Homework 04

Generalized Linear Models

Tingrui Huang

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?

```
#Clean data
risky_behaviors$fupacts = round(risky_behaviors$fupacts)
#Modeling
riskreg1 <- glm(fupacts~couples+women_alone, data = risky_behaviors, family = poisson())
display(riskreg1)
```

```
## glm(formula = fupacts ~ couples + women_alone, family = poisson(),
##     data = risky_behaviors)
##             coef.est coef.se
## (Intercept)  3.09     0.02
## couples     -0.32     0.03
## women_alone -0.57     0.03
## ---
##   n = 434, k = 3
##   residual deviance = 12925.5, null deviance = 13298.6 (difference = 373.1)
```

```
#Dispersion test
n <- nrow(risky_behaviors)
k <- length(riskreg1$coefficients)
yhat <- predict(riskreg1, type="response")
z <- (risky_behaviors$fupacts-yhat)/sqrt(yhat)
overdp_test <- sum(z^2/(n-k))
pchisq_test <- pchisq(sum(z^2),n-k)
overdp_test;pchisq_test
```

```
## [1] 44.13458
```

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

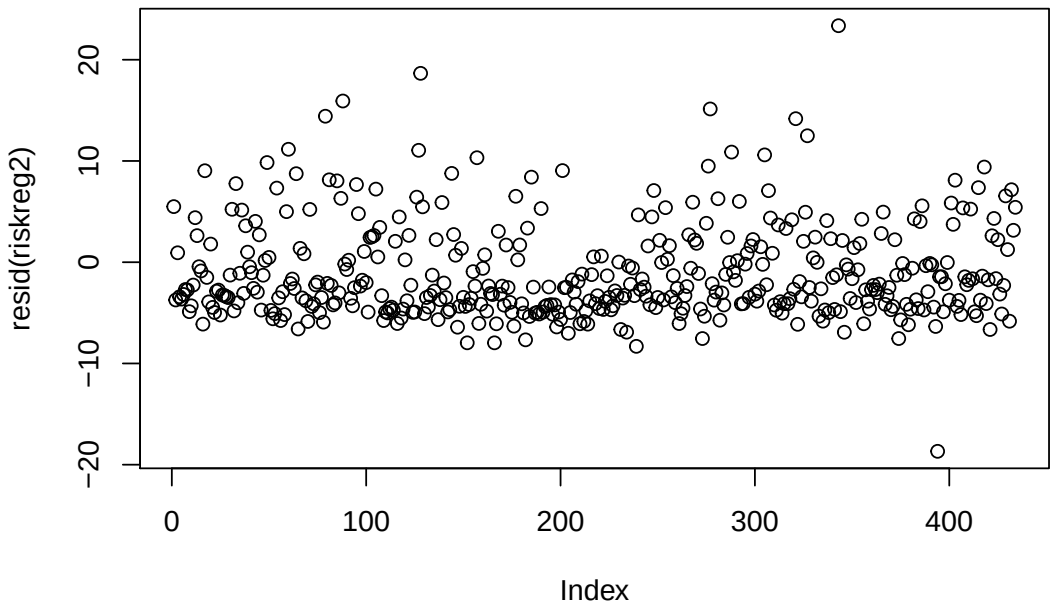
I would say it is not a perfect model due to its large residual deviance and AIC. But since the residual deviance is 300 lower than null deviance, I would say the model is fair. From the result of overdispersion test we can see there is overdispersion in the model.

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
#Model 2
riskreg2 <- glm(fupacts~women_alone+couples+bs_hiv+factor(sex)+bupacts, data = risky_behaviors, family =
summary(riskreg2)
```

```
##
## Call:
## glm(formula = fupacts ~ women_alone + couples + bs_hiv + factor(sex) +
##     bupacts, 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_alone    -0.6622159  0.0308962 -21.434  < 2e-16 ***
## couples        -0.4099761  0.0282298 -14.523  < 2e-16 ***
## bs_hivpositive -0.4383170  0.0353804 -12.389  < 2e-16 ***
## factor(sex)man -0.1086694  0.0237301  -4.579 4.66e-06 ***
## bupacts         0.0107789  0.0001738  62.013  < 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
```

```r
plot(resid(riskreg2))
```




```
#Check overdispersion
riskod <- glm(fupacts~women_alone+couples+bs_hiv+factor(sex)+bupacts, data = risky_behaviors, family = q
display(riskod)
```

```
## glm(formula = fupacts ~ women_alone + couples + bs_hiv + factor(sex) +
##     bupacts, family = quasipoisson, data = risky_behaviors)
##                 coef.est coef.se
## (Intercept)     2.90     0.13
## women_alone    -0.66     0.17
## couples        -0.41     0.15
## bs_hivpositive -0.44     0.19
## factor(sex)man -0.11     0.13
## bupacts         0.01     0.00
## ---
##   n = 434, k = 6
##   residual deviance = 10200.4, null deviance = 13298.6 (difference = 3098.2)
##   overdispersion parameter = 30.0
```

As we can see the residual deviance of 2nd model is way better than the first model. From the result table of QuasiPoisson we can see that the model is still overdispersed.

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

```
#Add offset to the model
interv_test <- risky_behaviors[risky_behaviors$bupacts>0,]
interv_reg <- glm(fupacts~women_alone+couples+bs_hiv+factor(sex), data = interv_test, family = quasipois
summary(interv_reg)
```

```
##
## Call:
## glm(formula = fupacts ~ women_alone + couples + bs_hiv + factor(sex),
```

```
##     family = quasipoisson, data = interv_test, offset = log(bupacts))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -16.315   -3.165   -1.072    2.218   21.552
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.03222    0.15314  -0.210  0.83349
## women_alone    -0.55581    0.20706  -2.684  0.00756 **
## couples        -0.40263    0.19078  -2.110  0.03542 *
## bs_hivpositive -0.32512    0.24316  -1.337  0.18193
## factor(sex)man -0.11843    0.16139  -0.734  0.46346
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 46.30972)
##
##     Null deviance: 10577  on 419  degrees of freedom
## Residual deviance: 10032  on 415  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

```
interv_reg2 <- glm(fupacts~factor(women_alone+couples)+bs_hiv+factor(sex), data = interv_test, family =
summary(interv_reg2)
```

```
##
## Call:
## glm(formula = fupacts ~ factor(women_alone + couples) + bs_hiv +
##     factor(sex), family = quasipoisson, data = interv_test, offset = log(bupacts))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -15.687   -3.156   -1.056    2.069   21.958
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -0.03383    0.15414  -0.220  0.82636
## factor(women_alone + couples)1 -0.47091    0.16916  -2.784  0.00562 **
## bs_hivpositive                 -0.30296    0.24274  -1.248  0.21270
## factor(sex)man                 -0.11781    0.16242  -0.725  0.46867
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 46.9077)
##
##     Null deviance: 10577  on 419  degrees of freedom
## Residual deviance: 10057  on 416  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

