




STAT 504

Analysis of Discrete Data

7.2.1 - Model Diagnostics

 [Printer-friendly version \(.../print/book/export/html/161/\)](#)

This section is dedicated to studying the appropriateness of the model. Do the model assumptions hold? This is done via various diagnostics, such as assessing the distribution of the residuals.

Let us begin with a bit of a review of Linear Regression diagnostics. The standard linear regression model is given by:

$$y_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = x_i^T \beta$$

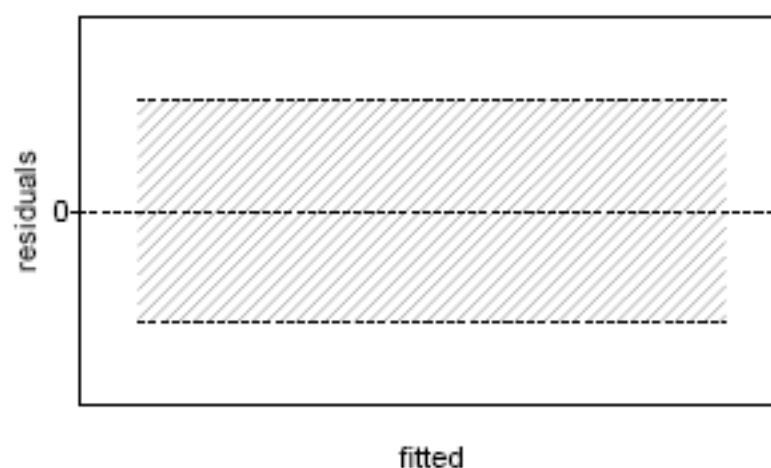
The two crucial features of this model are

- the assumed mean structure, $\mu_i = x_i^T \beta$, and
- the assumed constant variance σ^2 (homoscedasticity).

The most common diagnostic tool is the residuals, the difference between the estimated and observed values of the dependent variable. There are other useful regression diagnostics, e.g. measures of leverage and influence, but for now our focus will be on the estimated residuals.

The most common way to check these assumptions is to fit the model and then plot the residuals versus the fitted values $\hat{y}_i = x_i^T \hat{\beta}$.

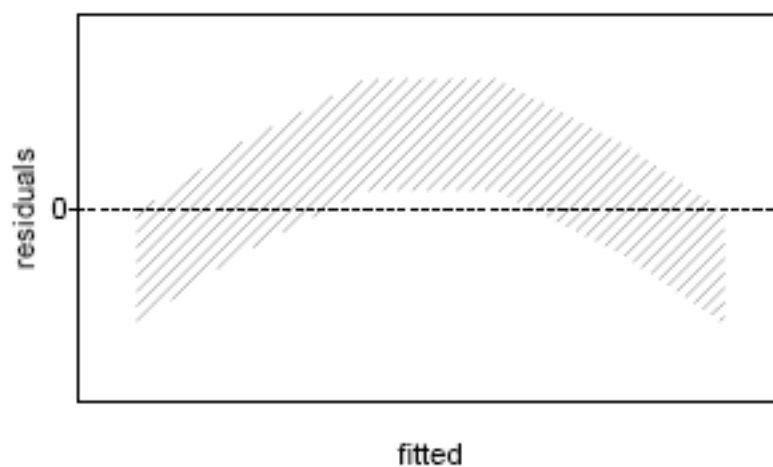
- If the model assumptions are correct, the residuals should fall within an area representing a horizontal band, like this:



- If the residuals have been standardized in some fashion (i.e., scaled by an estimate of σ), then we would expect

