

Logistic Regression - Model Significance and Goodness of Fit Measures

- For the model overall, a likelihood ratio test is performed, contrasting the model in question with a null model containing no predictors. This is analogous to the F-test in SLR/MLR.
- The test statistic

$$\Lambda = -2(\log\text{-likelihood}_{null} - \log\text{-likelihood}_{full})$$

is χ_{df}^2 where degrees of freedom $df =$ number of predictors.

- This same type of test is used for contrasting two nested models, say dropping out 3 predictors out of 7

$$LR\ Test = -2(\log\text{-likelihood}_{smaller\ model} - \log\text{-likelihood}_{bigger\ model})$$

is χ_3^2 (3 degrees of freedom) and tests whether the three parameters considered for dropping simultaneously have $\beta_j = 0$.

Multiple Logistic Regression Inference, uses, etc

From a multiple logistic regression analysis, we obtain several quantities:

- likelihood ratio (LR) test - overall test for any ‘significant’ predictors X of log odds ratio (analogous to the F -test in the linear model).
- We have Z -tests for individual β coefficients ($H_0 : \beta = 0$) for the X'_s , which are in fact tests for corresponding ORs ($H_0 : \text{OR} = 1.0$).
- Pseudo- R^2 measures proportion reduction in log-likelihood over null model. This is a useful measure, but more like another F -test than a measure of model explanatory power.

Multiple Logistic Regression Inference, uses, etc

- Predicted probabilities for individuals.

$$\Pr(Y = 1|X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \dots + \beta_q X_q)}{1 + \exp(\beta_0 + \beta_1 X_1 + \dots + \beta_q X_q)}$$

These can be used to develop classification algorithms, predict probability of response prospectively, etc.

Example: For all data records, Y's are either 1 (event, death, etc) or 0 (non-event, etc), but the model predicts $\text{Prob}(Y=1)$ for each record. Depending on cut-point for predicting/assigning a case to be an event, how many do you get right? Adjust the cut-point based on this to optimize classification.

Assumptions and Model Diagnostic

- **The assumptions:** No systematic bias in measurement assumption is still required. Uncorrelated observations assumption is still required. No strong multicollinearity is still required (and you can still use VIF to check for violation of assumption). Linearity (w.r.t link function) is still required. No influential observations is still required. Normality and constant variance assumptions are no longer required. The error term is Bernoulli distributed.
- Many of the approaches we discussed before for addressing assumption violations in MLR may still apply for logistic regression. Confounders should still be always included. Interaction terms should not be left alone without main effects. Categorical variables should be considered as a whole set. ...and so on.

Logistic Regression - Diagnostics

- Several diagnostic quantities, aiming to detect outliers and influential points, are defined (C&H 12.5). These borrow concepts from linear regression.
 - *Pearson residuals* and *deviance residuals* plotted against the predicted probabilities or an index measure are similarly assessed for large deviations.
 - A similar *Leverage* measure can be derived to identify the effect of particular observations.
 - The *DBETA* measure determines the change in regression coefficients when observation i is omitted, implicitly evaluating influence.

Logistic Regression - Model Significance and Goodness of Fit Measures

- A commonly used goodness of fit measure in logistic regression is the Hosmer-Lemeshow test. The test groups the n observations into groups (according to their estimated probability of event) and calculates the corresponding generalized Pearson χ^2 statistic. Usually deciles (10 groups) are used.

$$H = \sum_{g=1}^G \frac{(O_g - E_g)^2}{n_g(\hat{p}_g(1 - \hat{p}_g))},$$

where O_g is the number of events in group g and $E_g = n_g \times \hat{p}_g$ is the expected number of events

- In this type of test, a *large* p-value (> 0.05) indicates good correspondence between observed and predicted outcomes.

Logistic Regression - Model Significance and Goodness of Fit Measures

```
. logit sta typ age
```

```
Iteration 0:  log likelihood = -100.08048
Iteration 1:  log likelihood = -87.895217
. . .
Iteration 5:  log likelihood = -86.537821
```

```
Logistic regression                Number of obs    =        200
                                   LR chi2(2)         =        27.09
                                   Prob > chi2        =        0.0000
Log likelihood = -86.537821        Pseudo R2       =        0.1353
```

sta	Coef.	Std. Err.	z	P> z	[95\% Conf. Interval]	
typ	2.453535	.75257	3.26	0.001	.978525	3.928545
age	.0340162	.0106944	3.18	0.001	.0130556	.0549767
_cons	-5.508762	1.033511	-5.33	0.000	-7.534407	-3.483118

```
. estat gof, group(10) table
```

Logistic model for sta, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0362	0	0.5	20	19.5	20
2	0.0478	1	0.9	20	20.1	21
3	0.0812	1	1.2	18	17.8	19
4	0.1209	2	1.8	18	18.2	20
5	0.1916	2	3.1	18	16.9	20
6	0.2531	7	4.6	14	16.4	21
7	0.2936	6	6.1	16	15.9	22
8	0.3376	5	5.8	13	12.2	18
9	0.3846	8	7.3	12	12.7	20
10	0.5186	8	8.7	11	10.3	19

number of observations = 200
 number of groups = 10
 Hosmer-Lemeshow chi2(8) = 2.95
 Prob > chi2 = 0.9372

Note: See Canvas for the examples completed with R

Logistic Regression - Goodness of Fit Measures

- This result looks very positive - good fit. However, Pseudo- $R^2 = 13\%$, and is not very high.
- Pseudo- R^2 measures proportion reduction in log-likelihood over null model. This is a useful measure, but more like another F-test than a measure of model explanatory power.
- The Hosmer-Lemeshow g.o.f. test is more valuable as a means to identify major systematic variation that is not explained. Large p-value does not assure that prediction will be highly accurate, etc.
- AIC and BIC measures can also be used to select among models.

Summary – Logistic Regression Models

- Logistic regression can be extended to a multinomial outcome variable, where Y equals one of several mutually exclusive nominal categories (see C&H).
- Similarly, ordinal logistic regression permits an ordered categorical response variable.
- These models fit into a general class of models in which a “link” function of the mean of Y is modeled by a linear combination of predictors.