Homework 04

Generalized Linear Models

Kerui Cao

October 5, 2017

Data analysis

Poisson regression:

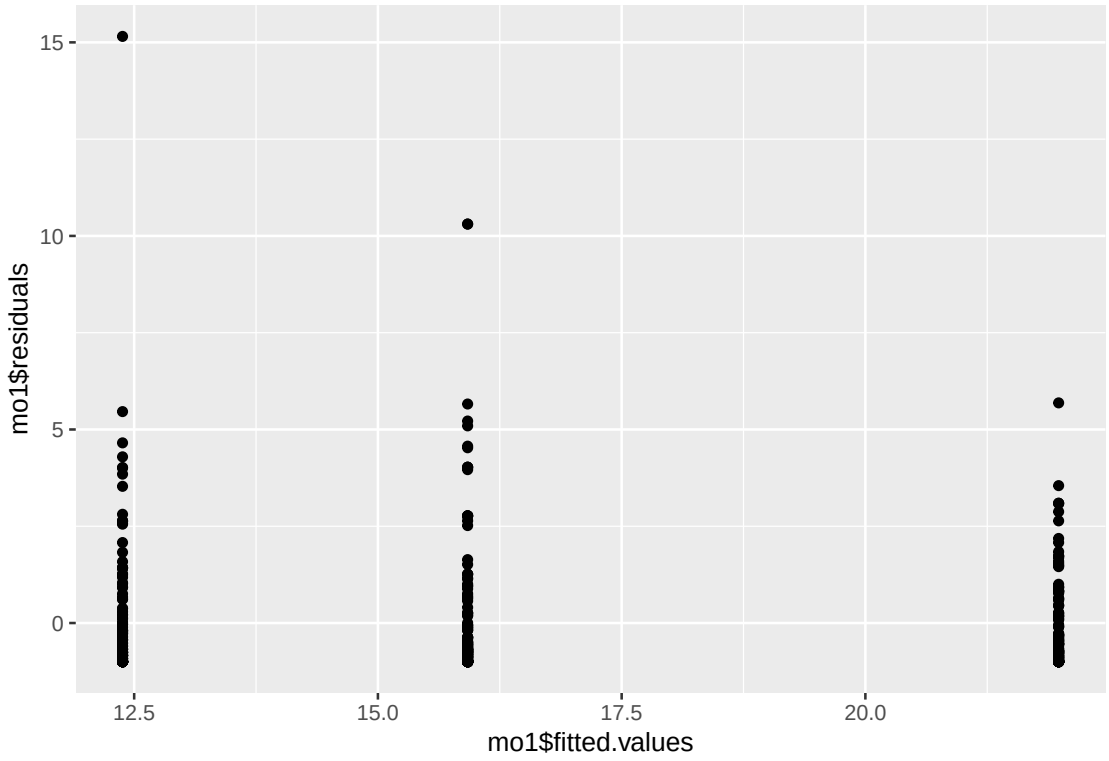
The folder `risky.behavior` contains data from a randomized trial targeting couples at high risk of HIV infection. The intervention provided counseling sessions regarding practices that could reduce their likelihood of contracting HIV. Couples were randomized either to a control group, a group in which 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?

The regression result is shown below:

```
##
## Call:
## glm(formula = fupacts ~ couples + women_alone, family = poisson,
##     data = da)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.6306  -4.9761  -3.2026   0.9829  27.1593
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.09024    0.01900  162.63   <2e-16 ***
## couples     -0.32263    0.02736  -11.79   <2e-16 ***
## women_alone -0.57409    0.03024  -18.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 13307  on 433  degrees of freedom
## Residual deviance: 12931  on 431  degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 6
```

Plot the residuals v.s. fitted value, we can tell there may exist overdispersion problem:



More over we can estimate the overdispersion factor:

```
yhat = predict(mo1,type = "response")
z = (da$fupacts - yhat)/sqrt(yhat)
cat ("overdispersion ratio is ", sum(z^2)/(434-2), "\n")
```

```
## overdispersion ratio is  44.04779
```

```
cat ("p-value of overdispersion test is ", pchisq (sum(z^2), 434-2), "\n")
```

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

So we can conclude that this model contains serious overdispersion problems.

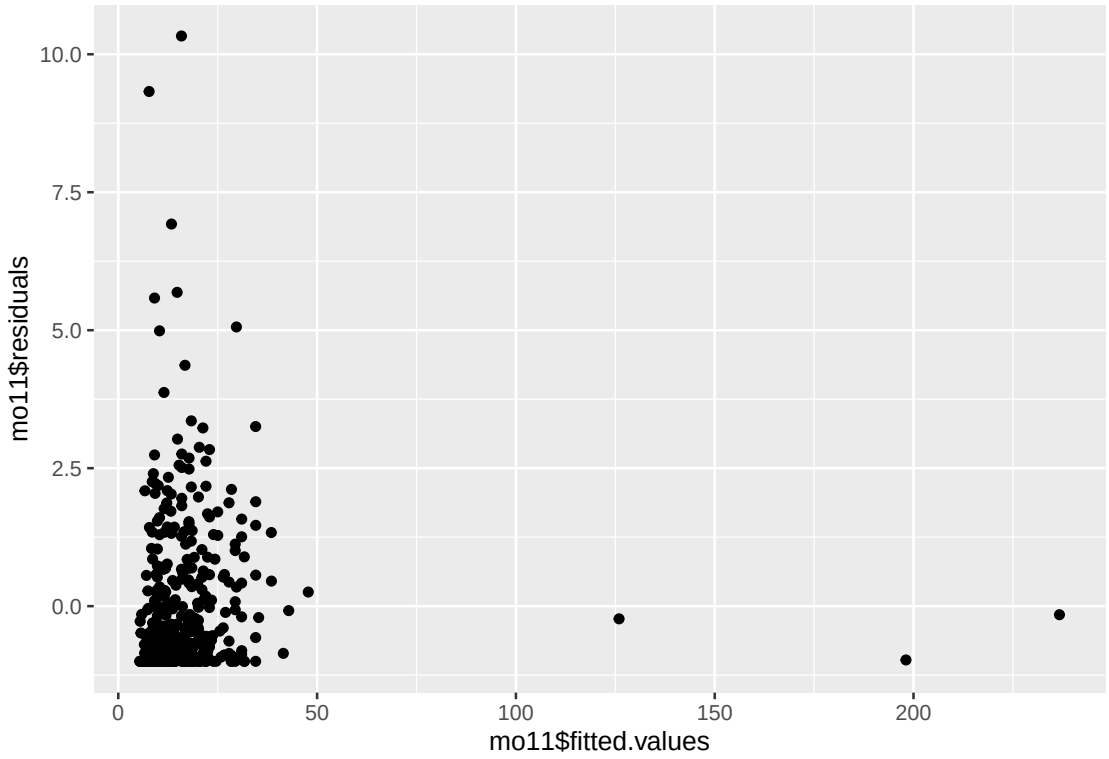
2. Next extend the model to include pre-treatment measures of the outcome and the additional pre-treatment variables included in the dataset. Does the model fit well? Is there evidence of overdispersion?

