Model fit

HL Chapter 5 – part 1

Model building up to this point

- Find the “best” model, i.e. find a model that is better than all the other models you tried

Better than lousy
may still be lousy

Goodness-of-fit

- Is the model you selected good?
- Or is it a lousy model that’s just a little better than all the other lousy models you tried?

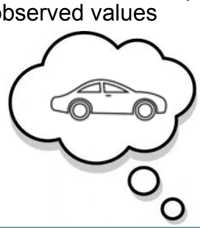
Definition: Covariate pattern

- Covariate pattern
  = a set of values for the model covariates
- Example: Assume the model contains age, gender and race
  - “age=30, Gender=Female, Smoking=No” is a covariate pattern

  - “age=50, Gender=Male, Smoking=Yes” is a covariate pattern
  - Etc.

What is model fit?

- Compare the observed outcome values (y) to those predicted by the model ($\hat{\pi}$)
- Determine how close the predicted values are to the observed values

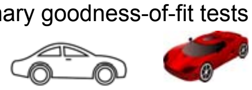


What is model fit?

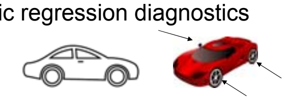
- Example:
  - Assume that the outcome values, y, for 5 subjects with covariate pattern
    “age=50, Gender=Male, Smoking=Yes”
    are 1, 1, 0, 1, 0
  - Average of y values = 3/5 = 0.6
  - $\hat{\pi}$ = probability of y predicted by the model for this covariate pattern should be near 0.6

Questions

- Does the model fit overall?
  - ➔ Summary goodness-of-fit tests



- Are there any individual observations that don't fit?
  - ➔ Logistic regression diagnostics



Summary goodness-of-fit tests

J = # covariate patterns, n = sample size

<u>Pearson X^2 test and Deviance test</u>

- Use when there are few different covariate patterns, J, i.e. when J << n
- Example:
  - n=200 study subjects, Outcome=y
  - Model covariates: Exposure (yes/no) and gender

J=4 covariate patterns
  - j=1: Exposed male; j=2: Exposed female; j=3: Unexposed male; j=4: Unexposed female

Summary goodness-of-fit tests

J = # covariate patterns, n = sample size

<u>Hosmer-Lemeshow test</u>

- Use when there are many different covariate patterns, J, i.e. when J ≈ n
- Example:
  - n=200 study subjects, Outcome=y
  - Model covariates: Age, systolic blood pressure, heart rate

J ≈ 200 covariate patterns
  - (Almost) everyone will have a different combination of age, systolic blood pressure and heart rate

Summary goodness-of-fit tests

J = # covariate patterns, n = sample size

<u>Osius-Rojek test</u>

- Use in all other cases if the sample size is "reasonably large"

The Pearson Chi-square test

Use when there are FEW covariate patterns (J) relative to the sample size (n)

- Calculates the difference between the observed and the predicted value for each covariate pattern
- Standardizes and squares each difference
- Adds the squared standardized differences over all covariate patterns

The Pearson Chi-square test

- If J << n, the resulting test statistic is X^2 distributed with J – p – 1 degrees of freedom

(J = # covariate patterns, p = # model covariates)

- P-value ≤ 0.05 ➔ evidence of lack of model fit
- P-value > 0.05 ➔ evidence of model fit

The Deviance test

Use when there are
FEW covariate patterns (J) relative to the sample size (n)

- Calculates the deviance for each covariate pattern and squares it
- Adds the squared deviances over all covariate patterns

The Deviance test

- If J << n, the resulting test statistic is X^2 distributed with J – p – 1 degrees of freedom

(J = # covariate patterns, p = # model covariates)

- P-value ≤ 0.05 ➔ evidence of lack of model fit
- P-value > 0.05 ➔ evidence of model fit

What if J≈n?

- IfJ≈n, the X^2 test assumption is violated

WHY?

