Homework 04

Generalized Linear Models

SHIYU ZHANG

October 5, 2017

Data analysis

Poisson regression:

The folder `risky.behavior` contains data from a randomized trial targeting couples at high risk of HIV infection. The intervention provided counseling sessions regarding practices that could reduce their likelihood of contracting HIV. Couples were randomized either to a control group, a group in which just the woman participated, or a group in which both members of the couple participated. One of the outcomes examined after three months was "number of unprotected sex acts".

1. Model this outcome as a function of treatment assignment using a Poisson regression. Does the model fit well? Is there evidence of overdispersion?

```
data1$fupacts <- round(data1$fupacts)
data1$couples <- factor(data1$couples)
data1$women_alone <- factor(data1$women_alone)

m1 <- glm(fupacts ~ women_alone, family=poisson, data=data1)
display(m1)
```

```
## glm(formula = fupacts ~ women_alone, family = poisson, data = data1)
##              coef.est coef.se
## (Intercept)   2.92     0.01
## women_alone1 -0.40     0.03
## ---
##   n = 434, k = 2
##   residual deviance = 13064.2, null deviance = 13298.6 (difference = 234.4)
```

```
summary(m1)
```

```
##
## Call:
## glm(formula = fupacts ~ women_alone, family = poisson, data = data1)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -6.093  -4.979  -3.304   1.237  27.150
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.92114    0.01368  213.58   <2e-16 ***
## women_alone1 -0.40367    0.02719  -14.84   <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: 13299  on 433  degrees of freedom
## Residual deviance: 13064  on 432  degrees of freedom
## AIC: 14393
##
## Number of Fisher Scoring iterations: 6
```

```
"the woman_alone factor appears to be statistically significant.
the model overall fits the data well."
```

```
## [1] "the woman_alone factor appears to be statistically significant.\nthe model overall fits the data
```

```
# to find the evidence of dispersion
library(AER)
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

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

```
##
## Attaching package: 'lmtest'
```

```
## The following object is masked from 'package:VGAM':
##
##     lrtest
```

```
## Loading required package: sandwich
```

```
## Loading required package: survival
```

```
##
## Attaching package: 'survival'
```

```
## The following objects are masked from 'package:faraway':
##
##     rats, solder
```

```
##
## Attaching package: 'AER'
```

```
## The following object is masked from 'package:VGAM':
##
##     tobit
```

```
dispersiontest(m1,trafo=1)
```

```
##
##  Overdispersion test
##
## data:  m1
## z = 4.9319, p-value = 4.072e-07
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##   alpha
## 41.9765
```

```
"from the dispersiontest output, we can say that the model has overdispersion"
```

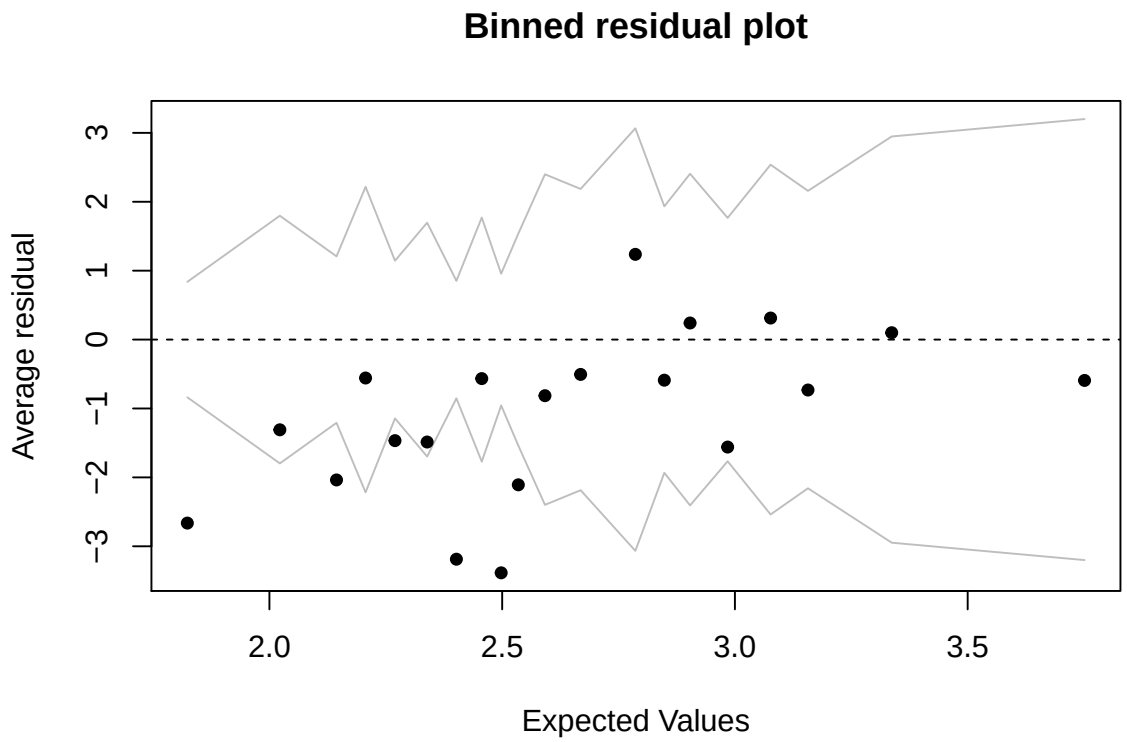
```
## [1] "from the dispersiontest output, we can say that the model has overdispersion"
```

2. Next extend the model to include pre-treatment measures of the outcome and the additional pre-treatment variables included in the dataset. Does the model fit well? Is there evidence of overdispersion?

```
# centralize the bupacts factor
data1$c.bupacts <- (data1$bupacts - mean(data1$bupacts)) / (2 * sd(data1$bupacts))
m2<- glm(fupacts ~ women_alone + sex + c.bupacts + couples + bs_hiv, family=poisson, data=data1)
display(m2)
```

```
## glm(formula = fupacts ~ women_alone + sex + c.bupacts + couples +
##     bs_hiv, family = poisson, data = data1)
##                coef.est coef.se
## (Intercept)     3.18     0.02
## women_alone1   -0.66     0.03
## sexman         -0.11     0.02
## c.bupacts       0.69     0.01
## couples1       -0.41     0.03
## bs_hivpositive -0.44     0.04
## ---
##   n = 434, k = 6
##   residual deviance = 10200.4, null deviance = 13298.6 (difference = 3098.2)
```

```
binnedplot(predict(m2), rstandard(m2))
```




```
" when the expected values is low, variance is much greater than 1, indicating overdispersion."
```

```
## [1] " when the expected values is low, variance is much greater than 1, indicating overdispersion."
```

```
library(AER)
dispersiontest(m2, trafo=1)
```

```
##
##  Overdispersion test
##
## data:  m2
## z = 5.5689, p-value = 1.282e-08
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##    alpha
## 28.65146
```

3. Fit an overdispersed Poisson model. What do you conclude regarding effectiveness of the intervention?

```
data1$c.bupacts <- (data1$bupacts - mean(data1$bupacts)) / (2 * sd(data1$bupacts))
m3<- glm(fupacts ~ women_alone + sex + c.bupacts + couples + bs_hiv, family=quasipoisson, data=data1)
display(m3)
```

```
## glm(formula = fupacts ~ women_alone + sex + c.bupacts + couples +
##     bs_hiv, family = quasipoisson, data = data1)
##                coef.est coef.se
## (Intercept)     3.18     0.12
## women_alone1   -0.66     0.17
## sexman         -0.11     0.13
## c.bupacts       0.69     0.06
## couples1       -0.41     0.15
## bs_hivpositive -0.44     0.19
## ---
##   n = 434, k = 6
##   residual deviance = 10200.4, null deviance = 13298.6 (difference = 3098.2)
##   overdispersion parameter = 30.0
```

```
"We can conclude that the intervention had a positive impact on decreasing unprotected sex happening.
for the women_alone coefficient, we can see an obviou decrease in unprotected sex acts of exp(0.66).
for the couples coefficient, it also shows a decrease impact (exp(0.41)) on the overall model "
```

```
## [1] "We can conclude that the intervention had a positive impact on decreasing unprotected sex happe
```

4. These data include responses from both men and women from the participating couples. Does this give you any concern with regard to our modeling assumptions?

