Homework 04

Generalized Linear Models

Name

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?

```
#I build a poisson regression between women_alone and fupacts and display it

risky_behaviors$fupacts <- round(risky_behaviors$fupacts)
fitsex <- glm(formula = fupacts ~ factor(women_alone), family = poisson, data = risky_behaviors)
display(fitsex)
```

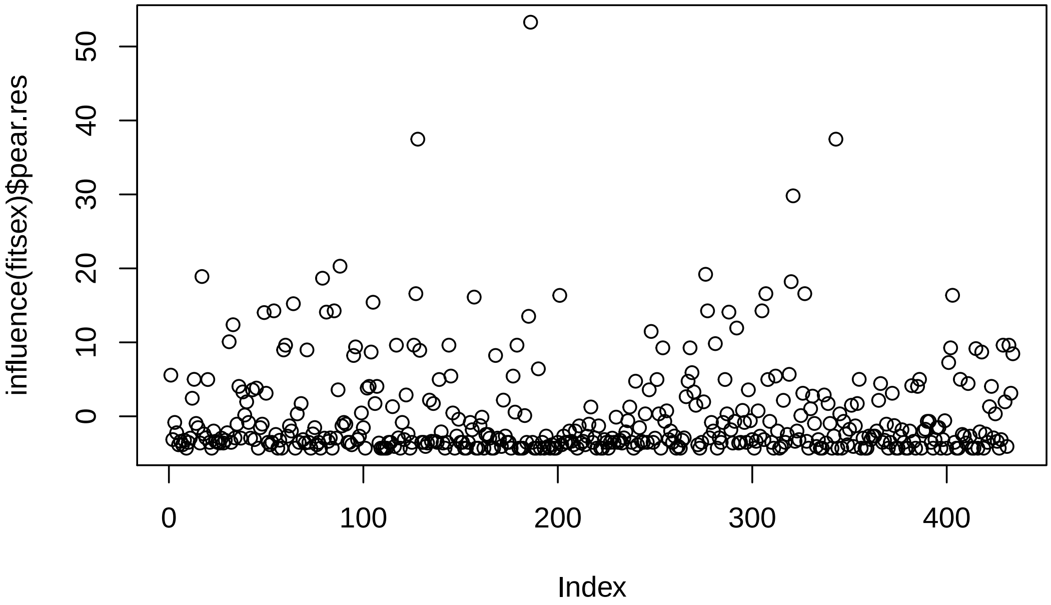
```
## glm(formula = fupacts ~ factor(women_alone), family = poisson,
##     data = risky_behaviors)
##                       coef.est coef.se
## (Intercept)            2.92     0.01
## factor(women_alone)1 -0.40     0.03
## ---
##   n = 434, k = 2
##   residual deviance = 13064.2, null deviance = 13298.6 (difference = 234.4)
```

```
# first i choose p-value of chi-square of fitsex to show

1 - pchisq(13064, 432)
```

```
## [1] 0
```

```
plot(influence(fitsex)$pear.res)
```



```r
#I do Anova to show whether this variable work in this formula

Anova(fitsex)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: fupacts
##                     LR Chisq Df Pr(>Chisq)
## factor(women_alone)   234.43  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# I do the mean and variance of outcome and input variable. I found that the mean and variance has huge

tapply(risky_behaviors$fupacts,risky_behaviors$women_alone, function(x)c(mean = mean(x),variance = var(
```

```
## $`0`
##     mean variance
##  18.5625 802.9229
##
## $`1`
##      mean  variance
##  12.39726 533.30316
```

