Lab 11: Generalized Linear Models

November 03, 2018

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 linearly to a set of terms based on predictor variables. Whereas linear regression was solved by ordinary least squares, GLM's are fit by a process called *iteratively reweighted least squares*, which overcomes the problem that transforming the data to make them linear also changes the variance.

In this lab, we will focus on Poisson regression, where the response variable consists of count data. Because count data must consist of positive integers, it is often assumed to have a Poisson distribution:

$$Y_i \sim Poisson(\lambda_i).$$

The Poisson distribution variable λ_i is modeled by

$$log(\lambda_i) = X_i\beta.$$

The mean response is related to the predictor variables through a 'log link'. And thus, the logarithmic expected number of incidents is modeled by the linear function of potential predictors. So, for example, $log(y_i) = \beta_0 + \beta_1 X_i$ or $y_i = exp^{(\beta_0+\beta_1 X_i)}$. This model is used to answer questions related to responses like (a) the number of cargo ships damaged by waves, (b) daily homicide counts in California, or (c) the environmental drivers of numbers of birds along an altitudinal gradient.

The learning goals of the lab are to:

- understand when to use Poisson regression;
- 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 a few problems to answer. *Submit your answers and your R-code to the class Sakai site under the Assignments folder before 11:55 pm on either Mon., Nov. 12 (Section 03) or Wed., Nov., 14 (Sections 01 and 02).*

More functions in R

- `glm()` - fits GLM's to data, with the user specifying 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 GLM's and carry out model comparisons. The function provides different test statistics. For models with known dispersion, the Chi-squared test or likelihood ratio Test (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 GLM's.
- `offset()` - including an offset allows modeling of a rate (e.g. number of birds per area), rather than just a count. For example, it allows the mean Poisson variable, μ, to be divided by effort, t: e.g. $log(\mu/t) = \beta_0 + \beta_1 X$ which is equivalent to $log(\mu) = \beta_0 + \beta_1 X + log(t)$.

- `deviance()` - returns the deviance of a fitted model object.

Poisson Regression

Stops by the NYPD

We will analyze data on the number of police stops of racial and ethnic minorities. 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. Let's rename the third column (*past.arrests*) to be *arrests* to make it shorter. Let's also take out one case where the number of arrests is 0.

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). Let's add *arrests.yr*, converting arrests into an annual figure instead of a 15-month figure.

```
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
```

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 `ggpairs()` and `summary()`. Let's also take a look at the number of stops by ethnic group, `s.eth`.

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

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 rate of stops, followed by Blacks and Whites.

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.

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. hours of counting birds or plot sizes for counting a rare plant).

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

Note that adding the offset changed the coefficient for Hispanics (*factor(eth)2*) to be positive relative to Blacks, 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 a reduction of 2 in deviance that would be expected if ethnicity had no explanatory power in the model.

To understand the coefficients, we exponentiate them and interpret them as multiplicative effects.

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

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)

 deviance(frisk2)-deviance(frisk3)
```

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 (74 is the increase in number of parameters in the model, and we would expect a decrease of 1 deviance for each added parameter).

Therefore, adding precinct has greatly improved the fit of the model to the data. In other words, accounting for precinct explains much of the variability in stops. It makes sense that certain precincts, perhaps lower income precincts, might have higher rates of stops than others. We can also compare the models using AIC.

