Homework 04

Generalized Linear Models

Guangyan Yu

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?

```
risky_behaviors$fupacts<-round(risky_behaviors$fupacts)
model<-glm(fupacts~factor(women_alone) + factor(couples),data=risky_behaviors, family = poisson)
display(model)
```

```
## glm(formula = fupacts ~ factor(women_alone) + factor(couples),
##     family = poisson, data = risky_behaviors)
##                       coef.est coef.se
## (Intercept)            3.09     0.02
## factor(women_alone)1 -0.57     0.03
## factor(couples)1     -0.32     0.03
## ---
##   n = 434, k = 3
##   residual deviance = 12925.5, null deviance = 13298.6 (difference = 373.1)
```

```
#overdispersion
predicted<-predict(model,type="response")
z<-(risky_behaviors$fupacts-predicted)/sqrt(predicted)
n<-nrow(risky_behaviors)
k<-length(model$coef)
cat("The overdispersion ratio is",sum(z^2)/(n-k))
```

```
## The overdispersion ratio is 44.13458
```

The deviance of this model is 373.1 smaller than the null model, so it fits the data better than null model.

The overdispersion ratio is 44.13458

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?

```
subsetrisks <- risky_behaviors[risky_behaviors$bupacts > 0,]
model<-glm(fupacts~factor(women_alone)+factor(couples)+factor(bs_hiv)+factor(sex),family=poisson, data=
display(model)
```

```
## glm(formula = fupacts ~ factor(women_alone) + factor(couples) +
```

```
##     factor(bs_hiv) + factor(sex), family = poisson, data = subsetrisks,
##     offset = log(bupacts))
##                         coef.est coef.se
## (Intercept)             -0.03     0.02
## factor(women_alone)1    -0.56     0.03
## factor(couples)1        -0.40     0.03
## factor(bs_hiv)positive -0.33     0.04
## factor(sex)man          -0.12     0.02
## ---
##   n = 420, k = 5
##   residual deviance = 10032.2, null deviance = 10577.1 (difference = 544.9)
```

```
#overdispersion
n<-nrow(subsetrisks)
k<-length(model$coef)
predicted<-predict(model,type="response")
z<-(subsetrisks$fupacts-predicted)/sqrt(predicted)
cat("The overdispersion ratio is",sum(z^2)/(n-k),"\n")
```

```
## The overdispersion ratio is 46.30971
```

The residuak deviance is smaller than the first model, so that this model is better than the first one.

The overdispersion ratio is 46.30971, so that this model is still overdispersed.

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

```
model_over<-glm(fupacts~factor(women_alone)+factor(couples)+factor(bs_hiv)+factor(sex),family=quasipois
display(model_over)
```

```
## glm(formula = fupacts ~ factor(women_alone) + factor(couples) +
##     factor(bs_hiv) + factor(sex), family = quasipoisson, data = subsetrisks,
##     offset = log(bupacts))
##                         coef.est coef.se
## (Intercept)             -0.03     0.15
## factor(women_alone)1    -0.56     0.21
## factor(couples)1        -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
```

```
#effectiveness of intervention
model2<-glm(fupacts~factor(women_alone+couples)+factor(bs_hiv)+factor(sex),family=quasipoisson, data=su
display(model2)
```

```
## glm(formula = fupacts ~ factor(women_alone + couples) + factor(bs_hiv) +
##     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
```

```
anova(model2,model_over)
```

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

Thourgh the quasipoisson model, the factor women_alone and couples are significant because 0 is not in their confidence intervals.

Through anova between model_over and model2, we know that the difference between the women alone group and the couples group looks significant.

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. We could have correlated errors since the couples data is recorded twice for fupacts if they are in the together group. Additionally, using only women may muddle our conclusions. The data may be telling us that working with women is a more effective treatment, or it may be telling us that working with one member of a couple, alone, is a more effective treatment. Without seeing outcomes of men being advised alone, we cannot distinguish between these two conclusions.

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)

```
nes5200<-read.dta("http://www.stat.columbia.edu/~gelman/arm/examples/nes/nes5200_processed_voters_realic
nes5200_dt <- data.table(nes5200)
  yr <- 1992
nes5200_dt_s<-nes5200_dt[ year==yr & presvote %in% c("1. democrat","2. republican")& !is.na(income)]
nes5200_dt_s<-nes5200_dt_s[,vote_rep:=1*(presvote=="2. republican")]
nes5200_dt_s$income <- droplevels(nes5200_dt_s$income)
#logit
model_logit<-glm(vote_rep~factor(gender) + factor(race) + factor(educ1),family = binomial(link = "logit"
display(model_logit)
```

```
## glm(formula = vote_rep ~ factor(gender) + factor(race) + factor(educ1),
##     family = binomial(link = "logit"), data = nes5200_dt_s)
##                                                 coef.est coef.se
## (Intercept)                                     -0.60     0.30
## factor(gender)2. female                         -0.11     0.13
## factor(race)2. black                            -2.74     0.37
## factor(race)3. asian                             0.65     0.51
## factor(race)4. native american                  -0.18     0.38
## factor(race)5. hispanic                         -0.48     0.30
## factor(educ1)2. high school (12 grades or fewer, incl  0.36     0.31
## factor(educ1)3. some college(13 grades or more,but no  0.56     0.32
## factor(educ1)4. college or advanced degree (no cases   0.72     0.31
## ---
```

```
##   n = 1179, k = 9
##   residual deviance = 1455.3, null deviance = 1591.2 (difference = 136.0)
```

```
#probit
model_probit<-glm(vote_rep~factor(gender) + factor(race) + factor(educ1),family = binomial(link = "probi
display(model_probit)
```

```
## glm(formula = vote_rep ~ factor(gender) + factor(race) + factor(educ1),
##     family = binomial(link = "probit"), data = nes5200_dt_s)
##                                                 coef.est coef.se
## (Intercept)                                     -0.38     0.18
## factor(gender)2. female                         -0.07     0.08
## factor(race)2. black                            -1.52     0.17
## factor(race)3. asian                             0.40     0.31
## factor(race)4. native american                  -0.11     0.24
## factor(race)5. hispanic                         -0.30     0.18
## factor(educ1)2. high school (12 grades or fewer, incl  0.23     0.19
## factor(educ1)3. some college(13 grades or more,but no  0.35     0.19
## factor(educ1)4. college or advanced degree (no cases   0.45     0.19
## ---
##   n = 1179, k = 9
##   residual deviance = 1455.2, null deviance = 1591.2 (difference = 136.0)
```