```
#ANOVA test
anova(interv_reg,interv_reg2)
```

```
## Analysis of Deviance Table
##
## Model 1: fupacts ~ women_alone + couples + bs_hiv + factor(sex)
## Model 2: fupacts ~ factor(women_alone + couples) + bs_hiv + factor(sex)
##   Resid. Df Resid. Dev Df Deviance
## 1       415      10032
## 2       416      10057 -1  -24.621
```

By adding the intervention, some of the predicting variables have become less significant. When comparing these two models, there isn't too much difference.

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? I think there will be concerns regarding our model assumptions because I have seen some weird things in the dataset. I found that some observations that "couples"=0, "women_only"=0 and "sex"=0. I would say this is very confusing, because "couple"=0 and "women_only"=0 together would indicate the patient is a male, but from the data we can see the patient is female. Therefore, these issues may cause problems with our assumptiona and the interpretation for the model.

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)

```
logit_reg <- glm(switch~arsenic+dist+assoc+educ, data = wells_dt, family=binomial(link="logit"))
probit_reg <- glm(switch~arsenic+dist+assoc+educ, data = wells_dt, family=binomial(link="probit"))
anova(logit_reg,probit_reg)
```

```
## Analysis of Deviance Table
##
## Model 1: switch ~ arsenic + dist + assoc + educ
## Model 2: switch ~ arsenic + dist + assoc + educ
##   Resid. Df Resid. Dev Df Deviance
## 1      3015     3907.8
## 2      3015     3909.7  0  -1.9178
```

```
coef_logit <- logit_reg$coefficients
coef_probit <- probit_reg$coefficients
1.6*coef_probit-coef_logit
```

```
##   (Intercept)       arsenic          dist         assoc          educ
##  2.157536e-02 -2.460870e-02  2.247523e-04 -3.132010e-03  9.273578e-05
```

The result from $1.6 * coef_probit - coef_logit$ is small enough to be seen as 0, and fromt he ANOVA test we can see the results of these two models are pretty the same.

Comparing logit and probit:

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

```
yyy <- rbinom(100,1,0.5)
xxx <- rnorm(100,40,8)
eee <- rnorm(100,5,0.6)
logit_reg2 <- glm(yyy~xxx+eee, family = binomial(link="logit"))
```

```
probit_reg2 <- glm(yyy~xxx+eee, family = binomial(link="probit"))
coef2_logit <- logit_reg2$coefficients
coef2_probit <- probit_reg2$coefficients
coef2_logit/coef2_probit
```

```
## (Intercept)         xxx         eee
##    1.601677    1.594292    1.585677
```

The difference in estimates are pretty close to 1.6.

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)

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v tibble  1.4.2     v purrr   0.2.5
## v tidyr   0.8.1     v dplyr   0.7.7
## v readr   1.1.1     v stringr 1.3.1
## v tibble  1.4.2     v forcats 0.3.0
```

```
## -- 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()
```

```
library(hett)
summary(hett::tlm(re78 ~ factor(treat)+age+educ+black+married, data = lalonde))
```

```
## Location model :
##
## Call:
```

```
## hett::tlm(lform = re78 ~ factor(treat) + age + educ + black +
##     married, data = lalonde)
##
## Residuals:
##      Min        1Q    Median        3Q       Max
## -26240.4   -7732.6     847.6    5797.6  106702.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1508.152    402.335  -3.748 0.000178 ***
## factor(treat)1 -3988.533    733.581  -5.437 5.48e-08 ***
## age              117.314      7.125  16.465  < 2e-16 ***
## educ             824.701     24.767  33.299  < 2e-16 ***
## black          -2239.136    238.345  -9.395  < 2e-16 ***
## married         6451.619    174.714  36.927  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Scale parameter(s) as estimated below)
##
##
## Scale Model :
##
## Call:
## hett::tlm(lform = re78 ~ factor(treat) + age + educ + black +
##     married, data = lalonde)
##
## Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0000  -1.5410  -0.5472   1.1842   5.8733
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 17.92793    0.01464    1225   <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 : 10 in 0.17
## Heteroscedastic t Likelihood : -199206.7
```

Coefficient of Treat : the predicted income of experimental treatment group will be less than the comparison group by 3988. Coeffiecient of age: for every one age older, the income will be increase by 117. Coefficient of EDUC: for every one more year in school, income will increase by 824. Coeffcient of Black: black people make 2239 dolloars less that non-black people. If a person is married, he or she will make 6451 dollars more than non-married people.

Robust linear regression using the t model:

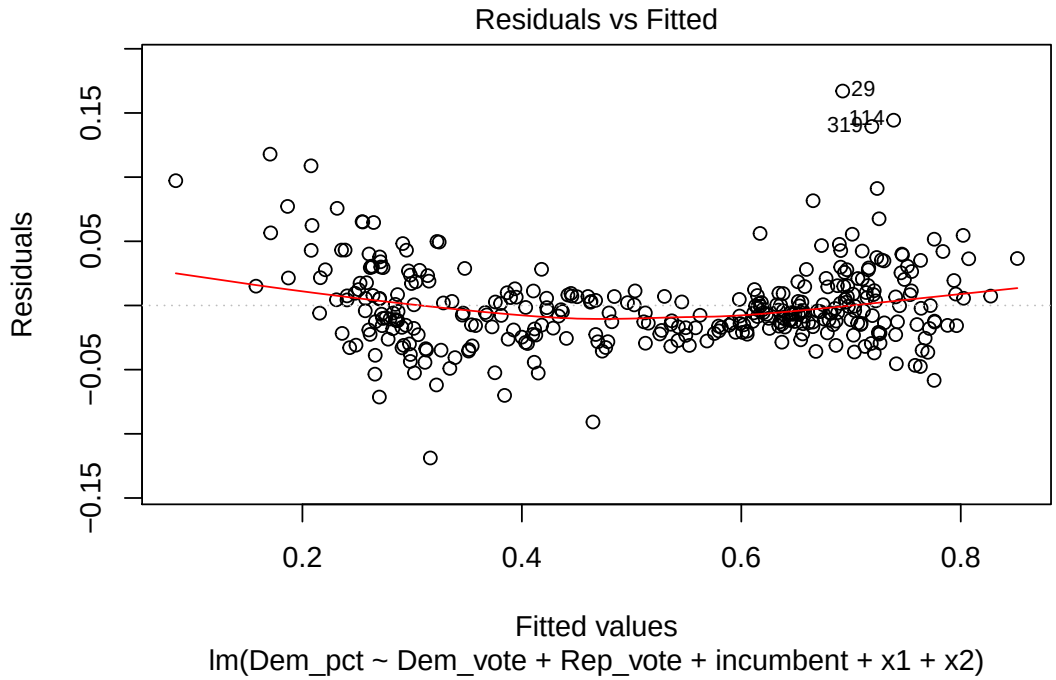
The csv file `congress` has the votes for the Democratic and Republican candidates in each U.S. congressional district in between 1896 and 1992, along with the parties' vote proportions and an indicator for whether the incumbent was running for reelection. For your analysis, just use the elections in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

```
#Filter for data in 1988 and contest=TRUE
con1988 <- congress %>% filter(year==1988) %>% filter(contested=="TRUE")
con88reg <- lm(Dem_pct~Dem_vote+Rep_vote+incumbent+x1+x2, data=con1988)
summary(con88reg)
```

```
##
## Call:
## lm(formula = Dem_pct ~ Dem_vote + Rep_vote + incumbent + x1 +
##     x2, data = con1988)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.118802 -0.017714 -0.005003  0.011762  0.167057
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.634e-01  1.104e-02  51.044  < 2e-16 ***
## Dem_vote     1.930e-06  7.145e-08  27.011  < 2e-16 ***
## Rep_vote    -2.506e-06  6.603e-08 -37.947  < 2e-16 ***
## incumbent    1.424e-02  3.812e-03   3.737 0.000218 ***
## x1           1.894e-04  8.174e-05   2.317 0.021092 *
## x2          -2.028e-04  1.195e-04  -1.698 0.090429 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03266 on 342 degrees of freedom
## Multiple R-squared:  0.9712, Adjusted R-squared:  0.9708
## F-statistic:  2308 on 5 and 342 DF,  p-value: < 2.2e-16
```

```
plot(con88reg, which=1)
```



