MAST30025: Linear Statistical Models

Solution to Week 11 Lab

1. We study the effect of various breeds and diets on the milk yield of cows. A study is conducted on 9 cows and the following data obtained:

	Diet		
Breed	1	2	3
1	18.8	16.7	19.8
	21.2		23.9
2	22.3	15.9	21.8
		19.2	

   (a) Express this as a two-factor model with no interaction in matrix form.
   Solution: $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$$\mathbf{y} = \begin{bmatrix} 18.8 \\ 21.2 \\ 16.7 \\ 19.8 \\ 23.9 \\ 22.3 \\ 15.9 \\ 19.2 \\ 21.8 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

   and $\boldsymbol{\varepsilon}$ is as expected.

   (b) Express this as a two-factor model with interaction in matrix form.
   Solution: $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$$\mathbf{y} = \begin{bmatrix} 18.8 \\ 21.2 \\ 16.7 \\ 19.8 \\ 23.9 \\ 22.3 \\ 15.9 \\ 19.2 \\ 21.8 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \xi_{11} \\ \xi_{12} \\ \xi_{13} \\ \xi_{21} \\ \xi_{22} \\ \xi_{23} \end{bmatrix}$$

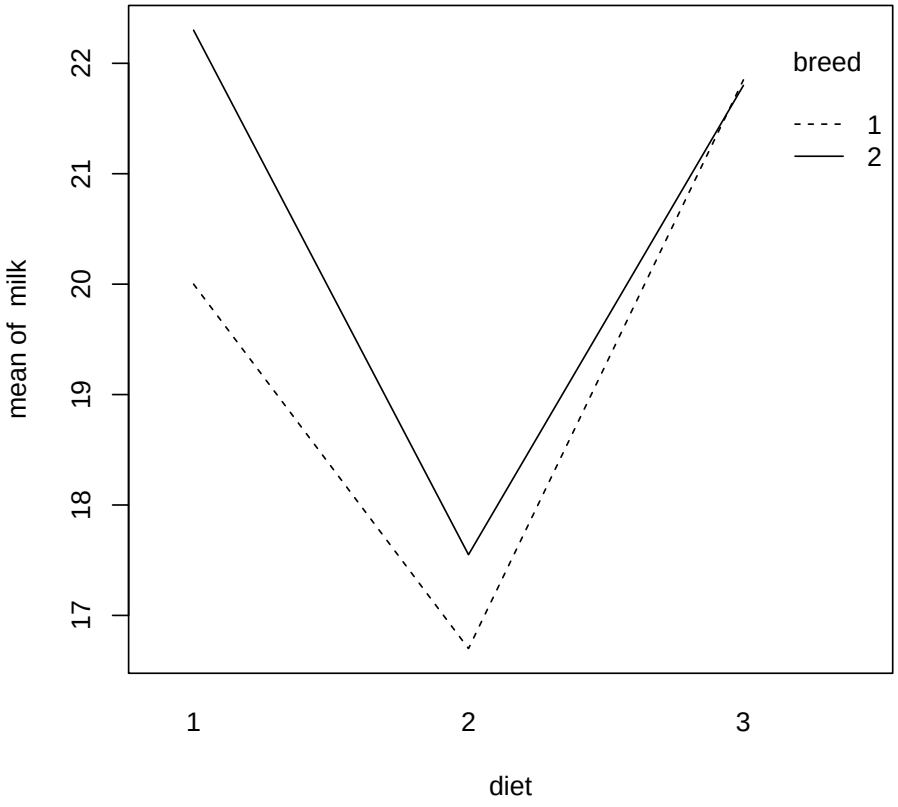
   and $\boldsymbol{\varepsilon}$ is as expected.

