr(>F)
area              3  40.000  13.333      5       0.0451975 *
factor(run)       3   24.000   8.000      3       0.1169598
factor(position)  3  216.000  72.000     27      0.0006987 ***
Residuals         6   16.000   2.667
```

```
summary(ab)
```

```
lm(formula = result ~ area + factor(run) + factor(position), data = fabric)

Coefficients:
                 Estimate    Std. Error    t value       Pr(>|t|)
(Intercept)       2.000e+01   1.291e+00   15.492     4.58e-06 ***
areaB            4.000e+00    1.155e+00    3.464     0.013400 *
areaC            3.000e+00    1.155e+00    2.598     0.040767 *
```

```
areaD             1.000e+00   1.155e+00    0.866     0.419753
factor(run)2       1.000e+00   1.155e+00    0.866     0.419753
factor(run)3       7.978e-16   1.155e+00 6.91e-16     1.000000
factor(run)4       3.000e+00   1.155e+00    2.598     0.040767 *
factor(position)2  1.000e+00  1.155e+00    0.866     0.419753
factor(position)3 -8.000e+00  1.155e+00   -6.928     0.000448 ***
factor(position)4 -5.000e+00  1.155e+00   -4.330     0.004928 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.633 on 6 degrees of freedom
Multiple R-squared: 0.9459,    Adjusted R-squared: 0.8649
F-statistic: 11.67 on 9 and 6 DF,  p-value: 0.003666
```

Position and area will affect the results of output.