HomeworkAnswers4.R

dbarron

Sun Mar 04 20:09:13 2018

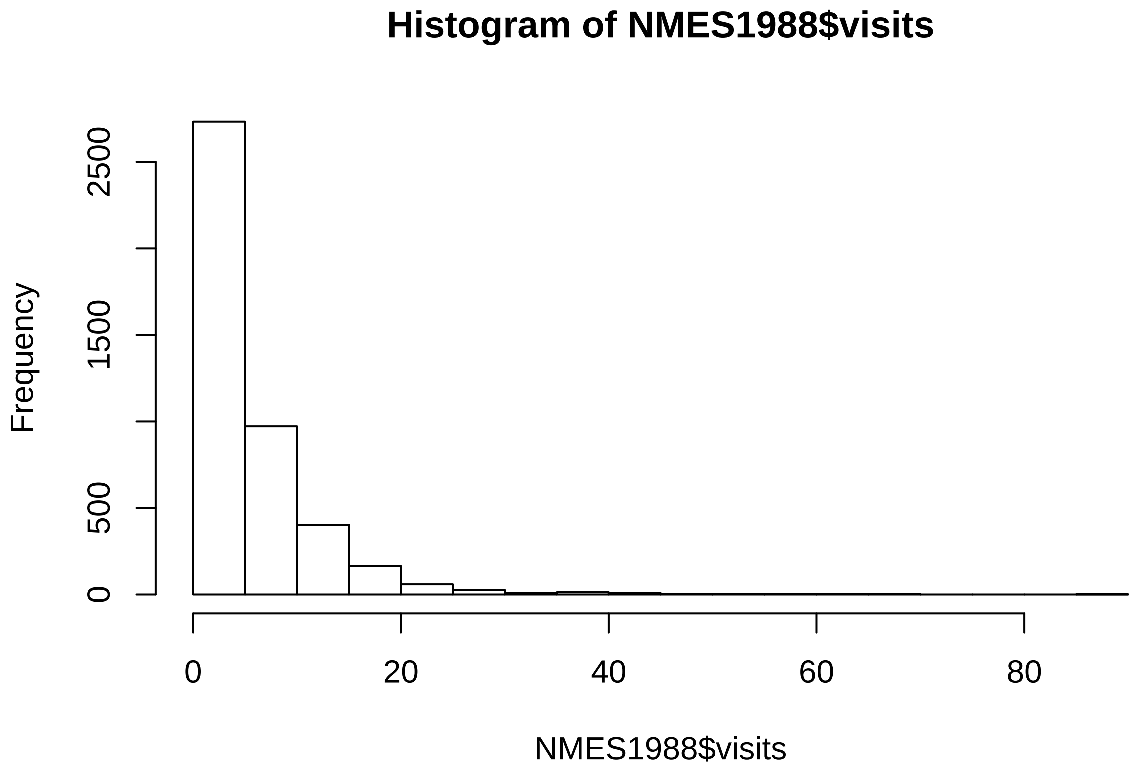
```r
library(AER)
```

```
## Loading required package: car

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival
```

```r
library(effects)
```

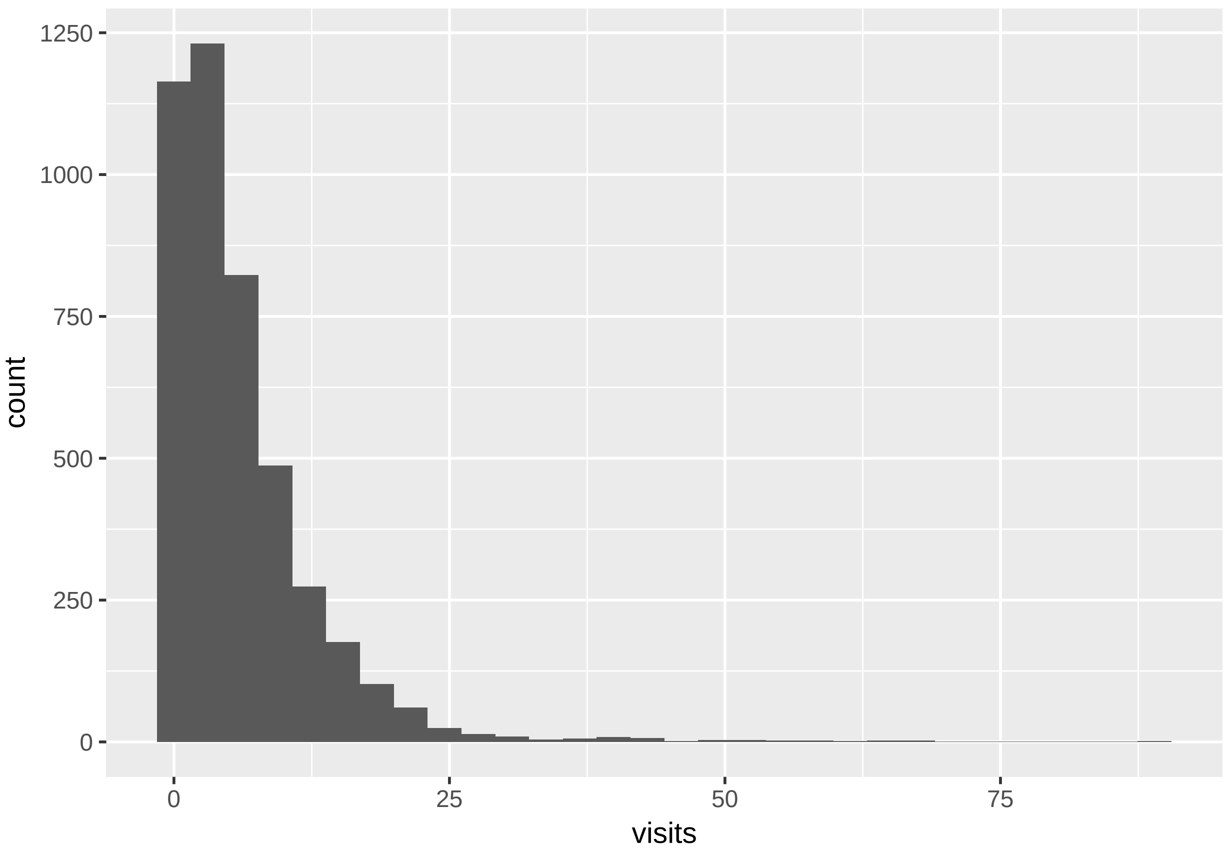
```
## Loading required package: carData

##
## Attaching package: 'carData'

## The following objects are masked from 'package:car':
##
##     Guyer, UN, Vocab

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```r
library(MASS)
data(NMES1988)

hist(NMES1988$visits)
```



```
## Or, if you prefer ggplot:

library(ggplot2)
ggplot(NMES1988, aes(x = visits)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
xtabs(~ visits, data = NMES1988)
```

```
## visits
##    0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17 
## 683 481 428 420 383 338 268 217 188 171 128 115  86  73  76  53  47  48 
##   18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35 
##   30  24  16  18  16  10  12   3   9   7   4   3   4   4   1   1   2   1 
##   36  37  38  39  40  41  42  43  44  47  48  49  50  51  53  55  56  58 
##    1   3   2   5   2   1   4   2   1   1   1   1   1   1   2   1   1   2 
##   61  63  65  66  68  89 
##    1   1   1   1   1   1
```

```
summary(NMES1988)
```

```
##      visits          nvisits           ovisits            novisits      
##  Min.   : 0.000   Min.   :  0.000   Min.   :  0.0000   Min.   :  0.0000  
##  1st Qu.: 1.000   1st Qu.:  0.000   1st Qu.:  0.0000   1st Qu.:  0.0000  
##  Median : 4.000   Median :  0.000   Median :  0.0000   Median :  0.0000  
##  Mean   : 5.774   Mean   :  1.618   Mean   :  0.7508   Mean   :  0.5361  
##  3rd Qu.: 8.000   3rd Qu.:  1.000   3rd Qu.:  0.0000   3rd Qu.:  0.0000  
##  Max.   :89.000   Max.   :104.000   Max.   :141.0000   Max.   :155.0000  
##    emergency          hospital            health        chronic     
##  Min.   : 0.0000   Min.   :0.000   poor     : 554   Min.   :0.000  
##  1st Qu.: 0.0000   1st Qu.:0.000   average  :3509   1st Qu.:1.000  
##  Median : 0.0000   Median :0.000   excellent: 343   Median :1.000  
##  Mean   : 0.2635   Mean   :0.296                    Mean   :1.542  
##  3rd Qu.: 0.0000   3rd Qu.:0.000                    3rd Qu.:2.000  
##  Max.   :12.0000   Max.   :8.000                    Max.   :8.000  
##       adl             region          age          afam         gender    
##  normal :3507   northeast: 837   Min.   : 6.600   no :3890   female:2628  
##  limited: 899   midwest  :1157   1st Qu.: 6.900   yes: 516   male  :1778  
##                 west     : 798   Median : 7.300                           
##                 other    :1614   Mean   : 7.402                           
##                                  3rd Qu.: 7.800                           
##                                  Max.   :10.900                           
##  married        school          income        employed   insurance 
##  no :2000   Min.   : 0.00   Min.   :-1.0125   no :3951   no : 985  
##  yes:2406   1st Qu.: 8.00   1st Qu.: 0.9122   yes: 455   yes:3421  
##             Median :11.00   Median : 1.6982                        
##             Mean   :10.29   Mean   : 2.5271                        
##             3rd Qu.:12.00   3rd Qu.: 3.1728                        
##             Max.   :18.00   Max.   :54.8351                        
##  medicaid  
##  no :4004  
##  yes: 402  
##            
##            
##            
##            
```

```r
base <- glm(visits ~ 1, data = NMES1988, family = poisson())

summary(base)
```

```
## 
## Call:
## glm(formula = visits ~ 1, family = poisson(), data = NMES1988)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -3.3984  -2.4580  -0.7821   0.8746  17.9001  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) 1.753434   0.006269   279.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 26943  on 4405  degrees of freedom
## Residual deviance: 26943  on 4405  degrees of freedom
## AIC: 39720
## 
## Number of Fisher Scoring iterations: 5
```

```r
vis.step <- step(base, scope = ~ hospital + health + chronic + gender + school + insurance,
                 direction='forward', trace = 1)
```

```
## Start:  AIC=39720.34
## visits ~ 1
## 
```

```
##             Df Deviance   AIC
## + chronic    1    24768 37547
## + hospital   1    25449 38228
## + health     2    25715 38497
## + insurance  1    26774 39553
## + school     1    26797 39576
## + gender     1    26878 39657
## <none>            26943 39720
##
## Step:  AIC=37546.91
## visits ~ chronic
##
##             Df Deviance   AIC
## + hospital   1    23982 36763
## + health     2    24391 37175
## + school     1    24524 37305
## + insurance  1    24526 37307
## + gender     1    24704 37485
## <none>            24768 37547
##
## Step:  AIC=36763.28
## visits ~ chronic + hospital
##
##             Df Deviance   AIC
## + school     1    23722 36505
## + health     2    23730 36516
## + insurance  1    23752 36535
## + gender     1    23915 36698
## <none>            23982 36763
##
## Step:  AIC=36505.47
## visits ~ chronic + hospital + school
##
##             Df Deviance   AIC
## + health     2    23382 36169
## + insurance  1    23609 36394
## + gender     1    23650 36435
## <none>            23722 36505
##
## Step:  AIC=36169.44
## visits ~ chronic + hospital + school + health
##
##             Df Deviance   AIC
## + insurance  1    23244 36033
## + gender     1    23316 36105
## <none>            23382 36169
##
## Step:  AIC=36033.15
## visits ~ chronic + hospital + school + health + insurance
##
##          Df Deviance   AIC
## + gender  1    23168 35959
## <none>         23244 36033
##
```