It is clear that the coefficents of probit regression is close to logistic regression coefficients divided by 1.6.

Comparing logit and probit:

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

```
#logit
model_logit<-glm(vote_rep~factor(gender) + factor(race) + factor(educ1) + factor(ideo7),family = binomia
#probit
model_probit<-glm(vote_rep~factor(gender) + factor(race) + factor(educ1) + factor(ideo7),family = binom
display(model_logit)
```

```
## glm(formula = vote_rep ~ factor(gender) + factor(race) + factor(educ1) +
##     factor(ideo7), family = binomial(link = "logit"), data = nes5200_dt_s)
##                                                 coef.est coef.se
## (Intercept)                                     -17.20   386.48
## factor(gender)2. female                           0.11     0.15
## factor(race)2. black                             -2.82     0.39
## factor(race)3. asian                              0.51     0.64
## factor(race)4. native american                    0.00     0.46
## factor(race)5. hispanic                          -0.72     0.35
## factor(educ1)2. high school (12 grades or fewer, incl   0.60     0.37
## factor(educ1)3. some college(13 grades or more,but no   1.10     0.38
## factor(educ1)4. college or advanced degree (no cases    1.49     0.38
## factor(ideo7)2. liberal                          13.98   386.48
## factor(ideo7)3. slightly liberal                 14.19   386.48
## factor(ideo7)4. moderate, middle of the road     15.89   386.48
## factor(ideo7)5. slightly conservative            16.79   386.48
## factor(ideo7)6. conservative                     17.32   386.48
## factor(ideo7)7. extremely conservative           17.50   386.48
## ---
```

```
##   n = 1134, k = 15
##   residual deviance = 1100.9, null deviance = 1535.2 (difference = 434.3)
```

```
display(model_probit)
```

```
## glm(formula = vote_rep ~ factor(gender) + factor(race) + factor(educ1) +
##     factor(ideo7), family = binomial(link = "probit"), data = nes5200_dt_s)
##                                                 coef.est coef.se
## (Intercept)                                     -6.08    95.67
## factor(gender)2. female                          0.05     0.09
## factor(race)2. black                            -1.55     0.19
## factor(race)3. asian                             0.29     0.36
## factor(race)4. native american                  -0.03     0.27
## factor(race)5. hispanic                         -0.40     0.21
## factor(educ1)2. high school (12 grades or fewer, incl  0.38     0.22
## factor(educ1)3. some college(13 grades or more,but no  0.66     0.23
## factor(educ1)4. college or advanced degree (no cases   0.87     0.23
## factor(ideo7)2. liberal                          4.25    95.67
## factor(ideo7)3. slightly liberal                 4.30    95.67
## factor(ideo7)4. moderate, middle of the road     5.28    95.67
## factor(ideo7)5. slightly conservative            5.84    95.67
## factor(ideo7)6. conservative                     6.16    95.67
## factor(ideo7)7. extremely conservative           6.26    95.67
## ---
##   n = 1134, k = 15
##   residual deviance = 1107.2, null deviance = 1535.2 (difference = 428.0)
```

Tobit model for mixed discrete/continuous data:

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

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

Robust linear regression using the t model:

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

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

```
data<-na.omit(congress)
data<-data[data$year==1988 & data$contested==TRUE,]
model_linear<-lm(Dem_pct~x1+x2+factor(incumbent)+Dem_vote+Rep_vote, data=data)
summary(model_linear)
```

```
##
## Call:
## lm(formula = Dem_pct ~ x1 + x2 + factor(incumbent) + Dem_vote +
##     Rep_vote, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.118337 -0.017364 -0.005168  0.011862  0.157290
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.499e-01  1.194e-02  46.045  < 2e-16 ***
## x1                  1.827e-04  8.178e-05   2.234 0.026134 *
## x2                 -2.194e-04  1.199e-04  -1.830 0.068125 .
## factor(incumbent)0  2.372e-02  7.860e-03   3.018 0.002737 **
## factor(incumbent)1  2.821e-02  7.616e-03   3.704 0.000247 ***
## Dem_vote            1.928e-06  7.136e-08  27.022  < 2e-16 ***
## Rep_vote           -2.512e-06  6.612e-08 -37.995  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03262 on 341 degrees of freedom
## Multiple R-squared:  0.9714, Adjusted R-squared:  0.9709
## F-statistic:  1929 on 6 and 341 DF,  p-value: < 2.2e-16
```

R squre is 0.97, which is close to 1, so that the model fits data kind of 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.

