Goodness of fit: problems

- The deviance has unclear asymptotic distribution when:
  - There are continuous predictors
  - The data is ungrouped

- The horseshoe crab dataset: what happens when the number of crabs observed grows?
Solutions

• Possible solutions:
  – Comparison with models that have additional terms
  – Artificial grouping

• The second approach leads to Hosmer-Lemeshow statistic

• The idea -group the data according to the FIT-TED probabilities

• There are $i$ partition groups, $i = 1, \ldots, g$ and $j$ observations in each group, $j = 1, \ldots, n_j$
• Let $y_{ij}$ be the binary outcome

• Then compute the $X^2$ based on this grouping:

$$X^2_{SM} = \sum_{j=1}^{g} \frac{(\sum_j y_{ij} - \hat{\pi}_{ij})^2}{(\sum_j \hat{\pi}_{ij}) \left[ 1 - \left( \sum_j \hat{\pi}_{ij} \right) / n_i \right]} \quad (1)$$
Solutions

- Based on simulations, the distribution is approximately $\chi^2_{g-2}$ for $g$ groups of data
- Does not have good power for detecting particular type of lack of fit
Models with Categorical Predictors

- ANOVA-type model:
  \[ \log \frac{\pi_i}{1 - \pi_i} = \alpha + \beta_i \]

- \( I \) categories and \( I - 1 \) nonredundant parameters

- The meaning of \( \beta_i \)
Dummy variables and coding

- \[
    \log \frac{\pi_i}{1 - \pi_i} = \alpha + \beta_1 x_1 + \ldots + \beta_{I-1} x_{I-1}
    \]

- Coding: treatment vs. effect coding
- Interpretation: differences only are meaningful
- Ordinal factor implies linear logit model:
  \[
  \log \frac{\pi_i}{1 - \pi_i} = \alpha + \beta x_i
  \]
Cochran-Armitage test

- Serves to test independence between ordered categories
- Based on linear probability model:
  \[ \pi_i = \alpha + \beta x_i \]
- Pearson statistic
  \[
  \chi^2_I = \frac{1}{p(1-p)} \sum n_i (p_i - p)^2
  = \frac{1}{p(1-p)} \sum n_i (p_i - \hat{\pi}_i)^2
  + \frac{b^2}{p(1-p)} \sum n_i (x_i - \bar{x})^2
  = \chi^2_L + z^2
  \]
- Partitioned \( \chi^2_I \): \( \chi^2_L \) tests goodness of fit, \( z^2 \) measures independence
Multiple logistic regression

• Model

\[
\log \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta_1 x_1 + \ldots + \beta_p x_p
\]

• Data is displayed by the multiway contingency table

• Conditional independence of the two variables

• No interaction between categorical variables: additivity on the logit scale

• AIDS-AZT example

<table>
<thead>
<tr>
<th>Race</th>
<th>AZT use</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>Yes</td>
<td>14</td>
<td>93</td>
</tr>
<tr>
<td>White</td>
<td>No</td>
<td>32</td>
<td>81</td>
</tr>
<tr>
<td>Black</td>
<td>Yes</td>
<td>11</td>
<td>52</td>
</tr>
<tr>
<td>Black</td>
<td>No</td>
<td>12</td>
<td>43</td>
</tr>
</tbody>
</table>
Goodness of fit

- Likelihood-ratio Statistic:
  \[ G^2(M_0|M_1) = -2(L_0 - L_1) \]
  \[ = \text{Deviance}(M_0) - \text{Deviance}(M_1) \]

- Useful for model comparison