1819Q2Quiz2

Lídia Montero

December, 19th 2018

Contents

1 Third party Insurance Data

Third party insurance is a compulsory insurance for vehi-cle owners in Australia. It insures vehicle owners against injury caused to other drivers, passengers or pedestrians, as a result of an accident. This data set records the number of third party claims in a twelve-month period between 1984 and 1986 in each of 176 geographical areas (local gov-ernment areas) in New South Wales, Australia. Areas are grouped into 13 statistical divisions. Other recorded variables are the number of accidents, the number of people killed or injured and population.

```
## Warning: package 'car' was built under R version 3.4.4

## Loading required package: carData

## Warning: package 'carData' was built under R version 3.4.4

## Warning: package 'FactoMineR' was built under R version 3.4.4

## Warning: package 'factoextra' was built under R version 3.4.4

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.4.4

## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ

## Warning: package 'missMDA' was built under R version 3.4.4

##          lga            sd         claims          accidents
##  Albury (C):  1   SD-1   :38   Min.   :   0.00   Min.   :  17.0
##  Armidale  :  1   SD-9   :26   1st Qu.:  47.75   1st Qu.: 165.2
##  Ashfield  :  1   SD-7   :20   Median : 136.50   Median : 420.5
##  Auburn    :  1   SD-10  :16   Mean   : 586.69   Mean   :1153.1
##  Ballina   :  1   SD-12  :16   3rd Qu.: 553.50   3rd Qu.:1217.2
##  Balranald :  1   SD-11  :14   Max.   :6524.00   Max.   :9416.0
##  (Other)   :170   (Other):46
##        ki           population      pop_density          f.bigcity
```

```
##  Min.   :  11.0   Min.   :   253   Min.   :   0.000   BigC-NO :157
##  1st Qu.: 127.0   1st Qu.:  4390   1st Qu.:   0.975   BigC-YES: 19
##  Median : 288.5   Median : 16739   Median :   6.750
##  Mean   : 650.6   Mean   : 47334   Mean   : 570.932
##  3rd Qu.: 793.0   3rd Qu.: 54092   3rd Qu.: 181.550
##  Max.   :4201.0   Max.   :368045   Max.   :8709.800
##
##      claimp           accidentp            kip                 f.hcla
##  Min.   :   0.000   Min.   :   0.342   Min.   :   0.222   Cluster-1:152
##  1st Qu.:   6.354   1st Qu.:  18.059   1st Qu.:  12.003   Cluster-2: 24
##  Median :  10.513   Median :  28.021   Median :  18.754
##  Mean   :  93.812   Mean   : 211.652   Mean   : 117.690
##  3rd Qu.:  19.554   3rd Qu.:  45.452   3rd Qu.:  34.509
##  Max.   :3724.719   Max.   :6567.416   Max.   :4331.461
##
```

```
##  [1] "lga"         "sd"          "claims"      "accidents"   "ki"
##  [6] "population"  "pop_density" "f.bigcity"   "claimp"      "accidentp"
## [11] "kip"         "f.hcla"
```

```
## [1] 1013.539
```

Consider a model for the number of claims in an area as a function of the number of accidents. A scatterplot of claims against accidents and boxplot of involved variables are shown

1.1 Point 1

Claims is considered the target variable. Determine the most promising variables for forescasting purposes of the selected target.

Output from condes() procedure in FactoMineR library is included. The number of accidents, killed peple, population and population_density are numeric variables directly associated to the target: from more intense to less intense.

Factors, from more to intese to less intense, f.hcla (cluster), statistical division (sd) and f.bigcity are globally associated to the number of claims (target).

Number of claims for Cluster-2 is 1185 units over the mid-point (1449) of the mean number of claims per LGA (263.25 and 2635.13) in each cluster, while Cluster-1 is -1185 units the mid-point. LGA with big-cities have significantly more accidents. Statistical district 1 is 1324 units over the grand mean of claims while SD-12, SD-7 and SD-9 are significantly under the grand mean (465).

```r
# Point 1
dim(df)
```

```
## [1] 176  12
```

```r
names(df)
```

