Example 1.2: Fever in early pregnancy and risk of fetal death

- The Danish National Birth Cohort Study recruited pregnant women 1997-2002 for telephone interviews scheduled to take place in gestational weeks 12-16.
- Here: data on women recruited before 31 March 1999, interviewed before week 17, and who were still pregnant at week 17.
- Study relation between risk of fetal death (or risk of “small for gestational age”, SGA) and explanatory variables
- Outcome variables (fetal death and SGA): binary,
- Both categorical and quantitative explanatory variables relevant.

Models for binary data

So far, the risk of a failure, \( p_i = \text{pr}(y_i = 1 \mid x_i) \) for given covariates, \( x_i \) has been described using the logit link:

\[
\text{logit}(p_i) = \log \left( \frac{p_i}{1 - p_i} \right) = a + bx_i.
\]

This provided \( b \)-estimates which are \( \log(\text{odds ratios}) \) and predicted probabilities

\[
p_i = \frac{\exp(a + bx_i)}{1 + \exp(a + bx_i)},
\]

which stayed nicely between 0 and 1.

Figure 1: Left: the logit function \( \ell = \text{logit}(p) = \log(p/(1 - p)) \). Right: the logistic function \( p = \exp(\ell)/(1 + \exp(\ell)) \).
Other link functions

Logistic regression

\[ \text{logit}(p_i) = LP_i \]

is by far the most common regression model for binary data. However, other link functions are sometimes used:

\[
\begin{align*}
\text{probit}(p_i) & = LP_i, \\
\log(p_i) & = LP_i, \\
p_i & = LP_i.
\end{align*}
\]

Here, \(LP_i\) is the linear predictor and \(\text{probit}(p)\) the \(p\)th percentile in the standard Normal distribution (e.g. \(z_{0.975} = 1.96\) if \(p = 0.975\)).

Comments to link functions

The logit and probit models can be thought of as arising from a “latent” variable, \(y_i^*\), “i’s health”, such that \(y_i = 1\) whenever \(y_i^*\) is below some threshold, \(c\): \((y_i = 1) \iff (y_i^* \leq c)\).

For the probit model the distribution of the latent variable is standard Normal, for the logit model it follows a standard “logistic” distribution (these two distributions have quite similar shapes).

The probit model also has the nice property that \(0 < p_i < 1\).

The log and identity links do not have this property.

Then why use them?

Because estimates have simple interpretations as log(relative risks) or risk differences.

Data

Table 1: Fever in pregnancy study: distribution of fetal death by smoking and by number of fever episodes in early pregnancy.

<table>
<thead>
<tr>
<th>Category</th>
<th>Women</th>
<th>Fetal Deaths</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>No smokers</td>
<td>8647</td>
<td>81</td>
<td>0.94</td>
</tr>
<tr>
<td>1–10 cigarettes/day</td>
<td>1760</td>
<td>19</td>
<td>1.08</td>
</tr>
<tr>
<td>11+ cigarettes/day</td>
<td>1371</td>
<td>19</td>
<td>1.39</td>
</tr>
<tr>
<td>No fever episodes</td>
<td>9693</td>
<td>98</td>
<td>1.01</td>
</tr>
<tr>
<td>1 fever episode</td>
<td>1872</td>
<td>20</td>
<td>1.07</td>
</tr>
<tr>
<td>2 fever episodes</td>
<td>183</td>
<td>1</td>
<td>0.55</td>
</tr>
<tr>
<td>3+ fever episodes</td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>11778</td>
<td>119</td>
<td>1.01</td>
</tr>
</tbody>
</table>

Results from simple models with 1 covariate

Table 2: Fever in pregnancy study: effects on fetal death of smoking and number of fever episodes in early pregnancy estimated using different link functions.

