Homework 04

Generalized Linear Models

Xuan Zhu

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?

No. When we do the deviance goodness of fit test, the p-value is 0.The null hypothesis is that our model is correctly specified, and we have strong evidence to reject that hypothesis. So we have strong evidence that our model fits badly.

Mean and variance are obviously different.

```
#convert women_alone into factor
risky_behaviors$women_alone <-as.factor(risky_behaviors$women_alone)
#round y
risky_behaviors$fupacts <- round(risky_behaviors$fupacts)
sexmodel <- glm(data=risky_behaviors,family=poisson,fupacts~women_alone)
summary(sexmodel)
```

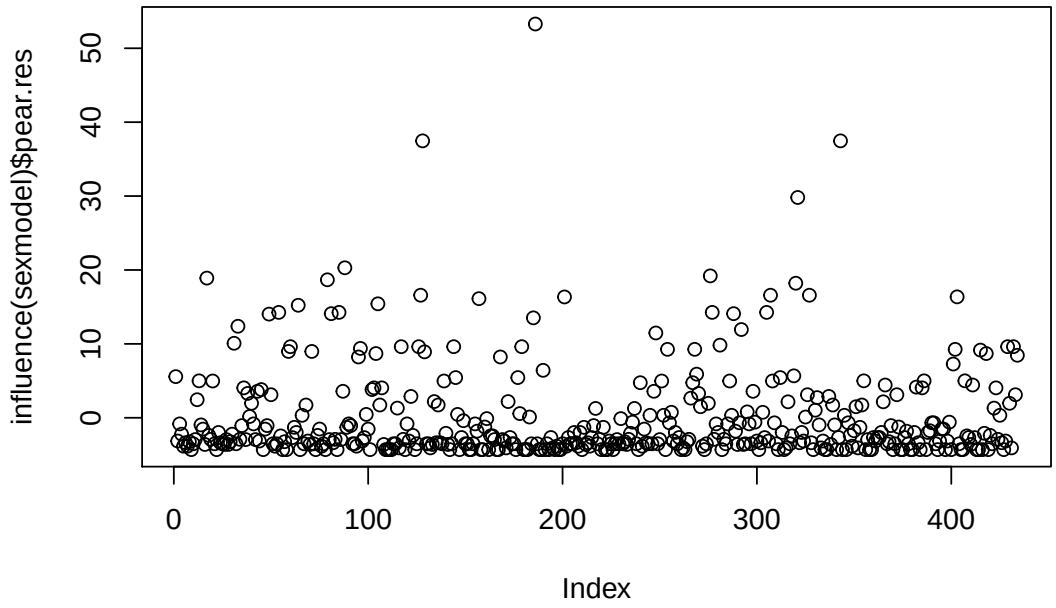
```
##
## Call:
## glm(formula = fupacts ~ women_alone, family = poisson, data = risky_behaviors)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -6.093  -4.979  -3.304   1.237  27.150
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.92114    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
```

```
1-pchisq(13065,432)
```

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

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




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

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

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?

Our model has been improved, but the fitness is still not goog enough.

Yes.

```
sexmodel2 <-glm(data=risky_behaviors,family=poisson,fupacts~women_alone+sex+couples+bupacts+bs_hiv)
summary(sexmodel2)
```

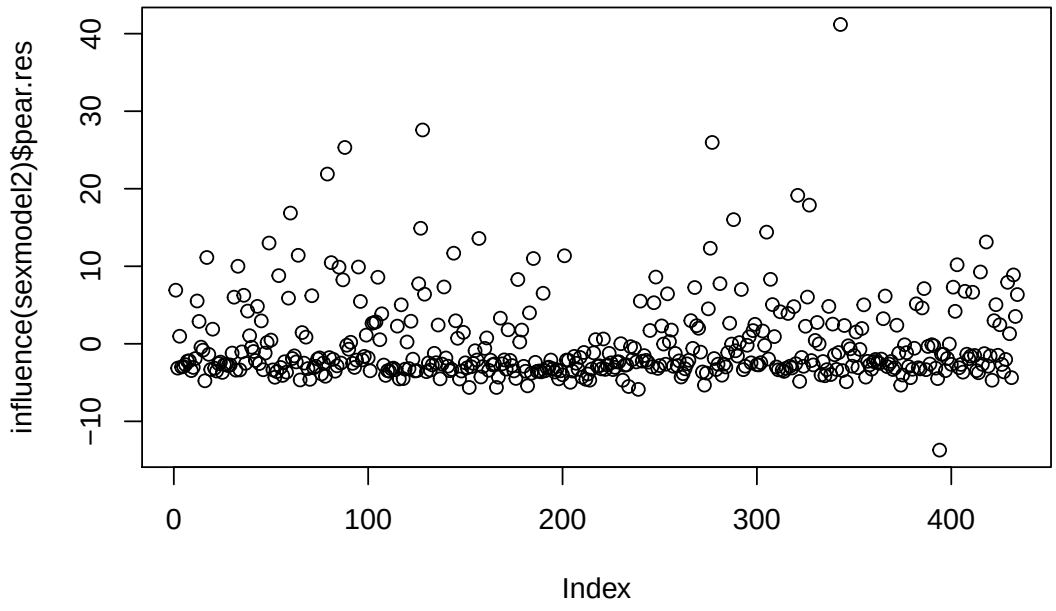
```
##
## Call:
## glm(formula = fupacts ~ women_alone + sex + couples + bupacts +
##     bs_hiv, family = poisson, data = risky_behaviors)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -18.679   -4.305   -2.511    1.368   23.361
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.8957952  0.0232074 124.779  < 2e-16 ***
## women_alone1   -0.6622159  0.0308962 -21.434  < 2e-16 ***
## sexman         -0.1086694  0.0237301  -4.579 4.66e-06 ***
## couples        -0.4099761  0.0282298 -14.523  < 2e-16 ***
## bupacts         0.0107789  0.0001738  62.013  < 2e-16 ***
## bs_hivpositive -0.4383170  0.0353804 -12.389  < 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: 10200  on 428  degrees of freedom
## AIC: 11537
##
## Number of Fisher Scoring iterations: 6
```

```
1-pchisq(10200,428)
```