```
## Step:  AIC=35959.23
## visits ~ chronic + hospital + school + health + insurance + gender
```

```
summary(vis.step)
```

```
##
## Call:
## glm(formula = visits ~ chronic + hospital + school + health +
##     insurance + gender, family = poisson(), data = NMES1988)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.4055  -1.9962  -0.6737   0.7049  16.3620
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.028874   0.023785  43.258   <2e-16 ***
## chronic          0.146639   0.004580  32.020   <2e-16 ***
## hospital         0.164797   0.005997  27.478   <2e-16 ***
## school           0.026143   0.001843  14.182   <2e-16 ***
## healthpoor       0.248307   0.017845  13.915   <2e-16 ***
## healthexcellent -0.361993   0.030304 -11.945   <2e-16 ***
## insuranceyes     0.201687   0.016860  11.963   <2e-16 ***
## gendermale      -0.112320   0.012945  -8.677   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 26943  on 4405  degrees of freedom
## Residual deviance: 23168  on 4398  degrees of freedom
## AIC: 35959
##
## Number of Fisher Scoring iterations: 5
```

```
# All those variables are statistically significant

## Test for overdispersion by doing negative binomial regression
vis.nb <- glm.nb(visits ~ hospital + health + chronic + gender + school + insurance, data = NMES1988)
summary(vis.nb)
```

```
##
## Call:
## glm.nb(formula = visits ~ hospital + health + chronic + gender +
##     school + insurance, data = NMES1988, init.theta = 1.206603534,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0469  -0.9955  -0.2948   0.2961   5.8185
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.929257   0.054591  17.022  < 2e-16 ***
## hospital         0.217772   0.020176  10.793  < 2e-16 ***
## healthpoor       0.305013   0.048511   6.288 3.23e-10 ***
## healthexcellent -0.341807   0.060924  -5.610 2.02e-08 ***
## chronic          0.174916   0.012092  14.466  < 2e-16 ***
## gendermale      -0.126488   0.031216  -4.052 5.08e-05 ***
## school           0.026815   0.004394   6.103 1.04e-09 ***
## insuranceyes     0.224402   0.039464   5.686 1.30e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.2066) family taken to be 1)
##
##     Null deviance: 5743.7  on 4405  degrees of freedom
## Residual deviance: 5044.5  on 4398  degrees of freedom
## AIC: 24359
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.2066
##           Std. Err.:  0.0336
##
##  2 x log-likelihood:  -24341.1070
```

