Homework 04

Generalized Linear Models

Jinfei Xue

October 6, 2018

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?

```
# First round fupacts
risky_behaviors$fupacts <- round(risky_behaviors$fupacts)
# Creat the possion regression model 1
r_1 <- glm(fupacts ~ women_alone + couples, data = risky_behaviors, fam
ily = poisson)
summary(r_1)
```

```
##
## Call:
## glm(formula = fupacts ~ women_alone + couples, family = poisson,
##     data = risky_behaviors)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.6285  -4.9794  -3.2015   0.9847  27.1502
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.08960    0.01901  162.55   <2e-16 ***
## women_alone -0.57212    0.03023  -18.93   <2e-16 ***
## couples     -0.32243    0.02737  -11.78   <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: 12925  on 431  degrees of freedom
## AIC: 14256
##
## Number of Fisher Scoring iterations: 6
```

```
Anova(r_1)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: fupacts
##             LR Chisq Df Pr(>Chisq)
## women_alone   367.92  1  < 2.2e-16 ***
## couples       138.69  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the regression result, we can see all the three coefficients are significant. Besides, the Anova table shows all of the model terms are significant.

```
n <- nrow(risky_behaviors)
k <- length(r_1$coef)
yhat <- predict(r_1, type = "response")
z <- (risky_behaviors$fupacts - yhat) / sqrt(yhat)
over_dispersion<-sum(z^2)/(n-k)
over_dispersion
```

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

```
pchisq(sum(z^2), n - k)
```

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

The overdispersion ratio is 44.13. And the p-value of the overdispersion test is 1, indicating that the probability is essentially zero that a random variable from a χ^2_{431} distribution would be as large as 19022. Therefore, it looks like there is evidence of 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?

```
# in order to consider offset variable, we select:
subsetrisks <- risky_behaviors[risky_behaviors$bupacts > 0,]
# Model 2
r_2 <- glm(fupacts ~ women_alone + couples + factor(bs_hiv) + factor(se
x),
           offset = log(bupacts), data = subsetrisks, family = poisson)
summary(r_2)
```

```
##
## Call:
## glm(formula = fupacts ~ women_alone + couples + factor(bs_hiv) +
##     factor(sex), family = poisson, data = subsetrisks, offset = log
(bupacts))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -16.315   -3.165   -1.072    2.218   21.552
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -0.03222    0.02250  -1.432    0.152
## women_alone             -0.55581    0.03043 -18.267  < 2e-16 ***
## couples                 -0.40263    0.02804 -14.362  < 2e-16 ***
## factor(bs_hiv)positive -0.32512    0.03573  -9.099  < 2e-16 ***
## factor(sex)man          -0.11843    0.02372  -4.994 5.92e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 10577  on 419  degrees of freedom
## Residual deviance: 10032  on 415  degrees of freedom
## AIC: 11356
##
## Number of Fisher Scoring iterations: 6
```

```
Anova(r_2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: fupacts
##                LR Chisq Df Pr(>Chisq)
## women_alone      342.70  1  < 2.2e-16 ***
## couples          206.32  1  < 2.2e-16 ***
## factor(bs_hiv)    88.90  1  < 2.2e-16 ***
## factor(sex)       24.95  1  5.869e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see this model fits better than the first one. Besides, there exist a small negative intercept in the regression model, which is reasonable.

```
n <- nrow(subsetrisks)
k <- length(r_2$coef)
yhat <- predict(r_2, type = "response")
z <- (subsetrisks$fupacts - yhat) / sqrt(yhat)
over_dispersion<-sum(z^2)/(n-k)
over_dispersion
```

```
## [1] 46.30971

pchisq(sum(z^2), n - k)

## [1] 1
```

The overdispersion ratio is 46.31. And the p-value of the overdispersion test is 1, indicating that it looks like there is still evidence of overdispersion.

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

According to the previous question, an overdispersed Poisson model is the following one:

```
# Model 3
r_3 <- glm(formula = fupacts ~ women_alone + couples +
      factor(bs_hiv) + factor(sex), family = quasipoisson,
    data = subsetrisks, offset = log(bupacts))
display(r_3)

## glm(formula = fupacts ~ women_alone + couples + factor(bs_hiv) +
##     factor(sex), family = quasipoisson, data = subsetrisks, offset =
 log(bupacts))
##                         coef.est coef.se
## (Intercept)             -0.03     0.15
## women_alone             -0.56     0.21
## couples                 -0.40     0.19
## factor(bs_hiv)positive -0.33     0.24
## factor(sex)man          -0.12     0.16
## ---
##   n = 420, k = 5
##   residual deviance = 10032.2, null deviance = 10577.1 (difference =
 544.9)
##   overdispersion parameter = 46.3

# Model 4: interaction
r_4 <- glm(formula = fupacts ~ factor(women_alone + couples) +
      factor(bs_hiv) + factor(sex), family = quasipoisson,
    data = subsetrisks, offset = log(bupacts))
display(r_4)

## glm(formula = fupacts ~ factor(women_alone + couples) + factor(bs_hi
v) +
##     factor(sex), family = quasipoisson, data = subsetrisks, offset =
 log(bupacts))
##                                 coef.est coef.se
## (Intercept)                     -0.03     0.15
## factor(women_alone + couples)1 -0.47     0.17
## factor(bs_hiv)positive          -0.30     0.24
## factor(sex)man                  -0.12     0.16
## ---
```

```
##   n = 420, k = 4
##   residual deviance = 10056.8, null deviance = 10577.1 (difference =
 520.3)
##   overdispersion parameter = 46.9
```

The coefficients of treatment interaction and bs_hiv are less significant. And the overdispersion parameter is larger, which means less overdispersion.

```
anova(r_4, r_3,test="Chisq")

## Analysis of Deviance Table
##
## Model 1: fupacts ~ factor(women_alone + couples) + factor(bs_hiv) +
factor(sex)
## Model 2: fupacts ~ women_alone + couples + factor(bs_hiv) + factor(s
ex)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       416      10057
## 2       415      10032  1   24.621   0.4659
```

The likelihood ratio test statistic is $\chi^2 = 24.621$ with a p-value=0.4659. Hence, we do not have relatively strong evidence in favor of rejecting H_0. That is to say, the two models do not have significant 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?

