GLM #2 (version 2)
Residuals and analysis of fit

Paul E. Johnson <pauljohn@ku.edu>

April 10, 2016

Contents

1 Terms

canonical exponential distribution

$$f(y_i|\theta_i,\phi) = exp\left\{\frac{y_i\theta_i - c(\theta_i)}{\phi} + h(y_i,\phi)\right\} \quad (1)$$

This is the probability of observing a particular outcome, y_i, given parameters θ_i and ϕ.

linear predictor $\eta_i = X_i b$

link function $\eta_i = g(\mu_i)$

inverse link function $\mu_i = g^{-1}(\eta_i)$

q function $\theta_i = q(\mu_i)$

Variance function $V(\mu_i)$ such that $Var(y_i) = \phi V(\mu_i)$

2 Sample R output

Here is an example of a generalized linear model fitted with R. This predicts the number of "wasted ballots" in Georgia counties as a function of the urban/rural classification of the county, the percent African American, and the type of voting equiptment that is used.

```
> myPois1<- glm(undercount ~ rural+perAA+equip, family=poisson, data=gavote)
> summary(myPois1)
```

```
Call:
glm(formula = undercount ~ rural + perAA + equip, family = poisson,
    data = gavote)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-77.623  -11.794   -2.844    8.385  165.005

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.536222   0.010373  437.32   <2e-16 ***
ruralurban   1.216001   0.007586  160.30   <2e-16 ***
perAA        2.213451   0.020352  108.76   <2e-16 ***
equipOS-CC   0.754020   0.010677   70.62   <2e-16 ***
equipOS-PC   0.937110   0.011196   83.70   <2e-16 ***
equipPAPER  -1.504971   0.094463  -15.93   <2e-16 ***
equipPUNCH   1.556716   0.010217  152.37   <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: 184237  on 158  degrees of freedom
Residual deviance:  77702  on 152  degrees of freedom
AIC: 78893

Number of Fisher Scoring iterations: 6
```

3 Deviance

Notice that output from glm models indicating "deviance"?

3.1 Saturated model

In the GLM literature, a benchmark for "good fitting" models is established by the so-called "saturated" model. The "saturated model" allows the author to choose a predicted value μ_i for each observation. In most cases, the saturated model will fit perfectly, and $-2lnLikelihood(saturated\,model) = 0$.

3.2 Deviance is a comparison of the saturated and fitted models

The deviance is defined as the gap between the saturated model and the final fitted model. You can write that down as

$$scaled\,model\,deviance = -2ln\left[\frac{L(fitted)}{L(saturated)}\right]$$

which is the same as the difference in $-2lnL$ of the two models.

$$scaled\,model\,deviance = 2lnL(saturated\,model) - 2lnL(fitted\,model)$$

This HAS THE APPEARANCE of a likelihood ratio test and one is tempted to treat it as a χ^2 approximation. But it isn't, as I labor to explain below.

3.3 Details on calculating scaled deviance, D_M^{scaled}

Suppose that ϕ is known. Observe, the log likelihood for the fitted GLM is

$$l(\hat{\theta}, \phi|y) = \sum_{i=1}^{N} \frac{y_i\hat{\theta}_i - c(\hat{\theta}_i)}{\phi} + h(y_i, \phi) \tag{2}$$

and the log likelihood for the saturated model is

$$l(\breve{\theta}, \phi|y) = \sum_{i=1}^{N} \frac{y_i\breve{\theta}_i - c(\breve{\theta}_i)}{\phi} + h(y_i, \phi) \tag{3}$$

Note that when we subtract one from the other in order to calculate $D_M^{scaled} = 2l(\hat{\theta}_i) - 2l(\breve{\theta}_i)$, the h terms cancel themselves out.

$$D_M^{scaled} = 2\sum_{i=1}^{N} \frac{y_i(\hat{\theta} - \breve{\theta}) - c(\hat{\theta}_i) + c(\breve{\theta}_i)}{\phi} \tag{4}$$

Its "scaled" in the sense that the dispersion parameter is used in the denominator.

