LOGISTIC REGRESSION

Chapter 6

Nominal Dependent Variables

There are three main approaches to handling nominal dependent variables:

1. Discriminant analysis (Chapter 5).
   - Dependent variable has two or more categories

2. Logistic regression (Chapter 6).
   - Dependent variable has two categories.

3. Multinomial logistic regression.
   - Dependent variable has more than two categories.

Logistic Regression May Be Preferred . . .

Discriminant analysis:

- Only continuous independent variables.
- Finicky
- Needs equal variance–covariance matrices (Box's M test) across groups, and this assumptions is not met in many situations.
- Y can be nominal or ordinal.

Logistic regression:

- Continuous, ordinal, and nominal variables.
- Robust against assumption violations.
- Easier because it is similar to multiple regression, thus it is intuitively appealing.

Example

We have a research question that requires a nominal dependent variable.

E.g. Foreign market entry mode of service firms can be divided between high control (subsidiary) and low control (agents etc.).

Converting Metric Variables to Non-Metric

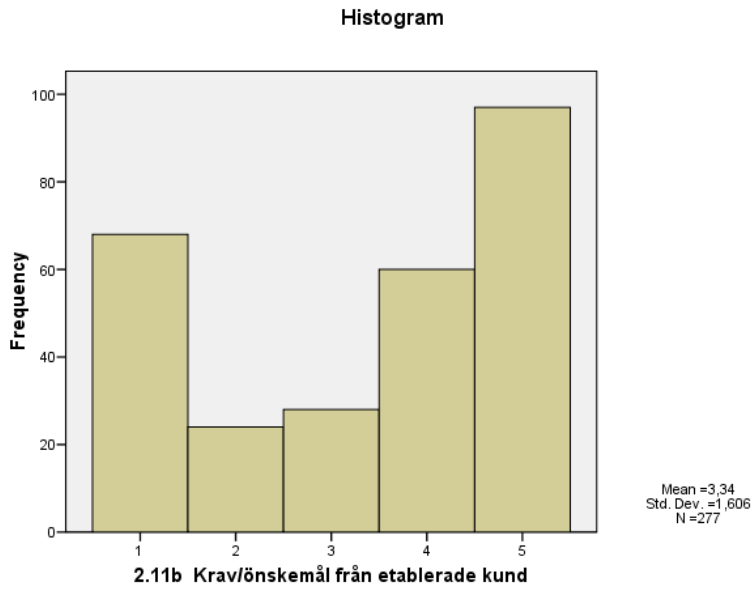
Did you start foreign operations because of demands from customers?

No 1 2 3 4 5 Yes

Polar extremes approach = compares only the extreme two groups and excludes the middle group.

Median split = dividing the data at the median.

Dichotomize a Variable



Reviewer: Why take a higher level variable and reduce it to a lower scale?

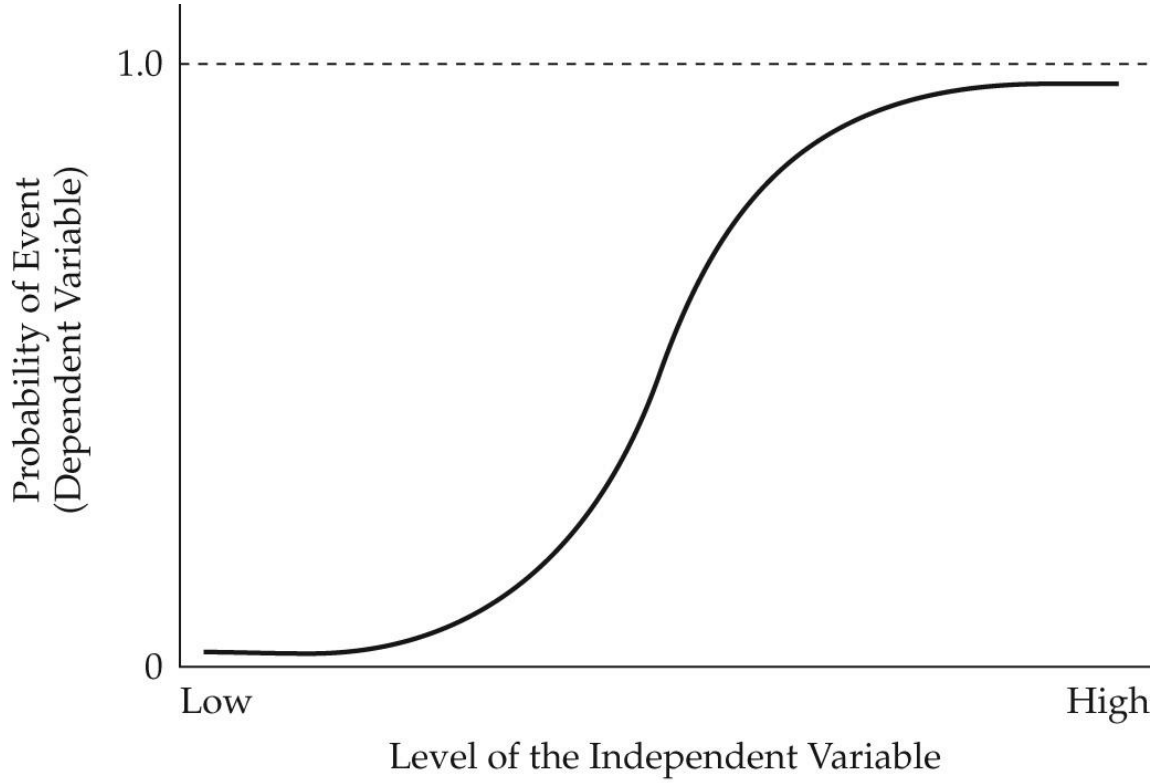
Answer: The variable was natural for dichotomization (I used polar extremes).

