ENV 710 – Applied Data Analysis
Fall 2014
Lab 11: Generalized Linear Models

Generalized linear models (GLM's) are an extension of regular linear models (i.e., linear regression and ANOVA) to situations where the probability model is not a normal distribution. As in linear regression, we are interested in the relationship between a response variable *Y* and a set of predictors *X*. The response probability distribution can be any member of the exponential family of distributions, which contains the normal distribution, the binomial distribution, and the Poisson distribution, among others. A nonlinear *link* function relates the mean of the response $\mu_i = E(Y_i)$ linearly to a set of terms based on predictor variables. Whereas linear regression was solved by ordinary least squares, the model coefficients, β, are computed by solving for the maximum likelihood.

In this lab, we will focus on Poisson regression, where the response variable consists of count data. This model is used to answer questions such as:

- Number of cargo ships damaged by waves;
- Daily homicide counts in California;
- Number of deaths due to SARS;
- Environmental drivers of numbers of birds along an altitudinal gradient?"

The learning goals of the lab are to:

- Understand the situations in which Poisson regression is used
- Learn to implement Poisson regression in **R**, including special topics like offsets for modeling rates
- Conduct model selection to find the minimum adequate model for a dataset and question of interest
- Interpret GLM coefficients and make predictions from the models.

At the end of the lab, there are two analyses to conduct. Please type your answers to each problem, including any requested graphs, in a Word document. *Submit your answers and your R-code to the class Sakai site under the folder Assignments before 5 pm on either Mon., Nov. 17 (Section 07) or Wed., Nov., 19 (Sections 01 and 02).*

More functions in R

In this lab, we introduce a few new **R** commands.

`glm()` – Used to fit GLMs, and must include specification of the probability distribution, called the family (e.g. family = poisson or family = binomial)
`logLik()` - Generates the log-likelihood of fitted models.
`AIC()` – Generates the Akaike Information Criterion for one or several fitted models.

`anova()` – Computes an analysis of deviance table to evaluate fits of generalize linear models and carry out model comparisons. The function provides different test statistics. For models with known dispersion, the *Chisq* test or *LRT* test is most appropriate. For models where the dispersion is estimated (Gaussian, quasibinomial, quasipoisson), the *F* test is most appropriate.
`lrtest()` – A general function from package *lmtest* for carrying out likelihood ratio tests, compares nested (generalized) linear models.

Poisson Regression

Stops by the NYPD
We will analyze data on the number of police stops of racial and ethnic minorities[1]. Previous studies have confirmed that police stop minorities more often than Whites relative to their proportion in the population. An alternative interpretation is that stop rates more accurately reflect rates of crimes committed by each ethnic group, or that stop rates reflect elevated rates in specific social areas such as neighborhoods or precincts. Here we look at data from pedestrian stops by the NY Police Department over a 15-month period. We compare stop rates by racial and ethnic groups, controlling for previous race-specific arrest rates.

The data can be found in the `frisks.txt` file in Sakai. Here we rename the third column (`past.arrests`) to be "`arrests`" because it is shorter. We also take out one case where the number of arrests is 0.

```
dat <- read.table("frisks.txt", skip = 6, header = T)
 names(dat)[3] <- "arrests"
  dat <- subset(dat[dat$arrests > 0, ])
dat$arrests.yr <- dat$arrests * 15/12
```

The data columns include: `stops` (number of police stops), `pop` (population of precinct), `arrests` (number of arrests in the precinct in the past year), `precinct` (identity of the precinct), `eth` (ethnicity: 1 = Black, 2 = Hispanic, 3 = White), `crime` (type of crime: 1 = violent, 2 = weapons, 3 = property, 4 = drug). Above I added `arrests.yr`, converting arrests into an annual figure instead of a 15-month figure.

Note that the number of police stops are *counts* of stops, thus it is appropriate to use GLM's with a Poisson probability distribution.

Take a look at the data using `pairs()` and `summary()`. Let's also take a look at the number of stops by ethnic group.

```
s.eth <- with(dat, tapply(stops, list(eth), sum))
a.eth <- with(dat, tapply(arrests.yr, list(eth), sum))
s.eth/a.eth
```

[1] Gelman et al. (2007) An analysis of the New York City Police Department's "Stop and Frisk" Policy in the Context of Claims of Racial Bias. J. American Statistical Association 102: 813-823.