```
AIC(frisk2, frisk3)
```

To Do

Check out the change in the coefficients of the ethnicities. How has adding `precinct` into the model changed their effects?

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. One sign of overdispersion is that the

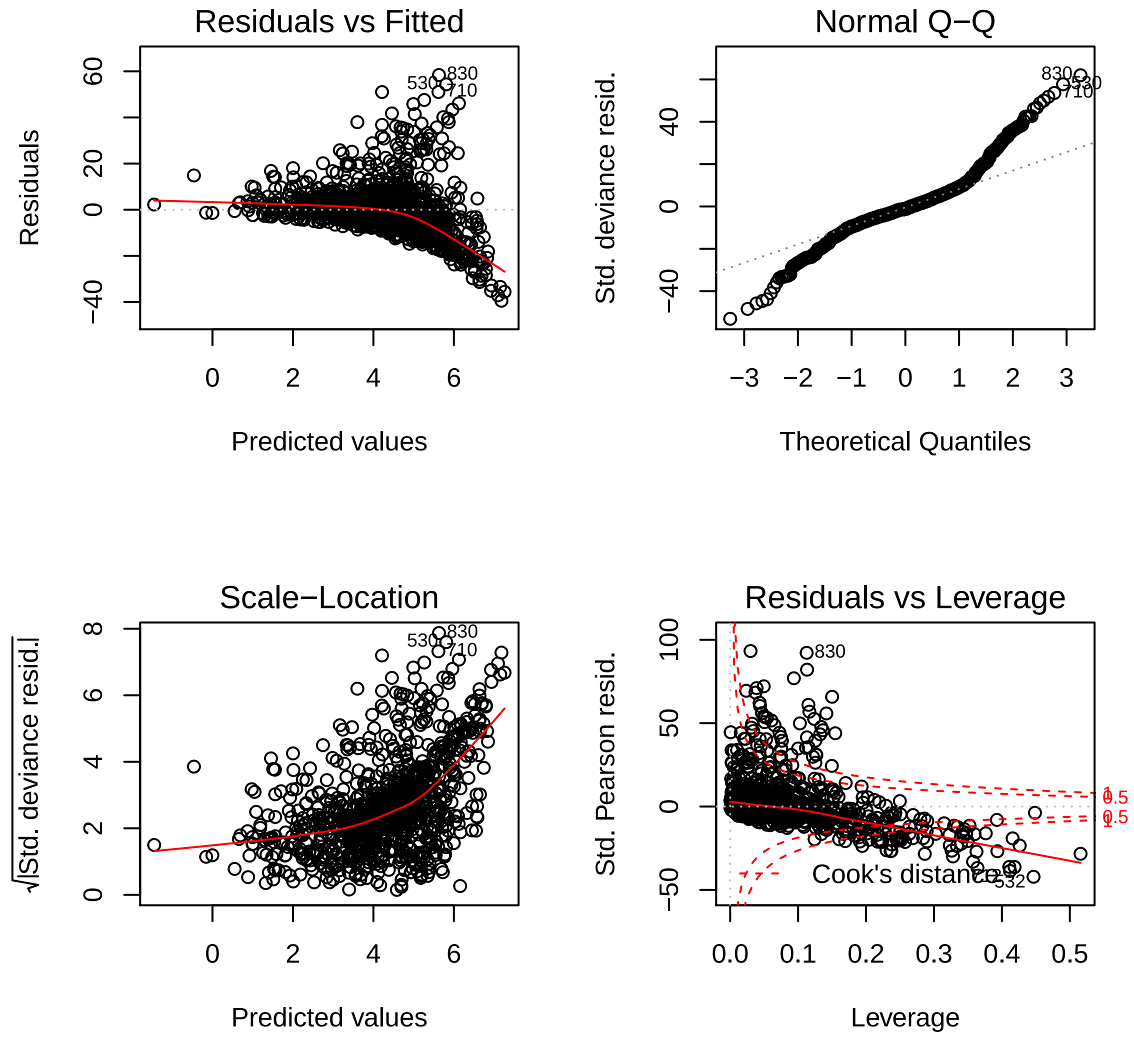

Figure 1: Diagnostic plots for the residuals of the `frisk3` model.

residual deviance is significantly higher than the residual degrees of freedom (which is true of `frisk3`). 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! For technical reasons that we won't explore in this class, overdispersion is calculated with the Pearson's χ^2 statistic:

$$\chi_p^2 = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{var(y_i)}$$

and

$$\hat{\phi} = \frac{\chi_p^2}{n-p}$$

We can calculate the overdispersion with the following function, `ovrdsp`, so that you just need to enter the response variable and model.