3.4 About unscaled deviance

Some books which discuss deviance are not referring to the scaled deviance, rather a version that ignores ϕ. The unscaled Deviance depends only on the numerator in 6.

$$D_M = 2 \cdot \sum_{i=1}^{N} y_i(\hat{\theta} - \breve{\theta}) - c(\hat{\theta}_i) + c(\breve{\theta}_i) \tag{5}$$

and so the scaled deviance is just

$$D_M^{scaled} = \frac{D_M}{\phi} \tag{6}$$

Venables and Ripley observe, and I have to agree, that it is sometimes confusing that some authors use the same term, "residual deviance" for D_M and D_M^{scaled}.

In all fairness, however, there are many distributions for which $\phi = 1$. If you are working with one of those distributions, as was common in early GLM research, then deviance and scaled deviance are the same thing (in Dobson, for example, there is no dispersion parameter for this reason). The parameter $\phi = 1$ in the Poisson and Binomial distributions.

The output from PROC GENMOD in SAS, as illustrated in Myers, Montgomery, and Vining, presents both the deviance and scaled deviance. R presents just scaled deviance.

4 Hypothesis tests

The two most widely used hypothesis tests are the Wald Chi-Square statistic and the Likelihood Ratio Test. The Wald Chi-Square test has some known flaws that should cause people to be cautious about it in some models (especially clearly in a binomial/logistic model).

4.1 Likelihood ratio test

This is a test for 2 nested models. There is a "full fitted" model and a "small model" that omits some variables. The likelihood ratio test indicates if the small model's fit is significantly worse than the big one.

$$-2ln\left[\frac{L(small\,model)}{L(full\,model)}\right] \sim \chi^2_{f-s}$$

That's the same as the difference in -2 times the log of the likelihood of the 2 models.

$$-2lnL(small\,model) + 2lnL(full\,model)$$

The full model has f variables and the subset has s variables and so the test value is compared against a χ^2 with $f - s$ degrees of freedom.

4.1.1 Compare 2 deviances of nested models: you have a likelihood ratio test

R does not report the likelihood values or $-2lnL$, but it reports deviance. Suppose the larger model has deviance

$$deviance_{full} = 2lnL(saturated\,model) - 2lnL(full\,model)$$

and the deviance of the model with some parameters omitted is:

$$deviance_{small} = 2lnL(saturated\,model) - 2lnL(small\,model)$$

Subtract the two deviances, and you will see a log likelihood test pop out.

$$deviance_{small} - deviance_{full} = 2LnL(full\,model) - 2lnL(small\,model)$$

The $L(saturated)$ terms will cancel out. So if you look at the difference of 2 deviances, you are actually calculating the LLR.

4.1.2 Analysis of variance in R.

In R, one can ask for the so-called "model chisquare" with the commands anova and drop1 (or Anova from the car package) will provide an LRT for the individual variables.

```
> anova(myPois1,test="Chisq")
```

```
Analysis of Deviance Table

Model: poisson, link: log

Response: undercount

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    158     184237
rural  1    60475       157     123762 < 2.2e-16 ***
perAA  1    19667       156     104095 < 2.2e-16 ***
equip  4    26393       152      77702 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> drop1(myPois1, test="Chisq")
```

```
Single term deletions

Model:
undercount ~ rural + perAA + equip
       Df Deviance    AIC   LRT  Pr(>Chi)
<none>      77702  78893
rural   1   105070 106259 27368 < 2.2e-16 ***
perAA   1    89603  90792 11901 < 2.2e-16 ***
equip   4   104095 105278 26393 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> library(car)
> Anova(myPois1)
```

```
Analysis of Deviance Table (Type II tests)

Response: undercount
      LR Chisq Df Pr(>Chisq)
rural    27368  1  < 2.2e-16 ***
perAA    11901  1  < 2.2e-16 ***
equip    26393  4  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

One can also estimate the small model and use anova to compare the difference.

