Two Way ANOVA

Practical 3

Pranav Gopalkrishna, 1940223

2021-08-20

# Initial function calls and some definitions

setwd("/Users/pranav/Documents/Study/computerScience/programming/r/data/")  
# install.packages('plyr')  
library(plyr)  
# Function to check if value lies in a certain range  
inRange = function(x, min, max)  
# min is included from range while max is excluded  
{  
 if(x >= min & x < max) {TRUE} else {FALSE}  
}  
  
# Function to return the class corresponding to the value  
# (based on the range and divisions)  
returnClass = function(x, min, max, divisions)  
# min is included from range while max is excluded  
{  
 step = (max - min) / divisions  
 i = 1  
 while(min < max)  
 {  
 if(inRange(x, min, min + step)){break}  
 min = min + step  
 i = i + 1  
 }  
 i  
}

# Data set

The data set I have chosen contains observations on various meteorological measurements taken for five different locations in Australia, collected over a period of around 8 years and 6 months (2008-12-01 to 2017-06-25).

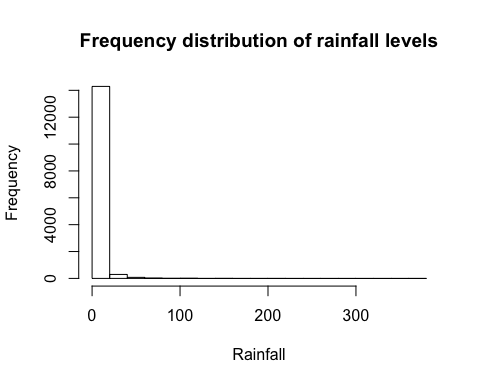
For this assignment, I will focus on three fields:

* Location
* Rainfall (for the day)
* Maximum temperature (for the day)

originalData = read.csv("weatherAustralia.csv")[c(2, 4, 5)]  
summary(originalData)

## Location MaxTemp Rainfall   
## Albury :3040 Min. : 6.80 Min. : 0.000   
## BadgerysCreek:3009 1st Qu.:19.40 1st Qu.: 0.000   
## Cobar :3009 Median :24.40 Median : 0.000   
## CoffsHarbour :3009 Mean :24.69 Mean : 2.386   
## Moree :3009 3rd Qu.:29.40 3rd Qu.: 0.200   
## Max. :47.30 Max. :371.000   
## NA's :62 NA's :342

maxObservations = length(originalData$Location)  
minRainfall = min(originalData$Rainfall, na.rm = TRUE)  
maxRainfall = max(originalData$Rainfall, na.rm = TRUE) + 1  
x1 = x2 = y = c()  
# Getting an idea for the distribution of rainfall values  
hist(originalData$Rainfall, main = "Frequency distribution of rainfall levels", xlab = "Rainfall", ylab = "Frequency")



I will be converting rainfall measurements into a factor type variable, which begins by dividing the range into class intervals, as done by the function "returnClass" in the loop below...

## REMOVING NULL VALUES

Removing null values and obtaining class interval levels for rainfall...

for(i in c(1:maxObservations))  
{  
 location = originalData$Location[i]  
 rainfall = originalData$Rainfall[i]  
 maxTemp = originalData$MaxTemp[i]  
 if(!is.na(location) & !is.na(rainfall) & !is.na(maxTemp))  
 {  
 x1 = c(x1, as.factor(location))  
 x2 = c(x2, returnClass(rainfall, minRainfall, maxRainfall, 20))  
 y = c(y, maxTemp)  
 }  
}  
x1 = as.factor(x1)  
x2 = as.factor(x2)  
myData = data.frame(x1, x2, y)

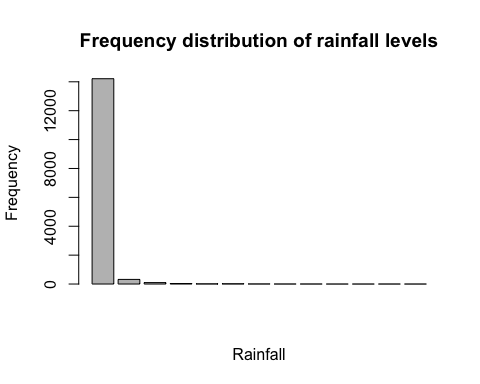
# Checking new data set  
head(myData)

