Overdispersion Models

Statistics 149

Spring 2006

Copyright ©2006 by Mark E. Irwin

Variance Assumptions

Wave Damage to Cargo Ships: As discussed last class, there was some evidence for lack of fit in this example.

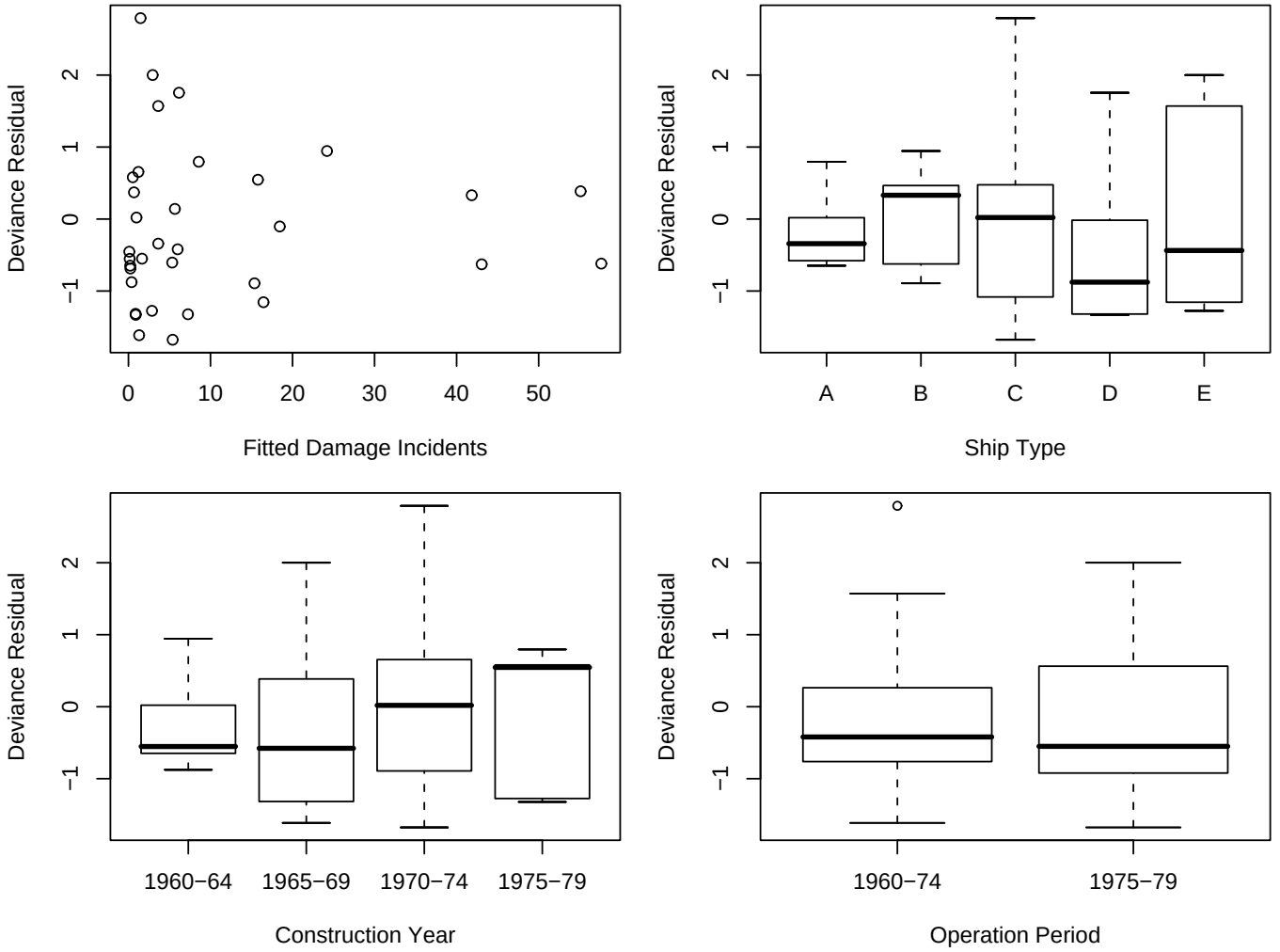
```
> summary(wave.glm)

    Null deviance: 146.328  on 33  degrees of freedom
Residual deviance:  38.695  on 25  degrees of freedom
AIC: 154.56

> pchisq(deviance(wave.glm), df.residual(wave.glm), lower.tail=F)
[1] 0.03951433
```