The number of stops per ethnic group (`s.eth`) seems to suggest that Blacks are stopped more often than Hispanics or Whites. However, when we divide these numbers by the total number of arrests (`a.eth`) for each of the ethnic groups in the previous year, Hispanics actually have the highest number of stops. So Hispanics, Blacks and then White are stopped most often after accounting for their relative crime rates.

First we fit a model with ethnicity as an indicator:

```
frisk1 <- glm(stops ~ factor(eth), family = poisson,
        data = dat)
 summary(frisk1)
```

This model is the equivalent of `s.eth` above, illustrating the number of stops per ethnic group with Black as the baseline to which Hispanic and White are compared. The coefficients for ethnicities 2 and 3 are both negative, lower than Black that is set to 0.

```
frisk2 <- glm(stops ~ factor(eth), family = poisson,
        offset = log(arrests.yr), data = dat)
```

We need to add an offset so that the counts can be interpreted relative to some baseline or "exposure". In other words, we want to interpret the results as the rate of stops to real arrests. If this rate is high it might represent racial profiling – stopping people because of their race when their actual crime incidence doesn't warrant it. In other applications, we could add an offset to account for different levels of effort (e.g. varying hours of counting birds or varying plot sizes for counting a rare plant).

```
summary(frisk2)
```

```
Call:
glm(formula = stops ~ factor(eth), family = poisson, data = dat,
    offset = log(arrests.yr))

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.811229   0.003784 -214.36   <2e-16 ***
factor(eth)2  0.070208   0.006061   11.58   <2e-16 ***
factor(eth)3 -0.161758   0.008558  -18.90   <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: 183981  on 898  degrees of freedom
Residual deviance: 183297  on 896  degrees of freedom
AIC: 188638
```

Note that adding the offset resulted in the coefficient for ethnicity 2 now being negative relative to the baseline category (Black), similar to our example above (`s.eth/a.eth`).

The two ethnicity coefficients are highly statistically significant, and the difference in

deviance between the null model (without `eth`) and this model is -684, much more than the 2 that would be expected if ethnicity had no explanatory power in the model.

The coefficients can be exponentiated and treated as multiplicative effects.

```
(coef <- exp(coefficients(frisk2)))

(Intercept) factor(eth)2 factor(eth)3
0.4443115    1.0727313    0.8506469
```

The intercept is the prediction if $X_1 = 0$ and $X_2 = 0$, which is the stop rate of Blacks relative to their arrest rate in each precinct. The coefficient of X_1 is the expected difference in y (on the logarithmic scale) when ethnicity is Hispanic. Thus, the expected multiplicative increase is $e^{0.07} = 1.07$, or a 7% increase in the rate of stops. The coefficient for X_2 is the expected difference in y when ethnicity is White: $e^{-0.16} = 0.85$, or a 15% decrease in stops.

Now let's add the 75 `precincts` to the model. There may be good reason to treat `precinct` as a random effect, but let's keep it as a fixed effect here.

```
frisk3 <- glm(stops ~ factor(eth) + factor(precinct),
        family = poisson, offset = log(arrests.yr),
        data = dat)
```

The decrease in deviance from `frisk2` to `frisk3` of 42,014 is huge – much larger than the decrease of 74 that would be expected if the precinct factor were random noise. Therefore, adding precinct has greatly improved the fit of the model to the data. We can also evaluate the different models using AIC. Check out the change in the coefficients of the ethnicities. How has adding `precinct` into the model changed their effects?

```
deviance(frisk2)-deviance(frisk3)
[1] 42014.04

> AIC(frisk2, frisk3)
       df      AIC
frisk2  3 188638.5
frisk3 77 146772.4
```

Under the Poisson distribution, the variance equals the mean. If this is true, then the residuals should be independent, each with a mean of 0 and standard deviation 1. Overdispersion is the case where the variance is much greater than the mean. With overdispersion, we expect the residuals to be much larger, reflecting the extra variation beyond what is predicted under the Poisson model. We can use the typical residual plots to verify.

```
par(mfrow=c(2,2))
plot(frisk3)
```