<table>
<thead>
<tr>
<th>Link Function</th>
<th>Logit</th>
<th>Probit</th>
<th>Log</th>
<th>Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 cigs/day</td>
<td>0 (ref)</td>
<td>0 (ref)</td>
<td>0 (ref)</td>
<td>0 (ref)</td>
</tr>
<tr>
<td>1–10 cigs/day</td>
<td>0.143 (0.256)</td>
<td>0.053 (0.096)</td>
<td>0.142 (0.254)</td>
<td>0.014 (0.003)</td>
</tr>
<tr>
<td>11+ cigs/day</td>
<td>0.396 (0.257)</td>
<td>0.150 (0.098)</td>
<td>0.392 (0.253)</td>
<td>0.0045 (0.003)</td>
</tr>
<tr>
<td>Intercept</td>
<td>-4.66 (0.11)</td>
<td>-2.35 (0.041)</td>
<td>-4.67 (0.11)</td>
<td>0.0094 (0.0010)</td>
</tr>
<tr>
<td>0 episodes</td>
<td>0 (ref)</td>
<td>0 (ref)</td>
<td>0 (ref)</td>
<td>0 (ref)</td>
</tr>
<tr>
<td>1 episode</td>
<td>0.056 (0.247)</td>
<td>0.0208 (0.092)</td>
<td>0.055 (0.244)</td>
<td>0.0006 (0.0026)</td>
</tr>
<tr>
<td>2 episodes</td>
<td>-0.620 (1.01)</td>
<td>-0.223 (0.350)</td>
<td>-0.615 (1.00)</td>
<td>-0.0046 (0.0056)</td>
</tr>
<tr>
<td>3+ episodes</td>
<td>-19.78 (35670)</td>
<td>-4.17 (4256)</td>
<td>-19.69 (34198)</td>
<td>-0.0101 (0.0010)</td>
</tr>
<tr>
<td>Intercept</td>
<td>-4.58 (0.10)</td>
<td>-2.32 (0.038)</td>
<td>-4.59 (0.10)</td>
<td>0.010 (0.0010)</td>
</tr>
</tbody>
</table>
Interpreting estimates for smoking

Logit link:
\[
\log\left(\frac{0.0108}{1-0.0108}\right) = 0.143, \quad \log\left(\frac{0.0094}{1-0.0094}\right) = 0.396,
\]

intercept logit(0.0094)=−4.66; log link:
\[
\log\left(\frac{0.0108}{0.0094}\right) = 0.142, \quad \log\left(\frac{0.0139}{0.0094}\right) = 0.392.
\]

Intercept log(0.0094)=−4.67. Identity link:
\[
0.0108 - 0.0094 = 0.0014, \quad 0.0139 - 0.0094 = 0.0045
\]

and the intercept is 0.0094.

Parameters for log and logit are close - Figure!

Interpreting estimates (ctd.)

Probit link: If the latent variable \( y^*_i \) for woman \( i \) (her “health”) is smaller than some threshold \( c \) then woman \( i \) experiences a fetal loss.

In the reference group (no smokers), \( y^*_i \) has a standard Normal distribution (mean 0).

In other smoking categories, “mean health”, that is, the mean value of \( y^*_i \) is reduced by 0.053 and 0.150, respectively.

Intercept −2.35: the 0.94 percentile in the standard Normal distribution corresponding to the relative frequency of fetal deaths in the reference group.

Estimates for fever episodes have a similar interpretation. Note the “strange” results for 3+ versus 0 (except for the model with identity link) due to 0 fetal deaths in the 3+ category.

One quantitative covariate

Study models including mother’s age in years (mean 29.6, SD=4.2, minimum=16):

Table 3: Fever in pregnancy study: effects on fetal death of mother’s age estimated using different link functions.

<table>
<thead>
<tr>
<th>Link Function</th>
<th>Logit</th>
<th>Probit</th>
<th>Log</th>
<th>Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>−4.967</td>
<td>−2.462</td>
<td>−4.974</td>
<td>0.0074</td>
</tr>
<tr>
<td>Age-25 (per 10 years)</td>
<td>0.726</td>
<td>0.269</td>
<td>0.718</td>
<td>0.0059</td>
</tr>
<tr>
<td>(SD)</td>
<td>(0.216)</td>
<td>(0.081)</td>
<td>(0.213)</td>
<td>(0.0018)</td>
</tr>
</tbody>
</table>

Mother’s age significant for all models.
**Interpretation**

Results for logit and log links are almost identical: Odds (risk) increases by \( \exp(0.72) \approx 2 \) for every 10 years increase of age.