```
"yes, the correlations between men and women should be much higher."
```

```
## [1] "yes, the correlations between men and women should be much higher."
```

Comparing logit and probit:

Take one of the data examples from Chapter 5. Fit these data using both logit and probit model. Check that the results are essentially the same (after scaling by factor of 1.6)

```
well <- read.table("http://www.stat.columbia.edu/~gelman/arm/examples/arsenic/wells.dat")
well$log.arsenic <- log(well$arsenic)
summary(well)
```

```
##      switch          arsenic           dist             assoc
##  Min.   :0.0000   Min.   :0.510   Min.   :  0.387   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.820   1st Qu.: 21.117   1st Qu.:0.0000
##  Median :1.0000   Median :1.300   Median : 36.761   Median :0.0000
```

```
##  Mean   :0.5752   Mean   :1.657   Mean   : 48.332   Mean   :0.4228
##  3rd Qu.:1.0000   3rd Qu.:2.200   3rd Qu.: 64.041   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :9.650   Max.   :339.531   Max.   :1.0000
##       educ         log.arsenic
##  Min.   : 0.000   Min.   :-0.6733
##  1st Qu.: 0.000   1st Qu.:-0.1985
##  Median : 5.000   Median : 0.2624
##  Mean   : 4.828   Mean   : 0.3139
##  3rd Qu.: 8.000   3rd Qu.: 0.7885
##  Max.   :17.000   Max.   : 2.2670
```

```
logit <- glm(switch ~ log.arsenic + dist + educ, family=binomial(link="logit"), data=well)
display(logit)
```

```
## glm(formula = switch ~ log.arsenic + dist + educ, family = binomial(link = "logit"),
##     data = well)
##             coef.est coef.se
## (Intercept)  0.32     0.08
## log.arsenic  0.89     0.07
## dist        -0.01     0.00
## educ         0.04     0.01
## ---
##   n = 3020, k = 4
##   residual deviance = 3878.2, null deviance = 4118.1 (difference = 239.9)
```

```
probit <- glm(switch ~ log.arsenic + dist + educ, family=binomial(link="probit"), data=well)
display(probit)
```

```
## glm(formula = switch ~ log.arsenic + dist + educ, family = binomial(link = "probit"),
##     data = well)
##             coef.est coef.se
## (Intercept)  0.19     0.05
## log.arsenic  0.54     0.04
## dist        -0.01     0.00
## educ         0.03     0.01
## ---
##   n = 3020, k = 4
##   residual deviance = 3878.3, null deviance = 4118.1 (difference = 239.8)
```

```
"From the two output of the two models, we can see that the coefficient of log.arsenic changes from 0.5
to 0.54, the coefficient of the distance remians the same (-0.01) and the one of education becomes 0.03
These are essentially the coefficients we would have scaling by 1.6 the coefficients of the logit model
```

```
## [1] "From the two output of the two models, we can see that the coefficient of log.arsenic changes fi
```

Comparing logit and probit:

construct a dataset where the logit and probit mod- els give different estimates.

Tobit model for mixed discrete/continuous data:

experimental data from the National Supported Work example are available in the folder `lalonde`. Use the treatment indicator and pre-treatment variables to predict post-treatment (1978) earnings using a tobit

model. Interpret the model coefficients.

- sample: 1 = NSW; 2 = CPS; 3 = PSID.
- treat: 1 = experimental treatment group (NSW); 0 = comparison group (either from CPS or PSID) - Treatment took place in 1976/1977.
- age = age in years
- educ = years of schooling
- black: 1 if black; 0 otherwise.
- hisp: 1 if Hispanic; 0 otherwise.
- married: 1 if married; 0 otherwise.
- nodegree: 1 if no high school diploma; 0 otherwise.
- re74, re75, re78: real earnings in 1974, 1975 and 1978
- educ_cat = 4 category education variable (1=<hs, 2=hs, 3=sm college, 4=college)

```
lalonde<-read.dta("http://www.stat.columbia.edu/~gelman/arm/examples/lalonde/NSW.dw.obs.dta")
```

Robust linear regression using the t model:

The csv file `congress` has the votes for the Democratic and Republican candidates in each U.S. congressional district in between 1896 and 1992, along with the parties' vote proportions and an indicator for whether the incumbent was running for reelection. For your analysis, just use the elections in 1986 and 1988 that were contested by both parties in both years.

```
congress<-read.csv("congress(1).csv",header=TRUE)
```

1. Fit a linear regression (with the usual normal-distribution model for the errors) predicting 1988 Democratic vote share from the other variables and assess model fit.

```
Dem=congress$Dem_vote
Pct=congress$Dem_pct
x1=congress$x1
x2=congress$x2
Rep=congress$Rep_vote
a1= glm (Pct ~ x1+x2+Rep+Dem, family=binomial(link="logit"))
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial
## glm!
```

```
display(a1)
```

```
## glm(formula = Pct ~ x1 + x2 + Rep + Dem, family = binomial(link = "logit"))
##             coef.est coef.se
## (Intercept) 0.19     0.04
## x1          0.01     0.00
## x2          0.00     0.00
## Rep         0.00     0.00
## Dem         0.00     0.00
## ---
##   n = 19984, k = 5
##   residual deviance = 2949.3, null deviance = 5870.2 (difference = 2920.8)
```

```
"the overall model's p-value is less than 0.05 which stands for
statistically significant. the model fits well."
```

```
## [1] "the overall model's p-value is less than 0.05 which stands for \nstatistically significant. the
```

2. Fit a t-regression model predicting 1988 Democratic vote share from the other variables and assess model fit; to fit this model in R you can use the `vglm()` function in the VGLM package or `tlm()` function in the hett package.

```
#a2=vglm(Pct~x1+x2+Rep+Dem, family=poisson, data=congress)
```

3. Which model do you prefer?

"from the deviance output, a2 (the second model) was much less than the first model. however, from the p-value for both models, the second model variable x2 has a p-value greater than 0.05. so i prefer the first model (a1)"

Robust regression for binary data using the robit model:

Use the same data as the previous example with the goal instead of predicting for each district whether it was won by the Democratic or Republican candidate.

1. Fit a standard logistic or probit regression and assess model fit.

```
b1= glm (Pct ~ x1+x2+Rep+Dem, family=binomial(link="logit"))
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial
## glm!
```

```
display(b1)
```

```
## glm(formula = Pct ~ x1 + x2 + Rep + Dem, family = binomial(link = "logit"))
##             coef.est coef.se
## (Intercept) 0.19     0.04
## x1          0.01     0.00
## x2          0.00     0.00
## Rep         0.00     0.00
## Dem         0.00     0.00
## ---
##   n = 19984, k = 5
##   residual deviance = 2949.3, null deviance = 5870.2 (difference = 2920.8)
```

```
b2= glm (Pct ~ x1+x2+Rep+Dem, family=binomial(link="probit"))
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial
## glm!
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
display(b2)
```

```
## glm(formula = Pct ~ x1 + x2 + Rep + Dem, family = binomial(link = "probit"))
##             coef.est coef.se
## (Intercept) 0.07     0.02
## x1          0.01     0.00
## x2          0.00     0.00
## Rep         0.00     0.00
## Dem         0.00     0.00
## ---
##   n = 19984, k = 5
##   residual deviance = 3086.5, null deviance = 5870.2 (difference = 2783.6)
```

```
"both of the models are statistically significant in general, from the deviance output,
first model(logit model) was less than the second model(probit). so the first model is
better."
```

```
## [1] "both of the models are statistically significant in general, from the deviance output,\nfirst m
```

2. Fit a robit regression and assess model fit.

