GLM dvisits dataset project

Vuyolwethu Zumani

15 April 2020

a)

Build a Poisson regression model with doctorco as the response and sex, age, agesq, income, levyplus, freepoor, freerepa, illness, actdays, hscore, chcond1, and chcond2 as possible predictor variables. Considering the deviance of this model, does this model fit the data?

```
library(faraway)
data(dvisits)

head(dvisits)
```

```
##   sex  age  agesq income levyplus freepoor freerepa illness actdays hscore
## 1   1 0.19 0.0361   0.55        1        0        0       1       4      1
## 2   1 0.19 0.0361   0.45        1        0        0       1       2      1
## 3   0 0.19 0.0361   0.90        0        0        0       3       0      0
## 4   0 0.19 0.0361   0.15        0        0        0       1       0      0
## 5   0 0.19 0.0361   0.45        0        0        0       2       5      1
## 6   1 0.19 0.0361   0.35        0        0        0       5       1      9
##   chcond1 chcond2 doctorco nondocco hospadmi hospdays medicine prescrib
## 1       0       0        1        0        0        0        1        1
## 2       0       0        1        0        0        0        2        1
## 3       0       0        1        0        1        4        2        1
## 4       0       0        1        0        0        0        0        0
## 5       1       0        1        0        0        0        3        1
## 6       1       0        1        0        0        0        1        1
##   nonpresc
## 1        0
## 2        1
## 3        1
## 4        0
## 5        2
## 6        0
```

```
modelglm_poisson <- glm(doctorco ~ sex + age + agesq + income + levyplus + freepoor + freerepa + illness
modelglm_poisson
```

```
##
## Call:  glm(formula = doctorco ~ sex + age + agesq + income + levyplus +
##     freepoor + freerepa + illness + actdays + hscore + chcond1 +
##     chcond2, family = poisson, data = dvisits)
```

```
##
## Coefficients:
## (Intercept)          sex          age        agesq       income     levyplus
##    -2.22385      0.15688      1.05630     -0.84870     -0.20532      0.12319
##    freepoor     freerepa      illness      actdays       hscore      chcond1
##    -0.44006      0.07980      0.18695      0.12685      0.03008      0.11409
##     chcond2
##     0.14116
##
## Degrees of Freedom: 5189 Total (i.e. Null);  5177 Residual
## Null Deviance:       5635
## Residual Deviance: 4380  AIC: 6737
```

```
summary(modelglm_poisson)
```

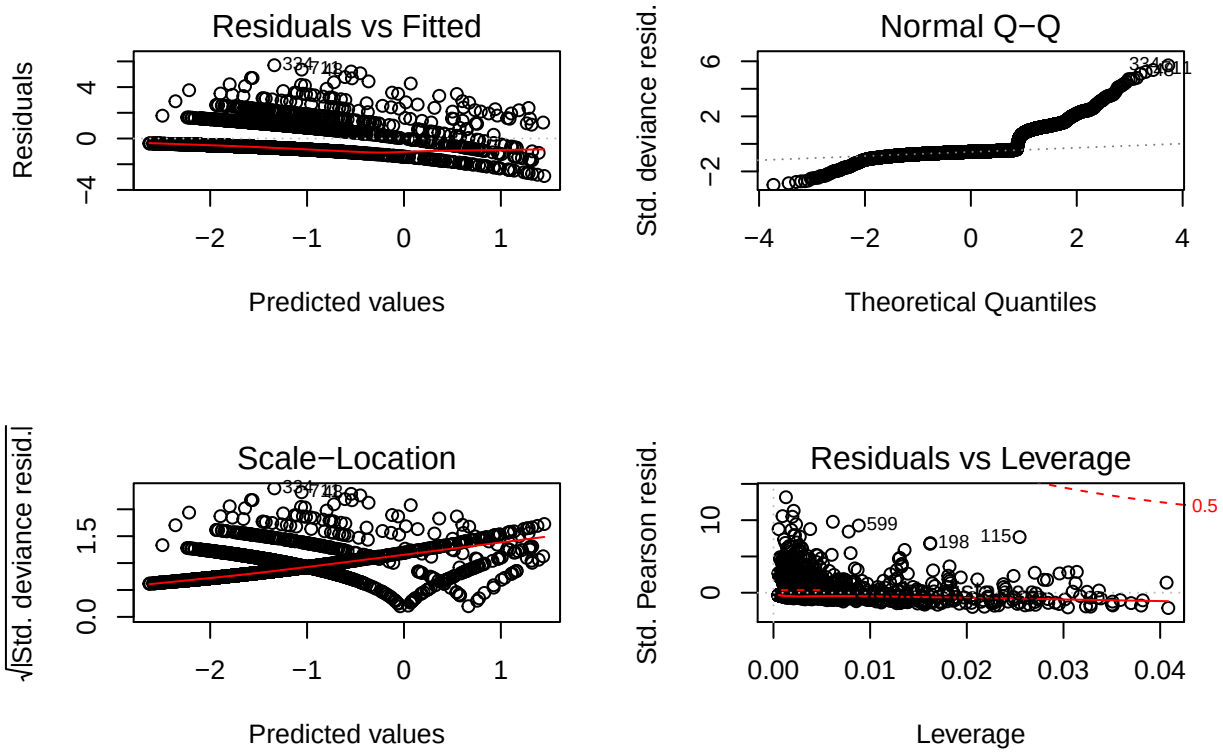
```
##
## Call:
## glm(formula = doctorco ~ sex + age + agesq + income + levyplus +
##     freepoor + freerepa + illness + actdays + hscore + chcond1 +
##     chcond2, family = poisson, data = dvisits)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9170  -0.6862  -0.5743  -0.4839   5.7005
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.223848   0.189816 -11.716   <2e-16 ***
## sex          0.156882   0.056137   2.795   0.0052 **
## age          1.056299   1.000780   1.055   0.2912
## agesq       -0.848704   1.077784  -0.787   0.4310
## income      -0.205321   0.088379  -2.323   0.0202 *
## levyplus     0.123185   0.071640   1.720   0.0855 .
## freepoor    -0.440061   0.179811  -2.447   0.0144 *
## freerepa     0.079798   0.092060   0.867   0.3860
## illness      0.186948   0.018281  10.227   <2e-16 ***
## actdays      0.126846   0.005034  25.198   <2e-16 ***
## hscore       0.030081   0.010099   2.979   0.0029 **
## chcond1      0.114085   0.066640   1.712   0.0869 .
## chcond2      0.141158   0.083145   1.698   0.0896 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5634.8  on 5189  degrees of freedom
## Residual deviance: 4379.5  on 5177  degrees of freedom
## AIC: 6737.1
##
## Number of Fisher Scoring iterations: 6
```