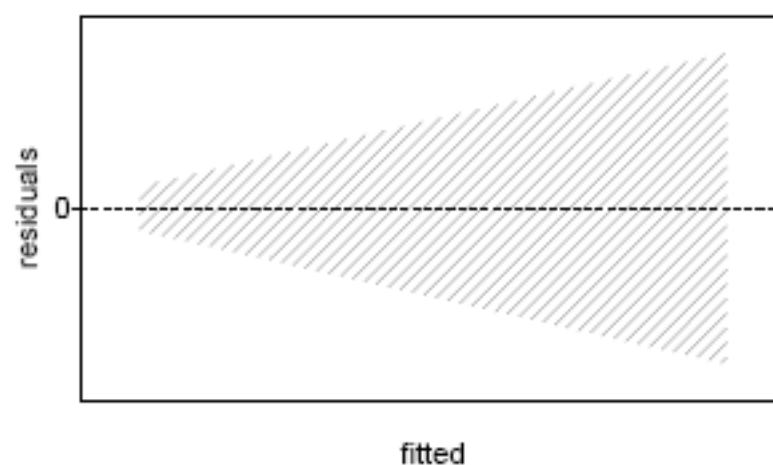
most of them to have values within ± 2 or ± 3 ; residuals outside of this range are potential outliers.

- If the plot reveals some kind of curvature—for example, like this,

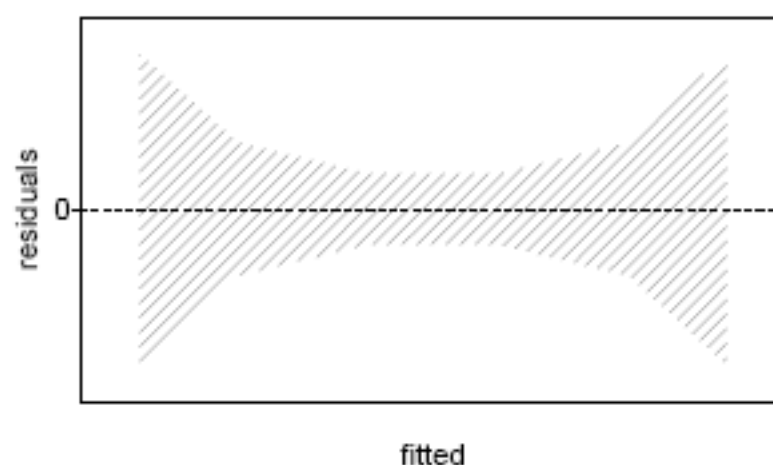


it suggests a failure of the mean model; the true relationship between μ_i and the covariates might not be linear.

- If the variability in the residuals is not constant as we move from left to right—for example, if the plot looks like this,



or like this,



then the variance $V(y_i)$ is not constant but changes as the mean μ_i changes.

Analogous plots for logistic regression.

The logistic regression model says that the mean of y_i is

$$\mu_i = n_i \pi_i$$

where

$$\log \left(\frac{\pi_i}{1 - \pi_i} \right) = x_i^T \beta$$

and that the variance of y_i is

$$V(y_i) = n_i \pi_i (1 - \pi_i).$$

After fitting the model, we can calculate the Pearson residuals (../220/)

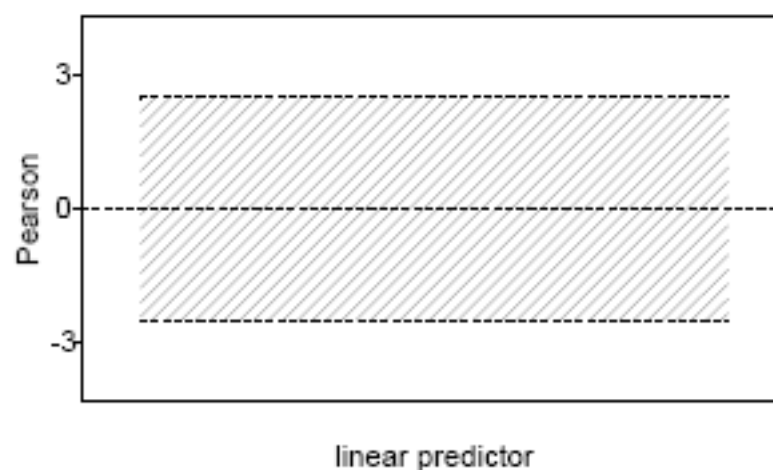
$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{V}(y_i)}} = \frac{y_i - n_i \hat{\pi}_i}{\sqrt{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}}$$

