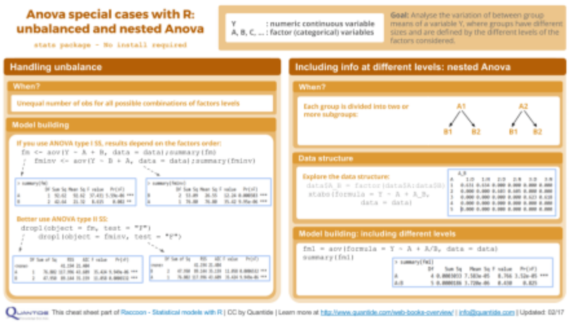
R news and tutorials contributed by (750) R bloggers

Raccoon | Ch 2.5 – Unbalanced and Nested Anova

February 21, 2017
By Quantide

(This article was first published on **R blog | Quantide - R training & consulting**, and kindly contributed to R-bloggers)



Download the Unbalanced and Nested Anova cheat sheet in full resolution: [Anova special cases]

This article is part of Quantide's web book "Raccoon – Statistical Models with R". Raccoon is Quantide's third web book after "Rabbit – Introduction to R" and "Ramarro – R for Developers". See the full project here.

The second chapter of Raccoon is focused on T-test and Anova. Through example it shows theory and R code of:

- 1-sample t-test
- 2-sample t-test and Paired t
- 1-way Anova
- 3-way Anova
- Unbalanced and Nested Anova

This post is the fifth section of the chapter, about Unbalanced Anova with one obs dropped and fixed effects Nested Anova.

Throughout the web-book we will widely use the ***package qdata****, containing about 80 datasets. You may find it here: https://github.com/quantide/qdata.*

Example: Brake distance 1 (unbalanced anova with one obs dropped)

Data description

We use the same data as in our previous article about 3-way Anova, but here one observation (row) has been removed. The detailed data description is in 3-way Anova.

Data loading

```
data(distance)
head(distance)
```

```
## # A tibble: 6 × 4
##     Tire  Tread      ABS Distance
##
## 1     GT     10  enabled 19.35573
## 2     GT    1.5 disabled 23.38069
## 3     MX    1.5  enabled 24.00778
## 4     MX     10  enabled 25.07142
## 5     LS     10 disabled 26.39833
## 6     GT     10  enabled 18.60888
```

```
str(distance)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    24 obs. of  4 variables:
##  $ Tire    : Factor w/ 3 levels "GT","LS","MX": 1 1 3 3 2 1 2 2 2 3 ...
```

```
##  $ Tread   : Factor w/ 2 levels "1.5","10": 2 1 1 2 2 2 1 1 2 2 ...
##  $ ABS     : Factor w/ 2 levels "disabled","enabled": 2 1 2 2 1 2 2 1 1 1 ...
##  $ Distance: num  19.4 23.4 24 25.1 26.4 ...
```

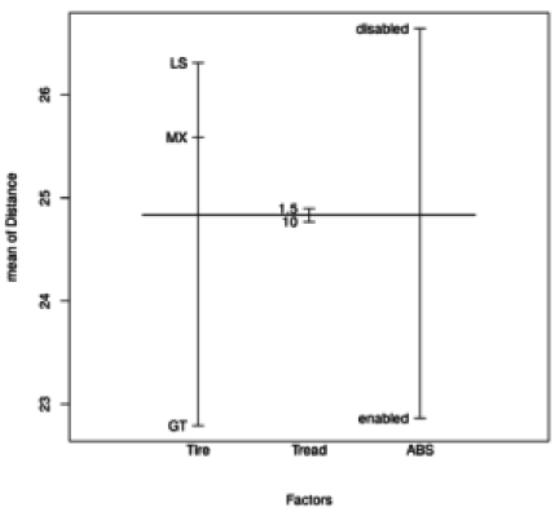
Let us drop one observation so that all the factor levels combinations do not contain the same number of observations. We drop the observation with the values "LS 10 enabled"

```
distance1 <- distance[-24,]
```