There exist some overdispersions in this model

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?

```r
# I change the column name "sex" to "man" and change woman = 0 and man =1 and change True of "bs_hiv" t

risky_behaviors$sex <-as.numeric(risky_behaviors$sex)
```

```r
colnames(risky_behaviors)[1] <- "man"
risky_behaviors$man <- risky_behaviors$man -1

risky_behaviors$bs_hiv <-as.numeric(risky_behaviors$bs_hiv)

risky_behaviors$bs_hiv <- risky_behaviors$bs_hiv -1
```

```r
#BUILD A NEW MODEL which use fupacts as outcome and  WOMEN_ALONG AND BS_HIV as inputs, i set the log(bu

fitsex2 <- glm(data = risky_behaviors, formula = fupacts ~ women_alone + bs_hiv + factor(bupacts), famil

display(fitsex2)
```

```
## glm(formula = fupacts ~ women_alone + bs_hiv + factor(bupacts),
##     family = poisson, data = risky_behaviors)
##                     coef.est coef.se
## (Intercept)           1.22     0.18
## women_alone          -0.46     0.03
## bs_hiv               -0.32     0.04
## factor(bupacts)1     -0.39     0.26
## factor(bupacts)2      0.68     0.19
## factor(bupacts)3      1.01     0.19
## factor(bupacts)4      0.57     0.20
## factor(bupacts)5      1.32     0.19
## factor(bupacts)6      1.20     0.20
## factor(bupacts)7      1.75     0.20
## factor(bupacts)8      0.88     0.20
## factor(bupacts)9      1.22     0.20
## factor(bupacts)10     1.62     0.18
## factor(bupacts)11   -13.21   284.66
## factor(bupacts)12     0.98     0.20
## factor(bupacts)13    -0.07     0.73
## factor(bupacts)15     1.68     0.18
## factor(bupacts)16     1.16     0.23
## factor(bupacts)17    -0.77     0.42
## factor(bupacts)18     0.88     0.24
## factor(bupacts)19    -0.21     0.73
## factor(bupacts)20     1.75     0.18
## factor(bupacts)22     1.77     0.28
## factor(bupacts)24     1.50     0.19
## factor(bupacts)25     2.12     0.20
## factor(bupacts)26   -13.07   284.66
## factor(bupacts)28     0.98     0.25
## factor(bupacts)30     1.63     0.19
## factor(bupacts)33     0.93     0.32
## factor(bupacts)34     0.53     0.36
## factor(bupacts)35     0.51     0.27
## factor(bupacts)36     2.30     0.18
## factor(bupacts)37   -13.21   284.66
## factor(bupacts)40     1.63     0.19
## factor(bupacts)45     2.34     0.18
## factor(bupacts)46     2.47     0.24
## factor(bupacts)47     2.26     0.22
## factor(bupacts)48     0.86     0.39
## factor(bupacts)49    3.15     0.21
## factor(bupacts)50    2.31     0.18
## factor(bupacts)56    2.65     0.24
## factor(bupacts)60    2.82     0.18
## factor(bupacts)64    3.02     0.23
## factor(bupacts)70    2.49     0.18
## factor(bupacts)78    3.16     0.21
## factor(bupacts)80    2.67     0.19
## factor(bupacts)84    3.97     0.19
## factor(bupacts)85    1.80     0.31
## factor(bupacts)87    0.57     0.44
## factor(bupacts)90    2.31     0.18
## factor(bupacts)99    3.04     0.23
## factor(bupacts)100   2.11     0.26
## factor(bupacts)228   3.35     0.20
## factor(bupacts)270   0.39     0.48
## factor(bupacts)300   4.54     0.19
## ---
##   n = 434, k = 55
##   residual deviance = 7614.9, null deviance = 13298.6 (difference = 5683.7)
```

```
#check overdispersion

tapply(risky_behaviors$fupacts,risky_behaviors$women_alone, function(x)c(mean = mean(x),variance = var(
```

```
## $`0`
##     mean variance
##  18.5625 802.9229
##
## $`1`
##      mean  variance
##  12.39726 533.30316
```