Probit model: Latent “health” decreases by 0.269 for every 10 years of age.

Identity link: risk increases by 0.0059 for every 10 years of age.

Note that (since intercept=0.0074 corresponding to age 25 years) the predicted risk

\[
0.0074 + 0.0059(\text{age} - 25)/10
\]

is negative for ages below 13 - not a great problem since minimum age in the data set was 16.

---

**Small for gestational age, SGA**

SGA means birth weight below some percentile for the actual gestational week, here the 5th percentile.

Table 4: Fever in pregnancy study: Distribution of small for gestational age (SGA) by parity and smoking.

<table>
<thead>
<tr>
<th>Cigarettes/Day</th>
<th>0</th>
<th>1–10</th>
<th>11+</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fraction</td>
<td>%</td>
<td>Fraction</td>
<td>%</td>
</tr>
<tr>
<td>Parity 0</td>
<td>223/3635</td>
<td>6.1</td>
<td>77/855</td>
<td>9.0</td>
</tr>
<tr>
<td>Parity 1+</td>
<td>115/4637</td>
<td>0.25</td>
<td>42/830</td>
<td>5.1</td>
</tr>
<tr>
<td>Total</td>
<td>338/8272</td>
<td>4.1</td>
<td>119/1685</td>
<td>7.1</td>
</tr>
</tbody>
</table>

---

**Results**

Table 5: Fever in pregnancy study: effects on small for gestational age of parity and smoking estimated using different link functions.

<table>
<thead>
<tr>
<th>Link Function</th>
<th>Logit</th>
<th>Probit</th>
<th>Log</th>
<th>Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parity 1+ vs. 0</td>
<td>-0.730 (0.089)</td>
<td>-0.344 (0.041)</td>
<td>-0.683 (0.084)</td>
<td>-0.0357 (0.0041)</td>
</tr>
<tr>
<td>Smoking 1–10 vs. 0</td>
<td>0.434 (0.111)</td>
<td>0.253 (0.053)</td>
<td>0.499 (0.103)</td>
<td>0.0271 (0.064)</td>
</tr>
<tr>
<td>Smoking 11+ vs. 0</td>
<td>0.836 (0.112)</td>
<td>0.407 (0.055)</td>
<td>0.768 (0.103)</td>
<td>0.0499 (0.0080)</td>
</tr>
<tr>
<td>LR test for no interaction</td>
<td>10.58</td>
<td>8.1</td>
<td>11.7</td>
<td>1.32</td>
</tr>
</tbody>
</table>

**Interpretation**

Risk of SGA depends significantly on parity and smoking for all four models (in the expected directions).

Note that presence of interaction is scale-dependent - only for the identity link do we see no significant interaction.
How to choose the link function?

Not obvious, but aspects like

• ease of interpretation,
• model fit,
• model simplicity

should be considered.

For case-control studies, there is no choice - only logit link works!

Case-control studies

For rare disease outcomes, a prospective study may be costly. Alternative: retrospective sampling of cases among those with \( y_i = 1 \) and controls among those with \( y_i = 0 \). Ascertain exposure subsequently.

Table 6: The basic two-by-two table for a case-control study with a single binary exposure, \( x \).

<table>
<thead>
<tr>
<th>Group</th>
<th>Controls ((y = 0))</th>
<th>Cases ((y = 1))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unexposed (x = 0)</td>
<td>( c_0 )</td>
<td>( d_0 )</td>
</tr>
<tr>
<td>Exposed (x = 1)</td>
<td>( c_1 )</td>
<td>( d_1 )</td>
</tr>
</tbody>
</table>

A subject with \( y_i = 1 \) is sampled as a case with probability \( q_d \) independent of exposure, and a subject with \( y_i = 0 \) is sampled as a control with probability \( q_c \) independent of exposure.