This does give me some concerns because the couples and women_alone variables interacting in women won't be i.i.d. And we could have correlated errors since the couples data is recorded twice for fupacts if they are in the together group.

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)

Take "wells in Bangladesh" data for example:

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

Then, we use logit and probit model to fit the data seperately.

```
# Logit model
r_logit = glm(switch ~ log.arsenic + dist + educ,
              family=binomial(link="logit"), data=wells)
display(r_logit)
```

```
## glm(formula = switch ~ log.arsenic + dist + educ, family = binomial
(link = "logit"),
##     data = wells)
##             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 = 2
39.9)
```

```
# Probit model
r_probit = glm(switch ~ log.arsenic + dist + educ,
               family=binomial(link="probit"), data=wells)
display(r_probit)
```

```
## glm(formula = switch ~ log.arsenic + dist + educ, family = binomial
(link = "probit"),
##     data = wells)
##             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 = 2
39.8)
```

According to the regression result, the logistic regression model is:

$Pr(swithch = 1) = logit^{-1}(0.32 + 0.89log.arsenic - 0.01dist + 0.04educ)$

The probit regression model is:

$Pr(swithch = 1) = \phi(0.19 + 0.54log.arsenic - 0.01dist + 0.03educ)$

```
coef_logit<-r_logit$coefficients
coef_probit<-r_probit$coefficients
diff<-1.6*coef_probit-coef_logit
cat("The differences relative to logit's coefficients are \n", diff)
```

```
## The differences relative to logit's coefficients are
##  -0.008558448 -0.01864176 0.0002012497 -0.0003712089
```

We can see that coefficients in probit model scaled by factor of 1.6 are approximately same as coefficients in logistic model.

Comparing logit and probit:

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

```
# Identify the features of variables
summary(wells)
```

```
##      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
```

```
# Construct a dataset
log.arsenic = runif(10,-0.6733,2.2670)
dist = runif(10,0.387,339.531)
educ = sample(0:17,10,replace = T)
predict_data = data.frame(log.arsenic,dist,educ)
```

```
# Predict fitted value seperately
predict(r_logit,predict_data)
```

```
##          1          2          3          4          5          6 
## -0.5273209 -1.1279511 -0.7712256 -2.1951716 -1.1180237 -0.9641385 
##          7          8          9         10 
##  0.9130341  2.0523809 -0.4225497 -1.0689286
```

```
predict(r_probit,predict_data)
```

```
##          1          2          3          4          5          6 
## -0.3212993 -0.6902924 -0.4712837 -1.3430871 -0.6808637 -0.5888022 
##          7          8          9         10 
##  0.5603810  1.2595410 -0.2562356 -0.6510883
```

We can see the logit and probit models 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$re78 = (lalonde$re78 - mean(lalonde$re78)) / sd(lalonde$re78)
tobit = vglm(re78 ~ educ + factor(treat) + factor(black) + factor(marri
ed) + age,
             tobit(Upper = 121174), data = lalonde)

## Warning in eval(slot(family, "initialize")): replacing response valu
es less
## than 'Lower' by 'Lower'

summary(tobit)

##
## Call:
## vglm(formula = re78 ~ educ + factor(treat) + factor(black) +
##     factor(married) + age, family = tobit(Upper = 121174), data = la
londe)
##
##
## Pearson residuals:
##                Min      1Q  Median     3Q    Max
## mu       -1.668 -0.8367 -0.1080 0.7033  12.77
## loge(sd) -1.008 -0.6089 -0.2709 0.1500 129.31
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept):1    -1.7285975  0.0426590 -40.521  < 2e-16 ***
## (Intercept):2    -0.1377315  0.0070147 -19.635  < 2e-16 ***
## educ              0.0835654  0.0025468  32.812  < 2e-16 ***
## factor(treat)1   -0.5338273  0.1070334  -4.987 6.12e-07 ***
## factor(black)1   -0.1493789  0.0251684  -5.935 2.94e-09 ***
## factor(married)1  0.5336860  0.0184733  28.890  < 2e-16 ***
## age               0.0126347  0.0007161  17.644  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  2
##
```

