BACHELOR DEGREE IN INFORMATICS (UPC). COURSE 18-19 Q2 – QUIZ1

Anàlisi de Dades i Explotació de la Informació (ADEI).

(Date: 30th/May/2019 14:00   Place: Room A5S108)

STUDENT NAME:	
DNI/PASSPORT:	

Lecturer: Lídia Montero Mercadé
Office: Edifici C5 D207
Norms: Calculator, statistical tables and R Studio reference documents are allowed. Internet access, emailing and chatting is strictly forbidden. Mobile phones should be switched off.
Quiz duration: 1h 00 min
Date for posting marks: Before June, 4th, 2019, to be posted at Subject's ATENEA WEB page.
Open-office: June, 4th, 2019 at 11:00 (C5-207).

Problem 1: All qüestions account for 1 point

MASS package in R contains insurance dataframe. It consists on claims from a British car insurance company in 1973:

`District`	district of policyholder (1 to 4): 4 is major cities (London).
`Group`	group of car (1 to 4), <1 litre, 1–1.5 litre, 1.5–2 litre, >2 litre.
`Age`	of driver in 4 ordered groups, <25, 25–29, 30–35, >35.
`Holders`	numbers of policyholders (pòlisses)
`Claims`	numbers of claims (sinistres)

Source: L. A. Baxter, S. M. Coutts and G. A. F. Ross (1980) Applications of linear models in motor insurance. *Proceedings of the 21st International Congress of Actuaries, Zurich* pp. 11–29

Data refer to 23,359 holders accounting for 3,151 sinisters in 1973. Total number of claims (Claims) is the target variable. Data is grouped into 64 classes.

```
> summary(baxter)
 District     Group         Age         Holders            Claims          logHolders
 1:16      <1l    :16   <25   :16   Min.   :   3.00   Min.   :  0.00   Min.   :1.099
 2:16      1-1.5l:16   25-29:16   1st Qu.:  46.75   1st Qu.:  9.50   1st Qu.:3.844
 3:16      1.5-2l:16   30-35:16   Median : 136.00   Median : 22.00   Median :4.912
 4:16      >2l    :16   >35   :16   Mean   : 364.98   Mean   : 49.23   Mean   :4.904
                                    3rd Qu.: 327.50   3rd Qu.: 55.50   3rd Qu.:5.791
                                    Max.   :3582.00   Max.   :400.00   Max.   :8.184
> dim(df)
[1] 64  6
> sd(df$Claims)
[1] 71.1624
```

1. Determine the most promising variables for forecasting purposes of the selected target.

You have to interpret condes() output. Grouped data for claims is included, since each register contains the total number of claims and the total number of holders. Groups are defined by factors District, Group and Age factor variables.

Numerical variables positively correlated to the total number of claims are Holders and logHolders, being more intense Holders correlation.

All factors (Age, Group and District) are meaningful to explain the total number of claims although Age relation intensity is higher than the ones for Group and District.

Age >35 holders have 80 claims over the overall mean and <25 show figures under the overall mean. District 1 holders class has 37 claims over the overall mean. Power Group 1-1.5l show average of claims over gross mean (41) and >2l under the mean (-31).

2. A model for Claims on Holders and Age is discussed. Fill the blank in the summary output from R.

```
> summary(m1)

Call: lm(formula = Claims ~ Holders + Age, data = df)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.463138     ( 1 )    2.112    0.039
Holders     0.110360   0.003243  34.026    ( 2 )
Age25-29    2.674287     ( 3 )    0.619    0.539
Age30-35    1.108561   4.332954   0.256    0.799
Age>35      6.183251   5.367633   1.152    0.254
---

Residual standard error: ( 4 ) on ( 5 ) degrees of freedom
Multiple R-squared:  0.9724,
F-statistic: 520.4 on ( 6 ) and 59 DF,  p-value: < 2.2e-16
```