I think the model is pretty good, since all variabels are statistically significant and the R-square of the more is 0.97, and the residual vs fitted plot is also not bad.

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.

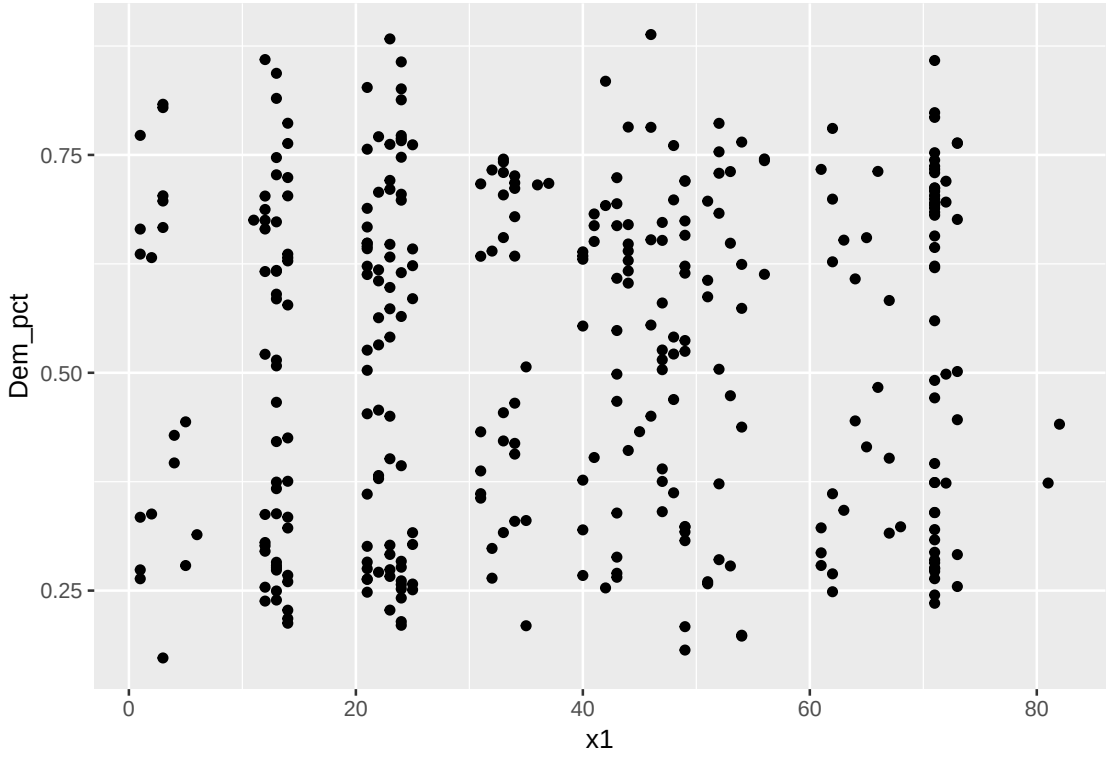
```
summary(hett::tlm(Dem_pct~Dem_vote+Rep_vote+incumbent+x1+x2, data=con1988))
```

```
## Location model :
##
## Call:
## hett::tlm(lform = Dem_pct ~ Dem_vote + Rep_vote + incumbent +
##     x1 + x2, data = con1988)
##
## Residuals:
##        Min          1Q      Median          3Q         Max
## -0.1172188  -0.0130060  -0.0007583   0.0150751   0.1611923
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.710e-01  8.371e-03  68.213  < 2e-16 ***
## Dem_vote     1.920e-06  5.418e-08  35.431  < 2e-16 ***
## Rep_vote    -2.611e-06  5.007e-08 -52.149  < 2e-16 ***
## incumbent    1.009e-02  2.891e-03   3.490 0.000547 ***
## x1           1.812e-04  6.199e-05   2.923 0.003696 **
## x2          -1.299e-04  9.059e-05  -1.433 0.152658
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

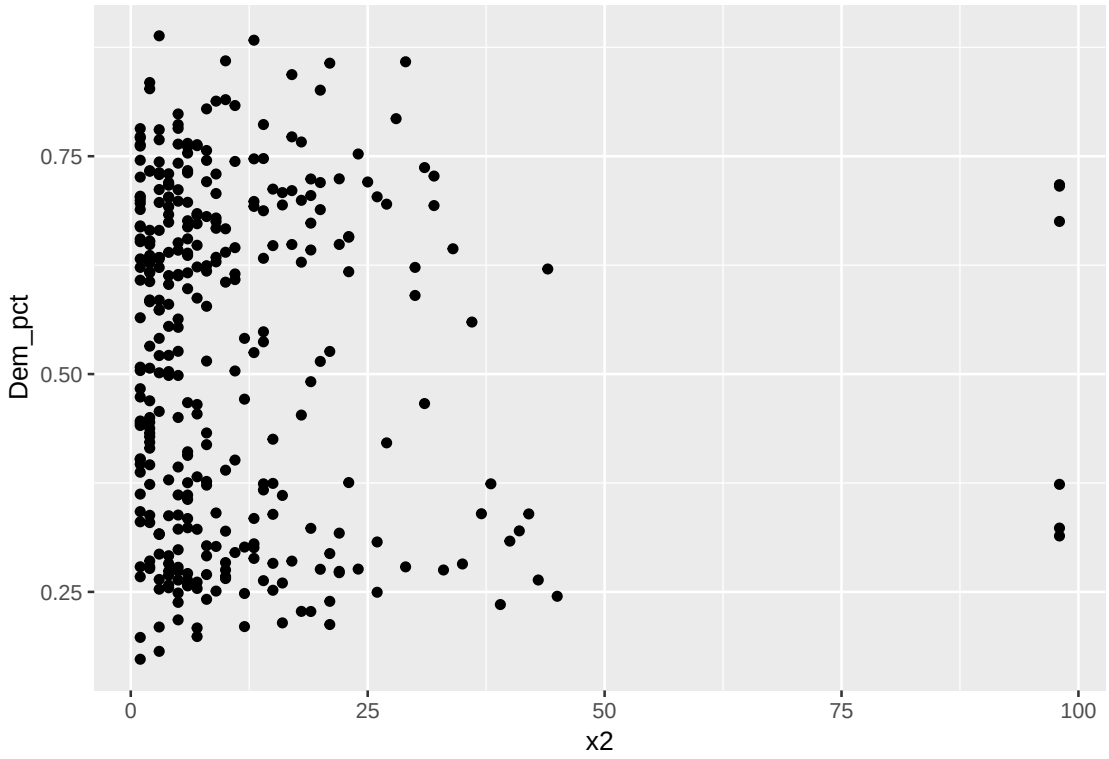
```
## (Scale parameter(s) as estimated below)
##
##
## Scale Model :
##
## Call:
## hett::tlm(lform = Dem_pct ~ Dem_vote + Rep_vote + incumbent +
##     x1 + x2, data = con1988)
##
## Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0000  -1.7044  -0.9393   1.4108   5.6391
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -7.8017     0.1072  -72.77   <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 : 21 in 0.03
## Heteroscedastic t Likelihood : 735.7843
```

3. Which model do you prefer?

