Logistic regression diagnostics

Biometry 755

Spring 2009

Assessing model fit

A good model is one that 'fits' the data well, in the sense that the values predicted by the model are in close agreement with those observed. In logistic regression, we obtain the predicted values (*event (1)* or *not event (0)*) using the fitted probabilities of the occurrence of an event.

From logit to probability in SAS

In the previous lecture, we modelled the log odds of remaining drug free as a function of AGE, NDRUGTX, IVHX, TREAT and SITE. The fitted model is

$$\ln\left[\frac{\text{Prob}(\text{DFREE}=1)}{1-\text{Prob}(\text{DFREE}=1)}\right] = -2.37 + 0.052 \times \text{AGE}$$
$$-0.062 \times \text{NDRUGTX} - 0.64 \times \text{IVH}$$
$$-0.79 \times \text{IVHX}_2 + 0.46 \times \text{TREAT}$$
$$+0.12 \times \text{SITE}$$

We also saw how to estimate the probability of remaining drug free for 12 months based on the fitted model. We can also request that SAS construct these probabilities for each subject in the data.

Estimated probabilities in SAS

```
proc logistic data = two descending;
    class ivhx (param = ref ref = 'Never');
    model dfree = age ndrugtx ivhx treat site;
    output out = fittedprobs pred = probs;
run;
quit;
```

id	age	beck	ivhx	ndrugtx	treat	site	dfree	probs
1	39	9.000	Recent	1	Long	A	Otherwise	0.32545
2	33	34.000	Previous	8	Long	A	Otherwise	0.20951
3	33	10.000	Recent	3	Long	A	Otherwise	0.23741
4	32	20.000	Recent	1	Short	A	Otherwise	0.17512
5	24	5.000	Never	5	Long	A	Remained drug free	0.27374
6	30	32.550	Recent	1	Long	A	Otherwise	0.23168

The Hosmer-Lemeshow GOF test

The Hosmer and Lemeshow goodness of fit (GOF) test measures how well the estimated model fits the observed data based on the following steps.

1. Arrange the observed data into ten groupings based on deciles of the estimated probabilities of an event. These are commonly referred to as *deciles of risk*. For the model shown on Slide 3, the 10th, 20th, . . ., 90th, 100th percentiles of the estimated probabilities are: 0.13, 0.16, 0.19, 0.22, 0.24, 0.28, 0.31, 0.35, 0.40, 0.58 (*I got these percentiles from PROC UNIVARIATE*).

The Hosmer-Lemeshow GOF test (cont.)

2. Sum the probabilities in each decile. This is the expected number of events within each decile.
3. Construct a usual 'observed - expected' chi-square statistic.

The null hypothesis for the Hosmer and Lemeshow test is that the data fit the model. Therefore you want to fail to reject the null hypothesis. In other words, you do not want to find evidence that the data differ significantly from the fitted model.

The Hosmer-Lemeshow GOF test in SAS

```
proc logistic data = one descending;
    class ivhx (param = ref ref = 'Never');
    model dfree = age ndrugtx ivhx treat site /lackfit
run;
quit;
```


The HL GOF test in SAS (cont.)

Partition for the Hosmer and Lemeshow Test

		dfree = Remained drug free		dfree = Otherwise	
Group	Total	Observed	Expected	Observed	Expected
1	58	6	5.73	52	52.27
2	58	8	8.59	50	49.41
3	58	8	10.18	50	47.82
4	58	14	11.94	44	46.06
5	58	8	13.26	50	44.74
6	58	24	15.16	34	42.84
7	58	13	17.03	45	40.97
8	58	22	19.18	36	38.82
9	58	20	21.74	38	36.26
10	53	24	24.19	29	28.81

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
12.9446	8	0.1138

Conclusion based on HL test

At $\alpha = 0.05$ we fail to reject the hypothesis that the data fit the model ($p = 0.1138$). Therefore we conclude that the model provides adequate fit.