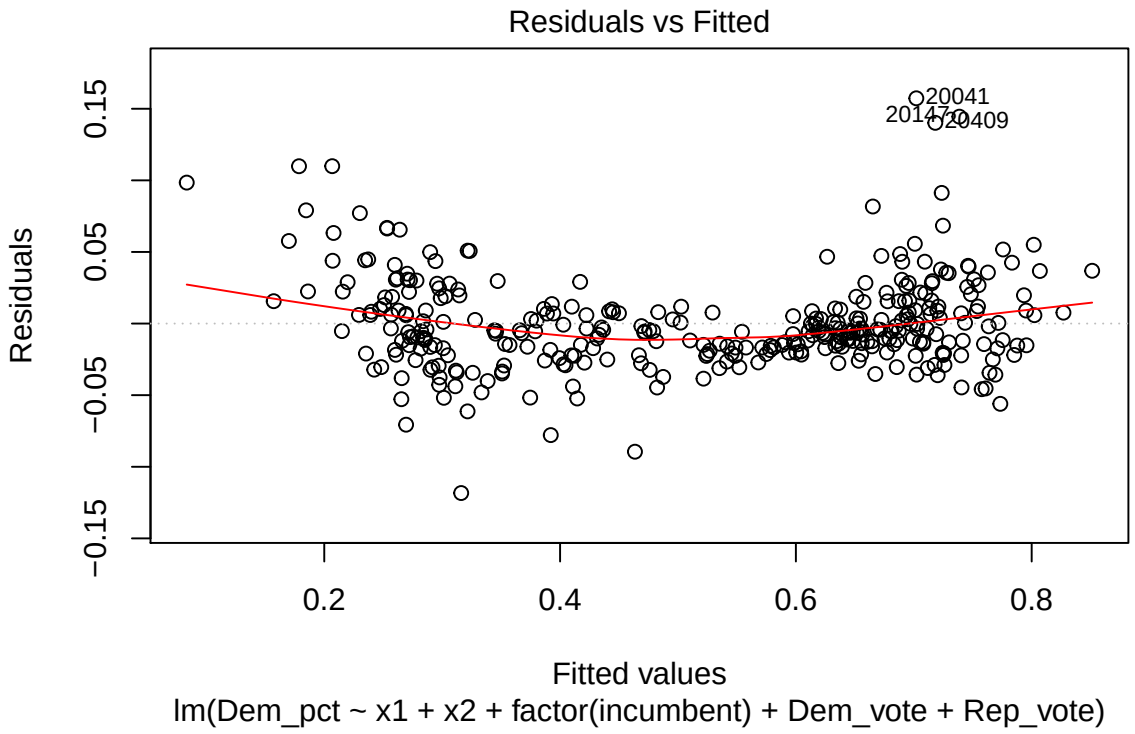
```
library(hett)
model_t<-tlm(Dem_pct~x1+x2+factor(incumbent)+Dem_vote+Rep_vote, data=data)
summary(model_t)
```

```
## Location model :
##
## Call:
## tlm(lform = Dem_pct ~ x1 + x2 + factor(incumbent) + Dem_vote +
##     Rep_vote, data = data)
##
## Residuals:
##        Min          1Q      Median          3Q         Max
## -0.1174482  -0.0131992  -0.0007138   0.0152583   0.1569731
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.630e-01  9.051e-03  62.204  < 2e-16 ***
```

```
## x1                   1.750e-04  6.198e-05   2.824 0.005029 **
## x2                  -1.340e-04  9.087e-05  -1.475 0.141214
## factor(incumbent)0   1.327e-02  5.956e-03   2.229 0.026494 *
## factor(incumbent)1   2.017e-02  5.772e-03   3.495 0.000536 ***
## Dem_vote             1.910e-06  5.408e-08  35.312  < 2e-16 ***
## Rep_vote            -2.622e-06  5.011e-08 -52.324  < 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 + factor(incumbent) + Dem_vote +
##     Rep_vote, data = data)
##
## Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0000  -1.7141  -0.9286   1.4079   5.6218
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -7.8056     0.1072  -72.81   <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 : 22 in 0.015
## Heteroscedastic t Likelihood : 735.9534
```

3. Which model do you prefer?

```
#residul plot of model_linear
plot(model_linear,which=1)
```



Because the residuals of linear regression is small and R-square is 0.97, so it is better to fit data by linear regression.

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.

```
#logistic
index<-1*(data$Dem_pct>=0.6)
newdata<-cbind(index,data)
model<-glm(index~x1+x2+factor(incumbent),data=newdata,family = binomial(link = logit))
display(model)
```

```
## glm(formula = index ~ x1 + x2 + factor(incumbent), family = binomial(link = logit),
##     data = newdata)
##                     coef.est coef.se
## (Intercept)        -5.19     1.08
## x1                  0.00     0.01
## x2                  0.01     0.01
## factor(incumbent)0  3.54     1.13
## factor(incumbent)1  6.63     1.03
## ---
##   n = 348, k = 5
##   residual deviance = 195.9, null deviance = 480.2 (difference = 284.3)
```