```
## Names of linear predictors: mu, loge(sd)
##
## Log-likelihood: -18490.16 on 37327 degrees of freedom
##
## Number of iterations: 12
##
## No Hauck-Donner effect found in any of the estimates
```

educ: With every 1 level increase in education level, one's expected z-score earning on 1978 would increase by 0.08 unit while holding all other variables in the model constant.

treat: If someone from NSW group, one's expected z-score earning on 1978 would be 0.53 unit lower than someone with the all same condition who from CPS or PSID group.

balck: If someone is black, one's expected z-score earning on 1978 would be 0.15 unit lower than someone with the all same condition who is not black.

married: If someone is married, one's expected z-score earning on 1978 would be 0.53 unit higher than someone with the all same condition who is not married.

age: With every 1 increase in the age, one's expected z-score earning on 1978 would increase by 0.012 unit while holding all the other variables in the model constant.

Robust linear regression using the t model:

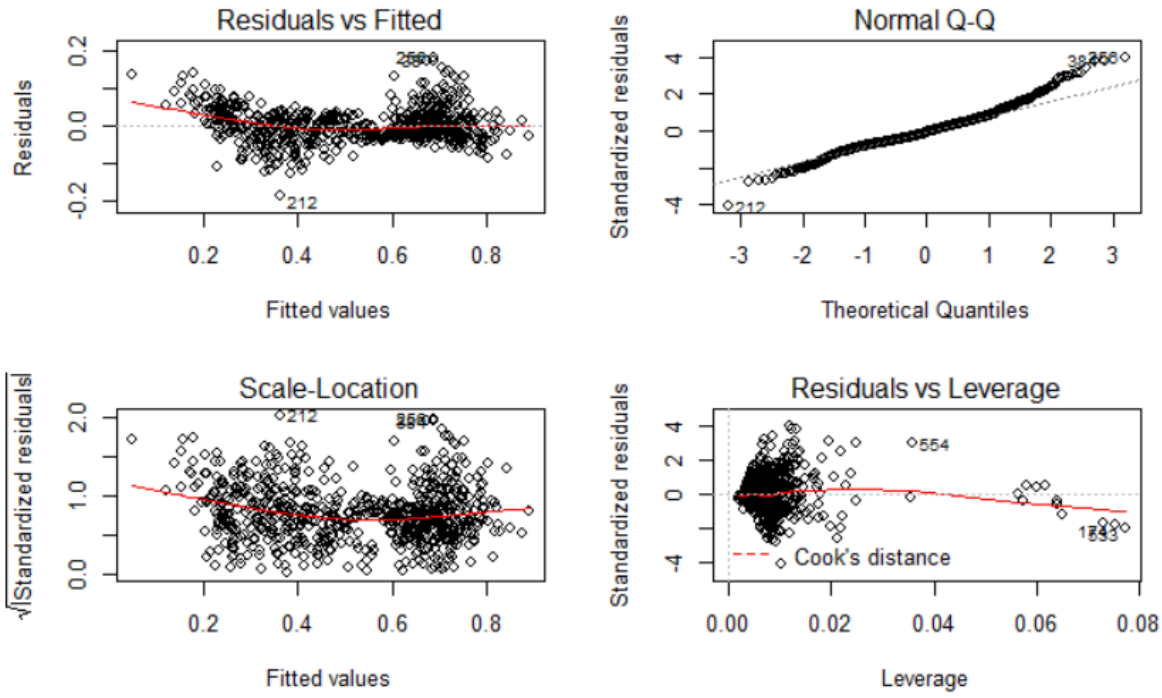
The csv file `congress` has the votes for the Democratic and Republican candidates in each U.S. congressional district in between 1896 and 1992, along with the parties' vote proportions and an indicator for whether the incumbent was running for reelection. For your analysis, just use the elections in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

```
sub_con <- congress %>% filter(year==1986 | year==1988, contested == TR
UE)
r_linear <- lm(Dem_pct ~ x1 + x2 + incumbent + Dem_vote + Rep_vote, dat
a = sub_con)
summary(r_linear)

##
## Call:
## lm(formula = Dem_pct ~ x1 + x2 + incumbent + Dem_vote + Rep_vote,
##     data = sub_con)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.186942 -0.027649 -0.002967  0.023232  0.179001
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.452e-01  7.738e-03  70.460   <2e-16 ***
## x1           1.847e-04  8.010e-05   2.306   0.0214 *
## x2          -2.058e-04  1.212e-04  -1.698   0.0899 .
## incumbent    3.094e-02  3.485e-03   8.881   <2e-16 ***
## Dem_vote     2.089e-06  6.576e-08  31.767   <2e-16 ***
## Rep_vote    -2.549e-06  6.018e-08 -42.361   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04566 on 697 degrees of freedom
## Multiple R-squared:  0.9433, Adjusted R-squared:  0.9429
## F-statistic:  2317 on 5 and 697 DF,  p-value: < 2.2e-16

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



According to the regression result, we can see all of the efficients are significant, adjusted R-squared is very close to 1 and p-value of F-statistic is very small, which means the model fit the data well.