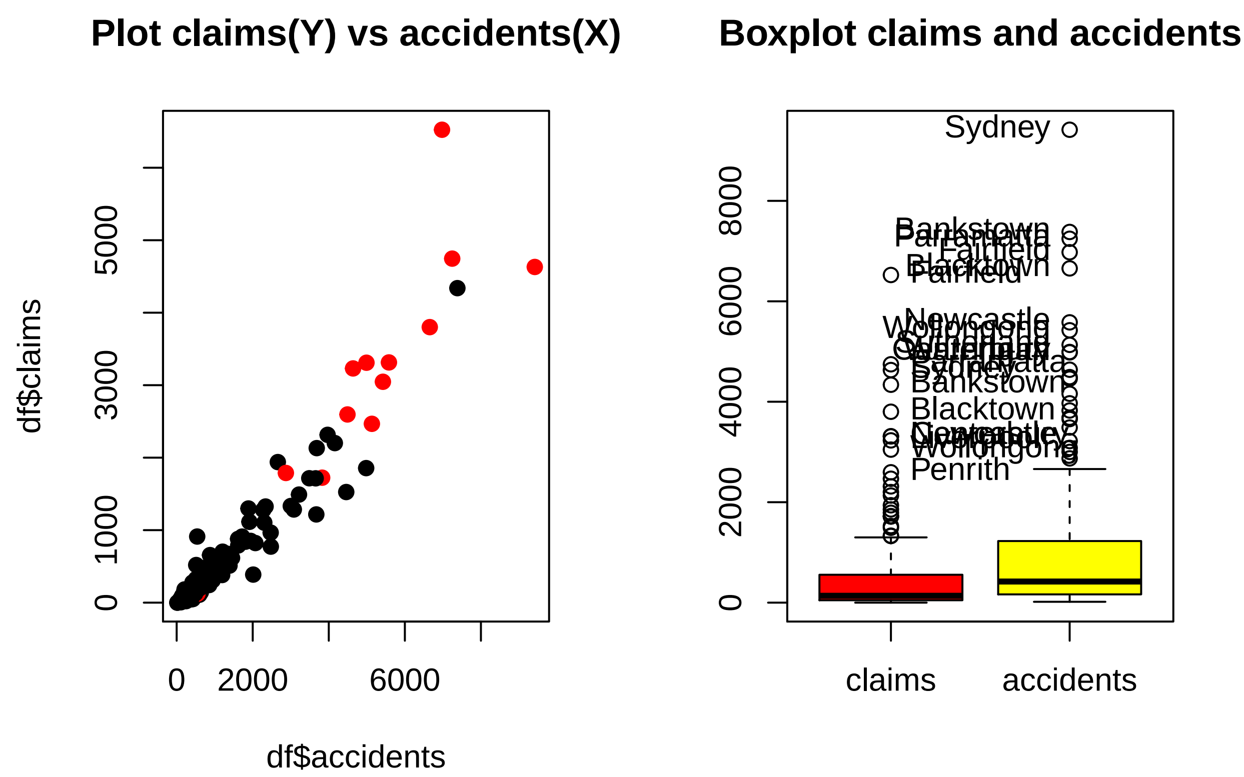
```
##  [1] "lga"         "sd"          "claims"      "accidents"   "ki"
##  [6] "population"  "pop_density" "f.bigcity"   "claimp"      "accidentp"
## [11] "kip"         "f.hcla"
```

```r
sd(df$claims)
```

```
## [1] 1013.539
```

```r
par(mfrow=c(1,2))
plot(df$accidents,df$claims,main="Plot claims(Y) vs accidents(X)",pch=19,col=(as.numeric(df$f.bigcity))
```

```
#legend(locator(n=1),legend=levels(df$f.bigcity), col=c(1,2),pch=19)
Boxplot(df[,3:4],main="Boxplot claims and accidents",col=heat.colors(2))
```




```
##  [1] "Fairfield"  "Parramatta" "Sydney"     "Bankstown"  "Blacktown"
##  [6] "Newcastle"  "Canterbury" "Liverpool"  "Wollongong" "Penrith"
## [11] "Sydney"     "Bankstown"  "Parramatta" "Fairfield"  "Blacktown"
## [16] "Newcastle"  "Wollongong" "Sutherland" "Canterbury" "Warringah"
```

```
par(mfrow=c(1,1))

names(df)
```

```
##  [1] "lga"          "sd"           "claims"       "accidents"    "ki"
##  [6] "population"   "pop_density"  "f.bigcity"    "claimp"       "accidentp"
## [11] "kip"          "f.hcla"
```

```
condes(df[,2:12],2)
```

```
## $quanti
##               correlation      p.value
## accidents       0.9587774 5.230857e-97
## ki              0.9468247 1.296340e-87
## population      0.6493865 1.872238e-22
## pop_density     0.4168224 8.692881e-09
##
## $quali
##                   R2      p.value
## f.hcla     0.6486442 2.257936e-41
```

```
## sd        0.4478551 6.621623e-16
## f.bigcity 0.3028704 2.530375e-15
##
## $category
##             Estimate      p.value
## Cluster-2  1185.9375 2.257936e-41
## SD-1       1324.2382 1.939927e-20
## BigC-YES    896.1644 2.530375e-15
## SD-12      -402.2191 2.991548e-02
## SD-7       -384.0066 1.750090e-02
## SD-9       -324.1950 1.482815e-02
## BigC-NO    -896.1644 2.530375e-15
## Cluster-1 -1185.9375 2.257936e-41
```

```
tapply(df$claims,df$f.hcla,mean)
```

```
## Cluster-1 Cluster-2
##   263.250  2635.125
```

```
mean(as.vector(tapply(df$claims,df$f.hcla,mean)))
```

```
## [1] 1449.188
```

```
mean(as.vector(tapply(df$claims,df$sd,mean)))
```

```
## [1] 465.6566
```

```
mean(df$claims)
```

```
## [1] 586.6875
```

1.2 Point 2

** A simple regression model for claims using the number of accidents is discussed. Fill the blanks.**

```
m1<-lm(claims~accidents,data=df)
summary(m1)
```

```
##
## Call:
## lm(formula = claims ~ accidents, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -968.38  -53.14   29.23   57.45 2576.60
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -78.71249   26.41044   -2.98  0.00329 **
## accidents     0.57705    0.01297   44.51  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 288.8 on 174 degrees of freedom
## Multiple R-squared:  0.9193, Adjusted R-squared:  0.9188
## F-statistic:  1981 on 1 and 174 DF,  p-value: < 2.2e-16
```

