Inference for a logit model

Susan Thomas
IGIDR, Bombay

September 17, 2008

Recap

- Setting up the model for the logit problem (how probable is it that a woman participates in the workforce, given her education.)
- Conditional expectations
- The logistic transformation
- MLE for the problem: Bernoulli + logistic transformation

Logit model results for being participating in the workforce

- Likelihood function is:

$$f_\theta(y_i) = \left(\frac{\exp \beta_0 + \beta_1 X}{(1 + \exp(\beta_0 + \beta_1 X))}\right)^{Y_i} \left(\frac{1}{(1 + \exp(\beta_0 + \beta_1 X))}\right)^{1-Y_i}$$

- Fitting the data to the model, we find that:

$$\beta_0 = -1.4, \beta_1 = 0.15$$

- We find that the value of the logl at $\hat{\beta}_0, \hat{\beta}_1$ is -4702.71.

Interpreting model results

- The log odds ratio with no education is $\beta_0 = -1.4$.
- We can extract the probability of workforce participation given no education from this:

$$
\begin{aligned}
\frac{\hat{p}(Y=1|X=0)}{\hat{p}(Y=0|X=0)} &= e^{-1.4} = 0.25 \\
\frac{\hat{p}(Y=1|X=1)}{\hat{p}(Y=0|X=1)} &= e^{-1.4+0.15} = 0.29 \\
\frac{\hat{p}(Y=1|X=10)}{\hat{p}(Y=0|X=10)} &= e^{-1.4+1.5} = 1.1 \\
\frac{\hat{p}(Y=1|X=20)}{\hat{p}(Y=0|X=20)} &= e^{-1.4+3.0} = 5.0
\end{aligned}
$$

Inference about chosen model parameter values

Inference about H_0

- A null of interest: education does not influence workforce participation.

$$H_0 : \beta_1 = 0$$

- How do we test this is not the truth, and that $\beta_1 = 0.15$ is? Use the LR test.
- First step: estimate the "restricted model" where β_1 is set forcibly to 0.
- The restricted model has the following values:

$$\beta_0 = 0.37, l = -4857.61$$

- Second step: LR test form: $-2 \log (L_R - L_U)$. Benchmark: $\chi^2(1)$. At 95% confidence, $\chi^2(1) = 3.84$.
- Inference: $l_R = -4857.61, l_U = -4702.71$

$$LR = -2 * (-4857.61 + 4701.71) = 312$$

Interpreting inference results

- $312 >> 3.84$.
  So, at a 95% confidence, we reject the null that $\beta_1 = 0$.
- Syntax: if the LR test was less than 3.84, the language would be that "we do not reject the null."
- By choosing a level of 5%, we accept that in 5% of hypothetical samples from the population, we reject a true hypothesis like $\beta_1 = 0$ by chance.
- With $LR = 312$, we do not find any support for the null $\beta = 0$.

Economic interpretation: Causation vs. correlation

- The empirical analysis establishes correlation between Y and X.
- However, we want to know whether education **causes** workforce participation particularly since economic policy can be founded on such analysis.
- For instance: if all the women were given one more year of education, would it increase the odds that they participate in the workforce?
- Ans: Not necessarily. Other factors could drive the choice of education – like a preference for studying, or a signal of ability.
  Without taking all these factors into account, we can't make the link to causality.

Tests for different forms of the alternative, H_a

- Example: $H_0 : \beta_1 = 0$ means no impact of education on workforce participation.
- Default alternative: $H_1 : \beta_1 \neq 0$.
  Another alternative: $H_1 : \beta_1 > 0$.
- How do we test the null under this alternative?
  Use the one-tailed test, rather than the usual "two-tailed" test.
- Two tailed tests have the critical region located symmetrically on both sides of the test-statistic distribution center.
  One tailed test have the critical region pooled all on one side of the distribution.
- These are also called "signed" tests. It refers to the nature of the alternative hypothesis.

Examples of “critical values” under two-tailed vs. one-tailed tests

- If the test statistic is gaussian distributed, critical values for different confidence levels are:
  1. 95% confidence, critical region 5%
     Two tailed test critical value: $x = 1.96$
     One tailed test critical value: $x = 1.645$
  2. 99% confidence, critical region 1%
     Two tailed test critical value: $x = 2.58$
     One tailed test critical value: $x = 2.33$

Setting up the signed LR test

- The LR-statistic is $\chi^2(1)$.
- A new test statistic ω is defined as follows:

$$\omega = \mathrm{sign}(\hat{\beta}_1)\sqrt{\mathrm{LR}}$$

where

$$\mathrm{sign}(x) = \begin{array}{ll} +1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{array}$$

- Then, $\omega \sim \mathrm{N}(0, 1)$ approximately.
- If theory says that $\beta_1 > 0$, then the critical region is chosen as $(\omega > \text{critical value})$.
  If theorhy says that $\beta_1 < 0$, then the critical region is chosen as $(\omega < -\text{critical value})$.
- Example, for a test at 95% confidence wrt a gaussian distribution, and critical value is 0.05 one-tailed, then

$$\omega > 1.65$$

Testing $H_a : \beta_1 > 0$ for education in workforce participation

- $l_R = -4857.61, l_U = -4702.71$

$$LR = -2 * (-4857.61 + 4701.71) = 312$$

- $\omega = +\sqrt{312} = 17.66$
- The 5% one-tailed test critical value for the gaussian is 1.645
  $\omega = 17.66 > 1.645$.
- Inference? H_0 is rejected even against a different alternative.

