EEP/IAS 118 - Introductory Applied Econometrics, Section 8

Pierre Biscaye and Jed Silver

October 2021

Agenda

- Categorical variables
- Bad controls
- Confidence intervals for predicted values
- Chow tests (a special kind of F-test)
- Linear Probability Models

Categorical variables

- Motivation: We want to incorporate qualitative information, such as the color of a car
- Idea: use a categorical variable - assign numbers to each possible category
- How to include in a regression and interpret the output?

Categorical variables: Example

Say we want to estimate how car color affects sales price at a local dealership. Our model:

$$price = \beta_0 + \beta_1 color + \beta_2 mileage + \beta_3 volume + u$$

where $color = \{red, blue, green, gray\}$

Categorical variables: Example

Say we want to estimate prices of cars sold at a local dealership. Our regression model:

$$price = \beta_0 + \beta_1 color + \beta_2 mileage + \beta_3 volume + u$$

where $color = \{red, green, blue, gray\}$

To actually run this as a regression, we need something like $color = 1$ for red, $color = 2$ for green, $color = 3$ for blue, and $color = 4$ for gray.

Question: Is the above sufficient? If so, why? If not, why not?

Categorical variables: Example

Say we want to estimate prices of cars sold at a local dealership. Our regression model:

$$price = \beta_0 + \beta_1 color + \beta_2 mileage + \beta_3 volume + u$$

where $color = \{red, green, blue, gray\}$

To actually run this as a regression, we need something like $color = 1$ for red, $color = 2$ for green, $color = 3$ for blue, and $color = 4$ for gray.

Question: Is the above sufficient? If so, why? If not, why not? No. It doesn't make sense to talk about a 1 unit increase in color in this case. **Next question:** How do we move to something interpretable?

From categorical to dummy

Our new model:

$$price = \beta_0 + \delta_0 green + \delta_1 blue + \delta_2 gray + \beta_1 mileage + \beta_2 volume + u \quad (1)$$

where

- $green = 1$ if car is green, 0 otherwise
- $blue = 1$ if car is blue, 0 otherwise
- $gray = 1$ if car is gray, 0 otherwise

From categorical to dummy

Our new model:

$$price = \beta_0 + \delta_0 green + \delta_1 blue + \delta_2 gray + \beta_1 mileage + \beta_2 volume + u \tag{2}$$

Questions:

- Q: Why isn't there something like $\delta_3 red$ in the above model?

- β_0 now provides information about outcomes for the omitted category

From categorical to dummy

Our new model:

$$price = \beta_0 + \delta_0 green + \delta_1 blue + \delta_2 gray + \beta_1 mileage + \beta_2 volume + u \quad (2)$$

Questions:

- Q: Why isn't there something like $\delta_3 red$ in the above model?
- A: It would be *perfectly multicollinear* with the other color dummies. Always *leave out* one of your categories.
- Q: How do we interpret the δ parameters?

- β_0 now provides information about outcomes for the omitted category

From categorical to dummy

Our new model:

$$price = \beta_0 + \delta_0 green + \delta_1 blue + \delta_2 gray + \beta_1 mileage + \beta_2 volume + u \tag{2}$$

Questions:

- Q: Why isn't there something like $\delta_3 red$ in the above model?
- A: It would be *perfectly multicollinear* with the other color dummies. Always *leave out* one of your categories.
- Q: How do we interpret the δ parameters?
- A: Effects *relative* to the omitted category—red cars
- β_0 now provides information about outcomes for the omitted category

Bad controls

- Often, a researcher will be concerned about omitted variable bias.
- Thought: If you have data on the variable, include it as a control.
- However, this may not be desirable variable to include as a control in the regression model, particularly when the variable considered for inclusion is itself an outcome of the independent variable of interest. If some of the effect of variable x_1 on y is through its effect on x_2, then including x_2 in the regression means the coefficient on x_1 no longer accurately captures its impact on y.

Bad control: Scenario

Imagine you are trying to analyze the impact of temperature on conflict (violence, crime, etc)

- Temperature is likely to affect conflict, but is probably correlated with other things that in turn affect conflict (any examples?)

Bad control: Scenario

Imagine you are trying to analyze the impact of temperature on conflict (violence, crime, etc)

- Temperature is likely to affect conflict, but is probably correlated with other things that in turn affect conflict (any examples?)
- One such example: per capita GDP
- Suppose you regress conflict on temperature and per capita GDP, and find that the former effect is not statistically significant at conventional levels, while the latter effect is. What do you conclude?