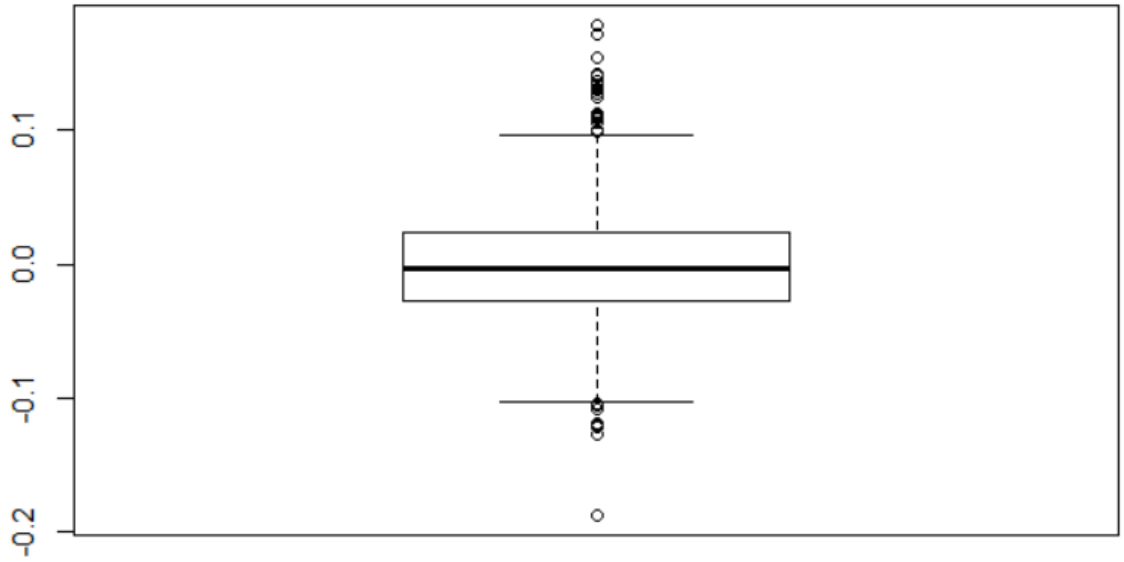
2. Fit a t-regression model predicting 1988 Democratic vote share from the other variables and assess model fit; to fit this model in R you can use the `vglm()` function in the VGLM package or `tlm()` function in the hett package.

```
r_t <- tlm( Dem_pct ~ x1 + x2 + incumbent + Dem_vote+ Rep_vote, data =
 sub_con)
summary(r_t)

## Location model :
##
## Call:
## tlm(lform = Dem_pct ~ x1 + x2 + incumbent + Dem_vote + Rep_vote,
##     data = sub_con)
##
## Residuals:
##       Min         1Q     Median         3Q        Max
## -0.198032  -0.020810  -0.002154   0.024724   0.169722
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.606e-01  6.472e-03  86.619  < 2e-16 ***
## x1           1.655e-04  6.701e-05   2.469   0.0138 *
## x2          -1.894e-04  1.014e-04  -1.868   0.0622 .
## incumbent    2.280e-02  2.915e-03   7.821 1.94e-14 ***
## Dem_vote     2.045e-06  5.501e-08  37.172  < 2e-16 ***
## Rep_vote    -2.695e-06  5.034e-08 -53.540  < 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_pct ~ x1 + x2 + incumbent + Dem_vote + Rep_vote,
##     data = sub_con)
##
## Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9999  -1.6199  -0.7946   1.1656   5.4461
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.93555    0.07543  -91.94   <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 : 18 in 0
## Heteroscedastic t Likelihood : 1200.77
```

3. Which model do you prefer?

```
boxplot(r_linear$residuals)
```



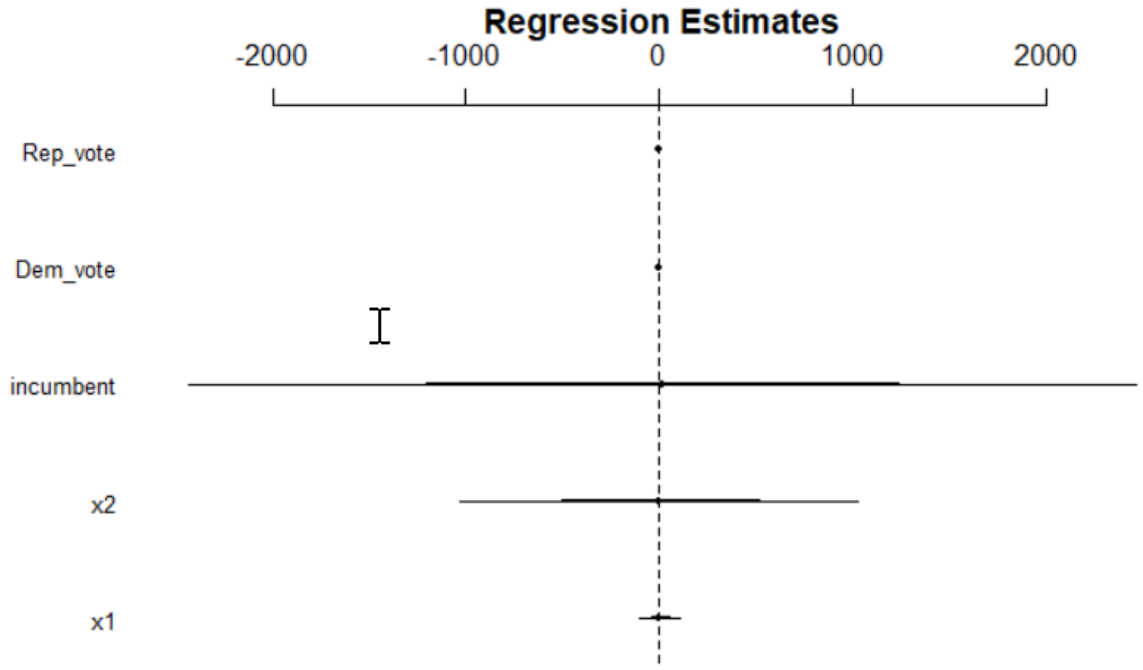
From the boxplot, we can see in linear regression model, there exist occasional very large errors, so it is more appropriate to use t-model rather than normal distribution for the errors.

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.

```
# Creat logistic model
sub_con$Dem_win <- ifelse(sub_con$Dem_vote>sub_con$Rep_vote, 1, 0)
r_logit <- glm(Dem_win ~ x1 + x2 + incumbent + Dem_vote+ Rep_vote,
           data=sub_con, family=binomial(link="logit"))

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

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

summary(r_logit)

##
## Call:
## glm(formula = Dem_win ~ x1 + x2 + incumbent + Dem_vote + Rep_vote,
##     family = binomial(link = "logit"), data = sub_con)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -1.308e-03  -2.000e-08   2.000e-08   2.000e-08   1.304e-03
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   10.05514 3871.30147   0.003    0.998
## x1             0.06440   51.51684   0.001    0.999
## x2            -0.63589  513.91909  -0.001    0.999
## incumbent     12.53249 1224.53346   0.010    0.992
## Dem_vote       0.02806    0.70364   0.040    0.968
## Rep_vote      -0.02818    0.70707  -0.040    0.968
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9.6655e+02  on 702  degrees of freedom
## Residual deviance: 5.7726e-06  on 697  degrees of freedom
## AIC: 12
##
## Number of Fisher Scoring iterations: 25

coefplot(r_logit)
```