or the deviance residuals (../220/). If the n_i 's are "large enough", these act something like standardized residuals in linear regression. To see what's happening, we can plot them against the *linear predictors*,

$$\hat{\eta}_i = \log \left(\frac{\pi_i}{1 - \pi_i} \right) = x_i^T \hat{\beta}_i$$

which are the estimated log-odds of success, for cases $i = 1, \dots, N$.

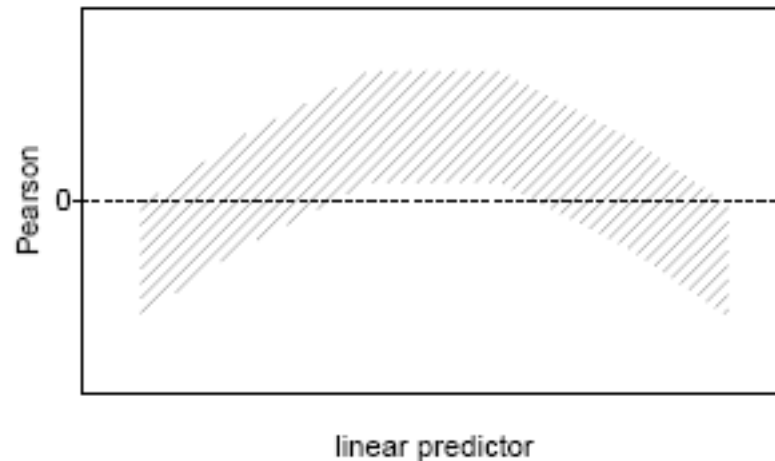
- If the fitted logistic regression model was true, we would expect to see a horizontal band with most of the residuals falling within ± 3 :



- If the n_i 's are small, then the plot might not have such a nice pattern, even if the model is true. In the extreme case of ungrouped data (all n_i 's equal to 1), this plot becomes uninformative. From now on, we will suppose that the n_i 's are not too small, so that the plot is at least somewhat meaningful.
- If outliers are present—that is, if a few residuals or even one residual is substantially larger than ± 3 — then X^2 and G^2 may be much larger than the degrees of freedom. In that situation, the lack of fit can be attributed to outliers, and the large residuals will be easy to find in the plot.

Curvature

- If the plot of Pearson residuals versus the linear predictors reveals curvature—for example, like this,



then it suggests that the mean model

$$\log \left(\frac{\pi_i}{1 - \pi_i} \right) = x_i^T \beta \quad (2)$$

has been mis-specified in some fashion. That is, it could be that one or more important covariates do not influence the log-odds of success in a linear fashion. For example, it could be that a covariate X ought to be replaced by a transformation such as \sqrt{X} or $\log X$, or by a pair of variables X and X^2 , etc.

To see whether individual covariates ought to enter in the logit model in a non-linear fashion, we could plot the *empirical logits*

$$\log \left(\frac{y_i + 1/2}{n_i - y_i + 1/2} \right) \quad (3)$$

versus each covariate in the model.

- If one of the plots looks substantially non-linear, we might want to transform that particular covariate.
- If many of them are nonlinear, however, it may suggest that the link function has been misspecified—i.e., that the left-hand side of (2) should not involve a logit transformation, but some other function such as
 - log,
 - probit (but this is often very close to logit), or
 - complementary log-log.

Changing the link function will change the interpretation of the coefficients entirely; the β_j 's will no longer be log-odds ratios. But, depending on what the link function is, they might still have a nice interpretation. For example, in a model with a log link, $\log \pi_i = x_i^T \beta$, an exponentiated coefficient $\exp(\beta_j)$ becomes a relative risk.

