Homework 04

Generalized Linear Models

Sky Liu

October 9, 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)
risky_m1 <- glm(fupacts ~ factor(women_alone) + factor(couples), data = risky_behaviors, family = poiss
display(risky_m1)
```

```
## 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)
```

```
n = 434; k = 3
yhat <- predict (risky_m1, type="response")
z <- (risky_behaviors$fupacts-yhat)/sqrt(yhat)
cat ("overdispersion ratio is ", sum(z^2)/(n-k), "\n")
```

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

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

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

The difference of devience between this model and the null model is 373.1. The estimated overdispersion factor is 44, and the p-value is 1, indicating that the post treatment data are overdispersed by a factor of 44, which is huge and statistically significant.

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?

```
risky_m2 <- glm(fupacts ~ factor(women_alone) + factor(couples) + factor(bs_hiv) + factor(sex) + bupacts
display(risky_m2)
```

```
## glm(formula = fupacts ~ factor(women_alone) + factor(couples) +
##     factor(bs_hiv) + factor(sex) + bupacts, family = poisson,
##     data = risky_behaviors)
##                        coef.est coef.se
## (Intercept)             2.90     0.02
## factor(women_alone)1   -0.66     0.03
## factor(couples)1       -0.41     0.03
## factor(bs_hiv)positive -0.44     0.04
## factor(sex)man         -0.11     0.02
## bupacts                 0.01     0.00
## ---
##   n = 434, k = 6
##   residual deviance = 10200.4, null deviance = 13298.6 (difference = 3098.2)
```

```
n = 434; k = 6
yhat2 <- predict (risky_m2, type="response")
z2 <- (risky_behaviors$fupacts-yhat2)/sqrt(yhat2)
cat ("overdispersion ratio is ", sum(z2^2)/(n-k), "\n")
```

```
## overdispersion ratio is  30.00404
```

```
cat ("p-value of overdispersion test is ", pchisq (sum(z2^2), n-k), "\n")
```

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

The difference of devience between this model and the last model is 2725.1. It has improved a lot. The estimated overdispersion factor is 30, and the p-value is 1, indicating that the post treatment data are overdispersed by a factor of 44, which is less huge than the last model but still huge and statistically significant.

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

```
risky_m3 <- glm(fupacts ~ factor(women_alone) + factor(couples) + factor(bs_hiv) + factor(sex) + bupacts
summary(risky_m3)
```

```
##
## Call:
## glm(formula = fupacts ~ factor(women_alone) + factor(couples) +
##     factor(bs_hiv) + factor(sex) + bupacts, family = quasipoisson,
##     data = risky_behaviors)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -18.679   -4.305   -2.511    1.368   23.361
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.8957952  0.1271206  22.780  < 2e-16 ***
## factor(women_alone)1   -0.6622159  0.1692369  -3.913 0.000106 ***
## factor(couples)1       -0.4099761  0.1546315  -2.651 0.008316 **
## factor(bs_hiv)positive -0.4383170  0.1937994  -2.262 0.024217 *
## factor(sex)man         -0.1086694  0.1299838  -0.836 0.403609
## bupacts                 0.0107789  0.0009521  11.321  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 30.00407)
##
```

```
##     Null deviance: 13299  on 433  degrees of freedom
## Residual deviance: 10200  on 428  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

```
n = 434; k = 6
yhat2 <- predict (risky_m3, type="response")
z2 <- (risky_behaviors$fupacts-yhat2)/sqrt(yhat2)
cat ("overdispersion ratio is ", sum(z2^2)/(n-k), "\n")
```

```
## overdispersion ratio is  30.00404
```

```
cat ("p-value of overdispersion test is ", pchisq (sum(z2^2), n-k), "\n")
```

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

From the summary we could see that the couple and women_alone coeffient are statistically significant. The result decreases for 1 - exp(-0.66), which is 48% if women come alone, and the result decreases for 1 - exp(-0.4099761), which is 34%, if couple participated.

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? There might be some overlapping in the cases men and women both participating as couples. In this case, the variables are no longer independent, which fails our modeling assumption.