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

```
plot(influence(sexmodel2)$pear.res)
```



```
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(sexmodel2,trafo=1)
```

```
##
##  Overdispersion test
##
## data:  sexmodel2
## 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?

It is effective to reduce unprotected sex behaviors. As we can see, the coef of women_alone has the expected sign and is significant.

```
oversexmodel <- glm(data=risky_behaviors,family=quasipoisson,fupacts~women_alone+sex+couples+bupacts+bs
summary(oversexmodel)
```

```
##
## Call:
## glm(formula = fupacts ~ women_alone + sex + couples + bupacts +
##     bs_hiv, family = quasipoisson, data = risky_behaviors)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -18.679   -4.305   -2.511    1.368   23.361
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.8957952  0.1271206  22.780  < 2e-16 ***
## women_alone1   -0.6622159  0.1692369  -3.913 0.000106 ***
## sexman         -0.1086694  0.1299838  -0.836 0.403609
## couples        -0.4099761  0.1546315  -2.651 0.008316 **
## bupacts         0.0107789  0.0009521  11.321  < 2e-16 ***
## bs_hivpositive -0.4383170  0.1937994  -2.262 0.024217 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 30.00407)
##
##     Null deviance: 13299  on 433  degrees of freedom
## Residual deviance: 10200  on 428  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

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, women_alone and couples are highly correlated, because they are i.i.d.

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)

```
#Use wells
wells <- read.table("http://www.stat.columbia.edu/~gelman/arm/examples/arsenic/wells.dat", header=TRUE)
wells_dt <- data.table(wells)
dist.d <- wells_dt$dist
dist.d[which(dist.d<100)] <- 1
dist.d[which(100 ==dist.d &100 < dist.d & dist.d<200)] <- 2
dist.d[which(dist.d>200)] <- 3
logit.dist <- glm(switch ~dist.d, data=wells_dt, family=binomial(link="logit"))
probit.dist <- glm(switch ~dist.d, data=wells_dt, family=binomial(link="probit"))
round(coef(logit.dist)[2] / coef(probit.dist)[2],2)
```

```
## dist.d
##    1.6
```

Comparing logit and probit:

construct a dataset where the logit and probit models give different estimates.

I constructed a dataset which contains many outliers. Think x as the distribution of people's income: 90% of people's revenue lie within the interval[10k,80k], and the rest 10% is within [200k,1000k]. And assume y as whether their kids can earn as much as their parents. This is only an example to help my reader get a sense of why large outliers exist.

```
set.seed(25)
A <- runif(900, min =10 , max =80 )
B <- runif(100,min=200,max=1000)
x <-c(A,B)
A1 <- rbinom(900,size=1,prob=0.7)
A2 <- rbinom(100,size=1,prob=0.9)
y <- c(A1,A2)
sim <-data.frame(income=x,success=y)
logit.sim <- glm(success ~income, data=sim, family=binomial(link="logit"))
probit.sim <- glm(success ~income, data=sim, family=binomial(link="probit"))
round(coef(logit.sim)[2] / coef(probit.sim)[2],2)
```

```
## income
##   1.84
```

Tobit model for mixed discrete/continuous data:

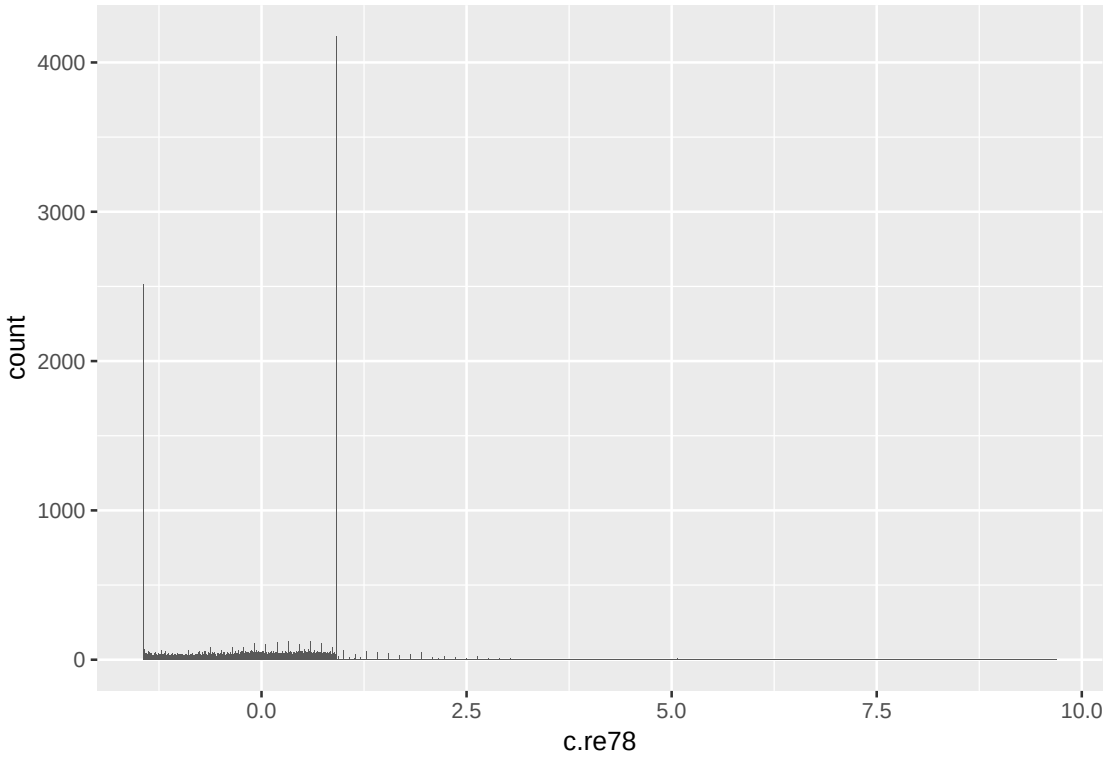
experimental data from the National Supported Work example are available in the folder `lalonde`. Use the treatment indicator and pre-treatment variables to predict post-treatment (1978) earnings using a tobit model. Interpret the model coefficients.

- 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)

```
##
## Attaching package: 'DescTools'
```

```
## The following object is masked from 'package:car':
##
##     Recode
```

```
## The following object is masked from 'package:data.table':
##
##     %like%
```




```
## [1] 25564.67
```

Interpretation:

Tobit regression coefficients are interpreted in the similar manner to OLS regression coefficients; however, the linear effect is on the uncensored latent variable, not the observed outcome.

For example, for one unit increase in age, there's a 0.0115 decrease in the predicted value of re78.

The coefficient labeled "(Intercept):1" is the intercept or constant for the model.

The coefficient labeled "(Intercept):2" is an ancillary statistic. If we exponentiate this value, we get a statistic that is analogous to the square root of the residual variance in OLS regression.

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.

```
##
## Attaching package: 'dplyr'
## The following object is masked from 'package:car':
##
##     recode
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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.

```
con1 <- lm(data=congress1,Dem_vote~x1+x2+incumbent+Rep_vote+contested+Dem_pct)
summary(con1)
```

```
##
## Call:
## lm(formula = Dem_vote ~ x1 + x2 + incumbent + Rep_vote + contested +
##     Dem_pct, data = congress1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47204  -6572    693   6105  75444
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.828e+05  1.123e+04 -16.279   <2e-16 ***
## x1            -4.064e+00  3.380e+01  -0.120    0.904
## x2             3.388e+01  5.213e+01   0.650    0.516
## incumbent      3.317e+03  1.463e+03   2.267    0.024 *
## Rep_vote       9.249e-01  4.975e-02  18.589   <2e-16 ***
## contestedTRUE  5.568e+04  6.408e+03   8.689   <2e-16 ***
## Dem_pct        2.612e+05  1.093e+04  23.896   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13460 on 353 degrees of freedom
## Multiple R-squared:  0.8141, Adjusted R-squared:  0.8109
## F-statistic: 257.6 on 6 and 353 DF,  p-value: < 2.2e-16
```

```
p1 <-predict(con1,congress2,level=.95,interval="prediction")
```

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.

