Generalized Linear Models

Generalized Linear Models (GLMs) are a suite of models where the distribution of a response variable can be any member of the exponential family of distributions (EFD). EFD includes the normal, inverse normal, exponential, gamma, beta, binomial, multinomial and other distributions. See Schabenberger and Pierce (2002) and also McCullagh and Nelder (1989) for good descriptions of distributions in the exponential family and what processes they might represent.

GLMs are an extension of linear models that allows for:

- A nonlinear link function, which is a model that transforms the response variable (*y)*, in order to predict the transformed y with a linear function of predictors (*x*'s).
- Unequal variances of the *y*'s (variance of the *y*'s is a function of the mean of *y*, given *x*)
- Errors are uncorrelated for ease of calculation of the likelihood

NOTE: If errors are correlated, Generalized Estimating Equations (GEEs) can be used

Many widely used statistical models are generalized linear models, including:

- classic linear models with normal errors,
- logistic and probit models for binary data
- Poisson regression for count data
- log-linear models for multinomial data.

Many other statistical models can be formulated as generalized linear models by selecting an appropriate link function and response probability distribution.

For the OLS model:

$$y_i = \mathbf{x_i}' \boldsymbol{\beta} + \varepsilon_i$$
$$\varepsilon_i \text{ are iid and} \sim N(0, \sigma_\varepsilon{}^2)$$
$$\mu_i = \mu_y \mid x_i = \mathbf{x_i}' \boldsymbol{\beta}$$

NOTE: In this notation, the *x* values for a particular observation are given as a column vector of values.

For generalized linear models, this is extended to:

$$\eta = \mathbf{x_i}' \boldsymbol{\beta} = g(\mu_i)$$
$$y_i = g^{-1}(\mu_i) + \varepsilon_i$$
$$V(y_i) = \frac{\phi V(\mu_i)}{w_i}$$
$$y_i \sim \text{any of the family of exponential distribution}$$
$$y_i \text{ are independent (or can use GEE)}$$

- ϕ is a constant, called the scale parameter
- V is a variance function (variance as some function of the mean)
- w_i is a weight for each observation (each observation is given a weight=1 by default).

Distributions, R Default Link Functions, and Variance Functions, Using glm() **function**

Distribution **DIST=**	**Default R** **Link Function** **LINK=**	**Variance Functions**
Normal	Identity $g(\mu_i) = \mu_i$ (no transformation to the y)	$V(\mu_i) = 1$ (equal variance)
Binomial (proportion)	Logit $g(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right)$	$V(\mu_i) = \mu_i(1-\mu_i)$ (like p(1-p) for the binomial)
Poisson	$\ln(\mu_i) = X\beta$	$V(\mu_i) = \mu_i$ (variance is equal to the mean for Poisson)
Gamma	$g(\mu_i) = \mu_i^{-1}$	$V(\mu_i) = \mu_i^2$
Negative Binomial	Log $g(\mu_i) = \ln(\mu_i)$	$V(\mu_i) = \mu_i + k\mu_i^2$

- Other link functions that can be used include: Other powers, the logarithm and probit as an alternative to logit (see R documentation)

Commonly Used Generalized Linear Models

1. OLS model: y is continuous, LINK=identity, DIST=normal

2. Logistic Regression: y is a proportion (or a 0, 1 Bernoulli variable), LINK=logit, DIST=binomial

3. Poisson Regression, log linear model: y is a count (no natural denominator, else use y as a proportion), LINK=log, DIST=Poisson

4. Count using Negative Binomial: y is a count (no natural denominator, else use y as a proportion), LINK=log, DIST=negbin

5. Gamma Model with log linear model: y is a positive continuous variable, DIST=gamma, LINK=log.

Under/Overdispersion

If the default variance for the specified distribution does not match the data then the data are *over-* or *underdispersed.* This can happen with the binomial, Poisson and other distributions (e.g., negative binomial). An overdispersion factor can be added to the variance function and an estimate of this found by MLE along with the other parameters. Alternatively, another distribution may be more appropriate (e.g., switch to negative binomial for count data) or use quasi-likelihood (quasi Poisson).

