Illustration: logistic regression

• Response: \( Y = 1 \) if infant develops bronchopulmonary dysplasia (BPD) by 28\(^{th} \) day of life; \( Y = 0 \) otherwise.

• Covariates:
  – birth weight (grams);
  – gestational age (weeks);
  – mother diagnosed with toxemia (0, 1).

• Sample: 223 infants.
SAS program for one covariate:

```sas
title 'Logistic regression analysis of BPD data';

options linesize = 80 pagesize = 21 nodate;

data bpd;
  infile 'bpd.txt' firstobs = 28;
  input BPD BirthWeight GestationalAge Toxemia;
  BirthWeight = BirthWeight / 100;
run;

proc genmod data = bpd descending;
  model BPD = BirthWeight / dist = binomial;
run;
```
SAS output:

Logistic regression analysis of BPD data

The GENMOD Procedure

Model Information

Data Set WORK.BPD
Distribution Binomial
Link Function Logit
Dependent Variable BPD
Observations Used 223

Response Profile

<table>
<thead>
<tr>
<th>Ordered Value</th>
<th>BPD</th>
<th>Total Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>76</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>147</td>
</tr>
</tbody>
</table>
Logistic regression analysis of BPD data

The GENMOD Procedure

PROC GENMOD is modeling the probability that BPD='1'.

Criteria For Assessing Goodness Of Fit

<table>
<thead>
<tr>
<th>Criterion</th>
<th>DF</th>
<th>Value</th>
<th>Value/DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviance</td>
<td>221</td>
<td>223.7206</td>
<td>1.0123</td>
</tr>
<tr>
<td>Scaled Deviance</td>
<td>221</td>
<td>223.7206</td>
<td>1.0123</td>
</tr>
<tr>
<td>Pearson Chi-Square</td>
<td>221</td>
<td>251.6647</td>
<td>1.1388</td>
</tr>
<tr>
<td>Scaled Pearson X2</td>
<td>221</td>
<td>251.6647</td>
<td>1.1388</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td></td>
<td>-111.8603</td>
<td></td>
</tr>
</tbody>
</table>

Algorithm converged.
Logistic regression analysis of BPD data

The GENMOD Procedure

Analysis Of Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Error</th>
<th>Confidence Limits</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>4.0343</td>
<td>0.6958</td>
<td>2.6706 5.3980</td>
<td>33.62</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>BirthWeight</td>
<td>1</td>
<td>-0.4229</td>
<td>0.0641</td>
<td>-0.5485 -0.2973</td>
<td>43.55</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Scale</td>
<td>0</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000 1.0000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: The scale parameter was held fixed.
R code:

bpd <- read.table("bpd.txt", skip = 27);
names(bpd) <- c("BPD", "BirthWeight", "GestationalAge", "Toxemia");

bpdGlm <- glm(BPD ~ I(BirthWeight / 100), data = bpd, family = binomial());
print(summary(bpdGlm));

pdf("bpd.pdf");
plot(bpd[, "BirthWeight"], bpd[, "BPD"]);

x <- seq(from = min(bpd[, "BirthWeight"]),
         to = max(bpd[, "BirthWeight"]), length = 40);
pred <- predict.glm(bpdGlm, data.frame(BirthWeight = x), se.fit = TRUE);
y <- pred$fit;
lines(x, exp(y) / (1 + exp(y)));
y <- pred$fit + qnorm(.025) * pred$se.fit;
lines(x, exp(y) / (1 + exp(y)), col = 2, lty = 2);
y <- pred$fit + qnorm(.975) * pred$se.fit;
lines(x, exp(y) / (1 + exp(y)), col = 2, lty = 2);
title("BPD data with fit and 95% pointwise CI");
dev.off();
R output:

Call:
glm(formula = BPD ~ I(BirthWeight/100), family = binomial(),
    data = bpd)

Deviance Residuals:
  Min      1Q  Median      3Q     Max
-1.9916  -0.7993  -0.4096   0.9242  2.4802

Coefficients:
                           Estimate Std. Error  z value Pr(>|z|)
(Intercept)               4.03429    0.69571   5.799 6.68e-09 ***
I(BirthWeight/100)        -0.42291    0.06408  -6.600 4.11e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 286.14  on 222  degrees of freedom
    Residual deviance: 223.72  on 221  degrees of freedom
    AIC: 227.72

Number of Fisher Scoring iterations: 4
BPD data with fit and 95% pointwise CI
SAS program for all covariates:

title 'Logistic regression analysis of BPD data';

options linesize = 80 pagesize = 21 nodate;

data bpd;
   infile 'bpd.txt' firstobs = 28;
   input BPD BirthWeight GestationalAge Toxemia;
   BirthWeight = BirthWeight / 100;
run;

proc genmod data = bpd descending;
   model BPD = BirthWeight GestationalAge Toxemia / dist = binomial;
run;
Logistic regression analysis of BPD data

The GENMOD Procedure

Analysis Of Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald 95% Confidence Limits</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>13.9361</td>
<td>2.9826</td>
<td>8.0904 19.7818</td>
<td>21.83</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>BirthWeight</td>
<td>1</td>
<td>-0.2644</td>
<td>0.0812</td>
<td>-0.4236 -0.1051</td>
<td>10.59</td>
<td>0.0011</td>
</tr>
<tr>
<td>GestationalAge</td>
<td>1</td>
<td>-0.3885</td>
<td>0.1149</td>
<td>-0.6137 -0.1634</td>
<td>11.44</td>
<td>0.0007</td>
</tr>
<tr>
<td>Toxemia</td>
<td>1</td>
<td>-1.3438</td>
<td>0.6075</td>
<td>-2.5345 -0.1531</td>
<td>4.89</td>
<td>0.0270</td>
</tr>
<tr>
<td>Scale</td>
<td>0</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000 1.0000</td>
<td>4.89</td>
<td>0.0270</td>
</tr>
</tbody>
</table>

NOTE: The scale parameter was held fixed.
Illustration: Poisson (log-linear) regression

- Response: $Y$ is number of cases of coronary heart disease (CHD);

- Covariates:
  - person-years of followup;
  - smoking exposure (0, 10, 20, 30);
  - blood pressure at least 140 (0, 1);
  - personality (1 = Type A, 0 = Type B).
• Data for 3,154 men aged 40–50, in 16 groups by covariates.