```
> myPois2<- glm(undercount ~ equip, family=poisson, data=gavote)
> summary(myPois2)
```

```
Call:
glm(formula = undercount ~ equip, family = poisson, data = gavote)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-67.028  -13.987   -4.957    7.116  205.451

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.437844   0.007666  709.34   <2e-16 ***
equipOS-CC   0.769648   0.010225   75.27   <2e-16 ***
equipOS-PC   1.226449   0.010805  113.50   <2e-16 ***
equipPAPER  -1.403604   0.094384  -14.87   <2e-16 ***
equipPUNCH   2.286368   0.009207  248.33   <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: 184237  on 158  degrees of freedom
Residual deviance: 112933  on 154  degrees of freedom
AIC: 114119

Number of Fisher Scoring iterations: 6
```

```
> anova(myPois2, myPois1,  test="Chisq")
```

```
Analysis of Deviance Table

Model 1: undercount ~ equip
Model 2: undercount ~ rural + perAA + equip
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       154     112933
2       152      77702  2    35231 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.1.3 Asymptotic χ^2

Comparison of this value from your models against the χ^2distribution is an asymptotic approximation. It would hold exactly if the sample size were infinite.

The term "model chi-square" is a colloquialism to refer to the likelihood ratio test that compares the "null" model–one that fits only the intercept, and the full fitted model. It is so commonly used that sometimes articles report it simply as $-2LLR$ and the reader is supposed to know that it is a particular likelihood ratio test.

4.1.4 Caution about χ^2 likelihood ratio tests when ϕ is estimated.

In a Poisson regression, the parameter ϕ is not estimated. But in a Gamma regression, it is.

If ϕ is estimated (through whatever consistent estimator one prefers), then the problem of interpreting the scaled deviance is raised. If the scaled deviance was not distributed as a χ^2 when ϕ is known, then how in the world could you proceed as though $D_M/\hat{\phi}$ is χ^2_{N-p}? See Venables and Ripley (p. 187).

As in OLS modeling, Venables and Ripley suggest that the significance of the impact of the omission of k parameters from a model that begins with p parameters is given by an approximate F test. Recall N is the sample size.

$$\frac{D_{restricted} - D_{full}}{\hat{\phi}(p-k)} \sim F_{p-k,N-k} \tag{7}$$

Of course, $(p-k)$ is the number of omitted variables in the restricted model. V&R caution about the use of this statistic, but I am a bit frustrated that they don't explain precisely how one should be cautious.

The anova and drop1 commands in R have an option to use an F test.

4.2 The Wald Test

4.2.1 Estimated Variance of $\hat{b}$

Supposing that $\phi_i = \phi$ for all observations, the weighted least squares estimate of b has variance-covariance matrix:

$$V(\hat{b}) = \phi^2 \cdot (X'WX)^{-1}$$

The matrix of weights is based on the estimates at the maximum likelihood estimate:

$$W = \begin{bmatrix} \frac{1}{\phi V(\mu_1)[g'(\mu_1)]^2} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\phi V(\mu_2)[g'(\mu_2)]^2} & 0 & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & \frac{1}{\phi V(\mu_{N-1})[g'(\mu_{N-1})]^2} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\phi V(\mu_N)[g'(\mu_N)]^2} \end{bmatrix} \tag{8}$$

Recall that $g'(\mu_i) = d\eta_i/d\mu_i$.

The different kinds of GLM have different values for ϕ, $V(\mu_i)$ and $g'(\mu_k)$. In a Normal model with an identity link, for example, the $V(\mu_i) = 1$ and $g'(\mu_i) = 1$,so the variance of the observations is all that remains, and if we assume that is fixed (as we do in OLS), then the W matrix plays no role in estimation whatsoever.

As usual, the standard errors $s.e.(b_k)$ are the square roots of the diagonal elements of $V(\hat{b})$.