1.3 Point 3

3. Check default diagnostic residuals and indicate what is shown in each available plot.

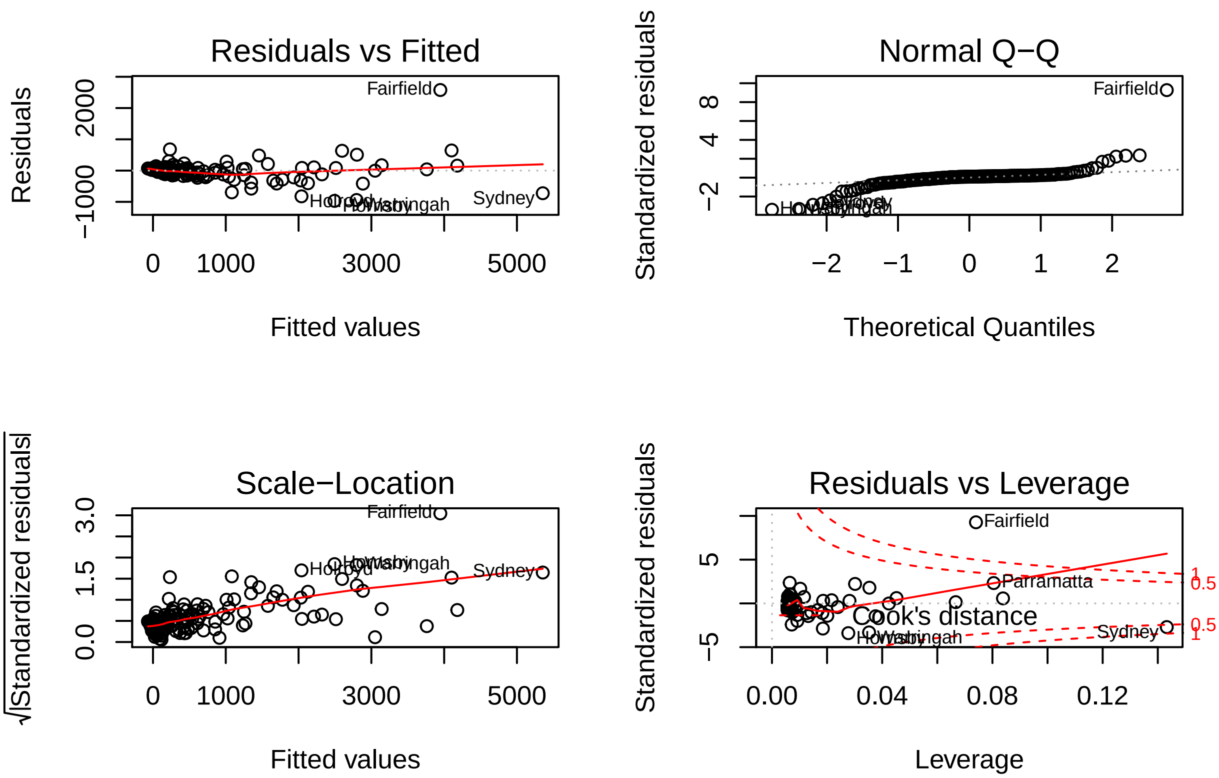
Upper-left plot displays a concentration of points for LGA with small predicted number of accidents. Large residuals can be seen for Fairfield and some other LGAs without identifier. Negative resisuals also appear, Sydney an unsual observation.

Upper-right shows standarized residuals fit to standard normal distribution that clearly fails. Fairfield is a residual outlier, but deviation from the line indicates some other problems in residual distribution.

Lower-left: the model is heterokedastic as indicated by the smoother. A transformation of the target would lead to improve in constant variance.

Lower-right: Sydney is an unusual observation (highest leverage), but the combination of leverage and missfit for Fairfield makes this LGA to become an influent data.

```
par(mfrow=c(2,2))
plot(m1,id.n=5)
```




```
par(mfrow=c(1,1))
```

1.4 Point 4

4. An alternative model (m2) using logarithm transformation for both variables in (m1) is calculated. Determine pros and cons of each model.

Upper-left plot displays a homogeneous prediction range of points for LGA . Large residuals can be seen for Brokenhill and some other LGAs without identifier. Negative resisuals are remarkable for Windouran, Severn, Cohargo and Unincorporated LGA areas.

Upper-right shows standarized residuals fit to standard normal distribution that fails, but it is much better than m1. Not remarkable residual outliers seem to be present despite deviation from the line that indicates problems with normality assumption for residuals.

Lower-left: Smoother is not flat due to observations with extreme outliers, but we are on the good way. Additional explanatory variables are needed.

Lower-right: Influent data can not be discussed according to the output. High leverage observations exist: Cohargo and Windouran.

```r
summary(df)
```

