Homework 04

Generalized Linear Models

Xinyi Wang

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)
#print(is.factor(risky_behaviors$women_alone))
m1 = glm(fupacts ~ factor(women_alone) + factor(couples), family=poisson, data=risky_behaviors)
display(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 = nrow(risky_behaviors)
k = length(m1$coef)
y = risky_behaviors$fupacts
yhat = predict(m1, type = "response")
z =  (y - yhat) / sqrt(yhat)
# head(z)
# z2 =  (y - yhat) / sd(yhat)
# head(z2)
cat("The overdispersion ratio is ", sum(z^2)/(n-k), "\n")
```

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

```
cat("And the p-value of the overdispersion test is", 1 - pchisq( sum(z^2), n - k), "\n")
```

```
## And the p-value of the overdispersion test is 0
```

The model is better than null since deviance is smaller. The data are overdispersed by a factor of 44.15, which is huge and also 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?

```
sub = risky_behaviors[risky_behaviors$bupacts > 0,]
m2 = glm(fupacts ~ factor(women_alone) + factor(couples) + factor(bs_hiv) + factor(sex), data=sub, offs
display(m2)
```

```
## glm(formula = fupacts ~ factor(women_alone) + factor(couples) +
##     factor(bs_hiv) + factor(sex), family = poisson, data = sub,
##     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)
```

```
n = nrow(sub)
k = length(m2$coef)
y = sub$fupacts
yhat = predict(m2, type = "response")
z =  (y - yhat) / sqrt(yhat)
cat("The overdispersion ratio is ", sum(z^2)/(n-k), "\n")
```

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

```
cat("And the p-value of the overdispersion test is", 1 - pchisq( sum(z^2), n - k), "\n")
```

```
## And the p-value of the overdispersion test is 0
```

This model fits better than the first since deviance is smaller. However m2 seems still overdispersed.

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

```
m3 = glm(fupacts ~ factor(women_alone) + factor(couples) + factor(bs_hiv) + factor(sex), data=sub, offs
display(m3)
```

```
## glm(formula = fupacts ~ factor(women_alone) + factor(couples) +
##     factor(bs_hiv) + factor(sex), family = quasipoisson, data = sub,
##     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
```

The treament appears less significant. And the difference between the women alone group and the couples group looks much less 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, this is a problem because the observations coming from the two elements of the couple won't be i.i.d. We are expecting an extremely high positive correlations between the answers of people part of the same couple. 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)

```
wells = read.table("http://www.stat.columbia.edu/~gelman/arm/examples/arsenic/wells.dat")
wells$log.arsenic = log(wells$arsenic)
#summary(wells)
logit = glm(switch ~ log(arsenic) + dist + educ, family=binomial(link="logit"), data=wells)
display(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 = 239.9)
```

```
probit = glm(switch ~ log(arsenic) + dist + educ, family=binomial(link="probit"), data=wells)
display(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 = 239.8)
```

In probit model, the coefficient of log.arsenic becomes 0.54 which is close to 0.89 / 1.6 = 0.5563; the coefficient of the distance stays -0.01 which is still close to -0.01 / 1.6 = -0.0062; and the one of education becomes 0.03 which is close to 0.04 / 1.6 = 0.0250. These are essentially the coefficients we would have scaling by 1.6 the coefficients of the logit model.

Comparing logit and probit:

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

```
arsenic = runif(10,0.51,9.65)
dist = runif(10,0.387,339.53)
educ = sample(0:17,10,replace = T)
```

```
predict_data = data.frame(arsenic,dist,educ)
predict(logit,predict_data)
```

```
##          1          2          3          4          5          6
## -0.2793828 -0.9565041 -1.7519715  0.7792656 -0.6474906 -0.8087058
##          7          8          9         10
##  2.0171333  1.0982073  1.4127797  1.3028728
```

```
predict(probit,predict_data)
```

```
##          1          2          3          4          5          6
## -0.1681844 -0.5814865 -1.0701667  0.4789101 -0.3935587 -0.4930034
##          7          8          9         10
##  1.2373514  0.6750315  0.8682661  0.7961523
```

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)

