Part II: Generalized Linear Models

Chapter II.3

Models for Binary Response

Logistic Regression with Categorical Predictors: Linear Logit Model

Logistic Regression Models Having All Predictors Categorical

II.3.31 Remark (contingency table)

In case all predictors $X_2, \dots, X_p$, for a binary response Y are categorical, the data are *grouped* in a p-way *contingency table* produced by cross-classifying the response and explanatory variables. The data form an $I_2 \times \dots \times I_p \times 2$ contingency table, where I_k is the number of categories of X_k, $k = 2, \dots, p$. The number of groups (i.e. table's cells) in the data set is $n = 2\prod_{k=2}^{p} I_k$.

II.3.32 Example

For a data set with a single predictor X of $I = 3$ levels, the data (left) form 3 groups, have weights m_i, $i = 1{,}2{,}3$, and can be expressed in a 3×2 contingency table form (right), where n_{i1} and n_{i2} is the number of successes $(Y = 1)$ and failures $(Y = 0)$ in the i-th group, respectively.

i	m_i	X	Y_i	n_{ij}
1	m_1	1	1	n_{11}
		1	0	n_{12}
2	m_2	2	1	n_{21}
		2	0	n_{22}
3	m_3	3	1	n_{31}
		3	0	n_{32}

	Y		
X	1	0	
1	n_{11}	n_{12}	$n_{1+} = m_1$
2	n_{21}	n_{22}	$n_{2+} = m_2$
3	n_{31}	n_{31}	$n_{3+} = m_3$

II.3.33 Remark

Let $\mathcal{M}$ be a binomial GLM for modeling a binary response in terms of $p-1$ categorical predictors having fixed number of categories. The likelihood ratio G^2 and Pearson's X^2 statistics for testing the fit of $\mathcal{M}$ based on a random sample, have both limiting χ^2_{df} distribution under $\mathcal{M}$, as the expected group counts (under $\mathcal{M}$) increase (s. Remark II.3.22).

II.3.34 Remark

Logit models are connected to *log–linear models*, as we shall see later on (s. II.5).

Logit Models for Two-Way Tables with a Binary Response

II.3.35 Logit Model

Consider an $I \times 2$ table with the column classification variable being the response Y (success-failure). Then, the probability of success, conditional on the level (row) i of the explanatory variable X is defined as:

$$\pi_i = P(Y = 1|x = i) = \frac{\pi_{i1}}{\pi_{i1} + \pi_{i2}},$$

while the *odds* of success will be $\frac{\pi_i}{1-\pi_i} = \frac{\pi_{i1}}{\pi_{i2}}$. Thus :

$$\text{logit}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i} = \log \frac{\pi_{i1}}{\pi_{i2}}.$$

The associated **logit model**

$$\text{logit}(\pi_i) = \log \left(\frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_{1(i)}, \ i = 1, \dots, I, \tag{II.5}$$

with parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_{1(2)}, \dots, \beta_{1(I)})' \in \mathbb{R}^I$ and $\beta_{1(1)} = 0$ (identifiability constraint).

II.3.36 Remark

In terms of success probabilities model (II.5) is expressed by:

$$\pi_i = \frac{\exp(\beta_0 + \beta_{1(i)})}{1 + \exp(\beta_0 + \beta_{1(i)})}, \; i = 1, \ldots, I.$$

II.3.37 Independence

In the set–up of II.3.35, the hypothesis of independence between X and Y is equivalent to

$$H_0 : \quad \pi_1 = \pi_2 = \ldots = \pi_I,$$

or, in terms of model (II.5), equivalent to

$$H_0 : \quad \beta_{1(1)} = \beta_{1(2)} = \ldots = \beta_{1(I)} = 0.$$

II.3.38 Remark (factor)

Note that though in a GLM every continuous predictor X_k contributes one column to the design matrix $\mathbf{X}$ and thus corresponds to one parameter β_k of the model, a categorical predictor X_q with I_q levels contributes $I_q - 1$ columns to $\mathbf{X}$, corresponding to dummy variables indicating the assignment to a category or not. For example, for $I_q = 3$, the corresponding columns would be $X_{q(2)}, X_{q(3)}$ and entries $(x_{iq(2)},x_{iq(3)})$ equal to (0,0), (1,0) or (0,1) for the i-th case (row of $\mathbf{X}$), correspond to $X_q(i)$ equal to 1, 2 or 3 respectively (a value (1,1) is not possible). Thus, to a categorical predictor of a model with I_q levels correspond $I_q - 1$ parameters.

