36-617: Applied Linear Models

Logistic Regression/GLMs: estimation & diagnostics

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Announcements…

- Take-home midterm due Wed 11:59pm
- No quiz today
  - Next Monday: class survey (for credit) instead of quiz
- HW05 will come out Weds, ***due Weds again***
- This week: More on logistic regression, GLMs
- Next week:
  - A brief introduction to nonparametric regression
  - Readings:
    - Sheather, Appendix on nonparametric smoothing
    - ISLR, Ch 7

Outline

- An example where AIC (and BIC) are not comparable across “model families”
- MLE’s and diagnostics for logistic regression
  - Finding MLE’s by Newton’s Method
  - Predicted values
  - Residuals
  - Goodness of Fit, Deviance Residuals
  - Hat Matrix, Standardized Residuals, Cook’s Distance
- Interpreting R’s casewise diagnostic plots
- Alternative residuals: DHARMa

An example of non-comparability of AIC/BIC

- Logistic regression model says

$$p_i = P[y_i = 1|X_i] = \frac{e^{X_i\beta}}{1+e^{X_i\beta}}$$

- Suppose we have i=1..N binomial outcomes with n_i trials per outcome and y_i successes per outcome. We can write the likelihood as

$$L_{bin}(\beta) = \prod_{i=1}^{N} \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i}$$

or

$$L_{ber}(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

where

$$n = \textstyle\sum_{i=1}^{N} n_i$$

$$L_{bin}(\beta) = \prod_{i=1}^{N} \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - p_i}$$ vs. $$L_{ber}(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-p_i}$$

- The log-likelihoods are

$$\ell_{ber}(\beta) = \log L_{ber}(\beta) = \sum_{i=1}^{N} \left[y_i \log(p_i/(1-p_i)) + n_i \log(1-p_i) \right]$$

and

$$\ell_{bin}(\beta) = \ell_{ber}(\beta) + \sum_{i=1}^{N} \log \binom{n_i}{y_i}$$

so

$$AIC_{ber} = -2\ell_{ber}(\hat{\beta}) + 2(p+1) = AIC_{bin} + 2\sum_{i=1}^{N} \log \binom{n_i}{y_i}$$

and

$$BIC_{ber} = -2\ell_{ber}(\hat{\beta}) + (p+1)\log N = BIC_{bin} + 2\sum_{i=1}^{N} \log \binom{n_i}{y_i} - 2(p+1)\log n/N$$

Example…

```
> data <- read.table("MissAmericato2008.txt",header=T)
> # arranged so that each row represents yi = # of top 10
> # finalists from each state in the US, in ni = 9 years
> glm.1 <- glm(cbind(Top10,9-Top10) ~ . - abbreviation,
+           data=data, family=binomial)
> newdata <- data[1,]
> for (r in 1:dim(data)[1]) {
+    row <- data[r,]
+    row$Top10 <- 1
+    reps <- data$Top10[r]
+    if (reps>0) for (j in 1:reps) {
+       newdata <- rbind(newdata,row)
+    }
+    row$Top10 <- 0
+    if(reps<9) for (j in 1:(9-reps)) {
+       newdata <- rbind(newdata,row)
+    }
+ }
> newdata <- newdata[-1,]
> # arranged so that each state is represented by 10 rows,
> # with a 1 or 0 indicating “top 10 finalist” or not, from that
> # state, in each of 10 years.
> glm.2 <- glm(Top10 ~ . - abbreviation, data=newdata,
+          family=binomial)
```

```
> round(cbind(coef(summary(glm.1)),coef(summary(glm.2))),2)
> #              --------- glm.1 -------    --------- glm.2 --------
                Est    SE     z     p   Est    SE     z     p
(Intercept)   -7.52  2.53 -2.97 0.00 -7.52  2.53 -2.97  0.00
LogPopulation  0.60  0.18  3.36  0.00  0.60  0.18  3.36  0.00
LogContestants 1.37  0.41  3.32  0.00  1.37  0.41  3.32  0.00
LogTotalArea  -0.36  0.14 -2.64  0.01 -0.36  0.14 -2.64  0.01
Latitude      -0.06  0.03 -2.15  0.03 -0.06  0.03 -2.15  0.03
Longitude      0.01  0.01  0.67  0.51  0.01  0.01  0.67  0.51
> round(data.frame(AIC=c(glm.1=AIC(glm.1),glm.2=AIC(glm.2)),
+ BIC=c(glm.1=BIC(glm.1),glm.2=BIC(glm.2))),2)
          AIC    BIC
glm.1 144.57 156.16
glm.2 420.83 446.24
```