```
#summary(lalonde)
lalonde$re78 = (lalonde$re78 - mean(lalonde$re78)) / sd(lalonde$re78)
tobit = vglm(re78 ~ educ + factor(treat) + factor(black) + factor(married) + age, tobit(Upper = 121174)
```

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

```
summary(tobit)
```

```
##
## Call:
## vglm(formula = re78 ~ educ + factor(treat) + factor(black) +
##     factor(married) + age, family = tobit(Upper = 121174), data = lalonde)
##
##
## 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 average earning on 1978 would increase by 0.08 while holding all other variables in the model constant.

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

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

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

age: With every 1 increase in the age, one's average earning on 1978 would increase by 0.012 while holding all 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.

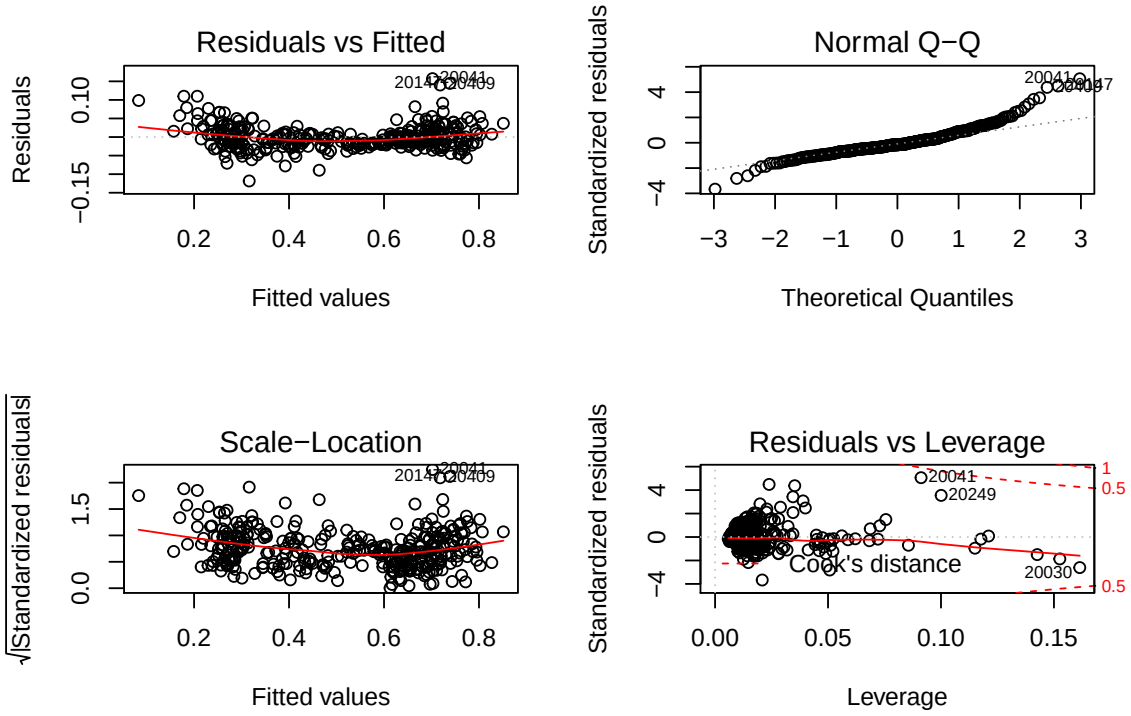
1. Fit a linear regression (with the usual normal-distribution model for the errors) predicting 1988 Democratic vote share from the other variables and assess model fit.

```
subcongress = congress[congress$year==1988 & congress$contested==TRUE,]
subcongress = na.omit(subcongress)
m1.congress = lm(Dem_pct ~ x1+x2+factor(incumbent)+Dem_vote+Rep_vote,data=subcongress)
summary(m1.congress)
```

```
##
## Call:
## lm(formula = Dem_pct ~ x1 + x2 + factor(incumbent) + Dem_vote +
##     Rep_vote, data = subcongress)
##
```

```
## 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
```

```
par(mfrow=c(2,2))
plot(m1.congress)
```



The model seems a good fit since all variables are significant and p-value is quite small.