   (c) Express the hypothesis that there is no interaction in terms of your parameters. Eliminate any redundancies.
   Solution: We know that we require $(a-1)(b-1) = 2$ hypotheses, so we take the obviously non-redundant hypotheses

$$\begin{aligned} (\xi_{11} - \xi_{12}) - (\xi_{21} - \xi_{22}) &= 0 \\ (\xi_{11} - \xi_{13}) - (\xi_{21} - \xi_{23}) &= 0. \end{aligned}$$

   (d) Input this data into R. Plot an interaction plot between breed and diet.
   Solution:

```
> milk <- data.frame(milk=c(18.8,21.2,16.7,19.8,23.9,22.3,15.9,19.2,21.8),
+                    diet=factor(c(1,1,2,3,3,1,2,2,3)),
+                    breed=factor(c(1,1,1,1,1,2,2,2,2)))
> with(milk, interaction.plot(diet, breed, milk))
```



(e) Test for the presence of interaction.

Solution:

```
> imodel <- lm(milk ~ breed * diet, data=milk)
> anova(imodel)
Analysis of Variance Table

Response: milk
           Df Sum Sq Mean Sq F value Pr(>F)
breed       1  0.174  0.1742  0.0312 0.8710
diet        2 36.204 18.1018  3.2460 0.1777
breed:diet  2  1.874  0.9372  0.1681 0.8527
Residuals   3 16.730  5.5767
```

There is clearly no interaction.

(f) What is the degrees of freedom used for the interaction test?

Solution: We use 2 and 3 degrees of freedom.

(g) From the interaction model, what is the estimated amount of milk produced from breed 2 and diet 3?

Solution:

```
> newcow <- data.frame(diet=factor(3),breed=factor(2))
> predict(imodel,newcow)
```

```
   1
21.8
```

```
> imodel$coeff
```

```
 (Intercept)       breed2        diet2        diet3 breed2:diet2 breed2:diet3
       20.00         2.30        -3.30         1.85        -1.45        -2.35
```

```
> c(1,1,0,1,0,1)%*%imodel$coeff
```

```
     [,1]
[1,] 21.8
```

(h) Fit an additive model. What is the estimated amount of milk produced from breed 2 and diet 3 now?

Solution:

```
> amodel <- lm(milk ~ breed + diet, data=milk)
> predict(amodel,newcow)
```

```
       1
22.52222
```

```
> amodel$coeff
```

```
(Intercept)      breed2       diet2       diet3
  20.422222    1.033333   -3.844444    1.066667
```

```
> c(1,1,0,1)%*%amodel$coeff
```

```
         [,1]
[1,] 22.52222
```

(i) Test the hypothesis (under the additive model) that the 2nd and 3rd diets are equivalent in terms of milk produced.

Solution:

```
> library(car)
> linearHypothesis(amodel, c(0,0,1,-1),0)
```

```
Linear hypothesis test

Hypothesis:
diet2 - diet3 = 0

Model 1: restricted model
Model 2: milk ~ breed + diet

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1      6 52.000
2      5 18.604  1    33.396 8.9752 0.03024 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We reject this hypothesis at a 5% level.

(j) Find a 95% confidence interval, under the additive model, for the amount of milk produced from breed 2 and diet 3. Use both matrix calculations and the `estimable` function from the `gmodels` package.

Solution: Using the design matrix:

```
> library(MASS)
> library(Matrix)
> n <- 9
> X <- matrix(0,n,6)
> X[,1] <- 1
> X[cbind(1:n,as.numeric(milk$breed)+1)] <- 1
> X[cbind(1:n,as.numeric(milk$diet)+3)] <- 1
```

```
> y <- milk$milk
> XtXc <- ginv(t(X) %*% X)
> b <- XtXc %*% t(X) %*% y
> r <- rankMatrix(X)
> s2 <- sum((y - X %*% b)^2)/(n - r)
> t <- c(1,0,1,0,0,1)
> mu23 <- t(t) %*% b
> wdth <- qt(.975, n - r)*sqrt(s2 * t(t) %*% XtXc %*% t)
> c(mu23 - wdth, mu23, mu23 + wdth)
```

