Logistic Regression

Pravat Uprety

Central Department of Statistics

Tribhuvan University

Models with dummy dependent variable

The model in which the dependent variable is a binary variable is also called binary choice model. In such a model, the dependent variable actually involves only two choices indicating presence or absence of an attribute or quality.

There are three important approaches to dealing with dummy dependent variable models or binary choice models and they are:

Linear Probability Model (LPM)

Logit Model

- Probit Model

Linear Probability Model

To know what determines car ownership status of the families in a locality and supposing that car ownership status of the families is determined by their family incomes. For this, following points need to be noted

i) Car status is a qualitative variable.

ii) However, car ownership status may be quantified by using a binary or dummy variable that is assigned value 1 if the family owns a car and value 0 if the family does not own a car.

iii) Family income is a usual quantitative variable

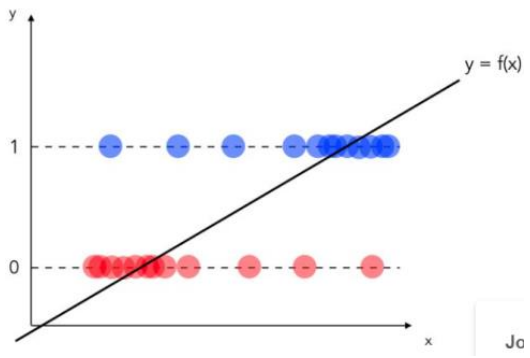
The model

$Y_i = \alpha + \beta X_i + \epsilon_i$ ----------(1)

- Where, X_i = Family income

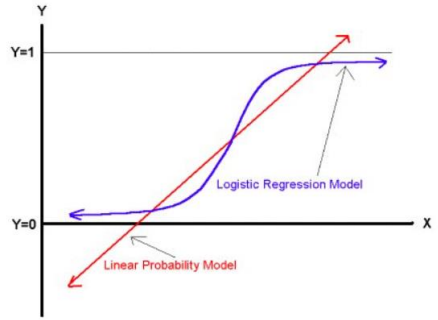
Y_i = 1 if the family owns a car

= 0 otherwise



Logit Model

- The linear probability model (LPM) reveals that when the dependent variable is a binary variable, a non linear specification of the model appears more appropriate. Specifically it seems appropriate to fit some kind of S – shaped (or sigmoid curve) to the observed data points as following graph.



The sigmoid curve has following features:

1) It resembles as elongated S.
2) The tails of the sigmoid curve level off before reaching p=0 or p =1 so that the problem of impossible values of estimated probability is avoided.
3) More importantly, the sigmoid curve, resemble the cumulative distribution function (CDF) of a random variables. So we can choose a suitable CDF to represent the sigmoid curve to capture 0-1 representation for the dependent variable.

Logistic Regression

- Binary logistic regression is a form of regression which is used when the dependent variable is a true or forced dichotomy and the independent variables are of any type.
- Logistic regression can be used to predict a categorical dependent variable on the basis of continuous and/or categorical independent variables; to determine the effect size of the independent variables on the dependent variable; to rank the relative importance of independent variables; to assess interaction effects; and to understand the impact of covariate control variables. The impact of predictor variables is usually explained in terms of odds ratios, which is the key effect size measure for logistic regression.

- Logistic regression applies maximum likelihood estimation after transforming the dependent into a logit variable.
- A logit is the natural log of the odds of the dependent equaling a certain value or not (usually 1 in binary logistic models, or the highest value in multinomial models).
- Logistic regression estimates the odds of a certain event (value) occurring. This means that logistic regression calculates changes in the log odds of the dependent, not changes in the dependent itself as does OLS regression.

Specification of logit model

$$P_i = P(Y_i=1) = F(z_i) = \frac{1}{1+e^{-z_i}}$$

Where P_i is the probability of $Y_i = 1$

$F(z_i)$ is the cdf of the cumulative logistic function

$Z_i = \alpha + \beta X_i$, is a predictor variable

e = 2.71828

$$1 - P_i = 1 - \frac{1}{1+e^{-z_i}} = \frac{1+e^{-z_i}-1}{1+e^{-z_i}} = \frac{e^{-z_i}}{1+e^{-z_i}} = \frac{\frac{1}{1+e^{-z_i}}}{\frac{e^{-z_i}}{1}} = \frac{1}{1+e^{z_i}}$$

$$\frac{P_i}{1-P_i} = \frac{1}{1+e^{-z_i}} / \frac{1}{1+e^{z_i}} = \frac{1+e^{z_i}}{1+e^{-z_i}} = \frac{1+e^{z_i}}{1+\frac{1}{e^{z_i}}} = \frac{1+e^{z_i}}{\frac{e^{z_i}+1}{e^{z_i}}} = e^{z_i}$$