Case-control studies: estimating odds ratio

As before: \( p_0 = \text{pr}(y_i = 1 \mid x_i = 0) \) and \( p_1 = \text{pr}(y_i = 1 \mid x_i = 1) \), leading to the odds ratio

\[
OR = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}.
\]

The case-control ratio among exposed \( \frac{d_1}{c_1} \) estimates

\[
\frac{p_1 q_d}{(1-p_1)q_c} \]

and that among the unexposed \( \frac{d_0}{c_0} \) estimates

\[
\frac{p_0 q_d}{(1-p_0)q_c},
\]

that is, the ratio between these \( \frac{d_1/c_1}{d_0/c_0} \) estimates

\[
\frac{\frac{p_1 q_d}{(1-p_1)q_c}}{\frac{p_0 q_d}{(1-p_0)q_c}} = OR.
\]
Case-control studies: logistic regression

Let

\[ b = \log(OR) = \log \left( \frac{p_1/(1-p_1)}{p_0/(1-p_0)} \right) \]

be the log(odds ratio) and \( a = \log \frac{p_0}{1-p_0} \) the log(odds) among non-exposed.

If the logistic regression model for the underlying population is

\[ \ell_i = \logit(p_i) = a + bx_i \]

then the resulting model for the sampled case-control data is:

\[ \tilde{\ell}_i = \tilde{a} + bx_i \]

with \( \tilde{a} = a + \log \frac{q_c}{q_d} \).

That is, from the case-control data we can estimate the slope, \( b \), but not the intercept, \( a \) provided that sampling probabilities for both cases and controls do not depend on exposure.

Similarly for multiple logistic regression:

\[ \ell_i = a + b_1 x_{i,1} + \cdots + b_n x_{i,n} \]

where the \( b \)'s but not \( a \) can be estimated if sampling probabilities for both cases and controls do not depend on any of the covariates \( x_1, \ldots, x_n \).

**Warning:** matched case-control studies. If controls are age-matched to controls then sampling probabilities for controls do depend on age, and age effects cannot be estimated.

Example: SGA and parity, smoking

All SGA cases (574) and (approximately) three non-SGA controls per case (1562 controls) were sampled:

Table 7: Fever in pregnancy study: distribution of sga cases and controls by parity and smoking: cases/controls.

<table>
<thead>
<tr>
<th>No Smokers</th>
<th>1–10 Cigarettes/Day</th>
<th>11+ Cigarettes/Day</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parity 0</td>
<td>223/498</td>
<td>77/113</td>
<td>58/86</td>
</tr>
<tr>
<td>Parity 1+</td>
<td>115/668</td>
<td>42/114</td>
<td>59/83</td>
</tr>
<tr>
<td>Total</td>
<td>338/1166</td>
<td>119/227</td>
<td>117/169</td>
</tr>
</tbody>
</table>

Results

Table 8: Fever in pregnancy study: logistic regression models for case-control sample and for the entire dataset.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Case-Control Sample</th>
<th>Full Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parity 0 vs 1+, No Smokers</td>
<td>0.956 (0.129)</td>
<td>0.944 (0.117)</td>
</tr>
<tr>
<td>Smoking 1–10 vs. 0, Parity 0</td>
<td>0.420 (0.168)</td>
<td>0.415 (0.138)</td>
</tr>
<tr>
<td>Smoking 1–10 vs. 0, Parity 1+</td>
<td>0.761 (0.207)</td>
<td>0.740 (0.184)</td>
</tr>
<tr>
<td>Smoking 11+ vs. 0, Parity 0</td>
<td>0.410 (0.188)</td>
<td>0.523 (0.155)</td>
</tr>
<tr>
<td>Smoking 11+ vs. 0, Parity 1+</td>
<td>1.418 (0.198)</td>
<td>1.247 (0.165)</td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.76 (0.10)</td>
<td>-3.67 (0.094)</td>
</tr>
</tbody>
</table>
Results

- Estimated log(odds ratios) quite similar
- Somewhat larger SD's for case-control study
- Intercepts differ by approximately

\[ \log \frac{q_d}{q_c} = \log \frac{1}{0.146} = 1.91 \]

where 0.146=1562/(11267-574), and 11267 is the number of women with information on gestational age at birth. Note that 100% (=1) of SGA women are sampled as cases.