Test for correctness of the link function

Hinkley (1985) suggests a nice, easy test to see whether the link function is plausible:

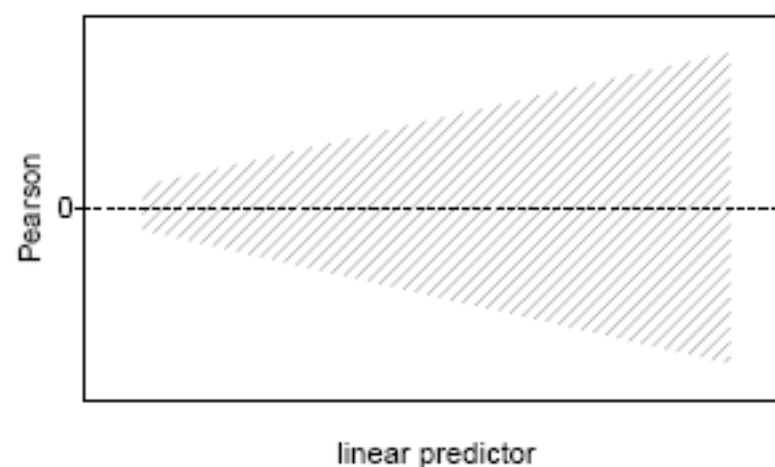
- Fit the model and save the estimated linear predictors

$$\hat{\eta}_i = x_i^T \hat{\beta}$$
- Add η_i^2 to the model as a new predictor and see if it's significant.

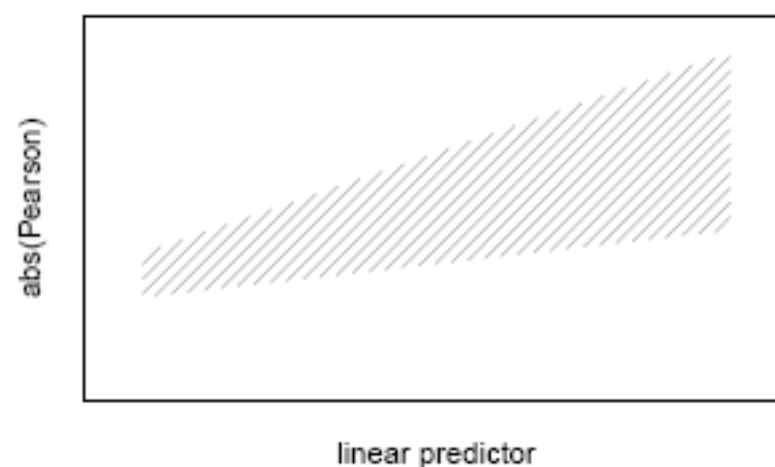
A significant result indicates that the link function is mis-specified. A nice feature of this test is that it applies even to ungrouped data (n_i 's equal to one), for which residual plots are uninformative.

Non-constant Variance

Suppose that the residual plot shows non-constant variance as we move from left to right:



Another way to detect non-constancy of variance is to plot the *absolute values* of the residuals versus the linear predictors and look for a non-zero slope:



Non-constant variance in the Pearson residuals means that the assumed form of the variance function,

$$V(y_i) = n_i \pi_i (1 - \pi_i)$$

is wrong and cannot be corrected by simply introducing a scale factor for overdispersion. Overdispersion and changes to the variance function will be discussed later.

Example

The SAS on-line help documentation provides the following quantal assay dataset. In this table, x_i refers to the log-dose, n_i is the number of subjects exposed, and y_i is the number who responded.

x_i	y_i	n_i
2.68	10	31
2.76	17	30
2.82	12	31
2.90	7	27
3.02	23	26

3.04	22	30
3.13	29	31
3.20	29	30
3.21	23	30


If we fit a simple logistic regression model, we will find that the coefficient for x_i is highly significant, but the model doesn't fit. The plot of Pearson residuals versus the fitted values resembles a horizontal band, with no obvious curvature or trends in the variance. This seems to be a classic example of overdispersion.