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.

```
m2.congress = tlm(Dem_pct ~ x1+x2+factor(incumbent)+Dem_vote+Rep_vote,data=subcongress)
summary(m2.congress)
```

```
## Location model :
```

```
##
## Call:
## tlm(lform = Dem_pct ~ x1 + x2 + factor(incumbent) + Dem_vote +
##     Rep_vote, data = subcongress)
##
## 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 = subcongress)
##
## 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 0.01
## Heteroscedastic t Likelihood : 735.9534
```

3. Which model do you prefer?

I prefer model 1 since it has high r-squared value which is 0.97.

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.

```
m3.congress = glm( winparty.index ~ x1 + x2 + factor(incumbent), data=subcongress, family=binomial(link=
summary(m3.congress)
```

```
##
## Call:
## glm(formula = winparty.index ~ x1 + x2 + factor(incumbent), family = binomial(link = "logit"),
##     data = subcongress)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7342  -0.2004  -0.1767   0.2420   2.9309
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          3.109471   0.685926   4.533 5.81e-06 ***
## x1                   0.006494   0.012796   0.507    0.612
## x2                   0.020445   0.015502   1.319    0.187
## factor(incumbent)0 -3.385032   0.680672  -4.973 6.59e-07 ***
## factor(incumbent)1 -7.443704   0.739656 -10.064  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 478.27  on 347  degrees of freedom
## Residual deviance: 105.69  on 343  degrees of freedom
## AIC: 115.69
##
## Number of Fisher Scoring iterations: 6
```

2. Fit a robit regression and assess model fit.
3. Which model do you prefer?

Q2&3 are not covered in the class.

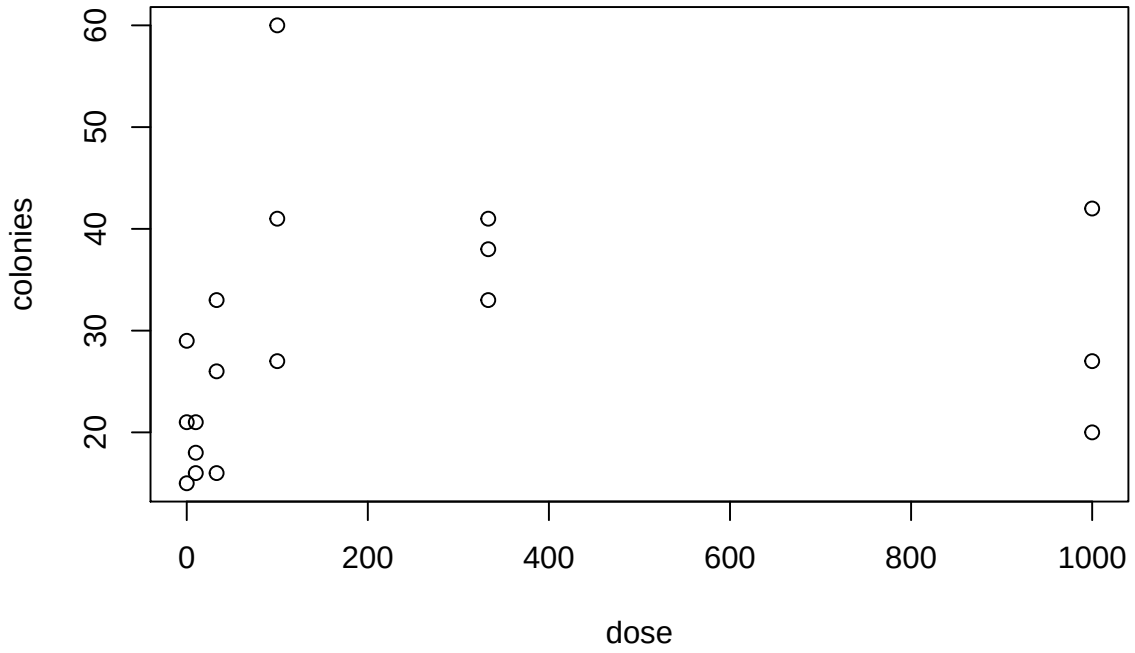
Salmonella

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.