```
b3= rlm (Pct ~ x1+x2+Rep+Dem)
summary(b3)
```

```
##
## Call: rlm(formula = Pct ~ x1 + x2 + Rep + Dem)
## Residuals:
##       Min        1Q    Median        3Q       Max
## -2.004785 -0.059940 -0.007836  0.062136  3.545832
##
## Coefficients:
##             Value     Std. Error t value
## (Intercept)    0.5349    0.0018   292.6222
## x1             0.0014    0.0000    33.2542
## x2            -0.0004    0.0001    -7.5222
## Rep            0.0000    0.0000  -212.9461
## Dem            0.0000    0.0000   164.0377
##
## Residual standard error: 0.08975 on 19979 degrees of freedom
##   (1327 observations deleted due to missingness)
```

3. Which model do you prefer?

```
#i prefer the first model, becasue the output from the first model is more
#clear and informative than the later one.""
```

Salmonellla

The `salmonella` data was collected in a salmonella reverse mutagenicity assay. The predictor is the dose level of quinoline and the response is the numbers of revertant colonies of TA98 salmonella observed on each of three replicate plates. Show that a Poisson GLM is inadequate and that some overdispersion must be allowed for. Do not forget to check out other reasons for a high deviance.

```
data(salmonella)
?salmonella
```

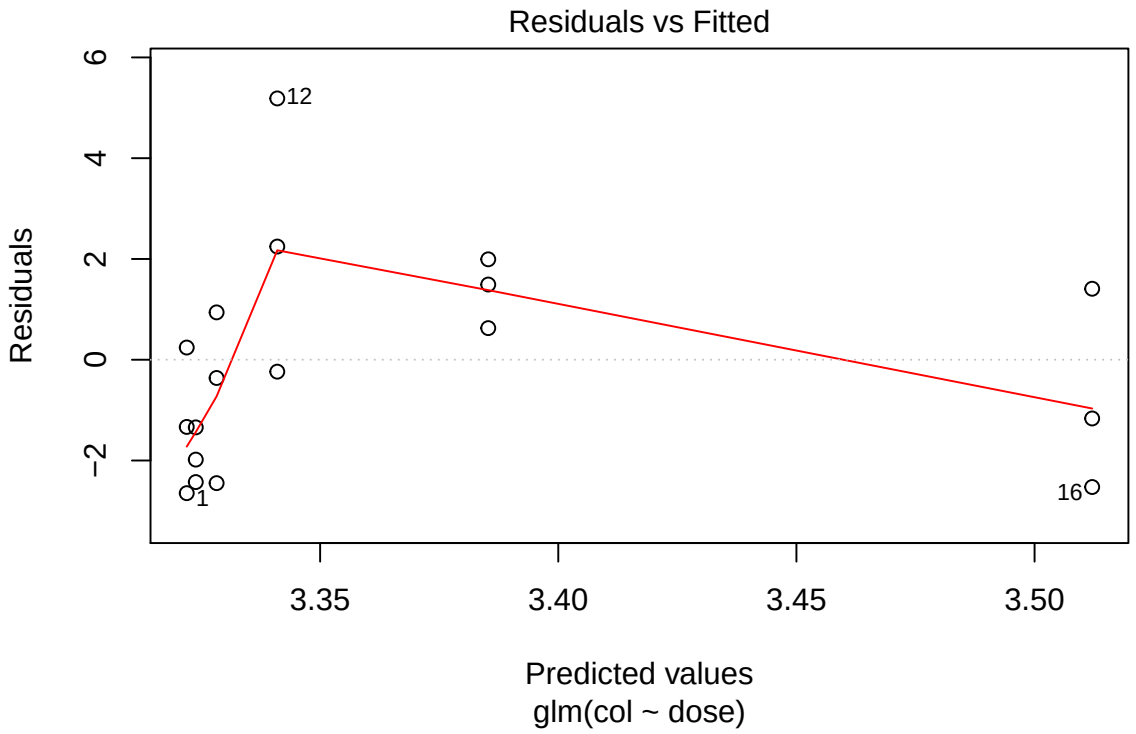
When you plot the data you see that the number of colonies as a function of dose is not monotonic especially around the dose of 1000.

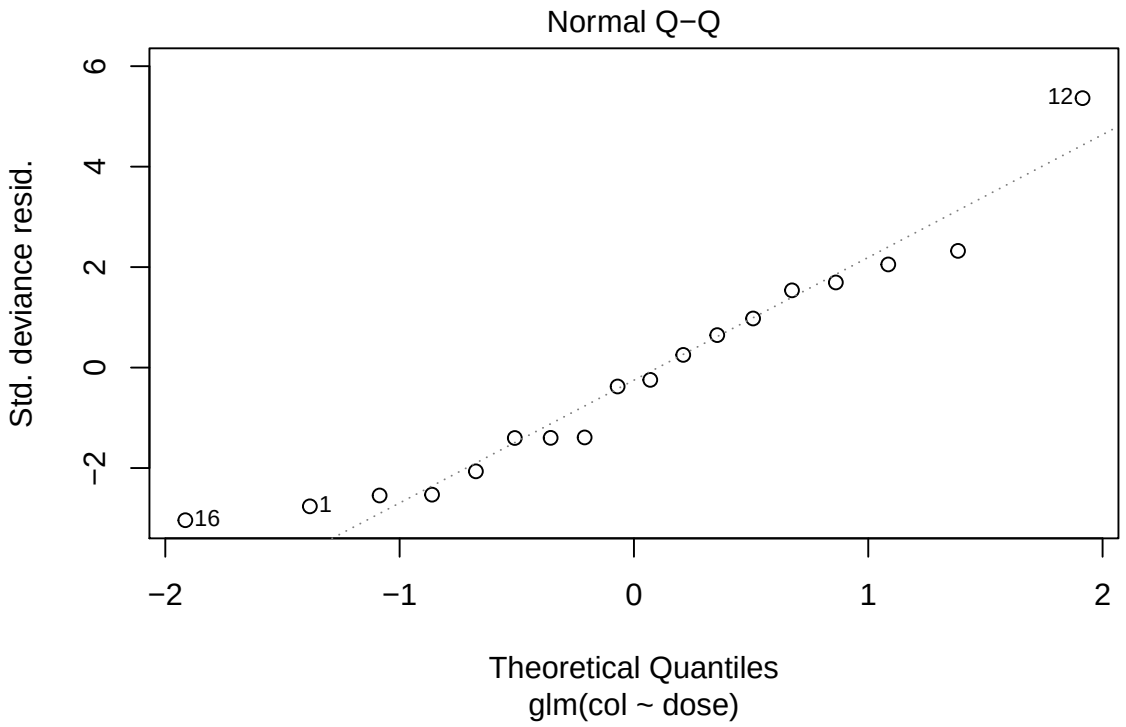
```
col=salmonella$colonies
dose=salmonella$dose
c1=glm(col~dose, data=salmonella, family=poisson)
display(c1)
```

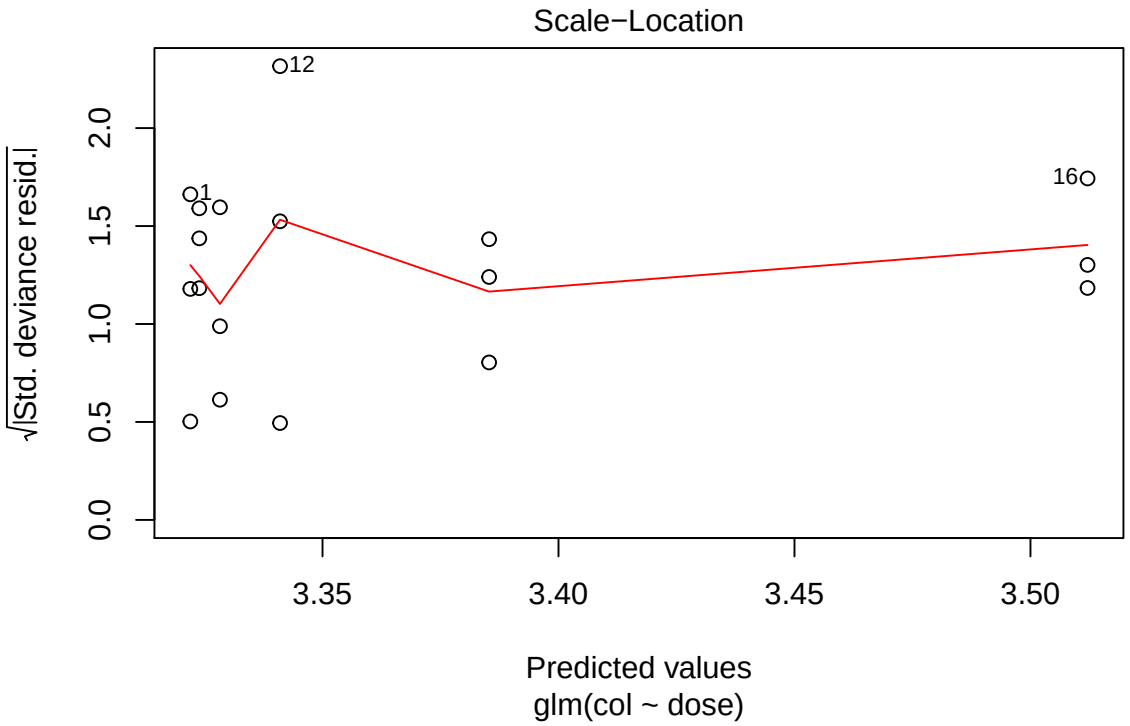
```
## glm(formula = col ~ dose, family = poisson, data = salmonella)
##             coef.est coef.se
## (Intercept) 3.32     0.05
## dose        0.00     0.00
## ---
```

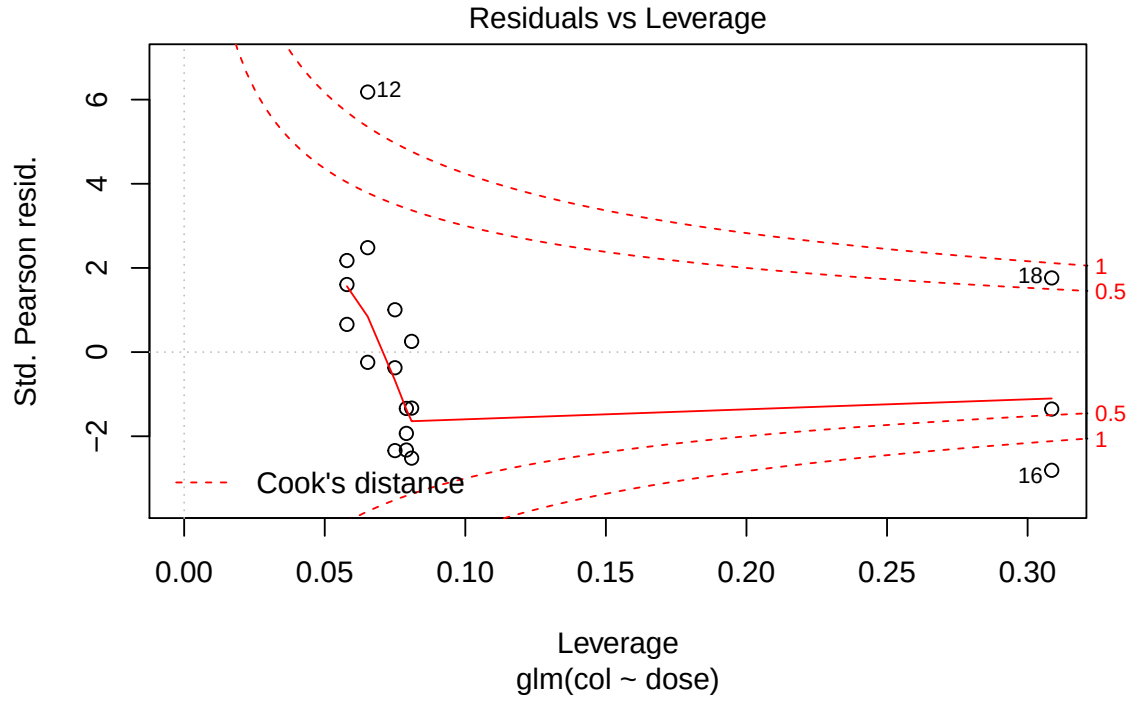
```
##    n = 18, k = 2
##    residual deviance = 75.8, null deviance = 78.4 (difference = 2.6)
```