```
# You can see that Theta is much more than twice its standard error.  Could do a likelihood ratio test.
# I'll illustrate how to write a function.

overdisp.test <- function(mod, alpha = 0.05){
  # mod is the result of running a glm.nb regression
  if (class(mod)[1] != 'negbin') stop('require model of class negbin\n')
  if (alpha < 0 | alpha > 1) stop('alpha must be in the range (0, 1)')
  # obtain Poissin regression results
  poisreg <- glm(formula = eval(mod$call$formula), data = eval(mod$call$data), family = poisson)
  llP <- logLik(poisreg)
  llNB <- logLik(mod)
  D <- 2 * (llNB - llP)
  cv <- qchisq(1 - (2 * alpha), df = 1)
  pval <- pchisq(D, df = 1, lower.tail = FALSE) / 2
  cat('Likelihood ratio test of H0: no overdispersion\n')
  cat('Test statistic: ', D, '\n')
  cat('Critical value of test statistic: ', cv, '\n')
  cat('p-value: ', pval, '\n')
  invisible(c(stat = D, critval = cv, pval = pval))
}

odt <- overdisp.test(vis.nb)
```

```
## Likelihood ratio test of H0: no overdispersion
## Test statistic:  11602.12
## Critical value of test statistic:  2.705543
## p-value:  0
```

```
# THere is a similar function in the package pscl, called odTest
pscl::odTest(vis.nb)
```

```
## Likelihood ratio test of H0: Poisson, as restricted NB model:
## n.b., the distribution of the test-statistic under H0 is non-standard
## e.g., see help(odTest) for details/references
##
## Critical value of test statistic at the alpha= 0.05 level: 2.7055
## Chi-Square Test Statistic =  11602.1184 p-value = < 2.2e-16
```

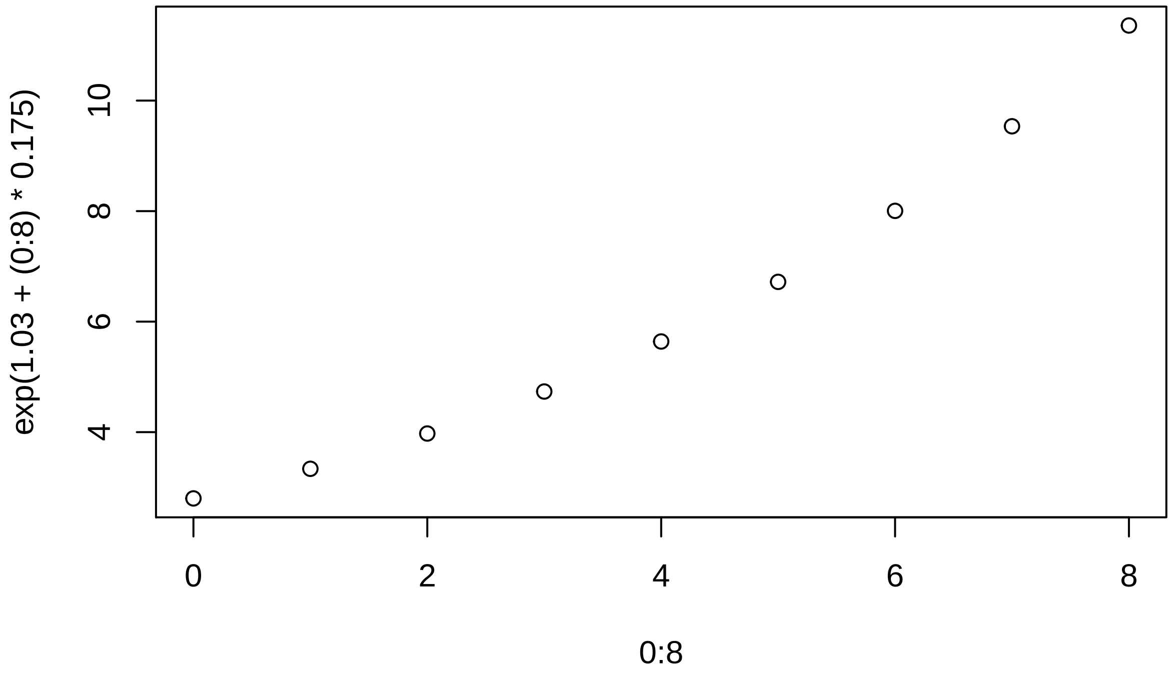
```
# There is very clear evidence of overdispersion

summary(vis.nb)
```

```
##
## Call:
## glm.nb(formula = visits ~ hospital + health + chronic + gender +
##     school + insurance, data = NMES1988, init.theta = 1.206603534,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0469  -0.9955  -0.2948   0.2961   5.8185
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.929257   0.054591  17.022  < 2e-16 ***
