Binary Logit/Probit

Murthi

Questions

- A medical researcher is interested in predicting the probability of getting a heart attack knowing the blood pressure, cholesterol, calorie intake, gender and physical activity
- Predict whether a household would subscribe to a package of premium channels
- A credit card issuing bank would like to predict the probability that a customer will default

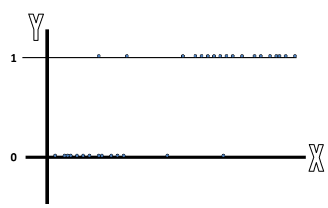
What other prediction problems can you think of in each area?

- Insurance industry
- Finance industry
- Healthcare industry
- Retail industry
- Web analytics

When to use these methods?

- In all the previous situations, notice that the dependent variable is discrete and binary.
- If Y is continuous then regression is used.
- If Y is binary and discrete, then logit/probit can be used.
- If Y has more than 2 levels and is discrete, then use
  - Multinomial Logit (MNL) or
  - Multinomial Probit (MNP)
  - e.g., Which brand will a person choose (A, B, C, or D)
  - Which major will an MBA student choose?

Scatterplot of with Y=(0,1): Y = Hired-Not Hired; X= Experience

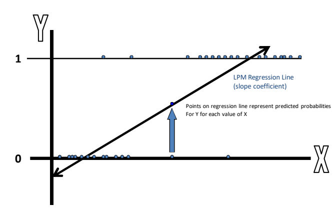


The Linear Probability Model (LPM)

If we estimate the slope using OLS regression:
Hired = α + β*Exper + e ;

- The result is called a "Linear Probability Model"
  - The predicted values are probabilities that Y equals 1;
  - The equation is linear – the slope is constant

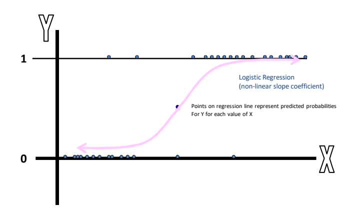
Picture of LPM



LPM Weaknesses

- The predicted probabilities can be greater than 1 or less than 0
- The error terms vary based on size of X-variable ("heteroskedastic") –
  - There may be models that have lower variance – more "efficient"
- The errors are not normally distributed because Y takes on only two values

S-shape - Logistic Regression



- Suppose our underlying dummy dependent variable depends on an unobserved utility index, *Y**
- *Y* is discrete—taking on the values 0 or 1 if someone buys a car, for instance
- Imagine a continuous variable *Y** that reflects a person's desire to buy the car
- *Y** would vary continuously with some explanatory variables like income, age

---

- Utility can be written as:
  - $Y_i^* = a + bX_{1i} + \varepsilon_i$
- If the utility index is "high enough," a person will buy a car
  - $Y_i = 1$, if $Y_i^* > 0$
- If the utility index is not "high enough," a person will not buy a car
  - $Y_i = 0$, if $Y_i^* <= 0$

$$\begin{aligned}
P_i &= \text{Prob}(Y_i = 1) \\
&= \text{Prob}(Y_i^* \geq 0) \\
&= \text{Prob}(\beta_0 + \beta_1 X_{1i} + \varepsilon_i \geq 0) \\
&= \text{Prob}(\varepsilon_i \geq -\beta_0 - \beta_1 X_{1i}) \\
&= 1 - F(-\beta_0 - \beta_1 X_{1i}) \text{ where } F \text{ is the c.d.f. for } \varepsilon \\
&= F(\beta_0 + \beta_1 X_{1i}) \text{ if } F \text{ is symmetric}
\end{aligned}$$

If cdf is Normal distribution, we get a probit model
If cdf is logistic distribution, then we get a logit model.