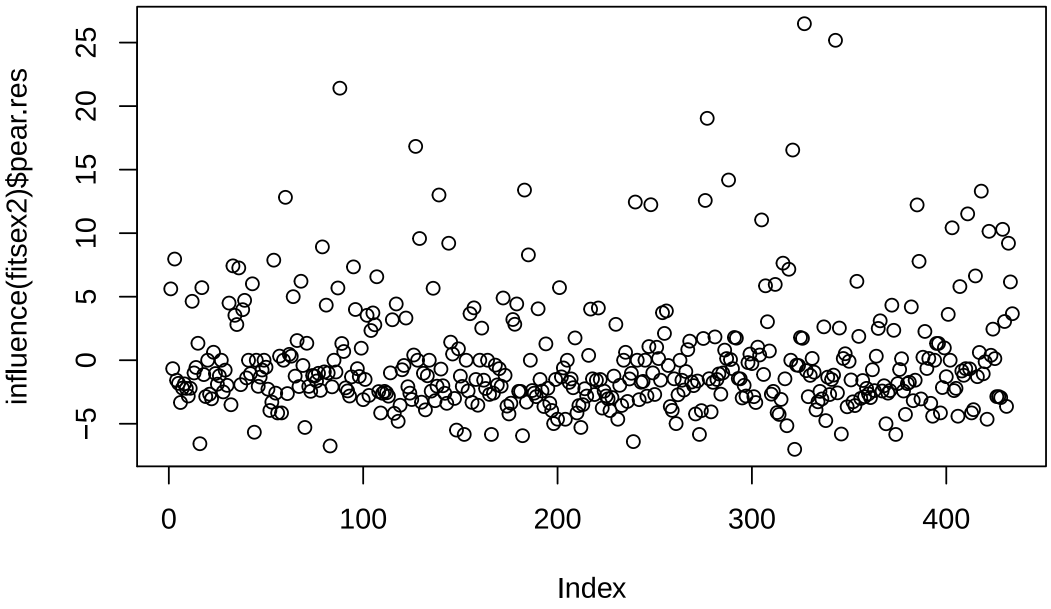
```
tapply(risky_behaviors$fupacts,risky_behaviors$bs_hiv, function(x)c(mean = mean(x),variance = var(x)))
```

```
## $`0`
##      mean  variance
##  18.38279 804.58220
##
## $`1`
##       mean   variance
##   9.907216 371.876718
```

```
#ANOVA TO DECIDED ABOUT P-VALUE FOR EACH VARIABLE COMPARE TO PREVIOUS VARIABLE, DOES IT REALLY WORK IN
Anova(fitsex2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: fupacts
##                 LR Chisq Df Pr(>Chisq)
## women_alone        241.1  1  < 2.2e-16 ***
## bs_hiv              75.5  1  < 2.2e-16 ***
## factor(bupacts)   5070.4 52  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
plot(influence(fitsex2)$pear.res)
```



```r
#Compare p of chisquare of this model and to check is that model work good
1- pchisq(10434, 379)
```

```
## [1] 0
```

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

```r
fitsex3 <- glm(data = risky_behaviors, formula = fupacts ~ man + bs_hiv + bupacts, family = quasipoisson

summary(fitsex3)
```

```
##
## Call:
## glm(formula = fupacts ~ man + bs_hiv + bupacts, family = quasipoisson,
##     data = risky_behaviors)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -17.836   -4.434   -2.697    1.404   22.729
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.5845683  0.1084436  23.833   <2e-16 ***
## man         -0.0915936  0.1322963  -0.692   0.4891
## bs_hiv      -0.5020824  0.1953270  -2.570   0.0105 *
## bupacts      0.0100417  0.0009264  10.839   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for quasipoisson family taken to be 31.26559)
##
##     Null deviance: 13299  on 433  degrees of freedom
## Residual deviance: 10686  on 430  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

```r
display(fitsex3)
```

```
## glm(formula = fupacts ~ man + bs_hiv + bupacts, family = quasipoisson,
##     data = risky_behaviors)
##             coef.est coef.se
## (Intercept)  2.58     0.11
## man         -0.09     0.13
## bs_hiv      -0.50     0.20
## bupacts      0.01     0.00
## ---
##   n = 434, k = 4
##   residual deviance = 10686.1, null deviance = 13298.6 (difference = 2612.5)
##   overdispersion parameter = 31.3
```

```r
yhat <- predict(fitsex2, type = "response")

z <-  ( risky_behaviors$fupacts - yhat)/ sqrt(yhat)

cat("the overdispersion ratio is", sum(z^2)/430 , "\n")
```

```
## the overdispersion ratio is 20.9504
```

```r
cat("the p-value of overdispersion test is ", pchisq(sum(z^2),430), "\n")
```

