Search this site...

STAT 504 Analysis of Discrete Data

6.2.1 - Fitting the Model in SAS

Lesson

There are different ways to do this depending on the format of the data. As before, for details you need to refer to SAS or R help. Here are some general guidelines to keep in mind. Please note that we make a distinction about the way the data are entered into SAS (or R).

If data come in a ***tabular form***, i.e., response pattern is with counts (as seen in the previous example).

`model y/n = x1 x2 /link=logit dist=binomial` [any other options you may want];

`model y/n = x1 x2 /` [put any other options you may want here];

- PROC GENMOD: We need a variable that specifies the number of cases that equals marginal frequency counts or number of *trials* (e.g. *n*), and the number of *events* (*y*)
- PROC LOGISTIC: We do need a variable that specifies the number of cases that equals marginal frequency counts

If data come in a ***matrix form***, i.e., *subject* × *variables* matrix with one line for each subject, like a database

`model y/n = x1 x2 / link = logit dist = binomial` [put any other options you may want here];

`model y = x1 x2 /` [any other options you may want];

- PROC GENMOD: We need a variable that specifies the number of cases that equals marginal frequency counts or number of *trials* (e.g. *n*), and the number of *events* (*y*)
- PROC LOGISTIC: We do NOT need a variable that specifies the number of cases that equals marginal frequency counts

fitting. The outputs from R will be essentially the same.

Resource Menu

Example 6-1: Student Smoking

Let's begin with collapsed 2x2 table:

	Student smokes	Student does not smoke
1–2 parents smoke	816	3203
Neither parent smokes	188	1168

Let's look at one part of smoke.sas :

```
1  data smoke;
2  input s $ y n ;
3  cards;
4  smoke 816 4019
5  nosmoke 188 1356
6  ;
7  proc logistic data=smoke descending;
8  class s (ref=first) / param=ref;
9  model y/n = s /scale=none;
10 run;
```

in the logistic step, the statement:

`descending`

insures that you are modeling a probability of an "event" which takes value 1, otherwise by default SAS models the probability of "nonevent"

`class S (ref=first) / param=ref;`

says that S should be coded as a categorical variable using the first category as the reference or zero group. (The first category is "nosmoke," because it comes before "smoke" in alphabetical order). You can vary this order by additional options provided by class and/or by entering data in a different order. See SAS online help for details, and the rest of smoke.sas for more options.

`model y/n = s /scale = none`

Because we have grouped data (i.e. multiple trials per line of the data set), the model statement uses the "event/trial" syntax, in which y/n appears on the left-hand side of the equal sign. The predictors go on the right-hand side, separated by spaces if there are more than one. An intercept is added automatically by default.

`scale=none`

option serves two purposes. One, it will give us the overall goodness-of-fit test statistics, deviance G^2 and Person X^2. It also enables us to specify a value for a dispersion parameter in order to correct for over- or under-dispersion. In this case, we are NOT controlling for either. *Overdispersion* is an important concept with discrete data. In the context of logistic regression, overdispersion occurs when there are discrepancies between the observed responses y_i and their predicted values $\hat{\mu}_i = n_i\hat{\pi}_i$ and these values are larger than what the binomial model would predict.

$$\mu_i(n_i - \mu_i)$$

continuous predictors.

Now let's review some of the output from the program smoke.sas that uses PROC LOGISTIC

Model Information

Model Information

Data Set	WORK.SMOKE
Response Variable (Events)	y
Response Variable (Trials)	n
Model	binary logit
Optimization Technique	Fisher's scoring

This section, as before, tells us which dataset we are manipulating, the labels of the response and explanatory variables and what type of model we are fitting (e.g. binarly logit), and type of scoring algorithm for parameter estimation. *Fisher scoring* is a variant of Newton-Raphson method for ML estimation. In logistic regression they are equivalent.

Response Profile

Response Profile

Ordered Value	Binary Outcome	Total Frequency
1	Event	1004
2	Nonevent	4371

From an explanatory categorical (i.e, class) variable S with 2 levels (0,1), we created one ***dummy variable*** (e.g. ***design variable***):

$X_1 = 1$ ("smoke") if parent smoking = at least one,
$X_1 = 0$ ("nosmoke") if parent smoking = neither

parent smoking = nosmoke is equal 0 and is the baseline.

Model Convergence Status

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Since we are using an iterative procedure to fit the model, that is to find the ML estimates, we need some indication if the algorithm converged.

Overall goodness-of-fit testing

Test: H_0: current model vs. H_A: saturated model

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	0.0000	0	.	.
Pearson	0.0000	0	.	.

Number of events/trials observations: 2

The Pearson statistic $X^2 = 27.6766$ is precisely equal to the usual X^2 for testing independence in the 2 × 2 table. And the deviance $G^2 = 29.1207$ is precisely equal to the G^2 for testing independence in the 2 × 2 table. *Thus by the assumption, the intercept-only model or the null logistic regression model states that student's smoking is unrelated to parents' smoking (e.g., assumes independence, or odds-ratio=1).* But clearly, based on the values of the calculated statistics, this model (i.e., independence) does NOT fit well. This example shows that analyzing a 2 × 2 table for association is equivalent to logistic regression with a single dummy variable. Later on we will compare these tests to the loglinear model of independence see smokelog.sas and smokelog.lst .

The goodness-of-fit statistics, X^2 and G^2, are defined as before in the tests of independence and loglinear models (e.g. compare observed and fitted values). For the χ^2 approximation to work well, we need the n_is sufficiently large so that the expected values $\hat{\mu}_i \geq 5$ and $n_i \hat{\mu}_i \geq 5$ for most of the rows i. We can afford to have about 20%

Statistical Inference and Estimation

Probability and Distribution

A Review of the Principles of Statistics

R Programs

Case Study

Analysis of Maximum Likelihood Estimates

Once an appropriate model is fitted, the success probabilities need to be estimated using the model parameters. Note that success probabilities are now **NOT** simply the ratio of observed number of successes and the number of trials. A model fit introduces a structure on the success probabilities. The estimates will now be functions of model parameters.