Case of conflicting inference between one-tailed and two-tailed tests

- In the earlier example, the test statistics were very far away from the critical values.
- It could be that the test-statistics are close to the critical value – in that case, the one-tailed and two-tailed test could give conflicting inference.
- Example, say the $\mathrm{LR} = 3.25$ (close to 3.84). Since $\omega = \sqrt{LR} = 1.8$ (close to 1.65).
  But LR "fails to reject" H_0 and ω rejects H_0.
- The one-tailed test is said to be more powerful than the two-tailed test.
- In such situations, rather than chose one result vs. the other, the objective is to
  strengthen the dataset,
  and thus, strengthen the test and inference.

Inference about the chosen model itself

Inference about the model

- Do we have the right model? Or is our model "misspecified"?
  For this, we need an alternative model itself.
- The model includes: independence, type of distribution used for Y, whether it is identical for each observation, the form of variation across observation.

Alternative model for workforce participation

- Example: $f(Y=1|X) = \pi(X)$
  It is not Bernoulli with the probability parameter as a function of (β_0, β_1) but $(\pi_0, \pi_1, \pi_2, \pi_3, \ldots, \pi_J)$.
  Where J is the number of categories of education used.
  With 20 years of education, $J = 20$.
- Then, the alternative for $f(Y_i)$ of effect of education on workforce participation is

$$\text{logit}(p(Y_i)) = \sum_{j=0}^{J} \pi_j \mathrm{I}_{(X_i=j)}$$

- Here, $\mathbf{I}_{(X_i=j)}$ is an indicator function, with

$$\mathrm{I}_{(X_i=j)} = 1, \text{ if } X_i = j, \text{ and } = 0, \text{ if } X_i \neq j$$

Alternative model for workforce participation

- Once the alternative is identified and formulated, we check whether we can calculate the log likelihood function for this model.
- If that can be done, we can apply the LR test framework to test the null of our original model against the alternative.
- Applying the alternative to the problem:

$$l_{(Y_1,Y_2,\ldots,Y_N|X_1,X_2,\ldots,X_N)}(\pi_0, \pi_1, \ldots, \pi_{20}) = -4688.92$$

- Now we can do inference.

Inference approach for model misspecification

- Inference step 1: identify the test.
  LR test.
- Inference step 2: to use the LR test, identify the "restricted" model and the "unrestricted model".
  $H_{\pi_0,\pi_1,\ldots,\pi_{20}}$ is the *unrestricted model*
  H_{β_0,β_1} is the *restricted model*
  Key point: the focus of inference is to determine whether the "restricted model" is a *significantly worse* description of the data than the "unrestricted model".
- Inference step 3: calculate the statistic.
  $L_{H_0} = -4702.7, L_{H_a} = -4688.92$

$$\text{LR} = -2(-4702.7 + 4688.92) = 27.59$$

- Inference step 4: compare with the benchmark distribution.
  $\chi^2(n)$ – what is n here?

Benchmark for model misspecification tests

- We know that variables that are sum of normal variate squared are generally distributed $\chi^2(n)$.
- The general definition is:
  Suppose $Z \sim N(0, 1)$. Then $Z^2 \sim \chi^2(1)$.
  If $Z_1, \ldots, Z_m$ are independently $N(0, 1)$, and
  $W_m = Z_1^2 + \ldots + Z_m^2$, then $W_m \sim \chi^2(m)$.
- In the misspecification LR-test, the degrees of freedom for the benchmark distribution are found as:
  the difference between the number of parameters of the H_0 model vs. H_a model.
- In our case, H_a has 21 parameters. H_0 has 2 parameters. Thus, the degrees of freedom for the relevant χ^2 distribution is $n = 19$.

Inference about the validity of the original model

- At 95%, $\chi^2(19) = 30.1$. At 99%, $\chi^2(19) = 36.2$
- The LR test is $27.59 < 30.1, 36.2$
- The restricted model H_{β_0,β_1} cannot be rejected against the unrestricted model $H_{\pi_0,\pi_1,\ldots,\pi_{20}}$
- Ie, the model where the probability of workforce participation, $\hat{f}(Y = 1|X)$ is
  a free-standing function of the years of education
  does not predict significantly better than
  as an output of a linear function ($\beta_0 + \beta_1 \times$ years of education)

Model misspecification tests – *goodness of fit* tests

- Principle is to stratify observations by some common factor as the alternative model.
- For each strata, the sample frequency is compared against the model predicted mean.
- Statistical literature refers to this as “testing the goodness of fit of the model”.
- Econometric literature refers to this as “testing the validity of the model”.