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

head(salmonella)

##   colonies dose
## 1       15    0
## 2       21    0
## 3       29    0
## 4       16   10
## 5       18   10
## 6       21   10
```

```r
# Creat possion regression model
rp_sal <- glm(colonies ~ dose, family = poisson, data = salmonella)
summary(rp_sal)
```

```
##
## 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
```

```r
# Compute overdispersion factor
n <- nrow(salmonella)
k <- length(rp_sal$coef)
yhat <- predict(rp_sal, type = "response")
z <- (salmonella$colonies - yhat) / sqrt(yhat)
over_dispersion<-sum(z^2)/(n-k)
over_dispersion
```

```
## [1] 5.087258
```

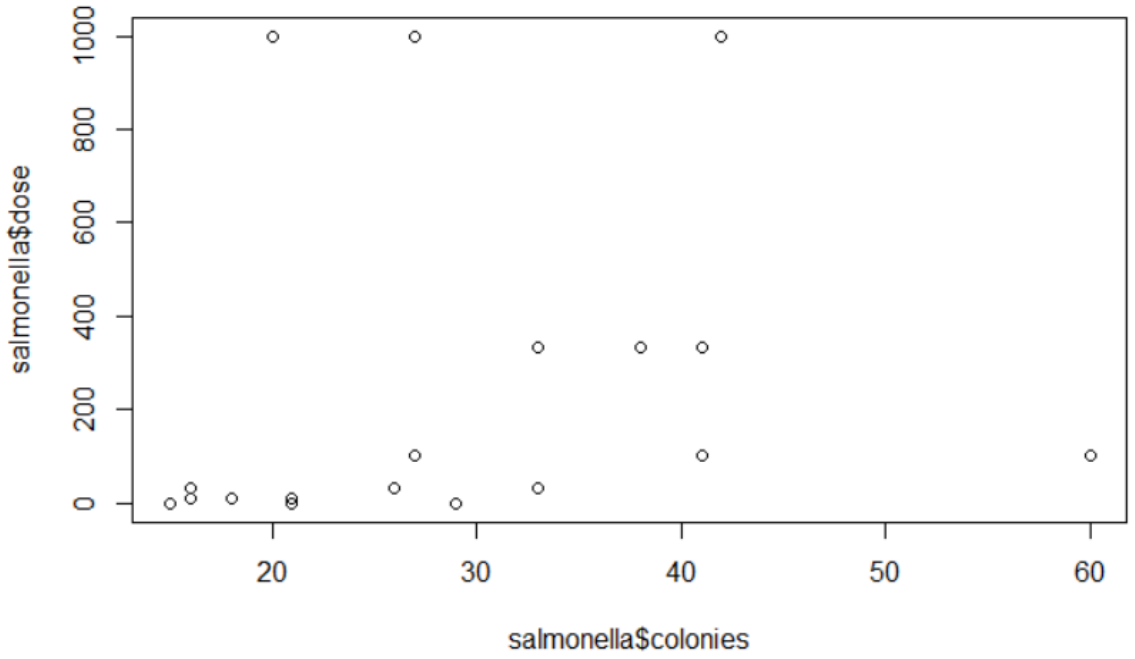
```r
pchisq(sum(z^2), n - k)
```

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

We can see the p-value of the overdispersion test is 1, indicating that it looks like there is evidence of overdispersion. Therefore, Poisson GLM is inadequate and that some overdispersion must be allowed for. We then consider the overdispersion by quasipoisson.

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.

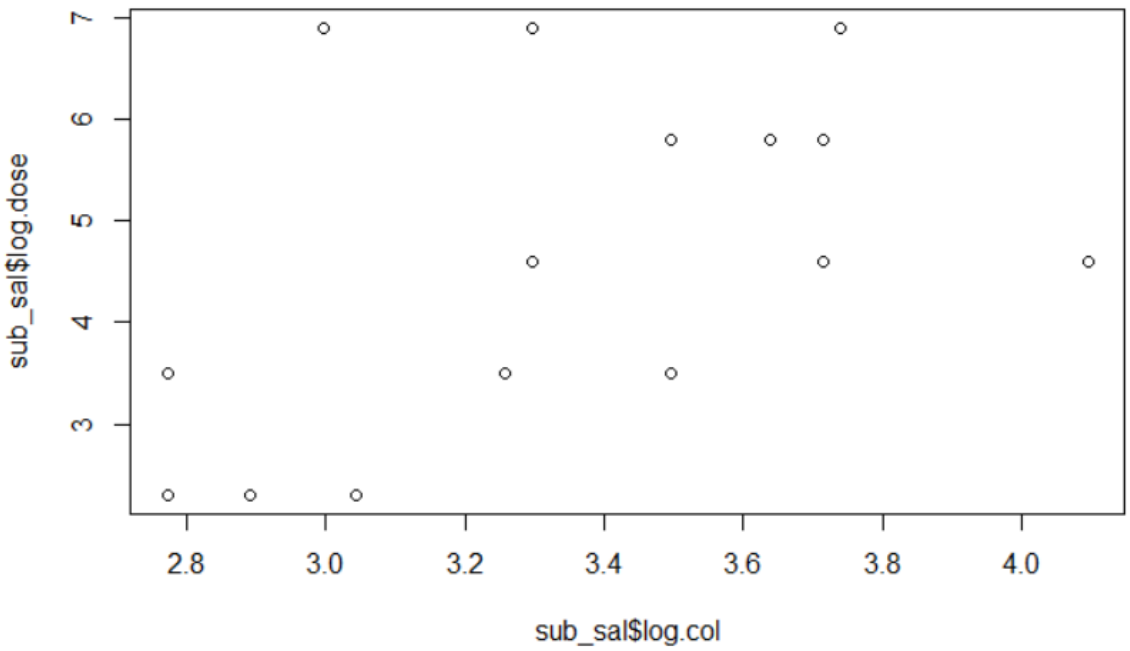
```r
plot(salmonella$colonies, salmonella$dose)
```



From the above plot, we can see that the number of colonies as a function of dose is not monotonic especially around the dose of 1000.

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.