As you can tell from the figures, the data are very overdispersed! We can calculate

the overdispersion with the following code:

```
n <- nrow(dat)
k <- 77
stops <- dat$stops
yhat <- predict(frisk3, type = "response")
z <- (stops - yhat)/sqrt(yhat)
cat("overdispersion ratio is", sum(z^2)/(n-k), "\n")
```

Overdispersion of 2 is considered a lot, so this is out of the ballpark. To handle overdispersion, we can use the quasipoisson "distribution". This essentially multiplies all the regression standard errors by the square root of the overdispersion $\sqrt{261} = 16.2$.

```
frisk4 <- glm(stops ~ factor(eth) + factor(precinct),
          family = quasipoisson, offset = log(arrests.yr),
          data = dat)
summary(frisk4)
```

Note that this doesn't change the coefficients, but increases the standard errors and reduces the p-values. Fortunately, this doesn't change our main inference, that the rate of stops for Whites is 34.3% lower than Blacks.

Abundance of Salamanders
Let's use a dataset from the *Sleuth3* package on the abundance of salamanders in relation to forest age and percent forest cover. The question of interest is the relationship of salamander numbers to *Forest Age* and *Percent Cover*.

```
require(Sleuth3)
saldat <- case2202
attach(saldat)
```

Let's take a look at the response variable (salamander numbers) in relation to our predictor variables (forest age and percent cover). Note that we add some noise to the *X*-variable with jitter so that we can see otherwise overlapping data points.

```
par(mfrow=c(2,2))
 plot(ForestAge, jitter(Salamanders), las=1,
      pch=21, bg="grey", cex=1.2, ylab = "Salamander
      Count")

  plot(ForestAge, jitter(log(Salamanders+0.1)),
      las=1, pch=21, bg="grey", cex=1.2, ylab =
      "log(Salamander Count)")

  plot(PctCover, jitter(Salamanders), las=1,
      pch=21, bg="grey", cex=1.2,ylab = "Salamander
      Count")

   plot(PctCover, jitter(log(Salamanders+0.1)),
```

```
las=1, pch=21, bg="grey", cex=1.2,
ylab = "log(Salamander Count)")
```

There is a rough break in canopy cover percentage separating closed canopy (>70%) from open canopy (<60%). It may be that mean salamander count only depends on this dichotomy, or more complex models and interactions might be in order. What are the mean numbers of salamanders for open and closed canopy?

```
mean(Salamanders[PctCover<60])
mean(Salamanders[PctCover>70])
```

Because of the apparent division in cover, we are going to add a categorical variable for closed canopy so that closed and open canopy are modeled separately. Note that we have to remove the attached dataframe and re-attach it so that it includes the new variable, *Closed*.

```
rm(saldat)
 saldat <- case2202
  saldat$Closed <- ifelse(saldat$PctCover>60, 1, 0)
attach(saldat)
```

Let's fit a model. We are going to start with a complex model, including quadratic terms and the interaction between *Forest Age* and *Percent Cover*.

```
glm1 <- glm(Salamanders ~ ForestAge + PctCover +
        I(ForestAge^2) + I(PctCover^2) +
        ForestAge*PctCover + factor(Closed) +
        ForestAge:factor(Closed) + PctCover:factor(Closed)
        + I(ForestAge^2):factor(Closed) +
        I(PctCover^2)*factor(Closed) +
        ForestAge:PctCover:factor(Closed),
        data = saldat, family=poisson)
```

That is a very complicated model. Note that all the coefficients that include *Forest Age* are not significant. Let's take *Forest Age* out, as it looks like it does not help explain salamander counts (this agrees with patterns that we saw when first plotting the data).

```
glm2<- glm(Salamanders ~ PctCover + I(PctCover^2) +
       factor(Closed) + PctCover:factor(Closed) +
       I(PctCover^2)*factor(Closed), data = saldat,
      family=poisson)
```

Our estimates of the two-way interactions between *Cover* and *PctCover* are both statistically significant, so it looks like we have the most reduced model. Let's check more formally.