- If J≈n, then each study subjects has his or her own covariate pattern (with a few exceptions)
- The person with the covariate pattern either has the outcome or doesn't have the outcome
- If we cross-classify outcome vs. exposure (i.e. covariate pattern), we have many zero cells
- This violates the X^2 test assumption that expected cell frequencies are "large"

The Hosmer-Lemeshow test

Use when there are
MANY covariate patterns (J) relative to the sample size (n)

Groups covariate patterns using 10 groups

<u>A. The deciles of risk method</u>

- Group 1 = 10% of study subjects with the lowest $\hat{\pi}$s
- Group 2 = 10% of study subjects with the next higher $\hat{\pi}$s
- …
- Group 10 = 10% of study subjects with the highest $\hat{\pi}$s

The Hosmer-Lemeshow test, cont.

<u>B. The fixed cutpoints method</u>

- Group 1 = all study subjects with $0 < \hat{\pi} \le 0.1$
- Group 2 = all study subjects with $0.1 < \hat{\pi} \le 0.2$
- …
- Group 10 = all study subjects with $0.9 < \hat{\pi} < 1.0$

The Hosmer-Lemeshow test, cont.

- Calculates the Pearson X^2 test based on groups rather than individuals
- The resulting test statistic is X^2 distributed with g-2 degrees of freedom
(g = # groups; in most cases g=10)

The Hosmer-Lemeshow test, cont.

In theory

- P-value ≤ 0.05 ➔ evidence of lack of model fit
- P-value > 0.05 ➔ evidence of model fit

- However, the Hosmer-Lemeshow test is not very powerful and in most cases a p-value below ≈0.25 is indicative of lack of fit

The Hosmer-Lemeshow test Problems

Fixed cutpoint method (B)

- Leads to a test statistic that does not adhere to the X^2(g-2) distribution very well

➔ P-values questionable
➔ Use deciles of risk method only

The Hosmer-Lemeshow test Problems

Deciles of risk method (A)

- After grouping, the expected cell frequencies may still be small
- The test is not very powerful, especially for n<400
- The test does not handle ties well (see next slides)

What if J<n?

- If J<n, ties occur
- A covariate pattern may be shared by several study subjects

- Each of these study subjects has the same value of $\hat{\pi}$
- I.e., $\hat{\pi}$ is tied for these study subjects

The Hosmer-Lemeshow test Dealing with ties – Example

- Hypothetical study, n=100
  - Persons 1-8 each have their own covariate pattern
  - Persons 9-14 have the same covariate pattern
  - Persons 15-20 each have their own covariate pattern
  - Etc.
- Deciles of risk:
  - Group 1 = 10% of study subjects with the lowest $\hat{\pi}$s

Example cont.

ID	Cov. pattern #	$\hat{\pi}$
1	1	0.08
2	2	0.10
3	3	0.11
4	4	0.12
5	5	0.13
6	6	0.15
7	7	0.16
8	8	0.17
9	9	0.18
10	9	0.18

10% with the lowest $\hat{\pi}$s

ID	Cov. pattern #	$\hat{\pi}$
11	9	0.18
12	9	0.18
13	9	0.18
14	9	0.18
15	10	0.23
16	11	0.25
17	12	0.26
18	13	0.27
19	14	0.28
20	15	0.30

10% with the next lowest $\hat{\pi}$s

Same covariate pattern and $\hat{\pi}$s

Option 1: Keep subjects 11-14 in group 2

- Pro
  - There are 10 groups each containing 10% of the observations
- Con
  - Persons with the same covariate pattern may be treated as different
  - Persons with different covariate patterns may be treated as if they were the same

Option 2: Move subjects 11-14 to group 1

- Pro
  - Persons with the same covariate pattern are not treated as different
- Con
  - More than 10% of subjects are in group 1; there aren't enough subjects left to have 10% in all subsequent groups
  - In extreme cases there may be only 8 or 9 groups
  - In these cases, the Hosmer-Lemeshow test almost always (possibly erroneously) indicates model fit
- **SAS uses Option 2**

The Osius-Rojek test

Use when there are
FEWER covariate patterns (J) than the sample size (n)
(but not too few)

- The Osius-Rojek test is a large sample normal approximation to the Pearson X^2 test
  ➔ results may be incorrect when the sample size is small
- Osius-Rojek test results are also questionable in the presence of very small or very large $\hat{\pi}$s
  ($\hat{\pi} < 10^{-5}$ or $\hat{\pi} > 1\text{-}10^{-5}$)