The deviance GOF test is marginally significant.

As suggested last time the variance assumption

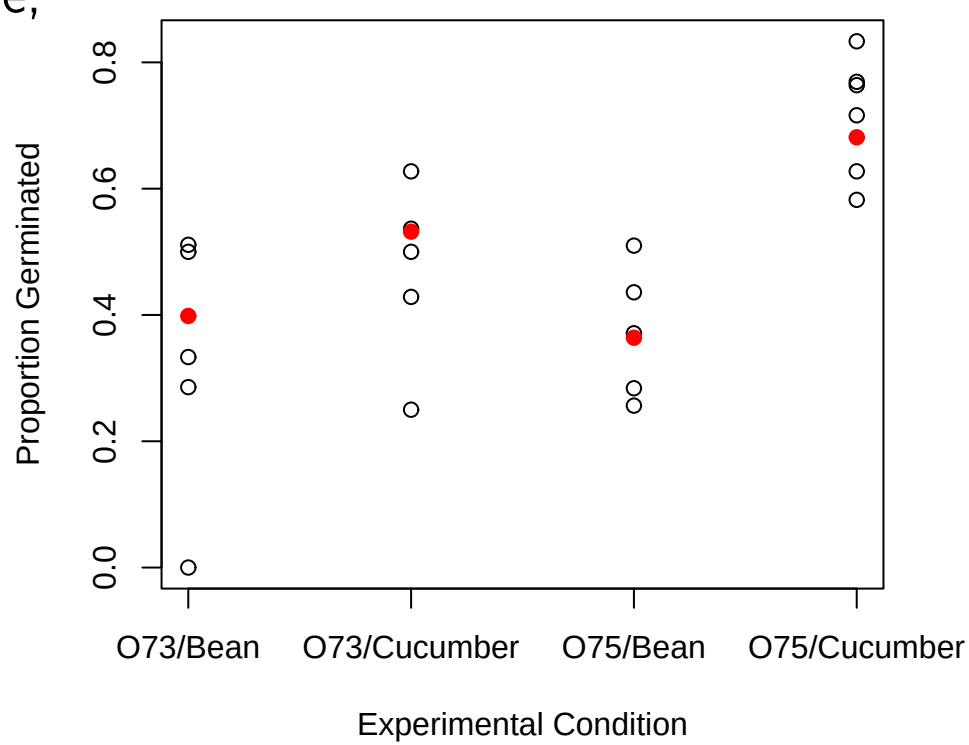
$$\mathrm{Var}(Y_i|X_i) = \mu_i$$

may not be reasonable.