ROC analysis

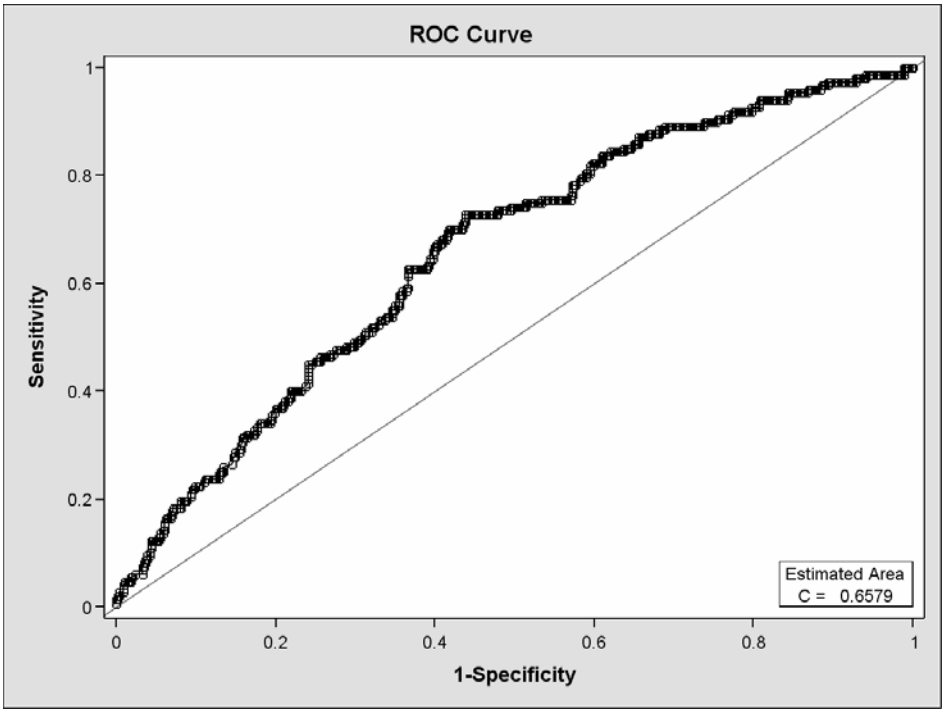
If the purpose of the logistic regression is to construct a predictive model, then an ROC (short for Receiver Operating Characteristics) curve is a useful graphical assessment of fit.

```
ods rtf file='E:\Logistic\ROC.rtf';
ods graphics on;
ods select ROCCurve;

proc logistic data = one descending;
    class ivhx (param = ref ref = 'Never');
    model dfree = age ndrugtx ivhx treat site;
    graphics ROC;
run;
quit;

ods graphics off;
ods rtf close;
```

ROC graphic



Assessing linearity in the logit

Recall the assumptions of logistic regression are

- Responses are observed values of independent Bernoulli random variables (comprises two assumptions in one)
- The model is linear in the logit for the predictors

We will discuss how to assess the validity of the linearity assumption. Such an assessment is only meaningful for continuous covariates.

Using LOESS to assess linearity

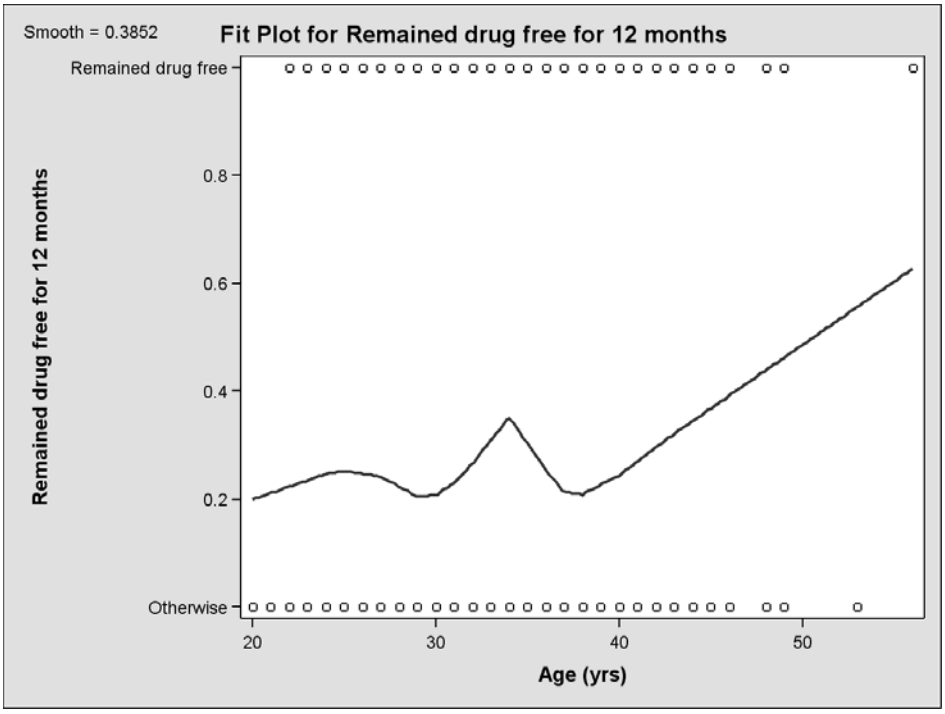
We will assess linearity in the logit for the covariate AGE in the IMPACT data.

```
ods html;
ods graphics on;

proc loess data = one;
    model dfree = age;
run;

ods graphics off;
ods html close;
```


Loess plot for AGE



Transformation for AGE?