Descriptives

As usual, we first plot the univariate effects:

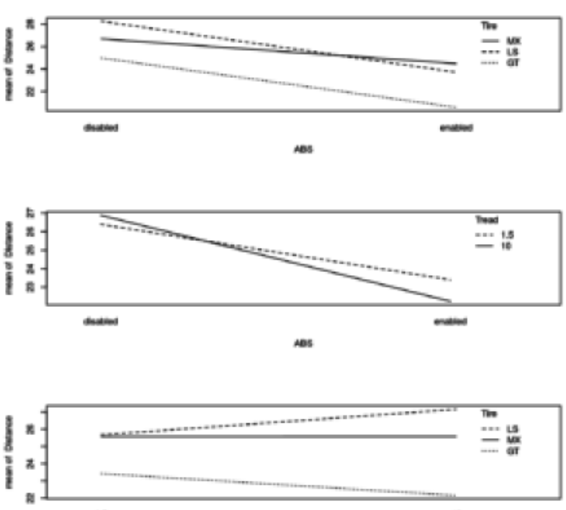
```
plot.design(Distance ~ ., data = distance1)
```



Univariate effects plot of unbalanced model

Secondly we look at the two-way interaction plot:

```
op <- par(mfrow = c(3, 1))
with(distance1, {
  interaction.plot(ABS, Tire, Distance)
  interaction.plot(ABS, Tread, Distance)
  interaction.plot(Tread, Tire, Distance)
  }
)
par(op)
```



Two-way interaction effects plots of unbalanced model

We notice that all effects do not seem to change with respect to the previous example of 3-way anova.

Inference and models

In this section is where we'll start noticing some differences with the balanced model. First let us fit the model with all interactions

```
fm <- aov(Distance ~ ABS * Tire * Tread, data = distance1)
summary(fm)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)    
## ABS             1  81.95   81.95  39.870 5.72e-05 ***
## Tire            2  47.95   23.98  11.665  0.00191 ** 
## Tread           1   0.19    0.19   0.092  0.76771    
## ABS:Tire        2   6.72    3.36   1.635  0.23898    
## ABS:Tread       1   3.26    3.26   1.588  0.23365    
## Tire:Tread      2   3.42    1.71   0.831  0.46107    
## ABS:Tire:Tread  2   4.99    2.50   1.215  0.33360    
## Residuals      11  22.61    2.06                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As none of the interactions are significant, we drop them one by one, starting from the three-way interaction:

```
fm <- update(fm, . ~ . -ABS:Tire:Tread)
summary(fm)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## ABS          1  81.95   81.95  38.594 3.16e-05 ***
## Tire         2  47.95   23.98  11.291  0.00144 ** 
## Tread        1   0.19    0.19   0.089  0.77049    
## ABS:Tire     2   6.72    3.36   1.583  0.24259    
## ABS:Tread    1   3.26    3.26   1.537  0.23691    
## Tire:Tread   2   3.42    1.71   0.805  0.46829    
## Residuals   13  27.60    2.12                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We continue on by dropping the two way interactions:

```
fm1 <- update(fm, .~ABS+Tire+Tread)
summary(fm1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## ABS          1  81.95   81.95  35.972 1.13e-05 ***
## Tire         2  47.95   23.98  10.524  0.00094 ***
## Tread        1   0.19    0.19   0.083  0.77694    
## Residuals   18  41.01    2.28                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

And we then make sure that the model without the interaction is actually better than that with interactions:

```
anova(fm, fm1)
```

```
## Analysis of Variance Table
## 
## Model 1: Distance ~ ABS + Tire + Tread + ABS:Tire + ABS:Tread + Tire:Tread
## Model 2: Distance ~ ABS + Tire + Tread
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     13 27.603                           
## 2     18 41.005 -5   -13.402 1.2624  0.337
```

As expected, the model with interactions is not significantly better than that without interactions.

The final model is hence the same as the model in the previous example of balanced Anova. However, notice that the sums of squares of the following two models (that we expect to be equal), are different:

```
fm <- aov(Distance ~ ABS + Tire, data = distance1)
summary(fm)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## ABS          1  81.95   81.95   37.80 6.57e-06 ***
## Tire         2  47.95   23.98   11.06 0.000653 ***
## Residuals   19  41.19    2.17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fminv <- aov(Distance ~ Tire + ABS, data = distance1)
summary(fminv)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Tire         2  53.09   26.55   12.24 0.000383 ***
## ABS          1  76.80   76.80   35.42 9.95e-06 ***
## Residuals   19  41.19    2.17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since `aov()` performs Type I SS ANOVA (we will see a wide explanation of *Types of Sum of Squares* in the Appendix) and this example uses data from unbalanced design, the previous 2 models give different results in terms of SS and respective p-values. In fact Type I SS ANOVA depends on the order in which factors are included in the model: `fm` is based on SS(ABS) and SS(Tire|ABS), whereas `fminv` is based on SS(Tire) and SS(ABS|Tire).

In order to avoid this problem, we may use Type II ANOVA: `drop1()` function allows to do this.

```
drop1(object=fm,test="F")
```

```
## Single term deletions
##
## Model:
## Distance ~ ABS + Tire
##        Df Sum of Sq     RSS    AIC F value    Pr(>F)
## <none>               41.194 21.404
## ABS     1    76.802 117.996 43.609  35.424 9.949e-06 ***
## Tire    2    47.950  89.144 35.159  11.058 0.0006532 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
drop1(object=fminv,test="F")
```

```
## Single term deletions
##
## Model:
## Distance ~ Tire + ABS
##        Df Sum of Sq     RSS    AIC F value    Pr(>F)
## <none>               41.194 21.404
## Tire    2    47.950  89.144 35.159  11.058 0.0006532 ***
## ABS     1    76.802 117.996 43.609  35.424 9.949e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

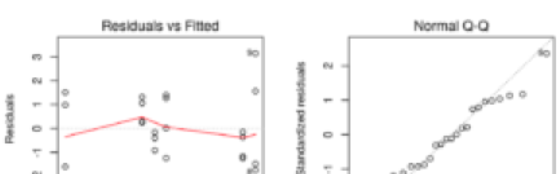
In this case, the results are equal. Alternatively, the function `Anova()` of the package `car` is available. `Anova()` allows Type II and III Sum of Squares too.

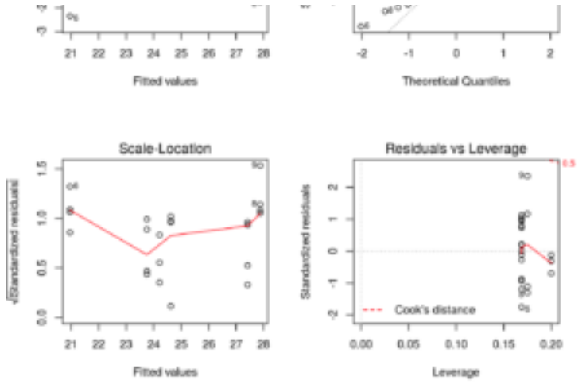
Notice that, until now, at least six types of sum of squares have been introduced in literature. However, there are open discussions among statisticians about the use and pros/cons of different Types of SS.

Residual analysis

Together with the model results, one should always provide some statistics/plots on the residuals.

```
op <- par(mfrow = c(2, 2))
plot(fm)
par(op)
```





Residual plots of unbalanced model

In this case, since the leverages are not constant (unbalanced design) 4th plot draws the leverages in x-axis.

Example: Pin diameters (Fixed effects Nested ANOVA)

Data description

The dataframe considered in this example contains data collected from five different lathes, each of them used by two different operators. The goal of the study is to evaluate if significant differences in the mean diameter of pins occur between lathes and/or operators. Notice that here we are concerned with the effect of operators, so the layout of experiment is nested. If we were concerned with shift instead of operator, the layout would have been the other way around.

Data loading