- Estimation of the β's typically done using a maximum likelihood estimator (MLE)

- Each outcome *Yi* has the density function *f*(*Yi*)
- $= P_i^{Yi} (1 - P_i)^{1 - Yi}$
- Each *Yi* takes on either the value of 0 or 1 with
- probability $f(0) = (1 - P_i)$ and $f(1) = P_i$

Likelihood function

$$\begin{aligned}
\ell &= f(Y_1, Y_2, \ldots Y_n) \\
&= f(Y_1) f(Y_2) \ldots f(Y_n) \\
&= P_1^{Y_1}(1 - P_1)^{1-Y_1} P_2^{Y_2}(1 - P_2)^{1-Y_2} \ldots P_n^{Y_n}(1 - P_n)^{1-Y_n} \\
&= \prod_{i=1}^{n} P_i^{Y_i}(1 - P_i)^{1-Y_i}
\end{aligned}$$

and

$$\ln \ell = \sum_{i=1}^{n} Y_i \ln P_i + (1 - Y_i) \ln(1 - P_i)$$

which, given $P_i = F(\beta_0 + \beta_1 X_{1i})$, becomes

$$\ln \ell = \sum_{i=1}^{n} Y_i \ln F(\beta_0 + \beta_1 X_{1i}) + (1 - Y_i) \ln(1 - F(\beta_0 + \beta_1 X_{1i}))$$

- For the logit model we specify

$$\text{Prob}(Y_i = 1) = F(\beta_0 + \beta_1 X_{1i}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i})}}$$

- Prob(*Yi* = 1) → 0 as $\beta_0 + \beta_1 X_{1i} \rightarrow -\infty$
- Prob(*Yi* = 1) → 1 as $\beta_0 + \beta_1 X_{1i} \rightarrow \infty$
- Thus, probabilities from the logit model will be between 0 and 1

- A complication arises in interpreting the estimated β's
- With a linear probability model, a β estimate measures the *ceteris paribus* effect of a change in the X variable on the probability *Y* equals 1
- In the logit model

$$\begin{aligned}
\frac{\partial \text{Prob}(Y_i = 1)}{\partial X_1} &= \frac{\partial F(\hat{\beta}_0 + \hat{\beta}_1 X_{1i})}{\partial X_1} \hat{\beta}_1 \\
&= \frac{\hat{\beta}_1 e^{-(\beta_0 + \beta_1 X_{1i})}}{[1 + e^{-(\beta_0 + \beta_1 X_{1i})}]^2}
\end{aligned}$$

Probit Model

- In the probit model, we assume the error in the utility index model is normally distributed
  - $\varepsilon_i \sim N(0, \sigma^2)$

$$\text{Prob}(Y_i = 1) = F\left(\frac{\beta_0 + \beta_1 X_{1i}}{\sigma}\right)$$

- Where *F* is the standard normal cumulative
- density function (c.d.f.)

$$\text{Prob}(Y_i = 1) = F\left(\frac{\beta_0 + \beta_1 X_{1i}}{\sigma}\right) = \int_{-\infty}^{\frac{\beta_0 + \beta_1 X_{1i}}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

Which is better – logit/probit?

- From an empirical standpoint logits and probits typically yield similar estimates of the relevant derivatives
  - Because the cumulative distribution functions for the two models differ slightly only in the tails of their respective distributions
- The derivatives are different only if there are enough observations in the tail of the distribution
- While the derivatives are usually similar, the parameter estimates associated with the two models are not
- Multiplying the logit estimates by 0.625 makes the logit estimates comparable to the probit estimates

Are stocks of larger firms favored?

Favored Stock		Less Favored Stock	
Success	Size	Success	Size
1	1	0	1
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	0	0	0
1	0	0	0

Contingency Table

Type of Stock	Large	Small	Total
Preferred	10	2	12
Not Preferred	1	11	12
Total	11	13	24

Basic Concepts

- Probability
  - Probability of being 'preferred' stock = 12/24 = 0.5
  - Probability that a company's stock is preferred given that the company is large = 10/11 = 0.909
  - Probability that a company's stock is preferred given that the company is small = 2/13 = 0.154

Odds and Probability

- Odds(Event) = Prob(Event)/(1-Prob(Event))
  O = p/(1-p)
- Prob(Event) = Odds(Event)/(1+Odds(Event))
  p = O/(1+O)

Concepts … contd.