We use the likelihood ratio test (LRT) to compare two models provided the simpler model is a species case of the more complex model (i.e., "nested"). The test is the ratio of two likelihood functions: the simpler model (*s*) has fewer parameter terms

than the complex model (*c*). The test statistic is distributed as a chi-squared random variable, with degree of freedom equal to the difference in the number of parameter between the two models. LRT can be presented as a difference in log-likelihoods which can be expressed in terms of deviance: $LRT = -2\ln(\mathcal{L}_S/\mathcal{L}_C) = -2(\ln(\mathcal{L}_S) - \ln(\mathcal{L}_C)) = -2\ln(\mathcal{L}_S) + 2\ln(\mathcal{L}_C) = deviance_S - deviance_C$.

In **R**, this is conducted using the `anova()` call. Note the use of the chi-squared statistic rather than the F-statistic used in multiple regression.

```
anova(glm1, glm2, test="Chisq")
```

Alternatively, there is also a LRT function in the *lmtest* package.

```
require(lmtest)
lrtest(glm1, glm2)
```

Or, you could calculate it yourself from the model deviances.

```
pchisq(glm2$deviance-glm1$deviance, df=glm2$df.residual-
       glm1$df.residual, lower.tail=F)
```

The LRT demonstrates that `glm2` fits better and that removing the variable *Forest Cover* lost nothing. The significance of the coefficient *Percent Cover*^{2} x *Closed* indicates that there is different curvature in the distributions of salamander counts in open and closed canopy forest. Fitting separate quadratics requires the inclusion of the remaining terms; but we do not try to interpret the lower order terms.

We need to examine model fitness, which we do by looking at the diagnostic plots.

```
plot(glm2)
```

A better plot can be found in the car package. We are concerned about residuals that have values more extreme than -2 or 2, which we seem to have.

```
require(car)
residualPlots(glm2)
```

As a rough goodness-of-fit test, we compare the residual deviance of the reduced model to a chi-square distribution with 41 degrees-of-freedom. It is highly significant, demonstrating that the model does not fit the observed data well.

```
pchisq(glm2$deviance, glm2$df.residual, lower.tail=F)
```

There is evidence of overdispersion (variance is much greater than the mean), so we will refit the model with quasipoisson.

```
glm2$deviance/glm2$df.residual

glm3<- glm(Salamanders ~ PctCover + I(PctCover^2) +
       factor(Closed) + PctCover:factor(Closed) +
```

```
 I(PctCover^2)*factor(Closed), data = saldat,
family=quasipoisson)
```

It is a bit difficult to understand the meaning of these parameters, so for pedagogical sake, let's look at a simpler model.

```
glm4 <- glm(Salamanders ~ PctCover, data = saldat,
      family=quasipoisson)
```

In this model, we would interpret *Percent Cover* as

```
exp(glm4$coef[2])
exp(confint(glm4))
```

which suggests that the mean salamander abundance increases by 3.3% with every unit change in *Percent Cover*. Or, we are 95% confident that it increases between 1.8 and 5.5 percent with each unit change. The interpretation of the results would not change if we used quasipoisson instead of a Poisson distribution.

Your assignment

Your assignment is to conduct two different analyses (see descriptions below).

For each regression, write a 1-page description of your analysis, results, and inference. Each write-up should include the following information:

- Null and alternative hypotheses of your tests
- A description of how you checked the assumptions of your statistical test
- Results of your statistical test, interpreting your test in 2-3 sentences that include the appropriate reporting of the statistics
- A figure that demonstrates the results of your test/model

Problem #1: Use the *frisk* data from above to model the number of stops by the NYPD by ethnicity and suspected crime. (Do not consider interactions, just main effects.) This time use population (`pop`) of each precinct, rather than the number of arrests, as the offset to model the stop rate. What is the dispersion for your model? What is your final model? What are your conclusions? Interpret the model coefficients in 1-2 sentences.

Problem #2: Use the data from the package *Sleuth3*, `ex2226`, which shows characteristics of terrestrial planets, gas giants, and dwarf planets in our solar system, including the number of moons. Larger planets have more moons, but what drives the relationship? Is it the volume or mass that are relevant, or both? Answer this question and derive a model for describing the mean number of moons as a function of planet size. Make sure to interpret the model coefficients in a couple of sentences. Provide graphs for the number of moons over the range of volumes and masses of the solar bodies.