Since there's only a single covariate, a good place to start is to plot the *empirical logits* as defined in equation (3) above versus X .

This is basically the same thing as a scatterplot of Y versus X in the context of ordinary linear regression. This plot becomes more meaningful as the n_i 's grow. With ungrouped data (all $n_i = 1$), the empirical logits will only take two possible values— $\log(1/3)$ and $\log 3/1$ —and the plot will not be very useful.

Using SAS (#)

Using R (#)



Here is the SAS program file assay.sas

(../sites/onlinecourses.science.psu.edu.stat504/files/lesson06/assay/index.sas) :

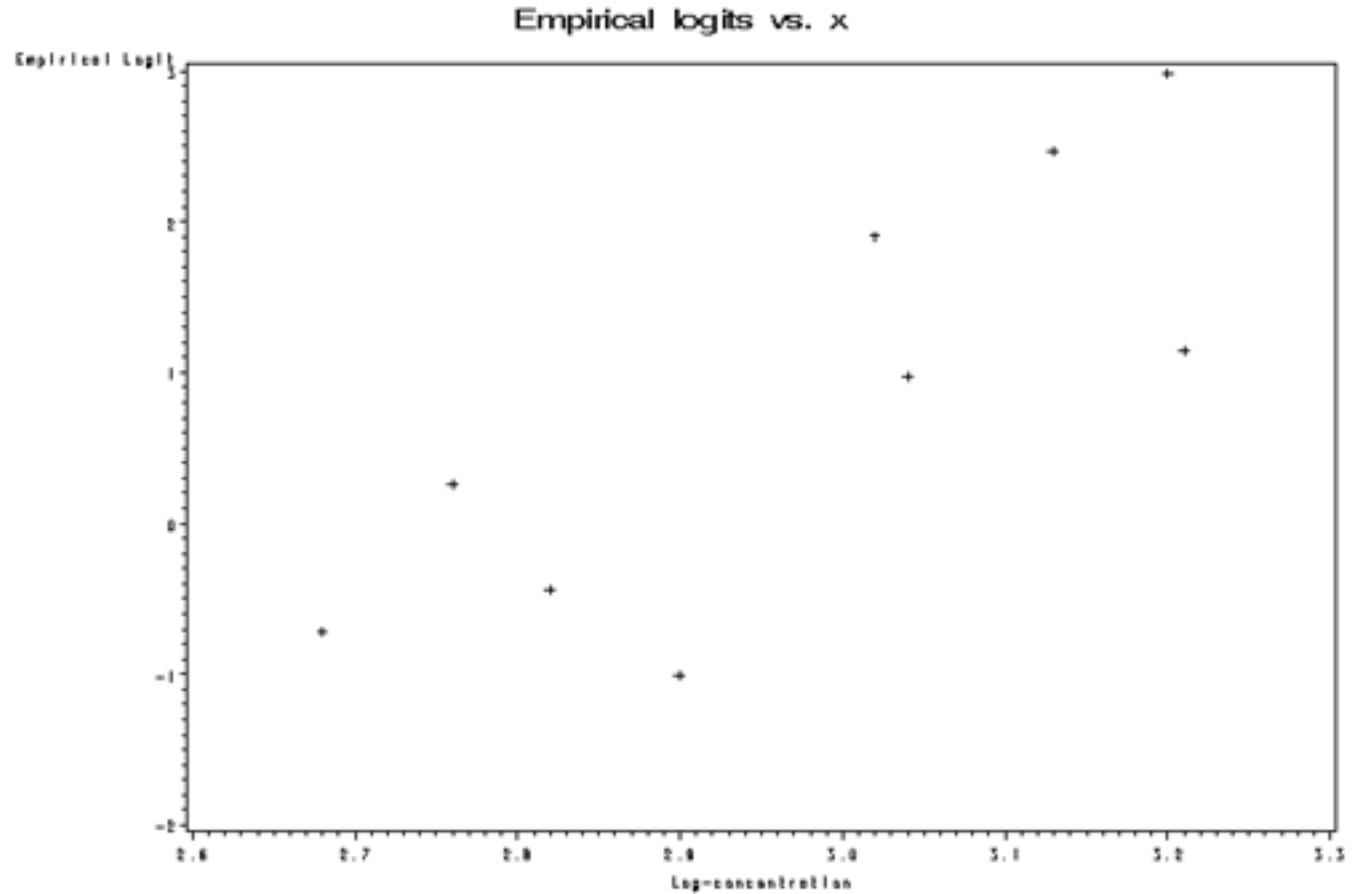
```

options nocenter nodate nonumber linesize=72;
data assay;
  input logconc y n;
  emplogit=log((y+.5)/(n-y+.5));
cards;
2.68 10 31
2.76 17 30
2.82 12 31
2.90 7 27
3.02 23 26
3.04 22 30
3.13 29 31
3.20 29 30
3.21 23 30
;
run;

axis1 label=('Log-concentration');
axis2 label=('Empirical Logit');

proc gplot data=assay;
  title 'Empirical logits vs. x';
  plot emplogit * logconc / haxis=axis1 vaxis=axis2;
run;