The residual deviance is quite high.This indicate that the Poisson is probably an ill-fit of the data since the residual deviance is very large.

b)

Plot the residuals and the fitted data - why are there lines of observations on the plot?

```
par(mfrow=c(2,2))
plot(modelglm_poisson)
```



The lines are there due to the fact that the responses are discrete continous numbers.

c)

Use backward elimination with a critical p-value of 5% to reduce the model as much as possible. Report your model.

```
step(modelglm_poisson, direction="backward")
```

```
## Start:  AIC=6737.08
## doctorco ~ sex + age + agesq + income + levyplus + freepoor +
##     freerepa + illness + actdays + hscore + chcond1 + chcond2
##
##             Df Deviance    AIC
## - agesq      1   4380.1 6735.7
## - freerepa   1   4380.3 6735.8
## - age        1   4380.6 6736.2
## <none>           4379.5 6737.1
```

```
## - chcond2   1   4382.4 6738.0
## - chcond1   1   4382.5 6738.0
## - levyplus  1   4382.5 6738.1
## - income    1   4385.0 6740.5
## - freepoor  1   4386.2 6741.8
## - sex       1   4387.4 6743.0
## - hscore    1   4388.1 6743.7
## - illness   1   4481.8 6837.4
## - actdays   1   4917.1 7272.7
##
## Step:  AIC=6735.7
## doctorco ~ sex + age + income + levyplus + freepoor + freerepa +
##     illness + actdays + hscore + chcond1 + chcond2
##
##            Df Deviance    AIC
## - freerepa  1   4381.0 6734.5
## <none>          4380.1 6735.7
## - age       1   4383.0 6736.5
## - chcond1   1   4383.2 6736.8
## - levyplus  1   4383.3 6736.9
## - chcond2   1   4383.5 6737.0
## - income    1   4385.0 6738.6
## - freepoor  1   4386.8 6740.4
## - sex       1   4388.0 6741.5
## - hscore    1   4389.1 6742.7
## - illness   1   4481.9 6835.4
## - actdays   1   4917.1 7270.7
##
## Step:  AIC=6734.53
## doctorco ~ sex + age + income + levyplus + freepoor + illness +
##     actdays + hscore + chcond1 + chcond2
##
##            Df Deviance    AIC
## <none>          4381.0 6734.5
## - levyplus  1   4383.4 6735.0
## - chcond1   1   4384.3 6735.9
## - chcond2   1   4384.7 6736.3
## - income    1   4386.7 6738.2
## - age       1   4387.1 6738.7
## - freepoor  1   4389.1 6740.6
## - sex       1   4389.5 6741.0
## - hscore    1   4390.2 6741.8
## - illness   1   4482.7 6834.2
## - actdays   1   4917.6 7269.2


##
## Call:  glm(formula = doctorco ~ sex + age + income + levyplus + freepoor +
##     illness + actdays + hscore + chcond1 + chcond2, family = poisson,
##     data = dvisits)
##
## Coefficients:
## (Intercept)          sex          age       income     levyplus     freepoor
##    -2.08906      0.16200      0.35513     -0.19981      0.08369     -0.46960
##     illness      actdays       hscore      chcond1      chcond2
```

```
##     0.18610      0.12661      0.03112      0.12110      0.15889
##
## Degrees of Freedom: 5189 Total (i.e. Null);  5179 Residual
## Null Deviance:       5635
## Residual Deviance: 4381  AIC: 6735
```