4.2.2 Wald χ^2 test for several coefficients

Consider two null hypotheses: $H_0 : b_1 = K_1, b_2 = K_2$.

$$W = [b_1 - K_1, b_2 - K_2] \begin{bmatrix} Var(\hat{b}_1) & Cov(\hat{b}_1, \hat{b}_2) \\ Cov(\hat{b}_1, \hat{b}_2) & Var(\hat{b}_2) \end{bmatrix}^{-1} \begin{bmatrix} b_1 - K_1 \\ b_2 - K_2 \end{bmatrix}$$

As the sample size goes to infinite, this is distributed as a χ^2 statistic. The "critical value" and "p-values" can be calculated approximately for samples.

4.2.3 Wald Test for a single coefficient looks like a t^2

Null hypothesis: $H_o : b_k = K$

When only a single coefficient is being tested, the above reduces to a much simpler thing:

$$W = [b_k - K] \left[Var(\hat{b}_k)\right]^{-1} [b_k - K] = \frac{(\hat{b}_k - K)^2}{Var(\hat{b}_k)}$$

If you replace $Var(\hat{b}_k)$ by the squared standard error, $s.e.(\hat{b}_k)^2$, then the Wald statistic looks like a squared t test from an ordinary regression

$$W = \left(\frac{\hat{b}_k - K}{s.e.(\hat{b}_k)} \right)^2$$

If the sample size were infinite, then that value would be distributed as a χ^2 statistic.

$$\left(\frac{\hat{b}_k - K}{s.e.(\hat{b}_k)} \right)^2 \sim approximately\, \chi^2$$

The χ^2 is the square of the *Normal* distribution. Because of that fact, when we are considering a single variable, it is equivalent to take the square root and the result is treated as a standardized Normal variable.

$$\frac{\hat{b}_k - K}{s.e.(\hat{b}_k)} \sim approximately Normal(0,1)$$

However, you sometimes see it reported as a t statistic.

4.2.4 Is that a t or a z statistic?

On 2006-02-05, I learned something new in the r-help list! For the Poisson and Binomial models, the parameter ϕ is assumed to be 1.0. That means the standard error does not depend on any unknowns that are estimated. In such a situation, the glm's summary command reports the significance tests as z statistics. It does that because z is the right test for a model in which the standard error is **known.**

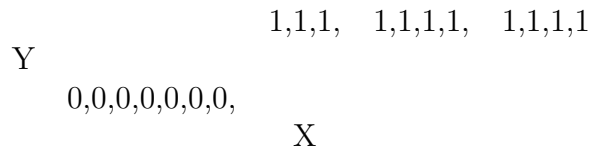
On the other hand, in models that estimate ϕ, then standard error is not known, but rather depends on estimate from the data. So the proper test is a t test. It is a slight wrinkle, and many programs get this wrong. Some always call the reported value a z or a t for all GLM.

Either way, we are still talking about an asymptotic approximation, however. And the asymptotic t is hardly different from the asymptotic z.

4.2.5 Warning about the Wald test: Hauck/Donner effect.

Sometimes, a binomial GLM will return estimates with a (deceptively subtle) warning about fitted probabilities being 0 or 1. That happens if you have "complete separation" or something close to it.

Complete separation is the problem in which the 0 and 1 dependent variable separates itself so that it appears the inputs perfectly predict outcomes. Suppose the data separates itself, as in this crude drawing:



Note, in the middle, the Y's are separated. The predicted probabilities ought to be 0 for the left set of data and 1 for the other.

Here the slope in the middle is infinite, cannot be estimated.

This can happen if you "dummy up" your variables so that a small population segment is represented by a category (or combination of categories) such as "Chinese-speaking Caucasian one-legged males who live in Dubuque, Iowa". Observations like that may have homogeneous Y's, all 0's or all 1's in your data. Then logistic regression breaks.

If you have truly complete separation or something close to it, then logit estimation fails. You probably should seriously rethink your data or your model. Other times, you just get a warning, and that's where judgment and prudence come into play.