```
[1] 18.82634 22.52222 26.21811
```

Alternatively we can use `estimable`. Note that like `linearHypothesis`, `estimable` requires that you express the estimated quantity in terms of the estimates R uses:

```
> library(gmodels)
> estimable(amodel, c(1,1,0,1), conf.int=0.95)
```

```
          Estimate Std. Error  t value DF     Pr(>|t|) Lower.CI Upper.CI
(1 1 0 1) 22.52222   1.437762 15.66477  5 1.927104e-05 18.82634 26.21811
```

(k) Find the same confidence interval under the interaction model.

Solution:

```
> estimable(imodel, c(1,1,0,1,0,1), conf.int=0.95)
```

```
              Estimate Std. Error  t value DF    Pr(>|t|) Lower.CI Upper.CI
(1 1 0 1 0 1)     21.8   2.361497 9.231434  3 0.002689148 14.28466 29.31534
```

(l) Why is the second interval wider than the first?

Solution: The second interval is wider than the first because we are attributing some degrees of freedom to the interaction term(s). The resulting loss in degrees of freedom for the residuals leads to greater error in our estimations.

2. We study the growth of peas when fed three different types of fertilizer. A study is conducted where the samples are divided into 6 "blocks", corresponding to different plots of land. The data is stored in the `npk` data frame in R. This data frame contains 5 variables:

   - block: label of the block of the sample
   - N: indicator (0/1) for the application of nitrogen
   - P: indicator (0/1) for the application of phosphate
   - K: indicator (0/1) for the application of potassium
   - yield: yield of peas in pounds/plot

(a) Fit an additive linear model with all variables; then repeat without the block variables. Does the fitted model change? Are the block variables significant?

Solution:

```
> blockmodel <- lm(yield ~ ., data = npk)
> summary(blockmodel)
```

```
Call:
lm(formula = yield ~ ., data = npk)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0000 -1.7083 -0.0833  2.2458  6.4833

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   53.800      2.450  21.955 8.13e-13 ***
block2         3.425      2.830   1.210  0.24483
block3         6.750      2.830   2.386  0.03068 *
```

```
block4        -3.900      2.830  -1.378  0.18831
block5        -3.500      2.830  -1.237  0.23512
block6         2.325      2.830   0.822  0.42412
N1             5.617      1.634   3.438  0.00366 **
P1            -1.183      1.634  -0.724  0.47999
K1            -3.983      1.634  -2.438  0.02767 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.002 on 15 degrees of freedom
Multiple R-squared:  0.7259,        Adjusted R-squared:  0.5798
F-statistic: 4.966 on 8 and 15 DF,  p-value: 0.003761
```

```
> amodel <- lm(yield ~ . - block, data = npk)
> summary(amodel)
```

```
Call:
lm(formula = yield ~ . - block, data = npk)

Residuals:
    Min      1Q  Median      3Q     Max
-9.2667 -3.6542  0.7083  3.4792  9.3333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   54.650      2.205  24.784   <2e-16 ***
N1             5.617      2.205   2.547   0.0192 *
P1            -1.183      2.205  -0.537   0.5974
K1            -3.983      2.205  -1.806   0.0859 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.401 on 20 degrees of freedom
Multiple R-squared:  0.3342,        Adjusted R-squared:  0.2343
F-statistic: 3.346 on 3 and 20 DF,  p-value: 0.0397
```

```
> anova(amodel,blockmodel)
```

```
Analysis of Variance Table

Model 1: yield ~ (block + N + P + K) - block
Model 2: yield ~ block + N + P + K
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     20 583.48
2     15 240.18  5     343.3 4.2879 0.01272 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The fitted model does not change, in the sense that the parameters corresponding to N, P and K are the same for the two models. This is because the design is *balanced*: the overall effect of the predictors of interest are observed in each individual block. However, the blocks themselves are significant: there is a difference in the yield of each block. Because the blocks have been carefully designed, this does not affect the fitted model itself when the blocks are removed from consideration.