Goodness of Fit Tests:

1. Asymptotic t-test for individual coefficients.
2. Likelihood ratio test for a full versus reduced model.

e.g. L0= likelihood with no variables, reduced model
L1= likelihood with some variables, full model

similar to F test for the MLR model, in this case. Test statistic for this is:

$$-2\log\left(L0 \big/ L1\right) = -2\log(L0) - (-2\log(L1))$$

- Compare this to a $\chi^2_{r,1-\alpha}$, where r is the difference in number of coefficients between L0 and L1 (number of constraints).
- Better models have higher likelihood (L) (or log likelihood, lnL), which is the same as saying the -2 lnL is smaller.

3. Pseudo R squared value, based on lnL of the model versus lnL for a "null model" with only the intercept (no explanatory variables). Several forms of these have been developed, in order to obtain a similar interpretation to R^2 for linear models.

4. Aikaike's Information Criterion (AIC). Smaller is better; it is based on $-2\log(L) + 2k$, where k is number of variables and response levels; gives a "penalty" for number of variables.

5. Schwarz Criterion (SC, BIC) - Smaller is better. Similar to AIC, but includes the number of response levels, the number of explanatory variables, and the frequency in each response level. The penalty term for including additional variables increases with increasing sample size.

Comparing Nested Models

- Use Deviance or Pearson's Chi Squared Statistic, for grouped or ungrouped data to compare unrestricted to restricted models, called, "Deviance partitioning", using a likelihood ratio test.
- This is like extra sums of squares used in partial F-tests for least squared fitted models.

Notes: Fitting Mortality Models Using Logistic Regression

Background:

In this session, two most commonly used GLMs, the logistic regression (binary response) and the Poisson regression (count response) are discussed. We introduce the logistic regression model using a mortality model example from forestry literature. Mortality is an essential component in predicting growth and yield of trees and in determining stand development patterns. Mortality varies by tree size, competition measures, site quality, and stand density.

Statistical Issues:

There are two main issues with mortality. 1) The response variable is a Bernoulli random variable, a tree is either live or dead. Because of the binary nature of the data, a binomial distribution which gives the probability of success (p, alive) versus failure (q=1-p, if tree died) is used.

The binomial distribution model can be expressed as:

$$\Pr(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}, y = 0,1,2,\ldots,n$$

$$E(Y) = np; \ \mathrm{var}(Y) = np(1-p)$$

We can now relate the underlying probability of success (p_i) to a set of linear predictor(s), $x_1, x_2, \ldots, x_k$, and make it more than one unknown constant. For example,

$$p_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}$$

2) A difficulty arises in that whereas p_i must lie between 0 and 1, the right hand side is unrestricted. The approach used to overcome this difficulty is to relate some function of p_i to the linear predictors. Such a function is referred to as a link function. For binomial variables this function is referred to as logistic regression.

$$\mathrm{logit}(p_i) = \log_e\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}$$

The probability of mortality is modeled as a function of Xs. The likelihood function for this model is proportional to:

$$L = \prod_{i=1}^{n} p_i^{\,y_i} (1-p_i)^{1-y_i}$$

$$L = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x_i}'\boldsymbol{\beta})}{(1+\exp(\mathbf{x_i}'\boldsymbol{\beta}))} \right)^{y_i} \left(1 - \frac{\exp(\mathbf{x_i}'\boldsymbol{\beta})}{(1+\exp(\mathbf{x_i}'\boldsymbol{\beta}))} \right)^{1-y_i}$$

R uses the logarithm of the likelihood (log likelihood):

$$\ln L = \sum_{i=1}^{n} [y_i \ln(p_i)] + [(1-y_i)\ln(1-p_i)]$$

For the logistic regression model it is not possible to express the estimates in simple closed form and it is necessary to use an iterative procedure. Robinson (2010) states that for logistic regression and other GLMs, *"Using the method of moments or least squares is extremely difficult, because the things that we previously relied upon no longer necessarily mean anything. The technique by which we maximize the likelihood is called Fisher scoring, and is iterative*."

