


Logistic Regression

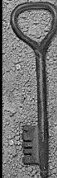
GOG 502/PLN 504 Youqin Huang 1



Binary Outcomes

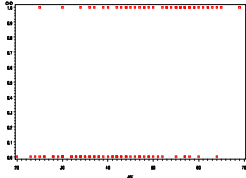
- ◆ Often we are interested in predicting whether or not some event will occur:
 - Will a family purchase a new car this year?
 - Will a customer default on a loan?
 - Will a patient survive a certain disease?
 - Will a household own a home?
 - Will a person move (migrate) or not?
- ◆ Outcomes (dependent variables) are binary and can be coded 0 (failure) and 1 (success).

GOG 502/PLN 504 Youqin Huang 2




Violation of OLS Assumptions

- ◆ In the case of binary dependent variables, most assumptions in linear regression are violated:
 - Normality: there are *only two possible* values, thus, the errors and Y_i cannot be normally distributed.
 - Linearity
 - Homoskedasticity



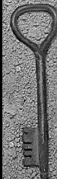
GOG 502/PLN 504 Youqin Huang 3



Violation of OLS Assumptions

- ◆ In the case of binary dependent variables, most assumptions in linear regression are violated
- ◆ The dependent variable is restricted to the range of (0, 1), while in OLS regression there is no bound and the DV ranges from $-\infty$ to ∞ .
- ◆ Solution: Logistic regression
 - does not make any assumption of normality, linearity, and homogeneity of variance for the independent variables.
 - The logistic transformation will expand the range from (0, 1) to $(-\infty, \infty)$.

GOG 502/PLN 504 Youqin Huang 4




Logistic Regression

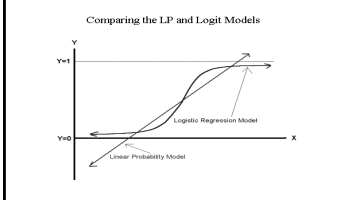
- ◆ Dependent variable is a binary variable ($Y_i = 0, 1$)
- ◆ Let π denote the proportion of success
- ◆ $P(Y_i = 1) = \pi_i$, $P(Y_i = 0) = 1 - \pi_i$
- ◆ We could estimate a linear probability model:

$$\pi = a + bX$$
 - Will have poor results because of severe violation of assumptions (e.g linearity, range restriction, constant standard deviation) and the limited range (0, 1)

GOG 502/PLN 504 Youqin Huang 5



Linear vs. Logistic Regression



- ◆ So, we need to transform logistic relationship into a linear relationship in order to apply linear regression – logistic transformation or logit

GOG 502/PLN 504 Youqin Huang 6

Jargon: Odds

- ◆ Probability of an event: π_i
- ◆ Odds: the ratio between probability of that event occurs and the probability of that event does not occur :

$$\frac{\pi_i}{1-\pi_i}$$
- ◆ E.g. $\pi=0.75$, the odds= $0.75/0.25=3$, meaning a success (e.g. move) is three times more likely as a failure (e.g. stay)
- ◆ Properties of odds:
 - Always positive (because $\pi_i < 1$)
 - Lower bound of zero, no upper bound, $(0, \infty)$

GOG 502/PLN 504 Youqin Huang 7

Jargon: The Logit

$$\text{Logit} = \text{Log}_e \left(\frac{\pi_i}{1-\pi_i} \right) = \ln \left(\frac{\pi_i}{1-\pi_i} \right)$$

- ◆ Odds ranges $(0, \infty)$; remove this lower bound by taking the natural logarithm of the odds.
- ◆ The natural logarithm of the odds is called the **logit**:
 - Its possible values are all real numbers: $-\infty$ to ∞ .
 - As π increases from 0 to 1, the odds increase from 0 to ∞ , and the logit increases from $-\infty$ to ∞ .

GOG 502/PLN 504 Youqin Huang 8

The Logistic Model

- ◆ Instead of modeling $\pi = a + bX$
- ◆ We will model $\text{Ln} \left(\frac{\pi}{1-\pi} \right) = a + bX$
- ◆ This model can be expanded to multiple Xs
- ◆ Like the linear model, b refers to whether the curve increases ($b > 0$) or decreases ($b < 0$) as X increases
- ◆ X has a **linear effect** on the **logit** (one unit increase in X, the logit increases b units), not the probability (π), nor the odds ($\pi/(1-\pi)$).

GOG 502/PLN 504 Youqin Huang 9

Logistic Regression

Figure 15.1 Linear and Logistic Regression Models for a (0, 1) Response; The Mean of Y Is the Probability π

- Logistic (1): $b > 0$; Logistic (2): $b < 0$
- $|b|$ determines the steepness of the curve

GOG 502/PLN 504 Youqin Huang 10

Logistic Regression: estimation

- ◆ Logistic regression does not try to minimize the sum of squares, but rather uses Maximum Likelihood Estimation (MLE).
 - Find a and b that make it the most likely that the observed pattern of events in the sample would have occurred: Maximizes the likelihood (L)
 - MLE is an iterative algorithm.
 - Starts with an initial arbitrary "guesstimate" of what the logit coefficients should be.
 - After this initial equation is estimated, the residuals are tested and a re-estimate is made with an improved function.
 - The process is repeated until convergence is reached (that is, until L does not change significantly).

