

Dummy Dependent Variables

Recap and Today

- Last time
 - Regression Assumptions
 - Problems with our models
 - Regression Users' Guide
- Today
 - Dummy Dependent Variables
 - Work through an example Quiz
 - Policy Report Workshop

Example

Habitation Sites & Culturally Modified Trees

- Outcome
 - Whether CMTs are present
 - 1=Yes CMTs are present, 0=CMTs not present
- Predictors
 - Elevation
 - Slope
 - Beach type
 - distance to water
 - distance to salmon

Linear Probability Model

OLS with dummy dependent variable

$$D_i = \beta_0 + \beta_1 X_{1i} + \varepsilon$$

Where D_i is a dummy outcome Variable.

$$E(D_i) = P_i$$

BUT:

Errors are heteroskedastic (violates #5)

Errors are non-normal (violates #7)

$\overline{R^2}$ not accurate

Predicts probabilities above 1 and below 0

Logit Model

$$\ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 X + \varepsilon$$

$$P_i = \frac{1}{1 + e^{-[\beta_0 + \beta_1 X + \varepsilon]}}$$

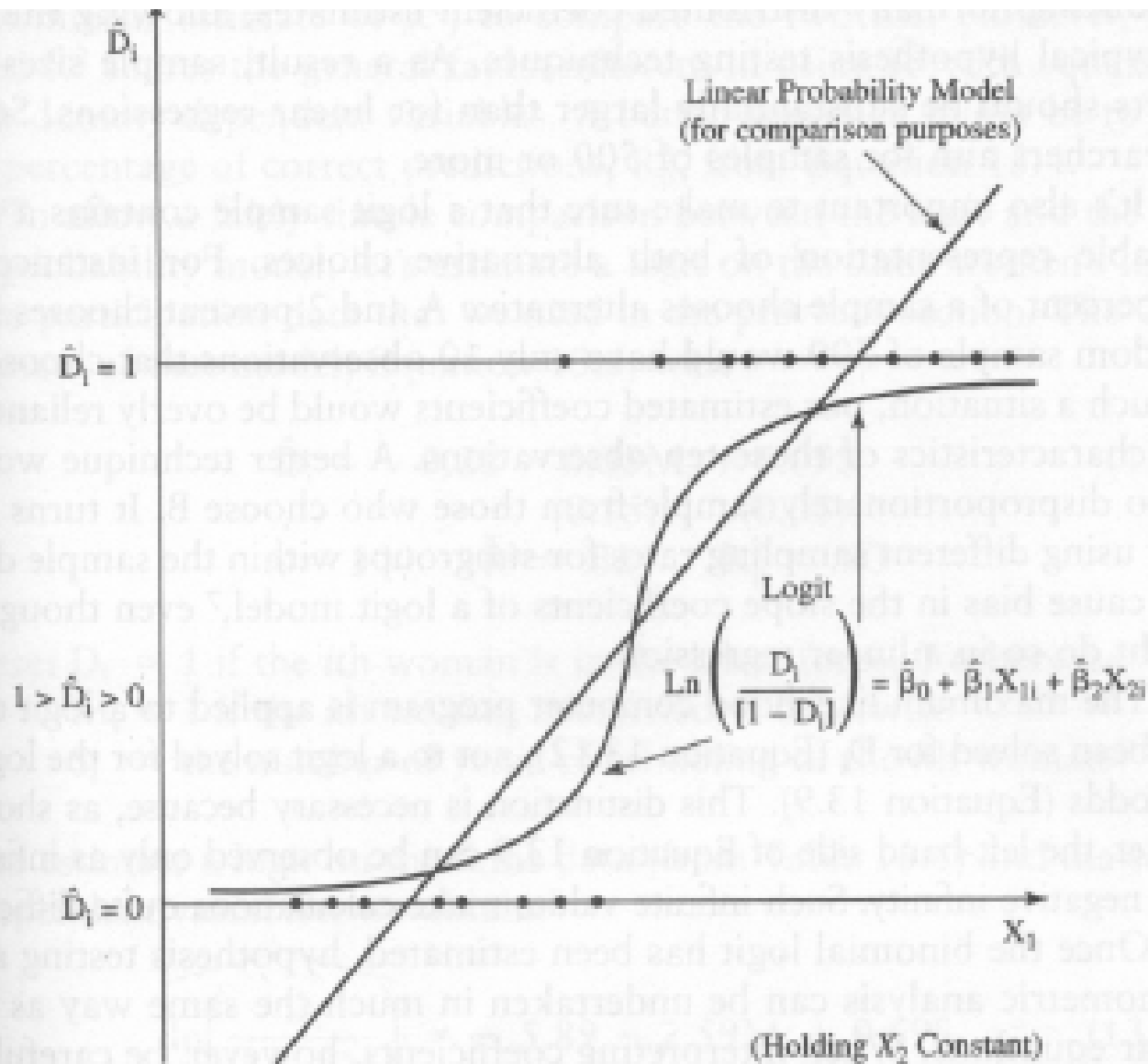


Figure 13.2 \hat{D}_i IS BOUNDED BY ZERO AND ONE IN A BINOMIAL LOGIT MODEL
 In a binomial logit model, \hat{D}_i is nonlinearly related to X_{1i} , so even exceptionally large or small values of X_{1i} , holding X_{2i} constant, will not produce values of \hat{D}_i outside the meaningful range of zero to one.

Predict the Probability

- Site 25
 - Elevation above sea level=1
 - Slope of land=20 degrees
 - Beach=0 (0=rocky, 1=sandy)
 - Distance to water=46 feet
 - Distance to salmon=450 feet

$$z = \beta_0 + \hat{\beta}_{elev}(elevation) + \hat{\beta}_{slope}(slope) + \hat{\beta}_{beach}(beach) + \hat{\beta}_{water}(water) + \hat{\beta}_{salmon}(salmon)$$
$$z = 3.2587 + (-0.3095)(1) + (0.0801)(20) + (-2.1536)(0) + (-0.0112)(46) + (0.0001)(450)$$
$$z = 4.081$$

$$P_i = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-4.081}} = 0.983$$

So, a site like site 25 has a 98.3% chance of having culturally modified trees.

20 feet closer to water?

$$Z=4.081+(-0.0112)(-20)=4.081+0.224=4.305$$

$$P_i = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-4.305}} = 0.987$$