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)

```
#President election data
nes5200<-read.dta("http://www.stat.columbia.edu/~gelman/arm/examples/nes/nes5200_processed_voters_realic
#saveRDS(nes5200,"nes5200.rds")
#nes5200<-readRDS("nes5200.rds")

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)
Nes <-  nes5200_dt_s %>%
       select(vote_rep,age,income,gender,race,educ1,partyid7,ideo,rlikes)
Nes <- na.omit(Nes)#clean rows with NAs
Nes$age <- Nes$age - mean(Nes$age)#center the age

Nes_glm_lo <- glm(vote_rep ~ age + gender  + race  + partyid7 + ideo + rlikes , family=binomial(link =
Nes_glm_pro <- glm(vote_rep ~ age + gender  + race  + partyid7 + ideo + rlikes , family=binomial(link =

display(Nes_glm_lo)
```

```
## glm(formula = vote_rep ~ age + gender + race + partyid7 + ideo +
##     rlikes, family = binomial(link = "logit"), data = Nes)
##                                        coef.est coef.se
## (Intercept)                            -3.52     0.48
## age                                     0.00     0.01
## gender2. female                          0.59     0.26
## race2. black                            -2.18     0.58
## race3. asian                             0.29     0.97
## race4. native american                   0.67     0.80
## race5. hispanic                          0.83     0.52
## partyid72. weak democrat                 1.01     0.47
## partyid73. independent-democrat          0.66     0.55
## partyid74. independent-independent       2.49     0.52
## partyid75. independent-republican        4.27     0.53
## partyid76. weak republican               3.63     0.49
## partyid77. strong republican             5.06     0.65
## ideo3. moderate ('middle of the road')  0.36     0.47
## ideo5. conservative                      1.50     0.29
## rlikes                                   0.75     0.07
## ---
##   n = 1132, k = 16
##   residual deviance = 460.4, null deviance = 1533.0 (difference = 1072.7)
```

```
display(Nes_glm_pro)
```

```
## glm(formula = vote_rep ~ age + gender + race + partyid7 + ideo +
##     rlikes, family = binomial(link = "probit"), data = Nes)
##                                         coef.est coef.se
## (Intercept)                             -1.87     0.24
## age                                      0.00     0.00
## gender2. female                          0.30     0.14
## race2. black                            -1.09     0.30
## race3. asian                             0.04     0.52
## race4. native american                   0.33     0.43
## race5. hispanic                          0.43     0.28
## partyid72. weak democrat                 0.49     0.24
## partyid73. independent-democrat          0.30     0.28
## partyid74. independent-independent       1.34     0.27
## partyid75. independent-republican        2.32     0.27
## partyid76. weak republican               1.98     0.25
## partyid77. strong republican             2.76     0.32
## ideo3. moderate ('middle of the road')  0.23     0.26
## ideo5. conservative                      0.78     0.16
## rlikes                                   0.41     0.04
## ---
##   n = 1132, k = 16
##   residual deviance = 459.7, null deviance = 1533.0 (difference = 1073.3)
```

The model results are about the same and the coefficients in a probit regression are about the logistic regression coefficients divided by 1.6.

Comparing logit and probit:

construct a dataset where 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)

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.

```
con_m1 <- lm(Dem_pct ~ x1 + x2 + incumbent + contested + Rep_vote, data = con_data)
summary(con_m1)
```

```
##
## Call:
## lm(formula = Dem_pct ~ x1 + x2 + incumbent + contested + Rep_vote,
##     data = con_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.291108 -0.041568  0.005002  0.045064  0.214917
##
## Coefficients: (1 not defined because of singularities)
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.890e-01  9.809e-03  70.246   <2e-16 ***
## x1             2.547e-04  1.252e-04   2.034   0.0423 *
## x2            -1.039e-04  1.894e-04  -0.549   0.5834
## incumbent      1.012e-01  4.212e-03  24.020   <2e-16 ***
## contestedTRUE         NA         NA      NA       NA
## Rep_vote      -2.156e-06  9.207e-08 -23.412   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07139 on 698 degrees of freedom
## Multiple R-squared:  0.8611, Adjusted R-squared:  0.8603
## F-statistic:  1082 on 4 and 698 DF,  p-value: < 2.2e-16
```

The R^2 is 86%, which means 86% of variations are explained by this model.

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.

```
#con_m2 <- vglm(Dem_pct ~ x1 + x2 + incumbent + contested + Rep_vote, data = con_data)
#summary(con_m2)
```

3. Which model do you prefer?

Robust regression for binary data using the robit model:

Use the same data as the previous example with the goal instead of predicting for each district whether it was won by the Democratic or Republican candidate.