1 std error = 6.46/2.11	2 Clearly << 0.05 since t value is very large	3 std error = 2.67/0.62
4 $s^2 = \frac{RSS}{n-p} = TSS\left(1-R^2\right)/\left(n-p\right) = 71.1624^2(64-1)*(1-0.9724)/59$	5 n-p=64-5=59	6 p-1=4

```
> summary(m1)

Call:
lm(formula = Claims ~ Holders + Age, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-42.051  -5.629  -2.240   6.377  39.363 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.463138   3.060832   2.112    0.039 *  
Holders     0.110360   0.003243  34.026   <2e-16 ***
Age25-29    2.674287   4.323184   0.619    0.539    
Age30-35    1.108561   4.332954   0.256    0.799    
Age>35      6.183251   5.367633   1.152    0.254    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.21 on 59 degrees of freedom
Multiple R-squared:  0.9724,	Adjusted R-squared:  0.9706 
F-statistic: 520.4 on 4 and 59 DF,  p-value: < 2.2e-16
```

3. Write the prediction equation for model (m2) and predict the expected number of claims for a 50 years old person.

Model Interpretation:

Log(Claims+0.5)=-0.95+0.85log(Holders) for i=1 Age<25

Log(Claims+0.5)=(-0.95-0.071)+0.85log(Holders) for i=2 Age25-29

Log(Claims+0.5)=(-0.95-0.14)+0.85log(Holders) for i=3 Age30-35

Log(Claims+0.5)=(-0.95-0.075)+0.85log(Holders) for i=4 Age>35

Prediction for one person in the 4th group of age: Holders=1 thus log(Holders)=0

Log(Claims+0.5)= (-0.95-0.075)=-1.25 -> Claims ~ exp(-1.25) = 0.2865

Using all decimals, the output from R is replicated:

```
> predict(m2,newdata=data.frame(Holders=1,Age=">35"))
        1
-1.026834
> exp(predict(m2,newdata=data.frame(Holders=1,Age=">35")))
        1
0.3581392
```

4. A new model containing all available factors and the total number of holder is calculated. Determine the significant net-effects of included variables.

Question refers to m3, according to Anova(m3) output only Group factor net-effect is significant once the rest of factors and log(Holders) are included. Log(Holders) covariate, the logarithm of the class size has also a relevant net-effect.

5. Taking a look to residual diagnostics in model (m3), a new model (m4) is proposed. Discuss the meaning of the new modelling for target and comment pros and cons of both modelling options.

Target is Claims, the same in m3 and m4, adding a 0.5 to avoid problems with logarithms. (m3) contains main effects for all available variables (log(Holders), Group, Age and District), but (m4) contains all main effects plus double interactions between factors (Group:Age, Group:District and Age:District) and factor interactions with covariate log(Holders) (Group:Log(Holders), District:Log(Holders) and Age:Log(Holders)).

(m3) is easier and R2 is 0.9589 (it explains 96% of target variability), some non-rellevant variables are present that should be removed.

(m4) is a complex model with R2 = 0.9852 and many interactions are not significant according to a stepwise procedure monitored using BIC and returning m5 model with 3 double interactions. From my point of view complexity is not worth since it provides a tidy explanatory benefit (less than 3%).

6. A new model with interactions between factors and covariate is proposed (m4 and m5). Determine significant net-effect interactions that are worth to retain. Justify your answer.

Model (m5) is obtained by a stepwise procedure monitored using BIC from m4. According to net-effects output provided by Anova(m5) method net-effects for interaction log(Holders):District is clearly worth since pvalue is less than 0.05 threshold. District:Age and District:Group interactions are slightly over the regular 0.05 threshold being District:Group interaction more important. We have to retain all 3 double interactions.

7. Make a rough assessment of the quality of the model based on the first impression of the diagnosis of residuals for (m5).

Clearly, non-normal distributed residuals are present (see QQ plot) and variance is not constant. Some atypical classes are observed (residual outliers, 46 group). It seems to exist influent data according to Cook's distance curves (29 and 61 groups).

8. A new binary factor consisting on grouping District levels into Others and London is defined and a new model m7 is obtained. Justify according to the provided output pros and cons of m5 and m7.