The result of new model is shown below:

```
##
## Call:
## glm(formula = fupacts ~ couples + women_alone + bupacts + bs_hiv +
##     sex, family = poisson, data = da)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -18.692   -4.303   -2.516    1.367   23.360
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.8961453  0.0232025 124.821  < 2e-16 ***
## couples        -0.4103990  0.0282237 -14.541  < 2e-16 ***
```

```
## women_alone    -0.6643498  0.0309055 -21.496  < 2e-16 ***
## bupacts         0.0107834  0.0001738  62.043  < 2e-16 ***
## bs_hivpositive -0.4375972  0.0353724 -12.371  < 2e-16 ***
## sexman         -0.1084953  0.0237295  -4.572 4.83e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 13307  on 433  degrees of freedom
## Residual deviance: 10204  on 428  degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 6
```

Still plot the residuals against fitted value:



And also estimates the overdispersion factor:

```
## overdispersion ratio is  29.73432
```

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

According to the residuals plot and overdispersion factor, we can still conclude that this new model contains serious overdispersion problem.

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

There are several choices to deal with overdispersion, we can use quasi-Poisson model, or fit the negative-binomial model, here we choose the quasi-Poisson Model:

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

```
## glm(formula = fupacts ~ couples + women_alone + bupacts + bs_hiv +
##     sex, family = quasipoisson, data = da)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -18.692   -4.303   -2.516    1.367   23.360
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.8961453  0.1271110  22.784  < 2e-16 ***
## couples        -0.4103990  0.1546192  -2.654 0.008245 **
## women_alone    -0.6643498  0.1693108  -3.924 0.000102 ***
## bupacts         0.0107834  0.0009522  11.325  < 2e-16 ***
## bs_hivpositive -0.4375972  0.1937824  -2.258 0.024437 *
## sexman         -0.1084953  0.1299981  -0.835 0.404413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 30.01223)
##
##     Null deviance: 13307  on 433  degrees of freedom
## Residual deviance: 10204  on 428  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

According to the result of regression model, we can see that the coefficients of two intervention are significant, so we may conclude that the intervention is effective, but further analysis is recommended.

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?

Couple and Women_alone variables won‘t be i.i.d.

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)

I take the Well data as example, and fit logit and probit models, below is the coeffificients of two models and adjusted coefficienta of Probit model.

	Logit	Probit	Adj-Probit
(Intercept)	-0.2139325	-0.1211576	-0.1938521
arsenic	0.4683638	0.2771243	0.4433989
dist	-0.0089564	-0.0054520	-0.0087233
educ	0.0428201	0.0268409	0.0429454

Comparing logit and probit:

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

We constructed new arsenic,dist and educ data, which comply with exponential and uniform distribution, then predict through Logit and Probit model.

Logit	Probit
1.9178645	1.1349955
-0.2890498	-0.1655915
0.1441964	0.0901431
1.1300607	0.6817971
-0.7758053	-0.4593481
-0.3037436	-0.2157526
0.2679312	0.1444828
-0.0504610	-0.0437733
0.9852566	0.5763158
1.7866690	1.0457937

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)

Loaded the data and we can see that compared to other variables, re74, re75 and re78 are way too big, so we normalize them. The regression result is shown below:

```
##
## Call:
## vglm(formula = re78 ~ treat + age + educ + black + hisp + married +
##     nodegree + re74 + re75, family = tobit(Upper = 121174), data = da3)
##
## Pearson residuals:
##                 Min      1Q  Median      3Q    Max
## mu          -6.660 -0.5004 -0.2211  0.3214  20.66
## loglink(sd) -1.007 -0.7085 -0.2947 -0.1376 286.37
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept):1 -0.1478396  0.0438717  -3.370 0.000752 ***
## (Intercept):2 -0.4898960  0.0066244 -73.953  < 2e-16 ***
## treat          0.1468918  0.0730128   2.012 0.044234 *
## age           -0.0070426  0.0005737 -12.276  < 2e-16 ***
## educ           0.0351261  0.0027193  12.917  < 2e-16 ***
## black         -0.0496659  0.0187084  -2.655 0.007937 **
## hisp          -0.0237510  0.0221604  -1.072 0.283821
## married        0.0550733  0.0140489   3.920 8.85e-05 ***
## nodegree       0.0331824  0.0172164   1.927 0.053933 .
```