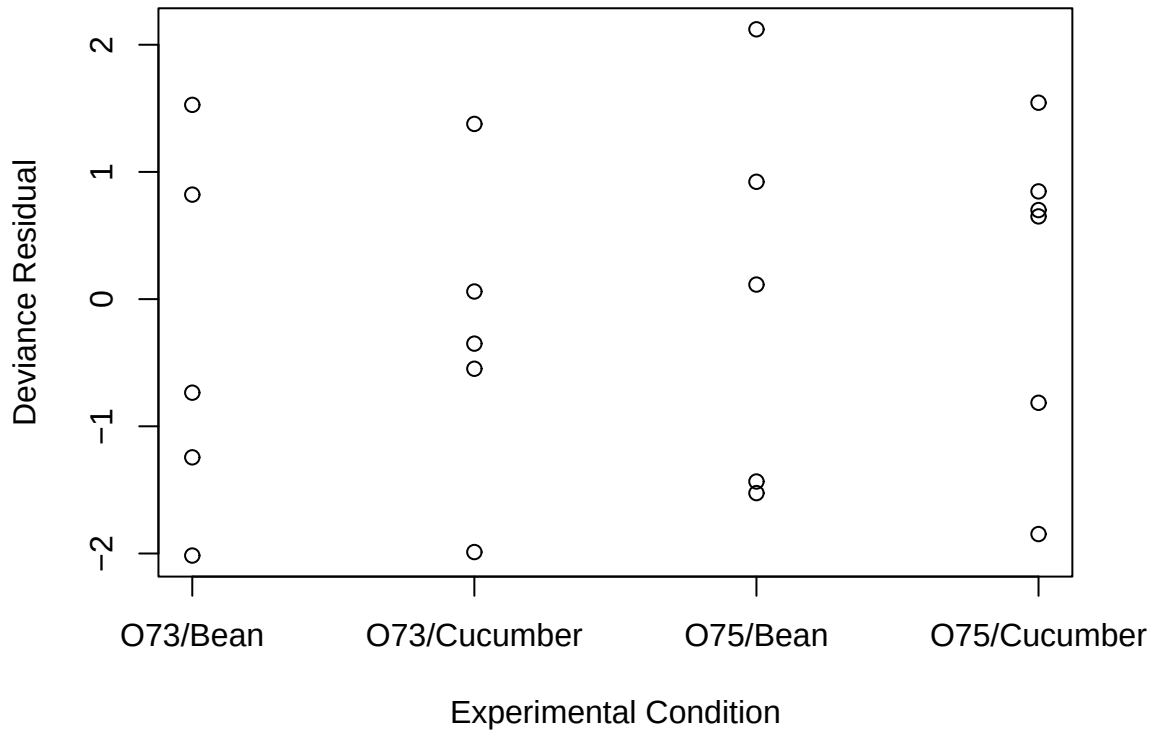
Maybe we want to do some thing else to model the variance.

Germination of Orobanche: Orobanche, commonly known as broomrape, is a genus of parasitic plants without chlorophyll that grow on the roots of flowering plants. In the course of research into factors affecting the germination of the seed of the species of the species Orobanche aegyptiaca, a batch of seeds was brushed onto a plate containing a 1/125 dilution of an extract prepared from the roots of either a bean or cucumber plant. The number of seeds which germinated was recorded. Two different varieties, O. aegyptiaca 75 and O. aegyptiaca 73, were studied.



```
> summary(orobanche.glm)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.01617  -1.24398   0.05995   0.84695   2.12123

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.4122     0.1842  -2.238   0.0252 *
Species075                 -0.1459     0.2232  -0.654   0.5132
ExtractCucumber             0.5401     0.2498   2.162   0.0306 *
Species075:ExtractCucumber  0.7781     0.3064   2.539   0.0111 *
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 98.719  on 20  degrees of freedom
Residual deviance: 33.278  on 17  degrees of freedom
AIC: 117.87

> pchisq(deviance(orobanche.glm), df.residual(orobanche.glm), lower.tail=F)
[1] 0.01039184
```

```
> sum(resid(orobanche.glm,type="pearson")^2)  # Pearson GOF
[1] 31.65114
> pchisq(sum(resid(orobanche.glm,type="pearson")^2),
     df.residual(orobanche.glm), lower.tail=F)
[1] 0.01662130
```



Both the Pearson and deviance Goodness of Fit tests suggests that there is a problem.

It is not clear how the systematic part of the model could be changed. We have all known factors, including all interactions, in the model, there don't appear to be any outliers.