- glm.1 fit the data as a binomial logistic regr.
- glm.2 fit the data as a Bernoulli logistic regr.
- The fit to the data is the same, estimated betas, SE’s, etc. all the same; AIC & BIC “should” reflect this
- AIC & BIC show different values for glm.1 & glm.2, because “normalizing constants” different

MLE's & Diagnostics for Binomial Logistic Regression…

- Let

$$\begin{aligned} i &= 1 \ldots N, \text{ total \# of observations} \\ n_i &= (\text{number of trials})_i \\ p_i &= P[(\text{success})|X_i] \\ y_i &= (\text{number of successes})_i \\ \mu_i &= E[y_i|X_i] \end{aligned}$$

- For Bernoulli, $n_i=1$, $\mu_i=p_i$, and $y_i = 0, 1$
- For Binomial, $n_i>1$, $\mu_i=n_ip_i$ and $y_i=0,1,\ldots,n_i$

For Binomial logistic regression

- $p_i = P[(\text{success})|X_i] = \dfrac{e^{X_i\beta}}{1+e^{X_i\beta}} = p_i(\beta)$
- $\mu_i = E[y_i|X_i] = n_i p_i = \mu_i(\beta)$

and

$$\begin{aligned} L_{bin}(\beta) &= \prod_{i=1}^{N} \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i} \propto \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{n_i-y_i} \\ &= \prod_{i=1}^{N} \left(\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right)^{y_i} \left(\frac{1}{1+e^{X_i\beta}}\right)^{n_i-y_i} = \prod_{i=1}^{N} e^{y_i X_i\beta}\left(1+e^{X_i\beta}\right)^{-n_i} \end{aligned}$$

so

$$\begin{aligned} \ell_{bin}(\beta) &= \sum_{i=1}^{N} y_i X_i\beta - n_i \log\left(1+e^{X_i\beta}\right) + C_{[we\ don't\ care]} \\ &= \sum_{i=1}^{N} y_i \sum_{s=0}^{p} X_{is}\beta_s - n_i \log\left(1+e^{\sum_{s=0}^{p} X_{is}\beta_s}\right) + C_{[we\ don't\ care]} \end{aligned}$$

To Maximize, set the gradient $\ell'_{bin}(\beta)$ to zero and solve for β .

$$\ell'_{bin}(\beta) = (\frac{\partial \ell}{\partial \beta_0}, \ldots, \frac{\partial \ell}{\partial \beta_p})^T$$

where

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_r} &= \sum_{i=1}^{N} y_i X_{ir} - n_i \left(\frac{e^{X_i \beta}}{1 + e^{X_i \beta}} \right) X_{ir} \\ &= \sum_{i=1}^{N} (y_i - n_i p_i) X_{ir} = \sum_{i=1}^{N} (y_i - \mu_i) X_{ir} \end{aligned}$$

which we can write in matrix form as

$$\ell'(\beta) = X^T (y - \mu(\beta))$$

Checking the maximum…

At the value β at which

$$\ell'(\beta) = X^T(y - \mu(\beta)) = 0$$

we can verify that we have achieved a maximum, by checking that

$$\ell''(\beta) = \left[\frac{\partial^2 \ell}{\partial \beta_r \partial \beta_s}\right]_{r,s=0}^{p}$$

is negative definite. We can calculate that

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \beta_r \partial \beta_s} &= \frac{\partial}{\partial \beta_s}\left[\sum_{i=1}^{N}\left(y_i - n_i \frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right)\right] X_{ir} \\ &= -\sum_{i=1}^{N} n_i X_{is} \frac{e^{X_i\beta}}{(1+e^{X_i\beta})^2} X_{ir} \;=\; -\sum_{i=1}^{N} n_i X_{is} p_i(1-p_i) X_{ir} \\ \text{so} \qquad \ell''(\beta) &= -X^T D X, \quad \text{where } D = \mathrm{diag}(n_i p_i(1-p_i), i = 1\ldots, N), \end{aligned}$$