```
## the p-value of overdispersion test is  1
```

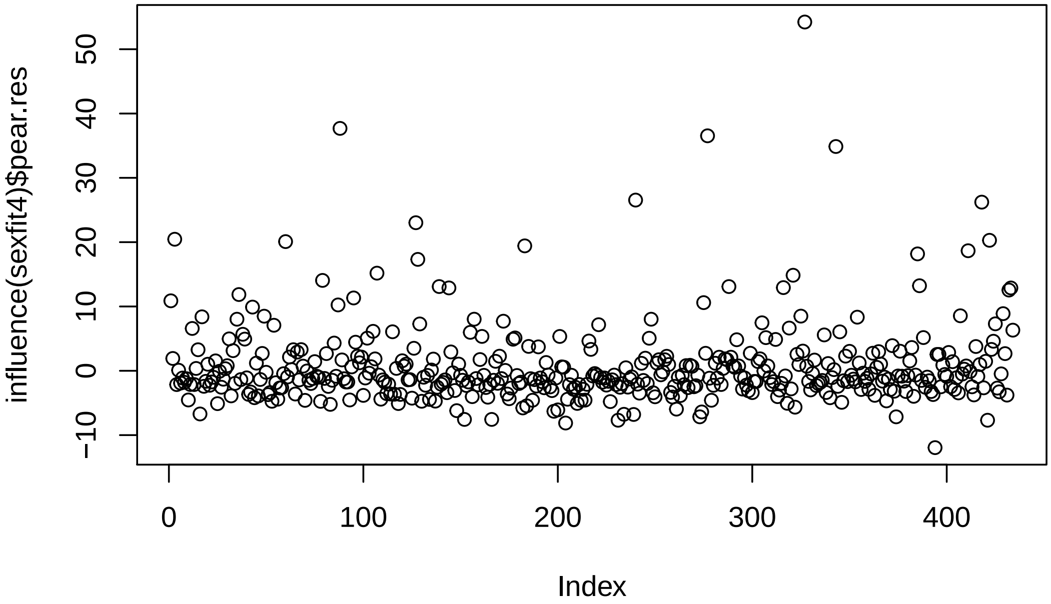
4. These data include responses from both men and women from the participating couples. Does this give you any concern with regard to our modeling assumptions?

```r
sexfit4 <- glm(data = risky_behaviors, formula = fupacts ~ man + bs_hiv + couples, family = poisson, of

summary(sexfit4)
```

```
##
## Call:
## glm(formula = fupacts ~ man + bs_hiv + couples, family = poisson,
##     data = risky_behaviors, offset = log(bupacts + 1))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -16.155   -3.143   -1.222    1.730   21.525
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.32189    0.01901 -16.934  < 2e-16 ***
## man         -0.10970    0.02365  -4.638 3.53e-06 ***
## bs_hiv      -0.36644    0.03525 -10.394  < 2e-16 ***
## couples     -0.14277    0.02504  -5.702 1.18e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 10434  on 433  degrees of freedom
## Residual deviance: 10232  on 430  degrees of freedom
## AIC: 11565
##
## Number of Fisher Scoring iterations: 6
```

```
plot(influence(sexfit4)$pear.res)
```



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)

```
dist100 <- wells_dt$dist/100
wells.fit1 <- glm(data = wells_dt, formula = switch ~ dist100, family = binomial(link = probit))
display(wells.fit1)
```

```
## glm(formula = switch ~ dist100, family = binomial(link = probit),
##     data = wells_dt)
##               coef.est coef.se
## (Intercept)  0.38     0.04
## dist100     -0.39     0.06
## ---
##   n = 3020, k = 2
##   residual deviance = 4076.3, null deviance = 4118.1 (difference = 41.8)
```

```
wells.fit2 <- glm(data = wells_dt, formula = switch ~ dist100, family = binomial(link = logit))
display(wells.fit2)
```

```
## glm(formula = switch ~ dist100, family = binomial(link = logit),
##     data = wells_dt)
##             coef.est coef.se
## (Intercept)  0.61     0.06
## dist100     -0.62     0.10
## ---
##   n = 3020, k = 2
##   residual deviance = 4076.2, null deviance = 4118.1 (difference = 41.9)
```

The coefficients of probit models is 1.6 times the coefficients of logit model

Comparing logit and probit:

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