GOG 502/PLN 504 Youqin Huang 11

Effect of X on the Odds

$$\text{Log}_e \left(\frac{\pi}{1-\pi} \right) = a + bX$$

- ◆ Take antilog: $\frac{\pi}{1-\pi} = e^{a+bX} = e^a (e^b)^X$
- ◆ Odds Ratio: $\frac{e^a (e^b)^{(X+1)}}{e^a (e^b)^X} = \frac{e^a (e^b)^X e^b}{e^a (e^b)^X} = e^b$
- ◆ X has a multiplicative effect of e^b on the **odds**
- ◆ One unit increase in x, the odds increase by a factor of e^b

GOG 502/PLN 504 Youqin Huang 12

Effect of X on the Odds

$$\text{Log}_e\left(\frac{\pi}{1-\pi}\right) = a + bX$$

$$\frac{e^a (e^b)^{(X+1)}}{e^a (e^b)^X} = \frac{e^a (e^b)^X e^b}{e^a (e^b)^X} = e^b$$

- E.g. homeownership and income (in \$1000): $b=0.05$, $e^{0.05}=1.05$; one unit (\$1000) increase in income, the odds of owning multiply by 1.05, or 5% more likely to own.

GOG 502/PLN 504 Youqin Huang 13

Effect of X on Probability (π)

$$\text{Log}_e\left(\frac{\pi}{1-\pi}\right) = a + bX$$

- Take antilog: $\frac{\pi}{1-\pi} = e^{a+bX}$
- Do some algebra:

$$\pi = e^{a+bX} (1-\pi) = e^{a+bX} - \pi e^{a+bX}$$

$$\pi = \frac{e^{a+bX}}{1 + e^{a+bX}} = \frac{e^a (e^b)^X}{1 + e^a (e^b)^X}$$

GOG 502/PLN 504 Youqin Huang 14

Logistic Regression: Hypothesis Test

$$\text{Ln}\left(\frac{\pi}{1-\pi}\right) = a + bX$$

- Global test: test for the overall model
 - Null hypothesis: all $\beta=0$; Alternative: at least one $\beta \neq 0$.
 - 1) Wald chi-square test
 - Similar to the t-test in OLS regression
 - 2) Likelihood-ratio test: compare the maximized likelihood when H_0 is true (L_0) to the maximized likelihood when H_0 is not true (L_1)
- Test for a specific coefficient: $\beta_i=0$ vs. $\beta_i \neq 0$

GOG 502/PLN 504 Youqin Huang 15

Example

- The effect of age on the likelihood of getting a chronic disease
- Dependent var: Have a chronic disease or not (1 vs. 0)
- Independent variable: age

GOG 502/PLN 504 Youqin Huang 16

Exploratory Analysis

CHD				
CHD	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	57	57.00	57	57.00
1	43	43.00	100	100.00

Staircase Statistics							
Variable	N	Mean	Std. Dev.	Sum	Minimum	Maximum	Label
AGE	100	44.38000	11.7233	4438	20.00000	69.00000	AGE
CHD	100	0.43000	0.49757	43.00000	0	1.00000	CHD

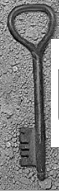
Probability=0.43
Odds=0.43/0.57=0.75

S-shape logistic curve

GOG 502/PLN 504 Youqin Huang 17

Analyze → Regression → Binary logistic

GOG 502/PLN 504 Youqin Huang 18




Step		B	S.E.	Wald	df	Sig.	Exp(B)
1 ^a	AGE	.111	.024	21.254	1	.000	1.117
	Constant	-5.309	1.134	21.935	1	.000	.005

a. Variable(s) entered on step 1: AGE.

- ◆ B=0.111: age has a positive effect on the likelihood of having a chronicle disease
- ◆ Odds ratio: $e^{0.111}=1.117$: A person with one additional year in age is 11.7% more likely to have a chronicle disease.
- ◆ For b, focus on whether it is larger than 0 or not; for odds ratio (exp(b)), focus on whether it is larger than 1 or not.

GOG 502/PLN 504 Youqin Huang 19




Hypothesis Test

Step		B	S.E.	Wald	df	Sig.	Exp(B)
1 ^a	AGE	.111	.024	21.254	1	.000	1.117
	Constant	-5.309	1.134	21.935	1	.000	.005

a. Variable(s) entered on step 1: AGE.

Age has a significant effect on having chronicle disease.

GOG 502/PLN 504 Youqin Huang 20




Goodness of Fit

- ◆ -2 Log Likelihood ratio (-2 LogL)
 - L_0 : maximum of likelihood for model with no covariates (H_0 is true), L_1 : for the model with covariates
 - Chi-square distribution, $df = \#$ of independent variable

$$-2\text{LogL} = -2\text{Log}\left(\frac{L_0}{L_1}\right)$$

- ◆ L: the larger, the better
- ◆ -2 Log L: the smaller, the better

GOG 502/PLN 504 Youqin Huang 21




Goodness of Fit

- ◆ R^2 in OLS is a very useful measure
- ◆ There is no equivalent in logistic regression
- ◆ Various "Pseudo" R^2
 - McFadden's R^2
 - Cox and Snell R^2
 - Nagelkerke R^2

$$R^2 = 1 - \frac{-2\log L_1}{-2\log L_0}$$

- Tends to be smaller than R^2 in OLS
- Does not refer to explained variation in Y

GOG 502/PLN 504 Youqin Huang 22




Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	107.353 ^a	.254	.341

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

- ◆ L: the larger, the better
- ◆ -2 Log L: the smaller, the better

GOG 502/PLN 504 Youqin Huang 23



Summary

- ◆ Concepts: odds, odds ratio, logit
- ◆ Logistic regression
- ◆ Interpretation of coefficients
- ◆ Hypothesis Test, Goodness of Fit

GOG 502/PLN 504 Youqin Huang 24