FINAL LAB

Barum Park

Department of Sociology
New York University

Apr. 10, 2018

Before we start ...

Any questions regarding last class?

!!! WARNING !!!

!!! PLEASE CONSULT YOUR TEXTBOOKS RATHER THAN USING THESE SLIDES TO STUDY !!!

THE TEXTBOOKS THAT YOU WERE ASSIGNED WENT THROUGH MANY REVISIONS. SO YOU CAN TRUST THEIR CONTENT

These slides, on the other hand, were created by a poor GRADUATE STUDENT from the top of his head !!

What follows will be quite technical. I couldn't find a better way to explain it. Please bear with me and interrupt me as often as possible.

The Problem of Variance in Logit/Probit Models

- Both Logit and Probit models can represented using a latent variable formulation:

$$y_i^* = \alpha + \beta x_i + \epsilon_i^*$$

$$y_i = \begin{cases} 1, & \text{if } y_i^* > \tau \\ 0, & \text{otherwise} \end{cases}$$

- Here y^* is a latent variable which we do not observe, τ is a threshold parameter determining when a value of y_i^* leads to an observed value of $y_i = 1$, and ϵ_i^* is the error term of the latent regression
- If we assume that $\epsilon_i^* \sim \text{Normal}(0, \sigma)$, we obtain the Probit model; if we assume $\epsilon_i^* \sim \text{Logistic}(0, \sigma)$, we obtain the Logit model.

What are we estimating?

Let us consider the probit model and let Φ be the cumulative distribution function of a standard Normal variable. Then we have

$$
\begin{aligned}
\Pr[y_i = 1] &= \Pr[y_i^* > \tau] = \Pr[\alpha + \beta x_i + \epsilon_i^* > \tau] \\
&= \Pr[\epsilon_i^* > \tau - \alpha - \beta x_i] \\
&= 1 - \Pr[\epsilon_i^* \leq \tau - \alpha - \beta x_i] \\
&= 1 - \Pr\left[\frac{\epsilon_i^*}{\sigma} \leq \frac{\tau - \alpha - \beta x_i}{\sigma}\right] \\
&= 1 - \Phi\left(\frac{\tau - \alpha - \beta x_i}{\sigma}\right) \\
&= \Phi\left(\frac{\alpha - \tau + \beta x_i}{\sigma}\right)
\end{aligned}
$$

- This leads to

$$\Phi^{-1}(p_i) = \left[\frac{\alpha - \tau}{\sigma}\right] + \left[\frac{\beta}{\sigma}\right] x_i$$

  where Φ^{-1} is the inverse cumulative distribution function of the standard Normal distribution and $p_i = \Pr[y_i = 1]$.
- Had we assumed $\epsilon^* \sim$ Logistic(0,1), all that would have changed is that we have Λ^{-1} instead of Φ^{-1}, where Λ^{-1} is the inverse cumulative distribution function of the standard Logistic distribution
- Note that

$$\Lambda^{-1}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right)$$

  our familiar formula.

What are we estimating? (cont.)

So, we have

$$\Pr[y_i = 1] = \Phi\left(\frac{\alpha - \tau}{\sigma} + \frac{\beta}{\sigma} x_i\right).$$

- This model is not identified: for example,
  - $(\alpha = 2, \tau = 1)$ and $(\alpha = 3, \tau = 2)$ lead to the same probability
  - so does $(\beta = 1, \sigma = 1)$ and $(\beta = 100, \sigma = 100)$
- As infinitely many parameter values lead to the exact same probability, we cannot estimate $\alpha, \tau, \beta, \sigma$ from the data

Identification of α

$$\Pr[y_i = 1] = \Phi\left(\frac{\alpha - \tau}{\sigma} + \frac{\beta}{\sigma} x_i\right)$$

- Given fixed values of σ and β, we can either estimate α by assuming $\tau = 0$, estimate τ by assuming $\alpha = 0$, or estimate the difference $\alpha' = \alpha - \tau$
- All methods lead to the same predicted probability and are also "substantively" equivalent, given that the latent variable does not have a "natural" scale. So, let us use

$$\Pr[y_i = 1] = \Phi\left(\frac{\alpha'}{\sigma} + \frac{\beta}{\sigma} x_i\right).$$