```
## re74           0.2833850  0.0105823  26.779  < 2e-16 ***
## re75           0.4312946  0.0103455  41.689  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: mu, loglink(sd)
##
## Log-likelihood: -13374.39 on 37323 degrees of freedom
##
## Number of Fisher scoring iterations: 12
##
## No Hauck-Donner effect found in any of the estimates
```

Treat: If someone from the New South Wales team, the expected z-score income in 1978 will be 0.14 units higher than someone from the CPS or PSID group.

Age: For every 1 year increase in age, the expected z-score income for 1978 will decrease by 0.007 units, while all other variables in the model remain unchanged.

Educ: For every 1 year increase in education, the expected z score in 1978 will increase by 0.035 units while keeping all other variables in the model unchanged.

Balck: If someone is black, the expected z-score income in 1978 will be 0.049 units lower than the non-black person in the same situation.

Hisp: If someone is Hispanish, the expected z-score income in 1978 will be 0.023 units lower than the non-black person in the same situation.

Married: If someone is married, the expected z-score income in 1978 will be 0.055 units higher than the income of the same but unmarried person.

Nodegree: If someone has no senior high school diploma, the expected z-score income in 1978 will be 0.033 units higher than the income of the same but has senior high school diploma.

re74: One standard deviation increase in revenue in 1974, the expected z-score income in 1978 will increase 0.28 units.

re75: One standard deviation increase in revenue in 1975, the expected z-score income in 1978 will increase 0.43 units.

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.

First we select observations of year 1986 and 1988, and delete observations with missing values.

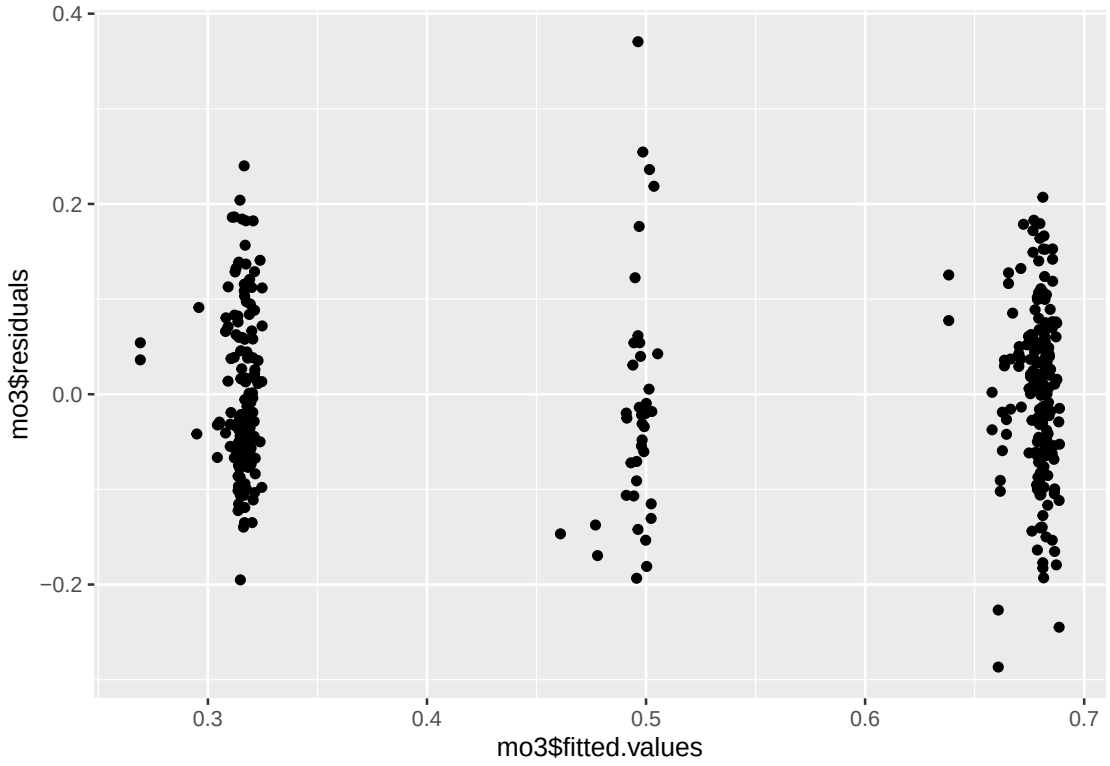
1. Fit a linear regression (with the usual normal-distribution model for the errors) predicting 1988 Democratic vote share from the other variables and assess model fit.

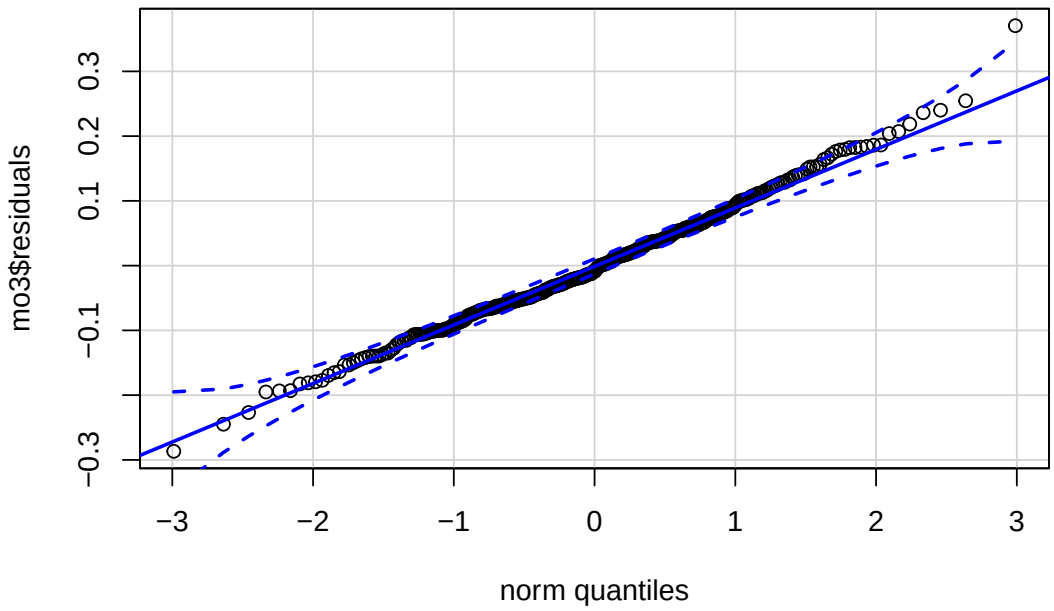
First fit the model and the result is shown below:

```
##
## Call:
## lm(formula = Dem_pct ~ x1 + x2 + incumbent, data = da4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.28674 -0.06218 -0.00861  0.05973  0.37051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.5078399  0.0103685  48.979   <2e-16 ***
## x1          -0.0001581  0.0002325  -0.680    0.497
## x2          -0.0004689  0.0003612  -1.298    0.195
## incumbent    0.1819812  0.0053659  33.914   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09404 on 354 degrees of freedom
## Multiple R-squared:  0.7687, Adjusted R-squared:  0.7668
## F-statistic: 392.3 on 3 and 354 DF,  p-value: < 2.2e-16
```

The R^2 of model is high, but the significance of coefficients are low, plot the residuals against the fitted value:





We can see the residuals of model comply with normal distribution, which comply with model assumption.

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.

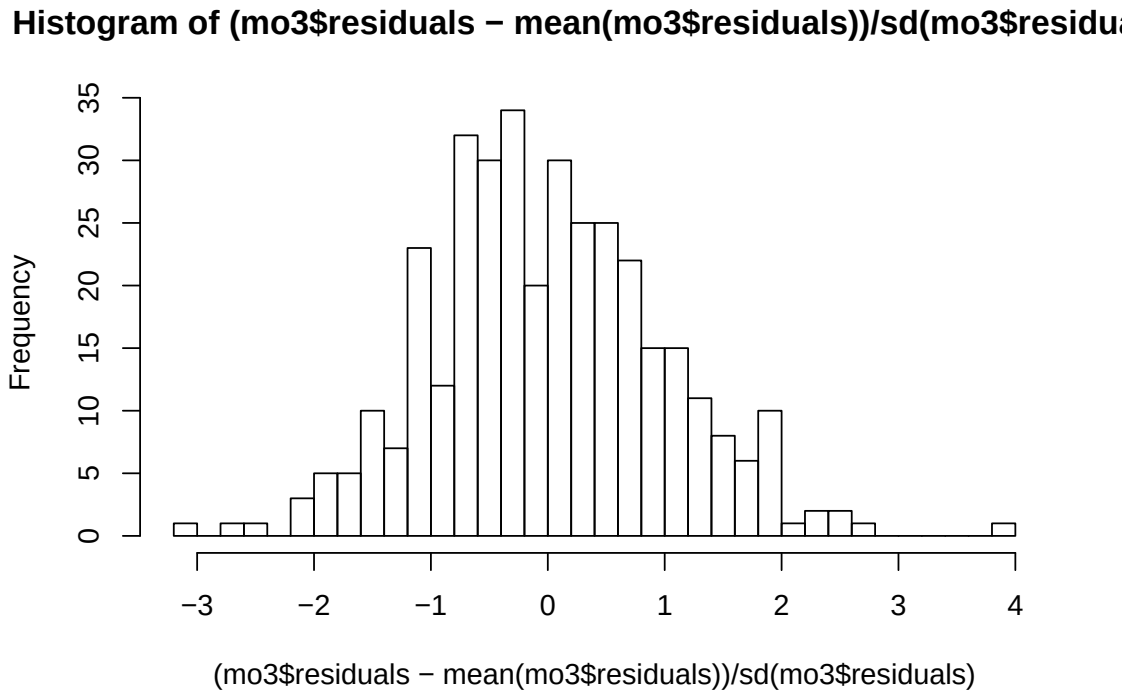
Use "tlm" fit Student-T regression and the result is shown below:

```
## Location model :
##
## Call:
## tlm(lform = Dem_pct ~ x1 + x2 + incumbent, data = da4)
##
## Residuals:
##       Min         1Q     Median         3Q        Max
## -0.299157  -0.055688  -0.002919   0.059408   0.375616
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4977343  0.0097466  51.067   <2e-16 ***
## x1          -0.0001007  0.0002185  -0.461    0.645
## x2          -0.0001815  0.0003395  -0.535    0.593
## incumbent    0.1895097  0.0050441  37.570   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Scale parameter(s) as estimated below)
##
##
## Scale Model :
```

```
##
## Call:
## tlm(lform = Dem_pct ~ x1 + x2 + incumbent, data = da4)
##
## Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0000  -1.5855  -0.5724   1.2487   5.2022
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.2573     0.1057  -49.74   <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 : 7 in 0
## Heteroscedastic t Likelihood : 331.1057
```

3. Which model do you prefer?

Plot the standardized residuals from simple linear regression model:



According to the histgram above, we can tell that there exists some outliers, which is more than 2 standard diviations away from the mean, so the Student-T regerssion may be better.

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.

First create variable representing whether Democrati wins,and then fit logistic model.

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

