Model Assessment

Statistics 149

Spring 2006

Logistic Regression for Binomial Responses

So far we've mainly focused on the case, from a **theory** point of view, on binary (Bernoulli trials) responses, not Binomial responses, i.e. looked at

$$Y_i \overset{ind}{\sim} Bin(1, \pi(X_i))$$

not

$$Y_i \overset{ind}{\sim} Bin(m_i, \pi(X_i)); \qquad m_i \geq 1$$

When discussing examples based on binomial sampling ($m_i > 1$), I've treated each of the m_i responses separately. While not justifying it yet, it is a valid thing to do.

So for observation Y_i, let $Z_{ij}, j = 1, \ldots, m_i$ be the individual trials such that

$$Y_i = \sum_{j=1}^{m_i} Z_{ij}$$

The likelihood based on the Z_{ij} is

$$\begin{aligned} L(\pi|\mathbf{Z}) &= \prod_{i=1}^{n}\prod_{j=1}^{m_i} \pi_i^{Z_{ij}}(1-\pi_i)^{1-Z_{ij}} \\ &= \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{m_i-Y_i} \end{aligned}$$

The likelihood based on the the Y_is is

$$\begin{aligned} L(\pi|\mathbf{Y}) &= \prod_{i=1}^{n} \binom{m_i}{Y_i} \pi_i^{Y_i}(1-\pi_i)^{m_i-Y_i} \\ &= C(\mathbf{Y}) \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{m_i-Y_i} \\ &= C(\mathbf{Y}) L(\pi|\mathbf{Z}) \end{aligned}$$

So the only difference in the likelihoods is the normalizing constant, which doesn't affect the maximization with respect to β.

One difference that does occur with $m_i > 1$ is that model assessment is easier.