## x1 x2 y  
## 1 1 1 22.9  
## 2 1 1 25.1  
## 3 1 1 25.7  
## 4 1 1 28.0  
## 5 1 1 32.3  
## 6 1 1 29.7

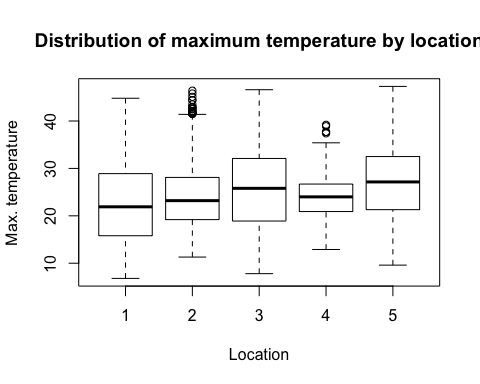
# Visually verifying the distribution of various fields

## Visually verifying the distribution of rainfall levels

Confirming whether there the levels created for rainfall are appropriate enough that the frequency distribution of rainfall matches the previously created histogram for the same.  
barplot(count(x2)$freq, main = "Frequency distribution of rainfall levels", xlab = "Rainfall", ylab = "Frequency")

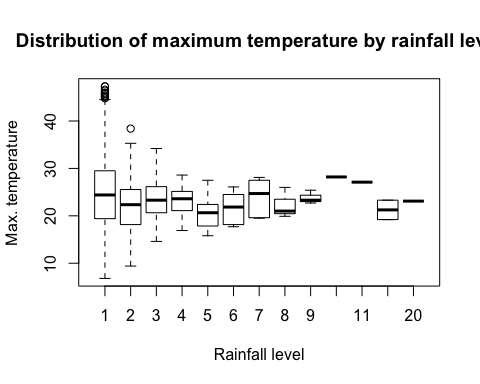


## Visually comparing maxTemp distribution w.r.t. location boxplot(y~x1, main = "Distribution of maximum temperature by location", xlab = "Location", ylab = "Max. temperature")



## Visually comparing maxTemp distribution w.r.t. rainfall level

boxplot(y~x2, main = "Distribution of maximum temperature by rainfall level", xlab = "Rainfall level", ylab = "Max. temperature")



# ANOVA

## Hypotheses

There are three null hypotheses:

1. There is no difference in mean maximum temperatures (y) between different locations (x1)
2. There is no difference in mean maximum temperatures (y) between different rainfall levels (x2)
3. There is no difference in mean maximum temperatures (y) between different interactions of location and rainfall levels (x2 and x1 together)

We will attempt to prove or disprove all of them individually.

## Using multiple linear regression model between the factors and response to perform ANOVA on the resulting estimated model...

Multiple linear regression model for two regressors...  
model = lm(y~x1 + x2)  
summary(model)

##   
## Call:  
## lm(formula = y ~ x1 + x2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.0769 -5.0935 -0.1716 4.6068 22.3065   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 22.6932 0.1220 186.079 < 2e-16 \*\*\*  
## x12 1.4003 0.1733 8.078 7.08e-16 \*\*\*  
## x13 3.1836 0.1724 18.469 < 2e-16 \*\*\*  
## x14 1.3609 0.1738 7.829 5.26e-15 \*\*\*  
## x15 4.2730 0.1744 24.504 < 2e-16 \*\*\*  
## x22 -2.5554 0.3776 -6.768 1.35e-11 \*\*\*  
## x23 -1.2390 0.6735 -1.840 0.0658 .   
## x24 -0.9725 1.1316 -0.859 0.3901   
## x25 -4.1887 1.9271 -2.174 0.0298 \*   
## x26 -2.1139 1.6702 -1.266 0.2056   
## x27 -0.5294 2.7243 -0.194 0.8459   
## x28 -1.7541 3.8527 -0.455 0.6489   
## x29 -0.2541 3.8527 -0.066 0.9474   
## x210 4.1459 6.6707 0.622 0.5343   
## x211 3.0459 6.6707 0.457 0.6480   
## x212 -2.8041 4.7177 -0.594 0.5523   
## x220 -0.9541 6.6707 -0.143 0.8863   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.67 on 14695 degrees of freedom  
## Multiple R-squared: 0.05248, Adjusted R-squared: 0.05145   
## F-statistic: 50.87 on 16 and 14695 DF, p-value: < 2.2e-16

Here, we see the estimates for the intercept and coefficients of the interactions of x1 and x2.