```
plot(c1)
```









Since we are fitting log linear model we should look at the data on log scale. Also becase the dose is not equally spaced on the raw scale it may be better to plot it on the log scale as well.

```
dose_log<-log(dose+1)
col_log<-log(col+1)
c2=glm(col_log ~ dose_log,family=poisson, data=salmonella)
```

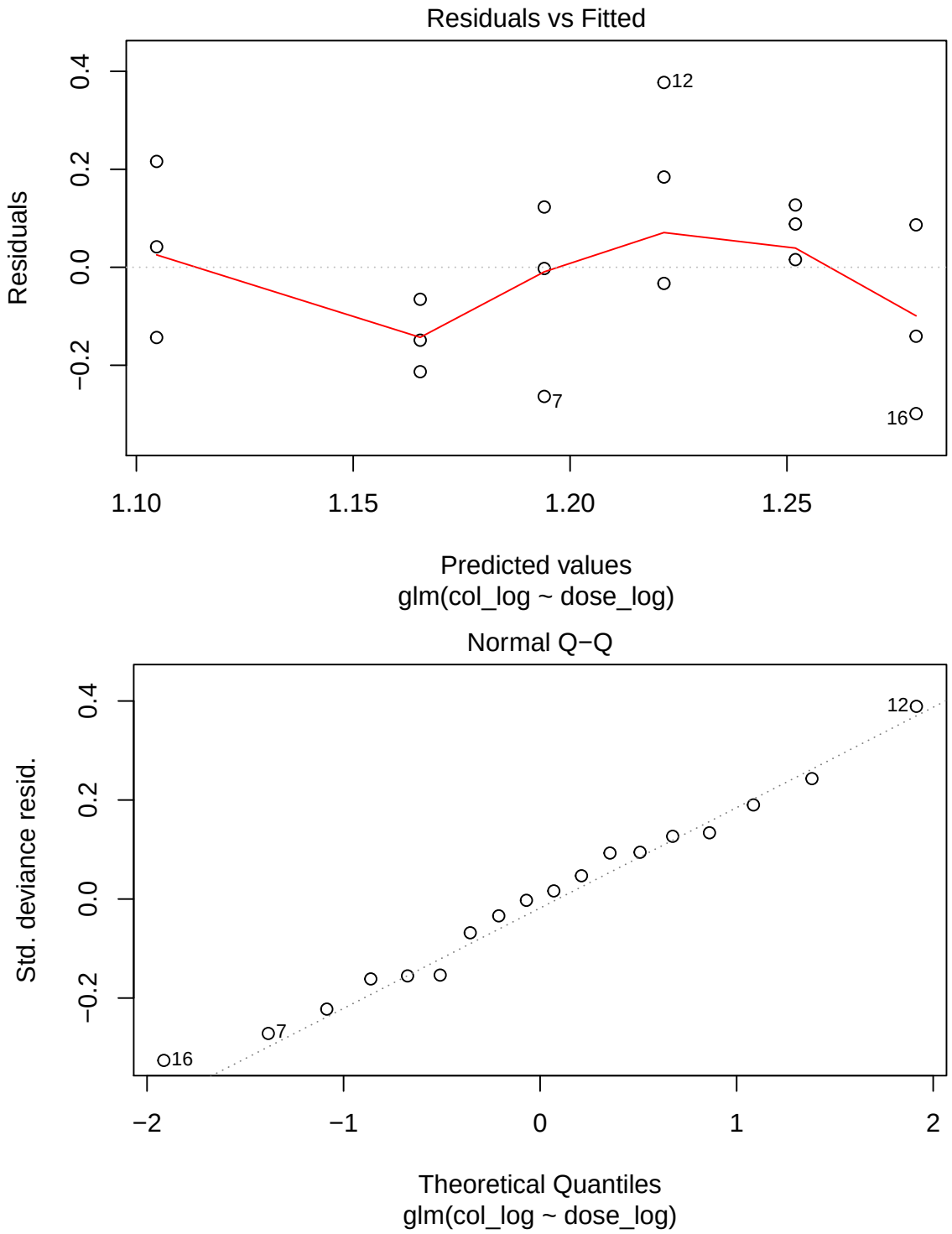
```
## Warning in dpois(y, mu, log = TRUE): non-integer x = 2.772589
```

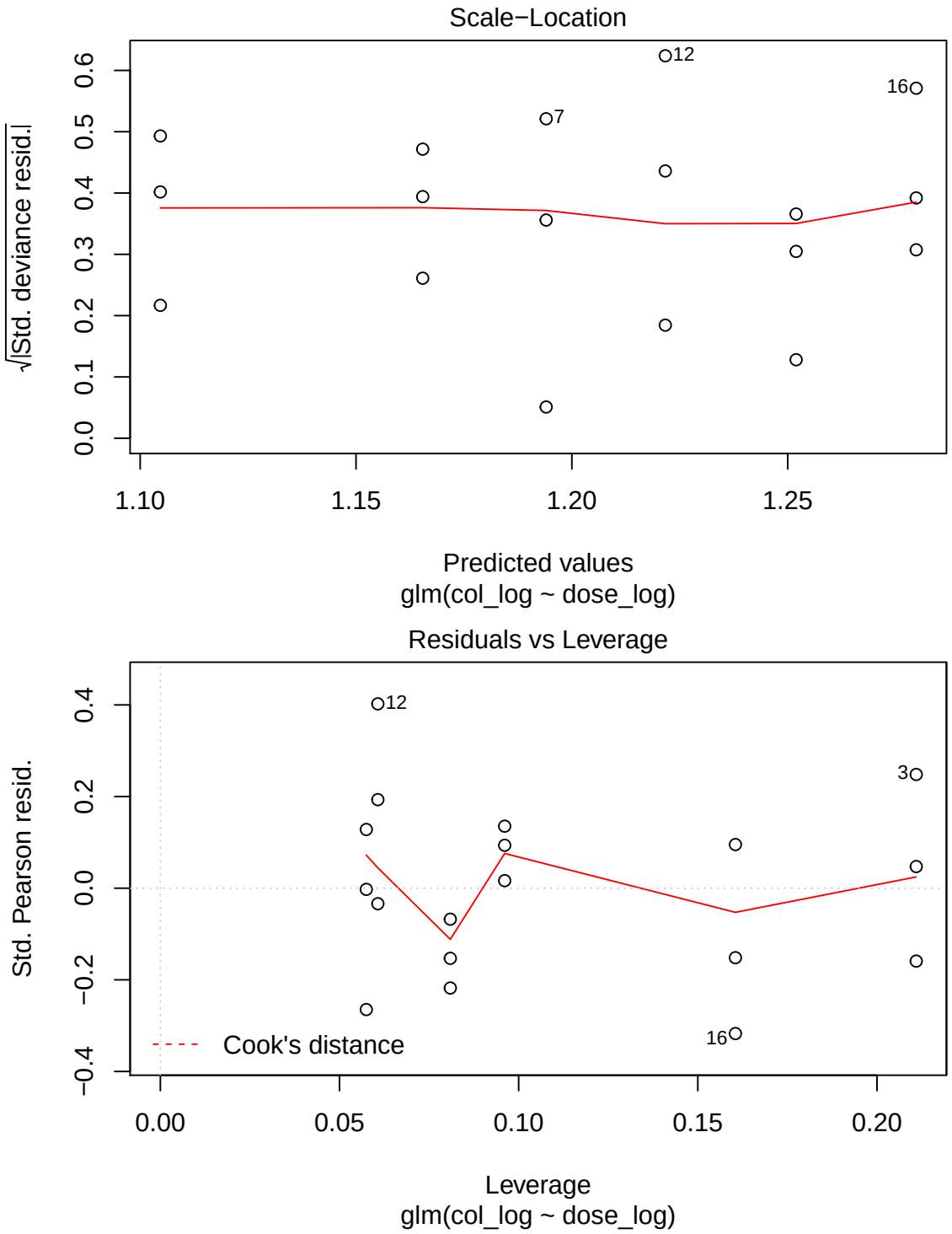
```
## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.091042
## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.401197
## Warning in dpois(y, mu, log = TRUE): non-integer x = 2.833213
## Warning in dpois(y, mu, log = TRUE): non-integer x = 2.944439
## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.091042
## Warning in dpois(y, mu, log = TRUE): non-integer x = 2.833213
## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.295837
## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.526361
## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.332205
## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.737670
## Warning in dpois(y, mu, log = TRUE): non-integer x = 4.110874
## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.526361
## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.663562
## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.737670
## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.044522
## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.332205
## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.761200
```

```
display(c2)
```

```
## glm(formula = col_log ~ dose_log, family = poisson, data = salmonella)
##             coef.est coef.se
## (Intercept) 1.10     0.26
## dose_log    0.03     0.06
## ---
##   n = 18, k = 2
##   residual deviance = 0.5, null deviance = 0.7 (difference = 0.2)
```

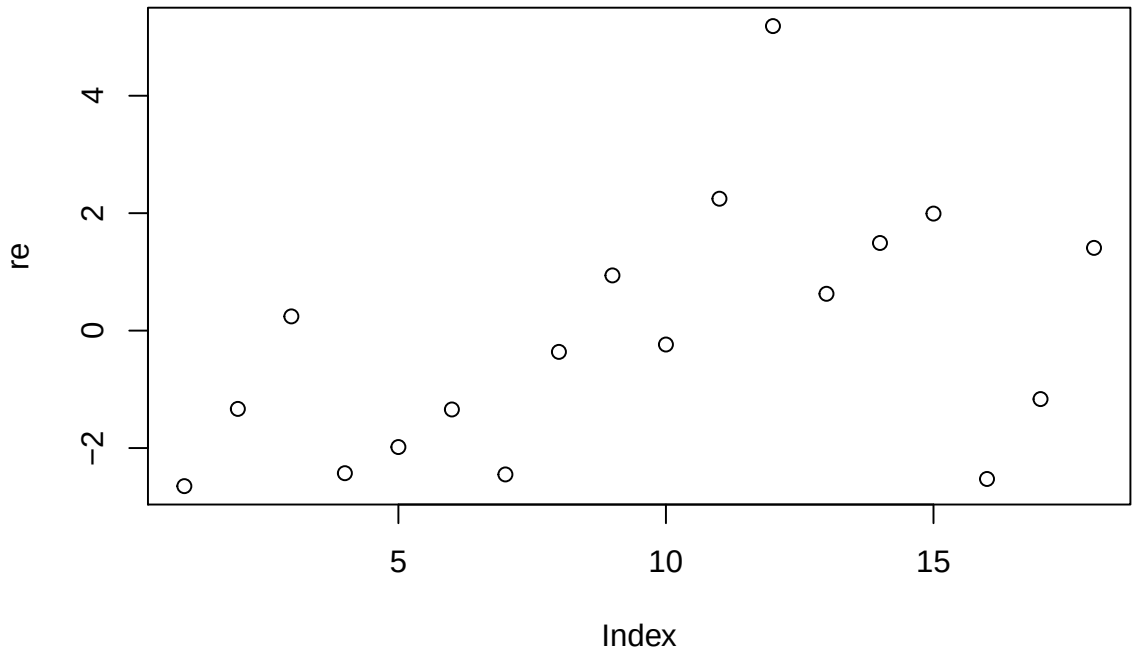
```
plot(c2)
```


Residuals vs Fitted
Residuals
0.4
0.2
0.0
−0.2
12
7
16
1.10
1.15
1.20
1.25
Predicted values
glm(col_log ~ dose_log)
Normal Q−Q
Std. deviance resid.
0.4
0.2
0.0
−0.2
12
7
16
−2
−1
0
1
2
Theoretical Quantiles
glm(col_log ~ dose_log)



This shows that the trend is not monotonic. Hence when you fit the model and look at the residual you will see a trend.