It is possible to check for things like

- Goodness of Fit
- Outliers
- Influential Points

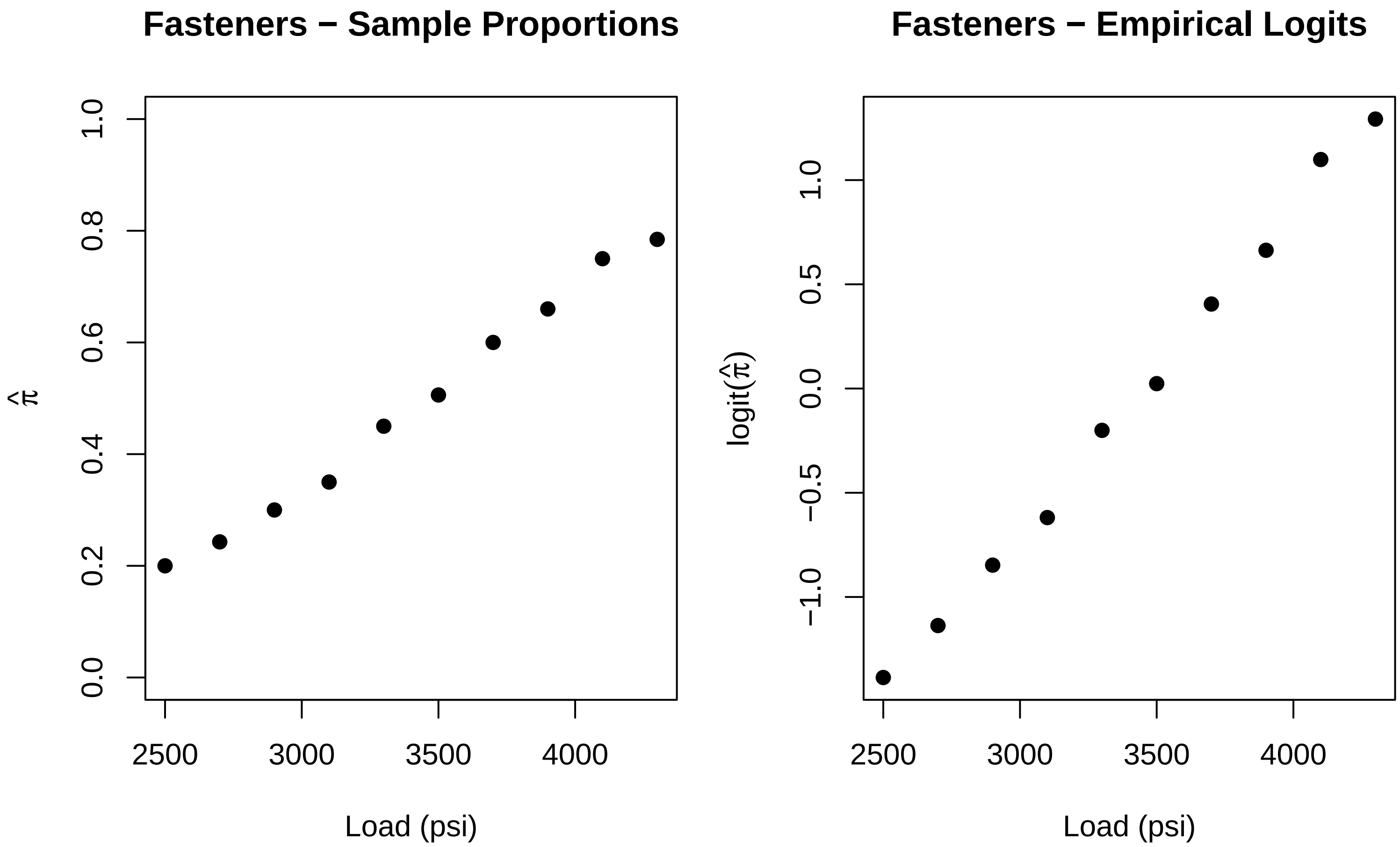
Model Assessment

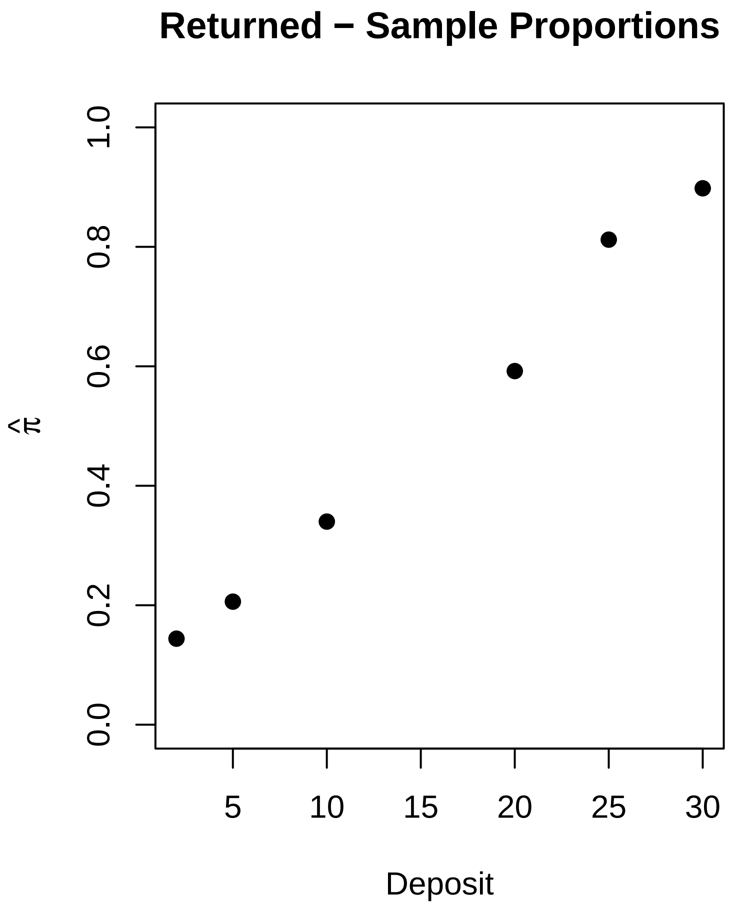
When $m_i > 0$, $\hat{p}_i = \frac{Y_i}{m_i}$ is an unbiased estimate of $\pi_i = \pi(X_i)$. So we can use these to help check the model.

Earlier we saw plots of $\hat{p}$ vs x_i and saw the S shaped relationship. However this really doesn't show whether

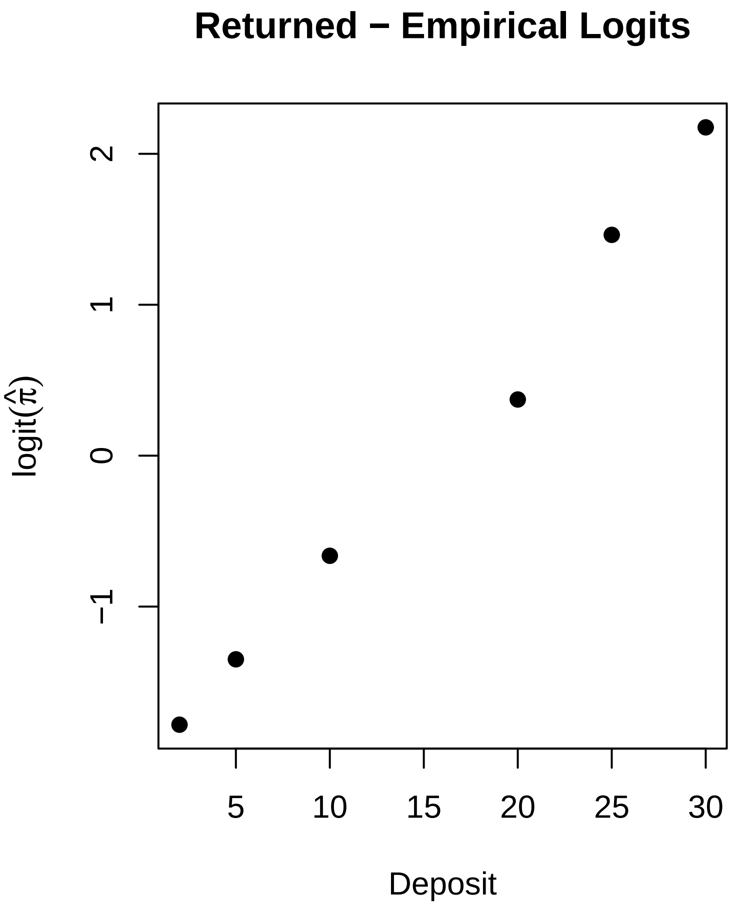
$$\operatorname{logit}(\pi(x_i)) = \beta_0 + \beta_1 x_i$$

is a reasonable model. One thing we can do instead is to plot $\operatorname{logit}(\hat{p}_i)$ vs x_i and check to see if this is approximately linear. For example, for the fastener and pop bottle deposit examples,


Fasteners − Sample Proportions
$\hat{\pi}$
0.0
0.2
0.4
0.6
0.8
1.0
2500
3000
3500
4000
Load (psi)
Fasteners − Empirical Logits
$\text{logit}(\hat{\pi})$
−1.0
−0.5
0.0
0.5
1.0
2500
3000
3500
4000
Load (psi)