If the problem isn't with the mean structure, maybe its with the variance. In the analysis, we are assuming that

$$\mathrm{Var}(Y_i|X_i) = m_i\pi_i(1-\pi_i)$$

Instead, lets assume that

$$\mathrm{Var}(Y_i|X_i) = \psi m_i\pi_i(1-\pi_i)$$

What happens to the Pearson GOF statistic under this assumption

$$\begin{aligned} X_p^2 &= \sum_{i=1}^{n} \frac{(Y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)} \\ &\approx \psi \sum_{i=1}^{n} \frac{(Y_i - m_i\hat{\pi}_i)^2}{\mathrm{Var}(Y_i)} \\ &\approx \psi\chi^2_{n-p} \end{aligned}$$

So the expected value of the Pearson GOF test is scaled by a factor of ψ. The deviance GOF statistic gets scaled in a similar fashion.

For Poisson count data, if we make the assumption

$$\mathrm{Var}(Y_i|X_i) = \psi\mu_i$$

we get the same scaling of the GOF statistics.

Where could the overdispersion come from

- Differences in the experimental conditions.

  In the Orobanche example, there were 5 or 6 trials under each of the 4 experimental treatments. It may be that there are difference between the different trials under the same treatment (temperature, humidity, etc), leading to different success probabilities.

  One way of thinking about this is the following two-stage model.

$$
\begin{aligned}
Y_i|p_i &\overset{ind}{\sim} Bin(m_i, p_i) \\
p_i &\overset{iid}{\sim} Beta(\alpha, \beta) \\
E[p_i] &= \frac{\alpha}{\alpha+\beta} = \pi \\
\mathrm{Var}(p_i) &= \frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta}\frac{1}{\alpha+\beta+1} = \pi(1-\pi)\gamma
\end{aligned}
$$

Then the marginal moments of Y_i are

$$E[Y_i] = E[E[Y_i|p_i]] = E[mp_i] = m\pi_i$$

$$\begin{aligned}
\mathrm{Var}(Y_i) &= E[\mathrm{Var}(Y_i|p_i)] + \mathrm{Var}(E[Y_i|p_i]) \\
&= E[m_i p_i(1-p_i)] + \mathrm{Var}(m_i p_i) \\
&= m_i(\pi - \mathrm{Var}(p_i) - \pi^2) + m_i^2 \pi(1-\pi)\gamma \\
&= m_i(\pi(1-\pi))(1-\gamma) + m_i^2 \pi(1-\pi)\gamma \\
&= m_i \pi(1-\pi)(1 + (m_i - 1)\gamma) \\
&= \psi m_i \pi(1-\pi)
\end{aligned}$$

where $\tau > 0$ is a function of α and β which implies $\psi > 1$. Y_i is said to have a Beta-Binomial distribution.

In the wave damage example, one of the observation times was extremely long. As conditions could change over time, this would introduce more variability if there where a shift in the type of ship used. In addition it is possible that there may be more variability among the ships for certain ship types (i.e. Type E vs Type A)

- Correlation between responses

  In binomial data, an important underlying assumption is that each of the individual Bernoulli trials is independent of the rest. If not, the variance will not be $m\pi(1-\pi)$.

  Suppose that for m Bernoulli trials, each with the same success probability π, the correlation between any pair of trials is ρ. Then the variance satisfies

$$
\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{Var}\left(\sum Z_i\right) \\
&= \sum_{i=1}^{m} \mathrm{Var}(Z_i) + 2\sum_{i<j} \rho\sqrt{\mathrm{Var}(Z_i)\mathrm{Var}(Z_j)} \\
&= m\pi(1-\pi) + m(m-1)\rho\pi(1-\pi) \\
&= m\pi(1-\pi)(1+(m-1)\rho) \\
&= \psi m\pi(1-\pi)
\end{aligned}
$$

In the Orobanche example, this might happen if a germinating seed produces a chemical that produces that promotes germination in other seeds.