Identification of β

- What about β?
- We can identify the ratios $\alpha^* = \alpha'/\sigma$ and $\beta^* = \beta/\sigma$, which leads then to

$$\Pr[y_i = 1] = \Phi\left(\alpha^* + \beta^* x_i\right)$$

- So, in a strict sense, when running a Logit/Probit in `STATA`, we are not estimating (α, β) but (α^*, β^*), and

$$(\alpha^*, \beta^*) = (\alpha', \beta) \text{ if and only if } \sigma = 1$$

and

$$(\alpha^*, \beta^*) = (\alpha, \beta) \text{ if and only if } \sigma = 1 \text{ and } \tau = 0$$

Summary

- Recall that we started from

$$y_i^* = \alpha + \beta x_i + \epsilon_i^*$$

$$y_i = \begin{cases} 1, & \text{if } y^* > \tau \\ 0, & \text{otherwise} \end{cases}$$

- We came to the conclusion that we are able to estimate only

$$\alpha^* = \frac{\alpha'}{\sigma} = \frac{\alpha - \tau}{\sigma} \text{ and } \beta^* = \frac{\beta}{\sigma}.$$

- Hence, when we are running `logit` or `probit` in STATA, we are estimating

$$\alpha^* \text{ and } \beta^*$$

- And we are estimating

$$\alpha \text{ and } \beta$$

  only if we assume that $\tau = 0$ and $\sigma = 1$.

Implications

- In most situations, this is no problem because
  1. the “latent” variable has no natural scale and
  2. we get the right probabilities, independent of how we parameterize our model (i.e., whether setting $\tau = 0$ and $\sigma = 1$ or not)
- Then why do we care?
- Because problems arise when we want to compare logit/probit coefficients across different models

Implications (cont.)

- Consider two *latent* regressions on the same outcome

$$y_i^* = \alpha + \beta x_i + \epsilon_i^*$$
$$y_i^* = \alpha + \beta x_i + \gamma z_i + \epsilon_i^{**}$$

- Call the first equation the short one and the second the long one. Notice that the value of α and β are the same in both equations
- Let $Var(\epsilon_i^*) = \sigma^*$ and $Var(\epsilon_i^{**}) = \sigma^{**}$ and note that it must be the case that $\sigma^* > \sigma^{**}$
- The intuitive reason is the following: you cannot explain less variance of your dependent variable by adding more variables. Accordingly, the residual variance must decrease as you enter more predictors into your model (given that you keep all the old ones).[1]

[1] This result needs a little bit of linear algebra but is simple. For the curious, it is proven in the appendix.

Implications (cont.)

- Let us focus on the β coefficient, which is of most interest.
- The important point is that because we are only able to estimate β/σ and not β directly, it follows we are estimating in the short regression

$$\beta^* = \frac{\beta}{\sigma^*}$$

and in the long regression

$$\beta^{**} = \frac{\beta}{\sigma^{**}}$$

- But as $\sigma^* > \sigma^{**}$, our estimates from the short regression will be consistently smaller in magnitude then the estimates from our long regression, even if the true parameter values are the same!
- Thus, to compare coefficients across models, we must ensure that the latent error variance is the same across models

Proportion Direct “Effect”

- Consider a path model, where $x \rightarrow y$, $x \rightarrow z$, and $z \rightarrow y$.
- Expressing this with equations, where y is binary, we have

$$\begin{aligned} y_i^* &= \alpha_0 + \alpha_1 x_i + \epsilon_i^* \\ z_i &= \beta_0 + \beta_1 x_i + \xi_i \\ y_i^* &= \gamma_0 + \gamma_1 x_i + \gamma_2 z_i + \epsilon_i^{**} \end{aligned}$$

- We are interested in the proportion of the “direct effect” (relative to the “total effect”), which is given as

$$\mu = \gamma_1 / \alpha_1$$

- To calculate μ, γ_1 and α_1 have to be comparable, which we know they are not, since $Var(\epsilon_i^*) \neq Var(\epsilon_i^{**})$. What should we do?

Solution

- The solution lies in deflating $Var(\epsilon_i) = \sigma^*$, so that $\sigma^* = \sigma^{**}$, without changing α_1 (where $\sigma^{**} = Var(\epsilon_i^{**})$).
- This can be done by adding variation to x_i that is independent of x_i itself
- The intuition is that adding a variable that is uncorrelated with x_i to your model will not change the coefficient of x_i (i.e., α_1)[2]

[2] Again, for the curious, the proof is in the appendix.