```
ggplot(con1988, aes(x=x1, y=Dem_pct))+geom_point()
```



```
ggplot(con1988, aes(x=x2, y=Dem_pct))+geom_point()
```



Although the outcomes from the linear regression seem to be really good, from the graphs above we can easily tell that there are truncation in variable x1 and x2. Therefore, I would say the tobit model could be better than the linear regression since tobit model is capable of dealing with truncations.

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.

```
con1988_s <- con1988 %>% mutate(win_dem = ifelse(con1988$Dem_pct>0.5,1,0))
reg_dem <- glm(win_dem~incumbent+x1+x2, data = con1988_s, family = binomial(link="probit"))
summary(reg_dem)
```

```
##
## Call:
## glm(formula = win_dem ~ incumbent + x1 + x2, family = binomial(link = "probit"),
##     data = con1988_s)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8919  -0.2278   0.1873   0.2130   2.7796
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)  0.204633   0.287722   0.711    0.477
## incumbent    1.980424   0.153083  12.937   <2e-16 ***
## x1          -0.002532   0.005987  -0.423    0.672
## x2          -0.009545   0.007477  -1.277    0.202
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 478.27  on 347  degrees of freedom
## Residual deviance: 106.66  on 344  degrees of freedom
## AIC: 114.66
##
## Number of Fisher Scoring iterations: 6
```

\color{blue Only the "incument" variable is statistically significant in this model. The residual deviance is way better than the null deviance, I would say the model is pretty good.

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
```

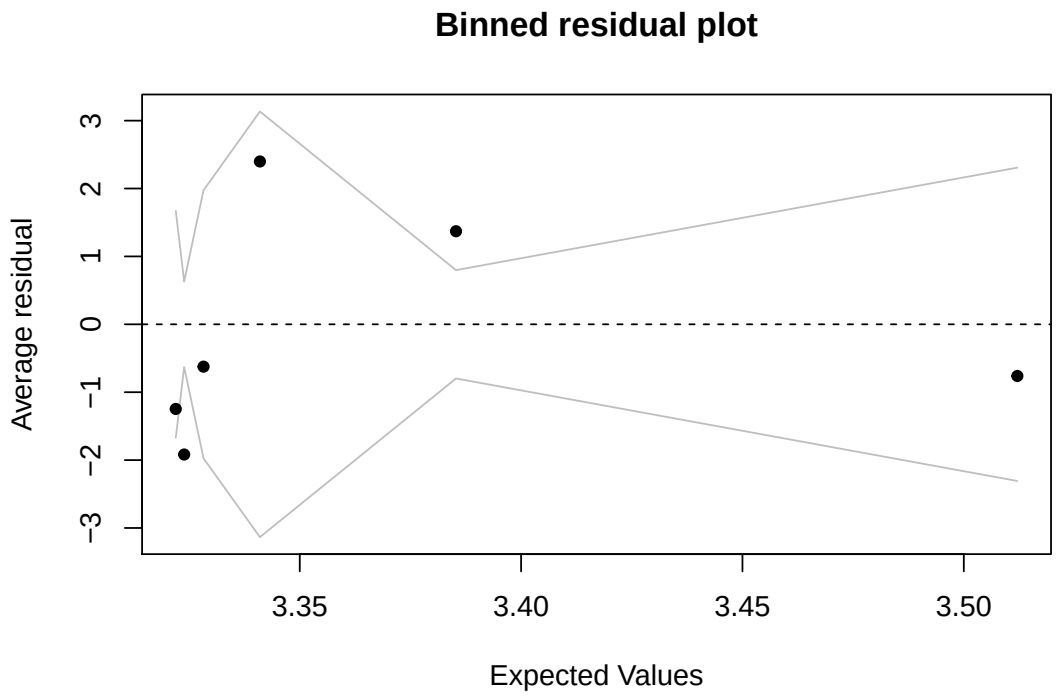
When you plot the data you see that the number of colonies as a function of dose is not monotonic especially around the dose of 1000.

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

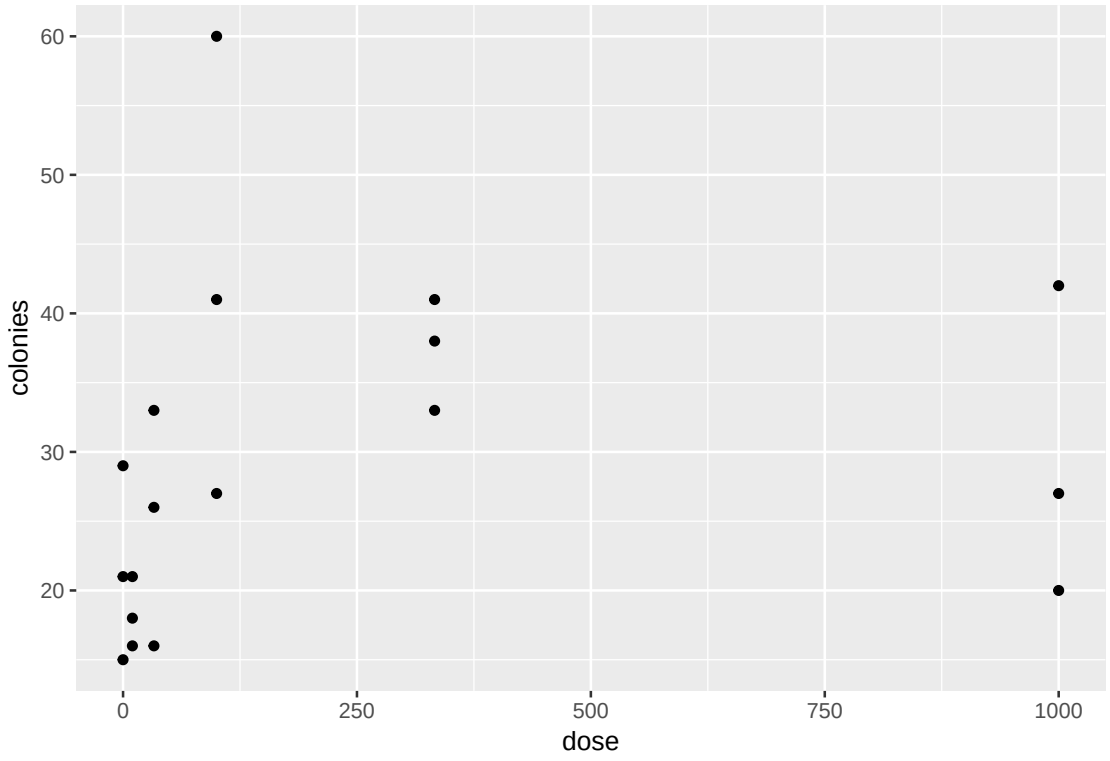
```
##
## Call:
## glm(formula = colonies ~ dose, family = poisson(), data = salmonella)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6482  -1.8225  -0.2993   1.2917   5.1861
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.3219950  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
```

```
binnedplot(predict(salreg),resid(salreg))
```