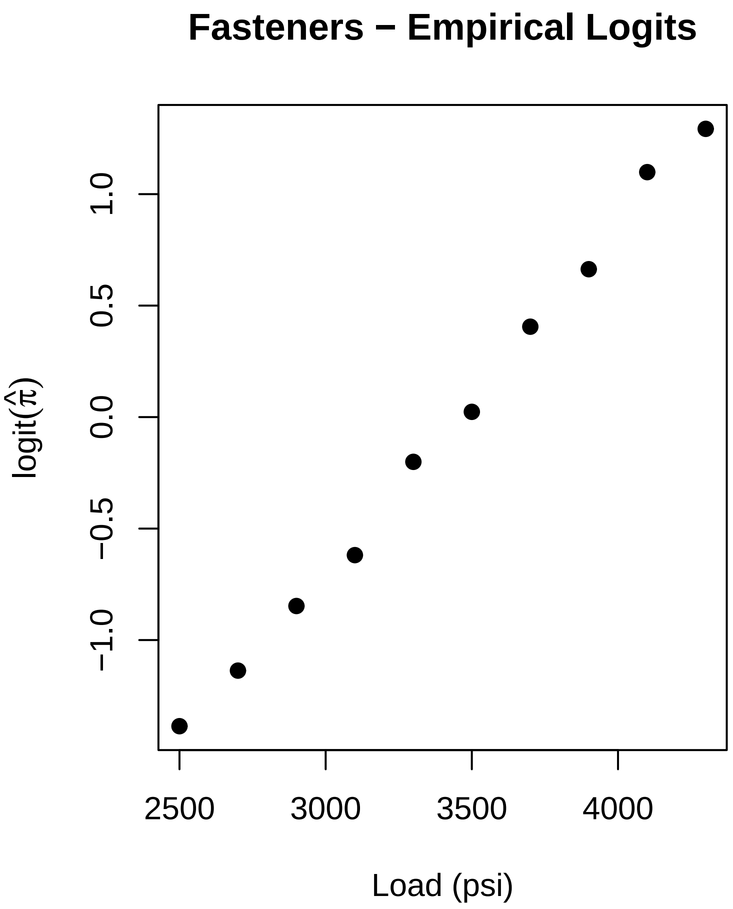


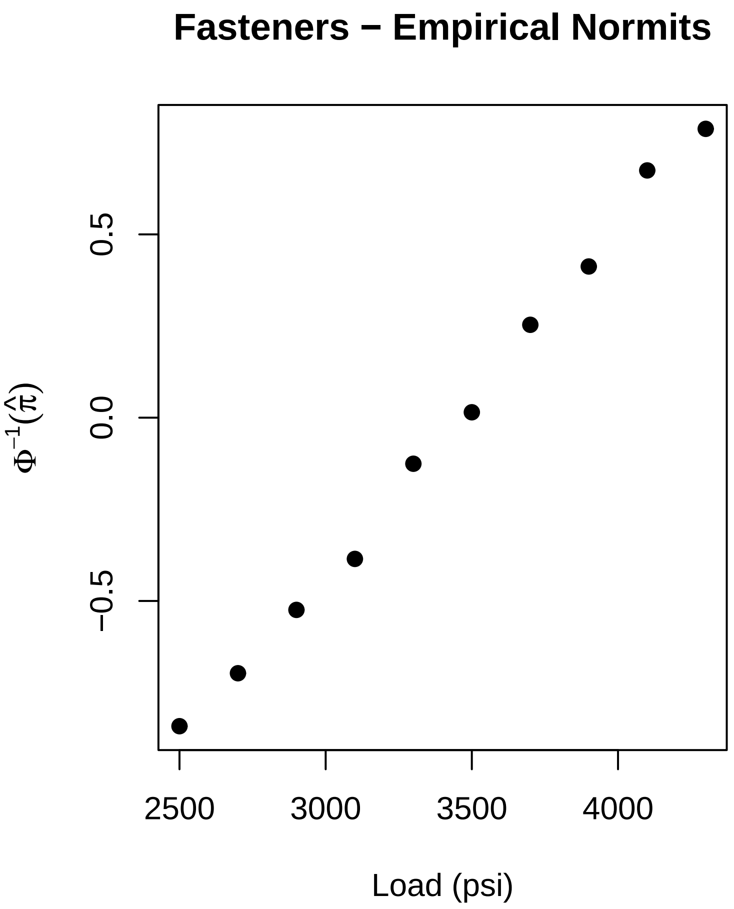
Returned − Sample Proportions



Returned − Empirical Logits

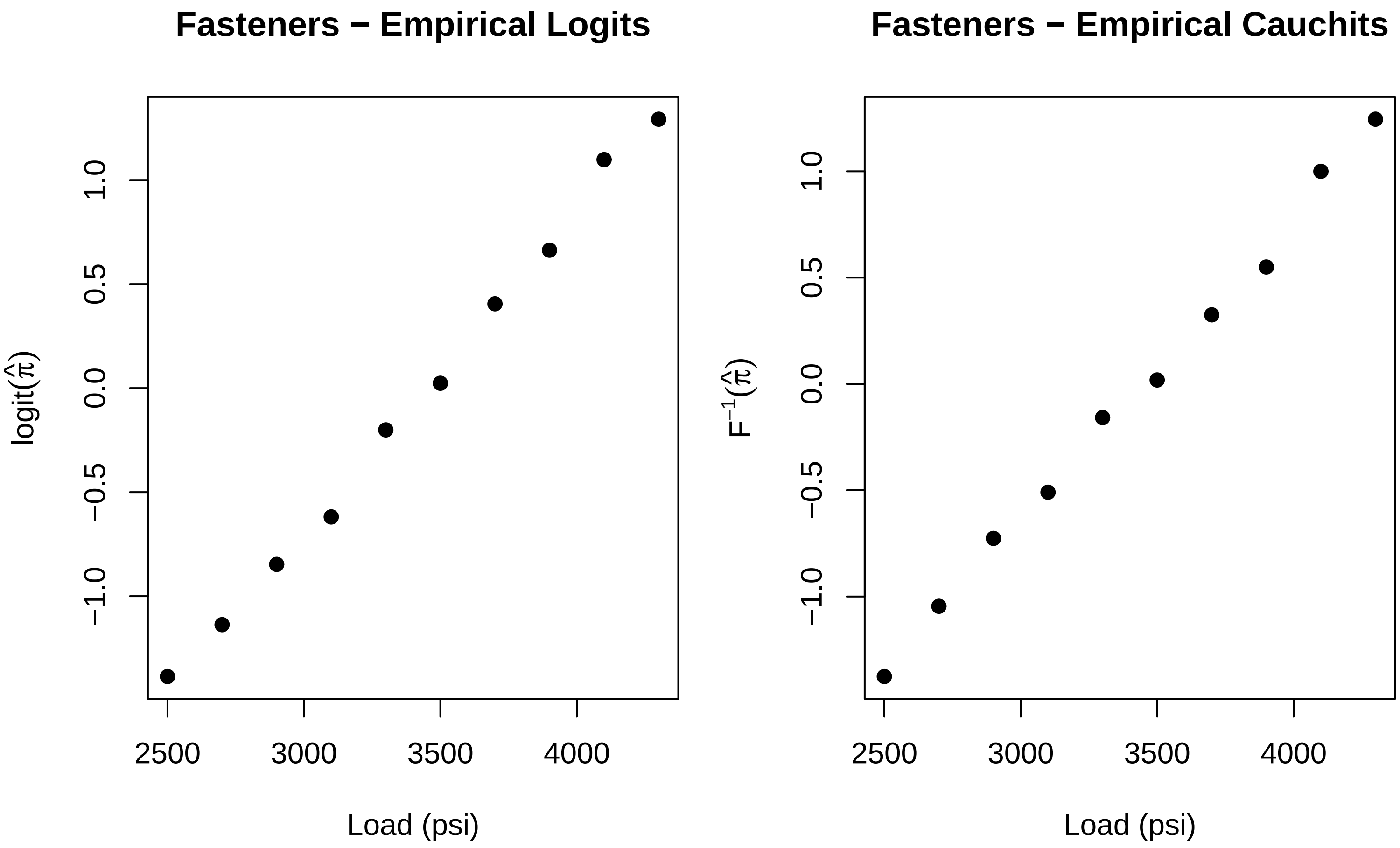
Note that this idea can be used for any link function. For example, if you were doing a probit regression, plot $\Phi^{-1}(\hat{p}_i)$ vs x_i and check for linearity.