```
# exclude ships with 0 months of service
sub_sal <- subset(salmonella, dose > 0)
# Log transformaiton
sub_sal$log.col <- log(sub_sal$colonies)
sub_sal$log.dose <- log(sub_sal$dose)
plot(sub_sal$log.col,sub_sal$log.dose)
```



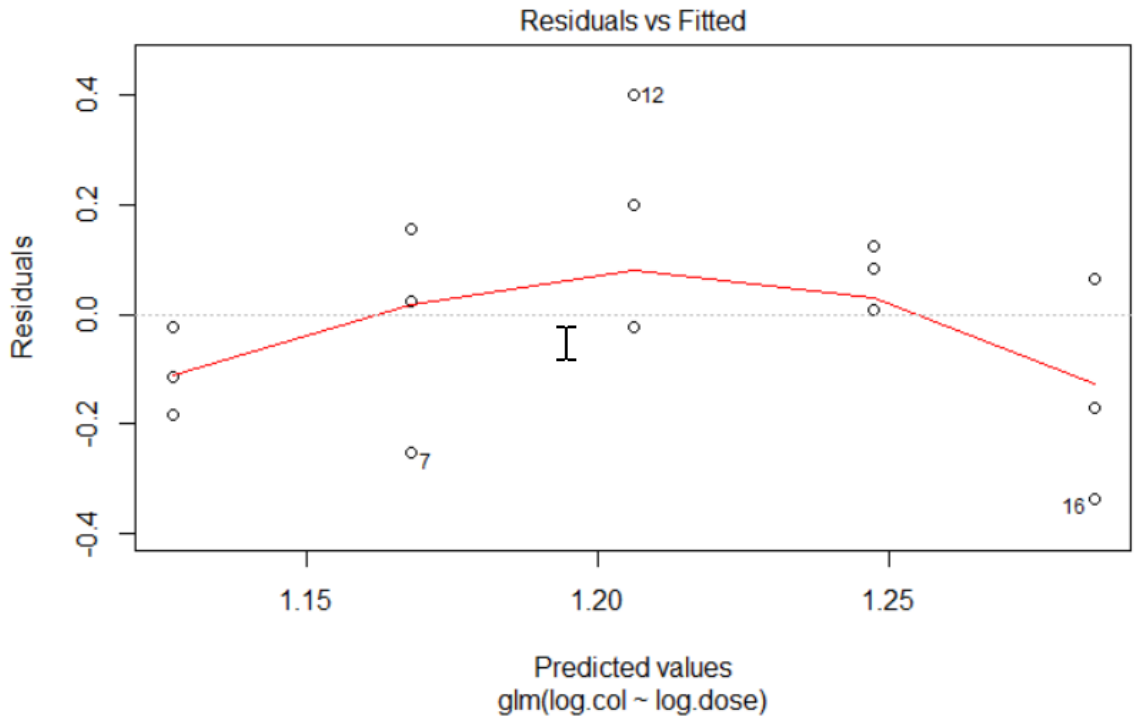
This shows that the trend is not monotonic. Hence when you fit the model and look at the residual you will see a trend.

```
rp2_sal <- glm(log.col ~ log.dose, family = poisson, data = sub_sal)

summary(rp2_sal)

##
## Call:
## glm(formula = log.col ~ log.dose, family = poisson, data = sub_sal)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.33609  -0.14163   0.00807   0.10307   0.39837
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.04759    0.43260   2.422   0.0155 *
## log.dose     0.03441    0.08673   0.397   0.6915
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 0.66005  on 14  degrees of freedom
## Residual deviance: 0.50242  on 13  degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 4
```

We can see the coefficient of log.dose is not significant.

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

```
plot(rp2_sal, which=1)
```



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.

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.

```
# Creat quasipoisson regression model
rq_sal <- glm(log.col ~ log.dose, family = quasipoisson, data = sub_sal)
summary(rq_sal)

##
## Call:
## glm(formula = log.col ~ log.dose, family = quasipoisson, data = sub_
sal)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.33609  -0.14163   0.00807   0.10307   0.39837
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.04759    0.08523  12.291 1.57e-08 ***
## log.dose     0.03441    0.01709   2.014   0.0652 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 0.03881927)
##
##     Null deviance: 0.66005  on 14  degrees of freedom
## Residual deviance: 0.50242  on 13  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

We can see the correction for overdispersion is to multiply all regression standard errors by $\sqrt{5.087279} = 2.2555$.