What Logistic Regression Does

- Using maximum likelihood estimation, logistic regression predicts the probability of an event occurring, in this case foreign market entry mode.

- Contrary to “typical” multiple regression, it does not use ordinary least squares estimation.

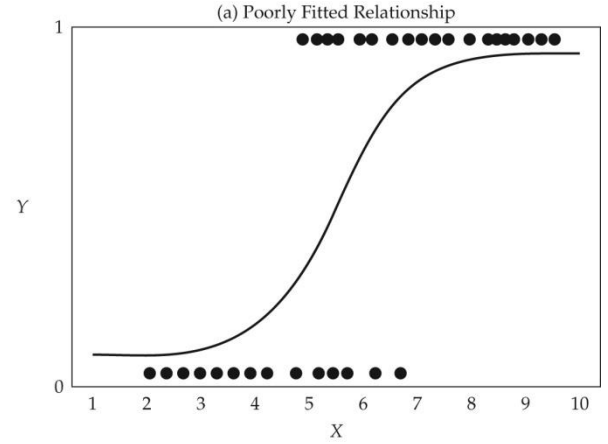
Logistic Curve



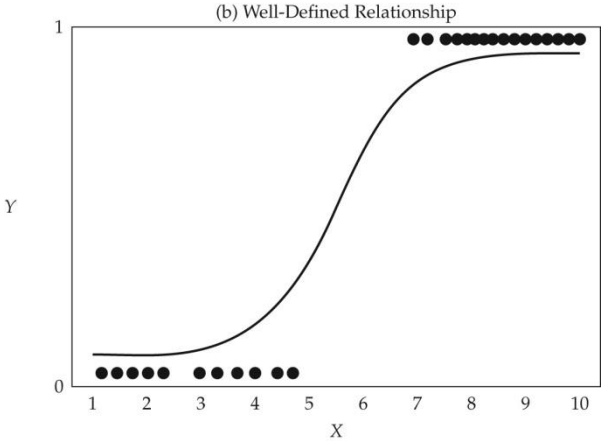
Predicts the probability of an event for levels of the independent variable(s).

Logistic Curve

Poor prediction



Good prediction



Sample Size

- Maximum Likelihood Estimation requires larger samples than OLS.
- Hosmer Lemeshow recommend at least 400 plus hold-out sample.
- 10 observations per estimated parameter.
- Non-metric nominal or ordinal variables create more groups.

Example

"Choice of Foreign Market Entry Mode in Service Firms“

Anders Blomstermo
D. Deo Sharma
James Sallis

(2006), *International Marketing Review*, Vol. 23 No. 2, pp. 211-229.

The Variables

Entry Mode: A dichotomous decision between a low control entry mode (coded 0), and a high control entry mode (coded 1).