Note it is hard to differentiate between logistic and probit regression here.

Similarly for Cauchit regression, plot $F^{-1}(\hat{p}_i)$ vs x_i.



where

$$F^{-1}(u) = \tan(\pi(\mu - 1/2))$$

Note that there can be a slight problem with this approach. Suppose that m_i is small and that $Y_i = 0$. In this case $\hat{p}_i = 0$ and $\text{logit}(0) = -\infty$, so this would be a bit difficult to plot.

Of course you get a similar problem when $Y_i = m_i$.

One solution is to tweak the data slightly for the plot. One idea is to add a small constant to the number of successes and number of failures in each observation. A common choice is to use 0.5, which gives

$$\widehat{\text{logit}_i} = \log \frac{Y_i + 0.5}{m_i - Y_i + 0.5}$$

This tends to shrink the empirical logits a bit towards 0, with more shrinkage occurring with the smaller sample sizes.

Note that this adjustment is only done for exploratory plotting, and not fitting the data. Maximum likelihood has no problem with $Y_i = 0$ or m_i.

Residual Analysis

As with linear regression, trying to find problems with plots of $\mathrm{logit}(\hat{p}_i)$ vs x_i can be difficult. As in linear regression, residual plots tend to be more useful.

The question is, what should we use for residuals. There are two common choices in binomial regression

- Deviance Residual

$$Dres_i = \mathrm{sign}(Y_i - m_i\hat{\pi}_i)\sqrt{2\left\{Y_i \log\frac{Y_i}{m_i\hat{\pi}_i} + (m_i - Y_i)\log\frac{m_i - Y_i}{m_i(1-\hat{\pi}_i)}\right\}}$$

  One motivation for these is the Deviance Goodness-of-Fit test (to be discussed later). These are also the residuals returned in **R** with the command `resid(glmobject)`.