- Odds
  - Odds of a preferred stock = 12/12 = 1
  - Odds of a preferred stock given that the company is large = 10/1 = 10
  - Odds of a preferred stock given that the company is small = 2/11 = 0.182

Logistic Regression

- Take Natural Log of the odds:
  - ln(odds(Preferred|Large)) = ln(10) = 2.303
  - ln(odds(Preferred|Small)) = ln(0.182) = -1.704
- Combining these relationships
  - ln(odds(Preferred|**Size**)) = -1.704 + 4.007***Size**
  - Log of the odds (or logit) is a linear function of size
  - The coefficient of size can be interpreted like the coefficient in regression analysis
  - i.e., For 1 unit increase in Size, change in log-odds=4.007
  - But what does it mean in plain english?

General Model

- ln(odds) = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_K$ (1)
- Recall:
  - Odds = p/(1-p)
- ln(p/1-p) = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_K$ (2)
- p = $\dfrac{e^{\beta_0+\beta_1 X_1}}{1+e^{\beta_0+\beta_1 X_1}}$
- OR
- p = $\dfrac{1}{1+e^{-(\beta_0+\beta_1 X_1)}}$

Interpretation

- If size=10, log-odds (LO)= -1.704 + 4.007*10
- If size=11, log-odds (LO)= -1.704 + 4.007*11
- Difference(LO_{11}-LO_{10})=4.007
- Take exp on both sides:
- Exp((LO_{11}-LO_{10}))=exp(4.007)
- $Odds_{11}$/$Odds_{10}$ = exp(4.007)
- For a unit change in size, Odds will increase by a multiple of 55 times.
- $Odds_{11}$ = exp(4.007)*$Odds_{10}$
- $Odds_{11}$ - $Odds_{10}$ = (exp(4.007)-1)$Odds_{10}$
- Percentage change in odds for a unit change in size =
- (exp(4.007)-1)*100

Interpretation of discrete variable coefficients

- If X is gender (M=1, F=0)
- $Odds_M$/$Odds_F$ =exp(β)
- Percentage increase in $Odds_M$ = (exp(β)-1)*100

Likelihood

- This is the joint probability of observing the 1s and 0s in the sample of data.
- For one observation y=1, lik = p_1
- For two observations y={1,0}, lik=p_1p_0
- If y={1011001}, lik=$p_1p_0p_1p_1p_0p_0p_1 = p_1^4\, p_0^3$
- Log-lik=log($p_1^4\, p_0^3$) = 4*log(p_1)+3*log(p_0)
- If there are 50 purchases among a sample of 300, loglik=50*log(50/300)+250*log(250/300)= -135.17
- Maximum value of log-lik = 0.

Interpretation of Results - Model Fit

Look at the –2 Log L statistic

- Null Model: Intercept only model (i.e. no X variables): 33.271
- Intercept and Covariates: 17.864
- Difference: 15.407 with 1 DF (p=0.0001)
- Means that the size variable is explaining a lot
- McFadden's R-sq = diff. in (-2LogL)/Null model's (-2logL)
  = 15.4/33.27 =46%

Do the Variables Have a Significant Impact?

- Like testing whether the coefficients in the regression model are different from zero
- Look at the output from Analysis of Maximum Likelihood Estimates
  - Loosely, the column Pr>Chi-Square gives you the probability of realizing the estimate in the Parameter estimate column if the estimate were truly zero – if this value is < 0.05 the estimate is considered to be significant

Model fit

- Akaike's Information Criterion (AIC), Schwartz's Criterion (SC or BIC)
  - like Adj-R^2 applies a penalty
  - there is a penalty for having additional covariates
- AIC = [-2log*L* + 2p]
- SC (or BIC) = [-2log*L* + plog(n)]
- BIC penalizes more heavily
- Model with lower AIC/BIC is better.

