APS 425 – Fall 2015

Qualitative Dependent Variables

Instructor: G. William Schwert
585-275-2470
schwert@schwert.ssb.rochester.edu

Topics

- Linear probability model
- Logit
- Probit

Dummy Dependent Variables

- If the variable you are trying to explain/predict is qualitative or discrete (e.g., 0 or 1), a regression model can be used

  $Y = \beta_0 + \beta_1 X + \varepsilon$

  The regression model gives the conditional mean of Y given X

  $E(Y|X) = \beta_0 + \beta_1 X$

  which can often be interpreted as the probability that Y=1 given X

Problems with Dummy Dependent Variables

- Since Y is either 1 or 0, the residuals are either $(1 - \beta_0 - \beta_1 X)$ or $(- \beta_0 - \beta_1 X)$
- Since E(u) = 0,
  - the probability that the residual will equal $(1 - \beta_0 - \beta_1 X)$
    - has to be $(\beta_0 + \beta_1 X)$
  - and the probability that the residual will be $(- \beta_0 - \beta_1 X)$
    - has to be $(1 - \beta_0 - \beta_1 X)$
- Therefore, the variance of the errors has to be
  - $\text{Var}(u) = (\beta_0 + \beta_1 X)\ [1 - \beta_0 - \beta_1 X]^2 + (1 - \beta_0 - \beta_1 X)\ [- \beta_0 - \beta_1 X]^2$
    $= (\beta_0 + \beta_1 X)\ [1 - \beta_0 - \beta_1 X]$
    $= E(Y| X)\ [1 - E(Y| X)]$

Problems with Dummy Dependent Variables

- Thus, there are a few known problems with the linear probability model
  - Residuals are heteroskedastic (Var(u) = E(Y| X) [1 - E(Y| X)])
  - There is no restriction that the prediction from the regression is between 0 and 1
- Nevertheless, in most cases with reasonable sample sizes the linear probability model works pretty well
  - You probably want to use White standard errors, since you know the residuals will be heteroskedastic

Epidural Data (A425_epidural.wf1)

- Labor & Delivery Service at a hospital question whether the use of epidural anesthetic prolongs childbirth
- Collect data on:
  - LABTIME (time in labor in minutes)
  - EPIDURAL (=1 if used)
  - Baby's weight in pounds (BABYWGHT)
  - Mother's weight in pounds (MW)
  - Term of pregnancy in days (TERM)
  - Mother's age in years (AGE)

Example: Use of Epidurals

- It looks like doctors are less likely to use epidurals with older mothers, but more likely with heavier mother's weight and older babies (TERM)

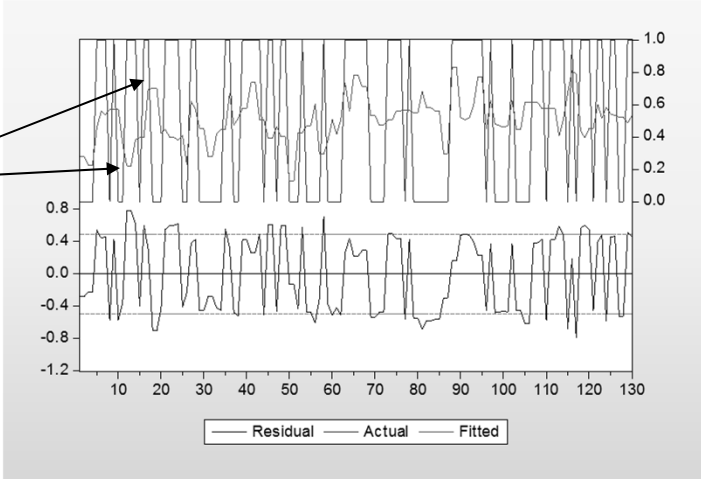
Dependent Variable: EPIDURAL
Method: Least Squares
Sample: 1 130
Included observations: 130
White heteroskedasticity-consistent standard errors & covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.264059	0.949846	-1.330804	0.1857
AGE	-0.020947	0.008397	-2.494505	0.0139
BABYWGHT	-0.010665	0.046025	-0.231722	0.8171
MW	0.004016	0.002574	1.560168	0.1212
TERM	0.044313	0.026689	1.660386	0.0993