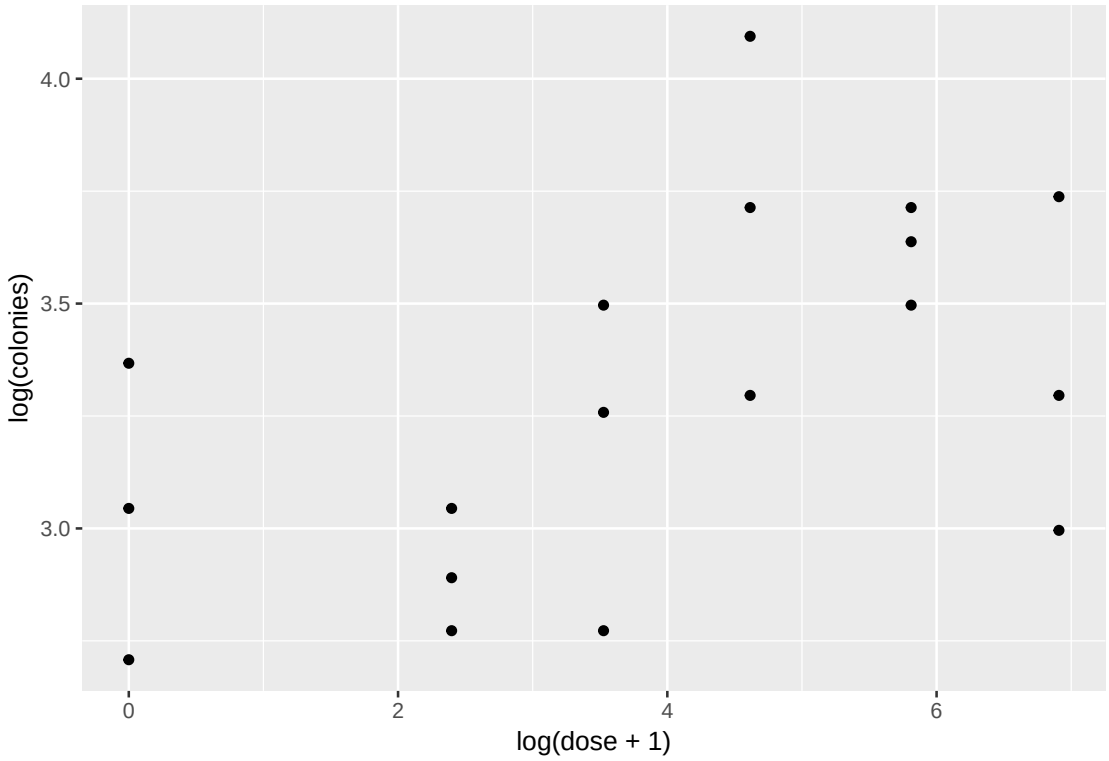

```
ggplot(salmonella, aes(x=dose, y=colonies))+geom_point()
```



The Poisson model doesn't fit the data well since the coefficient of dose is not significant and from the residual plot we see more than 60% residuals fall outside the boundary.

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.

```
ggplot(salmonella, aes(x=log(dose+1), y=log(colonies)))+geom_point()
```

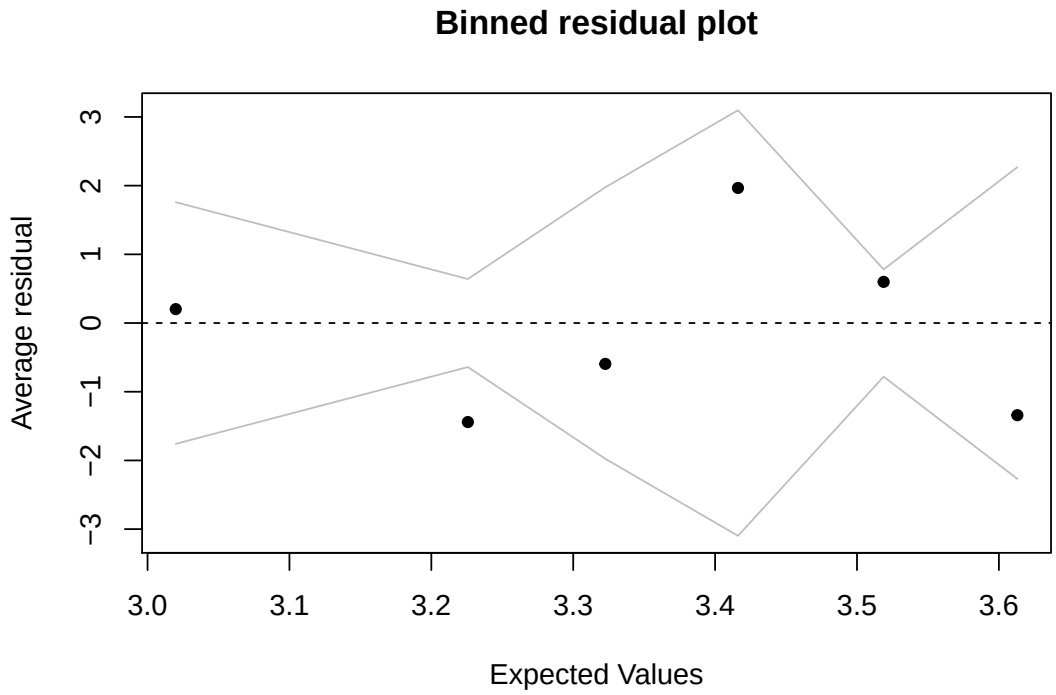


This shows that the trend is not monotonic. Hence when you fit the model and look at the residual you will see a trend.

```
salreg2 <- glm(colonies~log(dose+1), data = salmonella, family = poisson())
summary(salreg2)
```

```
##
## Call:
## glm(formula = colonies ~ log(dose + 1), family = poisson(), data = salmonella)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0764  -1.4488  -0.2306   0.9259   4.7212
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.01989    0.09712  31.095  < 2e-16 ***
## log(dose + 1)  0.08585    0.02018   4.255 2.09e-05 ***
## ---
## 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: 59.629  on 16  degrees of freedom
## AIC: 156.17
##
## Number of Fisher Scoring iterations: 4
```

```
binnedplot(predict(salreg2),resid(salreg2))
```



By adding log transformation on the variable "dose", the model has been improved a lot.