Predicted Probabilities and Observed Responses

- The response variable (success) classifies an observation into an *event* or a *no-event*
- A concordant pair is defined as that pair formed by an *event* with a PHAT higher than that of the *no-event*
- *Higher the concordance %, the better the model*
- Hit ratio = *% events correctly classified*

Classification

- For a set of new observations where you have information on size alone
- You can use the model to predict the probability that success = 1 i.e. the stock is favored
- If PHAT > 0.5 success = 1 else success=0

Logistic model

- Own price elasticity = (1-prob(j))*X_j*β
- Cross price elasticity = (-prob(j))*X_j*β

Model fit indicators

- **AIC -** Akaike Information Criterion.
- AIC = -2 Log L + 2((*k*-1) + *s*),
- where *k* is the number of levels of the dependent variable and *s* is the number of predictors in the model. The model with the smallest **AIC** is considered the best.
- **SC -** Schwarz Criterion.
- SC = - 2 Log L + ((*k*-1) + *s*)*log(Σ f_i),
- where f_i's are the frequency values of the i^{th} observation, and *k* and *s* were defined previously. Smallest **SC** is most desirable.
- **-2 Log L** - is used in hypothesis tests for nested models.

McFadden's R^2

- M_{full} = Model with predictors
- $M_{intercept}$ = Model without predictors

$$R^2 = 1 - \frac{\ln \hat{L}(M_{Full})}{\ln \hat{L}(M_{Intercept})}$$

$$R_{adj}^{\;2} = 1 - \frac{\ln \hat{L}(M_{Full}) - K}{\ln \hat{L}(M_{Intercept})}$$

- https://stats.idre.ucla.edu/sas/output/proc-logistic/

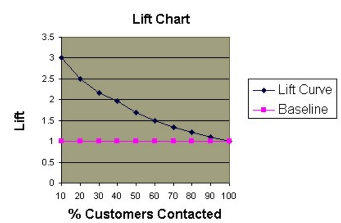
Direct Marketing

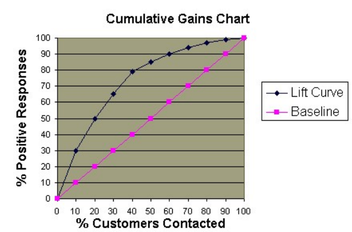
- A company wants to do a mail marketing campaign.
- costs = $1 for each item mailed.
- They have information on 100,000 customers.
- Overall response rate = 20%
- If all 100,000 customers are contacted then we will get 20,000 responses.

- Develop a logit model to predict response.
- Sort the customers based on probability of response and divide them into deciles.

Data to plot lift chart and Cumulative Gains chart

Cost	Contacts	Positive responses with model	Positive responses No model	lift	Cumulative gains chart = positive responses/total responses
10000	10000	6000	2000	3	6000/20000=30%
20000	20000	10000	4000	2.5	10000/20000 = 50%
30000	30000	13000	6000	2.166	13000/20000 = 65%
40000	40000	15800	8000	1.975	79%
50000	50000	17000	10000	1.70	85%
60000	60000	18000	12000	1.50	90%
70000	70000	18800	14000	1.343	94%
80000	80000	19400	16000	1.212	97%
90000	90000	19800	18000	1.10	99%
100000	100000	20000	20000	1.00	100%





Predicted	Actual	Label
0	0	True Negative (TN)
0	1	False Negative (FN)
1	0	False Positive (FP)
1	1	True Positive (TP)

	Predicted = 1	Predicted =0
Actual=1	10	2
Actual=0	3	35

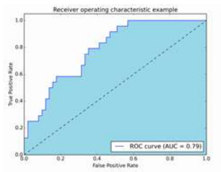
- TP+FN=P (number of actual 1s)
- TN+FP=N (number of actual 0s)

- FP rate=FP/N = fall-out = % neg. mis-classified
- TP rate = TP/P = Recall or hit rate = % pos. correctly classified
- Precision = TP/(TP+FP)
- Accuracy = (TP+TN)/(P+N)

- Percent concordance
- Pick a random 1 and a random 0
- Is the model predicted Prob(1)> Prob(0)? = concordance
- Ot
- Otherwise it is discordant.
- The AUC is the probability that the classifier will rank a random positive example higher than a randomly chosen negative example.
- (P(score(x+) > score(x-))

AUC: Receiver Operating Characteristic (ROC curve)

- Plot TPR against FPR to get ROC curve
- Dotted line indicates ROC curve of a random predictor.
- ROC >0.5 or ROC<0.5 is good.
- ROC=1 and ROC =0 are the best.
- Use ROCR package in R