II.3.39 Remark

The logit model (II.5) in II.3.35 is the binomial GLM with canonical link and in the classical GLM notation (s. Remark II.3.38), it is expressed as

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \sum_{c=2}^{I} \beta_c \mathsf{I}_d(c = i),\ i = 1, \ldots, I,$$

with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_I)' \in \mathbb{R}^I$ and I_d the indicator function. We adopt parameter notation of (II.5), since it is more informative in allocating parameters to specific factors, especially useful for models with multiple predictors.

Linear Logit Model for Binary Response

II.3.40 Linear Logit Model

In case the explanatory variable X is **ordinal**, one can assign (known) scores

$$x_1 \leq \ldots \leq x_I \quad (x_1 < x_I)$$

to its categories and consider the linear logit (LL) model:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \underbrace{\beta_1^* x_i}_{\beta_{1(i)}}, \; i = 1,\ldots,I, \tag{II.6}$$

with parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1^*)' \in \mathbb{R}^2$.

II.3.41 Remark

Note that the linear logit model (II.6) is a very parsimonious model, having just one parameter more than independence (s. II.3.37) while the logit model II.5 for two-way contingency tables is *saturated*.

II.3.42 Remark (odds ratio)

The odds for two levels of X, say i and i', are compared in terms of the *odds ratio*[a], considered for the 2×2 sub-table formed by rows i and i':

$$\theta_{ii'} = \frac{\pi_{i'}/(1-\pi_{i'})}{\pi_i/(1-\pi_i)} \ .$$

Under the linear logit model (II.6), it holds:

$$\log \theta_{ii'} = \text{logit}(\pi_{i'}) - \text{logit}(\pi_i) = \beta_1^*(x_{i'} - x_i) \ .$$

Thus, the linear logit model with *equidistant scores* for successive categories $(x_{i+1} - x_i = c)$ leads to

$$\theta_{i,i+1} = \theta = \exp(\beta_1^* c) \ , \quad i = 1, \ldots, I-1 \ ,$$

i.e. under this model the odds ratios for comparing successive levels of X are constant.

[a] The odds ratio (Yule, 1900, 1912) is a measure of association for 2×2 contingency tables and plays an important role in modeling of categorical data.

II.3.43 Remark

The idea of assigning scores to the categories of an ordinal explanatory variable (and reducing thus the parameters of the model) does directly extend to contingency tables with more than one ordinal explanatory variables, as well as in set-ups of explanatory variables of mixed type (continuous and categorical).

II.3.44 Example

Quality control for a sample of 500 products from 5 production machines, ordered from oldest to newest (A: old - E: new).

	Product. Machine					
Product	A	B	C	D	E	
defective	20	17	12	9	7	65
non-defective	80	83	88	91	93	435

If π_j is the *probability of a defective product for machine* j $(j = 1, \dots, 5)$, the independence model (I) applied on the table above, corresponds to testing $H_0\colon\ \pi_1 = \dots = \pi_5$.

☞ Verify that G^2(I)=10.51 (df=4, p-value=0.0327).

Does the logit of a defective product exhibit a linear trend with respect to the oldness order of the machines?

$$\text{logit}(\pi_j) = \log\left(\frac{\pi_j}{1-\pi_j}\right) = \log\left(\frac{\pi_{1j}}{\pi_{2j}}\right) = \beta_0 + \beta_1^* x_j, \quad x_j = j, \;\; j = 1, \ldots, 5\ .$$

The data need to be entered as:

```
> def <- c(20,17,12,9,7)
> nodef <- c(80,83,88,91,93)
> machine <- 1:5
> product <- data.frame(machine,def,nodef)
```

The model is then fitted by:

```
> lin.logit <- glm(cbind(def,nodef) ~ machine,
+          family=binomial(link="logit"), data=product)
```

```
> lin.logit
```

```
Call:
glm(formula = cbind(def, nodef)~machine, family = binomial(link = "logit"),
              data = product)
Coefficients:
(Intercept)  machine
-1.0383    -0.3111
Degrees of Freedom:  4 Total (i.e.  Null); 3 Residual
Null Deviance:  10.51
Residual Deviance:  0.1126 AIC: 25.13
```

This model (linear logit, LL) is of very good fit with $G^2 = 0.1126$ (p-value=0.9903, based on the χ_3 approximation under the LL model).