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



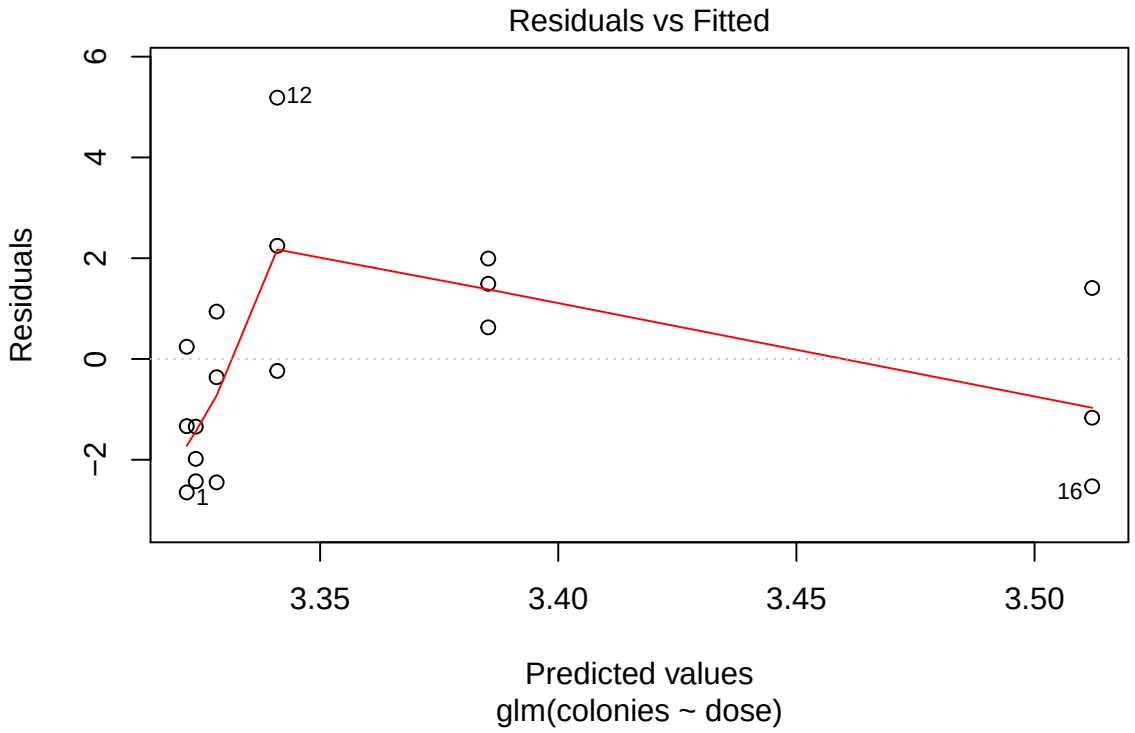
Since we are fitting log linear model we should look at the data on log scale. Also becase the dose is not equally spaced on the raw scale it may be better to plot it on the log scale as well.

```r
m1.salmonella = glm(colonies ~ dose, data = salmonella, family=poisson(link="log"))
summary(m1.salmonella)
```

```
##
## Call:
## glm(formula = colonies ~ dose, family = poisson(link = "log"),
##     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
```