SAS code

```
data Data1;
input disease n age;
datalines;
0 14 25
0 20 35
0 19 45
7 18 55
6 12 65
17 17 75 ;
ods graphics on;
proc logistic data=Data1 plots(only)=(roc(id=obs) effect);
model disease/n=age / scale=none clparm=wald clodds=pl rsquare;
units age=10;
run;
ods graphics off;
```

```
data roc;
   input alb tp totscore popind @@;
   totscore = 10 - totscore;
   datalines;
 3.0 5.8 10 0   3.2 6.3  5 1   3.9 6.8  3 1   2.8 4.8  6 0
 3.2 5.8  3 1   0.9 4.0  5 0   2.5 5.7  8 0   1.6 5.6  5 1
 3.8 5.7  5 1   3.7 6.7  6 1   3.2 5.4  4 1   3.8 6.6  6 1
 4.1 6.6  5 1   3.6 5.7  5 1   4.3 7.0  4 1   3.6 6.7  4 0
 2.3 4.4  6 1   4.2 7.6  4 0   4.0 6.6  6 0   3.5 5.8  6 1
 3.8 6.8  7 1   3.0 4.7  8 0   4.5 7.4  5 1   3.7 7.4  5 1
 3.1 6.6  6 1   4.1 8.2  6 1   4.3 7.0  5 1   4.3 6.5  4 1
 3.2 5.1  5 1   2.6 4.7  6 1   3.3 6.8  6 0   1.7 4.0  7 0
 3.7 6.1  5 1   3.3 6.3  7 1   4.2 7.7  6 1   3.5 6.2  5 1
 2.9 5.7  9 0   2.1 4.8  7 1   2.8 6.2  8 0   4.0 7.0  7 1
 3.3 5.7  6 1   3.7 6.9  5 1   3.6 6.6  5 1
 ;
```

```
ods graphics on;
  proc logistic data=roc plots=roc(id=prob);
    model popind(event='0') = alb tp totscore / nofit;
    roc 'Albumin' alb;
    roc 'K-G Score' totscore;
    roc 'Total Protein' tp;
    roccontrast reference('K-G Score') / estimate e;
  run;
  ods graphics off;
```

NOFIT option to prevent PROC LOGISTIC from fitting the model with three covariates.
The ROCCONTRAST statement implements the nonparameteric approach of DeLong, DeLong, and Clarke-Pearson (1988) to compare the three ROC curves.
REFERENCE option specifies that the K-G Score curve is used as the reference curve in the contrast
The E option displays the contrast coefficients, and the ESTIMATE option computes and tests each comparison

Rare events

- In a binary variable (response/no-response, good/bad, default/no-default, purchase/no-purchase, etc.) one of the two events is rare.
- In a sample of 1000 applicants, only 20 are selected – low event rate of 2%▪
- In a sample of 100,000 purchases from an online retailer, about 1800 are returned by the customer – low event rate of 1.8%
- Some real life examples:
  - Percentage of defaulters in credit card transactions
  - Goods returned in online retailing
- Why is this a problem for logistic regression?
- The usual maximum-likelihood estimation method is susceptible to ‘small sample bias’ and this bias is strongly dependent on the count (as opposed to percentage) of the rarer of the events.

Exact logistic regression

- Solutions
  - Exact logistic regression
  - Computationally intensive.
  - Good for small samples or unbalanced data with few covariates (<200)

```
Proc logistic data = a1 descending;
Freq cellcount; /* cellcount is the weight variable here */
Model y=x1 x2;
Exact x1/estimate=both;
run;
```

Alternative code:

```
Proc logistic data = a1 descending;
Model y/cellcount =x1 x2;
Exact x1/estimate=both;
run;
```

Penalized likelihood (Firth’s method)

- If you have a larger count of the rare events say 1000 in a sample of 100,000 you can use logistic regression using the penalized likelihood approach (Firth).

```
Proc logistic data = a1;
Class catvar1 catvar2/param=ref;
Model y = catvar1 catvar2 x1 x2/firth;
run;
```

Oversampling

- Create a sample with many examples of rare events (say 500). Match it with a sample of non-occurrence of the event cases in approximately 50-50% ratio or even 33-66% ratio.
- Then estimate the logit model.