1. Fit a standard logistic or probit regression and assess model fit.

```
con_data$Fac_Dem_pct <- ifelse(con_data$Dem_pct >= 0.5, 1, 0)

con_m3 <- glm(Fac_Dem_pct ~ x1 + x2 + incumbent + contested + Rep_vote, data = con_data, family = binom
summary(con_m3)
```

```
##
## Call:
## glm(formula = Fac_Dem_pct ~ x1 + x2 + incumbent + contested +
##     Rep_vote, family = binomial(link = "logit"), data = con_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.79229  -0.13139   0.06005   0.15574   2.77101
##
## Coefficients: (1 not defined because of singularities)
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    5.706e+00  9.467e-01   6.028 1.66e-09 ***
## x1             2.345e-03  9.817e-03   0.239    0.811
## x2            -1.088e-03  1.195e-02  -0.091    0.927
## incumbent      2.495e+00  2.426e-01  10.285  < 2e-16 ***
## contestedTRUE         NA         NA      NA       NA
## Rep_vote      -6.737e-05  1.005e-05  -6.705 2.02e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 966.55  on 702  degrees of freedom
## Residual deviance: 183.78  on 698  degrees of freedom
## AIC: 193.78
##
## Number of Fisher Scoring iterations: 7
```

```
con_m4 <- glm(Fac_Dem_pct ~ x1 + x2 + incumbent + contested + Rep_vote, data = con_data, family = binom
summary(con_m4)
```

```
##
## Call:
```

```
## glm(formula = Fac_Dem_pct ~ x1 + x2 + incumbent + contested +
##     Rep_vote, family = binomial(link = "probit"), data = con_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.79981  -0.10316   0.02676   0.12816   2.72388
##
## Coefficients: (1 not defined because of singularities)
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.020e+00  4.758e-01   6.347 2.19e-10 ***
## x1             2.999e-04  4.970e-03   0.060    0.952
## x2            -1.799e-03  6.244e-03  -0.288    0.773
## incumbent      1.388e+00  1.208e-01  11.488  < 2e-16 ***
## contestedTRUE         NA         NA      NA       NA
## Rep_vote      -3.489e-05  5.005e-06  -6.971 3.15e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 966.55  on 702  degrees of freedom
## Residual deviance: 181.23  on 698  degrees of freedom
## AIC: 191.23
##
## Number of Fisher Scoring iterations: 8
```

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.

```r
data(salmonella)
#?salmonella
data(salmonella)

sal_m1 <- glm(colonies ~ dose, data = salmonella, family = poisson)
display(sal_m1)
```

```
## 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)
```

```r
n = 18; k = 2
yhat3 <- predict (sal_m1, type="response")
```

```
z3 <- (salmonella$colonies-yhat3)/sqrt(yhat3)
cat ("overdispersion ratio is ", sum(z3^2)/(n-k), "\n")
```

```
## overdispersion ratio is  5.087258
```

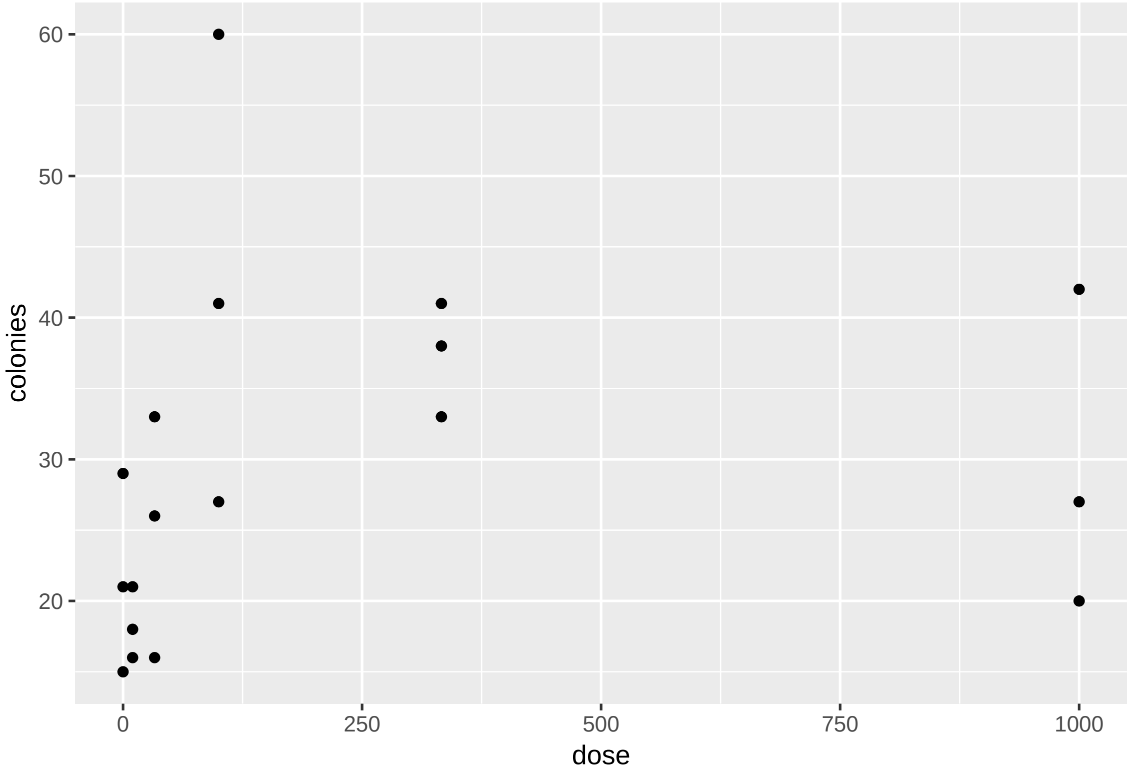
```
cat ("p-value of overdispersion test is ", pchisq (sum(z3^2), n-k), "\n")
```