R-squared	0.082606	Mean dependent var	0.507692
Adjusted R-squared	0.053249	S.D. dependent var	0.501875
S.E. of regression	0.488330	Akaike info criterion	1.442051
Sum squared resid	29.80825	Schwarz criterion	1.552341
Log likelihood	-88.73330	Hannan-Quinn criter.	1.486865
F-statistic	2.813883	Durbin-Watson stat	1.564774
Prob(F-statistic)	0.028153	Wald F-statistic	3.898354
Prob(Wald F-statistic)	0.005110		

Actual, Fitted, Residuals Plot

- Note that the predicted probabilities are between about 10% and 80%
- Actuals are either 1 or 0, so residuals look like a muted version of the actuals



Ignoring Heteroskedasticity

- Even using OLS standard errors would not give t-stats that are much different
- =>OLS and the linear probability model might work OK in this case

Dependent Variable: EPIDURAL
Method: Least Squares
Sample: 1 130
Included observations: 130

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.264059	1.033113	-1.223543	0.2234
AGE	-0.020947	0.009528	-2.198440	0.0298
BABYWGHT	-0.010665	0.045268	-0.235598	0.8141
MW	0.004016	0.002497	1.608504	0.1102
TERM	0.044313	0.028218	1.570397	0.1189

R-squared	0.082606	Mean dependent var	0.507692
Adjusted R-squared	0.053249	S.D. dependent var	0.501875
S.E. of regression	0.488330	Akaike info criterion	1.442051
Sum squared resid	29.80825	Schwarz criterion	1.552341
Log likelihood	-88.73330	Hannan-Quinn criter.	1.486865
F-statistic	2.813883	Durbin-Watson stat	1.564774
Prob(F-statistic)	0.028153		

Logit Model

- Suppose that you thought about a regression model for a "latent" variable Y^*

$$Y^* = \beta_0 + \beta_1 X + \varepsilon$$

Where Y* is not observed, but we can observe a dummy variable Y defined by

$Y = 1$ if $Y^* > 0$
$Y = 0$ if $Y^* < 0$

Logit Model

In the case of epidurals, Y^* might be thought of as the amount of pain and suffering that would be expected for the mother without anesthesia relative to a threshold that doctors would use to decide to administer an epidural

Logit Model

- $\text{Prob}(Y = 1) = \text{Prob}[\varepsilon > -\beta_0 - \beta_1 X]$
  $= 1 - F[-\beta_0 - \beta_1 X]$

where $F[\,.\,]$ is the cumulative distribution function for ε

- we usually assume that $\text{Var}(\varepsilon) = 1$ since this is arbitrary
- If the distribution of ε is symmetric, then $1 - F(-Z) = F(Z)$, so

$\text{Prob}(Y = 1) = F[\beta_0 + \beta_1 X]$

Logit Model

- The observed data *Y* are just realizations of a binomial process with probabilities given by the probability model
- So the "likelihood function" is

$$\prod_{y=1} P_i \prod_{y=0} (1 - P_i)$$

where $P_i = \text{Prob}(Y = 1) = F[\beta_0 + \beta_1 X]$

Logit Model

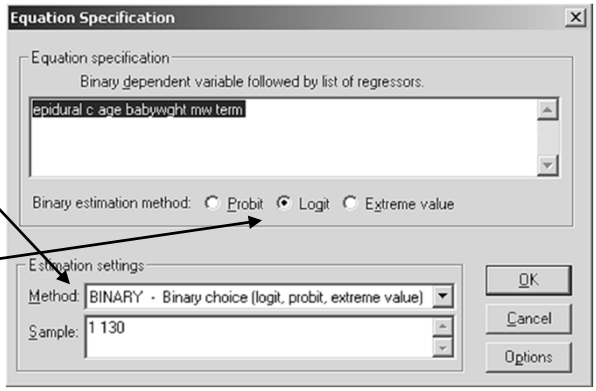
- If the distribution of the error term ε is logistic,
  - $F(Z) = \exp(Z)/ [1 – \exp(Z)]$
  - So, $\log \{F(Z) / [1 – F(Z)]\} = Z$