```
library(hett)
con2 <-tlm(data=congress1,Dem_vote~x1+x2+incumbent+Rep_vote+contested+Dem_pct)
summary(con2)
```

```
## Location model :
##
## Call:
## tlm(lform = Dem_vote ~ x1 + x2 + incumbent + Rep_vote + contested +
##     Dem_pct, data = congress1)
```

```
##
## Residuals:
##      Min        1Q    Median        3Q       Max
## -49576.1   -5802.8     389.2    5676.2   86928.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.913e+05  8.015e+03 -23.872   <2e-16 ***
## x1            -3.529e+01  2.413e+01  -1.462    0.144
## x2            -5.405e+01  3.721e+01  -1.452    0.147
## incumbent      1.216e+03  1.044e+03   1.164    0.245
## Rep_vote       8.075e-01  3.552e-02  22.734   <2e-16 ***
## contestedTRUE  8.072e+04  4.574e+03  17.646   <2e-16 ***
## Dem_pct        2.504e+05  7.805e+03  32.088   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Scale parameter(s) as estimated below)
##
##
## Scale Model :
##
## Call:
## tlm(lform = Dem_vote ~ x1 + x2 + incumbent + Rep_vote + contested +
##     Dem_pct, data = congress1)
##
## Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0000  -1.8039  -0.7791   1.1925   5.8092
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  17.9358     0.1054   170.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Scale parameter taken to be  2 )
##
##
## Est. degrees of freedom parameter:  3
## Standard error for d.o.f:  NA
## No. of iterations of model : 17 in 0.03
## Heteroscedastic t Likelihood : -3882.888
```

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

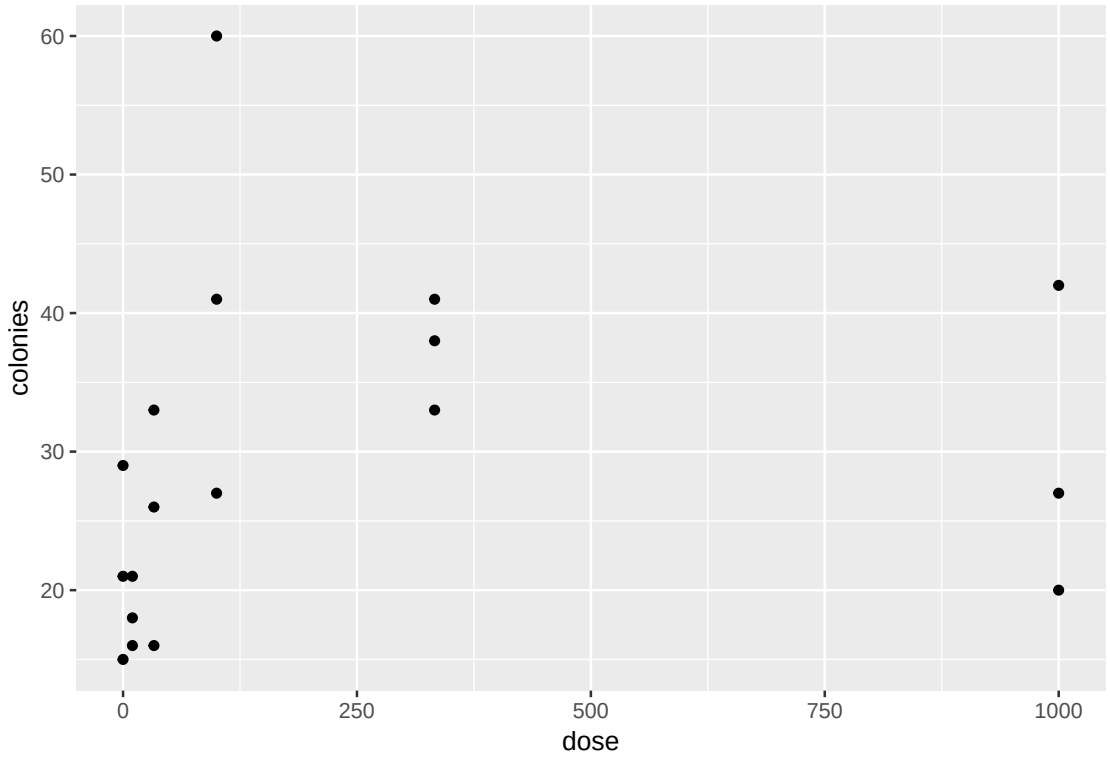
The `salmonella` data was collected in a salmonella reverse mutagenicity assay. The predictor is the dose level of quinoline and the response is the numbers of revertant colonies of TA98 salmonella observed on each of three replicate plates. Show that a Poisson GLM is inadequate and that some overdispersion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

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

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

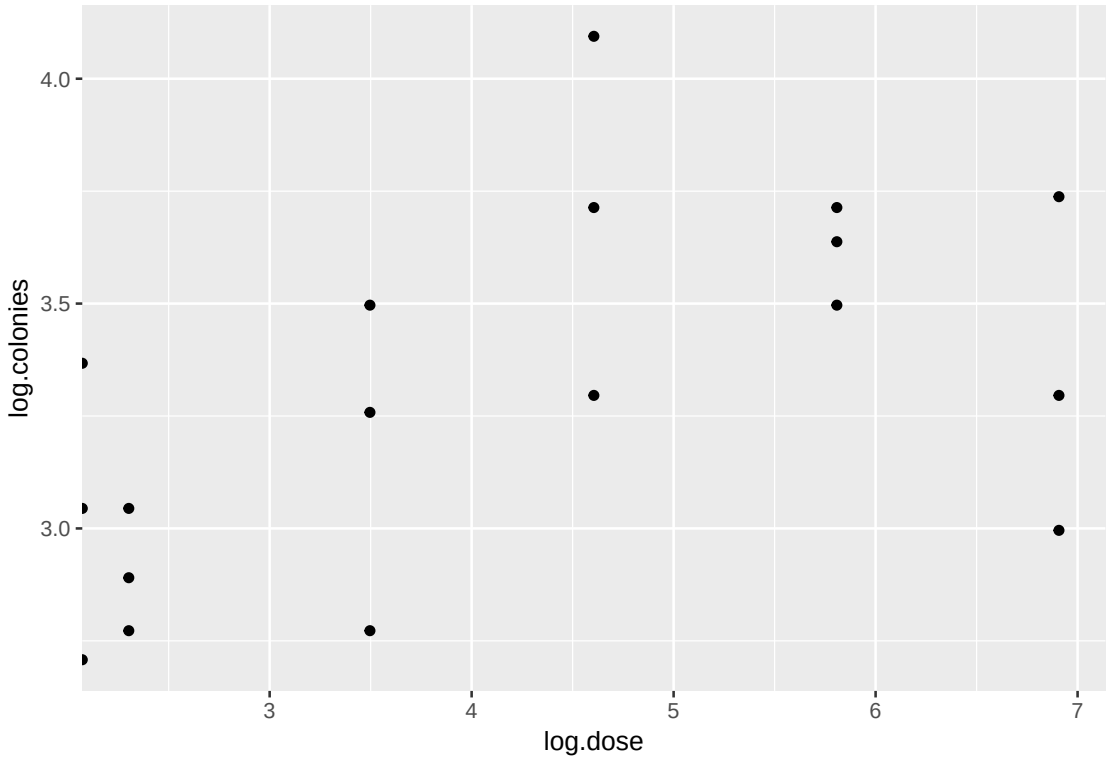
When you plot the data you see that the number of colonies as a function of dose is not monotonic especially around the dose of 1000.