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

The difference of devience between this model and the null model is 2.6. Not much of improvement. The estimated overdispersion factor is 5, and the p-value is 1, indicating that the post treatment data are overdispersed by a factor of 5, which is huge and statistically significant.

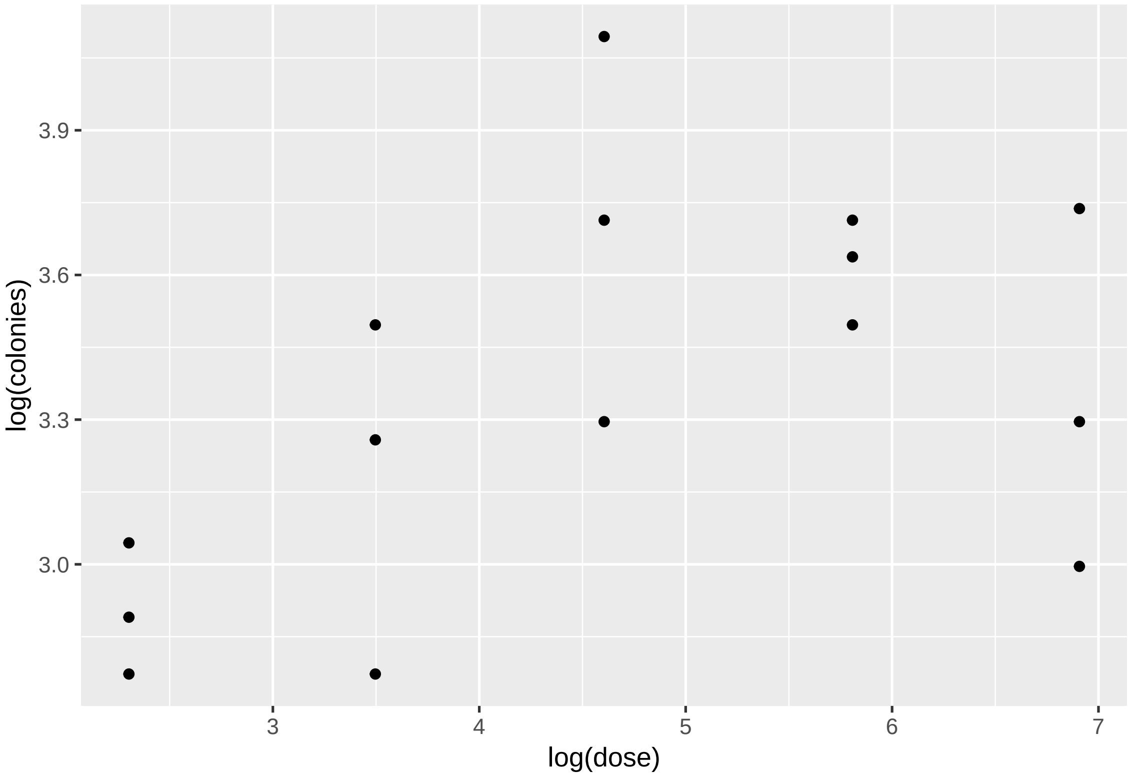
When you plot the data you see that the number of colonies as a function of dose is not monotonic especially around the dose of 1000.

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



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

```
lg_salmonella <- salmonella[salmonella$dose != 0,]
ggplot(lg_salmonella) + geom_point(aes(x = log(dose), y = log(colonies)))
```