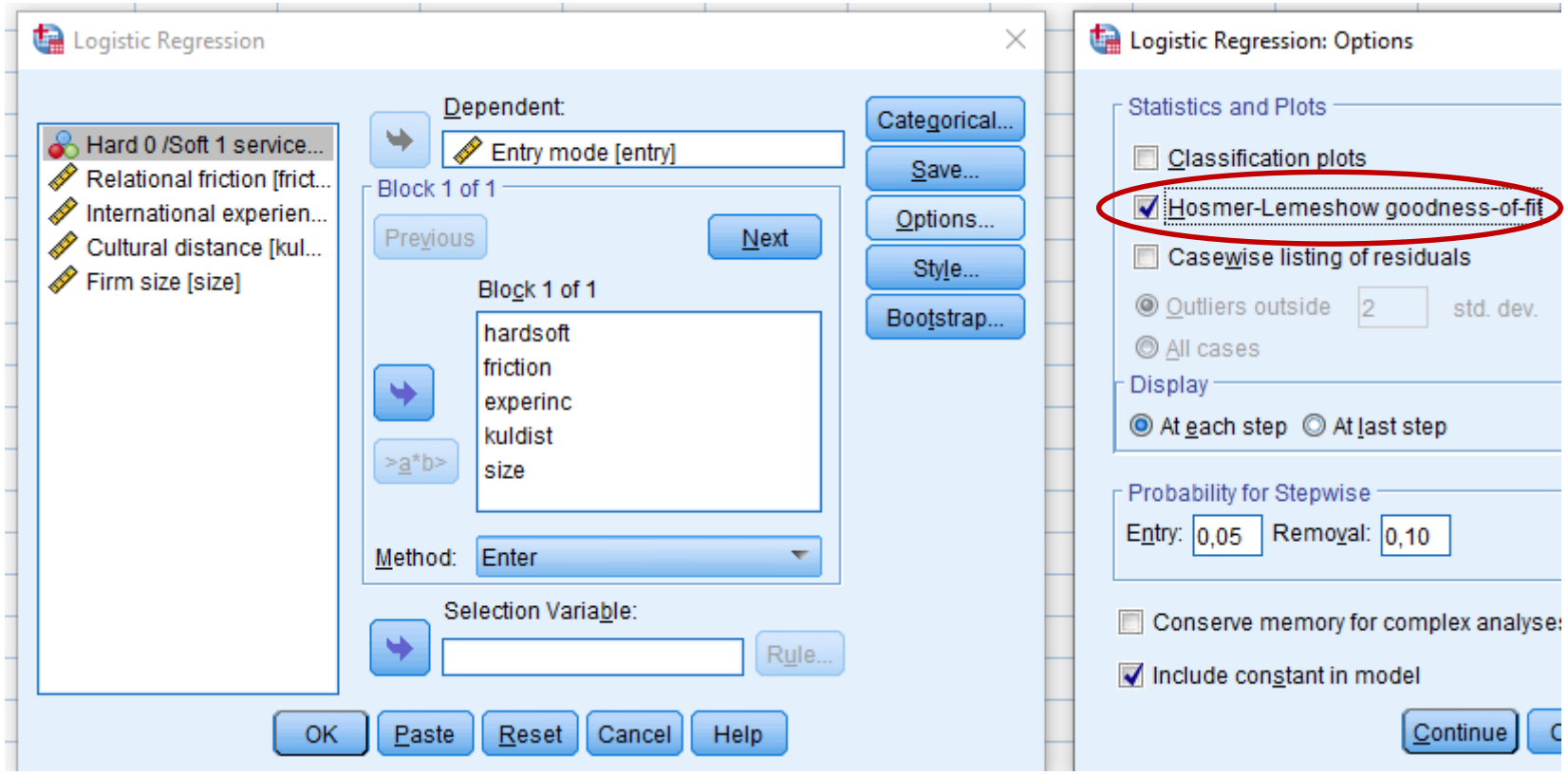
Hard Soft: A dichotomous variable coded 0 for hard services and 1 for soft services.

Relational friction: It is easier and cheaper work in high control relationships (partner versus subsidiary).

Experience: A firm's previous experience in international markets.

Cultural distance: Hofstede (80) Kogut & Singh (88).

Firm size: Number of employees.


Logistic Regression
Hard 0 /Soft 1 service...
Relational friction [frict...
International experien...
Cultural distance [kul...
Firm size [size]
Dependent:
Entry mode [entry]
Block 1 of 1
Previous
Next
Block 1 of 1
hardsoft
friction
experinc
kuldist
size
>a*b>
Method:
Enter
Selection Variable:
Rule...
Categorical...
Save...
Options...
Style...
Bootstrap...
OK
Paste
Reset
Cancel
Help
Logistic Regression: Options
Statistics and Plots
Classification plots
Hosmer-Lemeshow goodness-of-fit
Casewise listing of residuals
Outliers outside
2
std. dev.
All cases
Display
At each step
At last step
Probability for Stepwise
Entry:
0,05
Removal:
0,10
Conserve memory for complex analyses
Include constant in model
Continue

Interpretation

Do model calculations based on variables in analysis.

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	65	41,4
	Missing Cases	92	58,6
	Total	157	100,0
Unselected Cases		0	,0
Total		157	100,0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
Low control	0
High control	1

Dependent coding.