So $\ln\left(\frac{P_i}{1-P_i}\right) = Z_i = \alpha + \beta X_i$ ----------(2)

Hence, $\frac{P_i}{1-P_i}$ is called odds ratio in favour of the event occurring and ln ($\frac{P_i}{1-P_i}$) is the log odds ratio

(also called logit of p)

In model (2) we considered only one explanatory variable. But, of necessary, more explanatory variables can be added by supposing that Z_i is a linear function of a set of predictor variables.

$Z_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_3 X_{3i} \ldots\ldots\ldots\ldots + \beta_k X_{ki}$

$$\log\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_3 X_{3i} \ldots\ldots\ldots\ldots + \beta_k X_{ki}$$

The logistic regression model is a special case of a GLM. The random component for the (success, failure) outcomes has a binomial distribution. The link function of $\pi = P(Y = 1)$ is the *logit* function, $\log[\pi/(1-\pi)]$, symbolized by "logit(π)." Logistic regression models are often called *logit models*. Whereas $P(Y = 1)$ is restricted to the 0 to 1 range, the logit can be any real number. The real numbers are also the potential range for linear predictors. This model therefore does not have the structural limitation that the linear probability model has.

Estimation of logit model

It is not possible to estimate the logit (2) by OLS method for two reasons

This is a non linear model

$\ln(\frac{P_i}{1-P_i})$ is not a familiar quantity.

Logistic regression has many analogies to OLS regression:

logit coefficients correspond to b coefficients in the logistic regression equation;

the standardized logit coefficients correspond to beta weights;

and a pseudo R2 statistic is available to summarize the overall strength of the model.

Unlike OLS regression, however, logistic regression does not assume linearity of relationship between the raw values of the independent variables and raw values of the dependent; does not require normally distributed variables; does not assume homoscedasticity; and in general has less stringent requirements.

Assumptions of logistic regression

Results: Models the natural log of the odds (logits) of success probability

Associated Dependent Variable: Dichotomous

Probability Distribution: Binomial

Link: Logit

Estimator: Maximum likelihood

Assumptions: 1. Non-autocorrelation. $Cov(\varepsilon_{i,}, \varepsilon_j) = 0$ for $i \neq j$

2. Linear relationship between the predictors and the log odds (logit) of the dependent variable.

3. Absence of high partial multicollinearity.

4. Large samples.

5. Adequate expected cell frequencies.

Measuring goodness of fit

Effron's R^2

Effron revised the R^2 formula that is used in the context of OLS regression in such a way that the binary feature of dependent variable is taken

$$\text{Effron's } R^2 = 1 - \frac{n}{n_1 n_2} \Sigma(Y - \hat{Y})^2$$

McFadden's Pseudo R^2

McFadden's Pseudo R^2 is a very popular measure of goodness of fit in the context of binary dependent variable models. It is also known as the log-likelihood ratio index (LRI). It is called Pseudo R^2 because although there is no measure equivalent to R^2 in the maximum likelihood method, it tends to provide an interpretation similar to R^2 statistic.

The Pseudo R^2 formula is obtained by comparing the value of log likelihood of initial regression model (ln L) with the value of log likelihood that would have been obtained with only the intercept term in the regression model (ln L_0).

It is defined as

$$\text{Pseudo } R^2 = 1 - \frac{\ln L}{\ln L_0}$$

Examining the overall significance of regression

- For the purpose of examining the overall significance of logit/probit models, we may use the likelihood ratio statistic (LR-statistic). Two important feature of LR statistic are: (i) it uses variations in likelihood as a basis for test, and (ii) it is useful to assess the explanatory poser of the model, i.e., to understand overall significance of regression.

$$LR = 2 \ln \frac{L}{L_0} = 2 (\ln L - \ln L_0)$$

It has been shown that LR ~ χ^2 with degrees of freedom K (number of explanatory variables).

We reject null hypothesis if χ^2 > Critical (Tab) value of χ^2.

Odds: An odds is a ratio formed by the probability that an event occurs divided by the probability that the event does not occur. In binary logistic regression, the odds is usually the probability of getting a “1” divided by the probability of getting a “0”.