The backward elimination has not made any improvements.

d)

What sort of person would be predicted to visit the doctor the most under your selected model?

A person who may have some form of illness (illness); A person who is elderly (age); A person who has some form of income and can afford to see a private doctor (income); If a person had been inactive for some time (actdays); If a person is covered by the government to be able to see the doctor (freepoor); A person who may have bad health (hscore); People who may be living with chronic conditions (chcond1 & chcond2).

All of these are significant in the model and these are indeed some of the most significant reasons why people would seek to visit a doctor.

e)

For the last person in the dataset, compute the predicted probability distribution for their visits to the doctor, i.e., give the probability they visit 0,1,2, etc. times.

```r
# 5190 participated in the survey
predict(modelglm_poisson, dvisits[5190,], type="response")
```

```
##      5190
## 0.1533837
```

This implies that $\lambda = 0.16$ where λ is the mean of the poisson distribution.

```r
# Probability of 0 doctor's visits:
dpois(0, lambda = 0.153)
```

```
## [1] 0.8581297
```

```r
# Probability of 1 doctor's visits:
dpois(1, lambda = 0.153)
```

```
## [1] 0.1312938
```

```r
# Probability of 2 doctor's visits:
dpois(2, lambda = 0.153)
```

```
## [1] 0.01004398
```

```r
# Probability of 3 doctor's visits:
dpois(3, lambda = 0.153)
```

```
## [1] 0.0005122429
```

```r
# Probability of 4 doctor's visits:
dpois(4, lambda = 0.153)
```

```
## [1] 1.959329e-05
```

```r
# Probability of 5 doctor's visits:
dpois(5, lambda = 0.153)
```

```
## [1] 5.995548e-07
```

f)

Fit a comparable (Gaussian) linear model and graphically compare the fits. Describe how they differ.

```r
model_lm <- lm(doctorco ~ sex + age + agesq + income + levyplus + freepoor + freerepa + illness + actda
model_lm
```

```
##
## Call:
## lm(formula = doctorco ~ sex + age + agesq + income + levyplus +
##     freepoor + freerepa + illness + actdays + hscore + chcond1 +
##     chcond2, data = dvisits)
##
## Coefficients:
## (Intercept)          sex          age        agesq       income     levyplus
##    0.027632     0.033811     0.203201    -0.062103    -0.057323     0.035179
##    freepoor     freerepa      illness      actdays       hscore      chcond1
##   -0.103314     0.033241     0.059946     0.103192     0.016976     0.004384
##     chcond2
##    0.041617
```

```r
summary(model_lm)
```

```
##
## Call:
## lm(formula = doctorco ~ sex + age + agesq + income + levyplus +
##     freepoor + freerepa + illness + actdays + hscore + chcond1 +
##     chcond2, data = dvisits)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1352 -0.2588 -0.1435 -0.0433  7.0327
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.027632   0.072220   0.383  0.70202
## sex          0.033811   0.021604   1.565  0.11764
## age          0.203201   0.410016   0.496  0.62020
## agesq       -0.062103   0.458716  -0.135  0.89231
## income      -0.057323   0.033089  -1.732  0.08326 .
## levyplus     0.035179   0.024882   1.414  0.15748
## freepoor    -0.103314   0.052471  -1.969  0.04901 *
## freerepa     0.033241   0.038157   0.871  0.38371
## illness      0.059946   0.008357   7.173 8.39e-13 ***
## actdays      0.103192   0.003657  28.216  < 2e-16 ***
## hscore       0.016976   0.005190   3.271  0.00108 **
## chcond1      0.004384   0.023740   0.185  0.85349
## chcond2      0.041617   0.035863   1.160  0.24592
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7139 on 5177 degrees of freedom
## Multiple R-squared:  0.2018, Adjusted R-squared:    0.2
## F-statistic: 109.1 on 12 and 5177 DF,  p-value: < 2.2e-16
```

```
predict(model_lm, dvisits[5190,], type="response")
```

```
##      5190
## 0.1606531
```

They are fairly similar.