- Pearson Residual

  $$Pres_i = \frac{Y_i - m_i\hat{\pi}_i}{\sqrt{m_i\hat{\pi}_i(1-\hat{\pi}_i)}}$$

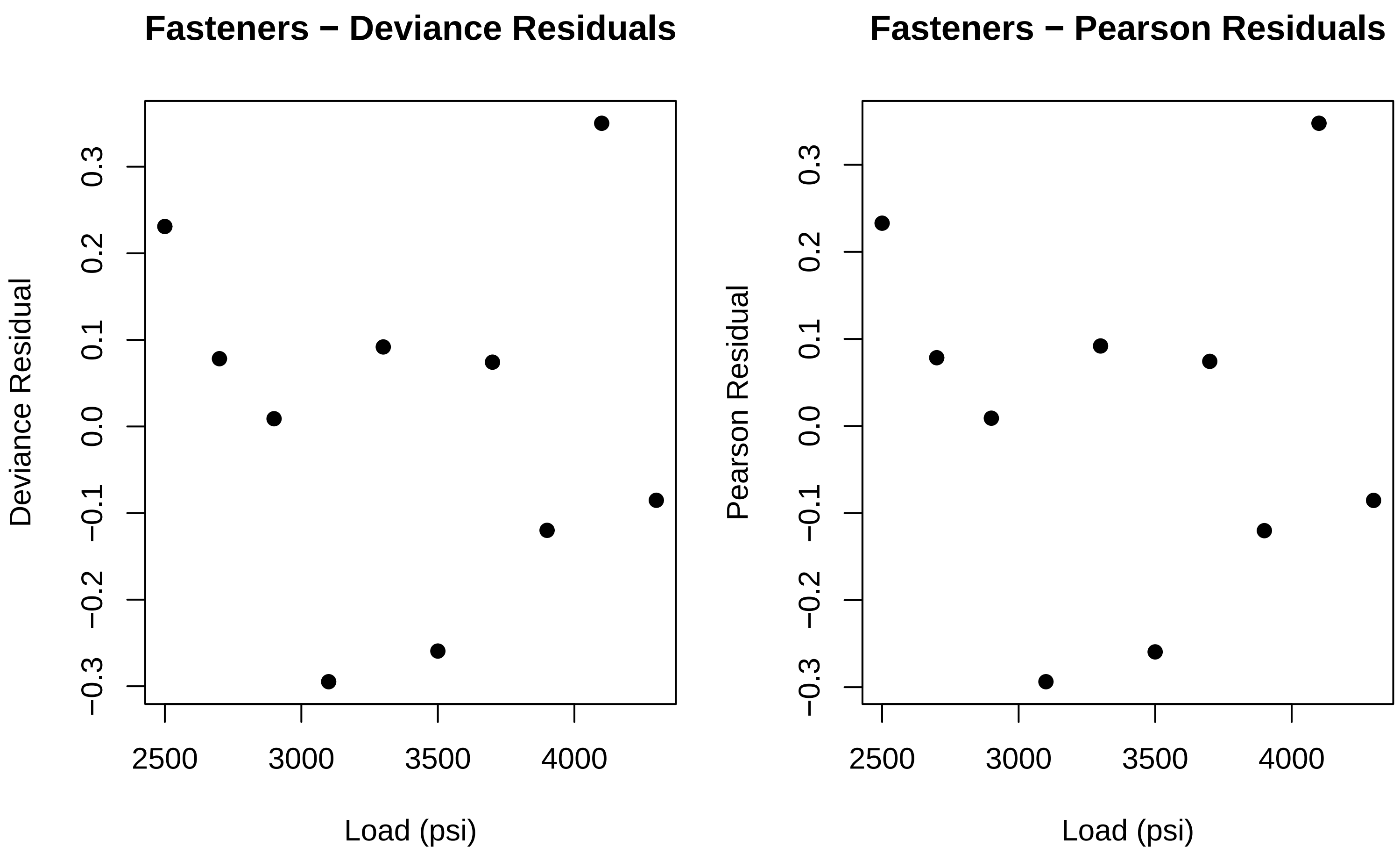
  These have a nice interpretation as

  $$\frac{Observed - Expected}{SE}$$

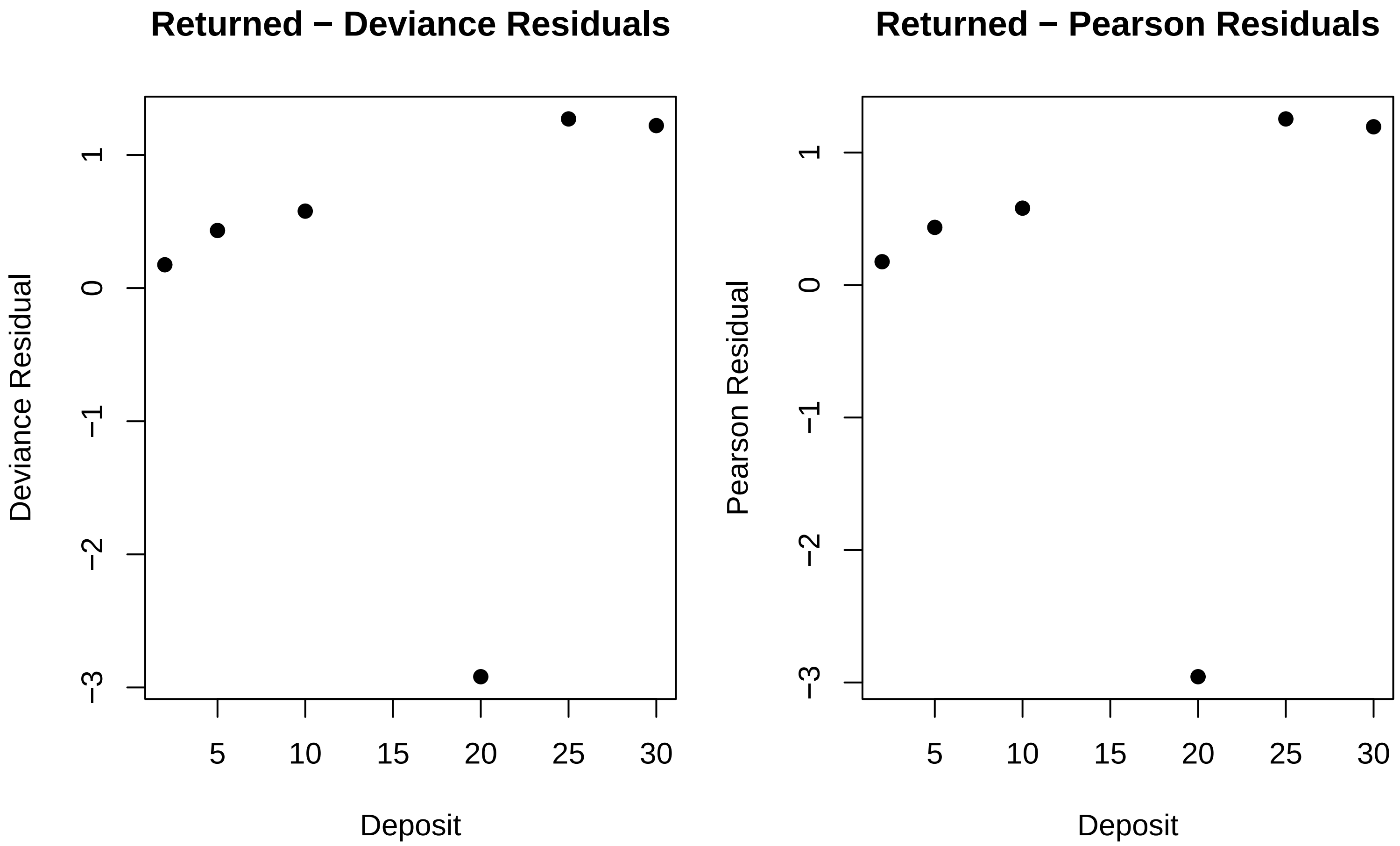
  So you can think of these like standardized residuals in linear regression.

For both type of residuals, if the model is correct, they act like they are draws from a $N(0,1)$ distribution, assuming that the m_is aren't too small.

Thus they can be used to check for problems with the choice of mean function and for outliers.