Bad control: Scenario

Imagine you are trying to analyze the impact of temperature on conflict (violence, crime, etc), controlling for GDP per capita

- Note that temperature affects economic productivity (per capita GDP), so it should really be an outcome variable, not a control
- Doesn't make sense to "hold GDP per capita constant" when estimating the relationship between temperature and conflict, when at least some part of temperature's effect is through income

Dealing with bad controls

What to do when faced with bad controls? Two possible ways

1. Focus on the reduced form relationship without controls (if the X variable is as good as randomly assigned): this tells you the effect if omitted variables are not a concern
2. Argue that variables are not as related as one might think (and provide evidence): address the concerns about OVB

Confidence Intervals for y

There are some instances where we may care about the predicted value of the dependent variable y given some characteristics x

We know that the estimated regression give us $\hat{y}$ which is our best guess for y for and given x. However, $\hat{y}$ is a random variable (just like $\hat{\beta}$) and therefore has uncertainty.

- We can quantify this uncertainty and create a confidence interval for $\hat{y}$ for any specific combination of x_j

Confidence Intervals for y

However, there are two types of CI that we may want to calculate:

1. A confidence interval for the **average** y given $x_1, ..., x_k$
2. A confidence interval for a **particular** y given $x_1, ..., x_k$

You can think of the difference as being the answer to these two questions:

1. How uncertain are we about the average income for people with certain characteristics x?
2. If we asked a particular person with certain characteristics x their income, what range would cover 95% of responses?

Confidence Intervals for average y

Recall that regression gives us an estimate of y given x:

$$\widehat{\mathbb{E}}[y|x_1, x_2, x_2] = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

- If we want the best estimate for a particular value of x_j, we just plug those values into the equation
- To get a CI, then we need to find the standard error for this prediction
- Recall that β_0 takes on the predicted value of y when all the x_j are zero

$$\hat{\beta}_0 = \hat{E}(y|x_1 = 0, x_2 = 0, x_3 = 0)$$

Confidence Intervals for average y

$$\hat{\beta}_0 = \hat{E}(y|x_1 = 0, x_2 = 0, x_3 = 0)$$

- Therefore, if we transform our x_j by subtracting the values (α_j) for which we want a prediction:

$$y = \beta_0 + \beta_1(x_1 - \alpha_1) + \beta_2(x_2 - \alpha_2) + \beta_3(x_3 - \alpha_3)$$

Then

$$\hat{\beta}_0 = \hat{E}(y|x_1 = \alpha_1, x_2 = \alpha_2, x_3 = \alpha_3)$$

When we run the regression with these transformed variables, $\hat{\beta}_0$ will then be the best prediction and Stata will produce the correct SE.

Confidence Intervals for average y

Steps for CI on **average** y for observation with characteristics $\{\alpha_1, \cdots, \alpha_k\}$:

1. Generate new variables: $\tilde{x}_j = x_j - \alpha_j$.
2. Run the regression of: $y = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{x}_1 + ... + \tilde{\beta}_k \tilde{x}_k + \tilde{u}$
3. Then $\hat{\mathbb{E}}[y|x_1 = \alpha_1, ..., x_k = \alpha_k] = \tilde{\beta}_0$ and the standard error for this estimate is $SE(\tilde{\beta}_0)$.
4. Plug these values into the formula for confidence intervals and interpret.

$$[\tilde{\beta}_0 - c \cdot SE(\tilde{\beta}_0) \ , \ \tilde{\beta}_0 + c \cdot SE(\tilde{\beta}_0)]$$

Confidence Intervals for average y: Example

For an example, let's use Woolridge's birthweight data. Let's say we want to find a prediction for average birthweight for babies with family income of 14.5 in \$ thousands ($ln(14.5) = 2.674$), mothers with 12 years of education, and with 2 older siblings ($parity = 3$)

Running the standard regression:

$$\begin{aligned}\widehat{bwght} &= 105.66 + 2.13ln(faminc) + 0.317meduc + 1.53parity \\ \hat{y} &= 105.66 + 2.13(2.674) + .317(12) + 1.53(3) \\ &= 119.75 \text{ ounces}\end{aligned}$$

Which is our best guess for $\hat{y}_{faminc=14.5,meduc=12,parity=3}$

Confidence Intervals for average y: Example

To get the SE of this prediction, we run:

$$bwght = \beta_0 + \beta_1(lfaminc - 2.674) +$$

$$\beta_2(meduc - 12) + \beta_3(parity = 3) + u$$

```
------------------------------------------------------------------------------
bwght |      Coef.   Std. Err.      t    P>|t|
------------+-----------------------------------------------------------------
lfaminc_0 |   2.131266   .6505986     3.28   0.001
meduc_0 |   .3171976   .2519682     1.26   0.208
parity_0 |   1.526144   .6119145     2.49   0.013
_cons |   119.6405   1.006928   118.82   0.000
------------------------------------------------------------------------------
```

Note, how now the "cons" takes on the predicted value and has a standard error!