• First model:

\[
\log \left( \frac{\mu_i}{\text{followup}_i} \right) = \beta_1 + \beta_2 \text{Smoking}_i
\]

Note: log(followup\_i) appears on the RHS, but with a coefficient of 1; this is called an offset.

• Can also include other covariates.
SAS program:

```
title 'Poisson regression analysis of CHD data';

options linesize = 80 pagesize = 21 nodate;

data chd;
  infile 'chd.txt' firstobs = 28;
  input Smoking BloodPressure Behavior CHD Followup;
  lFollowup = log(Followup);
run;

proc genmod data = chd;
  model CHD = Smoking
    / dist = Poisson offset = lFollowup /* dscale */;
run;

proc genmod data = chd;
  model CHD = Smoking Behavior BloodPressure
    / dist = Poisson offset = lFollowup /* dscale */;
run;
```
SAS output:

Poisson regression analysis of CHD data

The GENMOD Procedure

Model Information

Data Set WORK.CHD
Distribution Poisson
Link Function Log
Dependent Variable CHD
Offset Variable lFollowup
Observations Used 16

Criteria For Assessing Goodness Of Fit

<table>
<thead>
<tr>
<th>Criterion</th>
<th>DF</th>
<th>Value</th>
<th>Value/DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviance</td>
<td>14</td>
<td>89.3820</td>
<td>6.3844</td>
</tr>
<tr>
<td>Scaled Deviance</td>
<td>14</td>
<td>89.3820</td>
<td>6.3844</td>
</tr>
<tr>
<td>Pearson Chi-Square</td>
<td>14</td>
<td>104.2445</td>
<td>7.4460</td>
</tr>
</tbody>
</table>
Poisson regression analysis of CHD data

The GENMOD Procedure

Criteria For Assessing Goodness Of Fit

<table>
<thead>
<tr>
<th>Criterion</th>
<th>DF</th>
<th>Value</th>
<th>Value/DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaled Pearson X2</td>
<td>14</td>
<td>104.2445</td>
<td>7.4460</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td></td>
<td>463.2618</td>
<td></td>
</tr>
</tbody>
</table>

Algorithm converged.
Poisson regression analysis of CHD data

The GENMOD Procedure

Analysis Of Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Error</th>
<th>Confidence Limits</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-4.7993</td>
<td>0.0885</td>
<td>-4.9728 -4.6258</td>
<td>2939.54</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Smoking</td>
<td>1</td>
<td>0.0318</td>
<td>0.0056</td>
<td>0.0207 0.0428</td>
<td>31.88</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Scale</td>
<td>0</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000 1.0000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: The scale parameter was held fixed.
Poisson regression analysis of CHD data

The GENMOD Procedure

Model Information

Data Set WORK.CHD
Distribution Poisson
Link Function Log
Dependent Variable CHD
Offset Variable lFollowup
Observations Used 16

Criteria For Assessing Goodness Of Fit

<table>
<thead>
<tr>
<th>Criterion</th>
<th>DF</th>
<th>Value</th>
<th>Value/DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviance</td>
<td>12</td>
<td>21.2397</td>
<td>1.7700</td>
</tr>
<tr>
<td>Scaled Deviance</td>
<td>12</td>
<td>21.2397</td>
<td>1.7700</td>
</tr>
<tr>
<td>Pearson Chi-Square</td>
<td>12</td>
<td>22.1455</td>
<td>1.8455</td>
</tr>
</tbody>
</table>
Poisson regression analysis of CHD data

The GENMOD Procedure

Criteria For Assessing Goodness Of Fit

<table>
<thead>
<tr>
<th>Criterion</th>
<th>DF</th>
<th>Value</th>
<th>Value/DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaled Pearson X2</td>
<td>12</td>
<td>22.1455</td>
<td>1.8455</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td></td>
<td>497.3329</td>
<td></td>
</tr>
</tbody>
</table>

Algorithm converged.
Poisson regression analysis of CHD data

The GENMOD Procedure

Analysis Of Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Error</th>
<th>Wald 95% Confidence Limits</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-5.4202</td>
<td>0.1308</td>
<td>-5.6765 -5.1638</td>
<td>1716.79</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Smoking</td>
<td>1</td>
<td>0.0273</td>
<td>0.0056</td>
<td>0.0163 0.0383</td>
<td>23.72</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Behavior</td>
<td>1</td>
<td>0.7526</td>
<td>0.1362</td>
<td>0.4856 1.0195</td>
<td>30.53</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>BloodPressure</td>
<td>1</td>
<td>0.7534</td>
<td>0.1292</td>
<td>0.5001 1.0067</td>
<td>33.98</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Scale</td>
<td>0</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000 1.0000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: The scale parameter was held fixed.
Overdispersion:

- Use the dscale or pscale option on the model statement;

- Parameter estimates are unchanged, but $\phi$ is estimated by the deviance or Pearson's $\chi^2$, divided by degrees of freedom.

- Parameter standard errors are multiplied by $\sqrt{\phi}$. 