Caution

- Note that none of the goodness-of-fit tests are very powerful for sample sizes of less than approximately 400

The Stukel test

- Not a goodness-of-fit test
- Tests whether the model produces more or fewer small or large $\hat{\pi}$s than the standard logistic regression model assumes
- Does this by comparing the standard logistic regression model to a generalized logistic regression model with 2 extra parameters that allow for the tails (small or large $\hat{\pi}$s) to vary
- If neither extra parameter is significantly different from 0, the standard logistic regression model is OK

Example: Final GLOW500 model from chapter 4

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.7175	3.3217	0.2673	0.6051
PRIORFRAC	1	4.6117	1.8802	6.0163	0.0142
MOMFRAC	1	1.2465	0.3930	10.0630	0.0015
ARMASSIST	1	0.6441	0.2519	6.5370	0.0106
RATERISK 3 vs. 2,1	1	0.4690	0.2408	3.7935	0.0515
HEIGHT	1	-0.0467	0.0183	6.5005	0.0108
AGE	1	0.0573	0.0165	12.0578	0.0005
PRIORFRAC*AGE	1	-0.0553	0.0259	4.5423	0.0331
MOMFRAC*ARMASSIST	1	-1.2804	0.6230	4.2243	0.0398

Pearson chi-square, deviance and Hosmer-Lemeshow test

```
proc logistic descending data=glow500;
  model fracture=priorfrac momfrac armassist raterisk2
            height age priorfrac*age momfrac*armassist
        / scale=n aggregate lackfit;
run;
```

Perform Pearson chi-square and deviance test

Perform Hosmer-Lemeshow test

Model by covariate pattern rather than by subject

Pearson chi-square and deviance test - Results

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	469.6312	448	1.0483	0.2316
Pearson	442.3167	448	0.9873	0.5669

Number of Observations Read	500
Number of Observations Used	500

n

Number of unique profiles: 457 J

- P-values are >0.05 ➔ evidence of model fit
- But J=457 and n=500 (J≈n) ➔ test assumptions violated
- P-values must be interpreted with extreme caution

Hosmer-Lemeshow test results

Partition for the Hosmer and Lemeshow Test

Group	Total	FRACTURE = 1		FRACTURE = 0	
		Observed	Expected	Observed	Expected
1	50	3	3.31	47	46.69
2	50	4	4.86	46	45.14
3	50	7	6.28	43	43.72
4	51	11	8.08	40	42.92
5	50	8	9.60	42	40.40
6	50	12	11.44	38	38.56
7	50	9	14.34	41	35.66
8	50	19	17.69	31	32.31
9	50	25	21.90	25	28.10
10	49	27	27.50	22	21.50

Hosmer-Lemeshow test results

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
5.6582	8	0.6855

- J=457 similar to n=500
- Sample size adequate
- 10 groups
- Only two expected cell frequency < 5
- Test appropriate
- p>0.25 ➔ evidence of model fit

Osius-Rojek test results

pval
0.69515

Large p-value ➔ evidence of model fit

Reasonably large sample size (n=500) for large sample normal approximation
➔ test appropriate

Stukel test – Results

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	4.1659	3.5369	1.3873	0.2389
PRIORFRAC	1	6.5628	2.9014	5.1166	0.0237
MOMFRAC	1	1.9332	0.6910	7.8278	0.0051
ARMASSIST	1	0.8732	0.3615	5.8354	0.0157
raterisk2	1	0.6428	0.2925	4.8295	0.0280
HEIGHT	1	-0.0738	0.0276	7.1495	0.0075
AGE	1	0.0778	0.0317	6.0414	0.0140
priorfrac_age	1	-0.0788	0.0376	4.3907	0.0361
momfrac_armassist	1	-1.9541	0.8593	5.1714	0.0230
z1_j	1	-6.6173	3.2970	4.0282	0.0447
z2_j	1	-0.2534	0.3711	0.4665	0.4946

Shape of upper tail may be modeled in-adequately by this logistic model; ignore for now

Overall
p=0.0742