We call this the "logit" model, and

$$\text{Log} [P_i / (1-P_i)] = \beta_0 + \beta_1 X$$

Logit Model for Epidurals in Eviews

- In Eviews, instead of least squares in the "method" box, choose BINARY
- then highlight the "Logit" radio button



Logit Model for Epidurals in Eviews

- Note that the t-stats are very similar to what we saw with OLS (linear probability model)
- To compare the regression coefficients, you need to multiply the logit coefficients by .25
- Except the constant, where the adjustment is .25 β_L + .5
  - These are approximate adjustments

Dependent Variable: EPIDURAL
Method: ML - Binary Logit (Quadratic hill climbing)
Sample: 1 130
Included observations: 130
Convergence achieved after 5 iterations
Covariance matrix computed using second derivatives

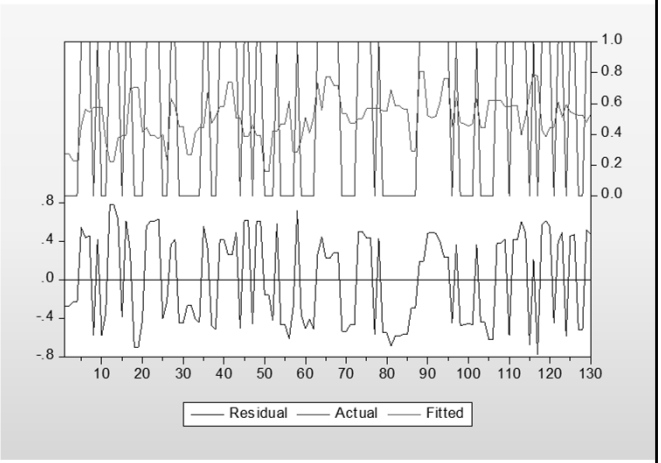
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-7.832855	4.655343	-1.682552	0.0925
AGE	-0.092133	0.042852	-2.150040	0.0316
BABYWGHT	-0.049414	0.190905	-0.258840	0.7958
MW	0.017350	0.010915	1.589540	0.1119
TERM	0.198494	0.126788	1.565557	0.1175

McFadden R-squared	0.062391	Mean dependent var	0.507692
S.D. dependent var	0.501875	S.E. of regression	0.488927
Akaike info criterion	1.376504	Sum squared resid	29.88115
Schwarz criterion	1.486793	Log likelihood	-84.47274
Hannan-Quinn criter.	1.421318	Deviance	168.9455
Restr. deviance	180.1875	Restr. log likelihood	-90.09375
LR statistic	11.24202	Avg. log likelihood	-0.649790
Prob(LR statistic)	0.023975		

Obs with Dep=0	64	Total obs	130
Obs with Dep=1	66		

Logit Model for Epidurals in Eviews

- Predictions from logit look very similar to the linear probability model

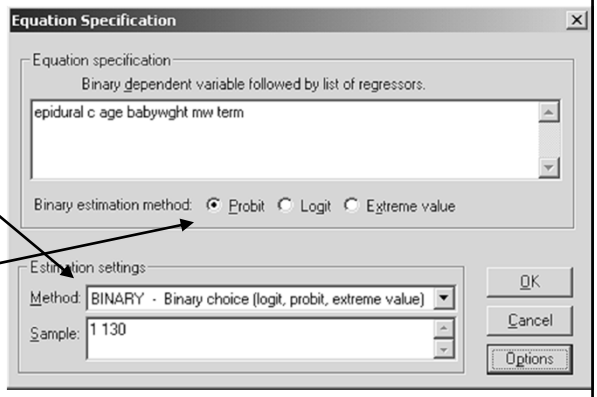


Probit Model

- If the distribution of the error term ε is normal,
  - F(Z) = cumulative normal
- It turns out that the logistic and normal distributions are close to each other, so you are unlikely to get much difference between logit and probit models
  - Differences are in the extreme tails of the distribution