## Performing ANOVA on previously obtained multiple regression model

anova(model)

## Analysis of Variance Table  
##   
## Response: y  
## Df Sum Sq Mean Sq F value Pr(>F)   
## x1 4 33688 8422.0 189.3311 < 2.2e-16 \*\*\*  
## x2 12 2516 209.7 4.7134 9.846e-08 \*\*\*  
## Residuals 14695 653673 44.5   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Here we can see that both the factors are significant i.e. the mean rainfall for , since their p-values are each less than the significance level 0.05. Hence, we can reject null hypotheses 1 and 2. Hence, we perform post hoc analysis (and confirm or reject the 3rd null hypothesis i.e. to estimate the effect of interaction and conclude if it is significant or not).

# POST HOC ANALYSIS

Here, we will try to find what pairs of factor levels show significantly different mean maximum temperature values.

# install.packages("lsmeans")  
library(lsmeans)

## Loading required package: emmeans

## Warning: package 'emmeans' was built under R version 3.6.2

## The 'lsmeans' package is now basically a front end for 'emmeans'.  
## Users are encouraged to switch the rest of the way.  
## See help('transition') for more information, including how to  
## convert old 'lsmeans' objects and scripts to work with 'emmeans'.

The "lsmeans" function from this package computes least-squares means for specified factors or factor combinations in a linear model. Optionally, it makes comparisons or contrasts among them.

**Least square estimates of y averaged over levels of x2**, i.e. the total least square estimates of y for a particular level of x1 is averaged. Here, x2 has 20 levels. So, for each level of x1, the estimates are added and divided by 20.

lse\_over\_x2 = lsmeans(model, "x1")

Same principle as the last one...

**Least square estimates of y averaged over levels of x1**, I.e. the total least square estimates of y for a particular level of x2 is averaged. Here, x1 has 5 levels. So, for each level of x1, the estimates are added and divided by 5.

lse\_over\_x1 = lsmeans(model, "x2")

## Least square estimates of y averaged for all replications of each interaction of x1 and x2. lse\_over\_interactions = lsmeans(model, c("x1", "x2")) FINDING SIGNIFICANTLY DIFFERENT PAIRS

Here, we compare pairs of factor levels with significantly different mean maximum temperature. First, we compare the factor levels of location alone, then factor levels for rainfall levels alone, and then the interaction i.e. the factor levels of location and rainfall levels together.

**Finding the pairs of x1 values (locations) with significantly different means**  
pairs(lse\_over\_x2)

## contrast estimate SE df t.ratio p.value  
## 1 - 2 -1.4003 0.173 14695 -8.078 <.0001  
## 1 - 3 -3.1836 0.172 14695 -18.469 <.0001  
## 1 - 4 -1.3609 0.174 14695 -7.829 <.0001  
## 1 - 5 -4.2730 0.174 14695 -24.504 <.0001  
## 2 - 3 -1.7833 0.174 14695 -10.271 <.0001  
## 2 - 4 0.0394 0.175 14695 0.225 0.9994  
## 2 - 5 -2.8727 0.176 14695 -16.364 <.0001  
## 3 - 4 1.8228 0.174 14695 10.459 <.0001  
## 3 - 5 -1.0894 0.175 14695 -6.238 <.0001  
## 4 - 5 -2.9121 0.176 14695 -16.547 <.0001  
##   
## Results are averaged over the levels of: x2   
## P value adjustment: tukey method for comparing a family of 5 estimates

Based on the p-values, every pair of locations have significantly different mean maximum temperatures except for 2-4 (i.e. BadgerysCreek and CoffsHarbour).