Solution

- Next, consider the regression

$$\begin{aligned} z_i &= \beta_0 + \beta_1 x_i + \xi_i \\ &= \hat{z}_i + \xi_i \end{aligned}$$

- We might decompose the variance of z_i as

$$Var(z_i) = Var(\hat{z}_i) + Var(\xi_i)$$

  where the first term on the right-hand side is the variation of z_i that co-varies with x_i and the second term is the variation of z_i that is independent of the variation of x_i
- Hence ξ_i is independent of x_i by construction.

- So, let's add ξ_i into our model to obtain

$$y_i^* = \delta_0 + \alpha_1 x_1 + \delta_2 \xi_i + \nu_i^*$$

- Notice that α_1 is exactly the the value that appeared in the regression $y_i^* = \alpha_0 + \alpha_1 x_i + \epsilon_i^*$.
- The last step is to show that

$$Var(\nu_i^*) = Var(\epsilon_i^{**})$$

  where ϵ_i^{**} is the error term from the regression $y_i^* = \gamma_0 + \gamma_1 x_i + \gamma_2 z_i + \epsilon_i^{**}$
- Then we can compare α_1 to γ_1 in order to calculate $\mu = \gamma_1 / \alpha_1$.

- So, let us compare the regressions

$$y_i^* = \delta_0 + \alpha_1 x_i + \delta_2 \xi_i + \nu_i^*$$
$$y_i^* = \gamma_0 + \gamma_1 x_i + \gamma_2 z_i + \epsilon_i^{**}$$

- Now, it must be the case that $\delta_2 = \gamma_2$. Intuitively speaking, this is because the coefficient of predictor, z_i, in a multiple regression of y_i on z_i and x_i can be obtained by 1) residualizing z_i with respect to x_i and 2) running a regression of y on the residualized version of z_i (recall the venn diagram I've shown you in an old lab).
- As ξ_i is the residualized version of z_i and x_i uncorrelated with ξ_i, the coefficients γ_2 and δ_2 have to be the same.[3]

[3] Again, see appendix.

- This leads to

$$
\begin{aligned}
y_i^* &= \delta_0 + \alpha_1 x_i + \gamma_2 \xi_i + \nu_i^* \\
y_i^* &= \gamma_0 + \gamma_1 x_i + \gamma_2 z_i + \epsilon_i^{**}
\end{aligned}
$$

- Now notice that α_1 is the coefficient from the regression $y_i^* = \alpha_0 + \alpha_1 x_i + \epsilon_i^*$. It represents, thus, the "total" effect of x_i on y_i^*. But we can decompose this total effect into a "direct" and "indirect" effect
- It turns out that[4]

$$\alpha_1 = \gamma_1 + \beta_1 \gamma_2$$

- Substituting this into the first equation, we obtain

$$
\begin{aligned}
\alpha_1 x_i + \gamma_2 \xi_i &= (\gamma_1 + \beta_1 \gamma_2) x_i + \gamma_2 (z_i - \beta_1 x_i) \\
&= \gamma_1 x_i + \gamma_2 z_i
\end{aligned}
$$

  Implying that $Var(\delta_0 + \alpha_1 x_i + \gamma_2 \xi) = Var(\gamma_0 + \gamma_1 x_i + \gamma_2 z_i)$. As $Var(y_i^*)$ is fixed, it follows that $Var(\nu_i^*) = Var(\epsilon_i^{**})$.

[4] Again, a heuristic proof is in appendix

- In other words, because $Var(\nu_i^*) = Var(\epsilon_i^{**}) = \sigma^2$, the regression coefficients that we are estimating in the two equations are

$$\alpha_1^* = \frac{\alpha_1}{\sigma}, \gamma_2^* = \frac{\gamma_2}{\sigma}$$

in the first equation and

$$\gamma_1^* = \frac{\gamma_1}{\sigma}, \gamma_2^* = \frac{\gamma_2}{\sigma}$$

in the second equation.

- The important point is that we are dividing all of these coefficients by the same constant. Thus,

$$\frac{\gamma_1^*}{\alpha_1^*} = \frac{\gamma_1/\sigma}{\alpha_1/\sigma} = \frac{\gamma_1}{\alpha_1} = \mu$$

Done.

Summary

- To summarize, when fitting logit/probit models, regression coefficients cannot be directly compared because the residual variance of the latent regression changes when different variables are added to the model.
- To compare how the coefficient of x_i changes when z_i is added to the model, we have to make sure that the residual variance is equal when regressing y on x and when regressing y on x *and* z
- This can be done with the following procedure:
  1. regress z_i on x_i to obtain the residuals ξ_i
  2. Run a logit/probit regression of y_i on x_i and ξ_i (this is the "total effect")
  3. Run a logit/probit regression of y_i on x_i and z_i (this is the "direct effect")
  4. Compare the coefficients of x_i in 2. and 3.