Odds ratio: An odds ratio is the ratio of two odds, such as the ratio of the odds for men and the odds for women. Odds ratios are the main effect size measure for logistic regression, reflecting in this case what difference gender makes as a predictor of some dependent variable. An odds ratio of 1.0 (which is 1:1 odds) indicates the variable has no effect. The further from 1.0 in either direction, the greater the effect.

Log odds: The log odds is the coefficient predicted by logistic regression and is called the “logit”. It is the natural log of the odds of the dependent variable equaling some value (ex., 1 rather than 0 in binary logistic regression). The log odds thus equals the natural log of the probability of the event occurring divided by the probability of the event not occurring: ln(odds(event)) = ln(prob(event)/prob(nonevent))

Logit: The “logit function” is the function used in logistic regression to transform the dependent variable prior to attempting to predict it. Specifically, the logit function in logistic regression is the log odds, explained above. The “logit” is the predicted value of the dependent variable. “Logit coefficients” are the b coefficients in the logistic equation used to arrive at the predicted value.

Parameter estimates: These are the logistic (logit or b) regression coefficients for the independent variables and the constant in a logistic regression equation, much like the b coefficients in OLS regression. Synonyms for parameter estimates are unstandardized logistic regression coefficients, logit coefficients, log odds-ratios, and effect coefficients. Parameter estimates are on the right-hand side of the logistic regression equation and logits are on the left-hand side. The logistic regression equation itself is:

$z = b0 + b1X1 + b2X2 + ... + bkXk$

Where z is the log odds of the dependent variable = ln(odds(event)). The "z" is the “logit”, also called the log odds.

- Exp(b) = the odds ratio for an independent variable = the natural log base e raised to the power of b. The odds ratio of an independent variable is the factor by which the independent variable increases or (if negative) decreases the log odds of the dependent variable The term “odds ratio” usually refers to odds ratios for independent variables.

TEST FOR INDIVIDUAL PREDICTORS

- In logistic regression there is an analogous statistics known as the Wald statistics. Wald statistics tell us whether the beta coefficient for that predictor is significantly different from zero. If the coefficient is significantly different from zero then we can say that the predictor is making a significant contribution to the prediction of the outcome (Y).

The Wald statistic is defined as

$$W = \frac{\hat{\beta}_i^{\,2}}{\{S.E(\hat{\beta}_i)\}^2}$$

Summarizing Predictive Power: Classification Tables

A *classification table* cross-classifies the binary outcome y with a prediction of whether $y = 0$ or 1. The prediction for observation i is $\hat{y} = 1$ when its estimated probability $\hat{\pi}_i > \pi_0$ and $\hat{y} = 0$ when $\hat{\pi}_i \leq \pi_0$, for some cutoff π_0. One possibility is to take $\pi_0 = 0.50$. However, if a low (high) proportion of observations have $y = 1$, the model fit may never (always) have $\hat{\pi}_i > 0.50$, in which case one never (always) predicts $\hat{y} = 1$. Another possibility takes π_0 as the sample proportion of 1 outcomes, which is $\hat{\pi}_i$ for the model containing only an intercept term.