Parameter estimation:

In R, GLMs are handled by the glm() function that provides computational algorithm for solving the MLE of model coefficients (β's). The MLE algorithm varies by the type of model family (e.g., EFD).

If OLS is used to fit these models:

- Resulting coefficients will be unbiased estimates of true coefficients
- The predicted values will not be constrained to be between 0 and 1.
- The variances of the coefficients will be biased estimates, since the variances of the errors will be unequal.

Fitting logistic model in R:

Mort_df1 <- glm(dead ~ dbh, family=binomial, data=DF)

The option family=binomial indicates that the data are binomial and a logit link is expected (unless otherwise specified, e.g., family=binomial(link="probit")).

Examine the plot diagnostics. Plots will have the same interpretation as do those from OLS.

par(las=1, mfrow=c(2,2), cex=0.75, mar=c(4,4,4,1))

plot(mort_df1)

summary(mort_df1)

```
Call:
glm(formula = dead ~ dbh, family = binomial, data = DF)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6706  -0.3380  -0.1480  -0.0644   3.3694

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.3525     0.1413   -9.57   <2e-16
dbh          -0.1000     0.0105   -9.51   <2e-16

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 892.69  on 2461  degrees of freedom
Residual deviance: 716.62  on 2460  degrees of freedom
AIC: 720.6

Number of Fisher Scoring iterations: 8
```

Notes:

1. **Residuals:** In GLMs, conventional residuals (observed - predicted) are not useful. GLMs have an equivalent called deviance residuals.

These are defined as

$$d_i = sign(y_i - \hat{y}_i)\sqrt{2y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + 2(n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - y_i}\right)}$$

The residual deviance, $D = \sum d_i^2$

The deviance residuals can also be standardized, by being divided by their estimated standard error.

When the model is true, the deviance residuals are approximately normally distributed. The quality of model fit is commonly assessed by deviance residuals.

There are numerous different kinds of residuals associated with GLMs. For example, The *Null deviance* is the deviance for a model with just a constant term, while *Residual deviance* is deviance of the fitted model (also the scaled deviance in case of a binomial model). These can be combined to give the proportion of deviance explained, a generalization of r^2, as:

$$r^2 = \frac{\text{Null deviance - Residual Deviance}}{\text{Null deviance}}$$

2. The parameters $\beta_0, \beta_1, \ldots, \beta_k$ are (usually) estimated using the method of maximum likelihood. The coefficients table is the same as we have seen in OLS, but we now use z as a basis for comparison instead of t. A Wald test based on the asymptotic normality of maximum likelihood estimators is used. Although the output is presented in the familiar form as in linear regression model, the linear relationship is between the logit of probability of mortality and dbh.

```
> summary(mort_df2)
Call:
glm(formula = dead ~ dbh + bal + ba + ccfl, family = binomial,
    data = DF)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4443  -0.3063  -0.1279  -0.0443   3.5411

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept) -2.62915    0.40394   -6.51  7.6e-11
dbh         -0.14084    0.01350  -10.43  < 2e-16
bal         -0.04014    0.01356   -2.96   0.0031
ba           0.05948    0.00955    6.23  4.6e-10
ccfl         0.00277    0.00187    1.48   0.1393

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 892.69  on 2461  degrees of freedom
Residual deviance: 672.08  on 2457  degrees of freedom
AIC: 682.1

Number of Fisher Scoring iterations: 8
```

3. Model interpretation:

The coefficients and intercept estimates give the following equation:

*logit(Pr (y=0))= 1.3525-0.10*dbh*

Intercept: In a linear model the intercept is the estimated mean response variable when the predictor is 0. In a logistic regression model, the estimated coefficient for intercept is the logit (log odds) of a tree with a dbh of zero being alive. The odds of being alive tree when dbh is zero is exp(-1.3525) =0.259. In terms of probability (p= exp(-1.3525) / 1+exp(-1.3525) =0.205). We do not put much emphasis as the estimated intercept corresponds to the logit of being alive when dbh is at the hypothetical value of zero.