```
ggplot(salmonella)+geom_point(mapping=aes(x=dose,y=colonies))
```



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.

```
log.colonies <- log(salmonella$colonies)
log.dose <- log(salmonella$dose)
ggplot(salmonella)+geom_point(aes(x=log.dose,y=log.colonies))
```

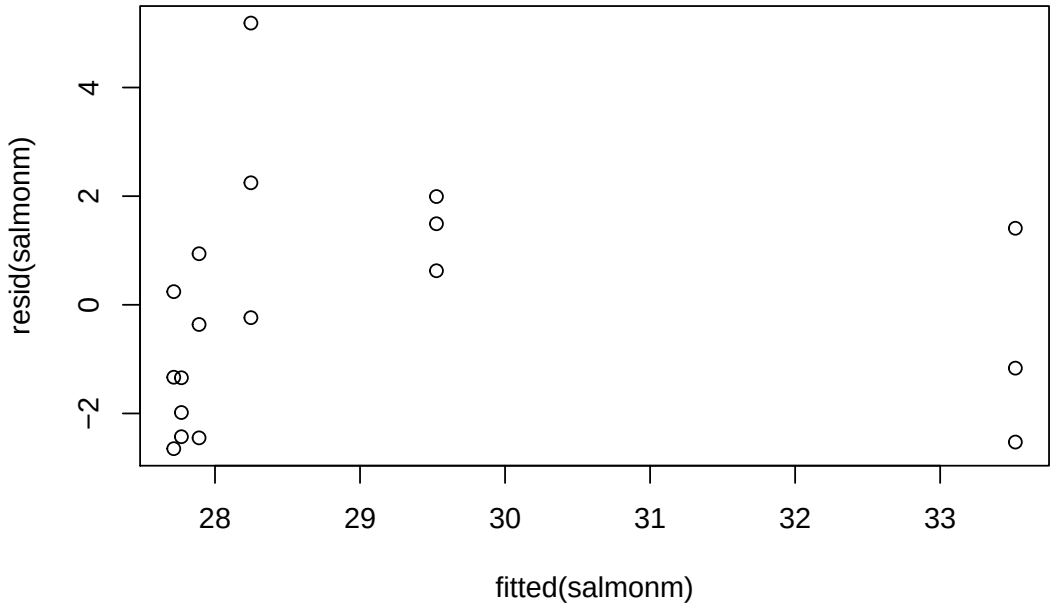


This shows that the trend is not monotonic. Hence when you fit the model and look at the residual you will see a trend.