(b) Fit a model with the fertilizer variables and all pairwise interaction terms. Are the interaction terms significant?

Solution:

```
> imodel <- lm(yield ~ (.-block)^2, data=npk)
> summary(imodel)
```

```
Call:
lm(formula = yield ~ (. - block)^2, data = npk)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8917 -3.2875  0.4583  3.4000  9.7083

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  52.6750     3.0114  17.492 2.64e-12 ***
N1            9.8500     3.9429   2.498    0.023 *
P1            0.4167     3.9429   0.106    0.917
K1           -1.9167     3.9429  -0.486    0.633
N1:P1        -3.7667     4.5529  -0.827    0.420
N1:K1        -4.7000     4.5529  -1.032    0.316
P1:K1         0.5667     4.5529   0.124    0.902
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.576 on 17 degrees of freedom
Multiple R-squared:  0.3968,        Adjusted R-squared:  0.184
F-statistic: 1.864 on 6 and 17 DF,  p-value: 0.146
```

```
> anova(amodel,imodel)

Analysis of Variance Table

Model 1: yield ~ (block + N + P + K) - block
Model 2: yield ~ ((block + N + P + K) - block)^2
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     20 583.48
2     17 528.58  3    54.898 0.5885 0.6308
```

The interaction terms are not significant.

(c) Perform variable selection using stepwise selection with AIC, starting from the model with no interaction terms (but considering them for inclusion). What do you find?

Solution:

```
> (finalmodel <- step(amodel, scope=~(.-block)^2, data=npk))

Start:  AIC=84.58
yield ~ (block + N + P + K) - block

       Df Sum of Sq    RSS    AIC
- P     1     8.402 591.88 82.926
<none>              583.48 84.583
+ N:K   1    33.135 550.34 85.180
+ N:P   1    21.282 562.20 85.691
- K     1    95.202 678.68 86.210
+ P:K   1     0.482 583.00 86.563
- N     1   189.282 772.76 89.326

Step:  AIC=82.93
yield ~ N + K

       Df Sum of Sq    RSS    AIC
<none>              591.88 82.926
+ N:K   1    33.135 558.75 83.543
- K     1    95.202 687.08 84.506
+ P     1     8.402 583.48 84.583
```

```
- N     1   189.282 781.16 87.586

Call:
lm(formula = yield ~ N + K, data = npk)

Coefficients:
(Intercept)           N1           K1
     54.058        5.617       -3.983
```

We find that our final model includes the variables corresponding to nitrogen and potassium, but not phosphate.

(d) What is the best treatment for peas, according to your final model? Find a 95% confidence interval for the yield of this treatment.

Solution: According to the final model, we should treat peas with nitrogen and not potassium (phosphate is unimportant). The confidence interval is:

```
> estimable(finalmodel, c(1,1,0), conf.int=0.95)
        Estimate Std. Error  t value DF Pr(>|t|) Lower.CI Upper.CI
(1 1 0)   59.675   1.876994 31.79286 21        0 55.77158 63.57842
```

3. A study was conducted to determine the effect of the size of the root system on the growth of Douglas-fir seedlings when they are planted out. Seedlings were obtained from three seed lots, and when they were planted out their root volume was classified as small (RV1), medium (RV2), or large (RV3). The heights of the seedlings were then measured at the end of the first growing season. The data from the experiment are given in the file `douglas.csv`.

(a) How has randomisation and blocking been used in the design of this experiment?

Solution: The data are blocked according to seed lots. `SeedLot` is the blocking factor, with three levels, and `RootVolume` is the treatment factor, with three levels and six replications.

