Machine Learning I Group Work

Adriana Ricklin, Yvonne Schaerli, Christina Sudermann, Carole Mattmann

5 Maerz 2020

Packages

```
library(ggplot2)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from 
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.6.2
```

Import and data cleaning

```
insurance <- read.csv("../01_data/insurance.csv", header=TRUE)
str(insurance)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

```
head(insurance)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

Linear models -> Christina

Linear models (GAM & Polynomial) -> Yvonne

Generalised Linear Models for count data -> Carole

Original data

The number of children an insured person has is analysed. We have the following data on children per person. The number of children ranges from 0 to 5 with a median of 1.

```
summary(insurance$children)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   0.000   1.000   1.095   2.000   5.000 
```

Poisson model

To model count data (number of children) the poisson model is used. An analysis performed beforehand showed that only the variables “charges” and “smoker” have a significant impact on the number of children.

```
glm.children <- glm(children ~ smoker+charges,
                   data=insurance,
                   family = "poisson")

summary(glm.children)
```

```
## 
## Call:
## glm(formula = children ~ smoker + charges, family = "poisson", 
##     data = insurance)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.8561  -1.4318  -0.1057   0.7768   2.9717  
## 
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -3.706e-02  4.213e-02  -0.880   0.3790    
## smokeryes   -3.239e-01  1.058e-01  -3.061   0.0022 ** 
## charges      1.419e-05  3.365e-06   4.217 2.48e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 2001.6  on 1337  degrees of freedom
## Residual deviance: 1984.1  on 1335  degrees of freedom
## AIC: 3879.4
## 
## Number of Fisher Scoring iterations: 5
```

To get the coefficients, the log transformation needs to be reversed:

```
exp(coef(glm.children))
```

```
## (Intercept)   smokeryes     charges 
##   0.9636169   0.7233085   1.0000142
```

Smoker (factor): The model shows that for the factor smoker (yes/no), a smoker has on average 72% of the number of children a non-smoker has. The more common-sense interpretation might be the other way around, that people who have 1 or more children smoke less, but for the moment we have no proof of that.

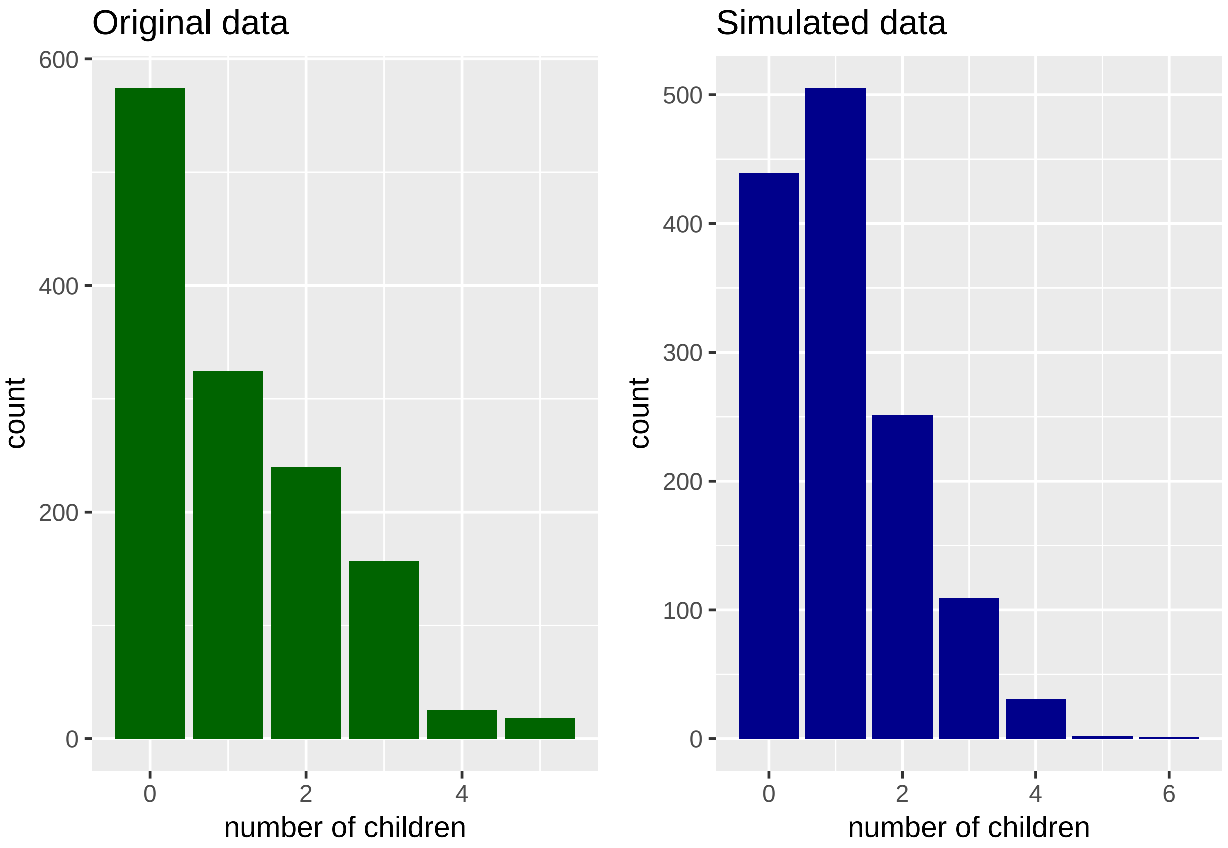
Charges: A person with higher charges will on average have more children. If charges are increased by 1000 dollars, the calculated number of children increases by 1.4%.

##Simulation of data and comparison

With the calculated model, data is simulated:

```
##      sim_1
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :1.000
##  Mean   :1.102
##  3rd Qu.:2.000
##  Max.   :6.000
```

The original and the simulated data are compared visually. The number of children from the simulated data (0-6) seem to be plausible. The distribution has a strong downwards trend starting at 1 like the original data. However the model does not seem to generate enough data with 0 children.



Generalised Linear Models for binomial data -> Carole

Cross Validation -> Carole