Logit models with Categorical predictors:

1. Dummy variables in logit models
2. Example: Alcohol and Infant Malformation
3. Cochran-Armitage Trend Test

Homework: Ch.5, Sec. 3
Problems: 5.5, 5.6, 5.9
 Dummy variables in Logit Models

Let us consider the response $Y$ with $Y=1$ or $Y=0$ outcomes, and a factor $X$ with $I$ categories.

Consider $I \times 2$ table

<table>
<thead>
<tr>
<th></th>
<th>$Y=1$</th>
<th>$Y=0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$Y_1$</td>
<td>$n_1-Y_1$</td>
</tr>
<tr>
<td>2</td>
<td>$Y_2$</td>
<td>$n_2-Y_2$</td>
</tr>
<tr>
<td></td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$I$</td>
<td>$Y_I$</td>
<td>$n_I-Y_I$</td>
</tr>
</tbody>
</table>

$Y_i$ are independent $\text{Bin}(1, n_i, \pi_i)$ with $\pi_i = \Pr(Y=1 | X=i)$

Logit model:

$$\log \left( \frac{\pi_i}{1-\pi_i} \right) = \alpha + \beta_i, \quad i=1, 2, \ldots, I$$
Assume for simplicity, that say, $\beta_i = 0$

\[ \logit(\pi_i) = \alpha + \beta_i, \quad i = 1, 2, \ldots, I-1 \]

So that

\[ \alpha = \logit(\pi_i) \]

and

\[ \beta_i = \logit(\pi_i) - \logit(\pi_1) \]

\[ = \log \left( \frac{\pi_i / (1 - \pi_i)}{\pi_1 / (1 - \pi_1)} \right) \]

is the log odds ratio of the pair of rows $i \neq 1$

i.e.

\[ \beta_i = \log \frac{\pi_i}{1 - \pi_i}, \quad \beta_i = \frac{\pi_i}{1 - \pi_i} \]

\[ \frac{\beta_i}{\beta_i} = e \]
Now let us use the dummy variables ($x_i = 1$ for observation in row $i$)
for rewriting our logit model:

$$\text{logit} (\pi_i) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{i-1} x_{i-1},$$

where

$x_i = 1$ for observation in row $i$,
and

$x_i = 0$, otherwise.

Remark #1. One can define the $X = 1$ for category 1,
and $X = -1$ for 2nd category.

i.e. $\beta_1 + \beta_2 = 0$

Note that

$$\exp (\hat{\beta}_2 - \hat{\beta}_1)$$

is the estimated odds of success ($Y = 1$)
in category a of $X$
divided by the estimated odds of success in category b of $X$. 
2. Alcohol & Infant Malformation

<table>
<thead>
<tr>
<th>Alcohol Consump.</th>
<th>Present Y=1</th>
<th>Absent Y=0</th>
<th>Observed</th>
<th>( \chi^2 )</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>48</td>
<td>17066</td>
<td>17114</td>
<td>48</td>
<td>.0028</td>
</tr>
<tr>
<td>&lt;1</td>
<td>38</td>
<td>14464</td>
<td>14502</td>
<td>38</td>
<td>.0026</td>
</tr>
<tr>
<td>1-2</td>
<td>5</td>
<td>788</td>
<td>793</td>
<td>.0063</td>
<td></td>
</tr>
<tr>
<td>3-5</td>
<td>1</td>
<td>126</td>
<td>127</td>
<td>.0079</td>
<td></td>
</tr>
<tr>
<td>&gt;6</td>
<td>1</td>
<td>37</td>
<td>38</td>
<td>.0263</td>
<td></td>
</tr>
</tbody>
</table>

\[
\chi^2 = \frac{\sum (O - E)^2}{E} = 32.481 \\
h = 32.574
\]

Testing Hypothesis \( H_0 : \beta_i = 0, \ i=1,2,\ldots,5 \)

The sample logits:

\[
Y_i = \begin{cases} 1 & \text{if } \text{Present} \\ 0 & \text{if } \text{Absent} \end{cases}
\]

\[
\log \frac{Y_i}{1-Y_i} = \alpha + \beta_i
\]

\[
\begin{array}{c|c|c}
\hline
Y_i & \log \frac{Y_i}{1-Y_i} & \alpha + \beta_i \\
\hline
5.87 & -5.87 & \\
5.94 & -5.94 & \\
5.06 & -5.06 & \\
4.84 & -4.84 & \\
3.61 & -3.61 & \\
\hline
\end{array}
\]

\[
\text{log}_e \frac{48}{11066} = \left[ -5.87 \right] \ i=1
\]
Test Statistic: chi-squared

\[ X^2 = \sum_{i,j} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \]

\[ X^2 \quad \text{d.f.} = 4 \]

Since \((2-1) \times (5-1) = 4\)

Here

\[ \hat{\mu}_{ij} = \frac{n_{i+} \cdot n_{+j}}{n} \]

with

\[ \hat{\mu}_{cc} = \frac{17114 \times 93}{32574} = 48.86 \]

Observed

\[ X^2_{\text{obs}} = \frac{(48 - 48.86)^2}{48.86} = 12.1 \]

\[ = 12.1 \]

\[ p\text{-value} = P(X^2 \geq 12.1) \approx 0.02 \]

Let \( G^2 = \sum n_{ij} \log \frac{n_{ij}}{\hat{\mu}_{ij}} \)

And \( G^2_{\text{obs}} = 6.2 \) with \( p\text{-value} = 0.19 \)
Linear Logit Model

Consider the ordered categories.