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

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.

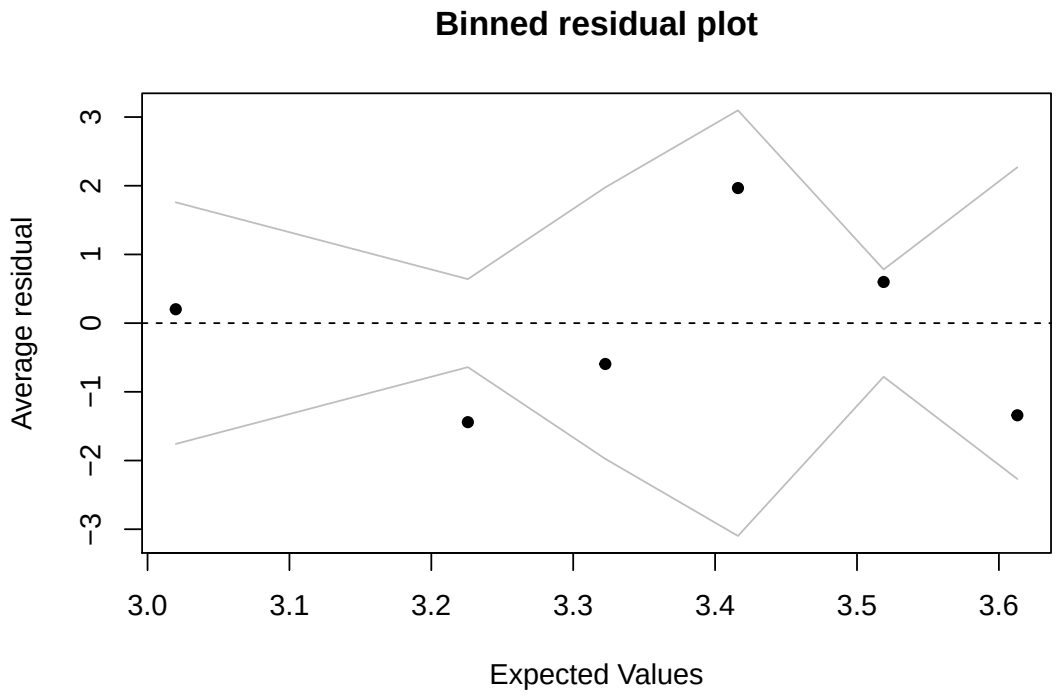
```
salreg3 <- glm(colonies~log(dose+1)+log(dose+1)^2+log(dose+1)^3+log(dose+1)^4, data = salmonella, family
summary(salreg3)
```

```
##
## Call:
## glm(formula = colonies ~ log(dose + 1) + log(dose + 1)^2 + log(dose +
##     1)^3 + log(dose + 1)^4, family = poisson(), data = salmonella)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0764  -1.4488  -0.2306   0.9259   4.7212
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.01989    0.09712  31.095  < 2e-16 ***
## log(dose + 1)  0.08585    0.02018   4.255 2.09e-05 ***
## ---
## 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: 59.629  on 16  degrees of freedom
## AIC: 156.17
##
## Number of Fisher Scoring iterations: 4
```

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.

```
binnedplot(predict(salreg3),resid(salreg3))
```



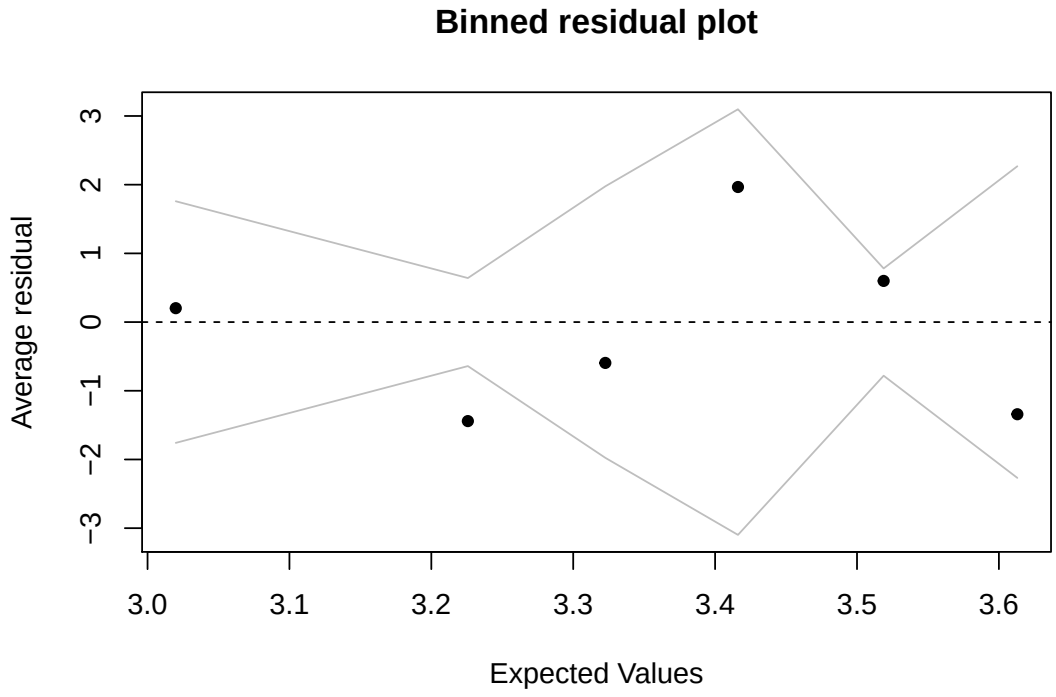
Dispite the fit, the overdispersion still exists so we'd be better off using the quasi Poisson model.

```
salreg4 <- glm(colonies~log(dose+1)+log(dose+1)^2+log(dose+1)^3+log(dose+1)^4, data = salmonella, family
summary(salreg4)
```

```
##
## Call:
## glm(formula = colonies ~ log(dose + 1) + log(dose + 1)^2 + log(dose +
##     1)^3 + log(dose + 1)^4, family = quasipoisson(), data = salmonella)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0764  -1.4488  -0.2306   0.9259   4.7212
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.01989    0.19311   15.64 4.09e-11 ***
## log(dose + 1)  0.08585    0.04012    2.14   0.0481 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for quasipoisson family taken to be 3.953875)
##
##     Null deviance: 78.358  on 17  degrees of freedom
## Residual deviance: 59.629  on 16  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

