Bernat Chiva

Ex. 14

```r
library(car)

## Loading required package: carData

library(HH)

## Loading required package: lattice

## Loading required package: grid

## Loading required package: latticeExtra

## Loading required package: multcomp

## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Loading required package: MASS

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser

## Loading required package: gridExtra

##
## Attaching package: 'HH'
```

```
## The following objects are masked from 'package:car':
##
##     logit, vif
```

```r
library(tables)
library(RcmdrMisc)
```

```
## Loading required package: sandwich
```

```r
library(doBy)
library(emmeans)
```

```
##
## Attaching package: 'emmeans'
```

```
## The following object is masked from 'package:HH':
##
##     as.glht
```

```r
library(readr)
comrect <- read_delim("C:/Users/berna/OneDrive/Desktop/UPC/S1/5. Models
Lineals/datasets/comrect.csv",
    ";", escape_double = FALSE, locale = locale(decimal_mark = ","),
    trim_ws = TRUE)
```

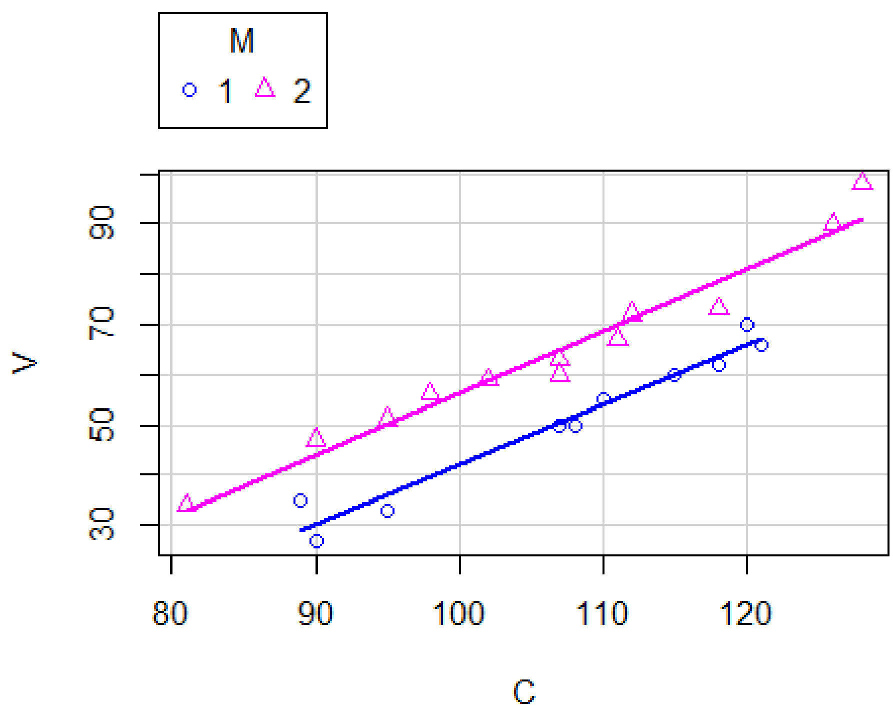
```
##
## -- Column specification --------------------------------------------------
------
## cols(
##   M = col_double(),
##   C = col_double(),
##   V = col_double(),
##   VV = col_double()
## )
```

We change the category of the factor as a factor and perform some simple descriptive statistics:

```
comrect$M<-as.factor(comrect$M)
summary(comrect)

##   M              C                    V                  VV
##   1:10   Min.    : 81.00   Min.    :27.00   Min.    :  7.00
##   2:12   1st Qu.: 95.75   1st Qu.:50.00   1st Qu.: 36.50
##          Median :107.50   Median :59.50   Median : 52.50
##          Mean    :106.73   Mean    :58.09   Mean    : 52.73
##          3rd Qu.:117.25   3rd Qu.:66.75   3rd Qu.: 65.00
##          Max.    :128.00   Max.    :98.00   Max.    :100.00


scatterplot(V~C|M,smooth=F,dat=comrect)
```