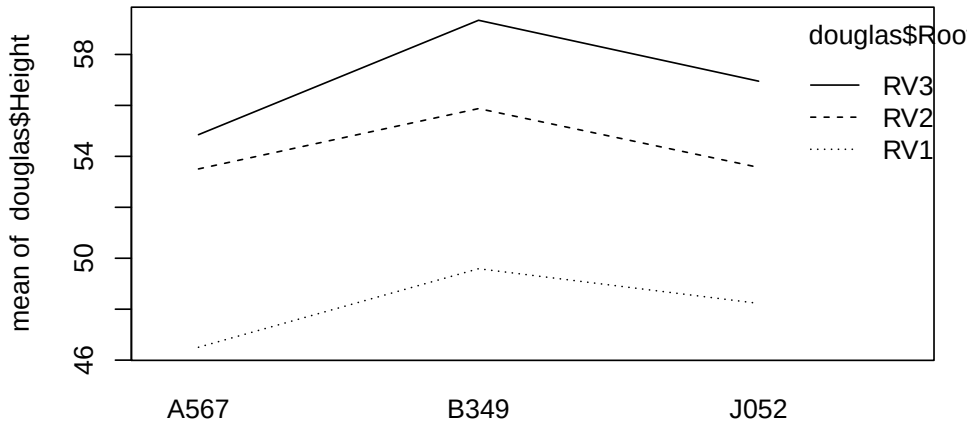
The allocation of `Plot` appears to have been randomised in the following way: each consecutive set of 9 plots have been randomly distributed among the possible `SeedLot` and `Height` combinations. Presumably this has been done to avoid any confounding between `SeedLot`, `Height`, and the physical location of the seedlings.

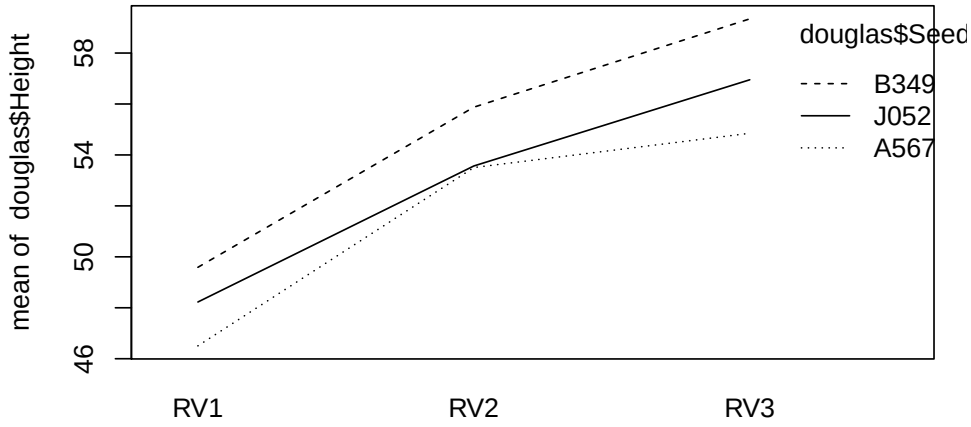
```
> douglas <- read.csv("../data/douglas.csv")
> idx <- order(douglas$SeedLot, douglas$RootVolume)
> douglas <- douglas[idx,]
> head(douglas)
   Plot RootVolume SeedLot Height
3     3        RV1    A567  46.97
13   13        RV1    A567  45.45
19   19        RV1    A567  47.34
29   29        RV1    A567  45.88
43   43        RV1    A567  44.08
53   53        RV1    A567  49.31
```

(b) Generate two interaction plots for the data. Is there any evidence of an interaction?

Solution:

```
> opar <- par(mfrow = c(2, 1), mar = c(3, 4, 1, 1))
> interaction.plot(douglas$SeedLot, douglas$RootVolume, douglas$Height)
> interaction.plot(douglas$RootVolume, douglas$SeedLot, douglas$Height)
> par <- opar
```





The lines are close to parallel, so there is little indication of an interaction.

(c) Fit a model with interaction to the data and use it to find the fitted means for each combination of factor levels.