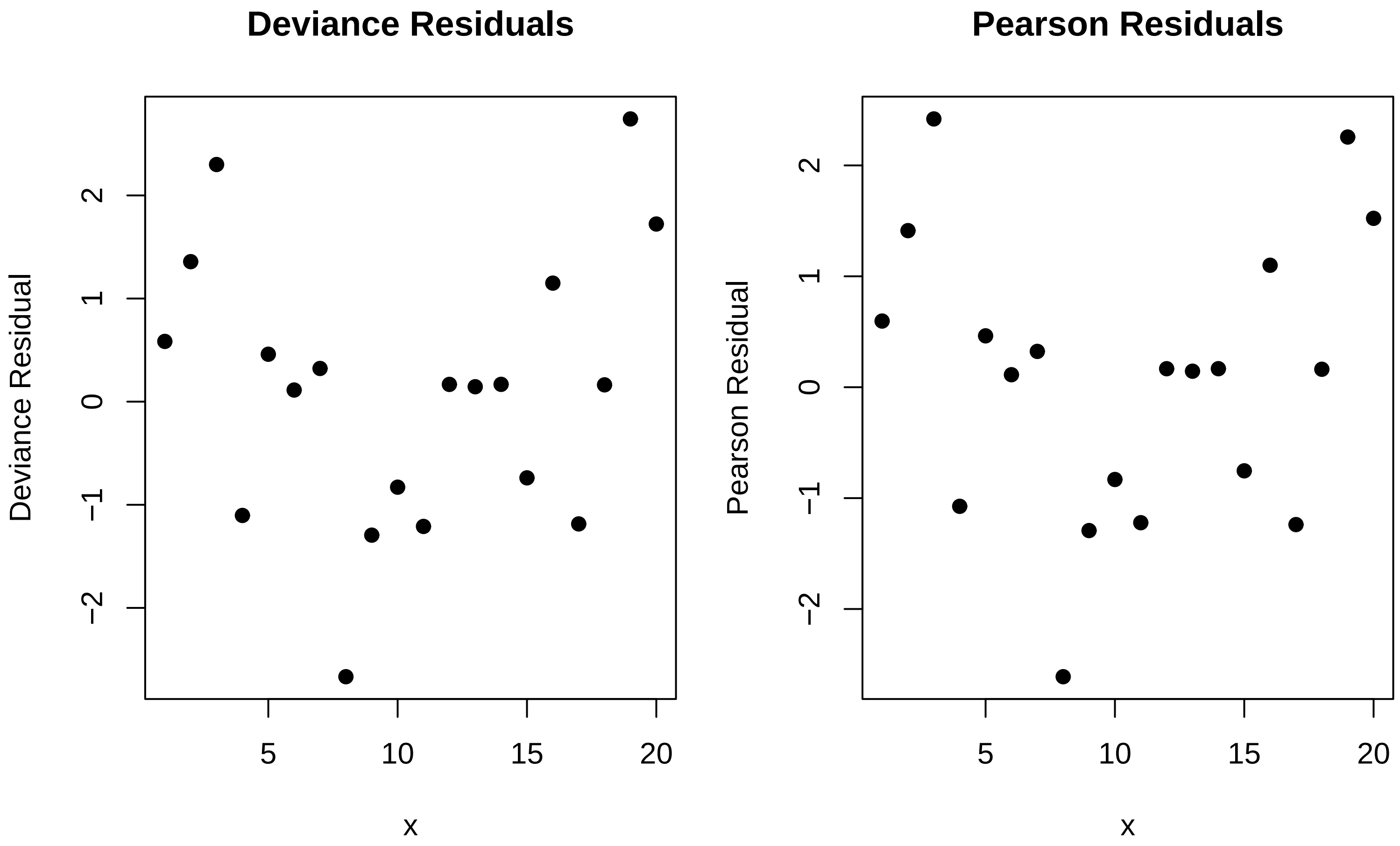
Everything seems to look nice here.



Returned − Deviance Residuals

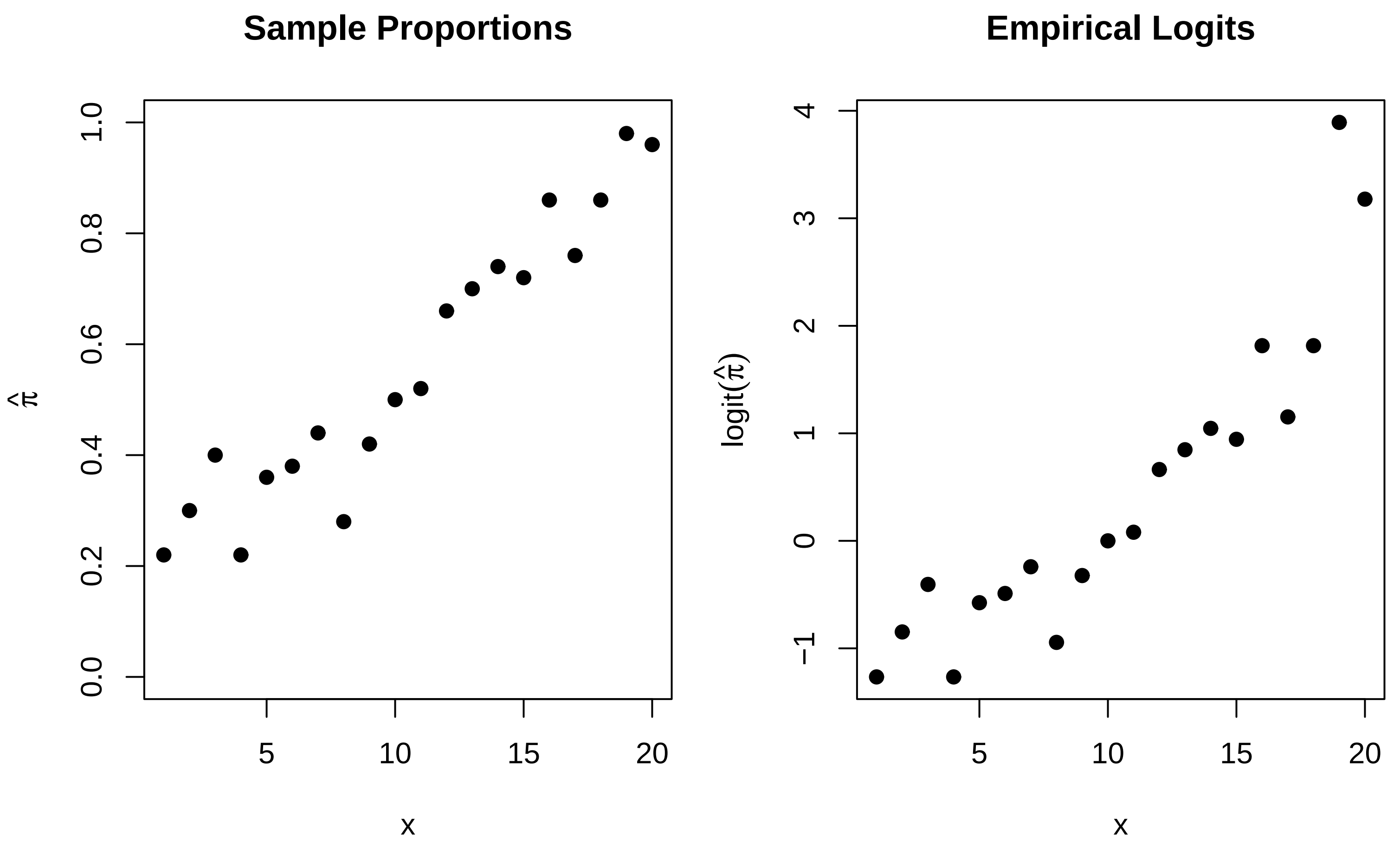
Returned − Pearson Residuals

So it looks like there is one outlier here.