```
re=resid(c1)
plot(re)
```



The lack of fit is also evident if we plot the fitted line onto the data.

```
#from the output, the residuals don't lies around a linear trend.
```

How do we adress this problem? The serious problem to address is the nonlinear trend of dose ranther than the overdispersion since the line is missing the points. Let's add a beny line with 4th order polynomial.

```
#we can solve this problem by collecting more sample datas.
```

The resulting residual looks nice and if you plot it on the raw data. Whether the trend makes real contextual sense will need to be validated but for the given data it looks feasible.

Dispite the fit, the overdispersion still exists so we'd be better off using the quasi Poisson model.

```
c4=glm(col~dose, data=salmonella, family = quasipoisson)
summary(c4)
```

```
##
## Call:
## glm(formula = col ~ dose, family = quasipoisson, data = salmonella)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6482  -1.8225  -0.2993   1.2917   5.1861
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.3219950  0.1218628  27.260 7.72e-15 ***
## dose        0.0001901  0.0002644   0.719    0.482
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 5.087279)
##
##     Null deviance: 78.358  on 17  degrees of freedom
## Residual deviance: 75.806  on 16  degrees of freedom
```

```
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

Ships

The `ships` dataset found in the MASS package gives the number of damage incidents and aggregate months of service for different types of ships broken down by year of construction and period of operation.

```
data(ships)
?ships
```

Develop a model for the rate of incidents, describing the effect of the important predictors.

```
d1=lm(incidents~type+year+period+service, data=ships)
summary(d1)
```

```
##
## Call:
## lm(formula = incidents ~ type + year + period + service, data = ships)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.5549  -3.4577  -0.0849   2.3616  16.6209
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.746e+01  1.856e+01  -1.479   0.1489
## typeB        8.622e+00  4.926e+00   1.750   0.0897 .
## typeC       -3.296e+00  3.662e+00  -0.900   0.3748
## typeD       -2.430e+00  3.663e+00  -0.663   0.5119
## typeE       -6.495e-01  3.663e+00  -0.177   0.8604
## year         1.528e-01  2.196e-01   0.696   0.4915
## period       3.124e-01  1.547e-01   2.020   0.0519 .
## service      1.102e-03  2.047e-04   5.385 6.47e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.322 on 32 degrees of freedom
## Multiple R-squared:  0.8035, Adjusted R-squared:  0.7605
## F-statistic: 18.69 on 7 and 32 DF,  p-value: 1.182e-09
```

```
"from the summary output, we can tell that the p-values of all the variables
(type,year,period,service) are greater than 0.05, which stands for not
statistically significant result. so we can say that these factors don't have
a strong impact on the rate of inccidents"
```

```
## [1] "from the summary output, we can tell that the p-values of all the variables\n(type,year,period,s
```

Australian Health Survey

The `dvisits` data comes from the Australian Health Survey of 1977-78 and consist of 5190 single adults where young and old have been oversampled.

```
data(dvisits)
?dvisits
```

1. 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?

```
e1=glm(doctorco~sex+age+agesq+income+levyplus+freepoor+freerepa+illness+actdays
       +hscore+chcond1+chcond2,family=poisson,data=dvisits)
summary(e1)
```

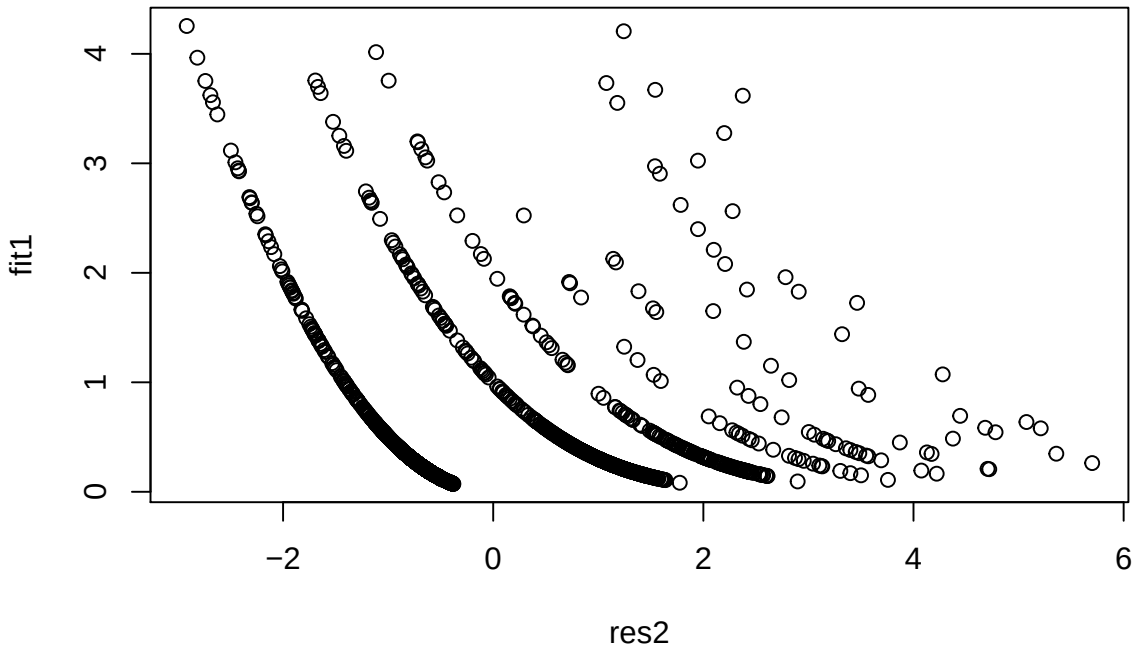
```
##
## Call:
## glm(formula = doctorco ~ sex + age + agesq + income + levyplus +
##     freepoor + freerepa + illness + actdays + hscore + chcond1 +
##     chcond2, family = poisson, data = dvisits)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9170  -0.6862  -0.5743  -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
```

```
"from the deviance as well as the p-value output , the model doesn't fit the data well "
```

```
## [1] "from the deviance as well as the p-value output , the model doesn't fit the data well "
```

2. Plot the residuals and the fitted values-why are there lines of observations on the plot?

```
fit1=fitted(e1)
res2=resid(e1)
plot(res2,fit1)
```