Beginning Block

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted: Entry mode: Low control	Predicted: Entry mode: High control	Percentage Correct
Step 0	Entry mode	Low control	39	0	100,0
		High control	26	0	,0
	Overall Percentage				60,0

n = 65

Low control group size 39

65

60%

High control group size 26

65

40%

Model Fit

Good model fit is indicated by a significant model chi-square statistic and an insignificant Hosmer-Lemeshow chi-square statistic.

This is comparable to the F-statistic in OLS regression

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	22,206	5	,000
	Block	22,206	5	,000
	Model	22,206	5	,000

Good!

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	11,800	7	,107

Good!

Comparing Models

Refer to the -2 Log likelihood. The lower the number the better the fit.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	65,286[a]	,289	,391

Quasi R^2 should not be thought of in the same way as OLS R^2. They are not the amount of explained variance. The only way that I would use them is to compare models.

Hit Rate (classification accuracy)

Profile misclassified observations to find a variable that may be able to discriminate.

Classification Table[a]

Observed			Predicted: Entry mode: Low control	Predicted: Entry mode: High control	Percentage Correct
Step 1	Entry mode	Low control	34	5	87,2
		High control	8	18	69,2
	Overall Percentage				80,0

Classification Accuracy

The classification accuracy is measured by the hit-rate:

The naive model (maximum chance) is simply the accuracy if all observations were placed into the largest group. In the current example, the percentage correct would need to exceed 60%. And, the rule of thumb is that it should exceed by a margin of 25%.

60% x 1.25 = 75%

Since 87.2% > 75%, the naïve (maximum chance) model exceeds prediction by a good margin.

Classification Table[a]

Observed			Predicted: Entry mode: Low control	Predicted: Entry mode: High control	Percentage Correct
Step 1	Entry mode	Low control	34	5	87,2
		High control	8	18	69,2
	Overall Percentage				80,0

60% Low control

40% High control

Classification Accuracy

Another assessment of classification accuracy is the random model (proportional chance). It estimates how well the model classifies observations to the small group. It is calculated as:

$\alpha^2 + (1-\alpha)^2$, where α is the proportion in the small group.

In the current example, the small group proportion is 40%

$.4^2 + (1 - .4)^2 = .52$. And, the rule of thumb is that it should exceed by a margin of 25%.

52% x 1.25 = 65%

Since 69.2 > 65, the random (proportional chance) model exceeds prediction by a good margin.

Hypothesis Tests

The Wald statistic indicates the significance of each estimated coefficient, providing tests for individual hypotheses. See the Sig. (p-value)

- In ”typical” multiple regression this is done with the t-tests.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Hard 0 /Soft 1 services	1,725	,699	6,098	1	,014	5,612
	Relational friction	-,532	,340	2,440	1	,118	,588
	International experience	-,455	,235	3,746	1	,053	,634
	Cultural distance	,638	,246	6,739	1	,009	1,893
	Firm size	,000	,000	,521	1	,471	1,000
	Constant	,907	1,654	,300	1	,584	2,476

a. Variable(s) entered on step 1: Hard 0 /Soft 1 services, Relational friction, International experience, Cultural distance, Firm size.

Coefficients

In logistic regression, the B coefficients measure the change in the odds of group membership on the dependent variable. They are difficult to directly interpret, so instead we look at the exponentiated logistic coefficient.

When the B is negative, the Exp(B) will be less than one. That means it reduces the odds of belonging to the dependent variable category coded 1.

When B is positive, the Exp(B) will be greater than one. That means it increases the odds of belonging to the dependent variable category coded 1.

Magnitude of the Relationship . . .

Metric variables, a one-unit change in the independent variable:

Exp(B) = 1, means no change in the odds.

Exp(B) = .5, means a 50% decrease in the odds of category 1.

Exp(B) = 2, means a 100% increase in the odds of category 1.

Dummy variables have the same interpretation as metric variables, but there is only the possibility of 0 or 1, not a one-unit change.

Interpretation

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Hard 0 /Soft 1 services	1,725	,699	6,098	1	,014	5,612
	Relational friction	-,532	,340	2,440	1	,118	,588
	International experience	-,455	,235	3,746	1	,053	,634
	Cultural distance	,638	,246	6,739	1	,009	1,893
	Firm size	,000	,000	,521	1	,471	1,000
	Constant	,907	1,654	,300	1	,584	2,476

a. Variable(s) entered on step 1: Hard 0 /Soft 1 services, Relational friction, International experience, Cultural distance, Firm size.

Soft service firms (coded 1) are 5.612 times more likely to choose a high control entry mode (coded 1) than hard service firms.

A one-unit increase in cultural distance will increase the odds of high control entry by 89.3%.

A two-unit increase in cultural distance will increase the odds of high control entry by 178.6%.