```
##
## Call:
## glm(formula = win ~ x1 + x2 + incumbent, family = binomial, data = da4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8664  -0.1722   0.1931   0.2406   2.9705
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.24785    0.54602   0.454   0.6499
## x1          -0.01042    0.01139  -0.915   0.3602
## x2          -0.02641    0.01457  -1.813   0.0699 .
## incumbent    3.92226    0.40477   9.690   <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: 494.94  on 357  degrees of freedom
## Residual deviance: 121.75  on 354  degrees of freedom
## AIC: 129.75
##
## Number of Fisher Scoring iterations: 6
```

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

Salmonellla

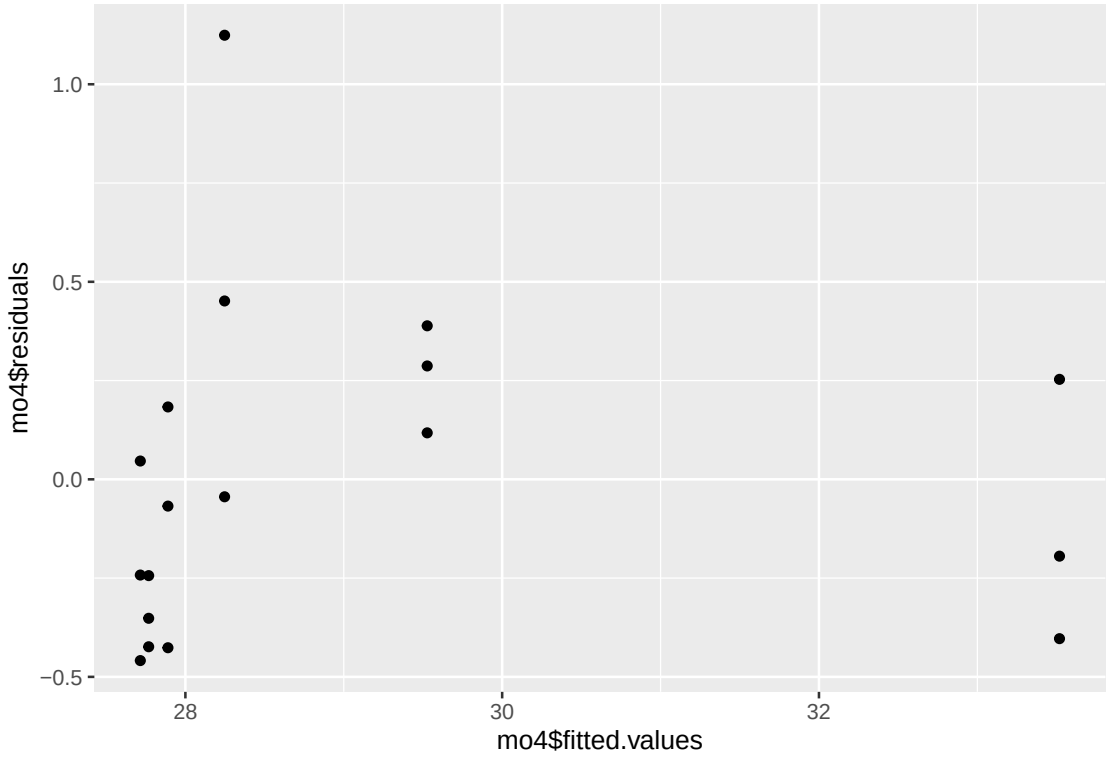
The `salmonella` data was collected in a salmonella reverse mutagenicity assay. The predictor is the dose level of quinoline and the response is the numbers of revertant colonies of TA98 salmonella observed on each of three replicate plates. Show that a Poisson GLM is inadequate and that some overdispersion must be allowed for. Do not forget to check out other reasons for a high deviance.

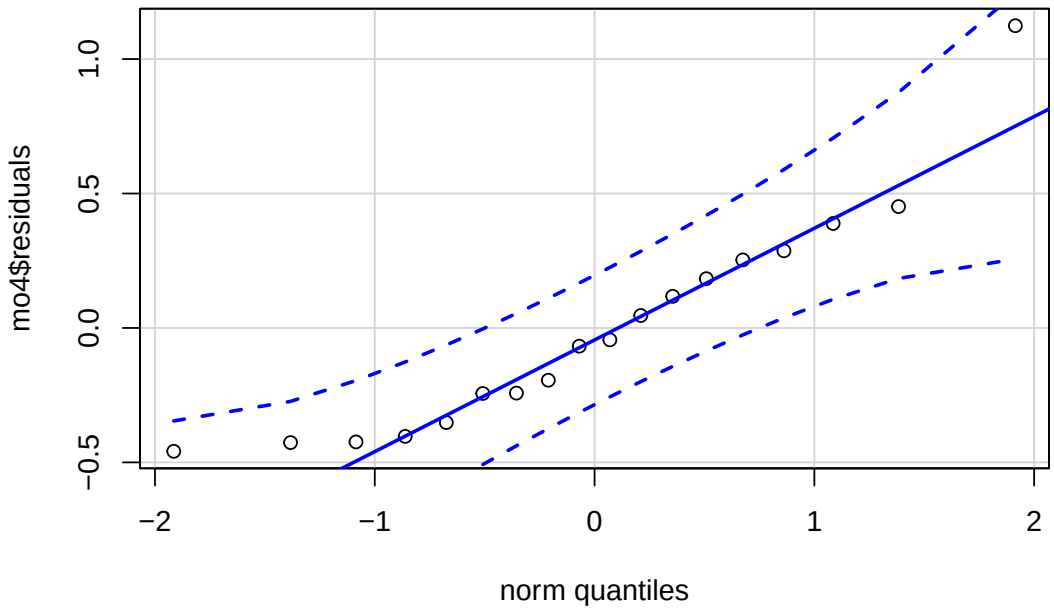
First fit a Poisson regression and check the model, below is the result of the model:

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