What are the parameter estimates? What is the fitted model?

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.8266	0.0786	540.2949	<.0001
s	smoke	1	0.4592	0.0878	27.3361	<.0001

The fitted model is $\text{logit}(\hat{\pi}_i) = \log \frac{\hat{\pi}_i}{1-\hat{\pi}_i} = \hat{\beta}_0 + \hat{\beta}_1 X_i = -1.837 + 0.459 smoke$

where *smoke* is a ***dummy variable*** (e.g. ***design variable***) that takes value 1 if at least one parents is smoking and 0 if neither is smoking *as discussed above*.

Estimated $\beta_0 = -1.827$ with a standard error of 0.078 is significant and it says that log-odds of a child smoking versus not smoking if neither parents is smoking (the baseline level) is -1.827 and it's statistically significant.

Estimated $\beta_1 = 0.459$ with a standard error of 0.088 is significant and it says that log-odds-ratio of a child smoking versus not smoking if at least one parent is smoking versus neither parents is smoking (the baseline level) is 0.459 and it's statistically significant. $exp(0.459) = 1.58$ are the estimated odds-ratios; compare with our analysis in Section 6.2.

Thus there is a strong association between parent's and children smoking behavior, and

$$\hat{\pi}_i = \frac{exp(-1.826 + 0.4592X_i)}{1 + exp(-1.826 + 0.4592X_i)}$$

For example, the predicted probability of a student smoking given that neither parent smokes is

$$P(Y_i = 1|X_i = 0) = \frac{exp(-1.826 + 0.4592 \times 0)}{1 + exp(-1.826 + 0.4592 \times 0)} = 0.14$$

and the predicted probability of a student being a smoker if at least one parent smokes is

$$P(Y_i = 1|X_i = 1) = \frac{exp(-1.826 + 0.4592(X_i = 1)}{1 + exp(-1.826 + 0.4592(X_i = 1))} = 0.20$$

By invoking the following option in MODEL, **output out=predict pred=prob** the PROC LOGISTIC will print the predicted probabilities in the output file:

Logistic regression for 2x2 table

Obs	s	y	n	prob
1	smoke	816	4019	0.20304
2	nosmoke	188	1356	0.13864

so you do NOT need to do this calculation by hand; but it maybe useful to try it out to see if you understand what's going on. In this model,β_0 is the log-odds of children smoking for no-smoking parents /(X_i = 0\). Thus *exp*(−1.8266) are the odds that a student smokes when the parents do not smoke.

Looking at the 2 × 2 table, the estimated log-odds for nonsmokers is

$$\left(\quad \right)$$

agrees exactly with the log-odds ratio from the 2 × 2 table (e.g. *ln*(1.58) = (816×1168) / (188×3203) = 0.459). That is *exp*(0.459) = 1.58 which is the estimated odds ratio of a student smoking.

To relate this to interpretation of the coefficients in a linear regression, you could say that for every one-unit increase in the explanatory variable X_1 (e.g. changing from no smoking parents to smoking parents), the odds of "success" $\pi_i/(1-\pi_i)$ will be multiplied by $exp(\beta_1)$, given that all the other variables are held constant.

This is not surprising, because in the logistic regression model β_1 is the difference in the log-odds of children smoking as we move from "nosmoke" (i.e. neither parent smokes) /(X_i = 0/) to "smoke" (i.e. at least one parent smokes) $X_i = 1$, and the difference in log-odds is a log-odds ratio.

Testing Individual Parameters

Testing the hypothesis that the probability of the characteristic depends on the value of the *j*th variable.

$$\text{Testing } H_0 : \beta_j = 0 \text{ versus } H_1 : \beta_j \neq 0.$$

The *Wald chi-squared statistics* $z^2 = (\hat{\beta}_j/\mathrm{SE}(\hat{\beta}_k))^2$ for these tests are displayed along with the estimated coefficients in the "Analysis of Maximum Likelihood Estimates"

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.8266	0.0786	540.2949	<.0001
s	smoke	1	0.4592	0.0878	27.3361	<.0001

The values indicate the significant relationship between the logit of the odds of student smoking in parents' smoking behavior.

Or, we can use the information from "Type III Analysis of Effects" section.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
s	1	27.3361	<.0001

Again, this information indicates that parents' smoking behavior is a significant factor in the model. We could compare z^2 to a chisquare with one degree of freedom; the *p*-value would then be the area to the right of z^2 under the χ^2_1 density curve.

A value of z^2 (Wald statistic) bigger than 3.84 indicates that we can reject the null hypothesis$\beta_j = 0$ at the .05-level.

$$\beta_1 : \left(\frac{0.4592}{0.0878}\right)^2 = 27.354$$

Confidence Intervals of Individual Parameters:

An approximate (1 − α) × 100% confidence interval forβ_j is given by

$$\hat{\beta}_j \pm z_{(\quad/)} \times SE(\hat{\beta}_j)$$

(*exp*(0.287112), *exp*(0.63128)) = (1.3325, 1.880)

Compare this with the output we get from PROC LOGISTIC:

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
s smoke	vs nosmoke	1.583	1.332	1.880

When fitting logistic regression, we need to evaluate the overall fit of the model, significance of individual parameter estimates and consider their interpretation. For assessing the fit of the model, we also need to consider the analysis of residuals. Definition of Pearson, deviance and adjusted residuals is as before, and you should be able to carry this analysis.

If we include the statement