Appendix: Why does residual variance decline when adding variables?

Consider two regressions

$$y = X\beta + \epsilon \text{ and } y = X\beta + z\gamma + u$$

Then,

$$\begin{aligned} u'u &= (y - X\beta - z\gamma)'(y - X\beta - z\gamma) \\ &= y'y - 2y'X\beta - 2y'z\gamma + \beta'X'X\beta + 2\beta'X'z\gamma + \gamma^2 z'z \\ &= \epsilon'\epsilon - 2y'z\gamma + 2\beta'X'z\gamma + \gamma^2 z'z \\ &= \epsilon'\epsilon - \gamma z'\Big(2(y - X\beta - \gamma z) + \gamma z\Big) \\ &= \epsilon'\epsilon - 2\gamma z'u - \gamma^2 z'z \\ &= \epsilon'\epsilon - \gamma^2 z'z \\ &\leq \epsilon'\epsilon \end{aligned}$$

Thus the residual variance can only get smaller when we add more variables to a regression. It will not change if either $\gamma = 0$ or $Var(z) = 0$.

Appendix: Why is the regression coefficient the same when adding uncorrelated variable?

Consider the regression

$$y = X\beta + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon$$

where the set of variables X_1 and X_2 are orthogonal to one another. The normal equations are given as

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix}$$

Using only the second equation, we get

$$X_1'X_1\beta_1 + X_1'X_2\beta_2 = X_1'y$$

and solving for β_1, we obtain

$$\beta_1 = (X_1'X_1)^{-1}(X_1'y - X_1'X_2\beta_2)$$

If the X_1 and X_2 are orthogonal (uncorrelated) $X_1'X_2 = 0$. Hence the second term in the parentheses disappears and we get

$$\beta_1 = (X_1'X_1)^{-1}X_1'y$$

which is the least-squares estimator for β_1 when running a regression of y on X_1.

Appendix: Frisch-Waugh Theorem

Again, consider the regression Consider the regression with normal equations

$$y = X\beta + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon, \quad \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix}$$

Recall from the last appendix that $\beta_1 = (X_1'X_1)^{-1}(X_1'y - X_1'X_2\beta_2)$. The second set of the normal equations leads to $X_2'X_1\beta_1 + X_2'X_2\beta_2 = X_2'y$. Substituting β_1 in to this second equation gives

$$\beta_2 = (X_2'M_1X_2)^{-1}(X_2'M_1y),$$

where $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$ which is often called the "residual maker" matrix as $M_1y = y - X_1\beta_1 = e_1$, which is the vector of residuals that are generated by regressing y on X_1 (without including X_2 in the model). As M_1 is idempotent (meaning that $M_1M_1 = M_1$) and symmetric, we have

$$\beta_2 = (\tilde{X}_2'\tilde{X}_2)^{-1}\tilde{X}_2'\tilde{y}$$

where $\tilde{X}_2 = MX_2$ and $\tilde{y} = M_2y$, implying that the regression coefficients, β_2, from a multiple regression of y on X_1 and X_2 is equal to the vector regression coefficients obtained when the X_1-residualized y (i.e., $\tilde{y}$) is regressed on X_1-residualized X_2. Note that by the properties of M_1, the step of residualizing y can be omitted.

Appendix: Decomposition of total into net and direct "effects"

A heuristic "proof." Assume without loss of generality that all variables that appear below are mean centered, so that $E[xy] = Cov(x,y)$ and $E[x^2] = Var(x)$. Now, consider the set of equations:

$$\begin{aligned} y &= \beta x + \epsilon \\ z &= \gamma x + \nu \\ y &= \delta x + \lambda z + \mu, \end{aligned}$$

where ϵ, ν, μ are error-terms with $E[\epsilon] = E[\nu] = E[\mu] = 0$.
Multiplying through by x for all equations and taking expectations, we obtain

$$\begin{aligned} Cov(x,y) &= \beta Var(x) \\ Cov(x,z) &= \gamma Var(x) \\ Cov(x,y) &= \delta Var(x) + \lambda Cov(x,z) = \delta Var(x) + \lambda\gamma Var(x) \\ &= (\delta + \lambda\gamma) Var(x) \end{aligned}$$

It follows that $\beta = (\delta + \lambda\gamma)$.