```



The relationship between the logits and X seems linear. Let's fit the logistic regression and see what happens.

See assay1.sas ([../sites/onlinecourses.science.psu.edu/stat504/files/lesson06/assay1/index.sas](http://sites/onlinecourses.science.psu.edu/stat504/files/lesson06/assay1/index.sas)) :

```
options nocenter nodate nonumber linesize=72;

data assay;
  input logconc y n;
  cards;
2.68 10 31
2.76 17 30
2.82 12 31
2.90 7 27
3.02 23 26
3.04 22 30
3.13 29 31
3.20 29 30
3.21 23 30
;
run;

proc logistic data=assay;
  model y/n= logconc / scale=none;
  output out=out1 xbeta=xb reschi=reschi;
run;

axis1 label=('Linear predictor');
axis2 label=('Pearson Residual');

proc gplot data=out1;
  title 'Residual plot';
  plot reschi * xb / haxis=axis1 vaxis=axis2;
run;
```

In `plot reschi * xb`, `reschi` are the Pearson residuals, and `xb` the *linear predictor*.

The output reveals that the coefficient of *X* is highly significant, but the model does not fit.

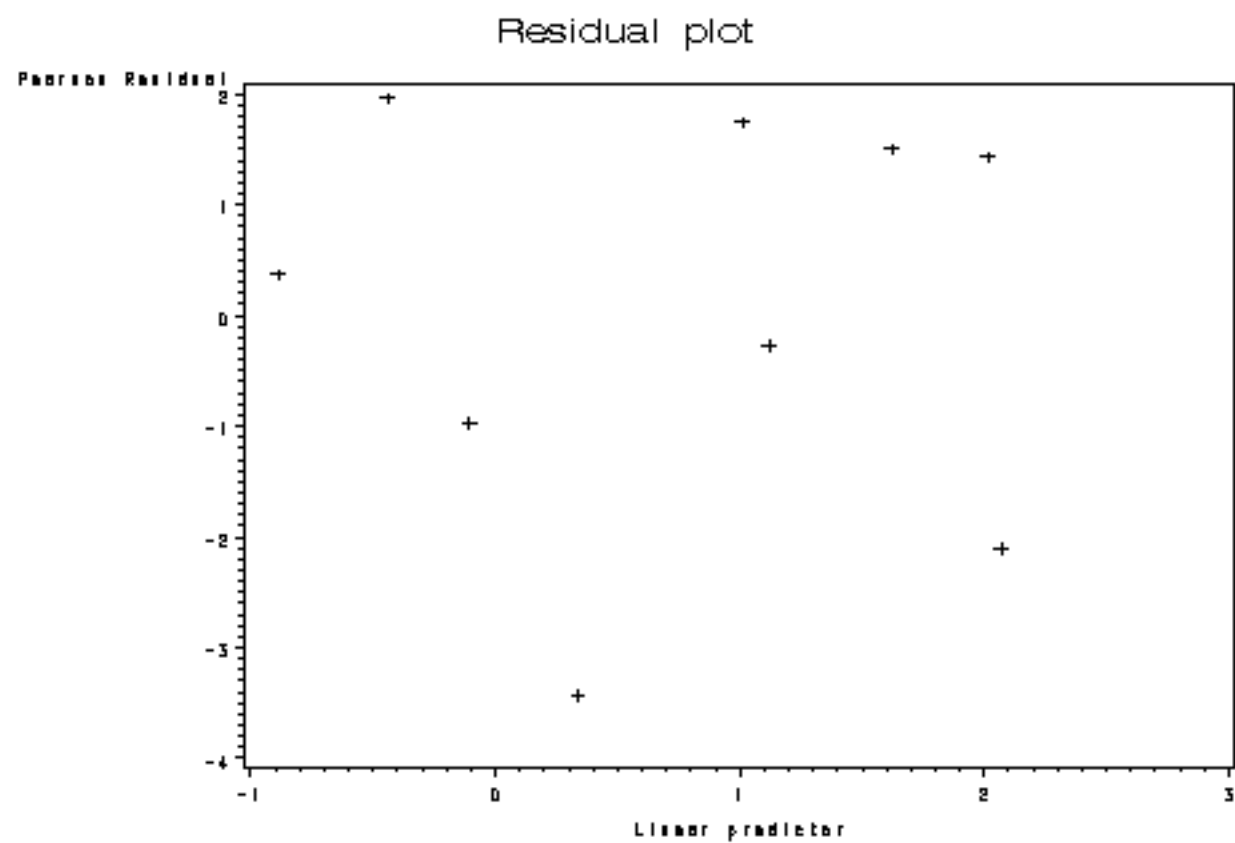
Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	29.3462	7	4.1923	0.0001
Pearson	28.5630	7	4.0804	0.0002

Number of events/trials observations: 9

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-15.8331	2.4371	42.2072	<.0001
logconc	1	5.5776	0.8319	44.9521	<.0001



Here is the R code `assay.R` ([../sites/onlinecourses.science.psu.edu.stat504/files/lesson06/assay/index.R](http://sites/onlinecourses.science.psu.edu/stat504/files/lesson06/assay/index.R)) that

corresponds to the SAS program `assay1.sas`:


```

options(contrasts=c("contr.treatment", "contr.poly"))

#### Logistic regression

r=c(10,17,12,7,23,22,29,29,23)
n=c(31,30,31,27,26,30,31,30,30)
logconc=c(2.68,2.76,2.82,2.90,3.02,3.04,3.13,3.20,3.21)
counts=cbind(r,n-r)
result=glm(counts~logconc,family=binomial("logit"))
summary(result,correlation=TRUE,symbolic.cor = TRUE)
result$coefficients

library()
plot.lm(result)

```

With `plot.lm(result)`, R will produce four diagnostic plots, including a residual plot, a QQ plot, a scale-location plot, and a residual vs leverage plot as well.