```
nes5200_dt_s$income_i <- as.integer(nes5200_dt_s$income)
nes5200_dt_s$partyid7_c = (as.integer(nes5200_dt_s$partyid7) - 5)

vote.fit1 <- glm(data = nes5200_dt_s, formula = vote_rep ~ income_i + partyid7_c, family = binomial(lin
display(vote.fit1)
```

```
## glm(formula = vote_rep ~ income_i + partyid7_c, family = binomial(link = probit),
##     data = nes5200_dt_s)
##             coef.est coef.se
## (Intercept) -0.12     0.15
## income_i     0.01     0.05
## partyid7_c   0.61     0.03
## ---
##   n = 1219, k = 3
##   residual deviance = 833.4, null deviance = 1651.1 (difference = 817.6)
```

```
vote.fit2 <- glm(data = nes5200_dt_s , formula = vote_rep ~ income_i + partyid7_c, family = binomial)
display(vote.fit2)
```

```
## glm(formula = vote_rep ~ income_i + partyid7_c, family = binomial,
##     data = nes5200_dt_s)
##             coef.est coef.se
## (Intercept) -0.22     0.28
## income_i     0.01     0.08
## partyid7_c   1.07     0.05
## ---
##   n = 1219, k = 3
##   residual deviance = 829.8, null deviance = 1651.1 (difference = 821.3)
```

```
Anova(vote.fit2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: vote_rep
##            LR Chisq Df Pr(>Chisq)
## income_i       0.01  1     0.9383    
## partyid7_c   788.46  1     <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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)

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.

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.

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.

3. Which model do you prefer?

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.

2. Fit a robit regression and assess model fit.

3. Which model do you prefer?

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.

```
library(faraway)
data(salmonella)
?salmonella
```

```
## starting httpd help server ... done
```

```
#here to do a poission regression between colonies and dose in salmonella.
sal.fit <- glm(data= salmonella, formula = colonies ~ factor(dose), family = poisson)
summary(sal.fit)
```

```
##
## Call:
## glm(formula = colonies ~ factor(dose), family = poisson, data = salmonella)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5737  -0.6820  -0.1110   0.6041   2.4989
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)        3.0758     0.1240  24.798  < 2e-16 ***
## factor(dose)10    -0.1671     0.1832  -0.912 0.361869
## factor(dose)33     0.1431     0.1695   0.844 0.398427
## factor(dose)100    0.6776     0.1523   4.449 8.62e-06 ***
## factor(dose)333    0.5441     0.1559   3.490 0.000484 ***
## factor(dose)1000   0.3142     0.1632   1.926 0.054099 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 78.358  on 17  degrees of freedom
## Residual deviance: 33.496  on 12  degrees of freedom
## AIC: 138.03
##
## Number of Fisher Scoring iterations: 4
```

```
display(sal.fit)
```

```
## glm(formula = colonies ~ factor(dose), family = poisson, data = salmonella)
##                  coef.est coef.se
## (Intercept)       3.08     0.12
## factor(dose)10   -0.17     0.18
## factor(dose)33    0.14     0.17
## factor(dose)100   0.68     0.15
## factor(dose)333   0.54     0.16
## factor(dose)1000  0.31     0.16
## ---
##   n = 18, k = 6
##   residual deviance = 33.5, null deviance = 78.4 (difference = 44.9)
```

```r
tapply(salmonella$colonies, salmonella$dose, function(x)c(mean = mean(x),variance = var(x)))
```

```
## $`0`
##     mean variance
## 21.66667 49.33333
##
## $`10`
##      mean  variance
## 18.333333  6.333333
##
## $`33`
##     mean variance
##       25       73
##
## $`100`
##      mean  variance
##  42.66667 274.33333
##
## $`333`
##     mean variance
## 37.33333 16.33333
##
## $`1000`
##      mean  variance
##  29.66667 126.33333
```

```r
Anova(sal.fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: colonies
##              LR Chisq Df Pr(>Chisq)
## factor(dose)   44.862  5  1.548e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

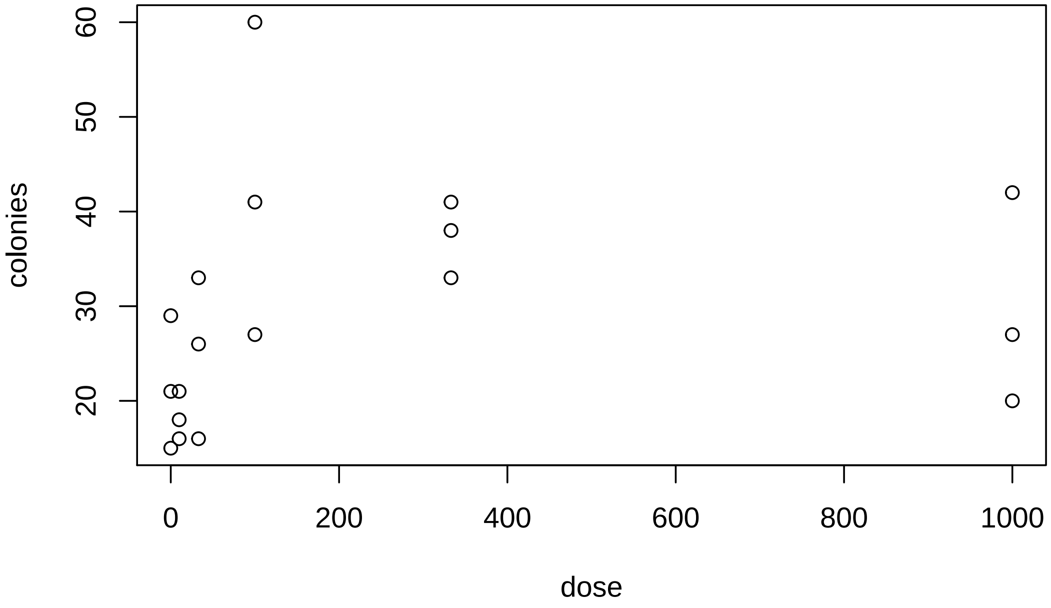
```r
1- pchisq(33.496, 12 )
```