```
data(diameters)
str(diameters)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    50 obs. of  3 variables:
##  $ Lathe   : Factor w/ 5 levels "1","2","3","4",..: 1 1 2 2 3 3 4 4 5 5 ...
##  $ Operator: Factor w/ 2 levels "D","N": 1 2 1 2 1 2 1 2 1 2 ...
##  $ Pin.Diam: num  0.125 0.124 0.118 0.116 0.123 0.122 0.126 0.126 0.118 0.125 ...
```

Descriptives

Let us first carry out some descriptive statistics and plots in order to get a glimpse of the data. The next few lines of code show descriptive statistics for each variable and the mean for each combination of Lathe and Operator factor levels.

```
summary(diameters)
```

```
##  Lathe  Operator    Pin.Diam
##  1:10   D:25     Min.   :0.114
##  2:10   N:25     1st Qu.:0.122
##  3:10            Median :0.125
##  4:10            Mean   :0.124
##  5:10            3rd Qu.:0.126
##                  Max.   :0.130
```

```
xtabs(formula=Pin.Diam~Lathe+Operator,data=diameters)
```

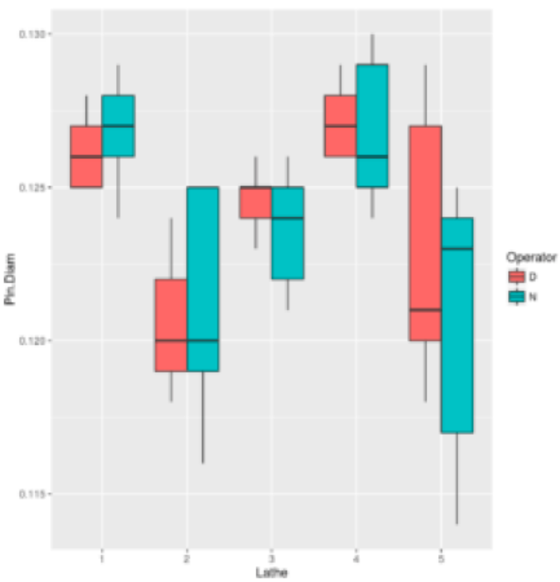
```
##      Operator
## Lathe     D     N
##     1 0.631 0.634
```

```
##     2 0.603 0.605
##     3 0.623 0.618
##     4 0.636 0.634
##     5 0.615 0.603
```

Above statistics are not completely well-advised, since the operators working in the same part of the day (day or night) are different (nested anova) for different lathes. Let us draw a box-plot for each Late x Operator combination.

```
ggp <- ggplot(data = diameters, mapping = aes(x=Lathe, y=Pin.Diam, fill=Operator)) +
  geom_boxplot()

print(ggp)
```



Boxplot of Pin.Diam by Lathe x Operator

It may seem natural to perform the following (incorrect) ANOVA to analyze diameters conditional on Lathe and Shift (i.e., considering Operator levels as equivalent to different shifts of working days) as for classical factorial layout.

```
fm1 <- aov(formula = Pin.Diam~Lathe*Operator, data = diameters)
summary(fm1)

##                Df    Sum Sq   Mean Sq F value   Pr(>F)
## Lathe           4 0.0003033 7.583e-05   8.766 3.52e-05 ***
## Operator        1 0.0000039 3.920e-06   0.453    0.505
## Lathe:Operator  4 0.0000147 3.670e-06   0.424    0.790
## Residuals      40 0.0003460 8.650e-06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above results give however an incorrect model for the data under study. In fact the actual data structure is the following:

```
diameters$Lathe_op <- factor(diameters$Lathe:diameters$Operator)
xtabs(formula=Pin.Diam~Lathe+Lathe_op,data=diameters)

##      Lathe_op
## Lathe   1:D   1:N   2:D   2:N   3:D   3:N   4:D   4:N   5:D   5:N
##     1 0.631 0.634 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
##     2 0.000 0.000 0.603 0.605 0.000 0.000 0.000 0.000 0.000 0.000
##     3 0.000 0.000 0.000 0.000 0.623 0.618 0.000 0.000 0.000 0.000
##     4 0.000 0.000 0.000 0.000 0.000 0.000 0.636 0.634 0.000 0.000
##     5 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.615 0.603
```