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
head(ships)
```

```
##   type year period service incidents
## 1    A   60     60     127         0
## 2    A   60     75      63         0
## 3    A   65     60    1095         3
## 4    A   65     75    1095         4
## 5    A   70     60    1512         6
## 6    A   70     75    3353        18
```

```
head(ships)
```

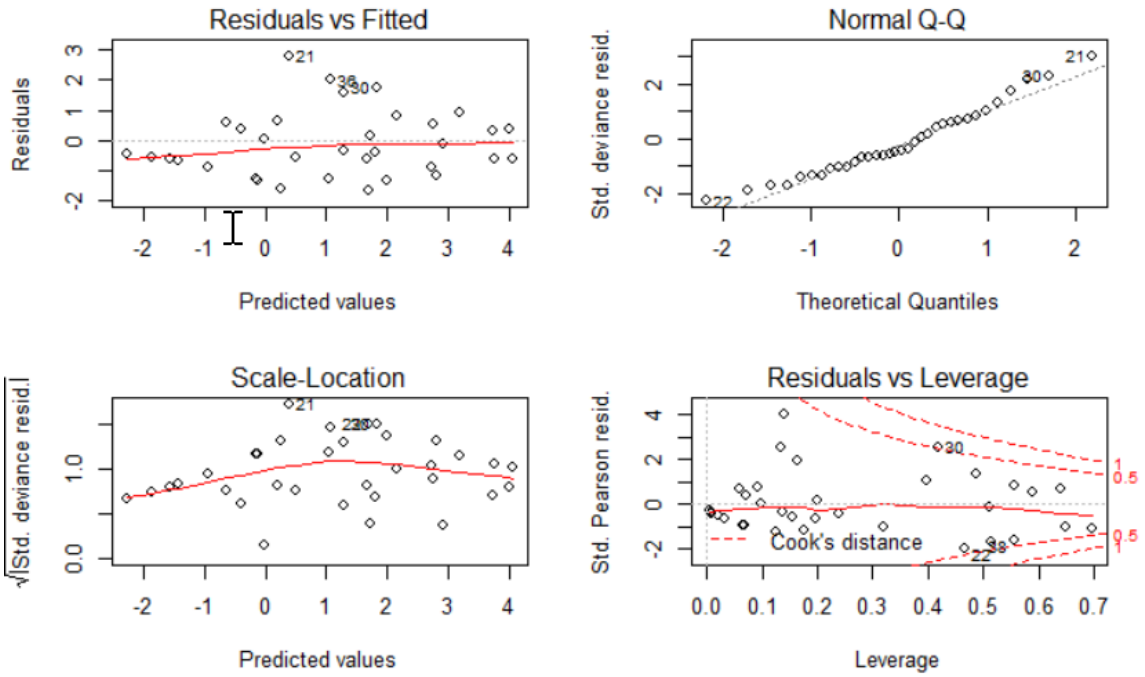
```
##   type year period service incidents
## 1    A   60     60     127         0
## 2    A   60     75      63         0
## 3    A   65     60    1095         3
## 4    A   65     75    1095         4
## 5    A   70     60    1512         6
## 6    A   70     75    3353        18
```

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

```r
# exclude ships with 0 months of service
ships2 <- subset(ships, service > 0)

# convert the period and year variables to factors
ships2$year <- as.factor(ships2$year)
ships2$period <- as.factor(ships2$period)
ships2$type <- as.factor(ships2$type)

# Creat a poisson regression model including all the variables
r_ships <- glm(incidents ~ type + year + period,
               family = poisson, data = ships2, offset = log(service))
summary(r_ships)
```

```
##
## Call:
## glm(formula = incidents ~ type + year + period, family = poisson,
##     data = ships2, offset = log(service))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6768  -0.8293  -0.4370   0.5058   2.7912
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.40590    0.21744 -29.460  < 2e-16 ***
## typeB       -0.54334    0.17759  -3.060  0.00222 **
## typeC       -0.68740    0.32904  -2.089  0.03670 *
## typeD       -0.07596    0.29058  -0.261  0.79377
## typeE        0.32558    0.23588   1.380  0.16750
## year65       0.69714    0.14964   4.659 3.18e-06 ***
## year70       0.81843    0.16977   4.821 1.43e-06 ***
## year75       0.45343    0.23317   1.945  0.05182 .
## period75     0.38447    0.11827   3.251  0.00115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 146.328  on 33  degrees of freedom
## Residual deviance:  38.695  on 25  degrees of freedom
## AIC: 154.56
##
## Number of Fisher Scoring iterations: 5
```

```r
par(mfrow=c(2,2))
plot(r_ships)
```



From the regression result and plots, we can see the model fits the data well, and the poisson model is:

$$log(\lambda)$$
$$= -6.41 - 0.54typeB - 0.69typeC - 0.08typeD + 0.33typeE + 0.70year65$$
$$+ 0.82year70 + 0.45year75 + 0.38period75 + log(service)$$

intercept:Since it is impossible that both type and year equal to 0, we will not try to interpret the constant term.

The coefficients of type: Compared to the baseline category typeA with the same year, we see that typeB has 54% fewer incidents, typeC has 69% fewer incidents and typeE has 33% more incidents, in proportion to service rates.

The coefficients of year: Compared to the baseline category year60 with the same type, we see that year65 has 70% more incidents, year70 has 82% more incidents and year75 has 45% more incidents, in proportion to service rates.

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`,

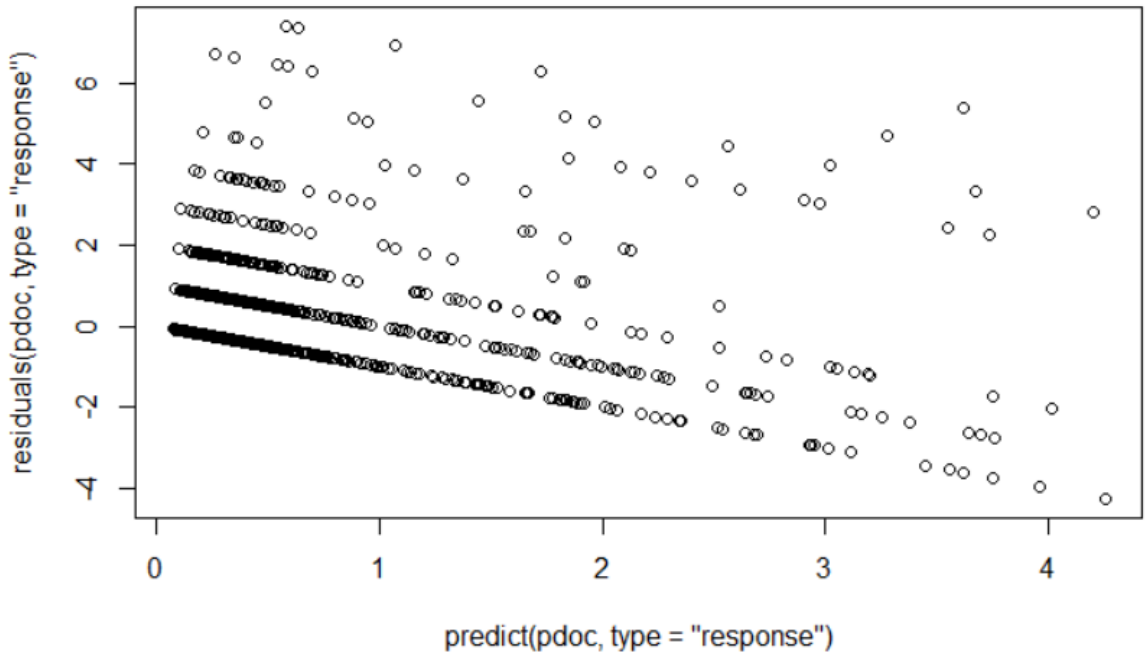
chcond1 and chcond2 as possible predictor variables. Considering the deviance of this model, does this model fit the data?