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

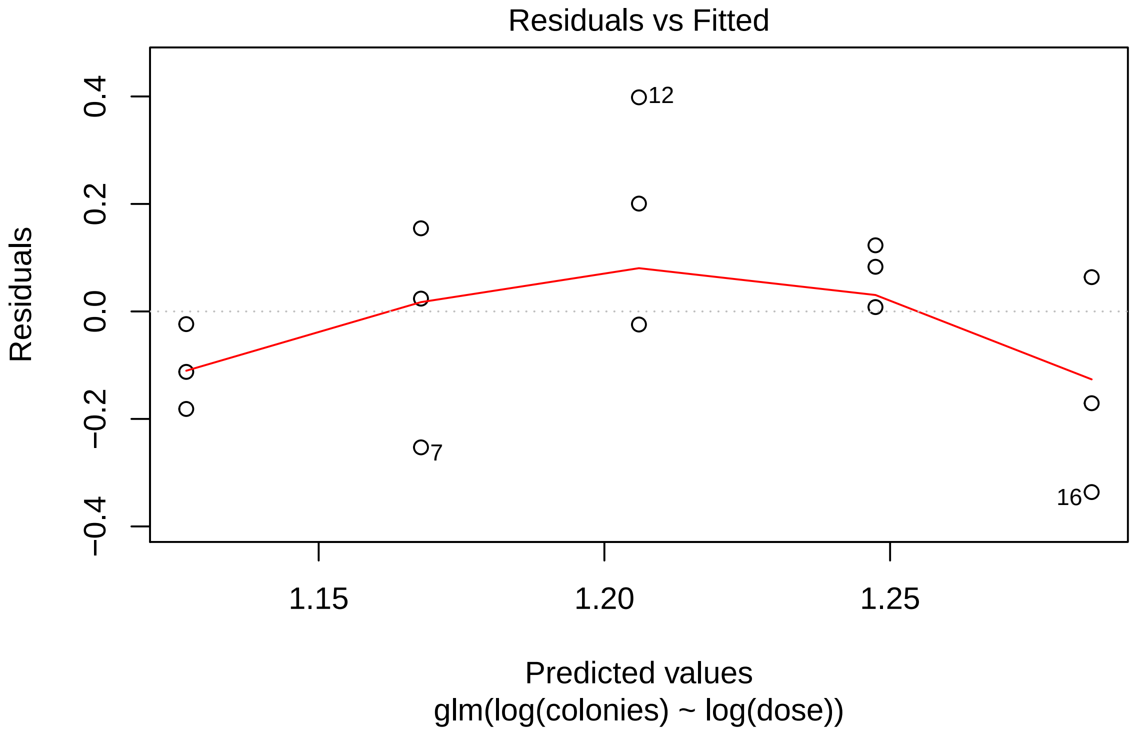
```
sal_m2 <- glm(log(colonies) ~ log(dose), data = lg_salmonella, family = poisson)
```

```
## Warning in dpois(y, mu, log = TRUE): non-integer x = 2.772589

## Warning in dpois(y, mu, log = TRUE): non-integer x = 2.890372

## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.044522

## Warning in dpois(y, mu, log = TRUE): non-integer x = 2.772589

## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.258097

## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.496508

## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.295837

## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.713572

## Warning in dpois(y, mu, log = TRUE): non-integer x = 4.094345

## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.496508

## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.637586

## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.713572

## Warning in dpois(y, mu, log = TRUE): non-integer x = 2.995732

## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.295837

## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.737670
```

```
display(sal_m2)
```

```
## glm(formula = log(colonies) ~ log(dose), family = poisson, data = lg_salmonella)
##             coef.est coef.se
## (Intercept) 1.05     0.43   
## log(dose)   0.03     0.09   
## ---
##   n = 15, k = 2
##   residual deviance = 0.5, null deviance = 0.7 (difference = 0.2)
```

