STAT 770 Oct. 26 Lecture 16

GLM model selection, checking, and alternatives

Reading and Topics for this lecture: Chapters 5, 7.

(1) Rationale for Stepwise Model Selection

(2) Computational Issues

(3) Checking GLM Goodness of Fit – Binning (Sec. 5.2.5)

(4) Hosmer-Lemeshow Test

(5) GLMs and 'Tests for Trend' in $I \times 2$ Tables (Sec. 5.3.4)

(6) Other Links, other Models (Secs. 7.1, 7.3)

Wilks Theorem & Variable Selection

For maximal set of covariates X_i (incl. interactions $X_{i,k_1} * X_{i,k_2}$ etc.), link and variance function $g(\mu), v(\mu)$, outcomes Y_i

consider $\beta^{(d)} \in \mathbb{R}^p$ with specified $(p-d)$-dim subvector $= \mathbf{0}$ versus $\beta^{(d+1)}$ with an extra nonzero coeff., $\beta^{(d-1)}$ with an extra 0

if H_0 holds that d coefficients are really non-0 :

$2 \log \left(L(\widehat{\beta}^{(d+1)})/L(\widehat{\beta}^{(d)}) \right) \leq \chi^2_{1,\alpha}$ with prob. $\approx 1 - \alpha$

$2 \log \left(L(\widehat{\beta}^{(d)})/L(\widehat{\beta}^{(d-1)}) \right) > \chi^2_{1,\alpha}$ with prob. $\gg \alpha$ *power*

Idea: $\log L(\widehat{\beta}^{(j)}) - \frac{j}{2} \chi^2_{1,\alpha}$ likely maximized at $j = d$

AIC, BIC, ... $\min_j \left\{ -\log L(\widehat{\beta}^{(j)}) + cj \right\}$ for $\begin{cases} c = 2, & AIC \\ c = \log n, & BIC \end{cases}$

Computational Issues in Penalized MLE

Objective Function (to minimize): $-\log L(\widehat{\beta}^{(j)}) + c \cdot j$

(1) In large data and covariate sets, exact maximization not possible over all sets of variables. (SAS does best subset selection by default only when $p \leq 11$) ``forward'' or ``backward'' or ``both'' all **greedy algorithm** searches

(2) Choosing c too low results in Overfitting, often the problem with AIC. BIC value $c = \log n$ is probably as high as one should go. Script `RscriptLec16.RLog` shows an example where a value in-between is best as judged by 20-fold cross-validation.

Diagnostics for Goodness of Fit

Predictive accuracy is not the same as model-adequacy. Checking goodness of fit assesses whether deviations from a model within a defined larger class of models are no more than might occur by chance, by patternlessness of residuals.

(1) LRTs do this explicitly in a model class.

(2) **Binning** allows non-model-based checks on grouped data.

Bins partition data, by covariate-defined cells $A_h = \{i : X_i \in C_h\}$ or by predictor intervals $A_h = \{i : \widehat{\beta}' X_i \in C_h\}$, $C_h = (a_h, a_{h+1}]$

Diagnostic GLM comparison of $\sum_{i \in A_h} Y_i$ versus $\sum_{i \in A_h} \widehat{\mu}_i$.

Illustrated in `RscriptLec16.RLog` *on Breast-cancer data.*

Hosmer-Lemeshow Test

In the setting where bins involve X partition only , put:

$$t_{y,h} = \sum_{i \in A_h} Y_i \ , \qquad \widehat{t}_{y,h} = \sum_{i \in A_h} \widehat{\mu}_i \ , \qquad n_h = |A_h|$$

Hosmer-Lemeshow Statistic: $\sum_{h=1}^{H} (\widehat{t}_{y,h} - t_{y,h})^2 / \Big[\widehat{t}_{y,h} (1 - \widehat{t}_{y,h}/n_h) \Big]$

Idea: $t_{y,h}/n_h$ represents true expected fraction of 1's in Group h, which is roughly the proportion for each $i \in A_h$; however $\widehat{t}_{y,h}/n_h$ is a fitted proportion using all d parameters in the fitted GLM! Degrees of freedom not clear $(\geq H - d)$.

$$\sum_{h=1}^{H} (\widehat{t}_{y,h} - t_{y,h})^2 / \widehat{t}_{y,h} \ \text{ only } \textit{resembles} \ \ X^2, \qquad (\leq \chi^2_{H-d}).$$

Logistic Regression in $I \times 2$ Tables

Data: $Y_{i1} \sim$ Binom(n_i, π_i), $1 \le i \le I$, $j = 1,,$ $\quad Y_{i2} = n_i - Y_{i1}$

Model: H_1 : logit$(\pi_i) = \alpha + \beta x_i$, $\qquad H_0 : \beta = 0$

predictor scores x_i *describe 'distances' between* i *levels*

This is a 'test for trend' with ordinal categories, also a GLM Logistic Regression (can use `glm`).