Finding MLE’s by Newton’s method

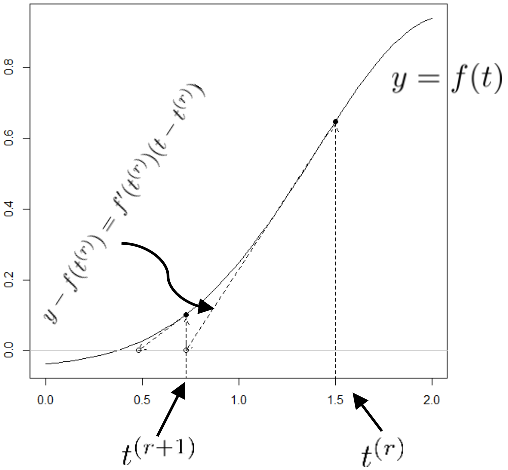
- $\ell'(\beta) = X^T(y - \mu(\beta)) \equiv 0$ seldom has a closed-form solution, so instead we want an iterative approach…

- Newton’s method for a single variable:

  Solve $f(t) = 0$ *iteratively:*

  - *Find tangent line at* $t^{(r)}$: $y - f(t^{(r)}) = f'(t^{(r)})(t - t^{(r)})$
  - *Find* $t^{(r+1)}$ *by solving* $y = 0$ *for* t:

$$t^{(r+1)} = t^{(r)} - \frac{f(t^{(r)})}{f'(t^{(r)})}$$



- We apply the multivariate version to $\ell'(\beta)$ and get

$$\begin{aligned} \beta^{(r+1)} &= \beta^{(r)} - [\ell''(\beta^{(r)})]^{-1}\ell'(\beta^{(r)}) \\ &= \beta^{(r)} + (X^T D(\beta^{(r)}) X)^{-1} X^T (y - \mu^{(r)}) \\ \text{where} \quad D(\beta^{(r)}) &= \text{diag}(n_i p_i^{(r)}(1 - p_i^{(r)}),\ i = 1, \ldots, N) \\ \text{and} \quad \mu^{(r)} &= n_i p_i^{(r)} = n_i \frac{e^{X_i \beta^{(r)}}}{(1 + e^{X_i \beta^{(r)}})} \end{aligned}$$

Summary (so far…)

- The log-likelihood for Binomial[1] logistic regression is (proportional to)

$$\ell(\beta) = \sum_{i=1}^{N} y_i X_i \beta - n_i \log\left(1 + e^{X_i \beta}\right)$$

- We can find the MLE's $\hat{\beta}$ by solving

$$\ell'(\hat{\beta}) = X^T(y - \mu(\hat{\beta})) \equiv 0$$

by iteration[2]

$$\beta^{(r+1)} = \beta^{(r)} + (X^T D(\beta^{(r)}) X)^{-1} X^T (y - \mu(\beta^{(r)}))$$

$$\begin{aligned} \text{where} \quad D(\beta^{(r)}) &= \text{diag}(n_i p_i^{(r)}(1 - p_i^{(r)}),\ i = 1, \ldots, N) \\ \text{and} \quad \mu(\beta^{(r)}) &= n_i p_i^{(r)} = n_i \frac{e^{X_i \beta^{(r)}}}{(1 + e^{X_i \beta^{(r)}})} \end{aligned}$$

- $SE(\hat{\beta})$ given[3] by the square roots of the diagonal elements of $[-\ell''(\hat{\beta})]^{-1} = (X^T D(\hat{\beta}) X)^{-1}$

---

[1] Same for Bernoulli, but with *N=n* and $n_i=1$

[2] With more work, can convert this to a *weighted least squares* calculation

[3] By std MLE theory (CLT for MLE's)

Predicted values for logistic regression

- Predicted or fitted values that may be useful:
  - raw: `fitted(glm.1)*glm.1$prior.weights`

    $$\hat{y}_i = n_i\hat{p}_i = n_i \frac{e^{X_i\hat{\beta}}}{1+e^{X_i\hat{\beta}}} = \mu_i(\hat{\beta})$$
  - response: `fitted(glm.1)` or `predict(glm.1,type="response")`

    $$\hat{p}_i = \frac{e^{X_i\hat{\beta}}}{1+e^{X_i\hat{\beta}}}$$
  - link: `predict(glm.1)` or `predict(glm.1,type="link")`

    $$X_i\hat{\beta} = \log \frac{\hat{p}_i}{1-\hat{p}_i}$$

  (there is another prediction type, `predict(…, type="terms")`, that is not very useful for us…)

- You can add SE's by adding `se.fit=TRUE` to the `predict()` arguments…

Logistic Regression Residuals[1]