By plotting the observations by the level of the factors it seems it doesn't exist interaction between the factor and the independent variable, as the lines are parallel. To be sure we would fit a model with interaction and check its significance:

1. Model with interaction for variable V:

```
mod1<-lm(V~C*M, comrect)

summary(mod1)

##
## Call:
## lm(formula = V ~ C * M, data = comrect)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6767 -1.9789  0.0376  1.3367  6.9744
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -77.49366   10.33232  -7.500 6.07e-07 ***
## C             1.19565    0.09575  12.487 2.65e-10 ***
## M2           10.45276   13.05628   0.801    0.434
## C:M2          0.03924    0.12133   0.323    0.750
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.474 on 18 degrees of freedom
## Multiple R-squared:  0.966,  Adjusted R-squared:  0.9603
## F-statistic: 170.4 on 3 and 18 DF,  p-value: 2.112e-13
```

The interaction coefficient does not show significance, meaning, as we predicted, that it does not exists any interaction between the factor and the independent variable.

2. Model without interaction for variable V:

```
#model without interaction
mod2<-lm(V~C+M, comrect)
summary(mod2)

##
## Call:
## lm(formula = V ~ C + M, data = comrect)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.503 -2.025 -0.088  1.529  7.296
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -80.11563    6.25264  -12.81 8.50e-11 ***
## C             1.22009    0.05741   21.25 1.05e-14 ***
## M2           14.64776    1.45307   10.08 4.62e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.391 on 19 degrees of freedom
## Multiple R-squared:  0.9658, Adjusted R-squared:  0.9622
## F-statistic: 268.2 on 2 and 19 DF,  p-value: 1.186e-14

anova(mod2)

## Analysis of Variance Table
##
## Response: V
##           Df Sum Sq Mean Sq F value    Pr(>F)
## C          1 4999.1  4999.1  434.82 1.489e-14 ***
## M          1 1168.3  1168.3  101.62 4.624e-09 ***
## Residuals 19  218.4    11.5
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Check intercept when C=0, at the origin
emm<-emmeans(mod2,~M|C,at=list(C=c(0)))
print(pairs(emm))
```

```
## C = 0:
##  contrast estimate   SE df t.ratio p.value
##  1 - 2       -14.6 1.45 19 -10.081 <.0001
```

a) Can we consider that the two lines are not statistically different?

No, the lines are different. We obtain a significant effect from M2, the second level of the factor, compared to the baseline M1, first level of the factor. This means that both levels are different from each other. The methodology that you can chose are different from each other if taking into account the results of the tests as response

b) Can the two lines be considered parallel lines?

The slope from the lines of the two methodologies is exactly the same. We can check that by checking the marginal means at three different point between methodologies. We find that the difference is always the same, 14.6, that is the value of the coefficient of the second methodology.

c) Can the two intercepts be considered not statistically different?

We have checked the marginal means using Tuckey method at coefficient of intelligence at 0 value, and the outcome is a significance difference of 14.6. This means that the intercepts are different for each methodology.

```
(emmt<-emmeans(mod2,~M|C,at=list(C=c(90, 105, 120))))
```

```
## C =  90:
##  M emmean    SE df lower.CL upper.CL
##  1   29.7 1.462 19     26.6     32.8
##  2   44.3 1.352 19     41.5     47.2
##
## C = 105:
##  M emmean    SE df lower.CL upper.CL
```

```
##  1   48.0 1.080 19     45.7     50.3
##  2   62.6 0.981 19     60.6     64.7
##
## C = 120:
##  M emmean    SE df lower.CL upper.CL
##  1   66.3 1.297 19     63.6     69.0
##  2   80.9 1.257 19     78.3     83.6
##
## Confidence level used: 0.95
```

```
print(pairs(emmt))
```

