Compare observed and fitted proportions

- Categorize/Group the continuous $x$-variables into intervals.
- Calculate the observed rate of success in each interval.
- Compare with the estimated proportion of success from the model.

![Observed and predicted probabilities](chart)

See towards the end of `f8a.R` for an example.
Pearson residuals

Simple standardization, since \( Y_i \sim Bin(1, p_i) \) with \( E(Y_i) = p_i \) and \( V(Y_i) = p_i(1 - p_i) \):

\[
\tilde{r}_i = \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}} \quad (\sim N(\cdot, \cdot) \text{ not even asymptotically!})
\]

In R: `infl<-influence(model)` then `infl$pear.res`.

However in general the problem with residual analysis for logistic regression is that such plots are not very revealing because of the binary nature of \( Y \).

Also the Leverage values, i.e. the diagonal elements of \( P = X(X'WX)^{-1}X'W \) are now depending both on \( X \) and \( Y \) and as such these are no more indicators of outliers w.r.t \( X \).
Residuals in logistic regression

**Standardized residuals**

Take leverages $v_{ii}$ from the diagonal elements of $P = X(X'WX)^{-1}X'W$:

$$ r_i = \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)(1 - v_{ii})}} \approx N(0, 1) \quad \text{(for large } n) $$

If $|r_i| > |\lambda_{\alpha/2}|$ it might be considered suspiciously large.

Plots of $r_i$ vs $i$ can be useful, although it’s sometimes more revealing to plot their squares, e.g. $r_i^2$ vs $i$.

Notice $v_{ii}$ can be obtained using `infl<-influence(model)` then `infl$hat`. 
Residual plots

The residuals in logistic regression always have a pattern but with few extreme values!

: Data, $\hat{p}_i$ and 95% CI

: $r_i^2$

: $r_i$ and 95% CI
Cook’s distance

There is a version of Cook’s distance for logistic regression:

\[ D_{i}^{Cook} = \frac{r_i^2}{p + 1} \cdot \frac{v_{ii}}{1 - v_{ii}} \]

Hosmer & Lemeshow consider influential cases those with

\[ D_{i}^{Cook} > 1. \]