The correct ANOVA to perform is thus the following:

```
fm1 <- aov(formula=Pin.Diam~Lathe+Lathe/Operator,data=diameters)
summary(fm1)

##                Df    Sum Sq   Mean Sq F value   Pr(>F)
```

```
## Lathe           4 0.0003033 7.583e-05   8.766 3.52e-05 ***
## Lathe:Operator  5 0.0000186 3.720e-06   0.430    0.825
## Residuals      40 0.0003460 8.650e-06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The / formula operator means that the levels of Operator factor are nested within the levels of Lathe factor. If we read the output we find that lathes seem to produce on average different products, whereas the difference between the means of the pins' diameter conditional on the operator does not seem to be significant. Although the final results of the two Anovas (the first of which is incorrect and the second one correct!) may seem similar nested Anova is the one to use because there is a control over the variability given by the (fixed) effect of the operators. Nested Anova is hence useful for reducing the general variability of the plan and getting more significant differences among the levels of the factors.

An equivalent model formulation for nested ANOVA is given by:

```
fm1a <- aov(formula=Pin.Diam~Lathe+Operator:Lathe,data=diameters)
summary(fm1a)
```

```
##                Df    Sum Sq   Mean Sq F value   Pr(>F)
## Lathe           4 0.0003033 7.583e-05   8.766 3.52e-05 ***
## Lathe:Operator  5 0.0000186 3.720e-06   0.430    0.825
## Residuals      40 0.0003460 8.650e-06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

or

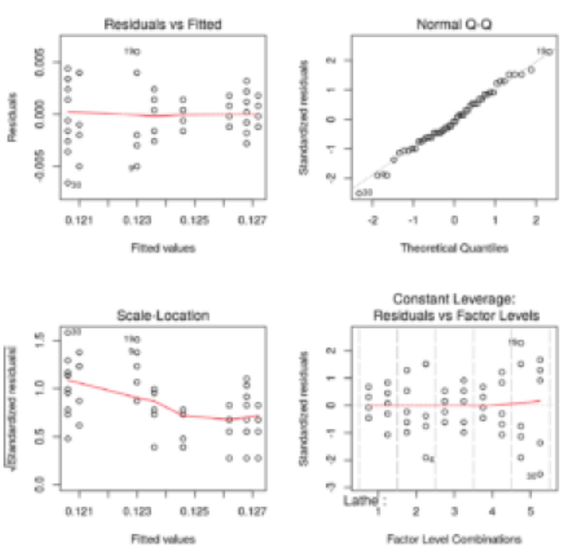
```
#alternative correct model
fm1b <- aov(formula=Pin.Diam~Lathe+Operator:Lathe_op,data=diameters)
summary(fm1b)
```

```
##                   Df    Sum Sq   Mean Sq F value   Pr(>F)
## Lathe              4 0.0003033 7.583e-05   8.766 3.52e-05 ***
## Operator:Lathe_op  5 0.0000186 3.720e-06   0.430    0.825
## Residuals         40 0.0003460 8.650e-06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residuals analysis

Finally, residuals may be plotted for model's diagnostics:

```
op <-  par(mfrow = c(2, 2))
plot(fm1)
par(op)
```



Residual plot of Late by Operator nested model

The post Raccoon | Ch 2.5 – Unbalanced and Nested Anova appeared first on Quantide – R training & consulting.

Recent popular posts

Most visited articles of the week

Sponsors

[Jobs for R users](#)

- [Data Scientist @ Garching bei München, Bayern, Germany](#)
- [Software Developer](#)
- [Senior Quantitative Analyst, Data Scientist](#)
- [R data wrangler](#)
- [Senior Data Scientist](#)
- [Manager, Statistical Consulting & Data Science](#)
- [Financial Controller](#)

Search & Hit Enter

[Full list of contributing R-bloggers](#)