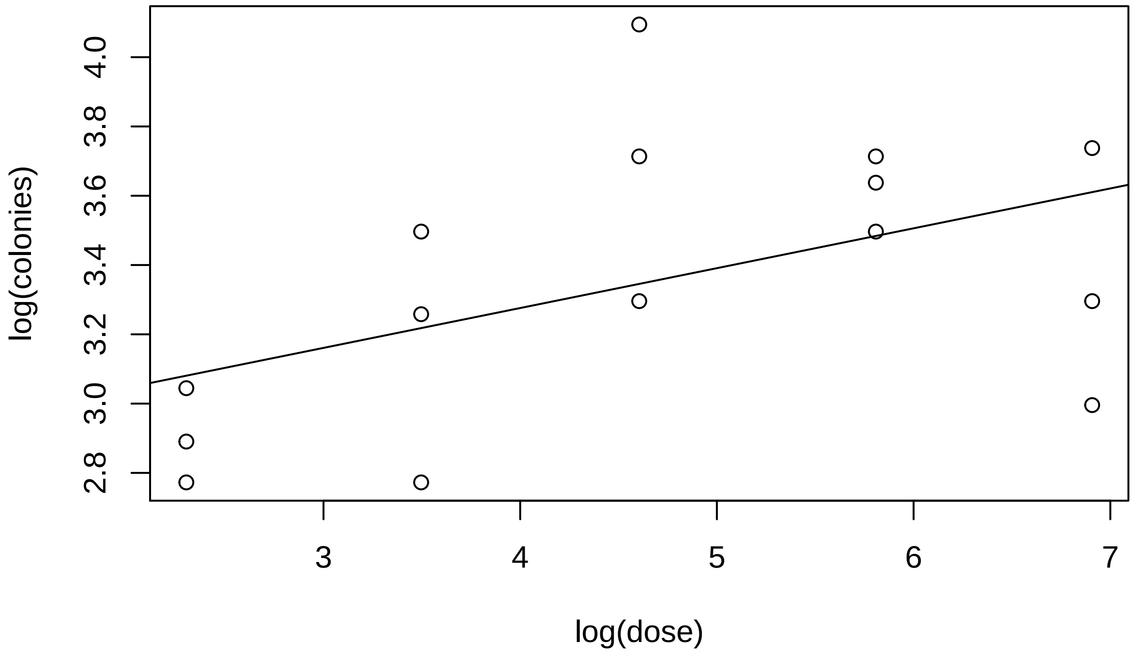
```
plot(sal_m2,which = 1)
```



The residuals are not evenly spreaded.

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

```
plot(x = log(lg_salmonella$dose), y = log(lg_salmonella$colonies), xlab = "log(dose)", ylab = "log(colon
```



The fitted line does not go accorss any points and the points are spreaded evenly.

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.

```
sal_m3 <- glm(colonies ~ dose, data = salmonella, family = quasipoisson)
display(sal_m3)
```

```
## glm(formula = colonies ~ dose, family = quasipoisson, data = salmonella)
##             coef.est coef.se
## (Intercept) 3.32     0.12
## dose        0.00     0.00
## ---
##   n = 18, k = 2
##   residual deviance = 75.8, null deviance = 78.4 (difference = 2.6)
##   overdispersion parameter = 5.1
```

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.

```
options(scipen=999)
ship_m1 <- glm(incidents ~ ., data = ships, family = poisson)
summary(ship_m1)
```

```
## 
## Call:
## glm(formula = incidents ~ ., 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) -5.706133761  1.220978309  -4.673       0.000002962378 ***
## typeB        0.813461790  0.202292452   4.021       0.000057898306 ***
## typeC       -1.204580820  0.327451699  -3.679             0.000234 ***
## typeD       -0.859523598  0.287518995  -2.989             0.002795 ** 
## typeE       -0.222555723  0.234784531  -0.948             0.343173    
## year         0.045191919  0.013410607   3.370             0.000752 ***
## period       0.060545698  0.008945311   6.768       0.000000000013 ***
## service      0.000059702  0.000007016   8.509 < 0.0000000000000002 ***
## ---
## 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
```

Types B, C, and D, year, period are important predictors.

The average number of incidents is increased by 124% if the number of type B ships increases by one unit.

The average number of incidents is decreased by 70% if the number of type C ships increases by one unit.

The average number of incidents is decreased by 57% if the number of type D ships increases by one unit.

The average number of incidents is increased by 4% if the year increases by one unit.

The average number of incidents is increased by 6% if the period increases by one unit.

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.

```r
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?

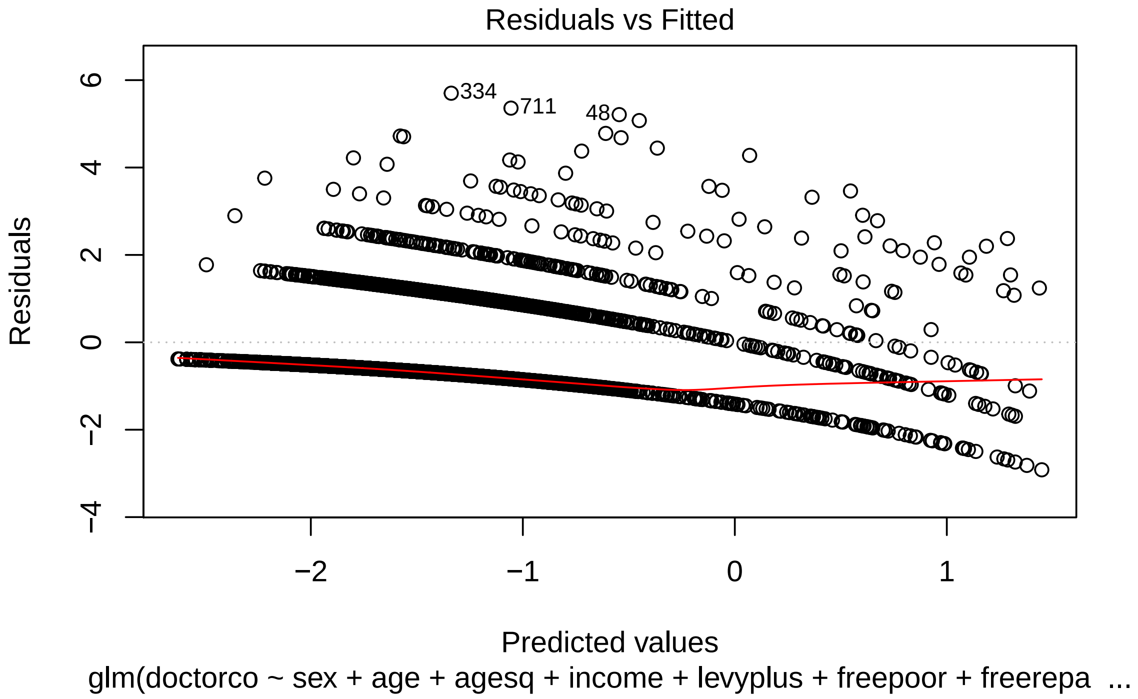
```r
dv_m1 <- glm(doctorco ~ sex + age + agesq + income + levyplus + freepoor + freerepa + illness + actdays
display(dv_m1)
```