(m7) is obtained after applying a stepwise BIC monitored procedure to the equivalent to model (m4) but considering London binary factor instead of District original factor. It has the value of reducing the number of levels of District to 2 (London and Otherwise) and simplifying the final model, less parameters and less complexity.

According to Fisher's test, that it can be applied since (m7) and (m5) models are nested models, both models are equivalent (pvalue > 0.05, in fact 0.17). BIC value for m7 is 63.99 << 95.28 =BIC(m5).

So, it works! Clearly District should be avoided and London binary factor used to continue with the modelling process. R2 is 0.97 (not included in the output).

9. Influent and atypical data analysis has to be discussed using influencePlot() output. Are there any atypical and/or influent data classes according to the output?

Large residual is shown for 46 group (Student residual -5.1), but a priori influent data according to large leverage is observed for groups 49 and mainly 61 that lays over 3 times the mean leverage in the sample. No boxplot for Cook's distance is included, but a 1.01 value for group 61 is remarkable and it is an influent observation for model (m7).

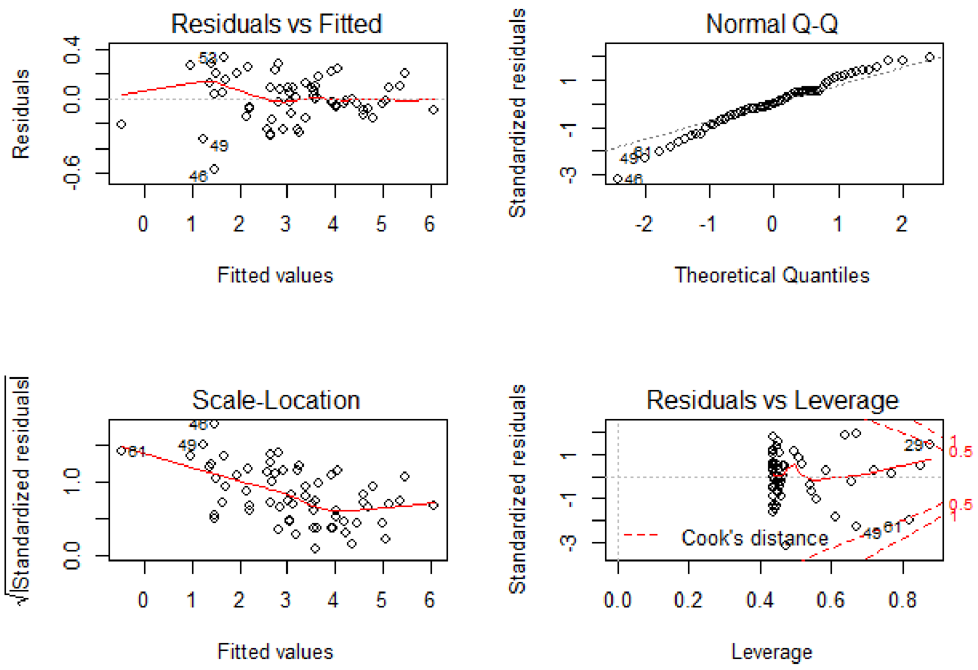
10. Indicate a 95% confidence interval for the expected number of claims according to (m7) for a person in the [30-35] age segment living in London and holding an insurance for a car in the most powerful group.

Output is given and 95% interval for log(Claims) according to model (m7) is -5.438462 -1.236674. To answer to the question you just have to calculate the exponential of both value and the expected number of claims for the given person is between 0.00435 and 0.29 accidents with a 95% confidence interval. Point estimate would be 0.0355 expected number of claims.