## hospital         0.217772   0.020176  10.793  < 2e-16 ***
## healthpoor       0.305013   0.048511   6.288 3.23e-10 ***
## healthexcellent -0.341807   0.060924  -5.610 2.02e-08 ***
## chronic          0.174916   0.012092  14.466  < 2e-16 ***
## gendermale      -0.126488   0.031216  -4.052 5.08e-05 ***
## school           0.026815   0.004394   6.103 1.04e-09 ***
## insuranceyes     0.224402   0.039464   5.686 1.30e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.2066) family taken to be 1)
##
##     Null deviance: 5743.7  on 4405  degrees of freedom
## Residual deviance: 5044.5  on 4398  degrees of freedom
## AIC: 24359
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.2066
##           Std. Err.:  0.0336
##
##  2 x log-likelihood:  -24341.1070
```

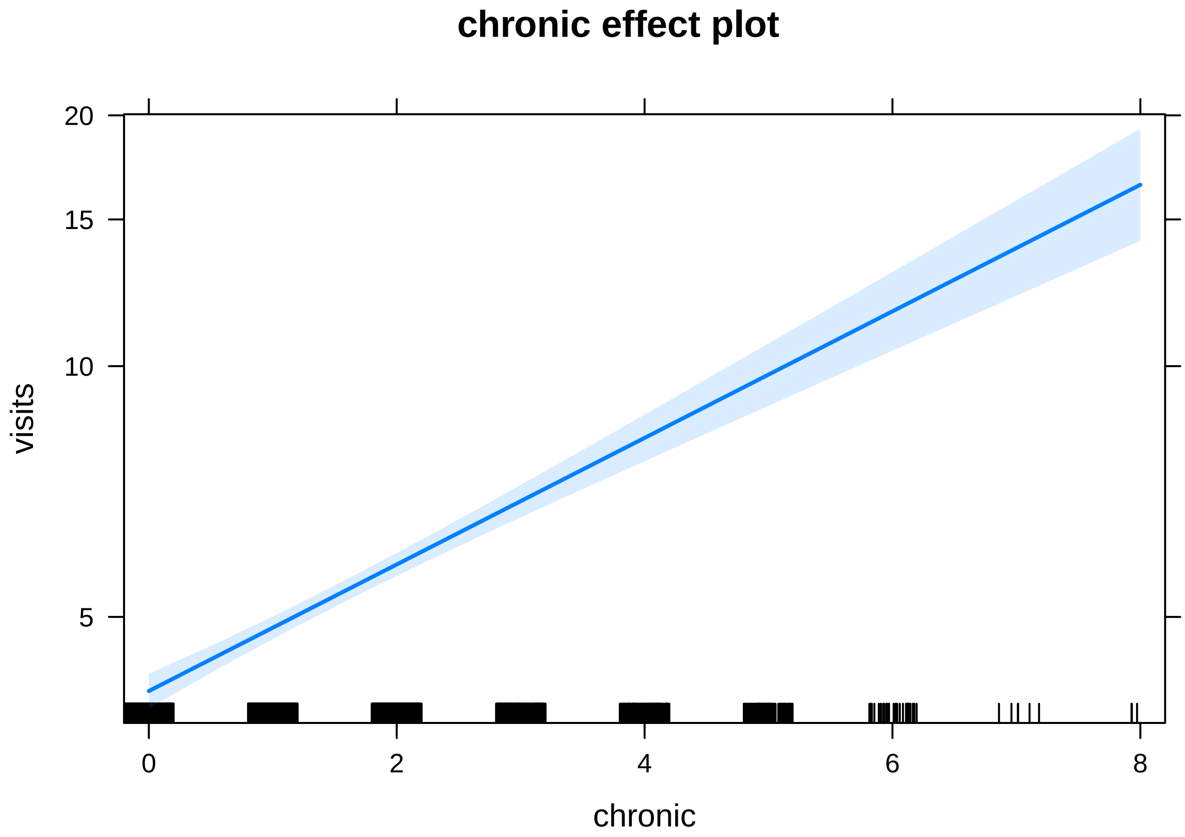
```
# You can compare the Poisson & negbin results side by side like this:
compareCoefs(vis.step, vis.nb)