The output reveals that the coefficient of X is highly significant, but the model does not fit.

```

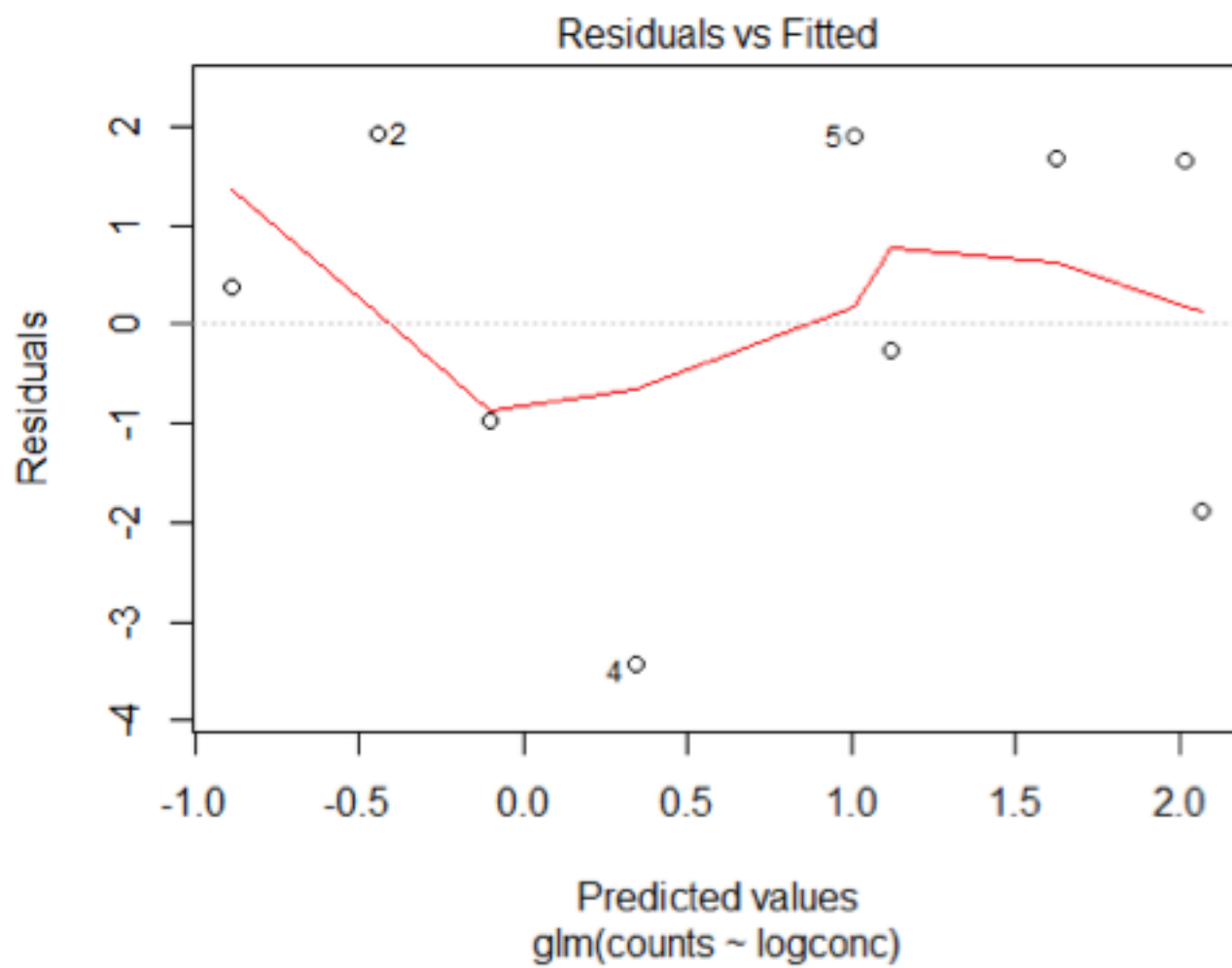
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -15.8339      2.4371  -6.497 8.20e-11 ***
logconc       5.5778       0.8319   6.705 2.02e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.631  on 8  degrees of freedom
Residual deviance: 29.346  on 7  degrees of freedom
AIC: 62.886

```

Here is the residual plot from R output:



The residuals plots in both SAS and R above do not show any obvious curvature or trends in the variance. And there are no other predictors that are good candidates for inclusion in the model. (It can be shown that a quadratic term for log-concentration will not improve the fit.) So it is quite reasonable to attribute the problem to overdispersion.