RESULTS

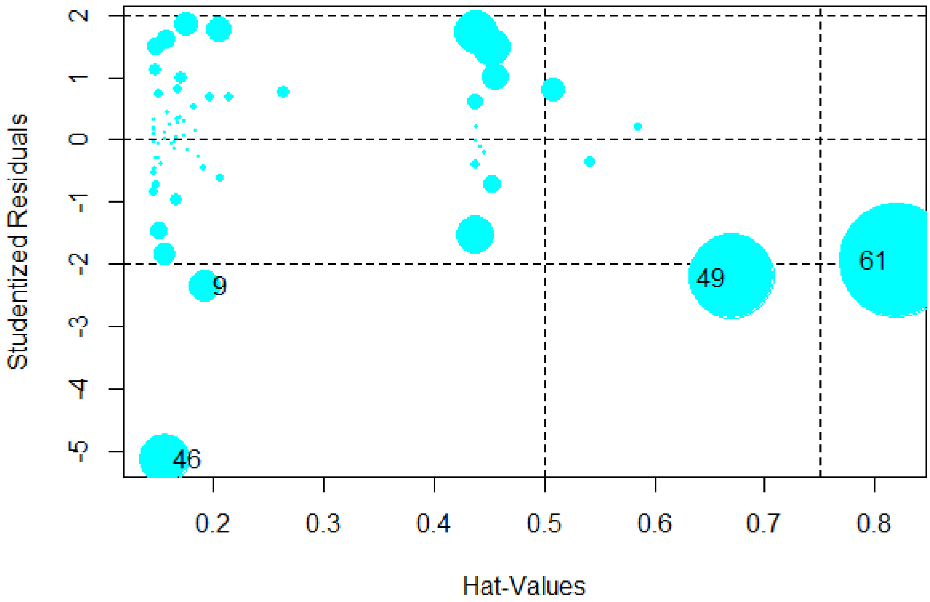
```
> condes(baxter,5)
$quanti
           correlation      p.value
Holders      0.9857701 9.887964e-50
logHolders   0.7741853 6.251833e-14

$quali
                R2      p.value
Age      0.4315501 1.834931e-07
Group    0.1459450 2.287953e-02
District 0.1235997 4.641178e-02

$category
        Estimate      p.value
>35     79.82812 5.058682e-09
1-1.5l  41.39062 6.227092e-03
1       37.07812 1.487515e-02
>2l    -30.54687 4.652338e-02
<25    -34.92187 2.217507e-02

> m2<-lm(log(Claims+0.5)~log(Holders)+Age,data=df)
> summary(m2)

Call: lm(formula = log(Claims + 0.5) ~ log(Holders) + Age, data = df)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.95149    0.17008  -5.594 6.03e-07 ***
log(Holders)  0.85354    0.04111  20.763  < 2e-16 ***
Age25-29     -0.07094    0.12277  -0.578    0.566
Age30-35     -0.13827    0.12720  -1.087    0.281
Age>35       -0.07535    0.16874  -0.447    0.657
---
Residual standard error: 0.3297 on 59 degrees of freedom
Multiple R-squared:  0.9386,   Adjusted R-squared:  0.9345
F-statistic: 225.6 on 4 and 59 DF,  p-value: < 2.2e-16

> m3<-lm(log(Claims+0.5)~log(Holders)+District+Group+Age,data=df)
> summary(m3)

Call:lm(formula = log(Claims + 0.5) ~ log(Holders) + District + Group +
    Age, data = df)
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -2.1369     0.7522  -2.841 0.006369 **
log(Holders)      1.0575     0.1627   6.502 2.86e-08 ***
District2         0.1715     0.1341   1.279 0.206609
District3         0.1564     0.2113   0.740 0.462606
District4         0.3136     0.3099   1.012 0.316208
Group1-1.5l       0.1020     0.1629   0.626 0.534054
Group1.5-2l       0.4176     0.1015   4.112 0.000137 ***
Group>2l          0.6284     0.2187   2.874 0.005827 **
Age25-29         -0.2619     0.1825  -1.435 0.157288
Age30-35         -0.3907     0.2251  -1.736 0.088419 .
Age>35           -0.6806     0.4931  -1.380 0.173290
---

Residual standard error: 0.2848 on 53 degrees of freedom
Multiple R-squared:  0.9589,   Adjusted R-squared:  0.9511
F-statistic: 123.6 on 10 and 53 DF,  p-value: < 2.2e-16

> Anova(m3)
Anova Table (Type II tests)
Response: log(Claims + 0.5)
             Sum Sq Df F value    Pr(>F)
log(Holders) 3.4279  1 42.2703 2.856e-08 ***
District     0.1718  3  0.7064 0.5524821
Group        1.8784  3  7.7210 0.0002269 ***
Age          0.2703  3  1.1110 0.3529063
Residuals    4.2980 53
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> m4<-
lm(log(Claims+0.5)~log(Holders)*(District+Group+Age)+(District+Group+Age)^2,data=df)
> summary(m4)

Call: lm(formula = log(Claims + 0.5) ~ log(Holders) * (District + Group +