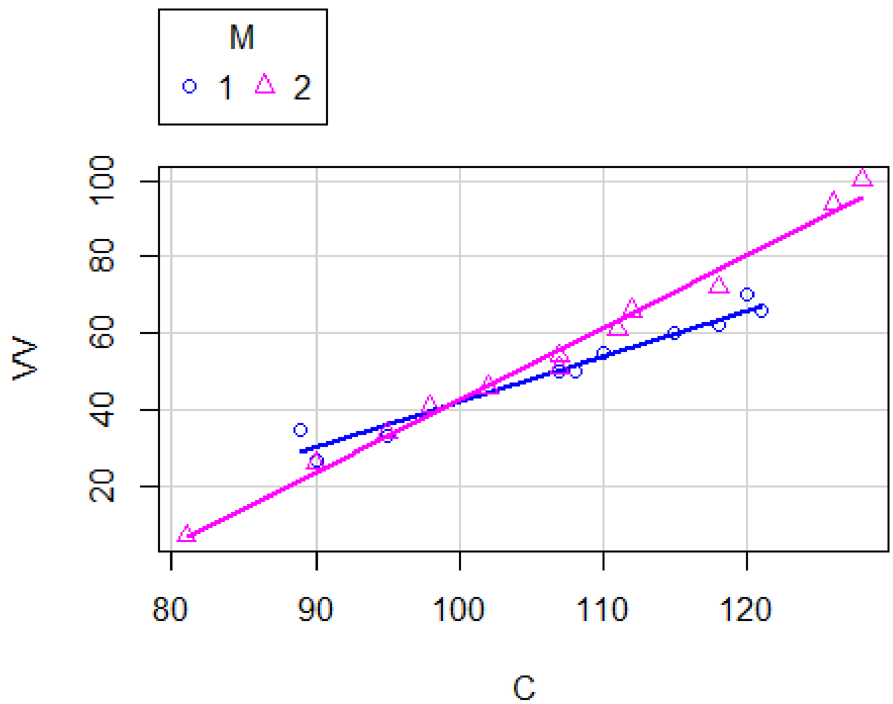
```
## C =  90:
##  contrast estimate   SE df t.ratio p.value
##  1 - 2       -14.6 1.45 19 -10.081 <.0001
##
## C = 105:
##  contrast estimate   SE df t.ratio p.value
##  1 - 2       -14.6 1.45 19 -10.081 <.0001
##
## C = 120:
##  contrast estimate   SE df t.ratio p.value
##  1 - 2       -14.6 1.45 19 -10.081 <.0001
```

d) For each one of the following values of the coefficient of intelligence c: 90, 105 y 120, which differences do it exists in the punctuation for each one of the methodologies?

The difference, as mentioned, is always 14.6 units form the first methodology and the second. This makes sense because the lines are parallel, the only difference is the intercept. For example, when the coefficient of intelligence is 90, the expects test punctuation using model 1 is 29.7, whereas the expected value using model 2 is 44.3, exactly 14.6 points higher.

<u>Conclusions:</u> Having fitted the model, we do not see any interaction between coefficient of intelligence and pedagogical methodology. Both variables have a significant impact on the punctuations of a tests. Furthermore, we can assume that the second methodology is better than the first, and although the relation between each methodology and coefficient of intelligence is the same, both regression have the same slope, the second methodology has a higher intercept, of about 14.6 points in the test higher, so we should expect better results by sticking to the second methodology.

```
#Descriptive statistics
scatterplot(VV~C|M,smooth=F,dat=comrect)
```