2. Fit a robit regression and assess model fit.

3. Which model do you prefer?

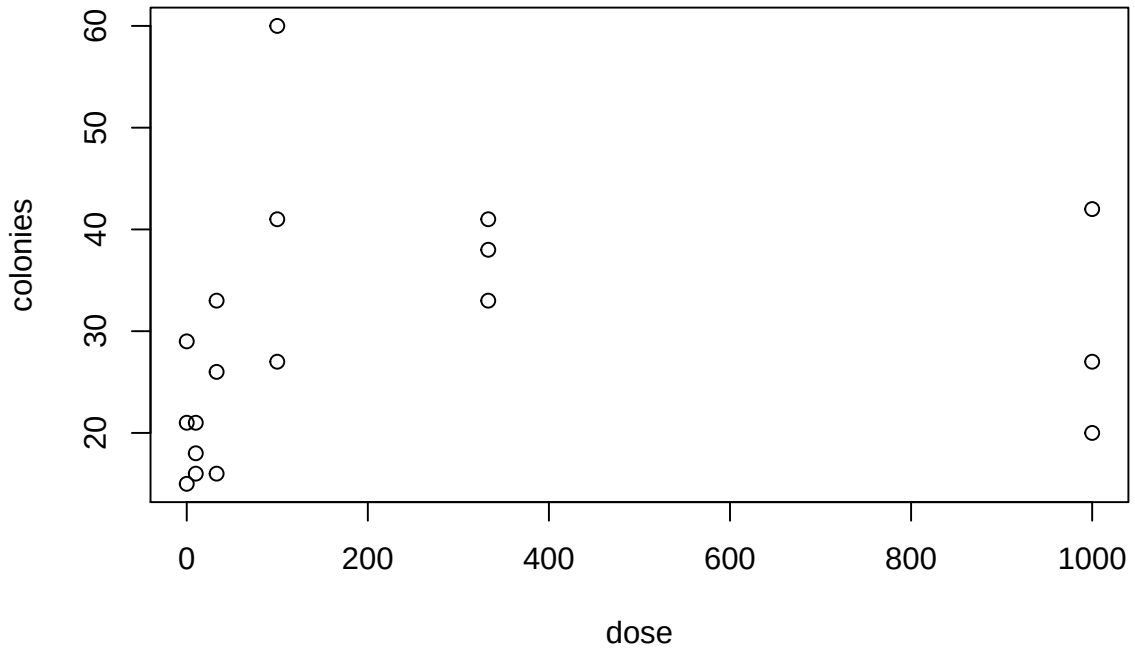
Salmonellla

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
```

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.

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

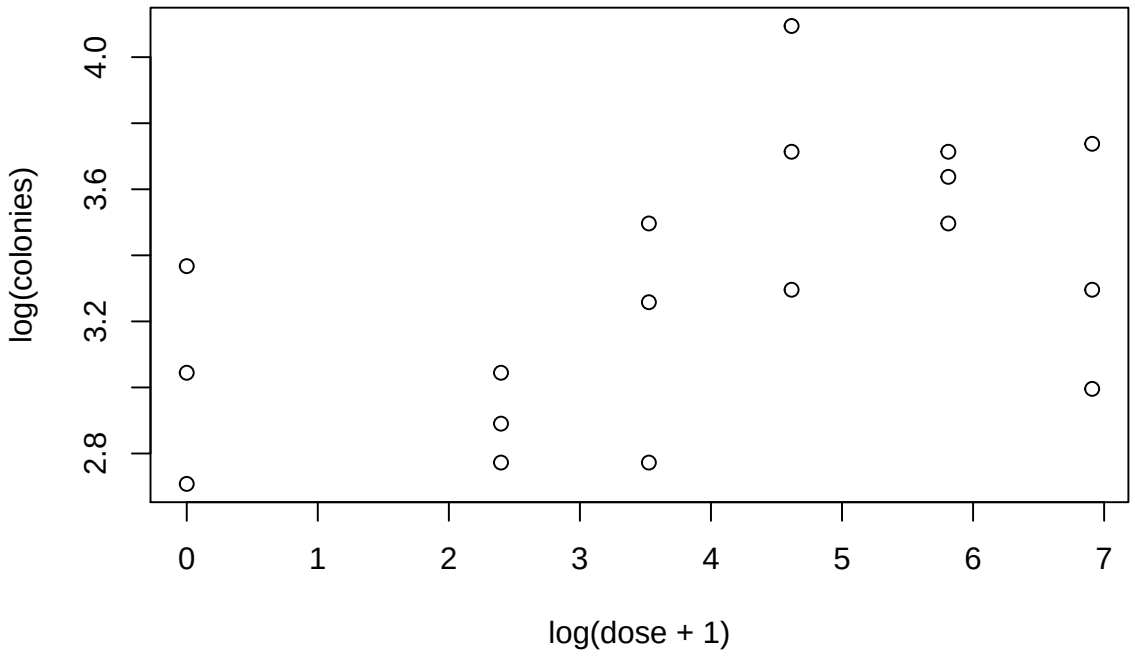


Since we are fitting log linear model we should look at the data on log scale. Also becase the dose is not equally spaced on the raw scale it may be better to plot it on the log scale as well.

```
model<-glm(colonies~dose,data=salmonella,family = poisson)
display(model)
```