```
## [1] 0.0008096159
```

The result displayed shows that the coeffiecients of different factors of dose. I calculate the mean and variance of colonies of different dose, it showed the variance and mean are not equal which means there will exist some overdispersion in this poisson regression model.

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.

```r
plot(colonies ~ dose, data = salmonella)
```

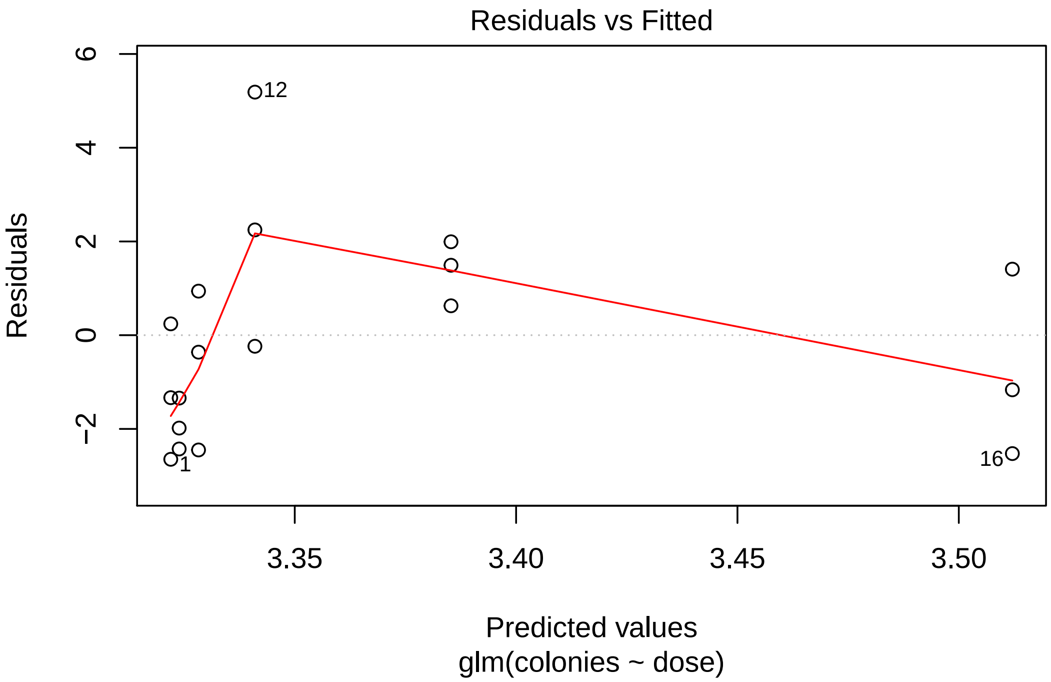


Since we are fitting log linear model we should look at the data on log scale. Also becase the dose is not equally spaced on the raw scale it may be better to plot it on the log scale as well.