```
## 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 = 1255.3)
```

The difference of deviance between this model and the null model is 1255.3, pretty big change. From this I think this model is a good fit.

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

```
plot(dv_m1, which = 1)
```



Becasue the number of doctor visits are discrete.

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

```
summary(dv_m1)
```

```
##
## 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 <0.0000000000000002 ***
## 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 <0.0000000000000002 ***
## actdays      0.126846   0.005034  25.198 <0.0000000000000002 ***
## 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
```

People with 5 or more illnesses in past 2 weeks and people with high number of days of reduced activities in paer two weeks due to the illness or injury will be the most likely to come to visit the doctor.

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.

```r
options(scipen=99)
mean_lp <- predict(dv_m1, dvisits[5190,], type = "response")
n <- c()
p <- c()
for (i in 0:10){
  n[i+1] <- i
  p[i+1] <- dpois(i, lambda = mean_lp)
}
pv <- cbind(n,p)
pv <- as.data.frame(pv)
kable(pv)
```

n	p
0	0.8578005
1	0.1315726

n	p
2	0.0100905
3	0.0005159
4	0.0000198
5	0.0000006
6	0.0000000
7	0.0000000
8	0.0000000
9	0.0000000
10	0.0000000

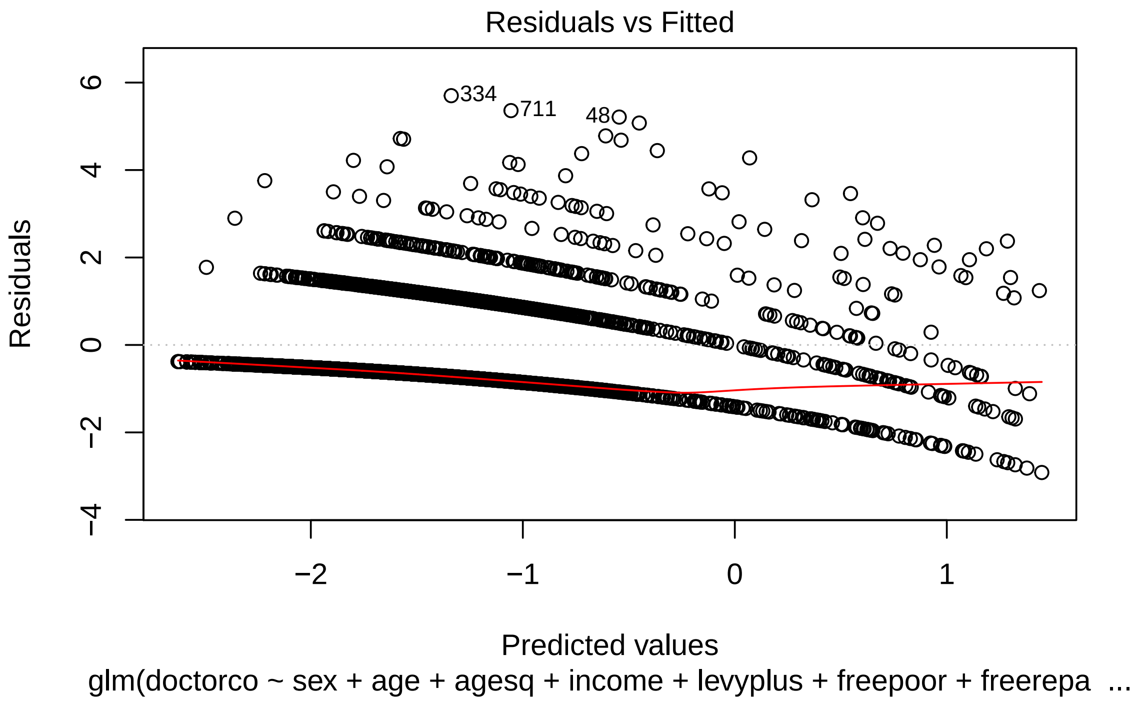
5. Fit a comparable (Gaussian) linear model and graphically compare the fits. Describe how they differ.

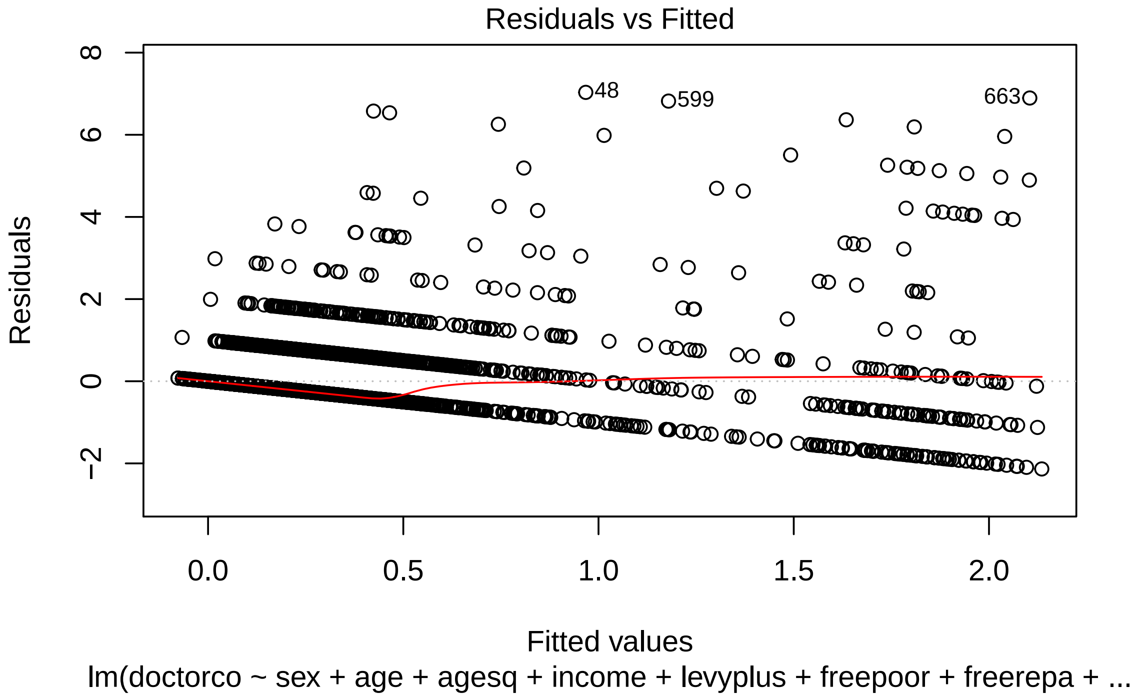
```
dv_m2 <- lm(doctorco ~ sex + age + agesq + income + levyplus + freepoor + freerepa + illness + actdays +
summary(dv_m2)
```