```
## glm(formula = colonies ~ dose, family = poisson, data = salmonella)
##             coef.est coef.se
## (Intercept) 3.32     0.05
## dose        0.00     0.00
## ---
##   n = 18, k = 2
##   residual deviance = 75.8, null deviance = 78.4 (difference = 2.6)
```

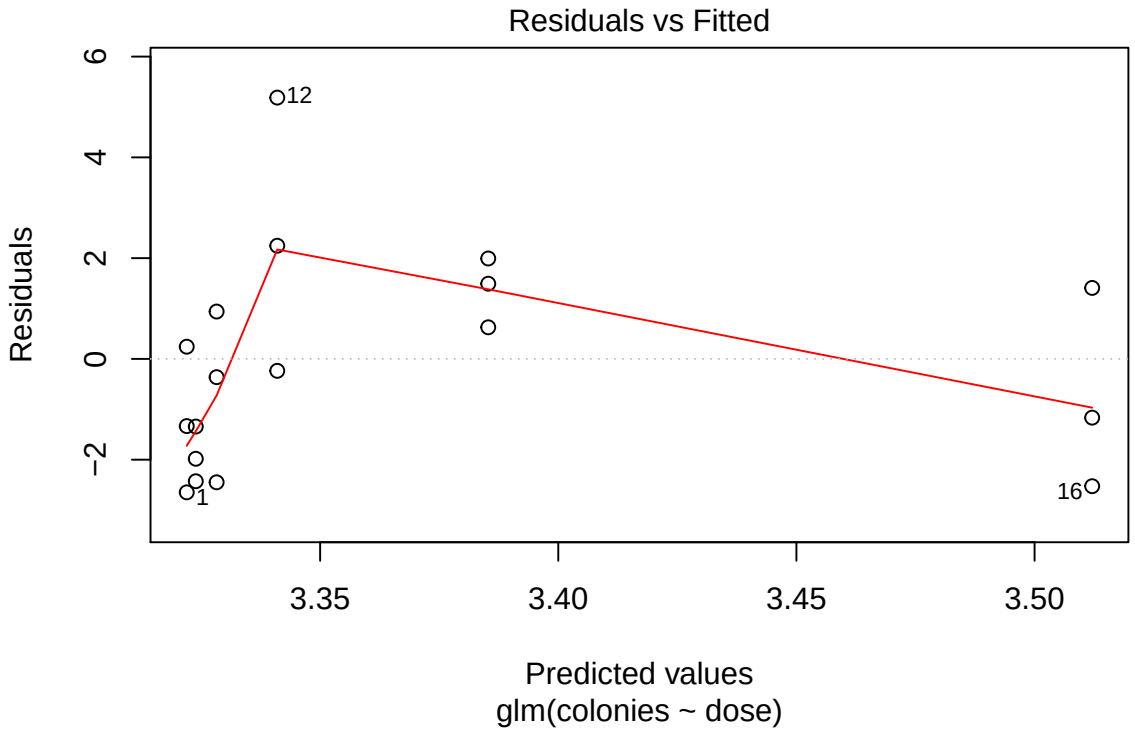
```
plot(log(colonies)~log(dose+1),data=salmonella)
```



Plot on the log scale decrease the right-skewness of the plot

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

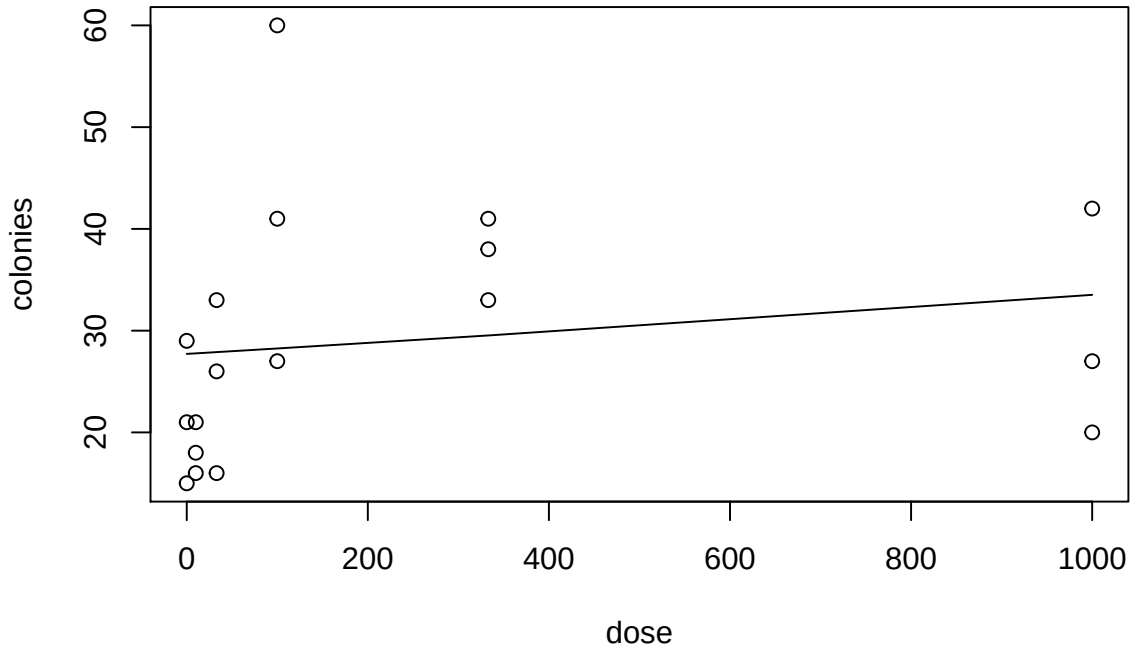
```
plot(model,which=1)
```



The residuals improve first and decrease as the dose increase.

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