```
sal.fit1 <- glm(data= salmonella, formula = colonies ~ dose, family = poisson(link = "log"))
```

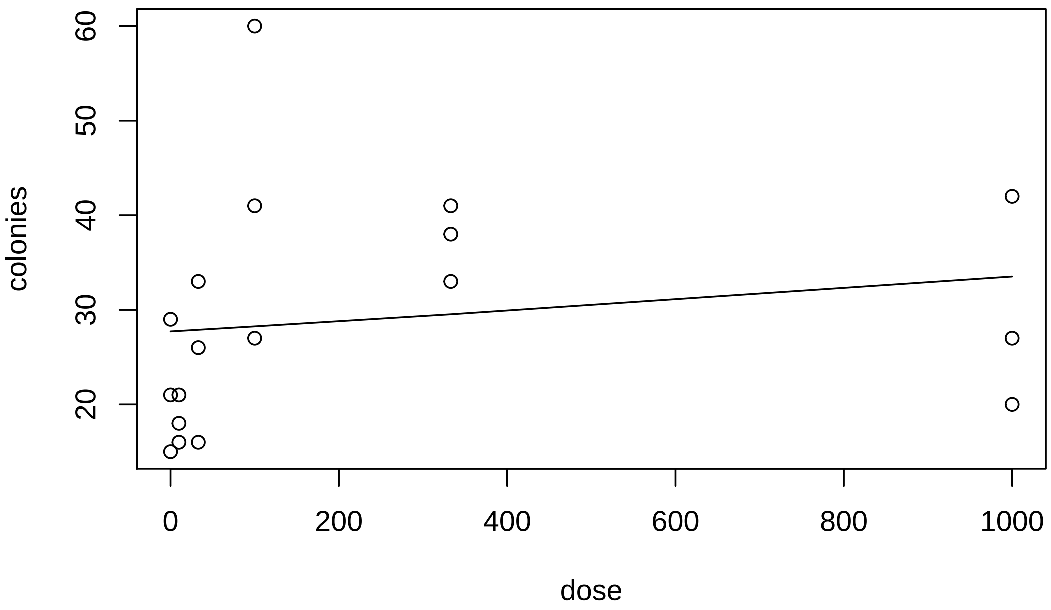
This shows that the trend is not monotonic. Hence when you fit the model and look at the residual you will see a trend.

```
plot(sal.fit1,which =1 )
```



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