Denote now \((x_1, x_2, \ldots, x_I)\)

the scores which describe the distances between categories of \(X\).

If we expect a monotone effect of \(X\) on \(Y\), then

\[
\text{logit} (u_i) = x + \beta x_i, \quad i=1,2,\ldots,I
\]

Let us show that this model fits better than independent model with \(\beta_i = 0\) in the Alcohol & Infant Malform. example.

Let \(x_1 = 0, x_2 = 0.5, x_3 = 1.5, x_4 = 4, x_5 = 7\)
The estimated multiplicative effect
\[ e^{\hat{\beta}} = e^{0.3166} = \boxed{1.37} \]

A unit increase in daily alcohol consumption leads to an increase of the odds of malformation by 37%.

**Testing Hypothesis**

\[ H_0: \text{logit}(\pi_i) = \alpha + \beta x_i \quad \text{with} \quad \beta > 0 \]

**Test Statistic:**

\[ X^2 \quad \text{d.f.} = 4 - 1 \quad \text{One parameter} \quad (\beta = 3 \quad \text{have to be estimated}) \]

\[ \hat{\beta} = 0.3166 \]

\[ X^2_{obs} = 2.05 \quad \text{p-value} \quad \text{is very large} \]

\[ x \quad \text{2.05} \]
3. **Cochran–Armitage Trend Test**

Let us consider again

A 1×2 table

with ordered rows (categories)

and I independent r.v.'s

\[ Y_i \sim \text{Bin}(n_i, \pi_i) \quad i = 1, 2, \ldots, I \]

**Null Hypothesis:**

\[ H_0: \text{independence of } X \text{ and } Y. \]

**Model:**

\[ \pi_i = \alpha + \beta x_i. \]

**Under** \( H_0, \beta = 0. \)

**Notations:**

\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{I} n_i x_i \]

\[ p_i = \frac{Y_i}{n_i}, \quad p = \frac{\sum Y_i}{\sum n_i} = \frac{\sum p_i}{n}. \]

Using least squares method:

\[ \hat{\pi}_i = p + b(x_i - \bar{x}). \]
where
\[
\beta = \frac{\sum_i n_i (p_i - \bar{p})(x_i - \bar{x})}{\sum_i n_i (x_i - \bar{x})^2}
\]

Denote the Pearson statistic for testing independence by
\[
X^2(I) = \frac{1}{p(1-p)} \sum_{i=1}^{I} n_i (p_i - \bar{p})^2 = \bar{z}^2 + X^2(L)
\]

with
\[
X^2(L) = \frac{1}{p(1-p)} \sum_{i=1}^{2} n_i (p_i - \pi_i)^2
\]

and
\[
\bar{z}^2 = \frac{\beta^2}{p(1-p)} \sum_{i=1}^{I} n_i (x_i - \bar{x})^2 = \left[ \frac{\sum (x_i - \bar{x})Y_i}{\sqrt{p(1-p)} \sum n_i (x_i - \bar{x})^2} \right]^2
\]
The statistic

\[ X^2(L) \sim \chi^2 \text{ with d.f. } = I - 2 \]

can be used to fit the model

\[ \pi_i = \alpha + \beta x_i \]

and

Statistic

\[ Z^2 \sim \chi^2 \text{ with d.f. } = 1 \]

can be used to test

\[ H_0 : \beta = 0 \]

for the linear trend in our model.

\[ Z^2 \] is called the

Cochran-Homitage Trend Test
Connections with $M^2$ test (score test)

$M_n^2 = \frac{(n-1) r^2}{n-1}$

with

$r = \text{the empirical correlation coeff.}

\text{Namely, when testing } H_0: \beta = 0

Z^2 = \text{the score test } M_n^2

\text{with } J=2 \text{ in } I \times J
\text{contingency table}

Note also that when } I=2

X^2(L) = 0

and

Z^2 = X^2(I)

Consider again Example: Alcohol & Infant Malformation

X^2(I) = 12.1

but the Cochran-Armitage trend test

$Z^2 = 6.6 \text{ with } p\text{-value } = .01$

$df = 1$

$H_0$ is rejected

i.e. (B > 0)
From output Table 5.4

Wald Statistic

\[ T^2 = \left( \frac{\hat{\beta}}{SE(\beta)} \right)^2 = \left( \frac{3.166}{.1254} \right)^2 = \]

\[ = 6.4 \]

p-value = .012

i.e. \( T^2 \) and \( Z^2 \) are similar in this example when testing \( H_0: \beta = 0 \)

in the linear logit model
95% Wald CI for $\beta$:

$$0.317 \pm 1.96 \times (0.125)$$

$$= (0.07, 0.56)$$

Likelihood Ratio 95% CI for $\beta$:

$$(-0.0187, 0.5236)$$

is better than Wald's CI.