```r
plot(colonies~dose,data=salmonella)
lines(salmonella$dose,predict.glm(model,type="response"))
```



It seems that the residulas of the plot are big.

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.

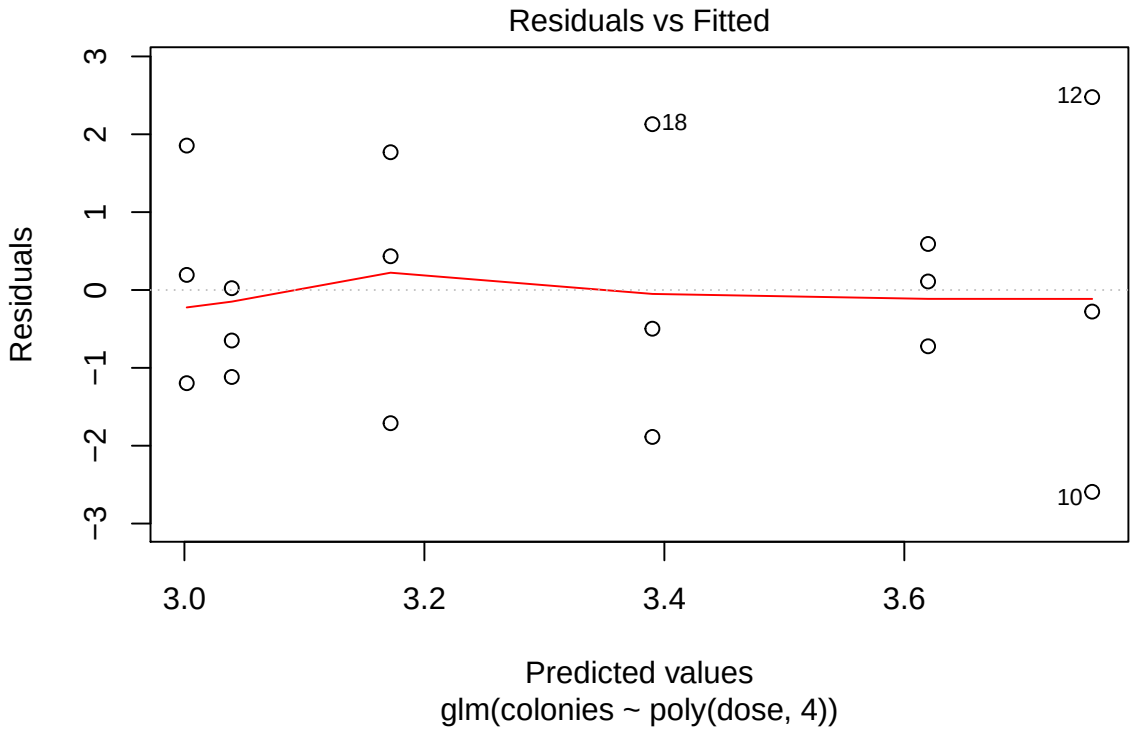
```r
model_new<-glm(colonies~poly(dose,4),data=salmonella,family = poisson)
summary(model_new)
```

```
##
## Call:
## glm(formula = colonies ~ poly(dose, 4), family = poisson, data = salmonella)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5928  -1.0187  -0.1270   0.5518   2.4771
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     3.32993    0.04547  73.226  < 2e-16 ***
## poly(dose, 4)1  0.38005    0.19014   1.999   0.0456 *
## poly(dose, 4)2 -0.85324    0.17657  -4.832 1.35e-06 ***
## poly(dose, 4)3  0.73745    0.17273   4.269 1.96e-05 ***
## poly(dose, 4)4  0.20857    0.20332   1.026   0.3050
## ---
## 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: 34.989  on 13  degrees of freedom
## AIC: 137.53
```

```
##
## 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.

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



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

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

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

```
##
##     Null deviance: 78.358  on 17  degrees of freedom
## Residual deviance: 34.989  on 13  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

Ships

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

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

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

```
ship_new<-ships[ships$service>0,]
model<-glm(incidents~type+factor(year)+factor(period),offset = log(service),data=ship_new,family=quasip
display(model)
```

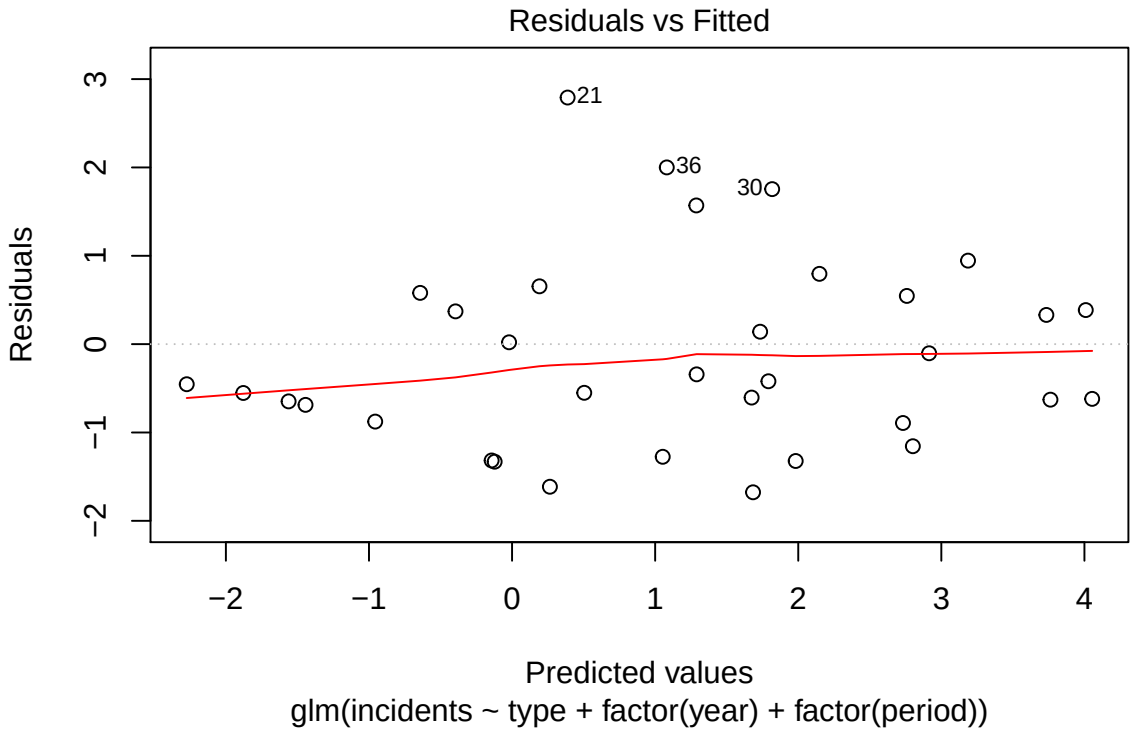
```
## glm(formula = incidents ~ type + factor(year) + factor(period),
##     family = quasipoisson(link = "log"), data = ship_new, offset = log(service))
##                   coef.est coef.se
## (Intercept)      -6.41     0.28
## typeB            -0.54     0.23
## typeC            -0.69     0.43
## typeD            -0.08     0.38
## typeE             0.33     0.31
## factor(year)65    0.70     0.19
## factor(year)70    0.82     0.22
## factor(year)75    0.45     0.30
## factor(period)75  0.38     0.15
## ---
##   n = 34, k = 9
##   residual deviance = 38.7, null deviance = 146.3 (difference = 107.6)
##   overdispersion parameter = 1.7
```