```
salmonm <- glm(data=salmonella,colonies~dose,family = poisson)
summary(salmonm)
```

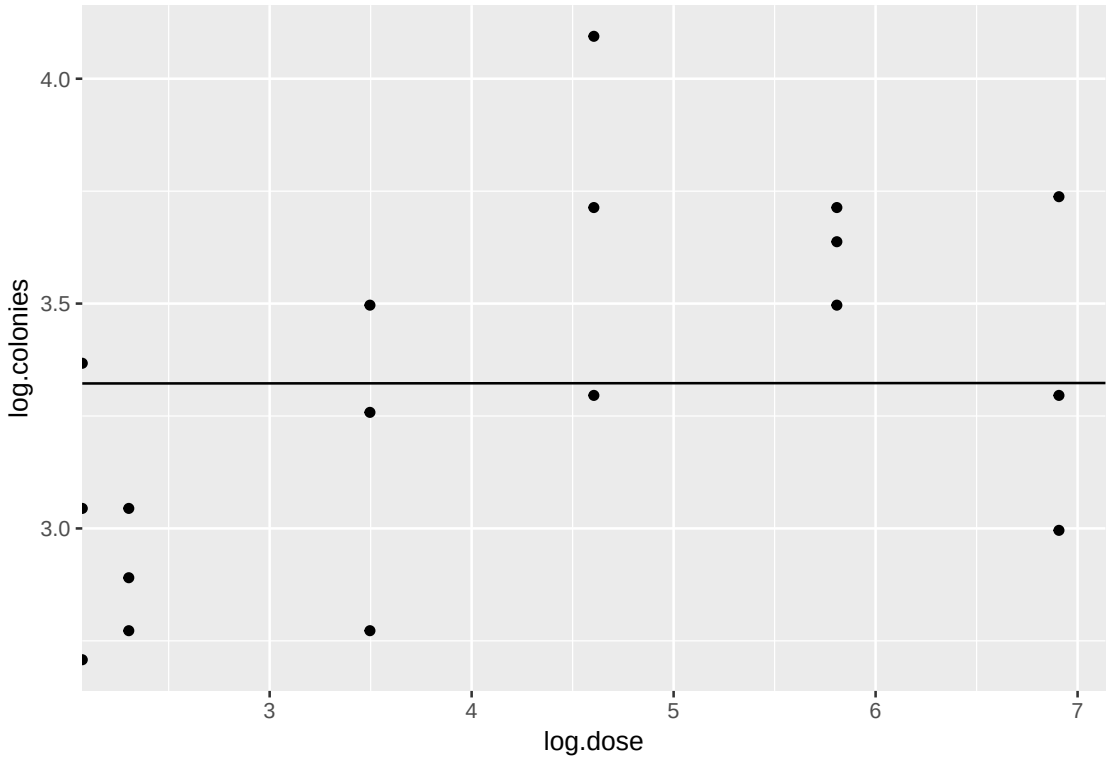
```
##
## Call:
## glm(formula = colonies ~ dose, family = poisson, data = salmonella)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6482  -1.8225  -0.2993   1.2917   5.1861
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.3219950  0.0540292  61.485   <2e-16 ***
## dose        0.0001901  0.0001172   1.622    0.105
## ---
## 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: 75.806  on 16  degrees of freedom
## AIC: 172.34
##
## Number of Fisher Scoring iterations: 4
```

```
plot(fitted(salmonm),resid(salmonm))
```



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

```
ggplot(salmonella,mapping=aes(x=log.dose,y=log.colonies))+geom_point()+geom_abline(intercept = 3.321995
```



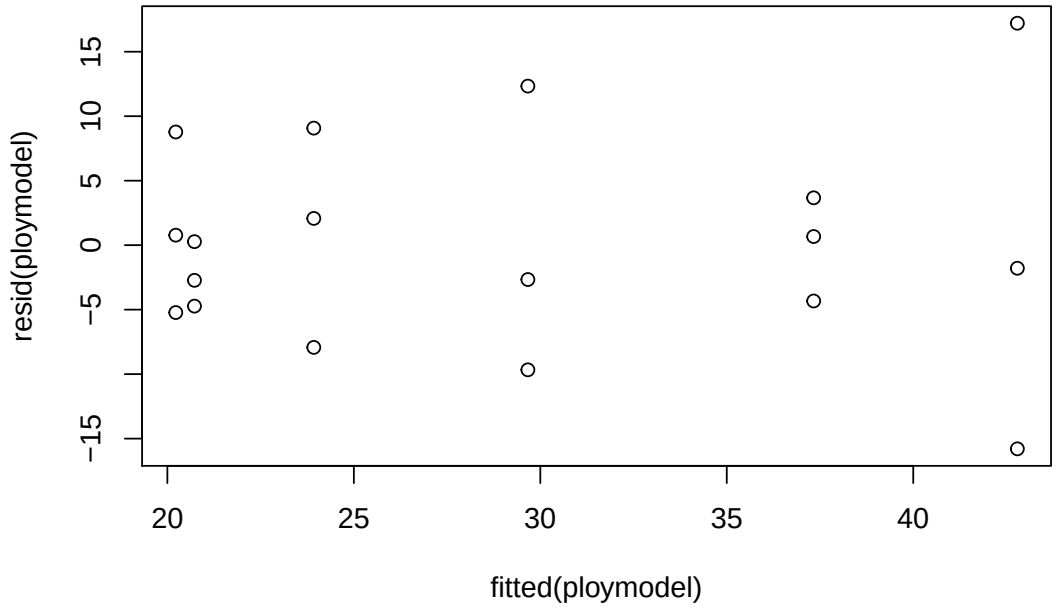
How do we adress this problem? The serious problem to address is the nonlinear trend of dose rather than the overdispersion since the line is missing the points. Let's add a beny line with 4th order polynomial.