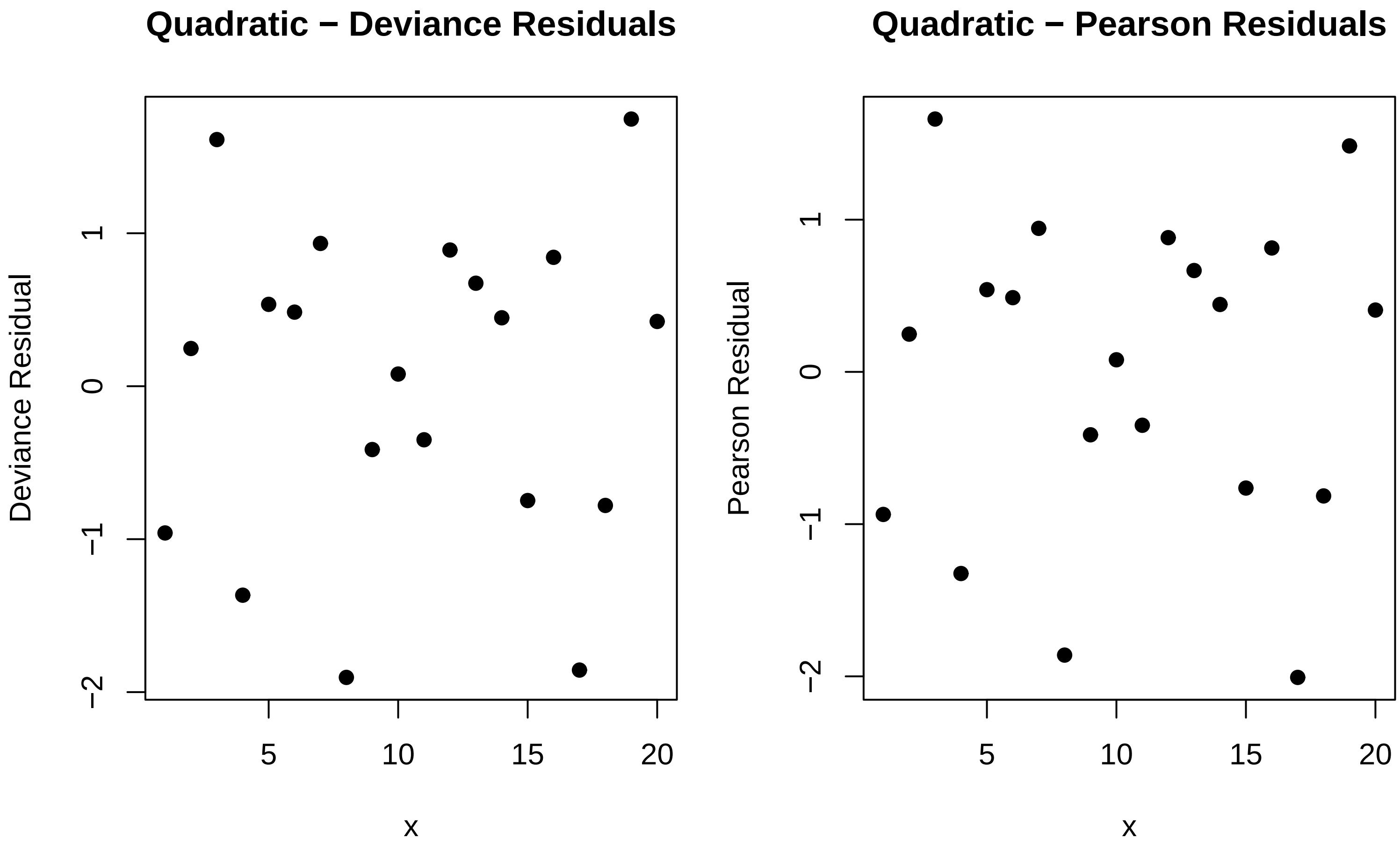


Deviance Residuals

Pearson Residuals



Looks like there may be some curvature here. Lets try fitting a quadratic.



The residuals look much better here.

```
> ysim2.glm <- glm(ymat ~ x + I(x^2), family=binomial())
> summary(ysim2.glm)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9038  -0.7555   0.3351   0.7157   1.7468

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.951140   0.240452  -3.956 7.63e-05 ***
x           -0.006128   0.055554  -0.110 0.912168
I(x^2)       0.009899   0.002796   3.540 0.000399 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 265.391  on 19  degrees of freedom
Residual deviance:  20.766  on 17  degrees of freedom
AIC: 106.21
```

Goodness of Fit Tests

One thing that would be nice is to get more evidence on whether a model actually fits the data that just what we can get from the residual analysis. When the m_i aren't too small, there are a couple of tests that we can do to examine this.

- Deviance Goodness-of-Fit Test

  What we really are interested in examining is the null hypothesis

$$H_0 : \text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i,p-1}$$

  In this null hypothesis some of the βs could be 0. We just don't want missing terms, such as a missing predictor or an x_j^2 type term.

  One possible alternative to compare this null with is

$$H_A : \text{logit}(\pi_i) = \alpha_i \qquad (i = 1, \ldots, n, \text{ with } n \text{ different parameters})$$

This model is sometimes referred to as the saturated model.

As these are nested models, we can do a drop of deviance test to see whether there is evidence that the hypothesized logistic model is adequate or not.

