

**22S:152**  
**Lab session 6**  
**Logistic Regression**  
**Autocorrelation**

Nov. 6, 2000

## 1 Logistic regression: risk factors for lung disease in low-birthweight infants

The following description is from Pagano, M and Gauvreau, K. (2000) *Principles of Biostatistics, 2nd ed.*, Duxbury. The study was reported in Van Marter, O.J. *et al.*, (1990) "Maternal Glucocorticoid Therapy and Reduced Ris of Bronchopulmonary Dysplasia," *Pediatrics*, Vol. 86, pp. 331-336.

Consider the population of low birth weight infants—in this case, defined as those weighing less than 1750 grams—who satisfy the following criteria: They are confined to a neonatal intensive care unit, they require intubation during the first 12 hours of life, and they survive for at least 28 days. In a sample of 223 such infants drawn from the underlying population, 76 were diagnosed with bronchopulmonary dysplasia (BPD), a chronic type of lung disease. The remaining 147 were not... We would estimate the probability that an infant in this population develops BPD by the sample proportion

$$\hat{p} = \frac{76}{223} = 0.341.$$

We might suspect that there are certain factors—both maternal and neonatal— that affect the likelihood that a particular infant will develop BPD. If we can classify a child according to these characteristics, it is possible that we could estimate his or her probability of developing the lung disease with greater precision than that afforded by the single value  $\hat{p}$ , and subsequently take measures to decrease this probability.

The dataset `bdp.dat` includes the following variables:

<code>bdp</code>	1 = presence of BPD in the child 0 = absence " " " " " "
<code>birthwt</code>	birthweight in grams
<code>gestage</code>	gestational age in weeks
<code>toxemia</code>	1 = mother had toxemia, 0 = mother did not
<code>steroid</code>	1 = mother was given steroid therapy, 0 mother was not

The following dataset will read in the dataset:

```
data bpd ;
infile 'bdp.dat' firstobs = 2 ;
* firstobs = 2 makes SAS skip 1st row of file;
input bpd birthwt gestage toxemia @50 steroid 1. ;
* @50 steroid 1. causes SAS to read in the steroid variable
correctly even though there is an extra character in the data
file right after it ;
run ;
```

We will not use the `toxemia` variable in this lab. We are interested in whether `birthwt`, `gestage`, and `steroid` are useful predictors of probability of `bdp`.

```
proc logistic descending ;
model bpd = birthwt gestage steroid /lackfit ;
run ;
```

The “descending” option on the “proc logistic” statement causes SAS to model the probability that `bdp` = 1 instead of the probability that `bdp` = 0.

## 2 Interpreting the output

Use your output to answer the following questions: Note that parameters in a logistic model

$$L_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$$

can be interpreted as follows:

$\beta_0$ : When  $X_1 = X_2 = X_3 = 0$ , the odds favoring  $Y = 1$  are  $e^{\beta_0}$ , and the probability that  $Y = 1$  is

$$P(Y = 1) = \frac{1}{1 + e^{-\beta_0}}$$

$\beta_1$ : Each one-unit increase in  $X_1$  multiplies the odds favoring  $Y = 1$  by  $e^{\beta_1}$ , if  $X_2$  and  $X_3$  are held constant. Another way to say this is that the odds favoring  $Y = 1$  change by  $100(e^{\beta_1} - 1)$  percent with each one-unit increase in  $X_1$ .

The probability that  $Y = 1$  changes by an amount that depends not only on  $\beta_1$  but also on all other terms in the equation:  $\beta_0, \beta_2, \beta_3, X_1, X_2$ , and  $X_3$ .

- Which predictors are significant at  $\alpha = .05$ ? At  $\alpha = .10$ ? (Refer to the Wald chi-square tests for the individual predictors.)
- What p-value do we obtain regarding the null hypothesis that all three coefficients are zero? (Refer to the chi-square test based on -2 times the log likelihood.)

3. Interpret each of the coefficients in terms of odds.

4. What probability of bpd do we predict for an infant of birthweight 1500 grams, gestational age 25 weeks, whose mother was not treated with steroids?

5. Does the Hosmer-Lemeshow test show evidence of poor fit of this model?

6. How can birthwt be such a significant predictor when its coefficient is so close to zero and the odds ratio is so close to 1? How might you change how you report its effect to improve understanding by nonstatisticians?

Percent Tied	0.2	Tau-a	0.303
Pairs	11172	c	0.835

Partition for the Hosmer and Lemeshow Test

Group	Total	bpd = 1		bpd = 0	
		Observed	Expected	Observed	Expected
1	22	2	0.67	20	21.33
2	22	1	1.33	21	20.67
3	23	2	2.14	21	20.86
4	22	4	3.25	18	18.75
5	22	4	4.70	18	17.30
6	23	7	7.63	16	15.37
7	22	10	9.18	12	12.82
8	22	10	12.21	12	9.79
9	22	16	15.52	6	6.48
10	23	20	19.37	3	3.63

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
4.4401	8	0.8154

The SAS System 7  
08:39 Tuesday, November 19, 2002

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	13.9168	2.9116	22.8470	<.0001
birthwt	1	-0.00274	0.000825	11.0529	0.0009
gestage	1	-0.3810	0.1120	11.5688	0.0007
steroid	1	-0.6277	0.3719	2.8483	0.0915

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
birthwt	0.997	0.996	0.999
gestage	0.683	0.548	0.851
steroid	0.534	0.258	1.107

Association of Predicted Probabilities and Observed Responses

Percent Concordant	83.4	Somers' D	0.671
Percent Discordant	16.4	Gamma	0.672

### 3 Logistic models in Insight

Go into Insight, using the bpd dataset. You can use the "Fit" option on the "Analyze" menu to do logistic regression. Specify the response and predictor variables as usual. Then click the "Method" button. For "Respons Dist.," choose "Binomial," and for "Link Function," choose "Logit." Then click the "Output" button, and choose "Output variables." Check "Predicted," "Linear predictor," and "Deviance residuals."