```
binnedplot(predict(salreg4),resid(salreg4))
```



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(MASS)
shipreg <- glm(incidents~factor(type)+year+period+service, data = ships, family = poisson())
shipreg2 <- MASS::glm.nb(incidents~factor(type)+year+period+service, data = ships)
summary(shipreg)
```

```
##
## Call:
## glm(formula = incidents ~ factor(type) + year + period + service,
```

```
##     family = poisson(), data = ships)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1013  -1.9648  -0.5380   0.9899   4.6212
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -5.706e+00  1.221e+00  -4.673 2.96e-06 ***
## factor(type)B  8.135e-01  2.023e-01   4.021 5.79e-05 ***
## factor(type)C -1.205e+00  3.275e-01  -3.679 0.000234 ***
## factor(type)D -8.595e-01  2.875e-01  -2.989 0.002795 **
## factor(type)E -2.226e-01  2.348e-01  -0.948 0.343173
## year           4.519e-02  1.341e-02   3.370 0.000752 ***
## period         6.055e-02  8.945e-03   6.768 1.30e-11 ***
## service        5.970e-05  7.016e-06   8.509  < 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: 730.25  on 39  degrees of freedom
## Residual deviance: 174.00  on 32  degrees of freedom
## AIC: 287.86
##
## Number of Fisher Scoring iterations: 6
```

```
summary(shipreg2)
```

```
##
## Call:
## MASS::glm.nb(formula = incidents ~ factor(type) + year + period +
##     service, data = ships, init.theta = 1.06309975, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9313  -1.2629  -0.2367   0.4180   1.7281
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -8.5486370  3.0827417  -2.773  0.00555 **
## factor(type)B  0.2232960  0.7052439   0.317  0.75153
## factor(type)C -0.8027046  0.6000777  -1.338  0.18100
## factor(type)D -0.8771070  0.6073869  -1.444  0.14872
## factor(type)E  0.1088337  0.5592650   0.195  0.84571
## year           0.0893065  0.0358743   2.489  0.01279 *
## period         0.0542139  0.0244804   2.215  0.02679 *
## service        0.0001097  0.0000288   3.808  0.00014 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.0631) family taken to be 1)
##
##     Null deviance: 107.480  on 39  degrees of freedom
## Residual deviance:  45.598  on 32  degrees of freedom
```

```
## AIC: 218.41
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.063
##           Std. Err.:  0.392
##
##  2 x log-likelihood:  -200.406
```

```
anova(shipreg,shipreg2)
```

```
## Analysis of Deviance Table
##
## Model 1: incidents ~ factor(type) + year + period + service
## Model 2: incidents ~ factor(type) + year + period + service
##   Resid. Df Resid. Dev Df Deviance
## 1        32    173.996
## 2        32     45.598  0    128.4
```

Taking type A as a baseline, Type B and E ships tend to have more accidents and Type C and D ships tend to have less accidents comparing with Type A ships, among these ships, Type D ships have least accidents. Year, Period and Service years all have positive correlation with accidents.

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
```

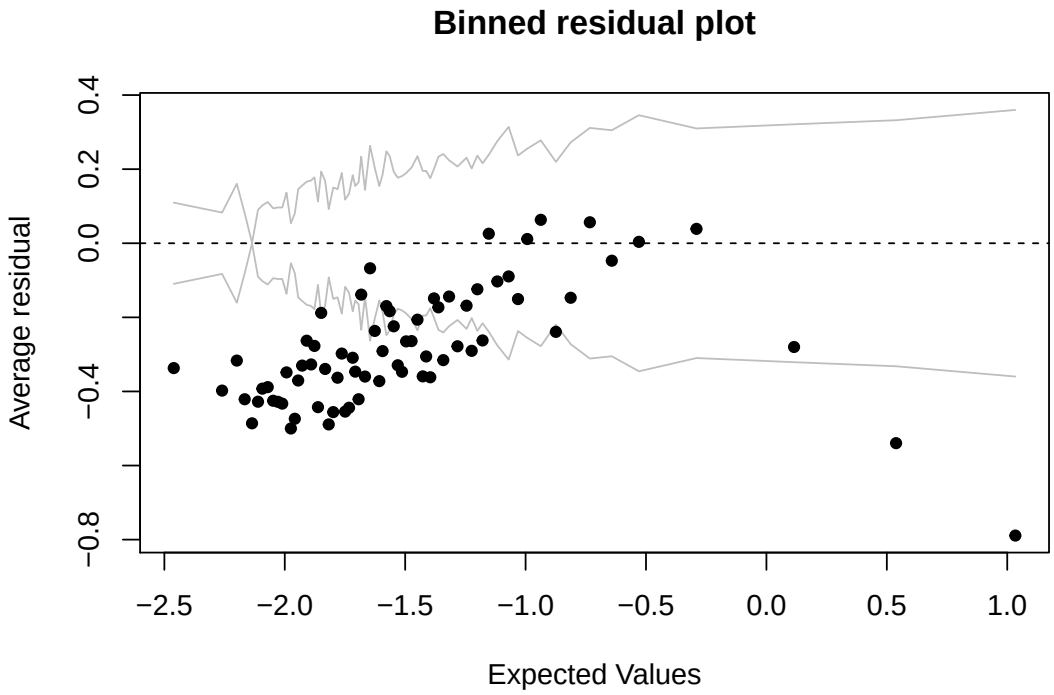
1. Build a Poisson regression model with `doctorco` as the response and `sex`, `age`, `agesq`, `income`, `levyplus`, `freepoor`, `freerepa`, `illness`, `actdays`, `hscore`, `chcond1` and `chcond2` as possible predictor variables. Considering the deviance of this model, does this model fit the data?

```
ahsreg <- glm(doctorco~sex+age+agesq+income+levyplus+freepoor+freerepa+illness+actdays+hscore+chcond1+ch
summary(ahsreg)
```