```
##
## Call:
## lm(formula = doctorco ~ sex + age + agesq + income + levyplus +
##     freepoor + freerepa + illness + actdays + hscore + chcond1 +
##     chcond2, data = dvisits)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1352 -0.2588 -0.1435 -0.0433  7.0327
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  0.027632   0.072220   0.383              0.70202
## sex          0.033811   0.021604   1.565              0.11764
## age          0.203201   0.410016   0.496              0.62020
## agesq       -0.062103   0.458716  -0.135              0.89231
## income      -0.057323   0.033089  -1.732              0.08326 .
## levyplus     0.035179   0.024882   1.414              0.15748
## freepoor    -0.103314   0.052471  -1.969              0.04901 *
## freerepa     0.033241   0.038157   0.871              0.38371
## illness      0.059946   0.008357   7.173    0.000000000000839 ***
## actdays      0.103192   0.003657  28.216 < 0.0000000000000002 ***
## hscore       0.016976   0.005190   3.271              0.00108 **
## chcond1      0.004384   0.023740   0.185              0.85349
## chcond2      0.041617   0.035863   1.160              0.24592
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7139 on 5177 degrees of freedom
## Multiple R-squared:  0.2018, Adjusted R-squared:    0.2
## F-statistic: 109.1 on 12 and 5177 DF,  p-value: < 0.00000000000000022
```

```
par(2,1);plot(dv_m1,which=1);plot(dv_m2,which=1)
```

```
## [[1]]
## NULL
##
## [[2]]
## NULL
```





```r
predict(dv_m2, dvisits[5190,], type = "response")
```

```
##      5190
## 0.1606531
```

The R^2 for the guassian linear model is only 20%, and the residuals are not evenly spreaded. Thus the Guassian linear model is not a good fit.