In the wave damage example, suppose multiple ships where in the same area at a time of a storm. The if one ship gets wave damage, I would expect the others to have an increased chance of damage.

Consequences of Overdispersion

In the earlier analyzes, the inferences performed are based on the assumption that $\psi = 1$. If in fact $\psi > 1$, the standard errors used are too small.

Consider the situation where we consider $Y_1, Y_2, \ldots, Y_n$ as iid draws from $P(\mu)$, but in reality $\mathrm{Var}(Y_i) = \psi\mu$. Then

$$\mathrm{Var}(\bar{Y}) = \frac{\psi\mu}{n}$$

but the analysis would use

$$\mathrm{Var}(\bar{Y}) = \frac{\mu}{n}$$

So for example, the calculated confidence interval for μ would be

$$\bar{Y} \pm z^*_{\alpha/2}\sqrt{\bar{Y}}$$

when instead it really should be

$$\bar{Y} \pm z^*_{\alpha/2}\sqrt{\psi\bar{Y}}$$

The interval would be centered at the right place, but would be too narrow by a factor of $\sqrt{\psi}$.

This situation follows through in the GLM regression analyzes. The reported standard errors and z-tests are off by a factor of $\sqrt{\psi}$, which implies that it is too easy to declare something significant if $\psi > 1$.

Similarly, the deviance based tests are also invalid as the statistics don't have the nominal χ^2 distributions, but scaled χ^2 distributions.

We need to account for overdispersion, if it exists, to have valid analyzes.

Quasi-Likelihood Analysis

One approach to dealing with overdispersion would be directly model the overdispersion with a likelihood based models. For example, use a beta-binomial model in the binomial case.

Another approach, which is easier to implement in the regression setting, is a quasi-likelihood approach. Instead of giving a full probability model, only moment assumptions will be made. A common approach is

- Systematic component ($E[Y_i|X_i] = \mu_i$):

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

  where $g(\cdot)$ is a link function.

- Dispersion component ($\mathrm{Var}(Y_i|X_i)$):

$$\mathrm{Var}(Y_i|X_i) = \psi V(\mu_i)$$

where $V(\cdot)$ is the variance function.

In general, the variance function can be quite general, when trying to deal with overdispersion, usually you keep the form matching the distribution you would like to use for modeling i.e.

- Binomial: $V(\pi) = \pi(1 - \pi)$
- Poisson: $V(\mu) = \mu$

R has families available in `glm` to handle these situation, `quasibinomial()`, `quasipoisson()`, and `quasi()` (general situation).

The `quasibinomial()` and `quasipoisson()` take the same link functions as `binomial()`, `poisson()` so any analysis discusses so far can be done with overdispersion accounted for. The family `quasi()` takes a range of link and variance functions (see `help(family)` to see all that are possible).

In one sense, the extra families `quasibinomial()` and `quasipoisson()` aren't needed, as the adjustment to the analyzes are simple.

First, the estimates of the βs don't change. To prove this would get into the details of the fitting algorithm. But the underlying idea uses the fact the fitting algorithm uses iteratively reweighted least squares. The algorithm doesn't calculations of the form

$$\hat{\beta}_{work} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$$

where the weight matrix $\mathbf{W}$ depends on the current guess for β.

If the overdispersion is accounted for, the form of weight matrix should be $\frac{1}{\psi}\mathbf{W}$. This gives

$$\begin{aligned}
\hat{\beta}^*_{work} &= (\mathbf{X}^T\tfrac{1}{\psi}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\tfrac{1}{\psi}\mathbf{W}\mathbf{Y} \\
&= \psi(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\tfrac{1}{\psi}\mathbf{W}\mathbf{Y} \\
&= \hat{\beta}_{work}
\end{aligned}$$

Its similar to the case that you don't need to know σ^2 in least squares to estimate β.