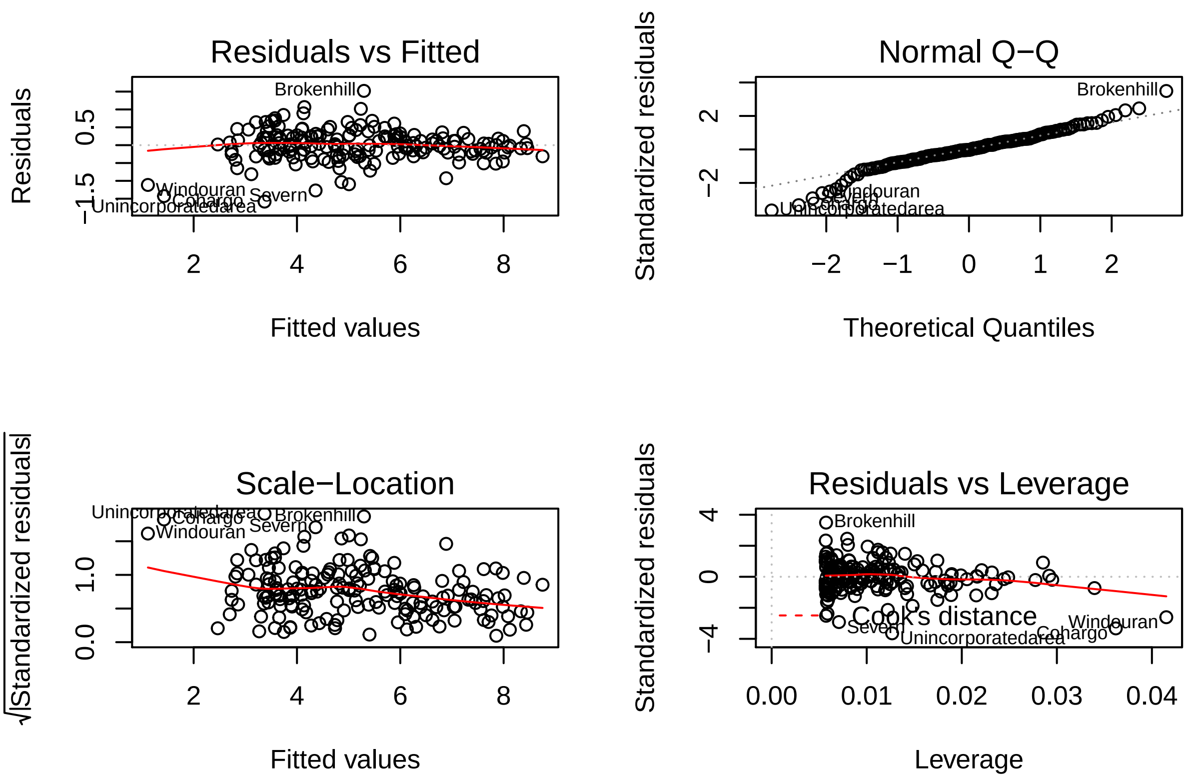
```
##          lga            sd         claims          accidents
##  Albury (C):  1   SD-1   :38   Min.   :   0.00   Min.   :  17.0
##  Armidale  :  1   SD-9   :26   1st Qu.:  47.75   1st Qu.: 165.2
##  Ashfield  :  1   SD-7   :20   Median : 136.50   Median : 420.5
##  Auburn    :  1   SD-10  :16   Mean   : 586.69   Mean   :1153.1
##  Ballina   :  1   SD-12  :16   3rd Qu.: 553.50   3rd Qu.:1217.2
##  Balranald :  1   SD-11  :14   Max.   :6524.00   Max.   :9416.0
##  (Other)   :170   (Other):46
##        ki           population      pop_density         f.bigcity
##  Min.   :  11.0   Min.   :   253   Min.   :   0.000   BigC-NO :157
##  1st Qu.: 127.0   1st Qu.:  4390   1st Qu.:   0.975   BigC-YES: 19
##  Median : 288.5   Median : 16739   Median :   6.750
##  Mean   : 650.6   Mean   : 47334   Mean   : 570.932
##  3rd Qu.: 793.0   3rd Qu.: 54092   3rd Qu.: 181.550
##  Max.   :4201.0   Max.   :368045   Max.   :8709.800
##
##      claimp            accidentp             kip                f.hcla
##  Min.   :   0.000   Min.   :   0.342   Min.   :   0.222   Cluster-1:152
##  1st Qu.:   6.354   1st Qu.:  18.059   1st Qu.:  12.003   Cluster-2: 24
##  Median :  10.513   Median :  28.021   Median :  18.754
##  Mean   :  93.812   Mean   : 211.652   Mean   : 117.690
##  3rd Qu.:  19.554   3rd Qu.:  45.452   3rd Qu.:  34.509
##  Max.   :3724.719   Max.   :6567.416   Max.   :4331.461
##
```

```r
m2<-lm(log(claims+1)~log(accidents),data=df)
summary(m2)
```

```
##
## Call:
## lm(formula = log(claims + 1) ~ log(accidents), data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58077 -0.21919 -0.01057  0.24563  1.52013
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.3117     0.1561  -14.81   <2e-16 ***
## log(accidents)   1.2093     0.0247   48.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.4366 on 174 degrees of freedom
## Multiple R-squared:  0.9323, Adjusted R-squared:  0.9319
## F-statistic:  2396 on 1 and 174 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(m2,id.n=5)
```




```r
par(mfrow=c(1,1))
```

1.5 Point 5

5. Write the prediction equation for model (m2) and predict the number of claims for a fictious LGA with a number of accidents in the mean of New South Wales LGAs.

Prediction equation: log(Y+1)=-2.3+1.21log(X) -> Y=(e^(-2.3))*(X^(1.21))-1

Point prediction: mean(accidents)=1153 Y = (exp(-2.3117))*(1153.097^(1.2093))-1=498.7535 claims

```r
mean(df$accidents)
```

```
## [1] 1153.097
```

```r
summary(m2)
```

```
##
## Call:
## lm(formula = log(claims + 1) ~ log(accidents), data = df)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58077 -0.21919 -0.01057  0.24563  1.52013
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.3117     0.1561  -14.81   <2e-16 ***
## log(accidents)   1.2093     0.0247   48.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4366 on 174 degrees of freedom
## Multiple R-squared:  0.9323, Adjusted R-squared:  0.9319
## F-statistic:  2396 on 1 and 174 DF,  p-value: < 2.2e-16
```

```r
exp(predict(m2,newdata=data.frame(accidents=mean(df$accidents))))-1
```

```
##        1
## 498.7004
```

1.6 Point 6

6. A new (m3) two-way anova model is stated for the logarithm of claims. Interpret the model indicating the predicted number of claims for each defined group. Do you have to consider the interactions in the modeling ?.

