HW10 ANOVA

Doug Nychka

2024-04-27

Setup

```
suppressMessages(library( fields))

setwd("~/Dropbox/Home/Teaching/MATH531/MATH-531/MATH531S2024/Assignments")
```

The PVC dataset

Loading the small example data set **pvc** from the textbook *Linear Models with R*, Julian Faraway.

```
load("pvc.rda")
head( pvc)
```

```
##   psize resin operator
## 1  36.2     1        1
## 2  36.3     1        1
## 3  35.3     2        1
## 4  35.0     2        1
## 5  30.8     3        1
## 6  30.6     3        1
```

```
# each combination of resin/operator has two replications
table( pvc$resin, pvc$operator)
```

```
##    
##      1 2 3
##    1 2 2 2
##    2 2 2 2
##    3 2 2 2
##    4 2 2 2
##    5 2 2 2
##    6 2 2 2
##    7 2 2 2
##    8 2 2 2
```

About the dataset (from Faraway's R package help file)

Production of PVC by operator and resin railcar

Description

Data from an experiment to study factors affecting the production of the plastic PVC, 3 operators used 8 different devices called resin railcars to produce PVC. For each of the 24 combinations, two samples were produced.

Dataset contains the following variables

- psize Particle size
- operator Operator number 1, 2 or 3
- resin Resin railcar 1-8

Source

R. Morris and E. Watson (1998) "A comparison of the techniques used to evaluate the measurement process" *Quality Engineering*, 11, 213-219

For reference the complete (and default) 2-way analysis. Note the use of the * in the formula equation.

```
fullFit<- lm(psize ~ operator*resin, pvc )

ANOVATable<- anova( fullFit)
ANOVATable
```

```
## Analysis of Variance Table
##
## Response: psize
##                 Df  Sum Sq Mean Sq
## operator         2  20.718  10.359
## resin            7 283.946  40.564
## operator:resin 14  14.335   1.024
## Residuals      24  35.480   1.478
##                 F value    Pr(>F)
## operator         7.0072   0.00401 **
## resin           27.4388 5.661e-10 ***
## operator:resin  0.6926   0.75987
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

```
# the complete sums of squares subtracting off the grand mean is
SSTotal<- sum( (pvc$psize - mean(pvc$psize))^2)
SSTotal
```

```
## [1] 354.4792
```

```
# compare to
sum( ANOVATable[,2])
```

```
## [1] 354.4792
```

Problem 1

- The fields function **stats** is a handy function that will find summary statistics of the columns of a matrix or a data frame E.g. `stats(pvc)` in this case the summaries for operator and resin are listed as NA. Why?
- For these data compute the 1-way analysis "by hand" (use R basic arithmetic and functions but not **lm**) for the **resin** factor and the response **psize**. Specifically under the model:

$$Y_{ik} = \mu + \alpha_i + e_{ik}$$

  Determine the OLS estimates for the grand mean, μ, and main effects, $\{\alpha_i\}$ $i = 1, ..., 8$, where the main effects are constrained to sum to 0.

- Fit this model using **lm**. Do you get the same results? Which parts are the same and which are different?
- Without using **lm** calculate the F statistic for the null hypothesis that all the $\{\alpha_i\}$ are zero. Check your answer against the results of using **lm** and the **anova** function.

Problem 2

- Now consider both factors, **resin** and **operator** and estimate "by hand" the parameters in the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

  Here i indexes the 8 resin levels and j indexes the 3 operator levels. α_i are the 8 main effects for resin and β_j are the 3 main effects for operator. Again assume that both estimates of the main effects are constrained to sum to 0.

- Find the F statistics for testing both for the resin main effects being zero and separately for the operator effects being zero. Compare your computation to the full ANOVA given above in the introduction – why are they slightly different?
- An engineer knows that the resin railcars have an effect but then asks: 'What is the size of different operators impact on **psize** separate from the resin effect?' What would you tell him?

Problem 3

Now consider the full model.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

where $\{\gamma_{ij}\}$ are the 24 interaction terms.

- List the number of constraints on the main effects and interactions needed to give an X matrix that has full rank
- Examine the **lm** fit of this full model given above. Which interactions does the software choose to leave out to create a full rank X matrix.
- Following the definition of an interaction we have the estimates:

$$\hat{\gamma_{ij}} = \overline{Y}_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$$

  where $\overline{Y}_{ij}$ is the group mean for resin level i and operator level j.

Create a table of 48 rows and with 6 columns: the data, the grand mean, the resin main effects,the operator main effects, the interaction effects, and the residuals. Call this `tablePVC` and list out the table. (use `\newpage` to create a separate page for this in your pdf.)

If done correctly, columns 2 through 6 should sum to the first. E.g. `rowSums(tablePVC[,2:6])` is equal to `tablePVC[,1]`. How are the sum of squares of *columns* 2 through 6 `colSums(tablePVC[,2:6])` related to the two way ANOVA table?

Here is an example of indexing that might help you to fill out the table with the estimates. Note you will need to find **alphaHat** and **betaHat** and **muijHat**, the 8×3 matrix of group means to make this code work.

```
indexResin<-as.numeric( pvc$resin)
indexOperator<-as.numeric( pvc$operator)
vecMeans<- muijHat[ cbind(indexResin, indexOperator)]
vecGrand<- rep( mean( pvc$psize), 48)
vecAlpha<- alphaHat[indexResin]
vecBeta<-  betaHat[indexOperator]
vecGamma<- vecMeans - vecAlpha - vecBeta + vecGrand
residual<- pvc$psize - vecMeans
```

- Is there statistical evidence that there is an interaction effect between resin and operators?