Slope: indicates for a one-unit increase in the dbh values, the expected change in logit (log odds) is -0.1. If we exponentiate the slope (exp(-0.1)), we have 0.905. Based on this we can say that for a one-unit increase in dbh, we expect to see about a 10% decrease in the odds of being alive.

We are interested in predicting the probability of tree mortality knowing that the relationship between the probability of mortality and dbh is not linear. For this model, the probability of tree mortality:

$$\Pr(\text{mort}) = 1 - \frac{e^{-1.3525-0.10\times dbh}}{1+e^{-1.3525-0.10\times dbh}}$$

[Temesgen will add interpretation of logistic models with multiple predictor variables, using Model 2 as example]

4. Confidence intervals for odds ratios:

An approximate 95% confidence interval for the estimated parameters is given by:

$$\hat{\beta} \pm 1.96 \times se(\hat{\beta}) = -1.3525 \pm 1.96 \times 0.1413 = (-1.6295, -1.0755)$$
$$\hat{\beta}_1 \pm 1.96 \times se(\hat{\beta}) = -0.10 \pm 1.96 \times 0.0105 = (-0.121, -0.079)$$

Hence an approximate 95% confidence interval for the odds ratio is given by:

$$\exp(\hat{\beta} \pm 1.96 \times se(\hat{\beta}))$$

A confidence interval that does not include 1 implies that the effect of the variable is (statistically) significant (Robinson 2010).

5. Generalized likelihood ratio test or Change in (residual) deviance

The residual deviance can be used to compare two nested models.

Note: the order in which terms are included in the model statement will affect the order in which they are tested.

mort_df2<- glm(dead ~ dbh + bal + ba + ccfl , family=binomial, data=DF)

Nested models comparison:

> anova(mort_df2, test= "Chisq")