Use fractional polynomials macro to investigate.

Pearson residuals

We analyze residuals to identify problems with the fitted model. The *Pearson residual*, r_j, is defined as follows:

$$r_j = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}$$

- j indexes a given covariate pattern (e.g. 40 year-olds with no prior drug treatments, recent history of injecting drug use, randomized to long arm at site B)
- y_j is the total number of positive responses for covariate pattern j
- m_j is the total number of observations with covariate pattern j
- $\hat{\pi}_j$ is the estimated probability of a positive response for covariate pattern j

Pearson residuals (cont.)

The summary measure based on the Pearson residuals is a chi-square statistic,

$$X^2 = \sum_{j=1}^{J} r_j^2 \sim \chi^2_{J-(k+1)}$$

where J is the total number of covariate patterns and k is the number of covariates in the model.

Deviance residuals

Another type of residual is the *deviance residual*, d_j. Its form is rather complicated, but the interested student can consult Hosmer and Lemeshow, *Applied Logistic Regression*, 2000, p. 146. A summary measure based on the deviance residuals is the *deviance*, and is defined as

$$D = \sum_{j=1}^{J} d_j^2 \sim \chi^2_{J-(k+1)}$$

where J is the total number of covariate patterns and k is the number of covariates in the model.

Residual plots in SAS

SAS provides a number of default plots based on the Pearson and deviance residuals that allow us to identify outlying observations and covariate patterns that are poorly fit by the model. We will focus on the following plots.

1. Pearson residuals (r_j) versus case number
2. Deviance residuals (d_j) versus case number
3. Change in Pearson Chi-square statistic versus case number (based on X^2 statistic shown on Slide 17)
4. Change in Deviance statistic versus case number (based on D statistic shown on Slide 18)

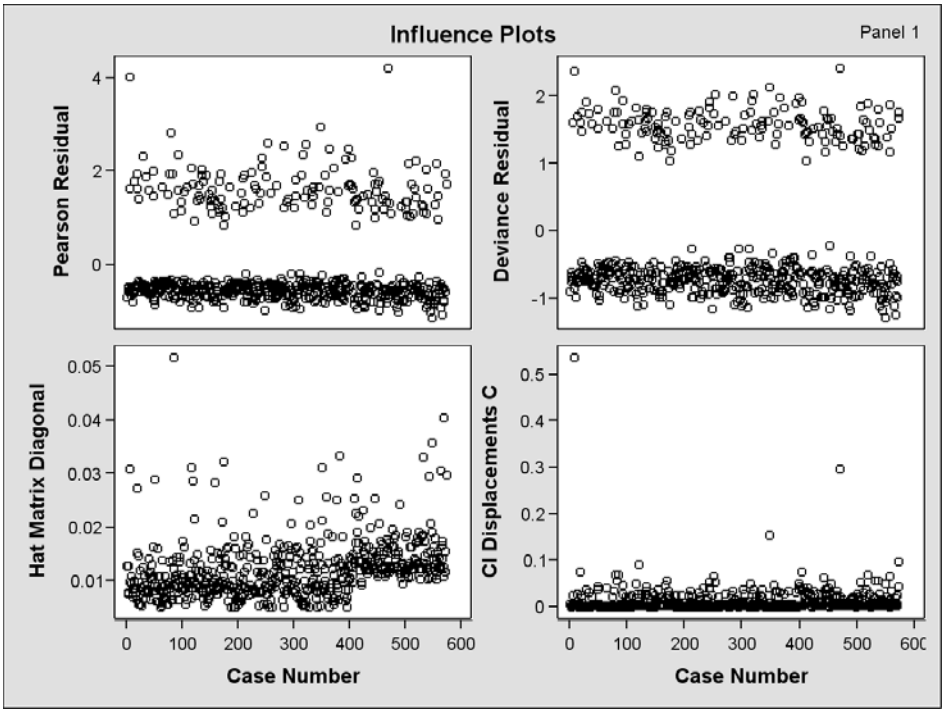
Residual plots in SAS (cont.)

```
ods html;
ods graphics on;

proc logistic data = two descending;
    class ivhx (param = ref ref = 'Never');
    model dfree = age ndrugtx ivhx treat site/influence
run;

ods graphics off;
ods html close;
```