Probit Model for Epidurals in Eviews

- In Eviews, instead of least squares in the "method" box, choose BINARY
- then highlight the "Probit" radio button



Probit Model for Epidurals in Eviews

- Again, the t-stats are very similar to what we saw with OLS (linear probability model) and to Logit
- To compare the regression coefficients, you need to multiply the probit coefficients by .4
- Except the constant, where the adjustment is .4 β_P + .5
  - These are approximate adjustments

Dependent Variable: EPIDURAL
Method: ML - Binary Probit (Quadratic hill climbing)
Sample: 1 130
Included observations: 130
Convergence achieved after 5 iterations
Covariance matrix computed using second derivatives

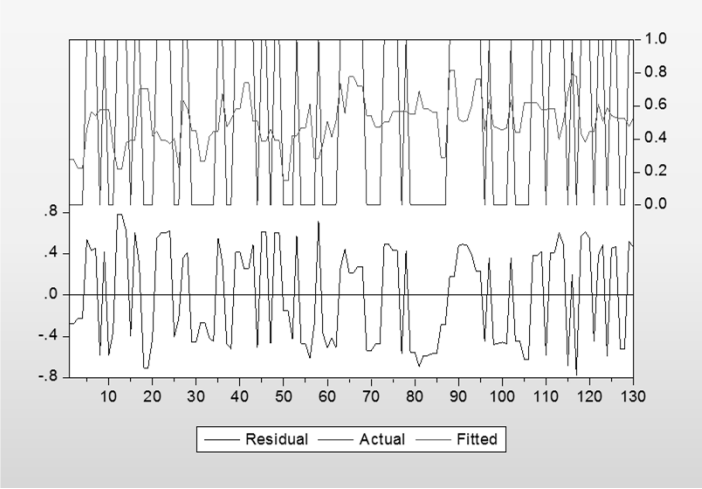
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-4.882704	2.829543	-1.725616	0.0844
AGE	-0.057577	0.026366	-2.183741	0.0290
BABYWGHT	-0.032537	0.118232	-0.275199	0.7832
MW	0.010644	0.006588	1.615683	0.1062
TERM	0.124892	0.076874	1.624632	0.1042

McFadden R-squared	0.062851	Mean dependent var	0.507692
S.D. dependent var	0.501875	S.E. of regression	0.488847
Akaike info criterion	1.375865	Sum squared resid	29.87148
Schwarz criterion	1.486155	Log likelihood	-84.43123
Hannan-Quinn criter.	1.420680	Deviance	168.8625
Restr. deviance	180.1875	Restr. log likelihood	-90.09375
LR statistic	11.32503	Avg. log likelihood	-0.649471
Prob(LR statistic)	0.023144		

Obs with Dep=0	64	Total obs	130
Obs with Dep=1	66		

Probit Model for Epidurals in Eviews

- Predictions from probit look very similar to the linear probability and logit models



Comparison of Linear Probability, Logit, and Probit Coefficients

	Linear	logit	logit-adjusted	probit	probit-adjusted
C	-1.264	-7.833	-1.458	-4.883	-1.453
AGE	-0.021	-0.092	-0.023	-0.058	-0.023
BABYWGHT	-0.011	-0.049	-0.012	-0.033	-0.013
MW	0.004	0.017	0.004	0.011	0.004
TERM	0.044	0.198	0.050	0.125	0.050

Thus, the three models give almost identical results
(once the coefficients are adjusted to be comparable)

Conclusions

- For most purposes, you are probably well advised to start with the linear probability model
  - It is easy to interpret the coefficients, R^2, etc., just as you would in a standard regression model
  - As long as the sample size is reasonable, the results are likely to be similar to the more complicated logit and probit models (in my experience)

Links

Epidural Data

http://schwert.ssb.rochester.edu/a425/a425_epidural.wf1

Return to APS 425 Home Page

http://schwert.ssb.rochester.edu/a425/a425main.htm