Two useful summaries of predictive power are

$$\text{Sensitivity} = P(\hat{y} = 1 \mid y = 1)$$

$$\text{Specificity} = P(\hat{y} = 0 \mid y = 0)$$

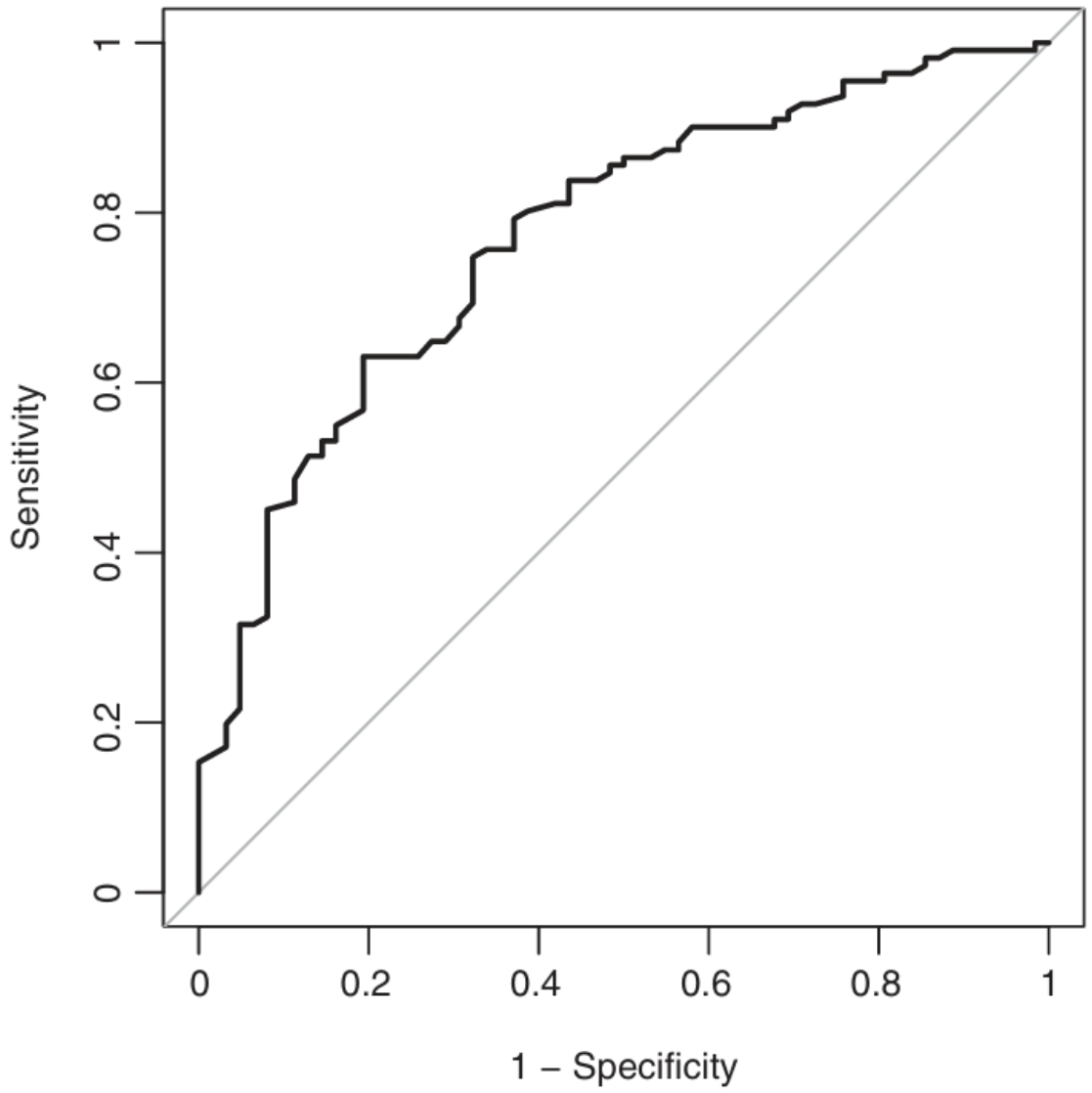
$$\begin{aligned} P(\text{correct classif.}) &= P(y = 1 \text{ and } \hat{y} = 1) + P(y = 0 \text{ and } \hat{y} = 0) \\ &= P(\hat{y} = 1 \mid y = 1)P(y = 1) + P(\hat{y} = 0 \mid y = 0)P(y = 0) \\ &= \text{sensitivity}[P(y = 1)] + \text{specificity}[1 - P(y = 1)], \end{aligned}$$

Summarizing Predictive Power: ROC Curves

A *receiver operating characteristic* (ROC) curve is a plot that shows the sensitivity and the specificity of the predictions for all the possible cutoffs π_0. This curve is more informative than a classification table, because it summarizes predictive power for all possible π_0.

The ROC curve plots sensitivity on the vertical axis versus (1 − specificity) on the horizontal axis. When π_0 gets near 0, almost all predictions are $\hat{y} = 1$; then, sensitivity is near 1, specificity is near 0, and the point for (1 − specificity, sensitivity) has coordinates near (1, 1). When π_0 gets near 1, almost all predictions are $\hat{y} = 0$; then, sensitivity is near 0, specificity is near 1, and the point for (1 − specificity, sensitivity) has coordinates near (0, 0). The ROC curve usually has a concave or nearly concave shape connecting the points (0, 0) and (1, 1).