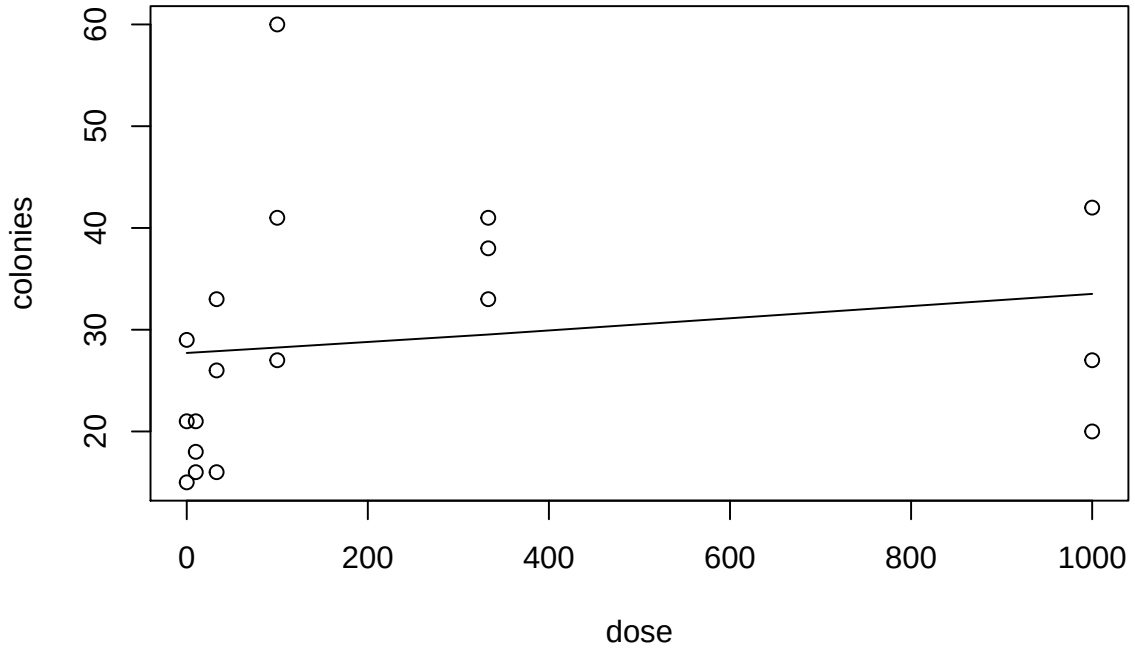
This shows that the trend is not monotonic. Hence when you fit the model and look at the residual you will see a trend.

```
plot(m1.salmonella,which=1)
```



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

```
plot(colonies ~ dose, data = salmonella)
lines(salmonella$dose, predict.glm(m1.salmonella, type="response"))
```



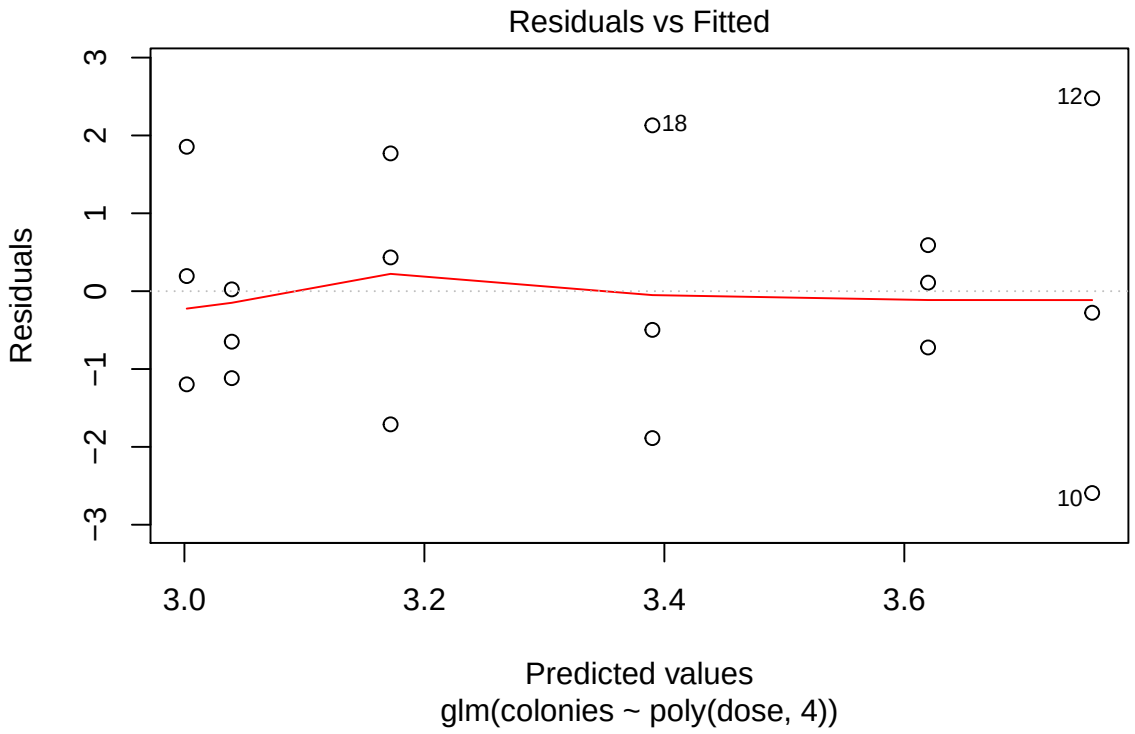
How do we adress this problem? The serious problem to address is the nonlinear trend of dose ranther than the overdispersion since the line is missing the points. Let's add a beny line with 4th order polynomial.

```r
m2.salmonella = glm(colonies ~ poly(dose,4), data = salmonella, family=poisson(link="log"))
summary(m2.salmonella)
```

```
##
## Call:
## glm(formula = colonies ~ poly(dose, 4), family = poisson(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 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.