Model (m3) explains 40% of number of claims variability, so it is not very good (sixty percent is still left). The interaction term between the 2 factors can not be omitted (pvalue 0.0178 in net-effect test output). Main-effects of factors should be both retained since interactions are significant.

Big-City-No and Cluster-1 : log(Y+1)=4.8 Big-City-No and Cluster-2 : log(Y+1)=4.8+0+2.59 Big-City-Yes and Cluster-1 : log(Y+1)=4.8-1.0827+0 Big-City-Yes and Cluster-2 : log(Y+1)=4.8-1.0827+2.59+1.76

The scale of accidents and predictions in accident scale need to exponentiate the former values in the logarithmic scale.

```r
table(df$f.bigcity,df$f.hcla)
```

```
##
##            Cluster-1 Cluster-2
##   BigC-NO        145        12
##   BigC-YES         7        12
```

```r
#interaction.plot(df$f.bigcity,df$f.hcla,df$claims)
#interaction.plot(df$f.hcla,df$f.bigcity,df$claims)
m3<-lm(log(claims+1)~f.bigcity*f.hcla,data=df)
data.frame(f.bigcity=c(rep(c("BigC-NO"),2),rep(c("BigC-YES"),2)),f.hcla=rep(c("Cluster-1","Cluster-2")))
```

```
##   f.bigcity    f.hcla
## 1   BigC-NO Cluster-1
## 2   BigC-NO Cluster-2
## 3  BigC-YES Cluster-1
## 4  BigC-YES Cluster-2
```

```r
predict(m3,newdata=data.frame(f.bigcity=c(rep(c("BigC-NO"),2),rep(c("BigC-YES"),2)),f.hcla=rep(c("Clust
```

```
##        1        2        3        4
## 4.803513 7.391007 3.720768 8.071936
```

```r
exp(predict(m3,newdata=data.frame(f.bigcity=c(rep(c("BigC-NO"),2),rep(c("BigC-YES"),2)),f.hcla=rep(c("C
```

```
##          1          2          3          4
##  121.93807 1621.33738   41.29609 3203.29835
```

```r
summary(m3)
```

```
##
## Call:
## lm(formula = log(claims + 1) ~ f.bigcity * f.hcla, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8035 -0.8338 -0.0636  1.0147  2.7669
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          4.8035     0.1088  44.130  < 2e-16 ***
## f.bigcityBigC-YES                   -1.0827     0.5072  -2.135   0.0342 *
## f.hclaCluster-2                      2.5875     0.3937   6.572 5.69e-10 ***
## f.bigcityBigC-YES:f.hclaCluster-2   1.7637     0.7373   2.392   0.0178 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.311 on 172 degrees of freedom
## Multiple R-squared:  0.3968, Adjusted R-squared:  0.3863
## F-statistic: 37.72 on 3 and 172 DF,  p-value: < 2.2e-16
```

```r
Anova(m3)
```

```
## Anova Table (Type II tests)
##
## Response: log(claims + 1)
##                  Sum Sq  Df F value  Pr(>F)
## f.bigcity          0.78   1  0.4540 0.50133
## f.hcla           148.07   1 86.1918 < 2e-16 ***
## f.bigcity:f.hcla   9.83   1  5.7222 0.01783 *
## Residuals        295.49 172
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.7 Point 7

7. A new model using numeric explanatory variables and available factors is being considered. Which are the significant variables for model building purposes? Comment goodness of fit..

Despite AIC criteria for selection of the best model m5 retains log(accidents):f.bigcity interaction. But net effects indicate that no significance according to Fisher test for forecasting purposes exits. Log(ki) and log(accidents) have very low pvalues in the net-effect test that it is affected in log(accidents) by a perturbation from the interaction. Main-effects in net-effects Anova(m5) testing should not be considered. Just, a new m6 model with the interaction has to be estimated and net-effects tests interpreted again (no output included in the exam).

```r
m4<-lm(log(claims+1)~(log(population)+log(ki)+log(accidents)+pop_density)*(f.bigcity+f.hcla), data=df)
m5<-step(m4)
```

```
## Start:  AIC=-292.43
## log(claims + 1) ~ (log(population) + log(ki) + log(accidents) +
##     pop_density) * (f.bigcity + f.hcla)
##
##                              Df Sum of Sq    RSS     AIC
## - log(population):f.bigcity  1   0.00194 28.178 -294.42
## - pop_density:f.bigcity      1   0.00928 28.186 -294.37
## - pop_density:f.hcla         1   0.07045 28.247 -293.99
## - log(population):f.hcla     1   0.07558 28.252 -293.96
## - log(accidents):f.hcla      1   0.10433 28.281 -293.78
## - log(ki):f.hcla             1   0.10594 28.282 -293.77
## <none>                                   28.176 -292.43
## - log(ki):f.bigcity          1   0.40384 28.580 -291.93
## - log(accidents):f.bigcity   1   0.45212 28.628 -291.63
##
## Step:  AIC=-294.42
## log(claims + 1) ~ log(population) + log(ki) + log(accidents) +
##     pop_density + f.bigcity + f.hcla + log(population):f.hcla +
##     log(ki):f.bigcity + log(ki):f.hcla + log(accidents):f.bigcity +
##     log(accidents):f.hcla + pop_density:f.bigcity + pop_density:f.hcla
##
##                            Df Sum of Sq    RSS     AIC
## - pop_density:f.bigcity     1   0.00981 28.188 -296.36
## - pop_density:f.hcla        1   0.06893 28.247 -295.99
## - log(population):f.hcla    1   0.07420 28.253 -295.96
## - log(accidents):f.hcla     1   0.10247 28.281 -295.78
## - log(ki):f.hcla            1   0.10483 28.283 -295.77
## <none>                                  28.178 -294.42
## - log(ki):f.bigcity         1   0.40290 28.581 -293.92
## - log(accidents):f.bigcity  1   0.45094 28.629 -293.63
##
## Step:  AIC=-296.36
## log(claims + 1) ~ log(population) + log(ki) + log(accidents) +
##     pop_density + f.bigcity + f.hcla + log(population):f.hcla +
##     log(ki):f.bigcity + log(ki):f.hcla + log(accidents):f.bigcity +
##     log(accidents):f.hcla + pop_density:f.hcla
##
##                            Df Sum of Sq    RSS     AIC
## - log(population):f.hcla    1   0.10230 28.290 -297.72
## - log(ki):f.hcla            1   0.11501 28.303 -297.64
## - log(accidents):f.hcla     1   0.11917 28.307 -297.62
## - pop_density:f.hcla        1   0.19117 28.379 -297.17
## <none>                                  28.188 -296.36
## - log(ki):f.bigcity         1   0.40643 28.595 -295.84
## - log(accidents):f.bigcity  1   0.45481 28.643 -295.54
##
## Step:  AIC=-297.72
## log(claims + 1) ~ log(population) + log(ki) + log(accidents) +
##     pop_density + f.bigcity + f.hcla + log(ki):f.bigcity + log(ki):f.hcla +
##     log(accidents):f.bigcity + log(accidents):f.hcla + pop_density:f.hcla
##
```