```
pdoc <- glm(doctorco ~ sex + age + agesq + income + levyplus +
              freepoor + freerepa + illness + actdays + hscore + chcond
1 + chcond2,
            family=poisson, data=dvisits)
deviance(pdoc)

## [1] 4379.515
```

We can see the deviance of this model is 4382.383, which is so large that we can conclude this model does not fit the data very well.

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

```
plot(predict(pdoc, type="response"), residuals(pdoc, type="response"))
```



The lines appear on the plot because the response reisuals are given by $y_i - \hat{y}_i$ and y_i only takes on finitely many values. Each line corresponds to a different possible value for y_i.

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

```
display(pdoc)

## glm(formula = doctorco ~ sex + age + agesq + income + levyplus +
##     freepoor + freerepa + illness + actdays + hscore + chcond1 +
##     chcond2, family = poisson, data = dvisits)
##             coef.est coef.se
## (Intercept) -2.22     0.19
```

```
## sex          0.16     0.06
## age          1.06     1.00
## agesq       -0.85     1.08
## income      -0.21     0.09
## levyplus     0.12     0.07
## freepoor    -0.44     0.18
## freerepa     0.08     0.09
## illness      0.19     0.02
## actdays      0.13     0.01
## hscore       0.03     0.01
## chcond1      0.11     0.07
## chcond2      0.14     0.08
## ---
##   n = 5190, k = 13
##   residual deviance = 4379.5, null deviance = 5634.8 (difference = 1
255.3)
```

So we can obtain:

$$log(\lambda)$$
$$= -2.22 + 0.16sex + 1.06age - 0.85agesq - 0.21income + 0.12levyplus$$
$$- 0.44freepoor + 0.08freerepa + 0.19illness + 0.13actdays + 0.03hscore$$
$$+ 0.11chcondl + 0.14chcond2$$

Because λ is the mean of the number of consultations with a doctor or specialist in the past 2 weeks, when λ increases, its mean also increases. Therefore, if a person is a female with large age, low income, covered by private health insurance fund for private patient in public hospital, not covered by government because low income, recent immigrant and unemployed but because of old-age or disability pension or invalid veteran or family of deceased veteran, with large number of illnesses in past 2 weeks and large number of days of reduced activity in past two weeks due to illness or injury, with high health questionnaire score and chronic conditions, then the person visits the doctor the most under my 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.

```
lamda=predict(pdoc,dvisits[5190,],type="response")
dpois(0:3, as.numeric(lamda))

## [1] 0.8578005057 0.1315726303 0.0100905496 0.0005159087
```

$$Pr(visit = 0) = 0.8578005057$$

$$Pr(visit = 1) = 0.1315726303$$

$$Pr(visit = 2) = 0.0100905496$$

$$Pr(visit = 3) = 0.0005159087$$

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

```
gdoc <- glm(doctorco ~ sex + age + agesq + income + levyplus +
              freepoor + freerepa + illness + actdays + hscore + chcond
1 + chcond2,
            family=gaussian, data=dvisits)
display(gdoc)

## glm(formula = doctorco ~ sex + age + agesq + income + levyplus +
##     freepoor + freerepa + illness + actdays + hscore + chcond1 +
##     chcond2, family = gaussian, data = dvisits)
##             coef.est coef.se
## (Intercept)  0.03     0.07
## sex          0.03     0.02
## age          0.20     0.41
## agesq       -0.06     0.46
## income      -0.06     0.03
## levyplus     0.04     0.02
## freepoor    -0.10     0.05
## freerepa     0.03     0.04
## illness      0.06     0.01
## actdays      0.10     0.00
## hscore       0.02     0.01
## chcond1      0.00     0.02
## chcond2      0.04     0.04
## ---
##   n = 5190, k = 13
##   residual deviance = 2638.3, null deviance = 3305.5 (difference = 6
67.1)
##   overdispersion parameter = 0.5
##   residual sd is sqrt(overdispersion) = 0.71

plot(predict(gdoc, type="response"), residuals(gdoc, type="response"))
```