Confidence Intervals for average y: Example

Using this output, the 95% confidence interval for the average birthweight for babies given family income of \$14,500, mothers with 12 years of education, and with 2 older siblings is:

$$\left[119.64 - 1.96(1.007), 119.64 + 1.96(1.007)\right] = \left[117.6653, 121.6158\right]$$

Confidence Intervals for a particular y

Now let's turn to how we can create a confidence interval of y for a *particular* individual with certain x characteristics.

- This is different (larger) than our CI for the *average* y in a sub-population with the same characteristics.
- This is because we need to account for both the variance in our calculation of $\hat{y}$ as well as the variance in the unobserved error term u.

Confidence Intervals for a particular y

Let's see how to think about this using our example. Let $bwght^0$ denote the value for which we want to construct a confidence interval:

$$bwght^0 = \beta_0 + \beta_1 lfaminc^0 + \beta_2 meduc^0 + \beta_3 parity^0 + u^0$$

Our best prediction of $bwght^0$ is $\widehat{bwght^0}$, where

$$\widehat{bwght}^0 = \hat{\beta}_0 + \hat{\beta}_1 lfaminc^0 + \hat{\beta}_2 meduc^0 + \hat{\beta}_3 parity^0$$

Now there is some error associated with using $\widehat{bwght}^0$ to predict $bwght^0$:

$$\begin{aligned}\hat{u}^0 = bwght^0 - \widehat{bwght}^0 &= \beta_0 + \beta_1 lfaminc^0 + \beta_2 educ^0 + \beta_3 par^0 + u^0 \\ &\quad - (\hat{\beta}_0 + \hat{\beta}_1 lfaminc + \hat{\beta}_2 meduc + \hat{\beta}_3 parity)\end{aligned}$$

Confidence Intervals for a particular y

To get a confidence interval, we need to quantify the variance of the error in this prediction:

$$
\begin{aligned}
Var(\hat{u}^0) &= Var(bwght^0 - \widehat{bwght}^0) \\
&= Var(\beta_0 + \beta_1 lfaminc^0 + \beta_2 educ^0 + \beta_3 parit^0 + u^0 - \widehat{bwght}^0) \\
&= Var(u^0 - \widehat{bwght}^0) \\
&= Var(\widehat{bwght}^0) + Var(u^0) \\
&= Var(\widehat{bwght}^0) + \sigma^2 \\
\widehat{Var(\hat{u}^0)} &= Var(\widehat{bwght}^0) + \hat{\sigma}^2 \\
&= Var(\widehat{bwght}^0) + \frac{\sum \hat{u}_i^2}{n-k-1} = Var(\widehat{bwght}^0) + \frac{SSR}{n-k-1}
\end{aligned}
$$

Confidence Intervals for a particular y

$$\widehat{Var(\hat{u}^0)} = Var(\widehat{bwght}^0) + \frac{SSR}{n-k-1}$$

There are thus two sources of variation in $\hat{u}^0$ which create uncertainty for predicting y for a particular observation:

1. The sampling error in $\widehat{bwght}^0$ which arises because we have estimated the population parameters (β).
2. The variance of the error in the population (u^0).

How to generate a confidence interval for the prediction?

- Compute the $Var(\widehat{bwght}^0)$ exactly as before
- Second we can compute $\frac{SSR}{n-k-1}$ from our regression output
- Then the 95% confidence interval for $bwght^0$:

$$\hat{y} \pm 1.96 \cdot se(\hat{u}^0)$$

Confidence Intervals for a particular y^i: Steps

1. Generate new variables: $\tilde{x}_j = x_j - \alpha_j$.
2. Run the regression of: $y = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{x}_1 + ... + \tilde{\beta}_k \tilde{x}_k + \tilde{u}$
3. Then $\hat{\mathbb{E}}[y|x_1 = \alpha_1, ..., x_k = \alpha_k] = \tilde{\beta}_0$ and the standard error for this estimate is $SE(\tilde{\beta}_0) = \sqrt{Var(\widehat{y^i})}$.
4. Get an estimate for the variance of $\hat{u} = \hat{\sigma}^2$ from the R regression output (you can call the object $sigma$ from the summary of your regression).
5. Compute the standard error: $\sqrt{SE(\hat{u}^i) = SE(\tilde{\beta}_0)^2 + \hat{\sigma}^2}$.
6. Plug these values into the formula for confidence intervals and interpret.

Chow test

- We have seen that interacting dummy variables with a) other dummies and b) continuous variables allows us to test whether different groups have different intercepts and different slopes, respectively.
- We may also wish to test the null that two groups follow the same regression function, against the alternative that one or more of the slopes or intercepts differ across groups.
- In other words, we may want to test that there are *any* differences in model parameters between two groups.
- Can do this with F test, but sometimes easier to use a Chow test.