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: dead

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                  2461        893
dbh   1    176.1      2460        717    <2e-16
bal   1      0.2      2459        716      0.67
ba    1     42.3      2458        674     8e-11
ccfl  1      2.1      2457        672      0.15
```

6. Over (and under) dispersion

Sometimes the underlying assumptions of the binomial distribution are not satisfied.

Generalized Linear Models
Exercise: Logistic Regression to Fit Mortality Model

Objective: To develop mortality equations to predict the probability of mortality in south west Oregon forests (SWO) and conduct model diagnostics.

Data: swodata. The data set *swodata* contains status variable "dead" (1=live or 0=dead).

Read in the data using the R script found in 'GLM_SWO.r'. Check to make sure the data have 50 stands, 2699 tree records.

In this exercise, three runs will be obtained and results will be compared.

1. Use the lm() function and fit a simple linear model to predict probability of mortality using tree size (dbh) and competition measures as predictors.
2. Use the glm() function and fit a logistic regression model to predict probability of mortality using tree size (dbh) as predictor variables. The linear combinations for the predictor variables can be given as:

$$\beta X = \beta_0 + \beta_1 \times DBH$$

$$P_m = \left[1 - \frac{1}{1 + e^{[\beta_0 + \beta_1 \times DBH]}}\right] \qquad [5]$$

3. Use the glm() function and fit a logistic regression model to predict probability of mortality using tree size and competition measures as predictors. The linear combinations for the predictor variables can be given as:

$$\beta X = \beta_0 + \beta_1 \times DBH + \beta_2 \times BAL + \beta_3 \times BA + \beta_4 \times CCFL$$

$$P_m = \left[1 - \frac{1}{1 + e^{[\beta_0 + \beta_1 \times DBH + \beta_2 \times BAL + \beta_3 \times BA + \beta_4 \times CCFL]}}\right] \qquad [6]$$

Based on the results:

1. For the lm() function,
   - are the predicted probabilities realistic?
   - what pattern do you see in the residual plot?
2. Compare the results for Model 5 versus 6, in terms of:
   - -2lnL, AIC, other criteria for goodness of fit
3. What are the results of the two likelihood ratio tests (compared to the null model with no variables) for Model 5?

4. Extract AIC's and compare models 5 and 6. Which model provided the lowest AIC?
5. What are the results of the two likelihood ratio tests for Model 5 (only with dbh) and Model 6 (with competition measures) for predicting mortality of DF? Which model would you select?
6. For the model selected in 5, use Wald test and examine if all the coefficients (or terms) are significant?
7. Check the GLM model assumptions. What is your conclusion?

Notes: Poisson Models for Count Data

In forestry, many response variables (y) are count of something. Examples include:

- Number of trucks arriving to a weighting station in a given day,
- Number of employees in a forestry firm,
- Number of snags or cavity trees/ha, etc.

A count variable is often assumed to have a Poisson distribution, which has a probability distribution function parameterized by only one parameter.

$$p(y) = \frac{\lambda^y e^{-\lambda}}{y!}; y = 0,1,2,...$$

The parameter λ is both the mean and variance of y. Given the intensity parameter λ, the Poisson distribution calculates the probability of y.

$$p\left(y=2|\lambda=1\right) = \frac{1^2 e^{-1}}{2!} = 0.18$$

Count variables are discrete, non-negative, and sometimes highly skewed. These variables can rarely be transformed to normality. The Poisson distribution is a member of the exponential family with a link function η(μ)=log(μ). As a result, GLM is used.

$$Y_i \sim \text{Pois}(\lambda_i) \quad [8]$$

The Poisson distribution variable λi is modeled by $\log(\lambda_i) = X_i\beta$ [9]

That is the logarithm of expected number of counts or incidents is modeled by a linear function of potential predictors. Equation 8 defines the probabilistic assumption on the response variable and the form of the likelihood function. The likelihood function is a function of the regression coefficients defined in Equation 9. The estimated model coefficients are the ones that maximize the likelihood function.

Fitting the model in R:

In R, the maximum likelihood estimator is implemented using the glm function. Suppose we wish to model the number of snags/ha for stand *i* in terms of a vector of predictors x_i. Assume y_i has Poisson distribution with expected value of λ_i:

$$\log\left(\lambda_i\right) = \mu + \beta x_i + e_i$$

The logarithm of expected number of snags is modeled by a linear combination of selected predictors. The "log-linear" link ensures that that $\lambda_i \geq 0$, regardless of the predictors on the right-hand side.

When fitting a Poisson regression model, we almost use the same syntax as fitting a linear regression using the lm() function. The Poisson model is fitted using the *glm* function.

modp <- glm(nsnag ~ ba+tph+bal, family=poisson, data= swodata
summary(modp)

```
Call:
glm(formula = nsnag ~ ba + tph + ba, family = poisson, data = swo)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.008  -0.824   0.139   0.587   6.732