While it is not needed for estimation, it is needed for inference. As in the least squares situation, we can estimate. There are two possible estimates

- Deviance estimate:

$$\hat{\psi}_D = \frac{\text{Residual Deviance}}{\text{Degrees of freedom}} = \frac{X^2}{df}$$

  This is the estimate suggested by Ramsey and Schafer.

- Pearson estimate:

$$\hat{\psi}_P = \frac{X_p^2}{df}$$

  This is the estimate implemented in **R** and recommended by McCullagh and Nelder.

Both these estimates based on the fact that under the correct model, both GOF statistics have approximate $\psi\chi^2_{df}$ distributions, which has expected value $\psi \times df$.

Inference in this case needs to make adjustments for the estimated $\hat{\psi}$. The first adjustment is that the standard errors must be multiplied by $\sqrt{\hat{\psi}}$, i.e. if $\mathrm{SE}_l(\hat{\beta}_j)$ is the standard error in the likelihood analysis,

$$\mathrm{SE}(\hat{\beta}_j) = \mathrm{SE}_l(\hat{\beta}_j)\sqrt{\hat{\psi}}$$

Another adjustment that is made on interference on single parameters is to use a reference t_{df} distribution instead of the $N(0,1)$ distribution. There is little theory to justify using the t_{df} instead of the $N(0,1)$. The motivation is more to mimic the methods of least squares. However it does have the advantage of being conservative, particularly in the small sample case.

For the Orobanche example

- $\psi = 1$:

```
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.4122     0.1842  -2.238   0.0252 *
Species075                 -0.1459     0.2232  -0.654   0.5132
ExtractCucumber             0.5401     0.2498   2.162   0.0306 *
Species075:ExtractCucumber  0.7781     0.3064   2.539   0.0111 *
---
(Dispersion parameter for binomial family taken to be 1)
Residual deviance: 33.278  on 17  degrees of freedom
```

- Overdispersion:

```
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                -0.4122     0.2513  -1.640   0.1193
Species075                 -0.1459     0.3045  -0.479   0.6379
ExtractCucumber             0.5401     0.3409   1.584   0.1315
Species075:ExtractCucumber  0.7781     0.4181   1.861   0.0801 .
---
(Dispersion parameter for quasibinomial family taken to be 1.862)
Residual deviance: 33.278  on 17  degrees of freedom
```

Including overdispersion in the model makes difference in the conclusion here. In the original analysis, the interaction looks to have fairly strong statistical significance. In the new model, the significance is now marginal.

For the main effects model, the conclusions don't change as much

- $\psi = 1$:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3919  -0.9948  -0.3744   0.9831   2.4766

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.7005     0.1507  -4.648 3.36e-06 ***
Species075       0.2705     0.1547   1.748   0.0804 .
ExtractCucumber  1.0647     0.1442   7.383 1.55e-13 ***
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 98.719  on 20  degrees of freedom
Residual deviance: 39.686  on 18  degrees of freedom
AIC: 122.28
```

• Overdispersion:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3919  -0.9948  -0.3744   0.9831   2.4766

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.7005     0.2199  -3.186  0.00512 **
Species075       0.2705     0.2257   1.198  0.24635
ExtractCucumber  1.0647     0.2104   5.061 8.14e-05 ***
(Dispersion parameter for quasibinomial family taken to be 2.128368)

    Null deviance: 98.719  on 20  degrees of freedom
Residual deviance: 39.686  on 18  degrees of freedom
AIC: NA
```

So extract appears to still be highly significant, which was evident in the original scatterplot. Species went marginally insignificant to quite insignificant in the overdispersion analysis. Again agreeing with the scatterplot, which suggested a small species effect if one existed.

A couple of other things to note. First note that there is no AIC given for the quasi-likelihood analysis. This makes sense as there is no likelihood function here.

Next, notice that residual summary is the same in both analyzes. In fact, the deviance and Pearson residuals in the quasi-likelihood analysis are defined the same way as before. Thus the can be examined for patterns and peculiarities, but can no longer be compared to a reference distribution to detect outliers.