Solution: Group means are given by the elements of $X\beta$. They come in blocks of six, corresponding to factor levels 11, 12, 13, 21, 22, 23, 31, 32, 33, where the first number is the index of `SeedLot` and the second is the index of `RootVolume`.

```
> imodel <- lm(Height ~ SeedLot * RootVolume, data=douglas)
> summary(imodel)

Call:
lm(formula = Height ~ SeedLot * RootVolume, data = douglas)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3383 -1.2537  0.0217  0.7733  4.0517

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                46.5050     0.7123  65.286  < 2e-16 ***
SeedLotB349                 3.0833     1.0074   3.061  0.00372 **
SeedLotJ052                 1.7217     1.0074   1.709  0.09433 .
RootVolumeRV2               7.0000     1.0074   6.949 1.21e-08 ***
RootVolumeRV3               8.3450     1.0074   8.284 1.34e-10 ***
SeedLotB349:RootVolumeRV2  -0.7133     1.4246  -0.501  0.61902
SeedLotJ052:RootVolumeRV2  -1.6600     1.4246  -1.165  0.25008
SeedLotB349:RootVolumeRV3   1.4083     1.4246   0.989  0.32817
SeedLotJ052:RootVolumeRV3   0.3783     1.4246   0.266  0.79179
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.745 on 45 degrees of freedom
Multiple R-squared:  0.8635,        Adjusted R-squared:  0.8393
F-statistic: 35.59 on 8 and 45 DF,  p-value: < 2.2e-16
```

```
> fitted(imodel)[6*(1:9)]
```

```
      53       47       54       49       52       46       48       51
46.50500 53.50500 54.85000 49.58833 55.87500 59.34167 48.22667 53.56667
      50
56.95000
```

(d) Find a 95% confidence interval for the difference in height between a seedling with large root volume (RV3) and a seedling with medium root volume (RV2). Suppose that the seedling came from seed lot B349.

Solution:

```
> library(gmodels)
> estimable(imodel, c(0,0,0,-1,1,-1,0,1,0), conf.int=0.95)
```

```
                     Estimate Std. Error  t value DF    Pr(>|t|) Lower.CI
(0 0 0 -1 1 -1 0 1 0) 3.466667   1.007378 3.441276 45 0.001260833 1.437703
                     Upper.CI
(0 0 0 -1 1 -1 0 1 0) 5.495631
```

(e) Test for the presence of an interaction at the 5% significance level. Would it be meaningful to check the significance of the main effects? Why?

Solution:

```
> amodel <- lm(Height ~ SeedLot + RootVolume, data=douglas)
> anova(amodel, imodel)
```

```
Analysis of Variance Table

Model 1: Height ~ SeedLot + RootVolume
Model 2: Height ~ SeedLot * RootVolume
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     49 147.67
2     45 137.00  4    10.674 0.8765 0.4855
```