This problem is related to a problem with test statitics in Logistic regression known as the Hauck-Donner effect. It gets surprisingly little treatment in the textbooks. I don't think I've ever seen it mentioned in an econometrics-style text. It is in *Modern Applied Statistics with S/R* by Venables and Ripley, but that treatment is somewhat brief. You will find this discussed in many statistically-oriented email lists, especially r-help, but also others. Here's an email post from Brian Ripley in the late 1990s that is more conversational:

http://www.math.yorku.ca/Who/Faculty/Monette/S-news/0049.html

I'd paraphrase the problem this way. Consider the ratio of the parameter estimate to the standard error as the "test statistic":

$$\frac{\hat{b}}{s.e.(\hat{b})}$$

Right now, I don't want to quibble right now about whether that's Normally or t distributed.

If $\hat{b}$ is huge, say nearly infinite, because of complete (or nearly complete separation) then the standard error is likely also to be massively huge, and the resulting test statistic can be small. Hauck and Donner showed that "Wald's test statistic decreased to zero as the distance between the parameter estimate and null value increases." Ironically, then, as your Null gets more and more wrong, the Wald stat gets smaller and smaller and you are less and able to reject a wrong Null.

Hence Ripley's point in the email cited above, which says that, paradoxically, if the value of $\hat{b}$ is either near zero or very far away from zero, the test statistic can be small.

The practical advice, then, is to run the model with all of the variables, and then run again with the questionable one removed, and conduct a likelihood ratio test. In the past, I had expected that such a test would reach the same conclusion as the Wald test. I think most people expect it will lead to the same conclusion. But Hauck & Donner show it is wrong to think that.

While browsing in Jstor for the Hauck-Donner article (1977) I found a few others you could also get. I kept all of these in a folder in case you want to see them.

"Wald's Test as Applied to Hypotheses in Logit Analysis," Walter W. Hauck, Jr.; Allan Donner *Journal of the American Statistical Association* Vol. 72, No. 360 (Dec., 1977), pp. 851-853

"A Reminder of the Fallibility of the Wald Statistic," Thomas R. Fears; Jacques Benichou; Mitchell H. Gail *The American Statistician* Vol. 50, No. 3 (Aug., 1996), pp. 226-227

"Understanding Wald's Test for Exponential Families," Nathan Mantel *The American Statistician* Vol. 41, No. 2 (May, 1987), pp. 147-148

"On the Use of Wald's Test in Exponential Families," Michael Vaeth *International Statistical Review / Revue Internationale de Statistique* Vol. 53, No. 2 (Aug., 1985), pp. 199-214

"Judging Inference Adequacy in Logistic Regression," Dennis E. Jennings *Journal of the American Statistical Association* Vol. 81, No. 394 (Jun., 1986), pp. 471-476

"A Note on Confidence Bands for the Logistic Response Curve," Walter W. Hauck *The American Statistician* Vol. 37, No. 2 (May, 1983), pp. 158-160

I found at least one political science article that cites Hauck-Donner:

"Issue Voting in Gubernatorial Elections: Abortion and Post-Webster Politics" Elizabeth Adell Cook; Ted G. Jelen; Clyde Wilcox *The Journal of Politics* Vol. 56, No. 1 (Feb., 1994), pp. 187-199

5 Residuals

We want a way to spot cases that don't fit a model. Residuals can do that, but there are many different kinds of residuals for GLMs.

If you were coming to this from an OLS perspective, you would expect that the residual would be $y_i - \hat{y}_i$. In the GLM framework, you are tempted to replace $\hat{y}_i$ with $\hat{\mu}_i$. However, that's not correct because $\hat{\mu}_i$ is a prediction about the mean, while y_i is an observed value. Think of a logit model, where the "predicted probability" is the $\hat{\mu}_i$. Is the residual equal to the difference between the observed 1 (or 0) and $\hat{\mu}_i$?

I don't think so.

Nobody intelligent does.

Several alternative residuals have been recommended. Notice that if you fit a model in R, and then want the residuals, you can specify 5 kinds of residuals.

```
help(residuals.glm)
```

Accessing Generalized Linear Model Fits

Description:

These functions are all 'methods' for class 'glm' or 'summary.glm' objects.