If the site were 20 feet closer, it would be slightly more likely to have culturally modified trees.

Interpreting Logit Coefficients

- Not directly interpretable, so transform them.
 - e^β (SPSS does this for you)
- A one unit change in X is associated with being $e^\beta(100\%)$ as likely to be $D_i=1$.

OR

If positive: $e^\beta(100\%)-(100\%)$ more likely to be $D_i=1$.

If negative: $100\%-e^\beta(100\%)$ less likely to be $D_i=1$.

Coefficient Interpretation Example 1 (continuous)

- What is the impact of elevation on the probability of having a CMT or not?

$$\hat{\beta} = -0.3095 \quad e^{\hat{\beta}} = e^{-0.3095} = 0.734$$

- So, the higher the elevation, the less likely a CMT.
- An additional foot of elevation is associated with being 73% as likely (or 27% less likely) to have CMTs.

Coefficient Interpretation Example 2 (dummy)

- What's the impact of a sandy beach on the likelihood of a CMT?

$$\hat{\beta} = -2.1536 \quad e^{\hat{\beta}} = e^{-2.1536} = 0.116$$

- So, a sandy beach is less likely to be the location of CMTs.
- It is 11.6% as likely, or 88% less likely.

Some Notes

- Model assumptions for OLS not a problem
- Concerned about influential cases and outliers. Plot residuals to identify (more than 3 SD away may be influential outlier).
- If you find outliers, rerun model without them to see how they influence the results.

Likelihood Ratio (LR) Tests

- Use instead of F-tests--Same hypotheses
- For Model, -2 Log Likelihood on output
- To Compare Models
 - Create a restricted and an unrestricted model
 - $LR = -2\log L_R - (-2\log L_U) = 2\log L_U - 2\log L_R$
 - Compare LR to critical χ^2 where degrees of freedom = number of regressors. If calculated LR is greater, then reject null--model has explanatory power

SPSS

ANALYZE>REGRESSION>BINARY LOGISTIC

Sample SPSS Output

Model: $\text{working} = \beta_0 + \beta_1(\text{Female}) + \beta_2(\text{Number of Children 17 or less}) + \beta_3(\text{Age}) + \beta_4(\text{Married}) + \beta_5(\text{More than a High School Education}) + \beta_6(\text{Own home}) + \beta_7(\text{King County Resident}) + \varepsilon$