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.70e+00   2.71e-02    62.7   <2e-16
ba          1.13e-02   4.59e-04    24.6   <2e-16
tph         9.85e-05   2.23e-06    44.2   <2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6190.3  on 2698  degrees of freedom
Residual deviance: 4398.1  on 2696  degrees of freedom
AIC: 15973
```

We can judge the goodness of the fit of a proposed model by checking the deviance of the model against a X^2 distribution with degrees of freedom equal to that of the model.

The proportion of deviance explained by the above model is:

$$r^2 = \frac{6190.3 - 4398.1}{6190.3} = 0.29$$

Because the fitted value is the estimated mean of the Poisson distribution, residuals from a Poisson regression model should have a variance equal to the predicted mean. As a result when plotting residuals against the fitted values, we expect to see a wedge-shaped pattern that is the residual variance increases at the same rate as the predicted mean increases.

plot(modp)

Over dispersion:

The Poisson distribution's variance is the same as its mean (ie., Var(y) = E(Y)). Because the Poisson distribution has only one parameter, it is not very flexible for empirical fitting purposes. If we use a dispersion parameter, ϕ, such that Var(y) = ϕ E(Y) = ϕy, we will have more flexible distribution. When ϕ =1 is the regular Poisson regression. When ϕ >1 is over dispersion, ϕ <1 is under dispersion.

The dispersion parameter may be estimated using (after Faraway 2006)

$$\hat{\phi} = \frac{X^2}{n-2} = \frac{\sum_i (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i}{n-p}$$

We estimate the dispersion parameter in our example:

dp <- sum(residuals(modp, typ="pearson")^2)/modp$df.res
[1] 1.58

We can adjust the standard errors and so forth in the summary as follows.

summary(modp, dispersion=dp)
When comparing Poisson models with overdispersion, an F-test rather than a X^2 test should be used.

drop1(modp, test="F")

```
Single term deletions

Model:
nsnag ~ ba + tph + ba
       Df Deviance   AIC F value  Pr(F)
<none>        4398 15973
ba      1     5033 16607     389 <2e-16
tph     1     6163 17737    1082 <2e-16
Warning message:

In drop1.glm(modp, test = "F") : F test assumes 'quasipoisson' family
```

The z-statistics from the summary() are less reliable and so the F-test is preferred.

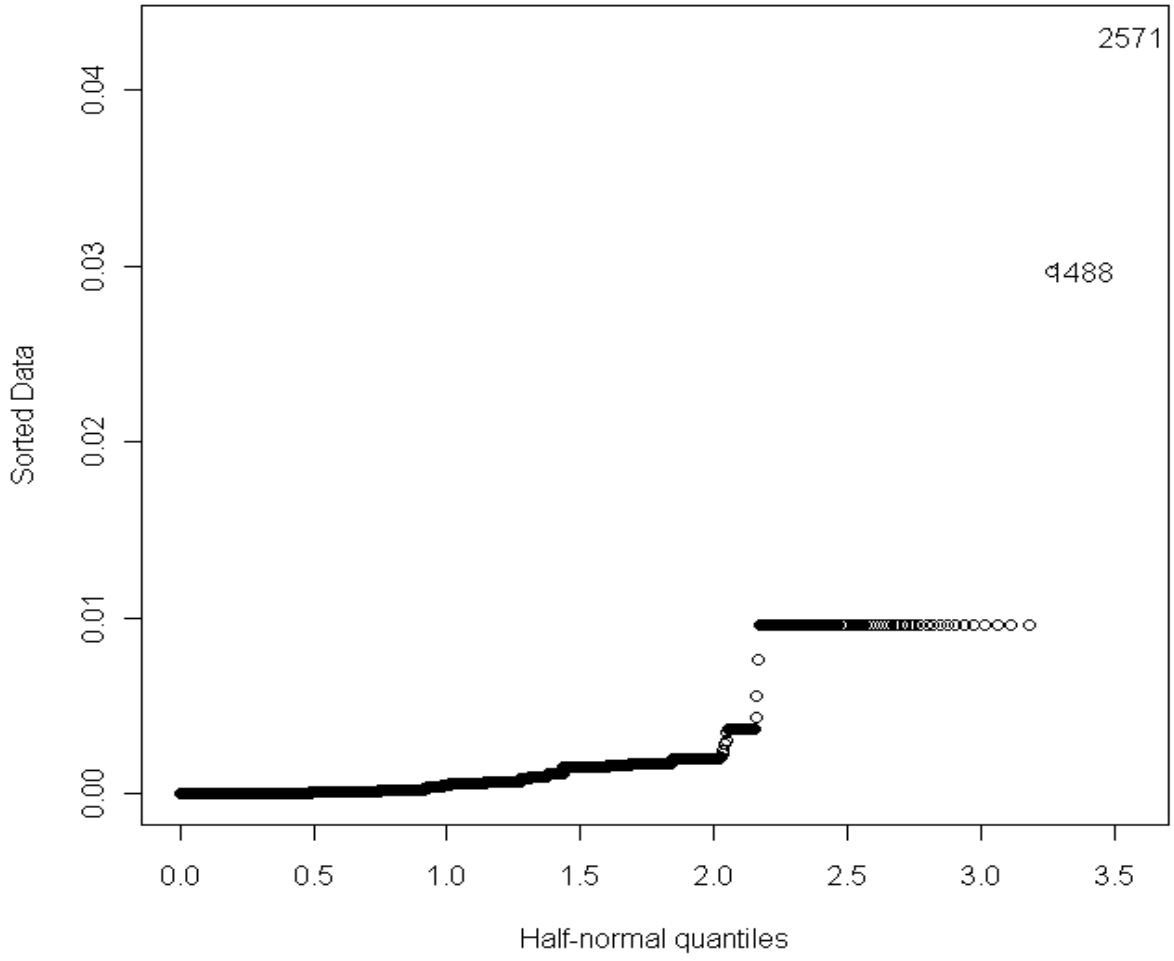
Model diagnostics:

A. Residual plots to examine nonlinearity
par(mfrow=c(1,1))
plot(residuals(modp) ~ predict(modp, type="response"), main="Nonlinearity Residual Plots",
xlab=expression(hat(mu)), ylab="Deviance residuals")

B. Residual plots to examine uniform variance
plot(residuals(modp) ~ predict(modp, type="link"), main="Uniform Variance Residual Plots",

xlab=expression(hat(eta)), ylab="Deviance residuals")

C. Graphical methods to detect outliers, examine leverage and influence using half-normal plots, which compares the sorted absolute residuals and the quantiles of the half-normal distribution.
halfnorm(cooks.distance(modp))



D. Investigate the nature of relationship between predictors and response variables.

Generalized Linear Model

Exercise: Poisson Regression

In this assignment you will again analyze the swodata set, using *nsnag* (number of snags/ha) as the response variable. Three models will be fit and results will be compared.

1. Use the lm() function and fit a multiple linear regression model to predict number of snags/ha (nsnag) using selected stand density measures as predictors.

2. Use the lm() function and fit a multiple linear regression model using the square root of the number of snags/ha (nsnag) as a response variable and selected stand density measures as predictor variables.

3. Use the glm() function and fit a Poisson regression to predict number of snags/ha (nsnag) using selected stand density measures as predictors.

Based on the results:

1. For the OLS,
   - are the predicted probabilities realistic?
   - what pattern do you see in the residual plot?
   - Did using the square root of the number of snags/ha (*nsnag*) as response variable improve the fit of the model?
2. Compare the results of the OLS and Poisson, in terms of -2lnL, AIC, and other criteria for goodness of fit.
3. Is the Poisson model an adequate fit? Is there over (or under) dispersion? What was the over (or under) dispersion parameter?
4. Does using dispersion parameter other than one have an effect on the regression parameter estimates?
5. Does using negative binomial distribution (instead of a Poisson distribution) to predict the number of snags/ha make any difference?

References:

Bates, D.M. and J.M. Chambers 1992. Statistical models in S, chapter 10 (Nonlinear models). Chapman and Hall, Boca Raton, FL.

Burnham, K.P. and D.R. Anderson. 2002. Model selection and multi-model inference. Springer, New York, 2nd edition.

Faraway, J.J. 2006. Extending the linear model with R. Generalized linear, mixed effects and nonparametrics regression models. Chapman & Hall/CRC. New York. 301 pp.

McCullagh, P. and J.A. Nelder. 1989. Generalized linear models, 2nd ed. Chapman & Hall, New York.

Robinson, A.P. 2010. icebreakeR. University of Melbourne Parkville, Vic. 3010. 187 pp.

Schabenberger, O. and F. J. Pierce. 2002. Contemporary statistical models for the plant and soil sciences. CRC Press. Boca Raton, Florida.