```
anova(model,test="F")
```

```
## Analysis of Deviance Table
##
## Model: quasipoisson, link: log
##
## Response: incidents
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev      F    Pr(>F)
## NULL                           33    146.328
## type            4   55.439        29     90.889 8.1961 0.0002289 ***
## factor(year)    3   41.534        26     49.355 8.1871 0.0005777 ***
## factor(period)  1   10.660        25     38.695 6.3039 0.0188808 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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




```
n<-dim(ship_new)[1]
k<-3
y_hat<-predict(model,type="response")
y<-(ship_new$incidents)
z<-(y-y_hat)/sqrt(y_hat)
cat("overdispersion ratio is ", sum(z^2)/(n-k), "\n")
```

```
## overdispersion ratio is  1.363718
```

Through the difference of residual deviance between null model and my model, it seems that the model is much better than the null model.

Through the anova, it seems that the coefficiants are all significant.

Through the residual plot, we can tell that the residuals scattered evenly.

Overdispersion ratio is 1.363718, is close to 1, so that the model do not have overdispersion problem.

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
```

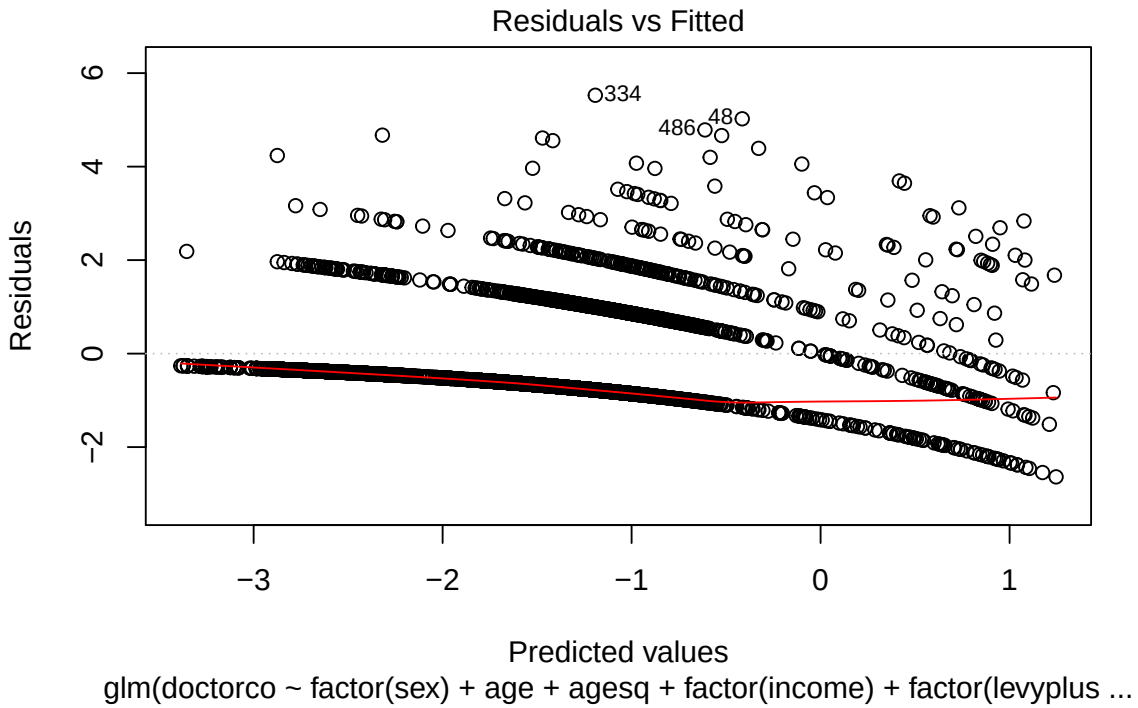
1. Build a Poisson regression model with `doctorco` as the response and `sex`, `age`, `agesq`, `income`, `levyplus`, `freepoor`, `freerepa`, `illness`, `actdays`, `hscore`, `chcond1` and `chcond2` as possible predictor variables. Considering the deviance of this model, does this model fit the data?

```
model1 = glm(doctorco ~ factor(sex) + age + agesq + factor(income) + factor(levyplus) + factor(freepoor
display(model1)
```