Usage:

S3 method for class 'glm': family(object, ...)

S3 method for class 'glm': residuals(object, type = c("deviance", "pearson", "working", "response", "partial"), ...)

Arguments:

object: an object of class 'glm', typically the result of a call to 'glm'.

type: the type of residuals which should be returned. The alternatives are: '"deviance"' (default), '"pearson"', '"working"', '"response"', and '"partial"'.

I'm focusing on deviance residuals and Pearson residuals

5.1 Unscaled Deviance Residuals

The (unscaled) deviance (5) is the source the deviance residual. That is so because the D_M can be written as a sum of N terms, one for each observation. Call an individual element in this sum D_i.

The *deviance residual* is defined as the square root of D_i with the special condition that its sign is the same as $(y_i - \hat{\mu}_i)$. If $(y_i - \hat{\mu}_i) > 0$, for example, we want the value to be $+\sqrt{D_i}$). Let's call this rd_i.

$$rd_i = [sign(y_i - \hat{\mu}_i)] \cdot \sqrt{D_i}$$

It is easy to see that one can compare the deviances of the observations in order to see which particular cases deviate from the fitted model.

5.2 Pearson's Residual

Standardize the residual according to the variance function, $V(\mu)$

$$rp_i = \frac{(y_i - \hat{\mu}_i)}{\sqrt{V(\hat{\mu}_i)}}$$

In my opinion, the Pearson residual is the most intuitive and logical residual. However, it is widely noted that rp_i is decidedly skewed in many GLM applications. Hence, it can be misleading to use this to identify outliers. It is widely recommended that, for diagnostic purposes–to spot outliers or influential observations–one ought to use the deviance residuals instead.

The Pearson residual is not completely useless, however, as we shall see in a minute.

6 Goodness of Fit

The goodness of fit is assessed by two different statistics, the deviance and the Pearson χ^2 statistic. Both of these are approximate for small samples, although for very large samples they are statistically equivalent. There appears to be a difference of opinion among authors about which ought to be used for small samples. Both agree neither is great.

6.1 What does Deviance mean for goodness of fit?

1. If the deviance is huge, then the model "doesn't fit very well." And if deviance is small, it "fits well."

2. Can we be more specific than that? Surprisingly, the answer is no! Frustratingly.

It is a widely used "rule of thumb" that a model does not fit too badly if $D_M^{scaled}$is not significantly greater than $N-p$. But people should be more cautious.

They often claim that the deviance is asymptotically distributed as a χ^2 variable.

$$D_M^{scaled} = \frac{D_M}{\phi} = -2ln\left(\frac{L(\hat{b})}{L(\check{b})}\right)/\phi \sim \chi^2_{N-p} \tag{9}$$

People act as though this is a likelihood ratio test, and consider deviance as an "approximate chi-square" variable. For an example, see Myers, Montgomery and Vining (2002, p. 113).

This χ^2 approximation is valid only for Normally distributed variables, however, and only if ϕ is known (not estimated from data).

This reasoning, while widely practiced, is technically wrong. David Firth's essay on GLM observes that it is not truly distributed as a χ^2 (p. 68). "These models are trivially nested, and it is tempting to conclude from sec. 3.5.1 that the deviance is distributed approximately as ϕ times a χ^2random variable if the fitted model holds. However, standard theory leading to the $\chi^2_{p_B-p_A}$ approximation for the null distribution of the log-likelihood ratio statistic is based on the limit as $n \to \infty$with p_A(clarification: the rank–number of rows of data–of model A) and p_B(the rank of model B) both fixed. If B is the saturated model, $p_B = n$ and so the standard theory does not apply. The deviance does not, in general, have an asymptotic χ^2 distribution in the limit as the number of observations increases; as a consequence, the distribution of the deviance may be far from χ^2, even if n is very large."