```r
plot(m2.salmonella,which=1)
```



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

```
m3.salmonella = glm(colonies ~ poly(dose,4), data = salmonella, family=quasipoisson(link = "log"))
summary(m3.salmonella)
```

```
##
## 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.

```
ships2 = subset(ships, service > 0)
ships2$year = as.factor(ships2$year)
ships2$period = as.factor(ships2$period)
m1.ships =  glm(incidents ~ type + year + period,family = poisson(link = "log"), data = ships2,
offset = log(service))
m2.ships = update(m1.ships, family = quasipoisson(link = "log"))
anova(m2.ships, 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 ***
## year    3   41.534        26     49.355 8.1871 0.0005777 ***
## period  1   10.660        25     38.695 6.3039 0.0188808 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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?

```
m1.dvisits = glm(doctorco ~ sex + age + agesq + income + levyplus + freepoor + freerepa + illness
                 + actdays + hscore + chcond1 + chcond2, family=poisson, data = dvisits)
summary(m1.dvisits)
```

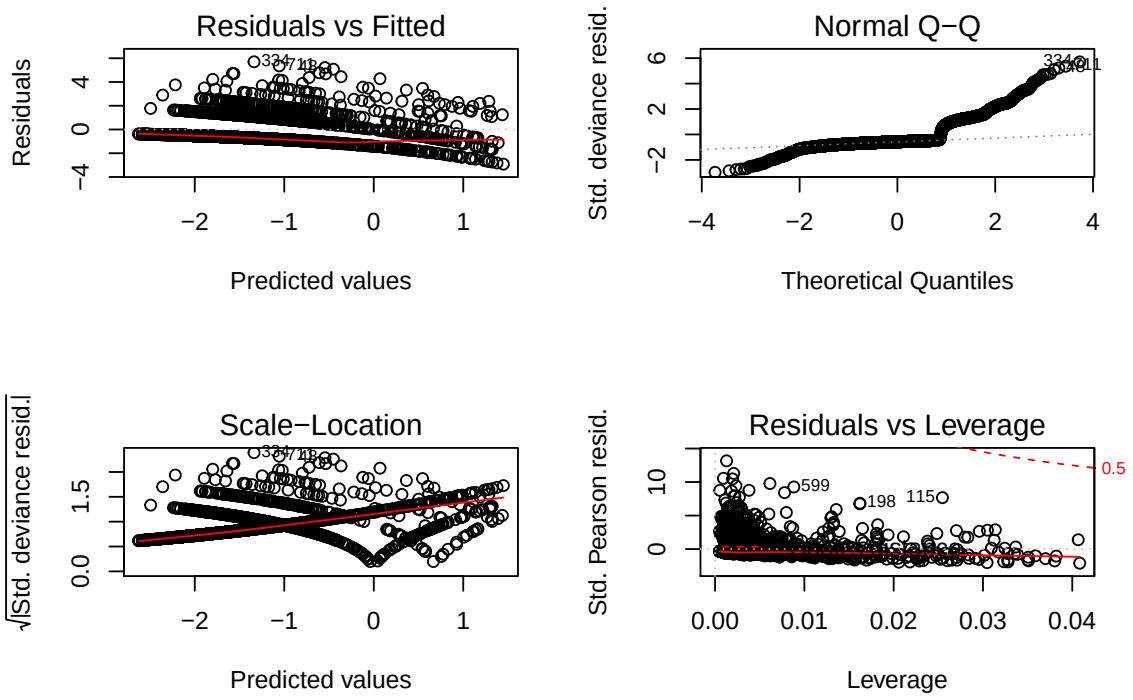
```
##
## Call:
## glm(formula = doctorco ~ sex + age + agesq + income + levyplus +
##     freepoor + freerepa + illness + actdays + hscore + chcond1 +
```