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step a 1	female	-.652	.066	98.096	1	.000	.521
	chldm17	.223	.030	56.173	1	.000	1.249
	married	.173	.068	6.484	1	.011	1.189
	Highschoolormore	.827	.083	98.220	1	.000	2.287
	Own	-.013	.172	.006	1	.938	.987
	king	.004	.085	.002	1	.965	1.004
	Constant	-.290	.088	10.957	1	.001	.748

a. Variable(s) entered on step 1: Own, king.

SPSS enters the mean in block 0, then all your variables in block 1

Block 1: Method = Enter

Block 2: Method = Enter

$$LR = -2 \log LR - (-2 \log LU) \\ = 2 \log LU - 2 \log LR \\ = 259.627 - 153.639 \\ = 105.988$$

Compare to chi-square at 2 df

Omnibus Tests of Model Coefficient

		Chi-square	df	Sig.
Step 1	Step	153.639	4	.000
	Block	153.639	4	.000
	Model	153.639	4	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	5409.183 ^a	.037	.050

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

Classification Table^a

Observed	LAST WEEK HAVE FT/PT JOB (FOR PAY)	Predicted		
		LAST WEEK HAVE FT/PT JOB (FOR PAY)		Percentage Correct
		0.NO	1.YES	
Step 1	LAST WEEK HAVE FT/PT JOB (FOR PAY)	0.NO	1.YES	27.7
		388	1906	83.1
	Overall Percentage			59.0

a. The cut value is .500

Variables in the Equation

Step		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	chldrn17	.218	.029	55.386	1	.000	1.243
	Highschoolormore	.835	.080	109.587	1	.000	2.304
	Own	.009	.169	.003	1	.958	1.009
	king	-.014	.084	.027	1	.870	.986
	Constant	-.535	.079	45.813	1	.000	.585

a. Variable(s) entered on step 1: chldrn17, Highschoolormore, Own, king.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	105.987	2	.000
	Block	105.987	2	.000
	Model	259.627	6	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	5303.196 ^a	.062	.083

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

Classification Table^a

Observed	LAST WEEK HAVE FT/PT JOB (FOR PAY)	Predicted		
		LAST WEEK HAVE FT/PT JOB (FOR PAY)		Percentage Correct
		0.NO	1.YES	
Step 1	LAST WEEK HAVE FT/PT JOB (FOR PAY)	0.NO	1.YES	38.6
		683	1085	78.1
	Overall Percentage	502	1792	60.9

a. The cut value is .500

Variables in the Equation

Step		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	chldrn17	.223	.030	56.173	1	.000	1.249
	Highschoolormore	.827	.083	98.220	1	.000	2.287
	Own	-.013	.172	.006	1	.938	.987
	king	.004	.085	.002	1	.965	1.004
	female	-.652	.066	98.096	1	.000	.521
	married	.173	.068	6.484	1	.011	1.189
	Constant	-.290	.088	10.957	1	.001	.748

a. Variable(s) entered on step 1: female, married.

TABLE B-8 The Chi-Square Distribution

Degrees of Freedom	Level of Significance (Probability of a Value of at Least as Large as the Table Entry)			
	10%	5%	2.5%	1%
1	2.71	3.84	5.02	6.63
2	4.61	5.99	7.38	9.21
3	6.25	7.81	9.35	11.34
4	7.78	9.49	11.14	13.28
5	9.24	11.07	12.83	15.09
6	10.64	12.59	14.45	16.81
7	12.02	14.07	16.01	18.48
8	13.36	15.51	17.53	20.1
9	14.68	16.92	19.02	21.7
10	15.99	18.31	20.5	23.2
11	17.28	19.68	21.9	24.7
12	18.55	21.0	23.3	26.2
13	19.81	22.4	24.7	27.7
14	21.1	23.7	26.1	29.1
15	22.3	25.0	27.5	30.6
16	23.5	26.3	28.8	32.0
17	24.8	27.6	30.2	33.4
18	26.0	28.9	31.5	34.8
19	27.2	30.1	32.9	36.2
20	28.4	31.4	34.2	37.6

Source: See Table B-7.