He goes on to argue, however, that there are cases in which the deviance is something like a χ^2 variable. "In situations where the information content of each observation is itself large, consideration of the limit as $n \to \infty$may be unnecessary. Such situations include Poisson models with large μ_i, binomial models with large m_i, and gamma models with small ϕ..."(p. 69) .

Venables and Ripley observe the same problem with the use of the "rule of thumb." They observe "sufficient (if not always necessary) conditions under which $\chi^2/\phi \sim \chi^2_{N-p}$ becomes approximately true are that the individual distributions for the components y_i should become closer to normal form and the link effectively closer to an identity link. The

approximation will *not* improve as the sample size N increases since the number of parameters under S also increases and the usual likelihood ratio approximation argument does not apply. Nevertheless, (it) may sometimes be a good approximation..." (p. 187).

6.2 Pearson's χ^2

As a measure of "badness of fit", the Pearson χ^2 is frequently recommended. For a categorical dependent variable, this indicator is quite reminiscent of the χ^2 statistic that is familiar from the analysis of crosstabulation tables. For details, consult Dobson (p. 125).

The Pearson χ^2statistic (a measure of "badness of fit") is calculated as a sum of squared Pearson residuals:

$$Pearson's\,\chi^2 = \sum_{i=1}^{N} rp_i^2$$

For large samples, Pearson's χ^2 is distributed as χ^2_{N-p}.

7 Overdispersion

Recall that $Var(y_i) = \phi \cdot V(\mu_i)$.

That is, the variance of observed y has 2 parts, one that is linked to the dispersion parameter and one that is linked to μ_i. Many of the "garden variety" GLMs (logistic regression, Poisson count) are designed with the assumption that $\phi = 1$. That is frequently violated in practice, hence there is over-dispersion.

Suppose you build a model around the premise that the distribution of y_i is Binomial or Poisson. In the structure of both of these distributions, there is a definite, mathematically clear, well known linkage between the observed variance and the value of the mean. The distribution-specific function $V(\mu_i)$ is the relationship between the value of the mean, μ_i, and the variance of observed y_i.

Poisson: $V(\mu_i) = \mu_i$

Binomial: $V(\mu_i) = n \cdot \mu_i(1-\mu_i)$ [where μ_i is the probability that a particular test succeeds and n is the number of tests conducted with the probability μ_i]

Models based on these distributions assume that the dispersion parameter $\phi = 1$ and the variance of y_i will follow $V(\mu_i)$.

If that is not true in practice, then one has "over dispersion" or "extra binomial variation" or such (see MM&V, p. 126).

As far as I know, there are 3 things that can be done about over-dispersion.

1. Apply a correction for the standard errors of the $b's$.

If over-dispersion is present, then the ML estimates of the b's are still asymptotically unbiased, but their variance does not match the variance/covariance matrix that is provided by computer programs. The "correction" for overdispersion is to multiply the $s.e.(\hat{b})$ by an estimate of ϕ (see Myers, Montgomery, and Vining, p. 128).

2. Change the model. One should revise the model to incorporate the right amount of variance. In the Negative Binomial model for count data, for example, one begins with a Poisson model and then incorporates a new source of randomness. This is what Gary King's Generalized Event Count approach is all about.

3. Adopt a quasi-likelihood framework. This is outside the scope of this handout, but, the gist is that one drops most of the details of the GLM and the exponential family, instead specifying only the mean and the variance of y_i. That means you can specify the right amount of variance for your model.

No matter which approach you take, it is necessary to estimate ϕ.

Digression on the Normally distributed dependent variable

You may wonder why overdispersion is not an issue in the Normal model. Well, it is, really. Notice for the normal

$V(\mu_i) = 1$

(That's a statement of homoskedasticity.) As a result, ϕ is responsible for *all of the variance* observed in a model that uses the Normal distribution. The familiar estimate from OLS, σ_e^2 , the *MSE* (Mean Square Error) is just an estimate of ϕ.