For a particular value of specificity, better predictive power corresponds to higher sensitivity. Therefore, the better the predictive power, the higher is the ROC curve. Because of this, the area under the curve provides a single value that summarizes predictive power. The greater the area, the better the predictive power. This measure of predictive power is called the *concordance index*. Consider all pairs of observations (i, j) such that $y_i = 1$ and $y_j = 0$. The concordance index estimates the probability that the predictions and the outcomes are *concordant*, which means that the observation with the larger y also has the larger $\hat{\pi}$. A concordance value of 0.50 means predictions were no better than random guessing.


Sensitivity
0
0.2
0.4
0.6
0.8
1
1 – Specificity
0
0.2
0.4
0.6
0.8
1

HOSMER- LEMESHOW TEST

- The goodness of fit (the degree of closeness of the model-predicted value to the corresponding observed value) is useful for applying to the regression model. The Hosmer-Lemeshow goodness of fit statistics (Hosmer & Lemeshow, 2000) proposes a Pearson's statistic to compare the observed and fitted counts for the partition.
- In general, the Hosmer-Lemeshow goodness of fit test divides subject into deciles based on predicted probabilities and computes a chi square from observed and expected frequencies.

Example and interpretation

Step	Chi-square	df	Sig.
1	1.228	3	.746

We observed from the example, that the Hosmer-Lemeshow chi-square statistic with 3 degree of freedom is 1.228 with p value 0.746. The large p value signifies that there is no difference between the observed and predicted values, implying that the model fits the data at an acceptable level.

Nagelkerke R2

- Coefficient of determination (R^2)is the proportion of the variation in the dependent variable that can be explained by predictors in the model. While coefficient of determination (R^2) is a very valuable measure of how well linear regression model fits, it is less useful in logistic regression.
- As logistic regression is not a linear model, we can't calculate R^2 directly as for linear regression model. However, test like -2log likelihood gives the pseudo R^2 (Nagelkerke R^2) statistics which are based on comparing the likelihood of the current model to the null model (without any predictors).A large pseudo R^2 indicate that more of the variation is explained by the model, from a minimum of 0 to maximum of 1. It should also be noted that pseudo R^2 values tend to be very low for logistic regression model, much lower than for linear regression model. This is because we are trying to predict the outcome where as the model only given us the probability of outcomes.

Logistic regression (command in R)

- ##Open MROZ.dta data file
- head(MROZ)

Linear Probability Model

- linprob<-lm(inlf~nwifeinc+educ+exper+I(exper^2)+age+kidslt6+kidsge6, data=MROZ)
- summary(linprob)
- ###for prediction.
- pred<-list(nwifeinc=c(100,0),educ=c(5,15), exper=c(0,10),age= c(20,52),kidslt6=c(2,0), kidsge6=c(0,0))
- predict(linprob,pred)

Logistic model

- ###Logistic regression
- glm(formula=inlf~nwifeinc+educ+exper+I(exper^2)+age+kidslt6+kidsge6,family=binomial(link=logit), data=MROZ)
- logit_femp<-glm(formula=inlf~nwifeinc+educ+exper+I(exper^2)+age+kidslt6+kidsge6,family=binomial(link=logit), data=MROZ)
- summary(logit_femp)

Logit model odds ratios

- exp(logit_femp$coefficients)

###Log likelihood

- logLik(logit_femp)

###McFadden's pseudo R2

1-logit_femp$deviance/logit_femp$null.deviance

library(lmtest)

lrtest(logit_femp)

Data File: Smoking

- Run the logistic regression taking smoker as a dependent variable and age, educ, income and pcigs79 as independent variables.

Data file: Smoking

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.745e+00  8.292e-01   3.311 0.000931 ***
age         -2.085e-02  3.739e-03  -5.577 2.44e-08 ***
educ        -9.097e-02  2.067e-02  -4.402 1.07e-05 ***
income       4.720e-06  7.170e-06   0.658 0.510356
pcigs79     -2.232e-02  1.247e-02  -1.789 0.073538 .
---
```

Meaning

- The variables age and education are highly statistically significant and have the expected signs.
- As age increases, the value of the logit decreases, that is, as people age, they are less likely to smoke.
- More educated people are less likely to smoke, perhaps due to the ill effects of smoking.
- The price of cigarettes has the expected negative sign and is significant at about the 10% level. The higher the price of cigarettes, the lower is the probability of smoking.
- Income has no statistically visible impact on smoking, perhaps because expenditure on cigarettes may be a small proportion of family income.

Meaning of coefficient

- holding other variables constant, if, for example, education increases by one year, the average logit value goes down by 0.09 , that is, the log of odds in favor of smoking goes down by about 0.09.

- Other coefficients are interpreted similarly.