Homework 04

Generalized Linear Models

Chaoqun Yin

October 5, 2018

Data analysis

Poisson regression:

The folder `risky.behavior` contains data from a randomized trial targeting couples at high risk of HIV infection. The intervention provided counseling sessions regarding practices that could reduce their likelihood of contracting HIV. Couples were randomized either to a control group, a group in which just the woman participated, or a group in which both members of the couple participated. One of the outcomes examined after three months was "number of unprotected sex acts". The variables bupacts: before treatment; fupacts: after treatment.

1. Model this outcome as a function of treatment assignment using a Poisson regression. Does the model fit well? Is there evidence of overdispersion?

```
risk <- risky_behaviors

#Fit the model
risk$fupacts <- round(risk$fupacts)
fit.1 <- glm(formula = fupacts ~ factor(women_alone) + factor(couples), data = risk, family = poisson)
summary(fit.1)
```

```
##
## Call:
## glm(formula = fupacts ~ factor(women_alone) + factor(couples),
##     family = poisson, data = risk)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.6285  -4.9794  -3.2015   0.9847  27.1502
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            3.08960    0.01901  162.55   <2e-16 ***
## factor(women_alone)1 -0.57212    0.03023  -18.93   <2e-16 ***
## factor(couples)1     -0.32243    0.02737  -11.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 13299  on 433  degrees of freedom
## Residual deviance: 12925  on 431  degrees of freedom
## AIC: 14256
##
## Number of Fisher Scoring iterations: 6
```

It seems that the model fits well. Because the coefficients are statistically significiant. And the duifference between residual deviance and null deviance is pretty large.

```
#Specify the n and the k
n = nrow(risk)
k = length(fit.1$coefficients)
#Calculate yhat
yhat <- predict(fit.1, type = "response")
#Calculate standarized residuals
z <- (risk$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 overdispersion ratio is big so there is an overdispersion.

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

```
fit.2 <- glm(fupacts ~ factor(sex) + couples + women_alone + factor(bs_hiv) + bupacts  , data = risk, fa
summary(fit.2)
```

```
##
## Call:
## glm(formula = fupacts ~ factor(sex) + couples + women_alone +
##     factor(bs_hiv) + bupacts, family = poisson, data = risk)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -18.679   -4.305   -2.511    1.368   23.361
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.8957952  0.0232074 124.779  < 2e-16 ***
## factor(sex)man          -0.1086694  0.0237301  -4.579 4.66e-06 ***
## couples                 -0.4099761  0.0282298 -14.523  < 2e-16 ***
## women_alone             -0.6622159  0.0308962 -21.434  < 2e-16 ***
## factor(bs_hiv)positive -0.4383170  0.0353804 -12.389  < 2e-16 ***
## bupacts                  0.0107789  0.0001738  62.013  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 13299  on 433  degrees of freedom
## Residual deviance: 10200  on 428  degrees of freedom
## AIC: 11537
##
## Number of Fisher Scoring iterations: 6
```

The AIC of the new model is smaller than the formal one, so the new model fits better.

```
#Specify the n and the k
n = nrow(risk)
```

```r
k = length(fit.2$coefficients)
#Calculate yhat
yhat <- predict(fit.2, type = "response")
#Calculate standarized residuals
z <- (risk$fupacts - yhat)/sqrt(yhat)
cat("Overdispersion ratio is ", sum(z^2)/(n-k), "\n")
```

```
## Overdispersion ratio is  30.00404
```

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

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

The overdispersion ratio is also big so there is an overdispersion. But it is lower than the medel one, so the new model fits better.

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

```r
#fit the model
fit.3 <- glm(fupacts ~ factor(sex) + couples + women_alone + factor(bs_hiv) + bupacts  , data = risk, fa
summary(fit.3)
```

```
##
## Call:
## glm(formula = fupacts ~ factor(sex) + couples + women_alone +
##     factor(bs_hiv) + bupacts, family = quasipoisson, data = risk)
##
## 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(sex)man          -0.1086694  0.1299838  -0.836 0.403609
## couples                 -0.4099761  0.1546315  -2.651 0.008316 **
## women_alone             -0.6622159  0.1692369  -3.913 0.000106 ***
## factor(bs_hiv)positive -0.4383170  0.1937994  -2.262 0.024217 *
## 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
```

Regarding the efficience of the intervention, the variables in the model become less statistically 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, it is kind of misleading. Because if we add women only as a controlling group, we should also add man only as a controlling group to balance whether it is significant that the coefficients

represent.

Comparing logit and probit:

Take one of the data examples from Chapter 5. Fit these data using both logit and probit model. Check that the results are essentially the same (after scaling by factor of 1.6)

```
#Take the well-switching as the example
wells <- read.table("http://www.stat.columbia.edu/~gelman/arm/examples/arsenic/wells.dat", header=TRUE)
wells_dt <- data.table(wells)

wells.logit <- glm(switch ~ log(dist),family = binomial(link = "logit"), data = wells_dt)
summary(wells.logit)
```

```
##
## Call:
## glm(formula = switch ~ log(dist), family = binomial(link = "logit"),
##     data = wells_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6365  -1.2795   0.9785   1.0616   1.2220
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.01971    0.16314   6.251 4.09e-10 ***
## log(dist)   -0.20044    0.04428  -4.526 6.00e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 4097.3  on 3018  degrees of freedom
## AIC: 4101.3
##
## Number of Fisher Scoring iterations: 4
```

```
wells.probit <- glm(switch ~ log(dist),family = binomial(link = "probit"), data = wells_dt)
summary(wells.probit)
```

```
##
## Call:
## glm(formula = switch ~ log(dist), family = binomial(link = "probit"),
##     data = wells_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6389  -1.2795   0.9794   1.0619   1.2196
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.6306     0.1007   6.262 3.79e-10 ***
## log(dist)    -0.1235     0.0274  -4.507 6.57e-06 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 4097.4  on 3018  degrees of freedom
## AIC: 4101.4
##
## Number of Fisher Scoring iterations: 4
```

```
coef(wells.logit)
```

```
## (Intercept)   log(dist)
##   1.0197146  -0.2004422
```

```
coef(wells.probit)*1.6
```

```
## (Intercept)   log(dist)
##   1.0089716  -0.1975926
```

From the results, the coefficients ssentially the same after the coefficients of probit regression scaling by factor of 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.
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.
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.
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.

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.

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.

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

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

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?
2. Plot the residuals and the fitted values-why are there lines of observations on the plot?
3. What sort of person would be predicted to visit the doctor the most under your 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.
5. Fit a comparable (Gaussian) linear model and graphically compare the fits. Describe how they differ.