- Residuals that correspond to fitted values:
  - Raw residuals: `resid(glm.1,type="response")` `*glm.1$prior.weights`

$$r_{raw,i} = y_i - \hat{y}_i = y_i - n_i\hat{p}_i = y_i - \mu_i(\hat{\beta})$$

  - Response residuals: `resid(glm.1,type="response")`

$$r_{resp,i} = y_i/n_i - \hat{p}_i = y_i/n_i - \frac{e^{X_i\hat{\beta}}}{1+e^{X_i\hat{\beta}}}$$

  - Pearson residuals: `resid(glm.1,type="pearson")`

$$r_{pearson,i} = \frac{y_i - \mu_i(\hat{\beta})}{\widehat{SE}(\hat{y}_i)} = \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}}$$

- Note that the Pearson residuals can be written in vector form as

$$r_{pearson} = D(\hat{\beta})^{-1/2}[y - \mu(\hat{\beta})]$$

---

[1] "Deviance" residuals are on the next page…

Logistic Regression Goodness of Fit

- Pearson Chi-squared statistic

$$P(X) = \sum_{i=1}^{N} \left(\frac{y_i - \hat{y}_i}{SE(\hat{y}_i)} \right)^2 = \sum_{i=1}^{N} \left(\frac{y_i - n_i \hat{p}_i}{\sqrt{n_i \hat{p}_i (1 - \hat{p}_i)}} \right)^2 = \sum_{i=1}^{N} r^2_{pearson,i}$$

- (Residual) Deviance statistic

$$\begin{aligned}
D(X) &= -2[\ell(\hat{\beta}) - \log L(\text{“saturated model”})] = 2[\log L(\text{“saturated model”}) - \ell(\hat{\beta})] \\
&= 2 \left[\log \prod_{i=1}^{N} \binom{n_i}{y_i} \left(\frac{y_i}{n_i} \right)^{y_i} \left(1 - \left(\frac{y_i}{n_i} \right) \right)^{n_i - y_i} - \log \prod_{i=1}^{N} \binom{n_i}{y_i} \hat{p}_i^{y_i} (1 - \hat{p}_i)^{n_i - y_i} \right] \\
&= 2 \left[\log \textstyle\prod_{i=1}^{N} y_i^{y_i} (n_i - y_i)^{n_i - y_i} - \log \prod_{i=1}^{N} (n_i \hat{p}_i)^{y_i} (n_i - n_i \hat{p}_i)^{n_i - y_i} \right] \\
&= 2 \sum_{i=1}^{N} y_i \log \left(\frac{y_i}{\hat{y}_i} \right) + (n_i - y_i) \log \left(\frac{n_i - y_i}{n_i - \hat{y}_i} \right), \qquad \hat{y}_i = n_i \hat{p}_i = \mu_i(\hat{\beta}) \\
&= \sum_{i=1}^{N} r^2_{deviance,i}, \qquad r_{deviance,i} = \mathrm{sgn}(r_{raw,i}) \sqrt{y_i \log \left(\frac{y_i}{\hat{y}_i} \right) + (n_i - y_i) \log \left(\frac{n_i - y_i}{n_i - \hat{y}_i} \right)}
\end{aligned}$$

- Both are $\approx \chi^2_{n-p-1}$ when the model is correct (small is good)
  - *D(X)* a little better than *P(X)*; need $n_i > 5$ or so for either to be trustworthy
  - $r_{deviance,i}$ tends to follow normal distribution better than other residuals

Logistic Regression Hat Matrix

- For logistic regression, define

$$H = D(\hat{\beta})^{1/2} X (X^T D(\hat{\beta}) X)^{-1} X^T D(\hat{\beta})^{1/2}$$

- This mostly works like a hat matrix
  - Sadly $Hy \neq \hat{y}$, but no matrix can satisfy this since logistic regression doesn't produce a linear fit[1]
  - But H still acts like a projection matrix:

$$\begin{aligned} H r_{pearson} &= D^{1/2} X (X^T D X)^{-1} X^T D^{1/2} [D^{-1/2} (y - \mu(\hat{\beta}))] \\ &= \text{(stuff)} X^T (y - \mu(\hat{\beta})) \;=\; \text{(stuff)} \ell'(\hat{\beta}) = 0 \\ H^T &= H \quad \text{(symmetric)} \\ H^2 &= H \quad \text{(idempotent)} \\ H D^{1/2} X &= D^{1/2} X \end{aligned}$$

---

[1] This projection property actually does work in the guts of weighted least squares, but we will skip this for now.