```

```
##
## Call:
## 1: glm(formula = visits ~ chronic + hospital + school + health +
##   insurance + gender, family = poisson(), data = NMES1988)
## 2: glm.nb(formula = visits ~ hospital + health + chronic + gender +
##   school + insurance, data = NMES1988, init.theta = 1.206603534, link =
##   log)
##                    Est. 1     SE 1   Est. 2     SE 2
## (Intercept)       1.02887  0.02378  0.92926  0.05459
## chronic           0.14664  0.00458  0.17492  0.01209
## hospital          0.16480  0.00600  0.21777  0.02018
## school            0.02614  0.00184  0.02682  0.00439
## healthpoor        0.24831  0.01784  0.30501  0.04851
## healthexcellent -0.36199  0.03030 -0.34181  0.06092
## insuranceyes      0.20169  0.01686  0.22440  0.03946
## gendermale       -0.11232  0.01295 -0.12649  0.03122
```

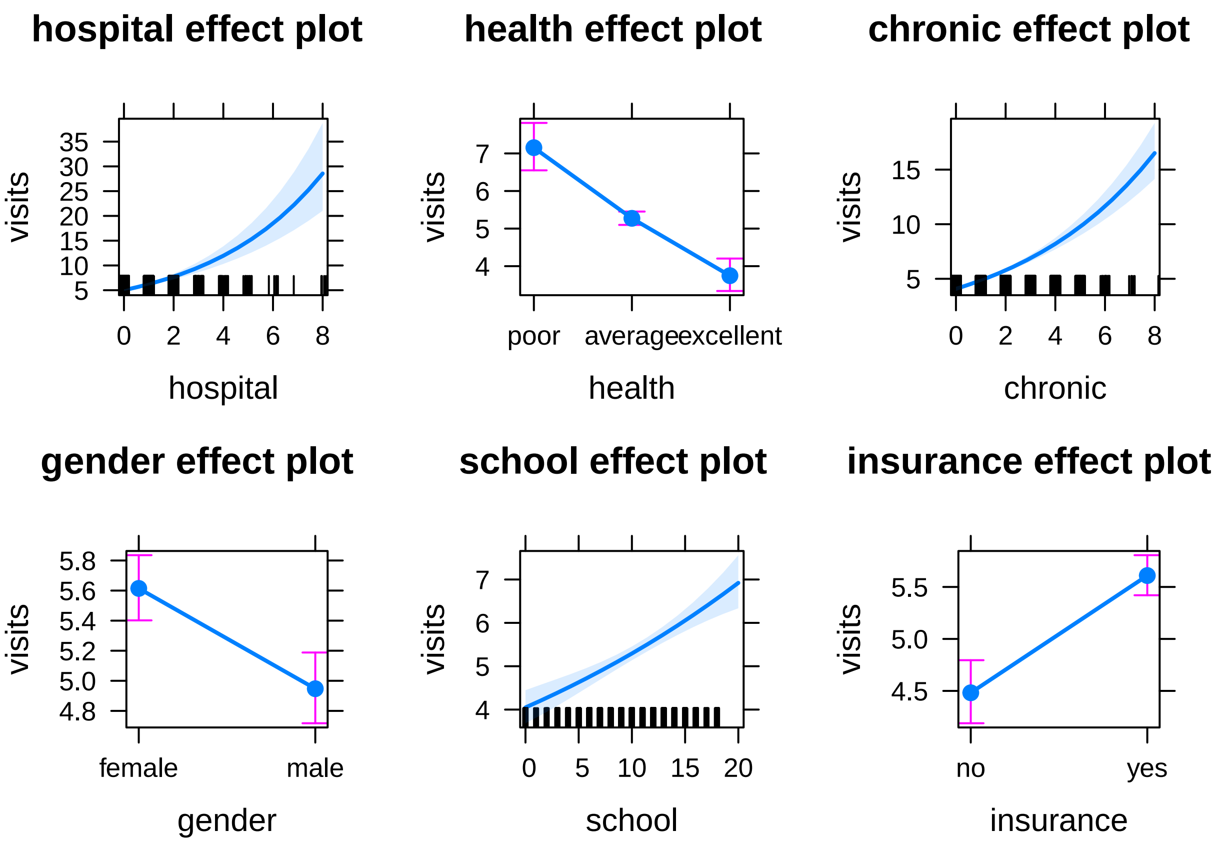
```