```
ploymodel <- lm(data=salmonella, colonies~poly(dose,4))
summary(ploymodel)
```

```
##
## Call:
## lm(formula = colonies ~ poly(dose, 4), data = salmonella)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7920  -4.6276  -0.7594   3.2708  17.2080
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      29.111      2.186  13.317 5.94e-09 ***
## poly(dose, 4)1    8.759      9.275   0.944   0.3622
## poly(dose, 4)2  -25.218      9.275  -2.719   0.0175 *
## poly(dose, 4)3   22.650      9.275   2.442   0.0297 *
## poly(dose, 4)4    8.238      9.275   0.888   0.3906
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.275 on 13 degrees of freedom
## Multiple R-squared:  0.5363, Adjusted R-squared:  0.3937
## F-statistic: 3.759 on 4 and 13 DF,  p-value: 0.03035
```

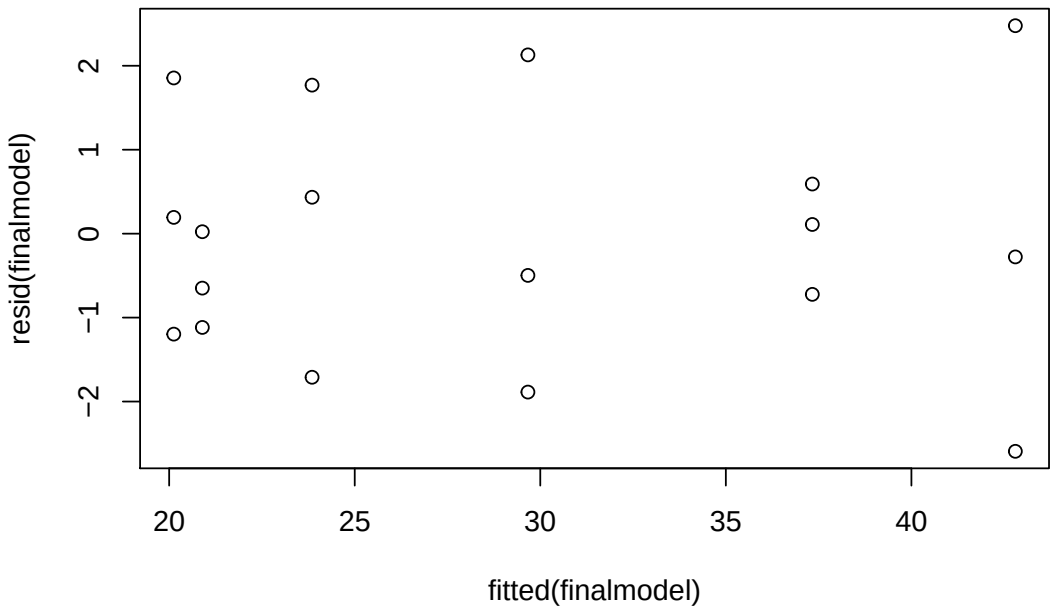
The resulting residual looks nice and if you plot it on the raw data. Whether the trend makes real contextual sense will need to be validated but for the given data it looks feasible.

```
plot(fitted(ploymodel),resid(ploymodel))
```



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

```
finalmodel <- glm(data=salmonella,colonies~poly(dose,4),family=quasipoisson)
plot(fitted(finalmodel),resid(finalmodel))
```



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.

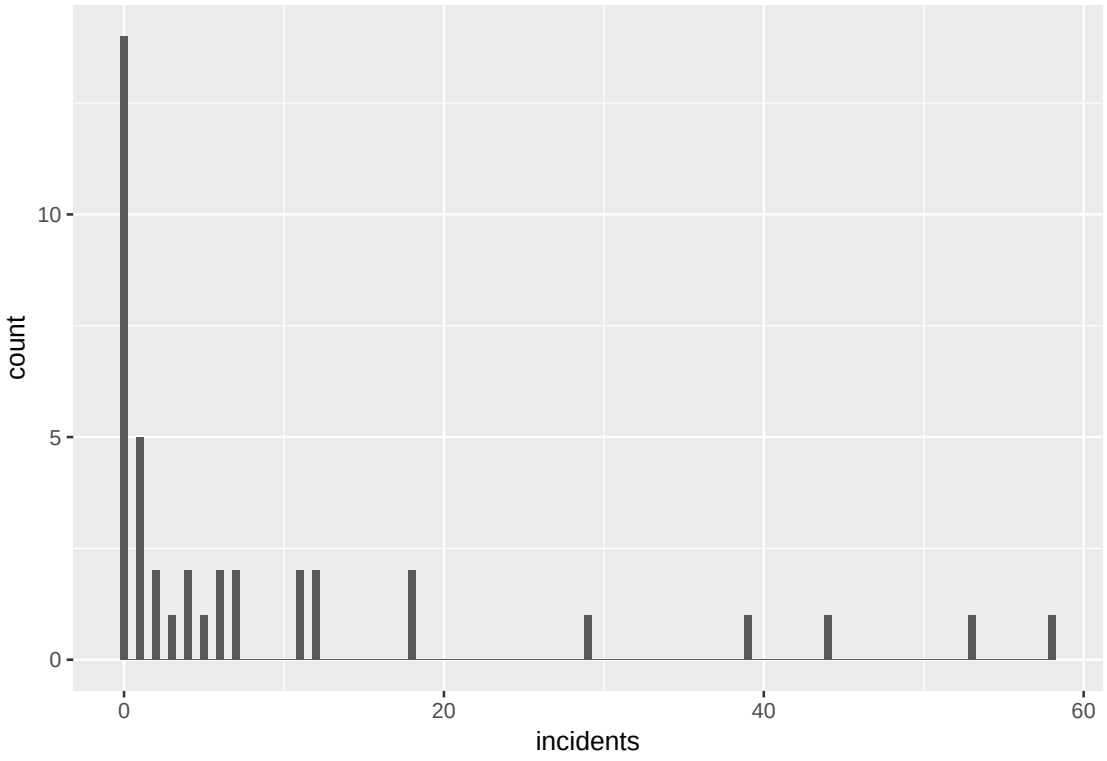
```
data(ships)
?ships
```

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