Score test is equivalent to Cochran-Armitage trend test (derived using OLS) with statistic

$$z^2 = \Big[\sum_{i=1}^{I} (x_i - \bar{x}) Y_{i,1} \Big]^2 \Big/ \Big[p(1-p) \sum_{i=1}^{I} n_i (x_i - \bar{x})^2 \Big]$$

where $p = Y_{+1}/n$, $\bar{x} = \sum_{i=1}^{n} n_i x_i / n$. More powerful than test for independence against H_1 alternatives.

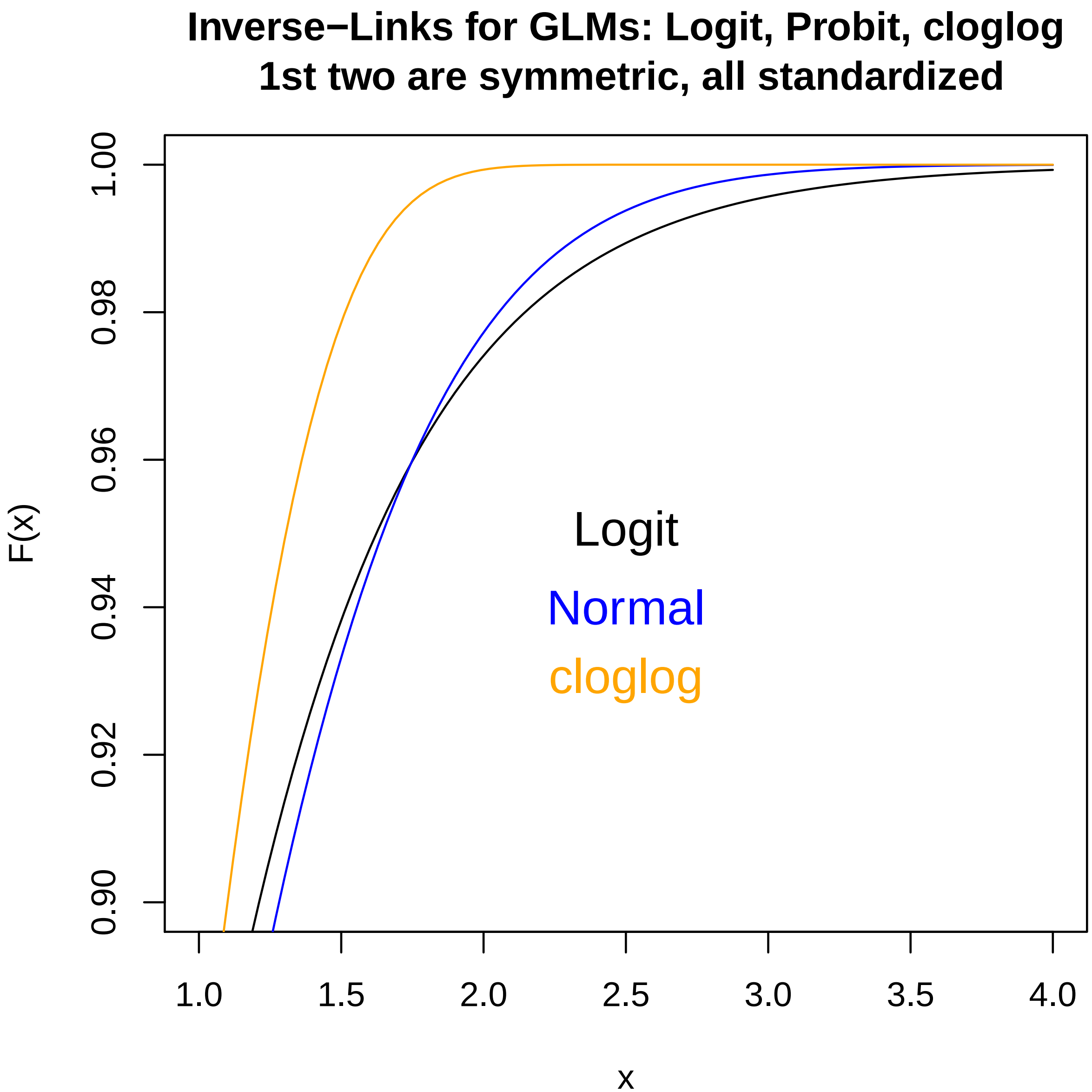
Other Models, Chapter 7

- `probit` and `cloglog` link binary-outcome GLMs

  Recall $g^{-1} = F$ could be any distribution function:
  $F = \Phi$ `probit`, $F(x) = 1 - \exp(-e^x)$ `cloglog`
  graphs on next page, example of fits in `RscriptLec16`

- Look at *conditional Logistic Regression* (sec. 7.3) next time

  Also look at (local) power and sample size
  formulas next time, Secs. 6.4 and 6.6.


Inverse−Links for GLMs: Logit, Probit, cloglog
1st two are symmetric, all standardized
F(x)
1.00
0.98
0.96
0.94
0.92
0.90
1.0
1.5
2.0
2.5
3.0
3.5
4.0
x
Logit
Normal
cloglog