    Age) + (District + Group + Age)^2, data = df)

…

Residual standard error: 0.3016 on 17 degrees of freedom
Multiple R-squared:  0.9852,   Adjusted R-squared:  0.9452
F-statistic: 24.61 on 46 and 17 DF,  p-value: 2.288e-09

> m5<-step(m4)
…

Step:  AIC=-159.59
log(Claims + 0.5) ~ log(Holders) + District + Group + Age + log(Holders):District +
    District:Group + District:Age

                        Df Sum of Sq    RSS     AIC
<none>                               1.9451 -159.59
- District:Age           9   1.11592 3.0610 -148.57
- District:Group         9   1.18381 3.1289 -147.17
- log(Holders):District  3   0.86227 2.8073 -142.10

> par(mfrow=c(2,2))
> plot(m5)
> par(mfrow=c(1,1))
>
```



```
> Anova(m5)
Anova Table (Type II tests)

Response: log(Claims + 0.5)
                       Sum Sq Df F value    Pr(>F)    
log(Holders)          2.73039  1 44.9201 1.445e-07 ***
District              0.17184  3  0.9424  0.431684    
Group                 1.98156  3 10.8668 4.433e-05 ***
Age                   0.35523  3  1.9481  0.141696    
log(Holders):District 0.86227  3  4.7287  0.007671 ** 
District:Group        1.18381  9  2.1640  0.052491 .  
District:Age          1.11592  9  2.0399  0.066960 .  
Residuals             1.94506 32                      
> df$London <- factor(ifelse(df$District!=4,0,1),labels=c("Other","London"))
> m6<-lm(log(Claims+0.5)~log(Holders)*(London+Group+Age)+(London+Group+Age)^2,data=df)
> summary(m6)

Call:
lm(formula = log(Claims + 0.5) ~ log(Holders) * (London + Group + 
    Age) + (London + Group + Age)^2, data = df)

…

Residual standard error: 0.2932 on 33 degrees of freedom
Multiple R-squared:  0.9729,	Adjusted R-squared:  0.9482 
F-statistic: 39.43 on 30 and 33 DF,  p-value: < 2.2e-16

> m7<-step(m6)
> anova(m7,m5)
Analysis of Variance Table

Model 1: log(Claims + 0.5) ~ log(Holders) + London + Group + Age + log(Holders):London 
+ London:Group + London:Age
Model 2: log(Claims + 0.5) ~ log(Holders) + District + Group + Age + 
log(Holders):District + District:Group + District:Age

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     48 3.3743                           
2     32 1.9451 16    1.4292 1.4696 0.1725

> BIC(m7,m5)
   df      BIC
m7 17 63.99292
m5 33 95.27760
> 
```

```
> influencePlot(m7,col="cyan",pch=19)
     StudRes       Hat      CookD
9  -2.348364 0.1932308 0.07545664
46 -5.120061 0.1565165 0.19932177
49 -2.207329 0.6708584 0.57433866
61 -1.945311 0.8193086 1.01363233
```



```
> predict(m7, newdata=data.frame(Holders=1,London="London",Group=">2l",Age="30-
35"),interval ="prediction",se.fit=T)
$fit
        fit       lwr       upr
  -3.337568 -5.438462 -1.236674
```