```
plot(colonies ~ dose, data = salmonella)
lines(x = salmonella$dose, y= predict.glm(sal.fit1, type = "response") )
```

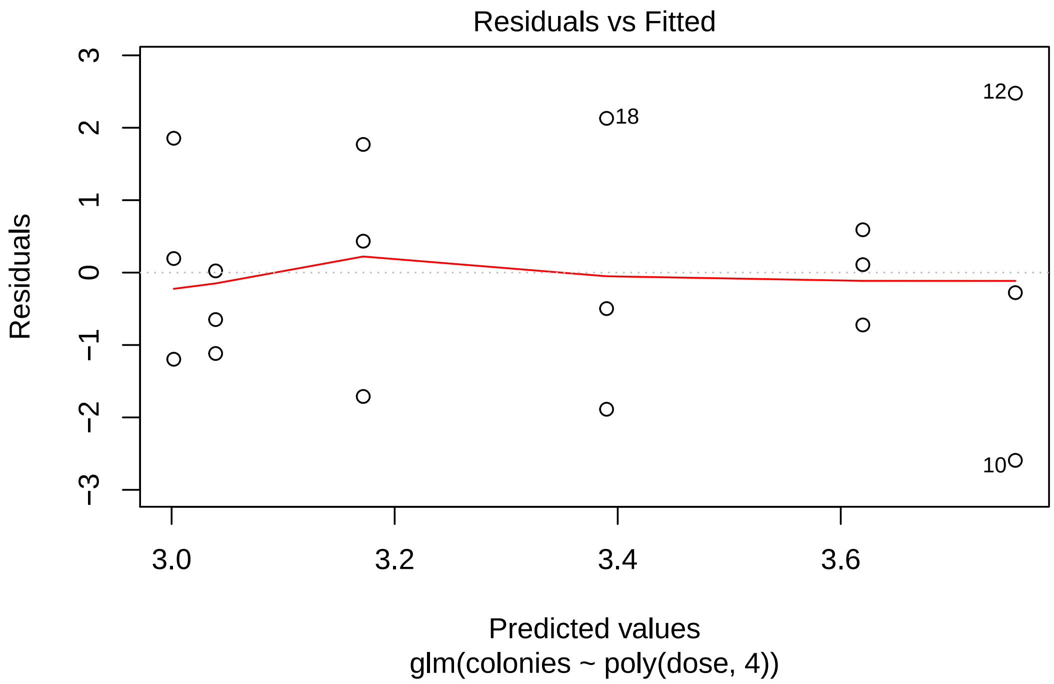


How do we adress this problem? The serious problem to address is the nonlinear trend of dose ranther than the overdispersion since the line is missing the points. Let's add a beny line with 4th order polynomial.

```
sal.fit2 <- glm(colonies ~ poly(dose, 4),data = salmonella, family = poisson(link = "log"))
```

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.

```
plot(sal.fit2,which =1 )
```



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

```
sal.fit3 <- glm(colonies ~ poly(dose, 4),data = salmonella, family = quasipoisson(link = "log"))
summary(sal.fit3)
```

```
##
## Call:
## glm(formula = colonies ~ poly(dose, 4), family = quasipoisson(link = "log"),
##     data = salmonella)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5928  -1.0187  -0.1270   0.5518   2.4771
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.32993    0.07494  44.434 1.38e-15 ***
## poly(dose, 4)1  0.38005    0.31334   1.213   0.2468
## poly(dose, 4)2 -0.85324    0.29098  -2.932   0.0117 *
## poly(dose, 4)3  0.73745    0.28466   2.591   0.0224 *
## poly(dose, 4)4  0.20857    0.33506   0.622   0.5444
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 2.715769)
##
##     Null deviance: 78.358  on 17  degrees of freedom
## Residual deviance: 34.989  on 13  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.

```
library(tidyverse)
```

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

## -- Attaching packages ---------------------------------------------------- tidyverse 1.2.1 --

## v tibble  1.4.2     v purrr   0.2.5
## v tidyr   0.8.1     v dplyr   0.7.6
## v readr   1.1.1     v stringr 1.2.0
## v tibble  1.4.2     v forcats 0.3.0

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

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

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

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

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

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

## -- Conflicts ------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::between()   masks data.table::between()
## x tidyr::expand()    masks Matrix::expand()
## x tidyr::fill()      masks VGAM::fill()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x dplyr::recode()    masks car::recode()
## x dplyr::select()    masks MASS::select()
## x purrr::some()      masks car::some()
## x purrr::transpose() masks data.table::transpose()
```

```
shipsnew <- ships %>% filter(service > 0)
```

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

```
ships2 <- subset(ships, service >0 )
ship.fit1 <- glm(data= shipsnew, incidents ~ factor(year) + factor(period) + type, offset = log(service

ship.fit.new <- glm(data= shipsnew, incidents ~ factor(year) + factor(period) + type, offset = log(servi

Anova(ship.fit.new)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: incidents
##                LR Chisq Df Pr(>Chisq)
## factor(year)    18.5733  3   0.000335 ***
## factor(period)   6.3039  1   0.012047 *
## type            13.9976  4   0.007303 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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?
2. Plot the residuals and the fitted values-why are there lines of observations on the plot?
3. What sort of person would be predicted to visit the doctor the most under your selected model?
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.