```
## (Intercept) 3.3219950  0.0540292  61.485   <2e-16 ***
## dose        0.0001901  0.0001172   1.622    0.105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 78.358  on 17  degrees of freedom
## Residual deviance: 75.806  on 16  degrees of freedom
## AIC: 172.34
##
## Number of Fisher Scoring iterations: 4
```

Then plot the residuals against the fitted values:





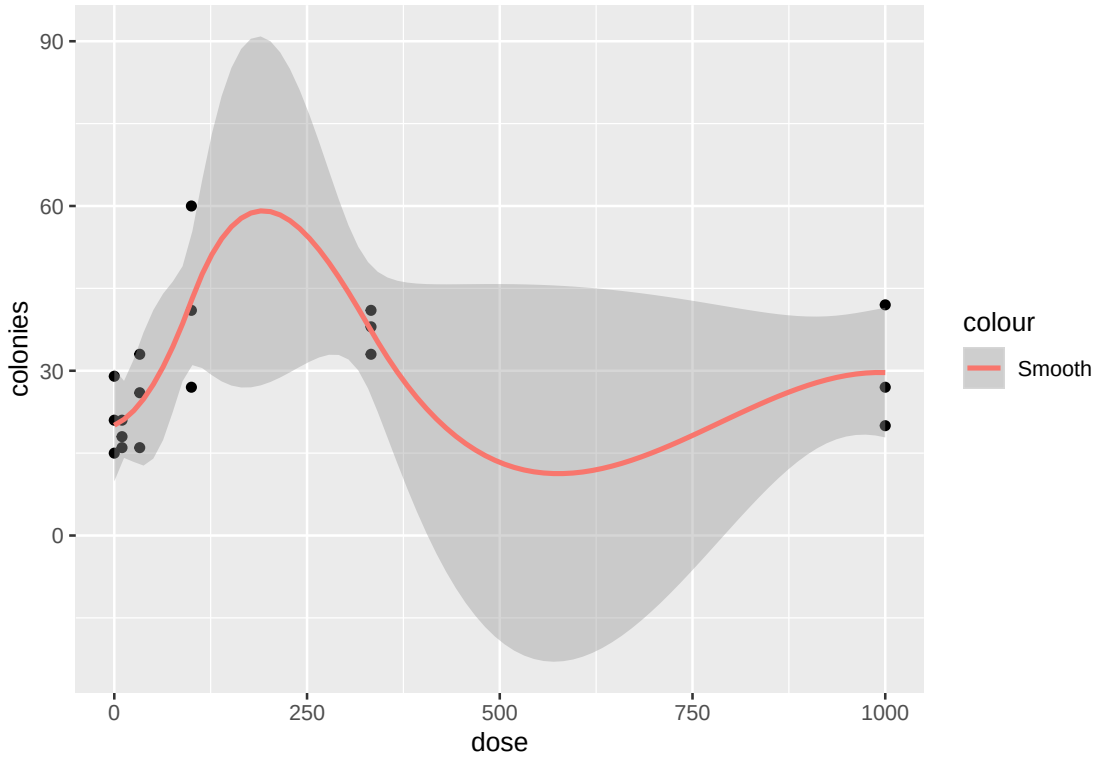
According to the residuals plot, the model seems fine and no obvious overdispersion, then estimate the overdispersion factor:

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

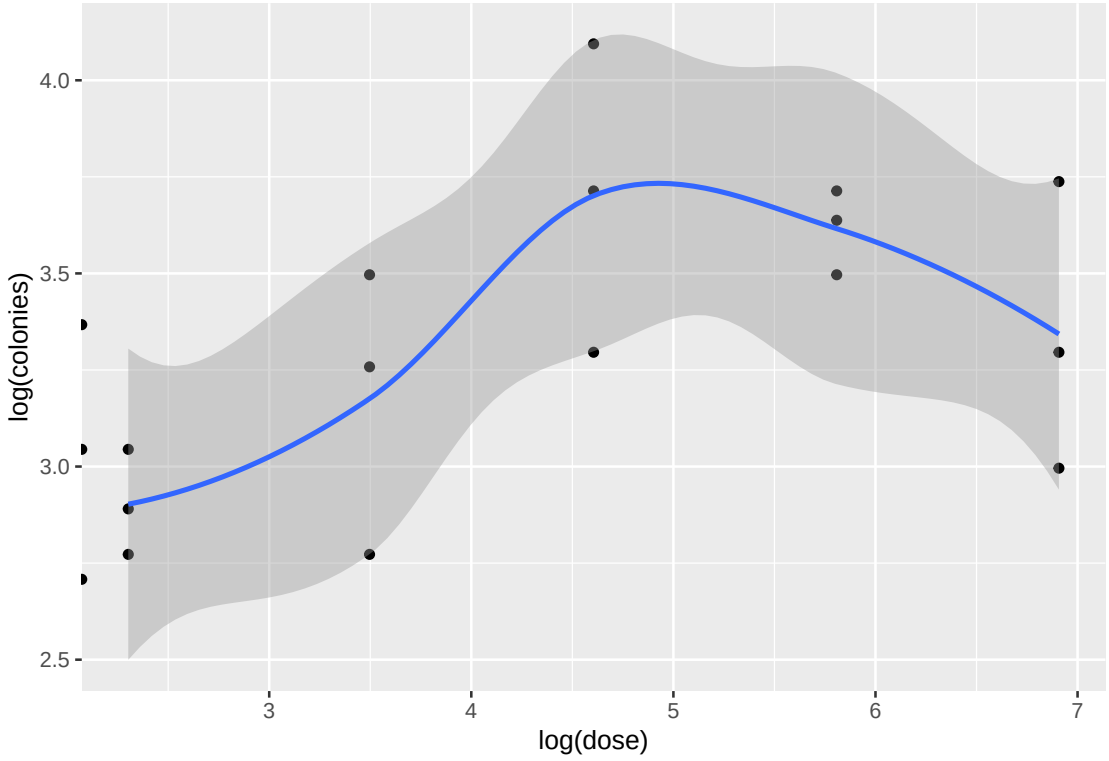
```
## p-value of overdispersion test is  1
```

According to the result above, the overdispersion factor shows that there exists overdispersion problem, so normal Poisson Regression does not fit.

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.