**Finding the pairs of x2 values (rainfall) with significantly different means**  
pairs(lse\_over\_x1)

## contrast estimate SE df t.ratio p.value  
## 1 - 2 2.5554 0.378 14695 6.768 <.0001  
## 1 - 3 1.2390 0.674 14695 1.840 0.8297  
## 1 - 4 0.9725 1.132 14695 0.859 0.9997  
## 1 - 5 4.1887 1.927 14695 2.174 0.6107  
## 1 - 6 2.1139 1.670 14695 1.266 0.9890  
## 1 - 7 0.5294 2.724 14695 0.194 1.0000  
## 1 - 8 1.7541 3.853 14695 0.455 1.0000  
## 1 - 9 0.2541 3.853 14695 0.066 1.0000  
## 1 - 10 -4.1459 6.671 14695 -0.622 1.0000  
## 1 - 11 -3.0459 6.671 14695 -0.457 1.0000  
## 1 - 12 2.8041 4.718 14695 0.594 1.0000  
## 1 - 20 0.9541 6.671 14695 0.143 1.0000  
## 2 - 3 -1.3164 0.767 14695 -1.716 0.8887  
## 2 - 4 -1.5829 1.189 14695 -1.331 0.9831  
## 2 - 5 1.6333 1.962 14695 0.833 0.9998  
## 2 - 6 -0.4415 1.710 14695 -0.258 1.0000  
## 2 - 7 -2.0261 2.749 14695 -0.737 0.9999  
## 2 - 8 -0.8013 3.870 14695 -0.207 1.0000  
## 2 - 9 -2.3013 3.870 14695 -0.595 1.0000  
## 2 - 10 -6.7013 6.681 14695 -1.003 0.9987  
## 2 - 11 -5.6013 6.681 14695 -0.838 0.9998  
## 2 - 12 0.2487 4.732 14695 0.053 1.0000  
## 2 - 20 -1.6013 6.681 14695 -0.240 1.0000  
## 3 - 4 -0.2666 1.313 14695 -0.203 1.0000  
## 3 - 5 2.9497 2.039 14695 1.447 0.9670  
## 3 - 6 0.8749 1.798 14695 0.487 1.0000  
## 3 - 7 -0.7097 2.804 14695 -0.253 1.0000  
## 3 - 8 0.5151 3.909 14695 0.132 1.0000  
## 3 - 9 -0.9849 3.909 14695 -0.252 1.0000  
## 3 - 10 -5.3849 6.704 14695 -0.803 0.9999  
## 3 - 11 -4.2849 6.704 14695 -0.639 1.0000  
## 3 - 12 1.5651 4.764 14695 0.329 1.0000  
## 3 - 20 -0.2849 6.704 14695 -0.043 1.0000  
## 4 - 5 3.2162 2.231 14695 1.441 0.9679  
## 4 - 6 1.1414 2.013 14695 0.567 1.0000  
## 4 - 7 -0.4431 2.947 14695 -0.150 1.0000  
## 4 - 8 0.7816 4.012 14695 0.195 1.0000  
## 4 - 9 -0.7184 4.012 14695 -0.179 1.0000  
## 4 - 10 -5.1184 6.764 14695 -0.757 0.9999  
## 4 - 11 -4.0184 6.764 14695 -0.594 1.0000  
## 4 - 12 1.8316 4.849 14695 0.378 1.0000  
## 4 - 20 -0.0184 6.764 14695 -0.003 1.0000  
## 5 - 6 -2.0748 2.547 14695 -0.815 0.9998  
## 5 - 7 -3.6593 3.335 14695 -1.097 0.9970  
## 5 - 8 -2.4346 4.306 14695 -0.565 1.0000  
## 5 - 9 -3.9346 4.306 14695 -0.914 0.9995  
## 5 - 10 -8.3346 6.942 14695 -1.201 0.9931  
## 5 - 11 -7.2346 6.942 14695 -1.042 0.9982  
## 5 - 12 -1.3846 5.094 14695 -0.272 1.0000  
## 5 - 20 -3.2346 6.942 14695 -0.466 1.0000  
## 6 - 7 -1.5846 3.193 14695 -0.496 1.0000  
## 6 - 8 -0.3598 4.196 14695 -0.086 1.0000  
## 6 - 9 -1.8598 4.196 14695 -0.443 1.0000  
## 6 - 10 -6.2598 6.875 14695 -0.911 0.9995  
## 6 - 11 -5.1598 6.875 14695 -0.751 0.9999  
## 6 - 12 0.6902 5.002 14695 0.138 1.0000  
## 6 - 20 -1.1598 6.875 14695 -0.169 1.0000  
## 7 - 8 1.2247 4.716 14695 0.260 1.0000  
## 7 - 9 -0.2753 4.716 14695 -0.058 1.0000  
## 7 - 10 -4.6753 7.204 14695 -0.649 1.0000  
## 7 - 11 -3.5753 7.204 14695 -0.496 1.0000  
## 7 - 12 2.2747 5.446 14695 0.418 1.0000  
## 7 - 20 0.4247 7.204 14695 0.059 1.0000  
## 8 - 9 -1.5000 5.446 14695 -0.275 1.0000  
## 8 - 10 -5.9000 7.701 14695 -0.766 0.9999  
## 8 - 11 -4.8000 7.701 14695 -0.623 1.0000  
## 8 - 12 1.0500 6.088 14695 0.172 1.0000  
## 8 - 20 -0.8000 7.701 14695 -0.104 1.0000  
## 9 - 10 -4.4000 7.701 14695 -0.571 1.0000  
## 9 - 11 -3.3000 7.701 14695 -0.428 1.0000  
## 9 - 12 2.5500 6.088 14695 0.419 1.0000  
## 9 - 20 0.7000 7.701 14695 0.091 1.0000  
## 10 - 11 1.1000 9.432 14695 0.117 1.0000  
## 10 - 12 6.9500 8.168 14695 0.851 0.9998  
## 10 - 20 5.1000 9.432 14695 0.541 1.0000  
## 11 - 12 5.8500 8.168 14695 0.716 1.0000  
## 11 - 20 4.0000 9.432 14695 0.424 1.0000  
## 12 - 20 -1.8500 8.168 14695 -0.226 1.0000  
##   
## Results are averaged over the levels of: x1   
## P value adjustment: tukey method for comparing a family of 13 estimates