Leverage and "standardized" resids

- h_{ii} = (ith diag element of H) is again a measure of leverage
  - $\sum_{i=1}^{N} h_{ii} = p+1; \quad 0 \leq h_{ii} \leq 1$
  - $h_{ii} > 2(p+1)/n$ is a common rule of thumb for "high leverage"
- Again use $(1-h_{ii})$ to correct for under-estimated standard errors:
  - Standardized Pearson Residuals: $s_{pearson,i} = \frac{r_{pearson_i}}{\sqrt{1-h_{ii}}}$

    `resid(glm.1, type="pearson")/sqrt(1 - hatvalules(object)))`
  - Deviance Residuals: $r_{deviance,i} = \text{sgn}(r_{raw,i})\sqrt{y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)}$

    `resid(glm.1, type="deviance")` or just `resid(glm.1)`
  - Standardized Deviance Residuals: $s_{deviance,i} = \frac{r_{deviance_i}}{\sqrt{1-h_{ii}}}$

    `rstandard(glm.1)`

Cook's Distance

- For ordinary regression, Cook's Distance was

$$D_i = \frac{r_{standardized,i}}{p+1} \cdot \frac{h_{ii}}{1-h_{ii}} = \frac{y_i - \hat{y}_i}{(p+1)\hat{\sigma}} \cdot \frac{h_{ii}}{(1-h_{ii})^2}$$

- We can imitate this for logistic regression

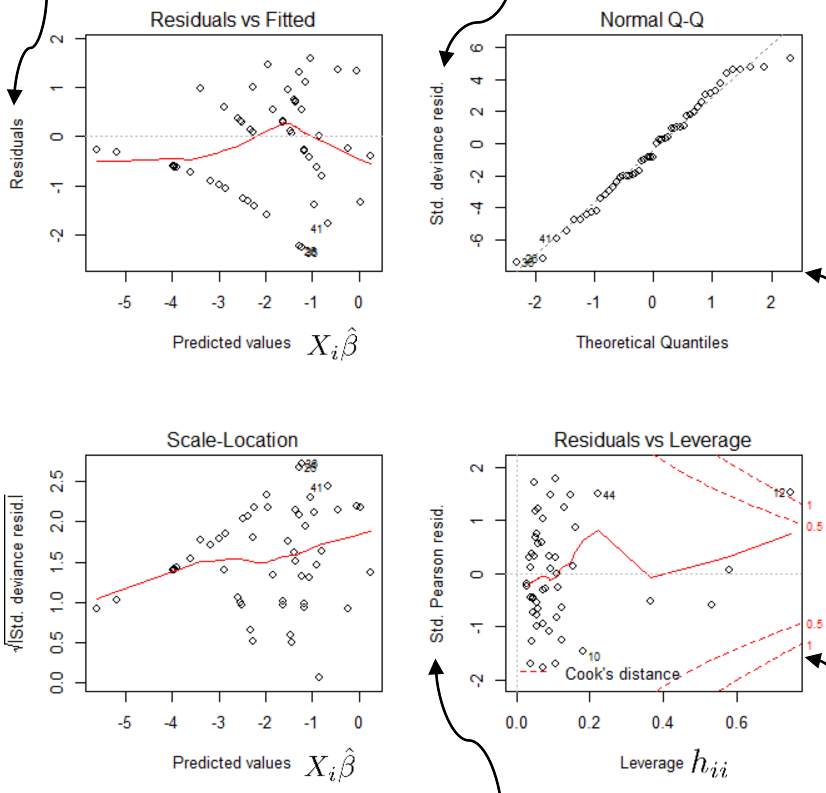
$$D_i = \frac{s_{pearson,i}}{p+1} \cdot \frac{h_{ii}}{1-h_{ii}} = \frac{y_i - n_i\hat{p}_i}{(p+1)\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}} \cdot \frac{h_{ii}}{(1-h_{ii})^2}$$

- Again, D_i gives us a measure of both
  - "outlierness" (how large is $y_i - n_i p_i$?)
  - "leverage" (how large is h_{ii} ?)