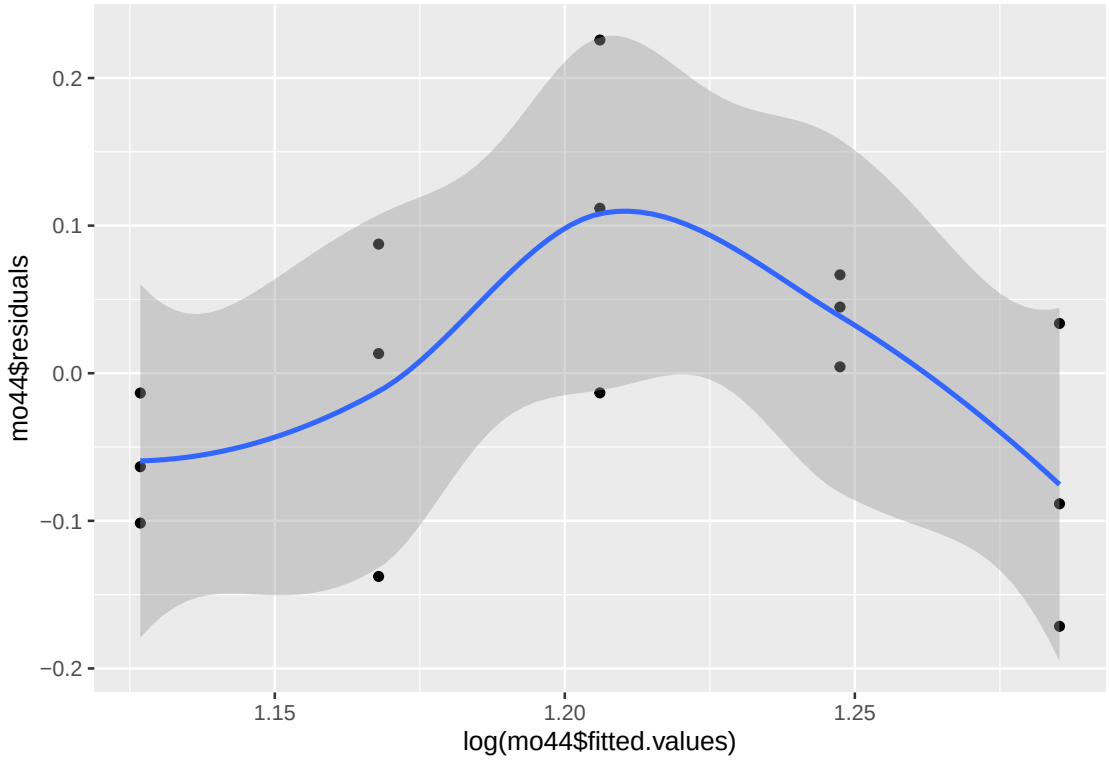
Since we are fitting log linear model we should look at the data on log scale. Also becase the dose is not equally spaced on the raw scale it may be better to plot it on the log scale as well.



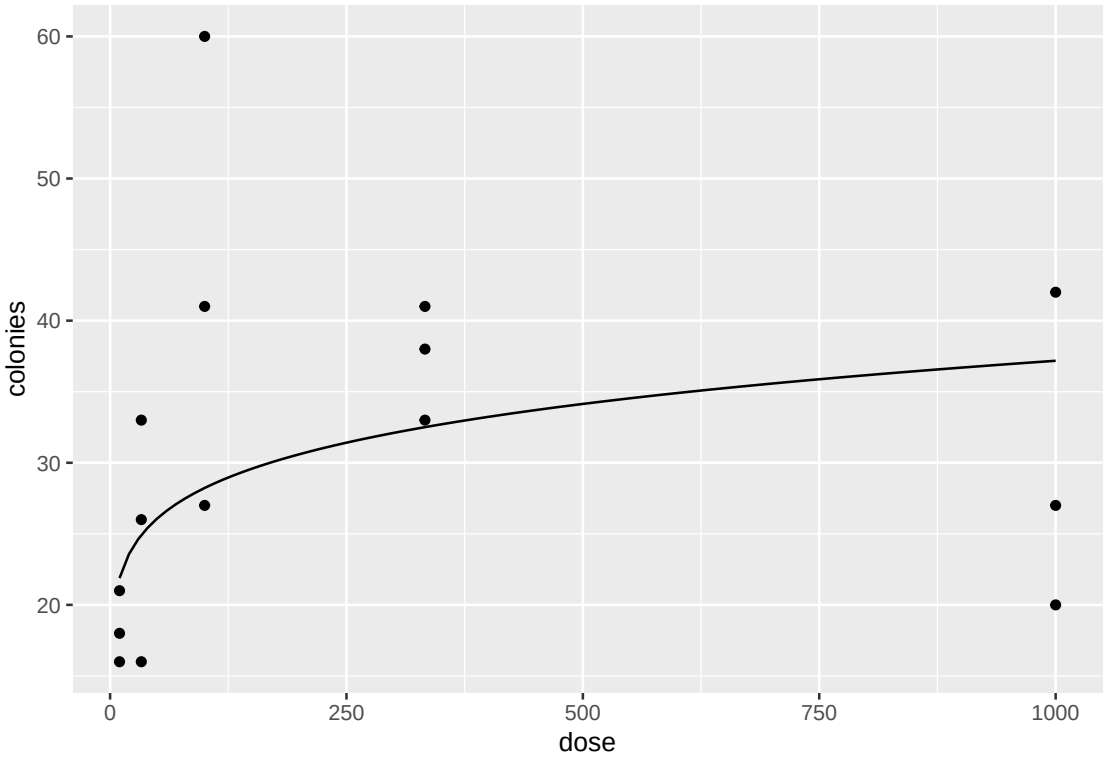
This shows that the trend is not monotonic. Hence when you fit the model and look at the residual you will

see a trend.

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



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



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

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.

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.

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

Treat type, year and period as categorical variables, and add an offset term $\log(service)$, below is the result:

```
##
## Call:
## glm(formula = incidents ~ offset(log(service)) + factor(type) +
##     factor(year) + factor(period), family = poisson, data = da5)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6768  -0.8293  -0.4370   0.5058   2.7912
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -6.40590    0.21744 -29.460  < 2e-16 ***
## factor(type)B    -0.54334    0.17759  -3.060  0.00222 **
```

```
## factor(type)C    -0.68740    0.32904  -2.089  0.03670 *
## factor(type)D    -0.07596    0.29058  -0.261  0.79377
## factor(type)E     0.32558    0.23588   1.380  0.16750
## factor(year)65    0.69714    0.14964   4.659 3.18e-06 ***
## factor(year)70    0.81843    0.16977   4.821 1.43e-06 ***
## factor(year)75    0.45343    0.23317   1.945  0.05182 .
## factor(period)75  0.38447    0.11827   3.251  0.00115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 146.328  on 33  degrees of freedom
## Residual deviance:  38.695  on 25  degrees of freedom
## AIC: 154.56
##
## Number of Fisher Scoring iterations: 5
```

We can tell from above result that type, year and period are all important variables.

THe intercept means that the average ratio of incidents of a type A ship is 0.0016518, which was constructed in year 1965 and servised between year 1960-1974, set this as base, most other types of ships will have lower ratio of incidents, and ships constructed at year other than 1960 have higher ratio of incidents, and ships servises between 1975-1979 had higher ratio of incidents.

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.

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?

The result of this model is shown below:

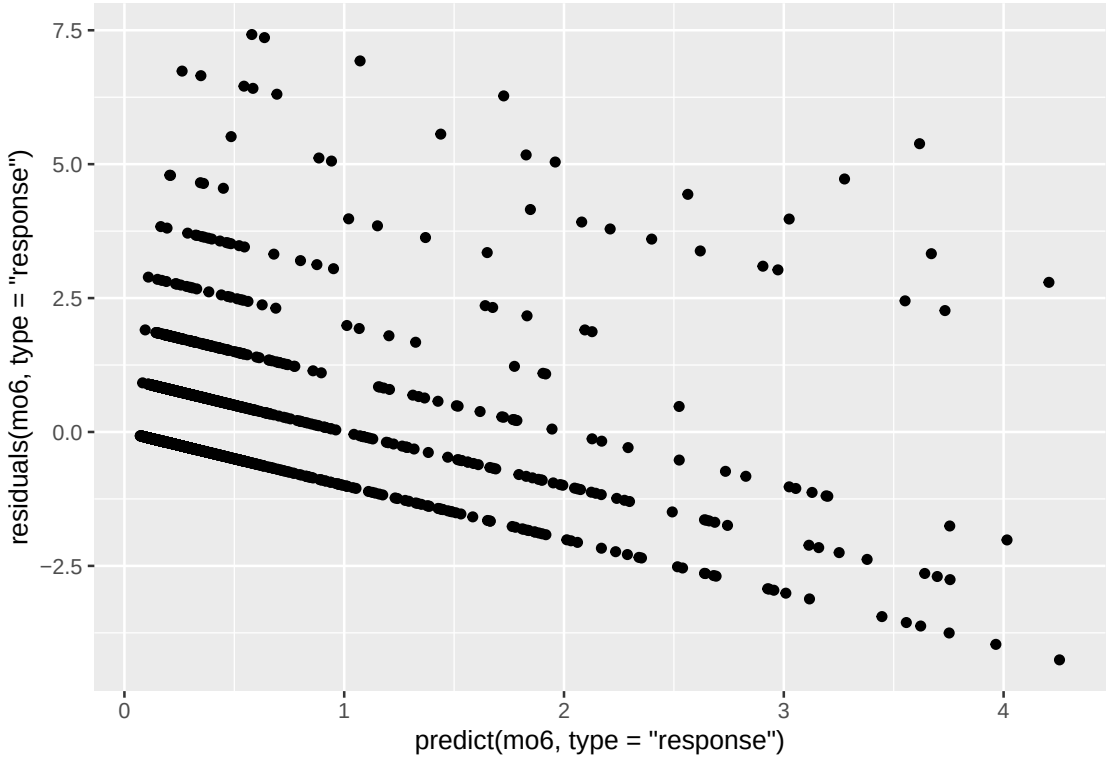
```
##
## Call:
## glm(formula = doctorco ~ sex + age + agesq + income + levyplus +
##     freepoor + freerepa + illness + actdays + hscore + chcond1 +
##     chcond2, family = poisson, data = da6)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9170  -0.6862  -0.5743  -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
```

```
## Deviance of this model is 4379.515
```

The deviance of this model is big, so this model does fit well.

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



These lines appear on the graph because the response residuals are given by $y_i - \hat{y}_i$ and y_i has only a limited number of values. Each lines corresponds to a different possible value of y_i.

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

A person who is a female with large age, low income, covered by private health insurance fund for private patient in public hospital, not covered by government, recent immigrant and unemployed disability pension or invalid veteran or family of deceased veteran, with large number of illnesses in past 2 weeks and large number of days of reduced activity in past two weeks due to illness or injury, with high health questionnaire score and chronic conditions, then the person visits the doctor the most under my selected model

4. For the last person in the dataset, compute the predicted probability distribution for their visits to the doctor, i.e., give the probability they visit 0,1,2, etc. times.

   The possibility of :

   Visiting 0 time : 0.8578005

   Visiting 1 time : 0.1315726

   Visiting 2 time : 0.0100905

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