The estimated under the LL model probabilities of defective products are

```
> lin.logit$fitted.values
1          2          3          4          5
0.20597   0.15990   0.12222   0.09257   0.06954
```

Reporting Predictive Power

II.3.45 Classification Table

The predictive power of a logistic regression model can be summarized in terms of a *classification table*, which is a 2×2 table cross-classifying the observed binary response Y with the prediction $\hat{Y}$, based on this model. For an observation i, $\hat{y}_i = \mathsf{I}_d(\hat{\pi}_i > C)$, for some cutoff point C, where I_d is the indicator function and $\hat{\pi}_i$ is the estimated success probability $(\pi_i = P(Y_i = 1))$.

☞ Commonly, it is set $C = 0.5$.

II.3.46 Example II.3.12 (continues)

Revisiting the cancer remission example, recall that the vector `y` contains the responses (y_i's) while the model fitted was saved under the `glm` object ttfamily logit.fit. The $\hat{y}_i$'s can easily be calculated based on the vector of $\hat{\pi}_i$'s, derived by applying the `fitted()` function, as shown below:

```
> predicted <- round(fitted(logit.fit),0);  xtabs(∼ predicted + y)
              y
 predicted    0   1
          0  16   5
          1   2   4
```

II.3.47 Remark (leave-one-out cross-validation)

The true misclassification probabilities are underestimated by predicting the same observations used to fit the model. The leave-one-out cross-validation reduces this bias, by predicting the success probability for every specific observation (out of n in total) on the prediction based on fitting the model on the other $n-1$ observations. Thus, the model is fitted $K=n$ times.

II.3.48 Example II.3.12 (continues)

Continuing the discussion on the cancer remission data set, the cross-validated prediction error is the percentage of missclassified observations equals $\frac{2+5}{27} = 0.259$. A bias-correction of it based on leave-one-out cross-validation (number of samples used $K=n$) can be obtained in the boot package by the `cv.glm()` function:

```
> library(boot)
> carems <- data.frame(LI,y) # cv.glm() requires the data to be in a data frame
> cost <- function(r, pi=0) mean(abs(r-pi) > 0.5)
> cv <- cv.glm(carems, logit.fit, cost, K = nrow(carems)); cv$delta
[1] 0.2592593 0.2770919
```

II.3.49 ROC curve

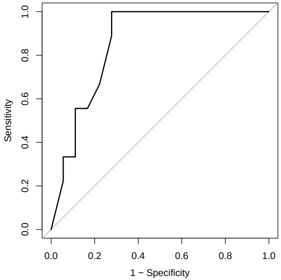
A *receiver operating characteristic* (ROC) curve is a plot of the *sensitivity* vs. the (1 − *specificity*) of the prediction for all possible cutoffs C, where

- prediction sensitivity $= P(\hat{Y} = 1 | Y = 1)$,
- prediction specificity $= P(\hat{Y} = 0 | Y = 0)$.

II.3.50 Example II.3.12 (continues)

```
> library(pROC);  roc.curve <- roc(y~fitted(logit.fit), data=carems)
```

```
> plot.roc(roc.curve, legacy.axes=TRUE)
> auc(roc.curve)
Area under the curve:  0.8549
```



II.3.51 Remark

- The ROC curve usually has a nearly concave shape connecting the points (0, 0) and (1,1).
- When the cutoff C is close to 0, then almost all predictions are equal to 1 and thus the sensitivity is near 1 and the specificity near 0. Analogously, for C close to 1, almost all predictions are 0 and hence sensitivity and specificity near to 0 and 1, respectively.
  ☞ The choice of C is based on 'balancing' sensitivity and specificity.
- The *area under the curve* (AUC) is a measure of the predictive power of the model:
  ☞ the greater the AUC, the higher the predictive power.
  AUC estimates the probability that the predicted and the observed responses are *concordant*.
  ☞ We expect that AUC$>$0.5, since AUC$<$0.5 means that predictions are worse than random guessing.