```
ovrdsp <- function(y, fit){
  phi <- sum((y-fit$fitted)^2/fit$fitted)/fit$df.residual
  cat("overdispersion ratio is", phi, "\n")
 }

ovrdsp(dat$stops, frisk3)
```

An overdispersion ratio of 2 is considered high, so this is out of the ballpark. To handle overdispersion, we can use the quasipoisson "distribution". This is not really a distribution. Rather, all the regression standard errors are rescaled through multiplication by the square root of the overdispersion $\sqrt{261} = 16.2$. Let's run the model.

```
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, reduces the statistics, and increases the *p*-values (making them more conservative). 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 how does forest age and percent cover affect numbers of salamanders?

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

Let's take a look at the number of salamanders in relation to the 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", xlab = "Forest Age")

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

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

  plot(PctCover, jitter(log(Salamanders+0.1)), las=1, pch=21,
       bg="grey", cex=1.2, ylab = "log(Salamander Count)", xlab = "Percent Forest Cover")
```

There is a rough break in percentage of canopy cover 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?

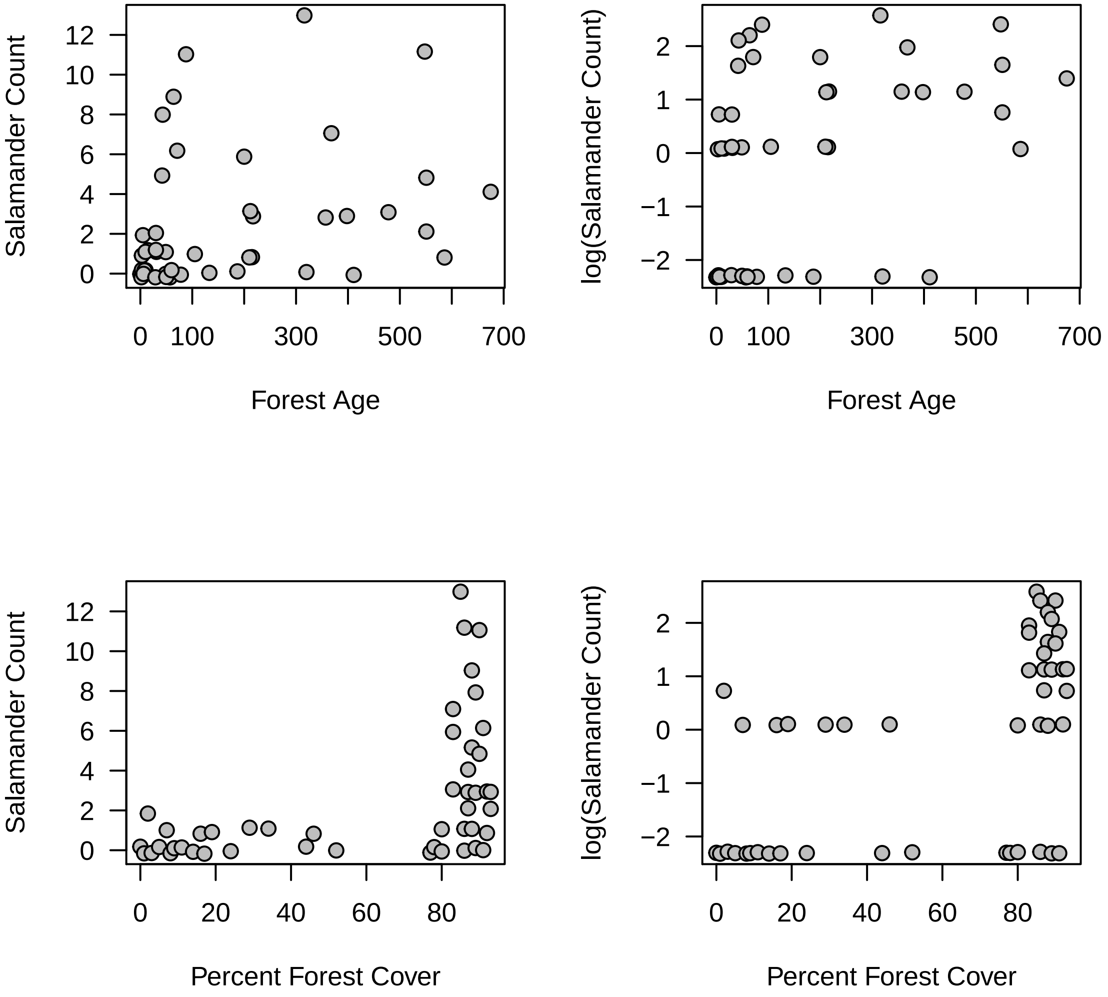

Figure 2: Plots of predictor variables, Forest Age and Percent Cover, versus the Salamander counts and the log-transformed Salamander counts.