```
library("pscl")
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
```

```
ggplot(ships)+geom_histogram(aes(x=incidents),binwidth=0.5)
```




```
summary(zeroinfl(incidents~type+period,dist = "poisson" , data=ships))
```

```
##
## Call:
## zeroinfl(formula = incidents ~ type + period, data = ships, dist = "poisson")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.8797 -0.8542 -0.4631  0.6926  3.5677
##
## Count model coefficients (poisson with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.611891   0.530002   3.041  0.00236 **
## typeB        1.462658   0.166795   8.769  < 2e-16 ***
## typeC       -1.548311   0.359199  -4.310 1.63e-05 ***
## typeD       -0.404381   0.289777  -1.395  0.16287
## typeE       -0.273502   0.235529  -1.161  0.24555
## period       0.007458   0.007302   1.021  0.30711
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.133764   3.724046   1.647   0.0995 .
## typeB       -1.564178   1.350175  -1.158   0.2467
## typeC       -1.264799   1.635145  -0.774   0.4392
## typeD        1.160169   1.113260   1.042   0.2973
## typeE       -0.004794   1.101299  -0.004   0.9965
## period      -0.099519   0.055189  -1.803   0.0714 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Number of iterations in BFGS optimization: 20
## Log-likelihood: -118.1 on 12 Df
```

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?

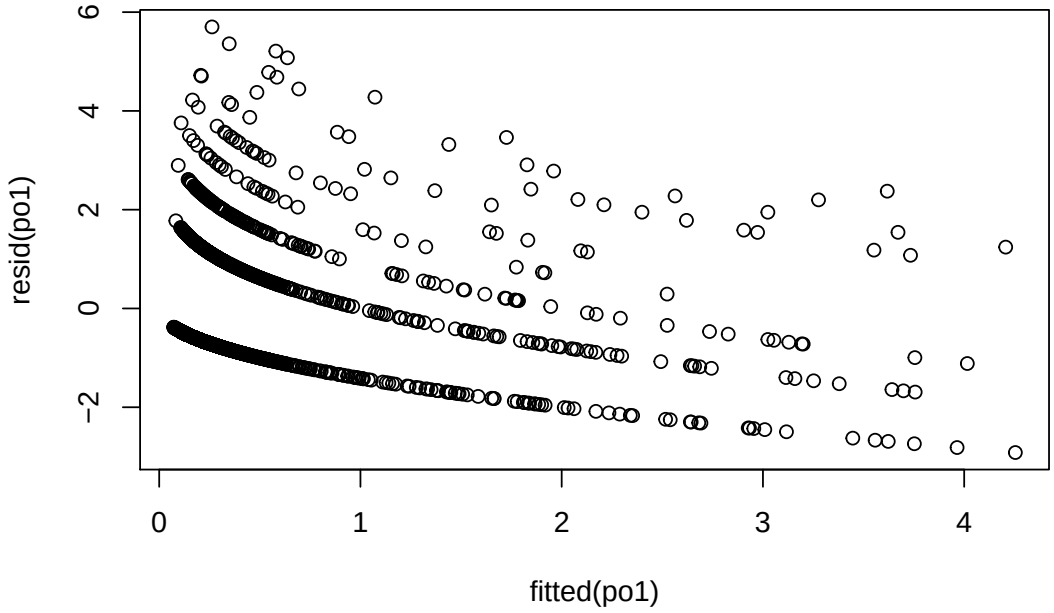
```
po1 <- glm(family=poisson,data=dvisits,doctorco~sex+age+agesq+income+levyplus+freepoor+freerepa+illness
summary(po1)
```

```
##
## 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
```

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

Observed values of y are repeated in the data.

```
plot(fitted(po1),resid(po1))
```



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.