# Notice that all the standard errors are larger

# Interpretation
# Number of chronic conditions. This varies from 0 to 8.  You can do an effect plot 'by hand' like this

plot(0:8, exp(1.03 + (0:8) * 0.175))
```



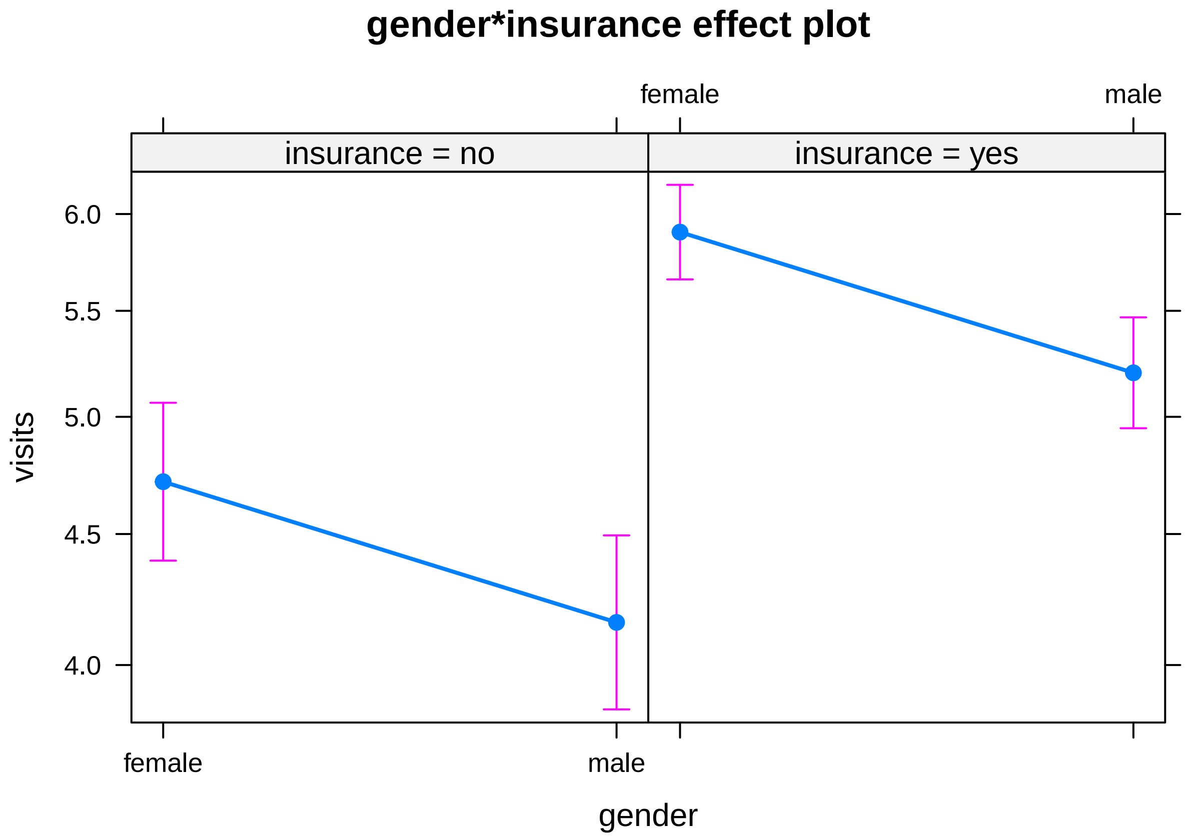
```
# Or using the effects package (which also adds means of other variables)

plot(Effect('chronic',  vis.nb))
```



```
# You can do them all at once like this:

plot(allEffects(vis.nb), type = 'response')
```



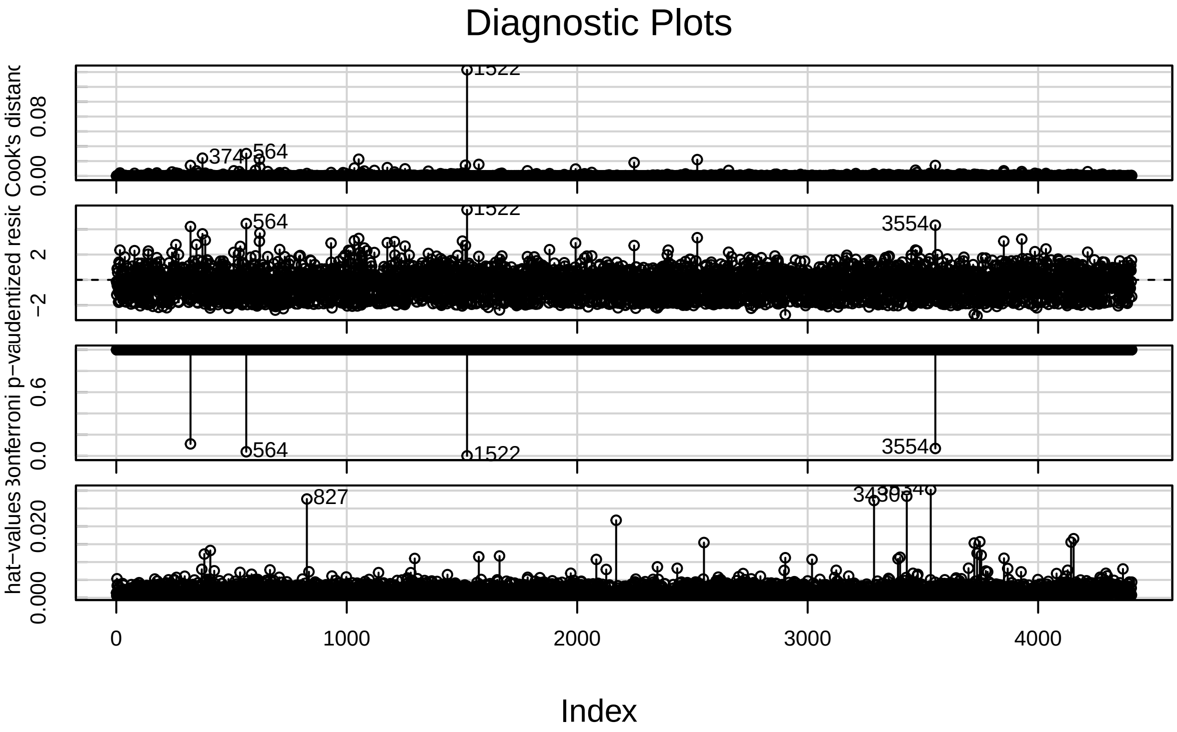
```
plot(Effect(c('gender','insurance'), vis.nb, multline = TRUE, type = 'response'))
```



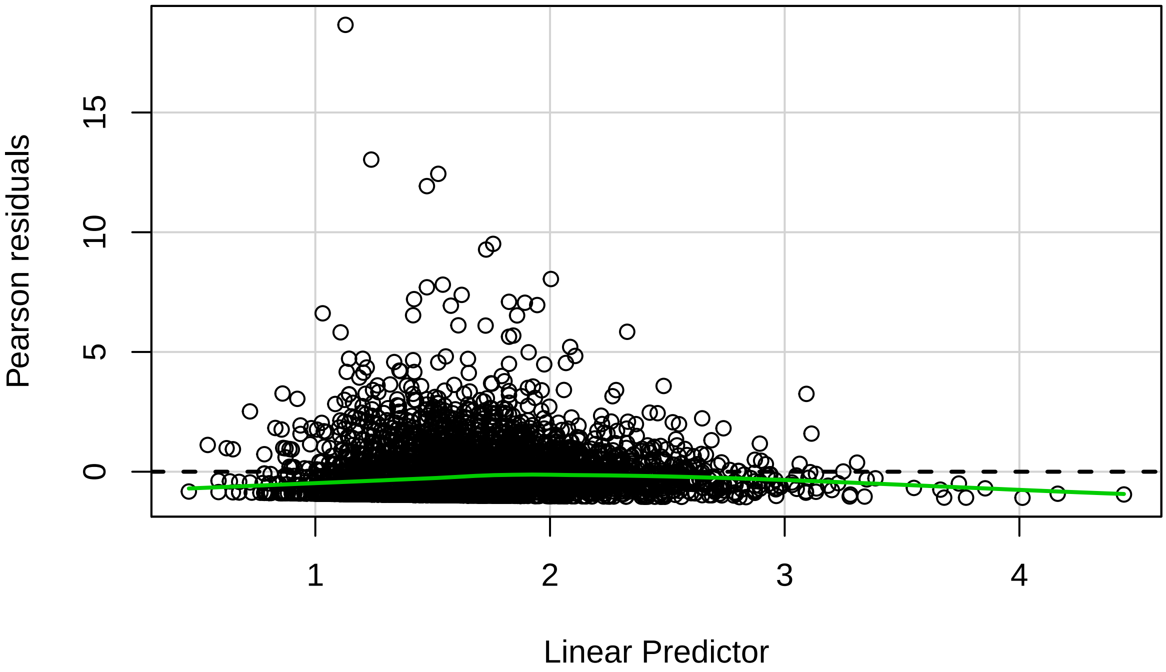
```
# So, for example we can see that women make on average 0.6 more visits than mens

# Check for outliers
outlierTest(vis.nb)
```

```
##      rstudent unadjusted p-value Bonferonni p
## 1522 5.541651         3.1715e-08   0.00013974
## 564  4.449918         8.8028e-06   0.03878500
```

```
influenceIndexPlot(vis.nb, id.n = 3)
```



```
residualPlot(vis.nb)
```


Pearson residuals
0
5
10
15
1
2
3
4
Linear Predictor