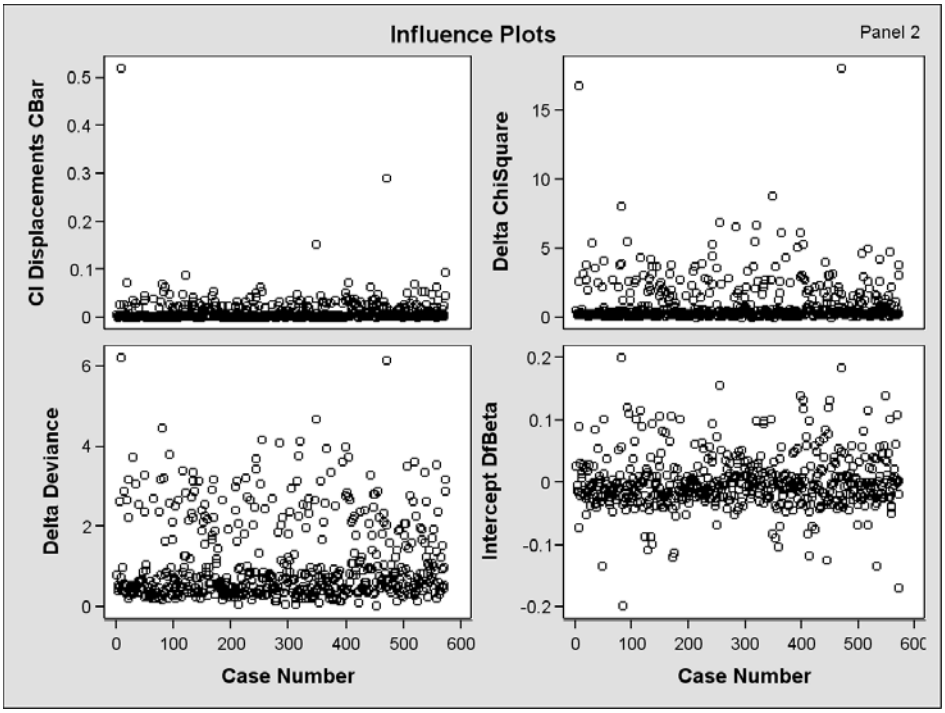
Pearson/deviance residual vs. case number



ΔX_j^2 and ΔD_j versus case number plots

1. For each covariate pattern, j, delete the observations corresponding to that covariate pattern.
2. Calculate the Pearson (or deviance) chi-square statistics with these observations deleted.
3. Calculate the decrease in the Pearson (or deviance) chi-square statistic due to the deletion of these observations. Call this quantity ΔX_j^2 (or ΔD_j).
4. Plot ΔX_j^2 (or ΔD_j) versus the index of each observation.

ΔX_j^2 and ΔD_j versus case number (cont.)



What values are “too big”?

In logistic regression we have to rely primarily on visual assessment, as the distribution of the diagnostics under the hypothesis that the model fits is known only in certain limited settings. In practice, an assessment of “large” is a judgement call based on experience and the particular set of data being analyzed.

Investigating poorly fit observations

From the plots of the Pearson and deviance residuals, and ΔX_j^2 and ΔD_j versus the case index, we identify two observations that appear to be poorly fit by the data. One has an index close to 0 and the other has an index close to 500. Both observations have Pearson residuals near 4, deviance residuals greater than 2, ΔX_j^2 values greater than 15 and ΔD_j values greater than 6.

In addition to the plots, the 'influence' option prints out the covariate values for each observation and the values of each statistic (i.e. r_j, d_j, ΔX_j^2, ΔD_j, etc.) to aid in their identification from the plots.

Investigating poorly fit observations (cont.)

Case Number	age	ndrugtx	ivhx Previous	ivhx Recent	treat	site
7	39	34	0	1	1	0
471	24	20	1	0	0	1

Case Number	Pearson Residual	Deviance Residual	Delta Deviance	Delta Chi-Square
7	4.0299	2.3863	6.2148	16.7604
471	4.2180	2.4221	6.1574	18.0818

Investigating poorly fit observations (cont.)

Clearly observations 7 and 471 are poorly fit by the model. We should compare the actual outcome for these observations against what is predicted by the model.

```
proc logistic data = one descending;
    class ivhx (param = ref ref = 'Never');
    model dfree = age ndrugtx ivhx treat site;
    output out = fittedprobs pred = probs;
run;

proc print data = fittedprobs;
    where id in (7,471);
run;
```


Investigating poorly fit observations (cont.)

Obs	id	age	beck	ivhx	ndrugtx	race	treat	site
7	7	39	19	Recent	34	White	Long	A
471	471	24	20	Previous	20	White	Short	B

Obs	dfree	probs
7	Remained drug free	0.058004
471	Remained drug free	0.053216

The covariate values for these subjects would indicate a poor prognosis, that is, a low probability (≈ 0.05) of remaining drug free for twelve months. However, they 'beat the odds' and remained drug free.