Similarly, the influence measures discussed earlier don't change when overdispersion is accounted for. This seems reasonable since the regression parameter estimates and thus the fitted values don't change. Also the effect of including $\sqrt{\hat{\psi}}$ in the calculation of the measures ends up getting canceled out.

A small section of the

- $\psi = 1$:

```
     dfb.1_ dfb.SO75 dfb.ExtC    dffit cov.r   cook.d    hat inf
1  -0.16802  -0.1977  0.31598 -0.49154 1.004 7.73e-02 0.1205
2  -0.04544  -0.0535  0.08546 -0.13293 1.449 6.21e-03 0.1915
3  -0.33454  -0.3936  0.62916 -0.97871 0.991 2.89e-01 0.2502
4   0.18388   0.2163 -0.34582  0.53795 1.085 9.36e-02 0.1575
```

- Overdispersion:

```
     dfb.1_ dfb.SO75 dfb.ExtC    dffit cov.r   cook.d    hat inf
1  -0.16802  -0.1977  0.31598 -0.49154 1.004 7.73e-02 0.1205
2  -0.04544  -0.0535  0.08546 -0.13293 1.449 6.21e-03 0.1915
3  -0.33454  -0.3936  0.62916 -0.97871 0.991 2.89e-01 0.2502
4   0.18388   0.2163 -0.34582  0.53795 1.085 9.36e-02 0.1575
```

When testing multiple regression parameters, the drop in deviance tests must also be modified. The approach is to mimic F-tests in linear regression. The drop in deviance F-test has the form

$$F = \frac{\text{Drop in deviance}/d}{\hat{\psi}}$$

when d is the difference in the number of parameters in the two models begin compared. This F statistic should be compared to a $F_{d,df}$ where df is the residual degrees of freedom from the full model.

As with the t procedures discussed earlier, there is little solid theory to justify this test.

To use this test in **R**, the `anova` function can be used, but the option, `test="F"`, must be used instead. For example,

```
> anova(wave.c.qglm, wave.qglm, test="F")
Analysis of Deviance Table

Model 1: Damage ~ Type + Operation
Model 2: Damage ~ Type + Construct + Operation
  Resid. Df Resid. Dev Df Deviance      F   Pr(>F)
1        28     70.103
2        25     38.695  3   31.408 6.1912 0.002708 **
```

In **R**'s implementation of the test, $\hat{\psi}$ is the Pearson estimate.

One additional comment on this example. Another mechanism that can lead to the appearance of overdispersion is missing predictors. If you can find components to add to the systematic component, it is usually preferable to a more complicated variance model.

In the wave example, adding the `Type:Construct` interaction seems to explain alot.

```
> summary(wave.qglm)

Call:
glm(formula = Damage ~ Type + Construct + Operation,
    family = quasipoisson(), data = wave2, offset = log(Service))

(Dispersion parameter for quasipoisson family taken to be 1.691)

    Null deviance: 146.328  on 33  degrees of freedom
Residual deviance:  38.695  on 25  degrees of freedom

> summary(wave2.qglm)

Call:
glm(formula = Damage ~ Type + Construct + Operation + Type:Construct,
    family = quasipoisson(), data = wave2, offset = log(Service))

(Dispersion parameter for quasipoisson family taken to be 1.336)

    Null deviance: 146.328  on 33  degrees of freedom
Residual deviance:  14.587  on 13  degrees of freedom
```

```
> anova(wave2.glm,test="Chisq")
Analysis of Deviance Table

Model: poisson, link: log

Response: Damage

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                             33    146.328
Type            4   55.439        29     90.889 2.629e-11
Construct       3   41.534        26     49.355 5.038e-09
Operation       1   10.660        25     38.695     0.001
Type:Construct 12   24.108        13     14.587     0.020
```

So considering this an example of overdispersion probably isn't valid.