```
##                            Df Sum of Sq    RSS     AIC
## - log(population)           1   0.02333 28.314 -299.58
## - log(accidents):f.hcla     1   0.12204 28.412 -298.96
## - pop_density:f.hcla        1   0.15867 28.449 -298.74
## - log(ki):f.hcla            1   0.17881 28.469 -298.61
## <none>                                  28.290 -297.72
## - log(ki):f.bigcity         1   0.41012 28.701 -297.19
## - log(accidents):f.bigcity  1   0.47153 28.762 -296.81
##
## Step:  AIC=-299.58
## log(claims + 1) ~ log(ki) + log(accidents) + pop_density + f.bigcity +
##     f.hcla + log(ki):f.bigcity + log(ki):f.hcla + log(accidents):f.bigcity +
##     log(accidents):f.hcla + pop_density:f.hcla
##
##                            Df Sum of Sq    RSS     AIC
## - log(accidents):f.hcla     1   0.12716 28.441 -300.79
## - pop_density:f.hcla        1   0.15344 28.467 -300.62
## - log(ki):f.hcla            1   0.17363 28.487 -300.50
## <none>                                  28.314 -299.58
## - log(ki):f.bigcity         1   0.41719 28.731 -299.00
## - log(accidents):f.bigcity  1   0.48179 28.796 -298.61
##
## Step:  AIC=-300.79
## log(claims + 1) ~ log(ki) + log(accidents) + pop_density + f.bigcity +
##     f.hcla + log(ki):f.bigcity + log(ki):f.hcla + log(accidents):f.bigcity +
##     pop_density:f.hcla
##
##                            Df Sum of Sq    RSS     AIC
## - log(ki):f.hcla            1   0.04810 28.489 -302.49
## - pop_density:f.hcla        1   0.19553 28.636 -301.58
## - log(ki):f.bigcity         1   0.29581 28.737 -300.97
## <none>                                  28.441 -300.79
## - log(accidents):f.bigcity  1   0.35758 28.798 -300.59
##
## Step:  AIC=-302.49
## log(claims + 1) ~ log(ki) + log(accidents) + pop_density + f.bigcity +
##     f.hcla + log(ki):f.bigcity + log(accidents):f.bigcity + pop_density:f.hcla
##
##                            Df Sum of Sq    RSS     AIC
## - pop_density:f.hcla        1   0.15655 28.645 -303.53
## - log(ki):f.bigcity         1   0.25661 28.746 -302.91
## - log(accidents):f.bigcity  1   0.32054 28.809 -302.52
## <none>                                  28.489 -302.49
##
## Step:  AIC=-303.53
## log(claims + 1) ~ log(ki) + log(accidents) + pop_density + f.bigcity +
##     f.hcla + log(ki):f.bigcity + log(accidents):f.bigcity
##
##                            Df Sum of Sq    RSS     AIC
## - pop_density               1   0.00706 28.653 -305.48
## - log(ki):f.bigcity         1   0.14475 28.790 -304.64
## - log(accidents):f.bigcity  1   0.19224 28.838 -304.35
## <none>                                  28.645 -303.53
## - f.hcla                    1   0.53468 29.180 -302.27
```

```
##
## Step:  AIC=-305.48
## log(claims + 1) ~ log(ki) + log(accidents) + f.bigcity + f.hcla +
##     log(ki):f.bigcity + log(accidents):f.bigcity
##
##                            Df Sum of Sq    RSS     AIC
## - log(ki):f.bigcity         1   0.18611 28.839 -306.34
## - log(accidents):f.bigcity  1   0.24050 28.893 -306.01
## <none>                                  28.653 -305.48
## - f.hcla                    1   0.54788 29.201 -304.15
##
## Step:  AIC=-306.34
## log(claims + 1) ~ log(ki) + log(accidents) + f.bigcity + f.hcla +
##     log(accidents):f.bigcity
##
##                            Df Sum of Sq    RSS     AIC
## <none>                                  28.839 -306.34
## - log(accidents):f.bigcity  1   0.41920 29.258 -305.80
## - f.hcla                    1   0.54973 29.388 -305.02
## - log(ki)                   1   2.92174 31.760 -291.36
```