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

```
step(e1)
```

```
## 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
##    -2.08906      0.16200      0.35513     -0.19981      0.08369
##    freepoor      illness      actdays       hscore      chcond1
##    -0.46960      0.18610      0.12661      0.03112      0.12110
##     chcond2
##     0.15889
##
## Degrees of Freedom: 5189 Total (i.e. Null);  5179 Residual
## Null Deviance:       5635
## Residual Deviance: 4381  AIC: 6735
```

```
e2=glm(formula = doctorco ~ sex + age + income + levyplus + freepoor + illness + actdays
       + hscore + chcond1 + chcond2, family = poisson, data = dvisits)
summary(e2)
```

```
##
## Call:
## glm(formula = doctorco ~ sex + age + income + levyplus + freepoor +
##     illness + actdays + hscore + chcond1 + chcond2, family = poisson,
##     data = dvisits)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0004  -0.6851  -0.5761  -0.4858   5.7284
##
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.089063   0.100811 -20.723  < 2e-16 ***
## sex          0.162000   0.055824   2.902  0.00371 **
## age          0.355131   0.143196   2.480  0.01314 *
## income      -0.199806   0.084328  -2.369  0.01782 *
## levyplus     0.083689   0.053544   1.563  0.11805
## freepoor    -0.469596   0.176360  -2.663  0.00775 **
## illness      0.186101   0.018260  10.191  < 2e-16 ***
## actdays      0.126611   0.005029  25.177  < 2e-16 ***
## hscore       0.031116   0.010065   3.092  0.00199 **
## chcond1      0.121100   0.066389   1.824  0.06814 .
## chcond2      0.158894   0.081762   1.943  0.05197 .
## ---
## 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: 4381.0  on 5179  degrees of freedom
## AIC: 6734.5
##
## Number of Fisher Scoring iterations: 6
```

```
"reject the variables whose p-value are greater than 0.05 - the levyplus, chcond1
and chcond2 variables"
```

```
## [1] "reject the variables whose p-value are greater than 0.05 - the levyplus, chcond1\nand chcond2 va
```

```
e3=glm(formula = doctorco ~ sex + age + income + freepoor + illness + actdays
       + hscore, family = poisson, data = dvisits)
summary(e3)
```

```
##
## Call:
## glm(formula = doctorco ~ sex + age + income + freepoor + illness +
##     actdays + hscore, family = poisson, data = dvisits)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9258  -0.6829  -0.5752  -0.4945   5.6960
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.051963   0.099522 -20.618  < 2e-16 ***
## sex          0.175529   0.055433   3.167  0.00154 **
## age          0.433532   0.137140   3.161  0.00157 **
## income      -0.171053   0.081926  -2.088  0.03681 *
## freepoor    -0.496325   0.175304  -2.831  0.00464 **
## illness      0.196008   0.017585  11.146  < 2e-16 ***
## actdays      0.127793   0.004899  26.088  < 2e-16 ***
## hscore       0.032433   0.009938   3.263  0.00110 **
## ---
## 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: 4388.1  on 5182  degrees of freedom
## AIC: 6735.7
##
## Number of Fisher Scoring iterations: 6
```

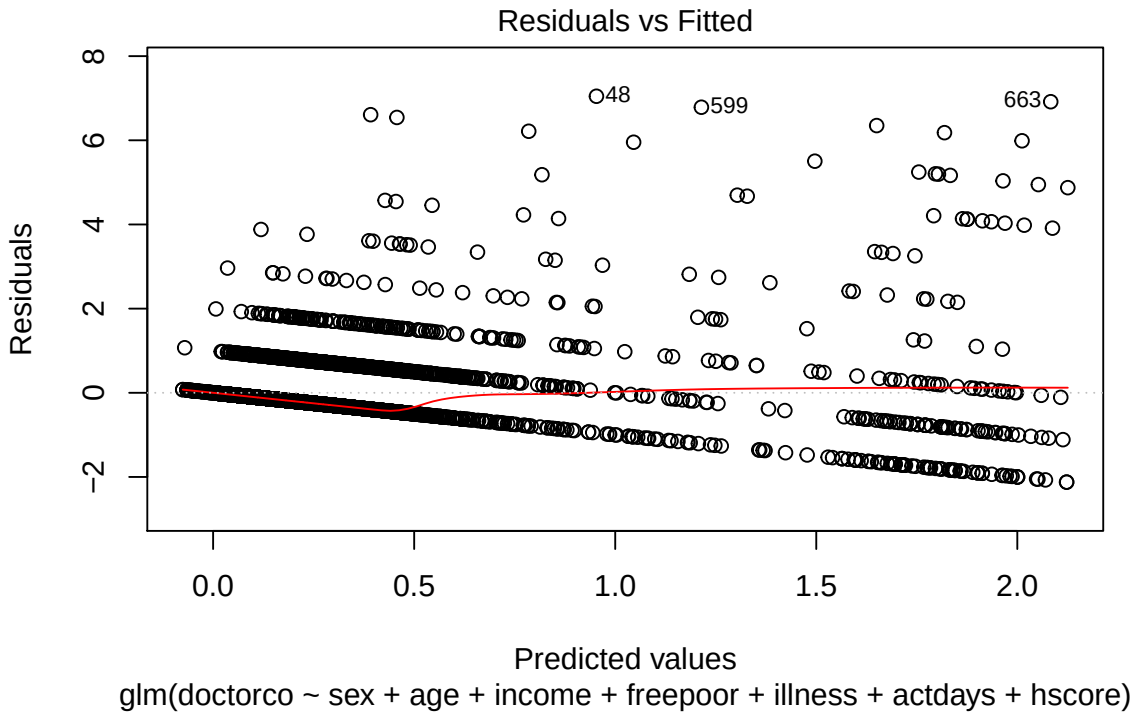
4. For the last person in the dataset, compute the predicted probability distribution for their visits to the doctor, i.e., give the probability they visit 0,1,2, etc. times.

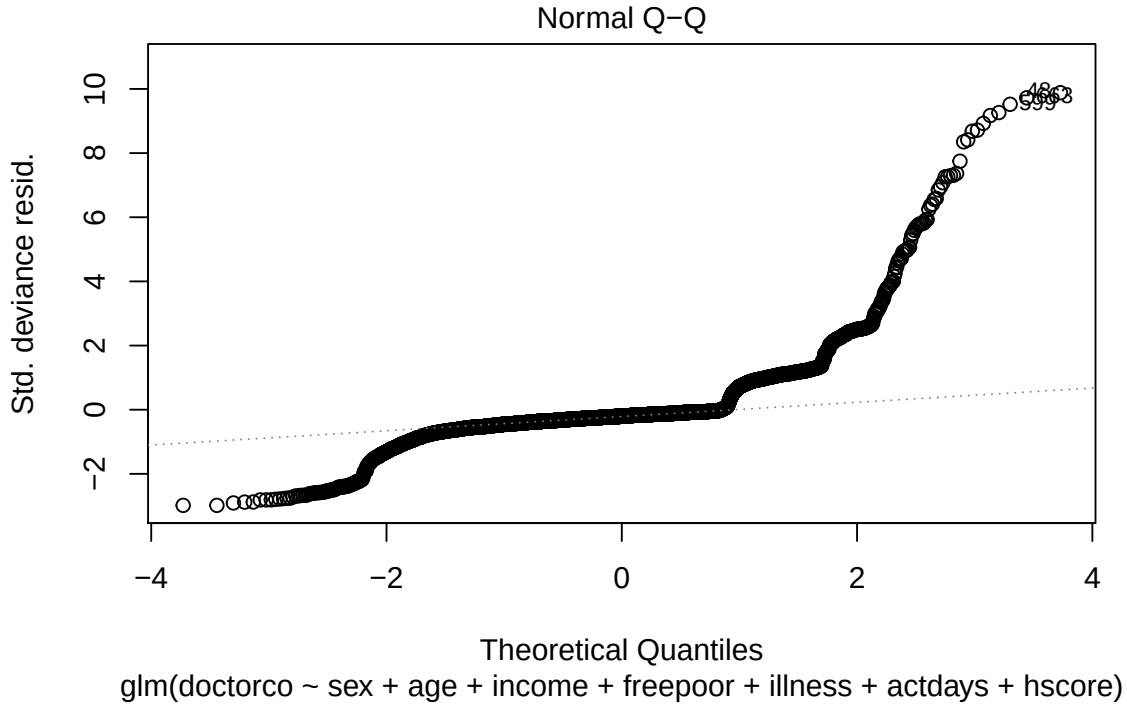
5. Fit a comparable (Gaussian) linear model and graphically compare the fits. Describe how they differ.