```
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*. This is one of the reasons why it is recommended to *not* attach datasets in the workspace.

```
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*.

To Do

Write out the model in terms of coefficients, β_j, and variables, X.

```
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)
summary(glm2)
```

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 degrees of freedom equal to the difference in the number of parameters between the two models.

LRT can be presented as a difference in log-likelihoods, which can be expressed in terms of deviance: $LRT = -2ln(\mathcal{L}_S/\mathcal{L}_C) = -2(ln(\mathcal{L}_S) - ln(\mathcal{L}_C)) = -2ln(\mathcal{L}_S) + 2ln(\mathcal{L}_C) = deviance_S - deviance_C$.

In R, this is conducted using the `anova()` call. Note the use of the χ^2 statistic rather than the F-statistic used in multiple regression.

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

Alternatively, the `lmtest` package also includes a LRT function.

```
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 *Forest Cover* was not contributing significantly to the model. 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 by looking at the diagnostic plots.

```
par(mfrow=c(2,2))
 plot(glm2)
```

An alternative 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
ovrdsp(Salamanders, glm2)

glm3<- glm(Salamanders ~ PctCover + I(PctCover^2) +
             factor(Closed) + PctCover:factor(Closed) +
             I(PctCover^2)*factor(Closed), data = saldat,
           family=quasipoisson)
```

The best way to understand this model would be to graph fits of the model to the data for closed and open forests, and over different levels of percent cover. The coefficients by themselves are difficult to understand because there are interactions that include quadratic terms.

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 $e^{0.0324}$ 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% with each unit change. The interpretation of the results would not change if we used quasipoisson instead of a Poisson distribution.

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

Problems

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:

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

Problem 1

Use the aircraft data, `AircraftDat.csv`, found on Sakai to evaluate the factors that led to damage of attack aircraft (bombers) during the Vietnam War. The database contains data from 30 strike missions involving two types of aircraft, the A-4 and the A-6. The regressor, x_1, is an indicator variable (A-4=0 and A-6=1), and the other regressors x_2 and x_3 are bomb load (in tons) and total months of aircrew experience. The response variable is the number of locations where damage was inflicted on the aircraft

Please do the following:

1. Look at a pairs (or `ggpairs`) plot of the data, and describe what you can interpret from it about the distributions of the variables and their relationships.
2. Develop models for every combination of variables, *excluding interactions*. In other words, run models that include all the IV's, each pair of IV's, and the individual IV's. Compare each of the models to the full model using the LRT (e.g. `anova(mod1, mod2, test="LRT")`). What do these comparisons demonstrate?
3. Interpret your final model. What is the final model? What are the effects of any remaining IV's in terms of aircraft damage? Are the data overdispersed? What does the diagnostic plot indicate?
4. Create a plot of the final model, i.e. showing the effects of the remaining predictors in relation to aircraft damage.

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? Answer this question by deriving a model for describing the number of moons, *Moons*, as a function of planet size, *Mass*, and distance from the plant, *Distance*. Make sure to interpret the model coefficients in a couple of sentences. Provide graphs to demonstrate the relationship between the number of moons and the independent variables. Note: if there is more than one continuous independent variable (IV) left in the model, then graph the response variable against IV_1, keeping IV_2 at its mean value. Then, graph the next plot over IV_2, keeping IV_1 at its mean value.