3. Model with interaction for variable VV:

We do not see a parallelism between the regression lines, so we will assume different slopes and interaction, we will fit a model with interaction:

```
mod3<-lm(VV~C*M, comrect)
summary(mod3)
```

```
##
## Call:
## lm(formula = VV ~ C * M, data = comrect)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7524 -1.7235 -0.1493  1.9500  6.0805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -77.49366    9.15403  -8.466 1.09e-07 ***
## C             1.19565    0.08483  14.095 3.64e-11 ***
## M2          -69.20305   11.56736  -5.983 1.17e-05 ***
## C:M2          0.69639    0.10750   6.478 4.30e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.077 on 18 degrees of freedom
## Multiple R-squared:  0.9828, Adjusted R-squared:  0.9799
## F-statistic: 342.3 on 3 and 18 DF,  p-value: 4.669e-16
```

```
anova(mod3)
```

```
## Analysis of Variance Table
##
## Response: VV
##           Df Sum Sq Mean Sq F value    Pr(>F)
## C          1 9178.7  9178.7 969.177 < 2.2e-16 ***
```

```
## M           1  149.7   149.7  15.812 0.0008852 ***
## C:M         1  397.5   397.5  41.967 4.296e-06 ***
## Residuals 18  170.5     9.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a) Can we consider that the two lines are not statistically different?

No, the lines are different. We obtain a significant effect from M2, the second level of the factor, compared to the baseline M1, first level of the factor. This means that both levels are different from each other. The methodology that you can chose have a different effect on the test punctuation.

b) Can the two lines be considered parallel lines?

The slope from the lines of the two methodologies is different. We can check that by checking the marginal means at three different point between methodologies. We find that the difference is always different and significant. There is a point between when the coefficient of intelligence has a value of 95 and a value of 100, that the line of the second methodology cuts the line of the first and performs better.

c) Can the two intercepts be considered not statistically different?

We have checked the marginal means from both lines when the coefficient of intelligence equal 0, and we have found a significance difference between methodologies. The mean of the first methodology has a difference of 69.2 points of the test compared to the second methodology.

```
emm2<-emmeans(mod3,~M|C,at=list(C=c(0)))
print(pairs(emm2))
```

```
## C = 0:
##  contrast estimate   SE df t.ratio p.value
##  1 - 2        69.2 11.6 18 5.983   <.0001
```

```
(emmt2<-emmeans(mod3,~M|C,at=list(C=c(90, 105, 120))))
```

```
## C =  90:
##  M emmean    SE df lower.CL upper.CL
##  1   30.1 1.761 18     26.4     33.8
##  2   23.6 1.393 18     20.7     26.5
##
```

```
## C = 105:
##  M emmean    SE df lower.CL upper.CL
##  1   48.0 0.993 18     46.0     50.1
##  2   52.0 0.892 18     50.1     53.8
##
## C = 120:
##  M emmean    SE df lower.CL upper.CL
##  1   66.0 1.452 18     62.9     69.0
##  2   80.3 1.270 18     77.7     83.0
##
## Confidence level used: 0.95
```

```
print(pairs(emmt2))
```

```
## C =  90:
##  contrast estimate   SE df t.ratio p.value
##  1 - 2        6.53 2.25 18  2.907  0.0094
##
## C = 105:
##  contrast estimate   SE df t.ratio p.value
##  1 - 2       -3.92 1.33 18 -2.936  0.0088
##
## C = 120:
##  contrast estimate   SE df t.ratio p.value
##  1 - 2      -14.36 1.93 18 -7.446  <.0001
```

d) For each one of the following values of the coefficient of intelligence c: 90, 105 y 120, which differences do it exists in the punctuation for each one of the methodologies?

We can see, given that the slope is different between regression lines, that the marginal means of these values are different. When the coefficient of intelligence is 90, the first methodology has a expected test score of 30.1, whereas the second methodology has an expected test punctuation of 23.6, 6.53 point lower. But if we check when the coefficient of intelligence is 105 and 120, students that have followed the second methodology perform better, with expected results 3.92 and 14.36 points higher.

Conclusions: Taking into account the observations of the variable VV, the model performs different than the first model of variable V, as in this one exists interaction. In practice this means that the variable coefficient of intelligence has an impact on the factor. The slopes of the regression lines of the two methodologies are different, meaning that they interact differently with the continuous variable.

In this case, the methodology one should pick depends on his/her coefficient of intelligence. This change of the better methodology depends on the value of the coefficient of intelligence and it happens between 95 and 100.