```
## glm(formula = doctorco ~ factor(sex) + age + agesq + factor(income) +
##     factor(levyplus) + factor(freepoor) + factor(freerepa) +
##     factor(illness) + actdays + hscore + factor(chcond1) + factor(chcond2),
##     family = poisson, data = dvisits)
##                     coef.est coef.se
## (Intercept)         -2.61     0.28
## factor(sex)1         0.15     0.06
## age                  0.89     1.02
## agesq               -0.45     1.10
## factor(income)0.01   0.40     0.31
## factor(income)0.06  -0.47     0.31
## factor(income)0.15  -0.02     0.22
## factor(income)0.25  -0.34     0.20
## factor(income)0.35  -0.32     0.20
## factor(income)0.45  -0.26     0.21
## factor(income)0.55  -0.44     0.21
## factor(income)0.65  -0.39     0.21
## factor(income)0.75  -0.37     0.21
## factor(income)0.9   -0.44     0.21
## factor(income)1.1   -0.55     0.23
## factor(income)1.3   -0.23     0.24
## factor(income)1.5   -0.38     0.23
## factor(levyplus)1    0.10     0.07
## factor(freepoor)1   -0.48     0.18
## factor(freerepa)1    0.08     0.10
## factor(illness)1     1.03     0.10
## factor(illness)2     1.30     0.11
## factor(illness)3     1.22     0.12
## factor(illness)4     1.36     0.13
## factor(illness)5     1.46     0.13
## actdays              0.12     0.01
## hscore               0.03     0.01
## factor(chcond1)1     0.02     0.07
## factor(chcond2)1     0.06     0.08
## ---
##   n = 5190, k = 29
##   residual deviance = 4252.5, null deviance = 5634.8 (difference = 1382.3)
```

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

```
plot(model1,which=1)
```



There are lines because doctorco is has several categories.

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

woman,old,low income,covered by private health insurance fund for private patient in public hospital,not covered bt government because low income,number of illness is high, Number of days of reduced activity in past two weeks due to illness or injury is high, bad health, have chronic condition.

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.

```
lambda<-predict(model1, dvisits[dim(dvisits)[1],], type="response")#lambda
print(paste0("Probability of 0 doctor's visits: ", dpois(0, lambda = lambda),3))
```

```
## [1] "Probability of 0 doctor's visits: 0.9175500813646233"
```

```
print(paste0("Probability of 1 doctor's visits: ", dpois(1, lambda = lambda),3))
```

```
## [1] "Probability of 1 doctor's visits: 0.07895345579073843"
```

```
print(paste0("Probability of 2 doctor's visits: ", dpois(2, lambda = lambda),3))
```

```
## [1] "Probability of 2 doctor's visits: 0.003396898059247683"
```

```
print(paste0("Probability of 3 doctor's visits: ", dpois(3, lambda = lambda),3))
```

```
## [1] "Probability of 3 doctor's visits: 9.74322260235948e-053"
```

```
print(paste0("Probability of 4 doctor's visits: ", dpois(4, lambda = lambda),3))
```

```
## [1] "Probability of 4 doctor's visits: 2.09596487052411e-063"
```

```
print(paste0("Probability of 5 doctor's visits: ", dpois(5, lambda = lambda),3))
```

```
## [1] "Probability of 5 doctor's visits: 3.60707656409883e-083"
```

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

```r
model2 = lm(doctorco ~ factor(sex) + age + agesq + factor(income) + factor(levyplus) + factor(freepoor)
summary(model2)
```

```
##
## Call:
## lm(formula = doctorco ~ factor(sex) + age + agesq + factor(income) +
##     factor(levyplus) + factor(freepoor) + factor(freerepa) +
##     factor(illness) + actdays + hscore + factor(chcond1) + factor(chcond2),
##     data = dvisits)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1502 -0.2741 -0.1502 -0.0178  7.0000
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.069988   0.110894   0.631   0.5280
## factor(sex)1         0.034410   0.021668   1.588   0.1123
## age                  0.099003   0.415248   0.238   0.8116
## agesq                0.062769   0.465195   0.135   0.8927
## factor(income)0.01   0.158340   0.145237   1.090   0.2757
## factor(income)0.06  -0.116557   0.113304  -1.029   0.3037
## factor(income)0.15   0.008991   0.092471   0.097   0.9225
## factor(income)0.25  -0.063853   0.085204  -0.749   0.4536
## factor(income)0.35  -0.062769   0.087624  -0.716   0.4738
## factor(income)0.45  -0.064514   0.088310  -0.731   0.4651
## factor(income)0.55  -0.110486   0.087294  -1.266   0.2057
## factor(income)0.65  -0.094907   0.087566  -1.084   0.2785
## factor(income)0.75  -0.083005   0.087940  -0.944   0.3453
## factor(income)0.9   -0.109831   0.086432  -1.271   0.2039
## factor(income)1.1   -0.126948   0.089753  -1.414   0.1573
## factor(income)1.3   -0.058359   0.099208  -0.588   0.5564
## factor(income)1.5   -0.092181   0.095739  -0.963   0.3357
## factor(levyplus)1    0.032808   0.025030   1.311   0.1900
## factor(freepoor)1   -0.113275   0.053250  -2.127   0.0334 *
## factor(freerepa)1    0.028622   0.039920   0.717   0.4734
## factor(illness)1     0.105760   0.026043   4.061 4.96e-05 ***
## factor(illness)2     0.181238   0.031145   5.819 6.27e-09 ***
## factor(illness)3     0.149058   0.038172   3.905 9.54e-05 ***
## factor(illness)4     0.230603   0.049928   4.619 3.96e-06 ***
## factor(illness)5     0.345984   0.054378   6.363 2.16e-10 ***
## actdays              0.102877   0.003671  28.023  < 2e-16 ***
## hscore               0.016639   0.005207   3.195   0.0014 **
## factor(chcond1)1    -0.001022   0.023938  -0.043   0.9660
## factor(chcond2)1     0.036995   0.035956   1.029   0.3036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7137 on 5161 degrees of freedom
## Multiple R-squared:  0.2048, Adjusted R-squared:  0.2005
## F-statistic: 47.47 on 28 and 5161 DF,  p-value: < 2.2e-16
```

```r
predict(model2, dvisits[5190,])
```

```
##      5190
## 0.1385791
```

Through the R squre, lm is also not fit the data well.