```
##
## 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
```

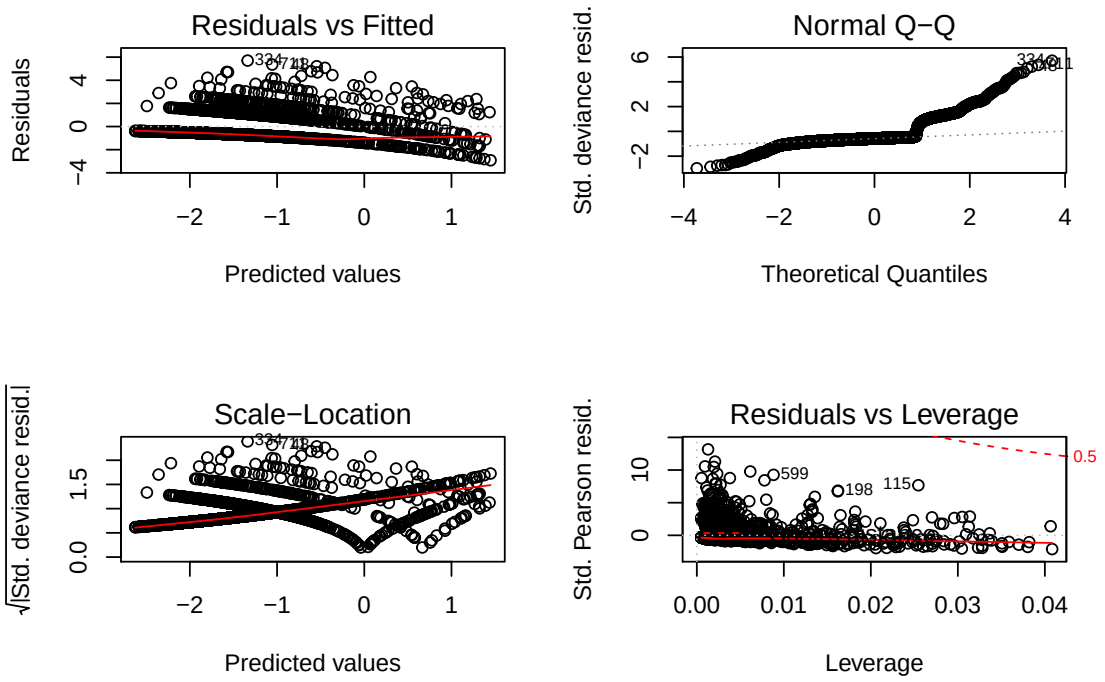
```
binnedplot(predict(ahsreg),resid(ahsreg))
```



From the summary table, we can see the residual deviance of this model is a lot smaller than the null deviance, and most of the variables in the model are statistically significant, from these points I would say the model is pretty good. However, when I took a look at the residual plot, almost half of the residuals are located outside the boundary, and this result led me to consider the model might not fit the data well.

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

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



Since the "doctorco" is a discrete variable, thus there are lines of observations on the plot.

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

```
summary(ahsreg)
```

```
##
## 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
```

```r
#Select the variables that are statistically significant and build another model
ahsreg2 <- glm(doctorco~sex+income+freepoor+illness+actdays+hscore+chcond2+chcond1, data = dvisits, fami
summary(ahsreg2)
```

```
##
## Call:
## glm(formula = doctorco ~ sex + income + freepoor + illness +
##     actdays + hscore + chcond2 + chcond1, family = poisson(),
##     data = dvisits)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8542  -0.6885  -0.5751  -0.4873   5.7388
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.930834   0.081374 -23.728  < 2e-16 ***
## sex          0.187024   0.055103   3.394 0.000689 ***
## income      -0.222436   0.079169  -2.810 0.004960 **
## freepoor    -0.566906   0.172943  -3.278 0.001045 **
## illness      0.189770   0.018164  10.448  < 2e-16 ***
## actdays      0.128010   0.004977  25.722  < 2e-16 ***
## hscore       0.028848   0.009996   2.886 0.003902 **
## chcond2      0.203258   0.079726   2.549 0.010789 *
## chcond1      0.169926   0.063554   2.674 0.007502 **
## ---
## 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: 4389.3  on 5181  degrees of freedom
## AIC: 6738.9
##
## Number of Fisher Scoring iterations: 6
```

A female person with relatively lower income, not covered by government, has more illness in the past 2 weeks, has more days of reduced activity, has a high score in health questionaire and has chronic conditions is more likely to visit doctors more than other people.

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.

```
View(dvisits)
prelast <- predict(ahsreg2, dvisits[5190,])
summary(prelast)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -1.986  -1.986  -1.986  -1.986  -1.986  -1.986
```

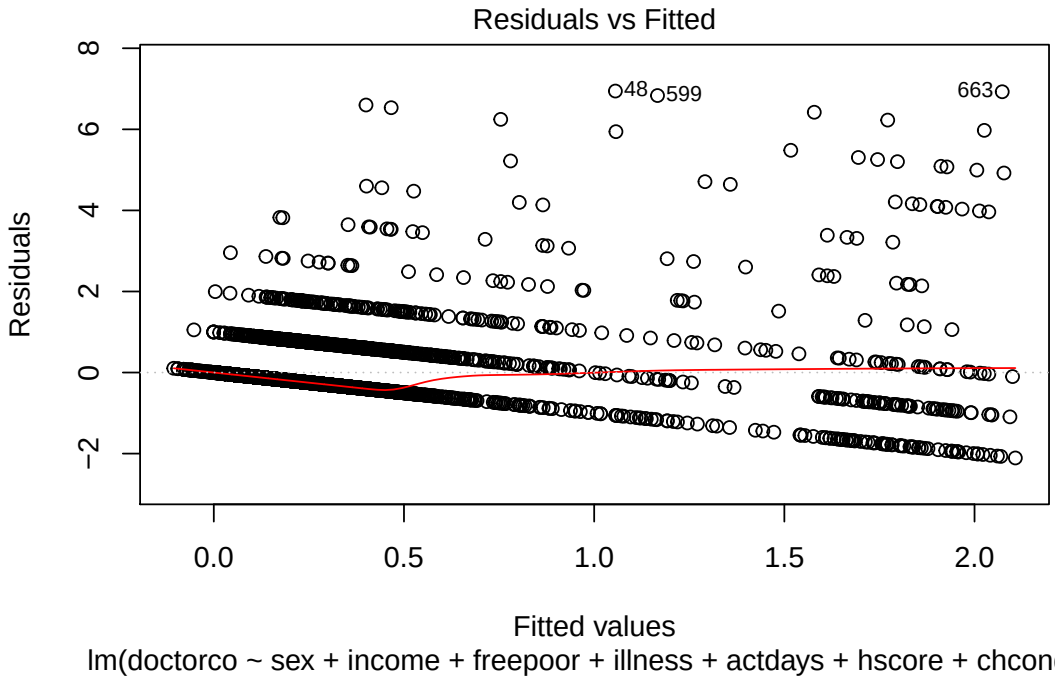
The average visits by this person is negative, so I would say it is very unlikely for this person to visit a doctor.

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

```
ahsreg3 <- lm(doctorco~sex+income+freepoor+illness+actdays+hscore+chcond2+chcond1, data = dvisits)
summary(ahsreg3)
```

```
##
## Call:
## lm(formula = doctorco ~ sex + income + freepoor + illness + actdays +
##     hscore + chcond2 + chcond1, data = dvisits)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1074 -0.2591 -0.1433 -0.0447  6.9442
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.109726   0.028374   3.867 0.000111 ***
## sex          0.050240   0.021025   2.390 0.016903 *
## income      -0.073427   0.028729  -2.556 0.010622 *
## freepoor    -0.148936   0.050228  -2.965 0.003039 **
## illness      0.061378   0.008327   7.371 1.96e-13 ***
## actdays      0.103697   0.003654  28.379  < 2e-16 ***
## hscore       0.015729   0.005169   3.043 0.002355 **
## chcond2      0.065552   0.034799   1.884 0.059657 .
## chcond1      0.027255   0.022751   1.198 0.230969
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7144 on 5181 degrees of freedom
## Multiple R-squared:    0.2,  Adjusted R-squared:  0.1987
## F-statistic: 161.9 on 8 and 5181 DF,  p-value: < 2.2e-16
```

```
# Fit for the Linear Reg.
plot(ahsreg3, which=1)
```




```
# Fit for the poisson reg.
plot(ahsreg2, which=1)
```