8 Estimating ϕ

There's a comment in Venables & Ripley (p. 185) that θ and ϕ are orthogonal. They observe

$$E\left[\frac{\partial l(\theta,\phi|y_i)}{\partial\theta\partial\phi}\right] = 0$$

That means that, generally speaking, the likelihood's response to changes in θ is not affected by changes in ϕ. Thus, ϕ and θ can be estimated separately.

And its a good thing. There's pretty much controversy over how to estimate ϕ, and if that tainted estimates of the b's, then the whole GLM exercise would be tedious and uncertain.

Roadmap

Now we'll compare 3 ways to estimate ϕ. All of these are "moment methods." That means we try to match up a summary statistic from the observations against a theoretically inspired quantity.

8.1 ML estimate of ϕ.

If you use the ML approach with the standard log likelihood, you arrive at:

$$\hat{\phi}_M = \frac{\sum_{i=1}^{N}(y_i - \hat{\mu}_i)^2}{N}$$

That is a biased estimate of ϕ.

The correction for bias is to adjust the denominator

$$\hat{\phi}_M = \frac{\sum_{i=1}^{N}(y_i - \hat{\mu}_i)^2}{N - p}$$

but that's not an ML estimator anymore. So you have to check the formula details when people talk about ML estimates.

I notice some SAS procedures use this by default, but the statistics books don't recommend it.

8.2 Scaled deviance estimate of ϕ

Venables & Ripley (p. 186) note that the scaled deviance is distributed as a χ^2 statistic with $N-p$ degrees of freedom.

$$\frac{D_M}{\phi} \sim \chi^2_{N-p}$$

The average value of a χ^2_{N-p} is $N-p$. According to V&R, the Deviance-based (they say customary) estimator for ϕ is

$$\hat{\phi} = \frac{D_M}{N-p} \tag{10}$$

8.3 ϕ Estimated as the average of Pearson residuals

Average of standardized rp_i , adjusting for degrees of freedom (after estimating p parameters, $df = N-p$).

$$\hat{\phi} = \frac{1}{N-p}\sum_{i=1}^{N}\frac{(y_i-\hat{\mu}_i)^2}{V(\hat{\mu}_i)} \tag{11}$$

8.4 What's the best way to estimate ϕ?

V&R prefer the estimator for ϕ that is based on Pearson's residuals. They say it has better small sample properties (this is especially important at least in the Binomial and Poisson cases).

> A common way to 'discover' over- or underdispersion is to notice that the residual deviance is appreciably different from the residual degrees of freedom... This can be seriously misleading. The theory is asymptotic. The estimate of ϕ used by summary.glm (if allowed to estimate the dispersion) is the (weighted) sum of the squared Pearson residuals divided by the residual degrees of freedom... This has much less bias than the other estimator sometimes proposed, namely the deviance (or sum of squared deviance residuals) divided by the residual degrees of freedom (V&R, MASS 4ed, p. 208).

They mention that in the MASS package's glm functions, the estimate of dispersion is based on the Pearson residuals.

Myers, Montgomery and Vining report on a simulation study of estimators of ϕ(p. 261). Their results are not completely decisive. The ML results are biased (as we already knew), but they are lower in variance. The Pearson and Deviance based estimates are similar, with a slight edge to the Pearson results.

9 Other Goodness of Fit indicators

Should you always choose the model with the highest log likelihood? Maybe not.

You should penalize your model for the number of fitted parameters, so something in the nature of an $adjusted R^2$ is needed. In that context, you will find people using various adjustments on the log likelihood. That is what the AIC and BIC are all about.

Akaike's Information Criterion (AIC). In R's AIC() documentation, the formula for AIC is given as

$$AIC = -2lnL + 2 * npar$$

This penalizes the fitted value of $= 2lnL$ (which is a positive value), and adds a penalty that depends on the number of fitted parameters.

Bayesian Information Criterion (BIC) replaces the value of 2 with the natural log of the sample size. That means the penalty for additional coefficients is more severe and the BIC is "worse" than the AIC.

$$BIC = -2lnL + ln(N) * npar$$