Under H_0,

$$\log L(\hat{\beta}) = C + \sum_{i=1}^{n} Y_i \log \hat{\pi}_i + (m_i - Y_i) \log(1 - \hat{\pi}_i)$$

and under H_A,

$$\log L(\hat{\alpha}) = C + \sum_{i=1}^{n} Y_i \log \hat{p}_i + (m_i - Y_i) \log(1 - \hat{p}_i)$$

So the drop in deviance test statistic is

$$\begin{aligned}
X^2 &= -2(\log L(\hat{\beta}) - \log L(\hat{\alpha})) \\
&= -2 \sum_{i=1}^{n} \{(Y_i \log \hat{\pi}_i + (m_i - Y_i) \log(1 - \hat{\pi}_i)) \\
&\qquad\qquad - (Y_i \log \hat{p}_i + (m_i - Y_i) \log(1 - \hat{p}_i))\} \\
&= 2 \sum_{i=1}^{n} Y_i \log \frac{\hat{p}_i}{\hat{\pi}_i} + (m_i - Y_i) \log \frac{1 - \hat{p}_i}{1 - \hat{\pi}_i} \\
&= 2 \sum_{i=1}^{n} Y_i \log \frac{Y_i}{m_i \hat{\pi}_i} + (m_i - Y_i) \log \frac{m_i - Y_i}{m_i - m_i \hat{\pi}_i}
\end{aligned}$$

This is compared to a χ^2_{n-p} distribution.

Note that this is sometimes referred to as the likelihood ratio goodness-of-fit test since it is a likelihood ratio test.

This statistics has a tie with the deviance residuals as

$$X^2 = \sum_{i=1}^{n} Dres_i^2$$

This test is easily conducted in **R**. The line for Residual Deviance in the `summary(glmobject)` gives information for this statistic.

For the example examined today

```
> summary(fasten.logit.glm)

    Null deviance: 112.83207  on 9  degrees of freedom
Residual deviance:   0.37192  on 8  degrees of freedom

> pchisq(deviance(fasten.logit.glm),
    df.residual(fasten.logit.glm), lower.tail=F)
[1] 0.999957
```

```
> summary(deposit.glm)

    Null deviance: 1108.171  on 5  degrees of freedom
Residual deviance:   12.181  on 4  degrees of freedom

> pchisq(deviance(deposit.glm), df.residual(deposit.glm),
    lower.tail=F)
[1] 0.01605229

> summary(ysim.glm)

    Null deviance: 265.391  on 19  degrees of freedom
Residual deviance:  33.822  on 18  degrees of freedom

> pchisq(deviance(ysim.glm), df.residual(ysim.glm),
    lower.tail=F)
[1] 0.01324954
```

```
> summary(ysim2.glm)

    Null deviance: 265.391  on 19  degrees of freedom
Residual deviance:  20.766  on 17  degrees of freedom

> pchisq(deviance(ysim2.glm), df.residual(ysim2.glm),
    lower.tail=F)
[1] 0.2369435
```

Note that you can have significant parameters in models that don't fit. For example, the quadratic example when only a linear term is fit, showed significant lack of fit. However the linear term was still significant.

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

Model: binomial, link: logit

Response: ymat

Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    19    265.391
x     1  231.569        18     33.822 2.711e-52
```

In this case, the linear term described much of the variability in the counts, but there was still some left to be explained by the quadratic term.

```
> anova(ysim2.glm, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: ymat

Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                      19    265.391
x       1  231.569        18     33.822 2.711e-52
I(x^2)  1   13.056        17     20.766 3.023e-04
```

Note that it is possible to have a model that doesn't show significant lack of fit, but can still have new variables added to the model that show statistical significant.

There is another way to think of this test. Consider the 2 $\times n$ table of observed counts

Y_1	Y_2	$\cdots$	Y_{n-1}	Y_n
$m_1 - Y_1$	$m_2 - Y_2$	$\cdots$	$m_{n-1} - Y_{n-1}$	$m_n - Y_n$
m_1	m_2	$\cdots$	m_{n-1}	m_n

and the corresponding table of expected counts, where the expected counts come from the logistic regression model

$m_1\hat{\pi}_1$	$m_2\hat{\pi}_2$	$\cdots$	$m_{n-1}\hat{\pi}_{n-1}$	$m_n\hat{\pi}_n$
$m_1 - m_1\hat{\pi}_1$	$m_2 - m_2\hat{\pi}_2$	$\cdots$	$m_{n-1} - m_{n-1}\hat{\pi}_{n-1}$	$m_n - m_n\hat{\pi}_n$
m_1	m_2	$\cdots$	m_{n-1}	m_n

So we can consider this Goodness-of-Fit test as comparing the observed counts with the expected counts with the statistic

$$X^2 = \sum_{\text{all cells}} 2O_i \log \frac{O_i}{E_i}$$

- Pearson Goodness-of-Fit Test

A common way of examining goodness of fit is with a Pearson Chi-square test. We can do the same thing here.

Working with the same observed and expected table, Pearson's Chi-square test has the form

$$X_p^2 = \sum_{\text{all cells}} \frac{(O_i - E_i)^2}{E_i}$$

This statistic is also compared to a χ^2_{n-p} distribution.

As with the Deviance Goodness-of-Fit test, this statistic can be tied residuals, Pearson residuals in this case as

$$X_p^2 = \sum_{i=1}^{n} Pres_i^2$$

Usually the test statistics give similar results. For the four examples considered

Test	Fastener	Deposit	Simulated - Linear	Simulated - Quadratic
X^2	0.372	12.19	33.82	20.77
X_p^2	0.371	12.29	31.18	20.35

Note that both of these tests require that the m_i to be large.

To exhibit what can happen in this case, lets consider the situation where $Y_i \overset{iid}{\sim} Bin(1, \pi)$.

In this case $\hat{\pi} = \bar{y}$ giving

$$X_p^2 = \sum \frac{(Y_i - \bar{y})^2}{\bar{y}(1 - \bar{y})} = n$$

and

$$X^2 = -2n \left\{ \bar{y} \log \bar{y} + (1 - \bar{y}) \log(1 - \bar{y}) \right\}$$

In the first case the distribution is degenerate and in the second it strongly depends on $\hat{\pi}$. For these to be valid goodness of fit tests, we need that the distribution not to depend on the parameter estimates (at least not strongly). This will be the case if the m_i are big.