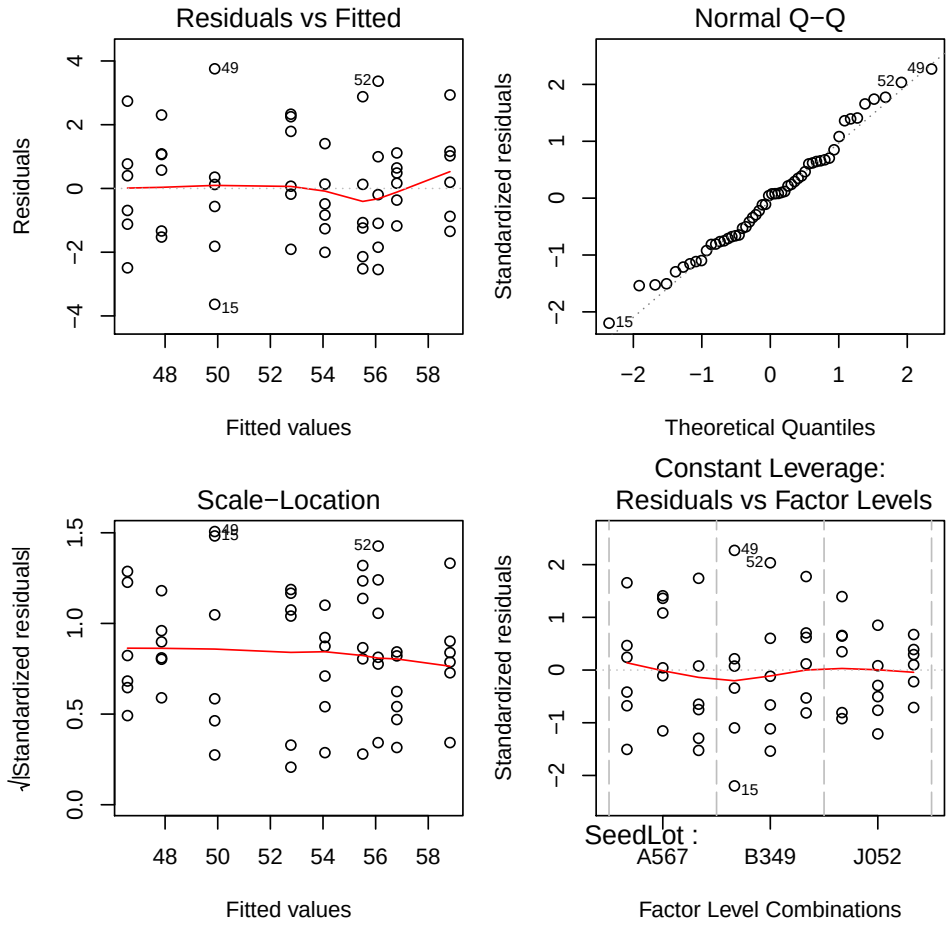
The large p-value indicates that we should retain the null hypothesis that there is no interaction.

In the absence of an interaction it does make sense to test for the significance of the main effects, under the additive model of course.

(f) Fit an additive model to the data using the `lm` command, and produce plots to justify the model assumption that the errors are normal and homoskedastic.

Solution:

```
> opar <- par(mfrow=c(2,2),mar=c(4,4,3,1))
> plot(amodel, which = 1)
> plot(amodel, which = 2)
> plot(amodel, which = 3)
> plot(amodel, which = 5)
> par <- opar
```



From the second plot, the residuals look normal. In the other three, there is no sign of heteroskedasticity or outliers.

4. Suppose that $\mathbf{y} \sim MVN(\mu\mathbf{1}, \Sigma)$, where

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix}.$$

For what values of ρ are the sample mean and sample variance independent?

Solution: Let J be the matrix of all 1s. We can write

$$\begin{aligned} \bar{y} &= \frac{1}{n}\mathbf{1}^T\mathbf{y}, \\ s^2 &= \frac{1}{n-1}\mathbf{y}^T(I - \tfrac{1}{n}J)\mathbf{y}. \end{aligned}$$

Thus they are independent iff $\mathbf{1}^T \text{Var } \mathbf{y}(I - \frac{1}{n}J) = 0$. The LHS is

$$\begin{aligned} \mathbf{1}^T((1-\rho)I + \rho J)(I - \frac{1}{n}J) &= \mathbf{1}^T\left((1-\rho)I + \rho J - \frac{1-\rho}{n}J - \rho J\right) \\ &= (1-\rho)\mathbf{1}^T(I - \tfrac{1}{n}J) \\ &= 0. \end{aligned}$$

So the sample mean and variance are independent for any ρ.

5. In the one-way classification model, show that any linear combination of $\bar{y}_1 - \bar{y}_\cdot, \ldots, \bar{y}_k - \bar{y}_\cdot$ can be written as a linear combination of $\bar{y}_1, \ldots, \bar{y}_k$. Does the converse hold?

   Solution: We have

$$\sum a_i(\bar{y}_i - \bar{y}_\cdot) = \sum a_i \bar{y}_i - \left(\sum a_i\right) \frac{1}{k} \sum \bar{y}_i = \sum (a_i - \bar{a})\bar{y}_i.$$

   The converse only holds for contrasts.