```r
Anova(m5)
```

```
## Anova Table (Type II tests)
##
## Response: log(claims + 1)
##                           Sum Sq  Df F value    Pr(>F)
## log(ki)                   2.9217   1 17.2232 5.244e-05 ***
## log(accidents)            3.6530   1 21.5337 6.917e-06 ***
## f.bigcity                 0.1371   1  0.8084   0.36987
## f.hcla                    0.5497   1  3.2406   0.07361 .
## log(accidents):f.bigcity  0.4192   1  2.4711   0.11781
## Residuals                28.8387 170
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# A new model without the interaction should be checked to assess factor significancies
m6<-lm(log(claims+1)~(log(ki)+log(accidents))+(f.bigcity+f.hcla), data=df)
Anova(m6)
```

```
## Anova Table (Type II tests)
##
## Response: log(claims + 1)
##                 Sum Sq  Df F value    Pr(>F)
## log(ki)         3.1702   1 18.5286 2.812e-05 ***
## log(accidents)  3.6530   1 21.3500 7.503e-06 ***
## f.bigcity       0.1371   1  0.8015    0.3719
## f.hcla          0.2981   1  1.7424    0.1886
## Residuals      29.2579 171
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.8 Point 8

8. Model (m5) is chosen for a detailed analysis. Comment goodness of fit and predict the number of claims for a fictitious LGA unit in cluster 2 having variables number of killed/injured and number of accidents on the mean.

Coefficient of determination indicates that the model explains 94% of target variability, nevertheless it is not optimal since the interaction term is redundant. A simplified model has to be stated.

In terms of interpretation: Prediction equation - BigCity-NO: log(Y+1)=-2.67+0.65log(650.62)+0.67log(1153.1)+0-0.257+0 -> Y=exp(0-2.67+0.65*log(650.62)+0.67*log(1153.1)+0-0.257+0)-1 = 392.6 claims

Prediction equation - BigCity-YES: log(Y+1)=-2.67+0.65log(650.62)+(0.67+0.106)log(1153.1)-0.823-0.257 -> Y=exp(0-2.67+0.65*log(650.62)+(0.67+0.106)*log(1153.1)-0.823-0.257)-1 = 362.1 claims

```r
summary(m5)
```

```
##
## Call:
## lm(formula = log(claims + 1) ~ log(ki) + log(accidents) + f.bigcity +
##     f.hcla + log(accidents):f.bigcity, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.78513 -0.22340  0.02513  0.23246  1.25766
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -2.67239    0.19545 -13.673  < 2e-16 ***
## log(ki)                            0.64768    0.15606   4.150 5.24e-05 ***
## log(accidents)                     0.66755    0.14134   4.723 4.84e-06 ***
## f.bigcityBigC-YES                 -0.82969    0.47579  -1.744   0.0830 .
## f.hclaCluster-2                   -0.25686    0.14269  -1.800   0.0736 .
## log(accidents):f.bigcityBigC-YES  0.10622    0.06757   1.572   0.1178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4119 on 170 degrees of freedom
## Multiple R-squared:  0.9411, Adjusted R-squared:  0.9394
## F-statistic: 543.6 on 5 and 170 DF,  p-value: < 2.2e-16
```

```r
mean(df$ki);mean(df$accidents)
```

```
## [1] 650.6193
```

```
## [1] 1153.097
```

```r
predict(m5,newdata=data.frame(f.bigcity=c("BigC-NO","BigC-YES"),f.hcla=c("Cluster-2"),ki=mean(df$ki),acc
```

```
##        1        2
## 5.972746 5.891935
```

```r
exp(predict(m5,newdata=data.frame(f.bigcity=c("BigC-NO","BigC-YES"),f.hcla=c("Cluster-2"),ki=mean(df$ki
```

```
##        1        2
## 391.5821 361.1052
```

1.9 Point 9

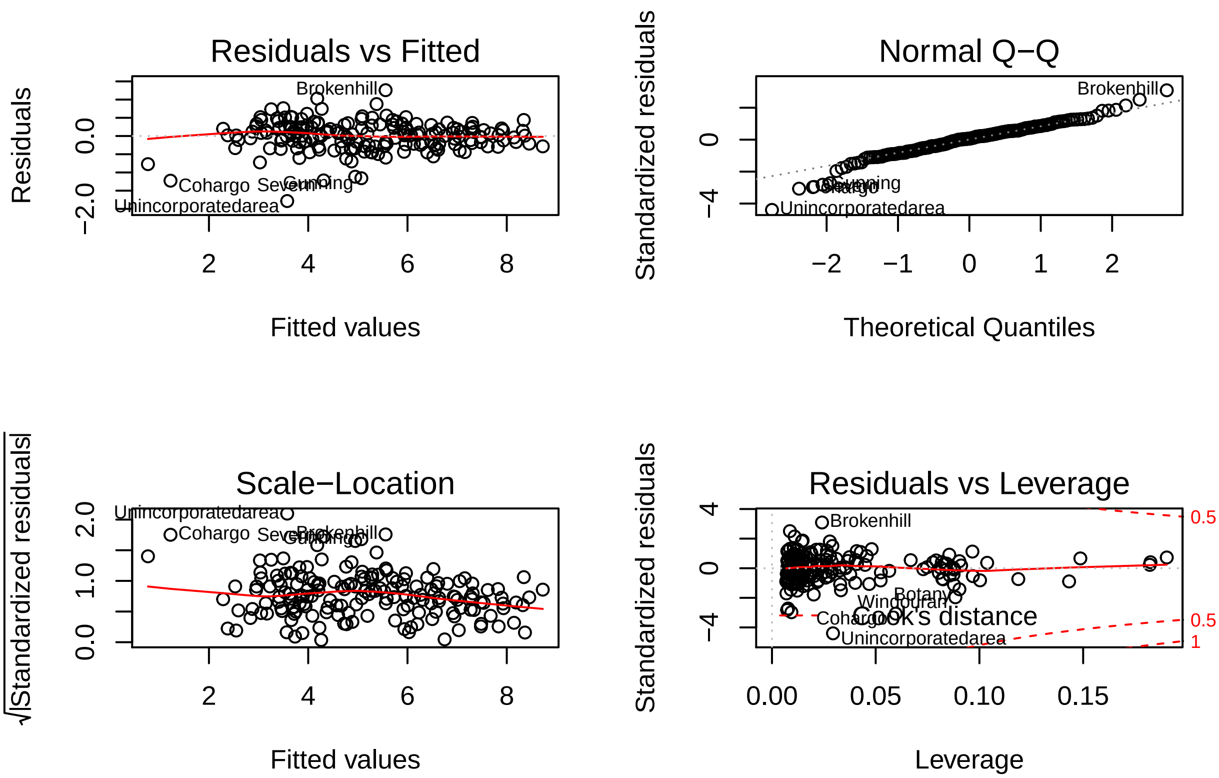
9. Make a rough assessment of the quality of the model based on the first impression of the diagnosis of residuals for (m5).