Interpreting casewise diagnostic plots for logistic regression

$r_{raw,i} = y_i - n_i p_i = y_i - \hat{y}_i$

$s_{deviance,i} = \text{sgn}(r_{raw,i})\sqrt{y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i-y_i}{n_i-\hat{y}_i}\right)} \Big/ \sqrt{1-h_{ii}}$



$s_{pearson,i} = \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}} \Big/ \sqrt{1-h_{ii}}$

- Raw residual plot nearly useless
    - Except for detecting extreme under- or over-prediction
    - Binned raw residuals somewhat useful
- Normal QQ plot useful
    - $s_{deviance,i}$ nearly normal when model holds
- Scale-location plot mostly useless
    - Since $Var(y_i) = n_i p_i(1-p_i)$ depends on location, always expect patterns here
- Plot of h_{ii}, D_i, $s_{pearson}$ quite useful
    - Less useful for Bernoulli logistic regression than for Binomial

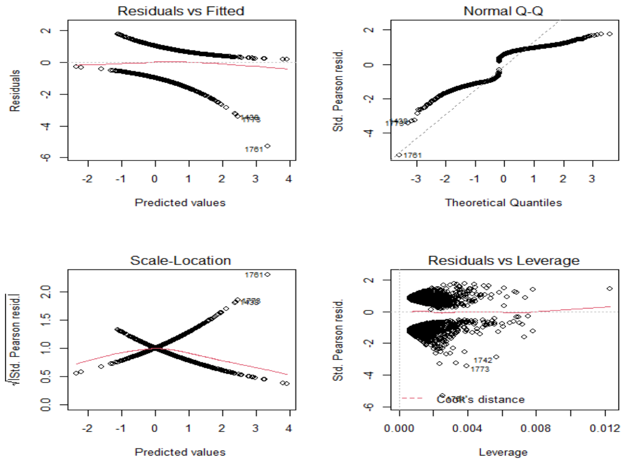
Alternative residuals for glm's

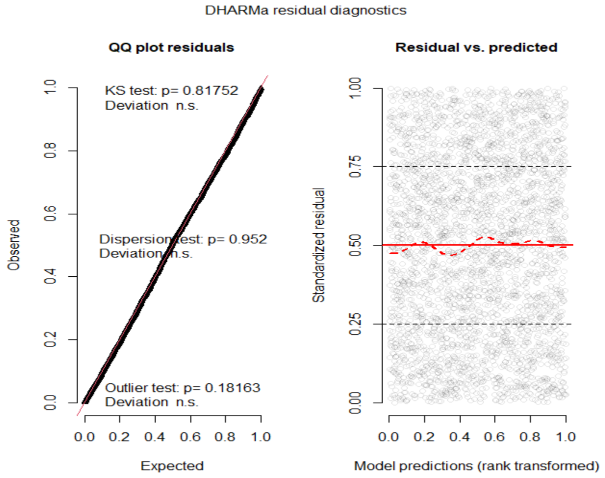
- Insight: If a continuous r.v. X has CDF
$$F_X(x) = P[X \leq x]$$
then $Y = F_X(X) \sim Unif(0,1)$, a *uniform distribution* (ex!)
- Approach ("parametric bootstrap"):
  - ☐ Fit the logistic regression model
  - ☐ Simulate many batches (say, 250) of new data from the fitted model
  - ☐ Use the simulated data to estimate $F_R(r)$ for the residuals $r_i, i = 1, \dots, n$
  - ☐ If the fitted model was "correct". then the transformed residuals will be uniformly distributed

See "DHARMa overview.pdf"
for some additional detail

Example: the wells data...

```
> library(DHARMa)

> data <- read.table("wells.dat",header=T)

> summary(glm.all <- glm(switch ~ .,
+ data=data,family=binomial))

             Est     SE      z  Pr(>|z|)
(Int)      -0.16   0.10  -1.57      0.12
arsenic     0.47   0.04  11.23      0.00
dist       -0.01   0.00  -8.57      0.00
assoc      -0.12   0.08  -1.61      0.11
educ        0.04   0.01   4.43      0.00

> par(mfrow=c(2,2))
> plot(glm.all)

> dev.new()

> simdata <- simulateResiduals(glm.all)
> plot(simdata)
```





A strategy for checking residuals of glm's

- Fit `mymodel <- glm(y ~ x1 + x2 + ..., data=mydata, family=binomial)` as usual
- Check the residuals vs leverage plot (lower right) from `plot(mymodel)` for high leverage or influential observations
- Use residual plot from DHARMa to check
  - Residuals Unif(0,1) (suggests model is good fit)?
  - Skewed left or right? Overdispersed? Underdispersed?
  - Outlying residuals?

Summary

- An example where AIC (and BIC) are not comparable across “model families”
- MLE’s and diagnostics for logistic regression
  - Finding MLE’s by Newton’s Method
  - Predicted values
  - Residuals
  - Goodness of Fit, Deviance Residuals
  - Hat Matrix, Standardized Residuals, Cook’s Distance
- Interpreting R’s casewise diagnostic plots
- Alternative residuals: DHARMa