Based on the p-values, only the pair of rainfall levels 1-2 (i.e. class intervals of rainfall measurements [0, 37.1) and [37.1, 74.2)) have significantly different mean maximum temperatures.

**Finding the pairs of interactions with significantly different means**  
pairs(lse\_over\_interactions)

## contrast estimate SE df t.ratio p.value  
## 1 1 - 2 1 -1.40029 0.173 14695 -8.078 <.0001  
## 1 1 - 3 1 -3.18362 0.172 14695 -18.469 <.0001  
## 1 1 - 4 1 -1.36086 0.174 14695 -7.829 <.0001  
## 1 1 - 5 1 -4.27300 0.174 14695 -24.504 <.0001  
## 1 1 - 1 2 2.55544 0.378 14695 6.768 <.0001  
## 1 1 - 2 2 1.15515 0.413 14695 2.795 0.8910  
## 1 1 - 3 2 -0.62819 0.417 14695 -1.506 1.0000  
## 1 1 - 4 2 1.19457 0.410 14695 2.916 0.8183  
## 1 1 - 5 2 -1.71756 0.416 14695 -4.126 0.0477  
## 1 1 - 1 3 1.23904 0.674 14695 1.840 1.0000  
## 1 1 - 2 3 -0.16124 0.694 14695 -0.232 1.0000  
## 1 1 - 3 3 -1.94458 0.696 14695 -2.792 0.8921  
## 1 1 - 4 3 -0.12182 0.689 14695 -0.177 1.0000  
## 1 1 - 5 3 -3.03395 0.694 14695 -4.372 0.0184  
## 1 1 - 1 4 0.97249 1.132 14695 0.859 1.0000  
## 1 1 - 2 4 -0.42780 1.145 14695 -0.374 1.0000  
## 1 1 - 3 4 -2.21113 1.145 14695 -1.930 1.0000  
## 1 1 - 4 4 -0.38837 1.135 14695 -0.342 1.0000  
## 1 1 - 5 4 -3.30051 1.145 14695 -2.882 0.8408  
## 1 1 - 1 5 4.18871 1.927 14695 2.174 0.9992  
## P value adjustment: tukey method for comparing a family of 65 estimates

Many records have been omitted due to the sheer number of records. However, we see that at least some interactions have a p-value less than 0.05, meaning at least some interactions have significantly different mean maximum temperatures, meaning that we can reject the 3rd null hypothesis as well.

# Conclusion

From the above data, we see that we may reject all the null hypotheses for a 0.05 significant level. In other words, each factor level (locations and rainfall levels)as well as their interactions has a significantly different mean effect on the response i.e. maximum daily temperatures.

**ALTERNATE APPROACH TO TWO-WAY ANOVA (UNUSED)**  
library(multcompView)  
model = aov(y~x1 + x2 + x1:x2, myData)  
print(model)

A balanced design is an experimental design where all cells (i.e. treatment combinations) have the same number of observations. This does not seem to be the case for us.

model = aov(y~x1 + x2, myData)  
Since interaction effect is statistically insignificant, we exclude it.  
tukey = TukeyHSD(x = model, conf.level = 0.95)  
tukey