Upper-left: A random pattern seems to be present, but some observations are far away from the cloud of point with remarkable lack of fit (Brokenhill, Cohargo). Residual outliers are present.

Upper-right: Normal distribution of residuals is not met. Negative tails show big discrepancies to normality. Brokenhill is an outlier.

Below-left: Variance seems to be constant despite the perturbations caused by unsual data (far from the cloud of points, as Windouran, Cohargo) Below-right: residual outliers and unsual data are present. Influent data can not be assessed with the plot.

```
par(mfrow=c(2,2))
plot(m5,id.n=5)
```



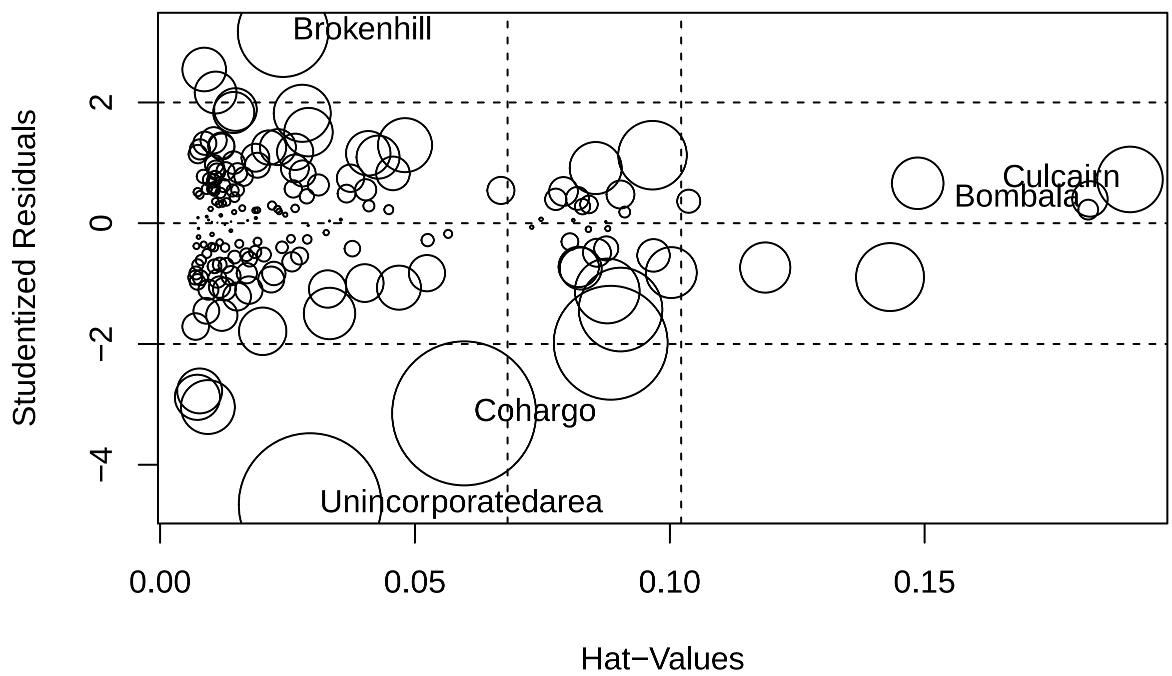

```
par(mfrow=c(1,1))
```

1.10 Point 10

10. Indicate whether observations 182, 366 and 254 are residual outliers or influent data or none of both. Justify your answer in terms of residuals, leverage and Cook's distance.

Bombala is unusual, large leverage. Residual is not large and causes no bad influence to parameter estimates Brokenhill shows lack of fit (residual outlier). No leverage and no influence. Cohargo shows lack of fit and remarkable large Cook's distance: it is an influent observation.

```r
influencePlot(m5)
```




```
##                      StudRes        Hat       CookD
## Bombala            0.4021433 0.18245341 0.006045007
## Brokenhill         3.1723657 0.02412814 0.039371972
## Cohargo           -3.1492424 0.05966103 0.099646679
## Culcairn           0.7252676 0.19031062 0.020663435
## Unincorporatedarea -4.6597088 0.02943100 0.097816859
```