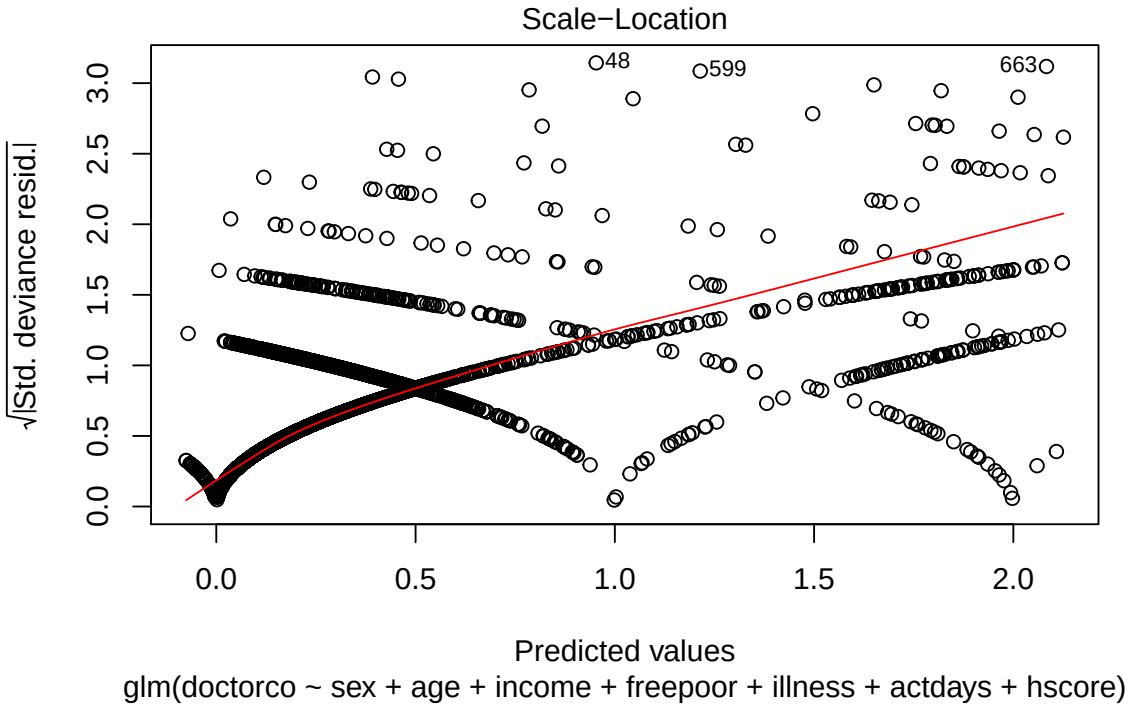
```
e4=glm(formula = doctorco ~ sex + age + income + freepoor + illness + actdays
       + hscore, family = gaussian,data = dvisits)
summary(e4)
```

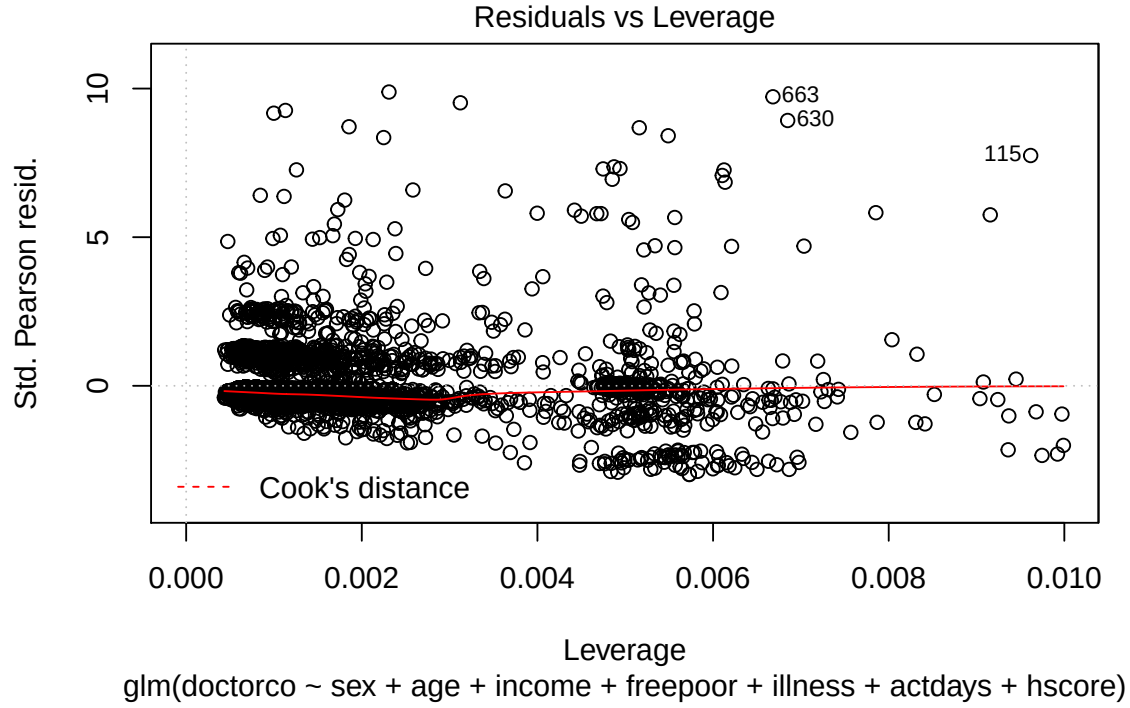
```
##
## Call:
## glm(formula = doctorco ~ sex + age + income + freepoor + illness +
##     actdays + hscore, family = gaussian, data = dvisits)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1226  -0.2586  -0.1456  -0.0453   7.0467
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.042149   0.035493   1.188 0.235074
## sex          0.038470   0.021261   1.809 0.070449 .
## age          0.186468   0.053427   3.490 0.000487 ***
## income      -0.052481   0.029458  -1.782 0.074880 .
## freepoor    -0.119439   0.051007  -2.342 0.019237 *
## illness      0.061894   0.007941   7.794 7.79e-15 ***
## actdays      0.103803   0.003614  28.720  < 2e-16 ***
## hscore       0.017575   0.005157   3.408 0.000660 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.5095055)
##
##     Null deviance: 3305.5  on 5189  degrees of freedom
## Residual deviance: 2640.3  on 5182  degrees of freedom
## AIC: 11239
##
## Number of Fisher Scoring iterations: 2
```

```
plot(e4)
```


Residuals vs Fitted
Residuals
Predicted values
glm(doctorco ~ sex + age + income + freepoor + illness + actdays + hscore)
48
599
663


Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
glm(doctorco ~ sex + age + income + freepoor + illness + actdays + hscore)


Scale−Location
48
599
663
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.5
1.0
1.5
2.0
Predicted values
glm(doctorco ~ sex + age + income + freepoor + illness + actdays + hscore)


Residuals vs Leverage
663
630
115
Std. Pearson resid.
10
5
0
Cook's distance
0.000
0.002
0.004
0.006
0.008
0.010
Leverage
glm(doctorco ~ sex + age + income + freepoor + illness + actdays + hscore)