```
##     chcond2, family = poisson, data = dvisits)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9170  -0.6862  -0.5743  -0.4839   5.7005
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.223848   0.189816 -11.716   <2e-16 ***
## sex          0.156882   0.056137   2.795   0.0052 **
## age          1.056299   1.000780   1.055   0.2912
## agesq       -0.848704   1.077784  -0.787   0.4310
## income      -0.205321   0.088379  -2.323   0.0202 *
## levyplus     0.123185   0.071640   1.720   0.0855 .
## freepoor    -0.440061   0.179811  -2.447   0.0144 *
## freerepa     0.079798   0.092060   0.867   0.3860
## illness      0.186948   0.018281  10.227   <2e-16 ***
## actdays      0.126846   0.005034  25.198   <2e-16 ***
## hscore       0.030081   0.010099   2.979   0.0029 **
## chcond1      0.114085   0.066640   1.712   0.0869 .
## chcond2      0.141158   0.083145   1.698   0.0896 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5634.8  on 5189  degrees of freedom
## Residual deviance: 4379.5  on 5177  degrees of freedom
## AIC: 6737.1
##
## Number of Fisher Scoring iterations: 6
```

The residual deviance is quite high. Probably not the best fit.

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

```
par(mfrow=c(2,2))
plot(m1.dvisits)
```



There are lines because the responses are discrete continuous numbers.

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

age, income, hscore, actdays and illness are statistically significant, which makes sense.

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.

```
predict(m1.dvisits, dvisits[5190,], type="response")
```

```
##      5190
## 0.1533837
```

```
#The mean amount of visits to the doctor for patient 5190 would be 0.16 visits. We will set lambda =0.1
print(paste0("Probability of 0 doctor's visits: ", round(dpois(0, lambda = 0.153),3)))
```

```
## [1] "Probability of 0 doctor's visits: 0.858"
```

```
print(paste0("Probability of 0 doctor's visits: ", round(dpois(1, lambda = 0.153),3)))
```

```
## [1] "Probability of 0 doctor's visits: 0.131"
```

```
print(paste0("Probability of 0 doctor's visits: ", round(dpois(2, lambda = 0.153),3)))
```

```
## [1] "Probability of 0 doctor's visits: 0.01"
```

```
print(paste0("Probability of 0 doctor's visits: ", round(dpois(3, lambda = 0.153),3)))
```

```
## [1] "Probability of 0 doctor's visits: 0.001"
```

```
print(paste0("Probability of 0 doctor's visits: ", round(dpois(4, lambda = 0.153),3)))
```

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

```
print(paste0("Probability of 0 doctor's visits: ", round(dpois(5, lambda = 0.153),3)))
```

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

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

```
m2.dvisits = lm(doctorco ~ sex + age + agesq + income + levyplus + freepoor + freerepa + illness + actda
summary(m2.dvisits)
```

```
##
## 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 8.39e-13 ***
## actdays      0.103192   0.003657  28.216  < 2e-16 ***
## 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: < 2.2e-16
```

```
predict(m2.dvisits, dvisits[5190,])
```

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

It appears that it isn't likely to be too different.