Chow test: Example using sleep75 data set

Suppose you have the following model of sleep (in minutes per week):

$$sleep = \beta_0 + \beta_2 age + \beta_4 totwrk + u$$

BUT you suspect that the relationship between $sleep$ and age and $totwrk$ is different if you have young kids vs not. There are two ways you could formally test this hypothesis:

- F-test: compare unrestricted and restricted regressions
- Chow test: compare restricted regression and regressions for subsamples of interest

F-test approach

If you suspect that this whole regression might be different if we ran it for only people with young kids, that's equivalent to saying that each of the βs is different depending on whether someone has young kids. What would the unrestricted regression be in this case?

F-test approach

We can rewrite restricted and unrestricted regressions as:

$$\text{Unrestricted} : sleep = \beta_0 + \beta_1 yngkids + \beta_2 age + \beta_3 yngkids * age + \beta_4 totwrk + \beta_5 yngkids * totwork + e$$

$$\text{Restricted} : sleep = \beta_0 + \beta_2 age + \beta_4 totwrk + e$$

The F-test that will tell us whether there is a significant difference between these two models:

$$H_0 : \beta_1, \beta_3, \beta_5 = 0$$

$$H_1 : \text{not } H_0$$

Practice Interpreting Interactions

$$sleep = \beta_0 + \beta_1 yngkids + \beta_2 age + \beta_3 yngkids * age + \beta_4 totwrk + \beta_5 yngkids * totwork + e$$

1. Interpret $\beta_1, \beta_3, \beta_5$.
2. What is the average number of minutes slept per week for individuals who are 50 years old, work 40 hours per week, and do not have young kids (write in terms of the β coefficients)?
3. What is the average number of minutes slept per week for individuals who are 30 years old, work 45 hours per week, and have a young child?

Practice Interpreting Interactions

$$sleep = \beta_0 + \beta_1 yngkids + \beta_2 age + \beta_3 yngkids * age + \beta_4 totwrk + \beta_5 yngkids * totwork + e$$

1. β_1: the average difference in minutes of sleep for people with young kids relative to those without young kids, holding constant age and total hours worked.
2. β_3: the difference in the change in minutes slept associated with being one year older for people who have young kids vs those who don't, holding total hours worked constant.
3. β_5: the difference in the change in minutes slept associated with working one more hour per week for people who have young kids relative to those who don't, holding age constant.

Practice Interpreting Interactions

$$sleep = \beta_0 + \beta_1 yngkids + \beta_2 age + \beta_3 yngkids * age + \beta_4 totwrk + \beta_5 yngkids * totwork + e$$

- 2) What is the average number of minutes slept per week for individuals who are 50 years old, work 40 hours per week, and do not have young kids (write in terms of the β coefficients)?

- 3) What is the average number of minutes slept per week for individuals who are 30 years old, work 45 hours per week, and have a young child?

Practice Interpreting Interactions

$$sleep = \beta_0 + \beta_1 yngkids + \beta_2 age + \beta_3 yngkids * age + \beta_4 totwrk + \beta_5 yngkids * totwork + e$$

- 2) What is the average number of minutes slept per week for individuals who are 50 years old, work 40 hours per week, and do not have young kids (write in terms of the β coefficients)?
- $\beta_0 + \beta_2(50) + \beta_4(40)$
- 3) What is the average number of minutes slept per week for individuals who are 30 years old, work 45 hours per week, and have a young child?
- $\beta_0 + \beta_1 + \beta_2(30) + \beta_3(30) + \beta_4(45) + \beta_5(45)$

Chow test approach

We get everything we need to compute the F-stat from running the following regressions (A and B) on different subsamples:

- $sleep = \beta_0 + \beta_1 age + \beta_2 totwrk (A : Have\ young\ kids\ only)$
- $sleep = \beta_0 + \beta_1 age + \beta_2 totwrk (B : No\ young\ kids\ only)$

Noting that:

- $SSR_{UR} = SSR_A + SSR_B$
- $q = k + 1$ the hypothesis that each beta is the same across the two groups involves $k + 1$ restrictions.
- The unrestricted model, which we can think of as having a group dummy variable and k interaction terms in addition to the intercept and variables themselves, has $n - 2(k + 1)$ degrees of freedom

Chow test approach

We can use these three facts to rewrite our F-statistic in a way so that we only need to run (1) Restricted Model, (2) Model A and (3) Model B instead of the usual restricted and unrestricted regressions:

$$F = \frac{\Big(SSR_{pooled} - (SSR_A + SSR_B)\Big)/q}{(SSR_A + SSR_B)/(n - 2(k+1))}$$

What this means is that we can calculate the F-statistic that tests whether or not each parameter in our original model (1) is different for household with and without young kids without actually running the unrestricted model.

Note: R example in notes

Linear Probability Model: Intro

Sometimes our y variable can be a dummy (i.e. only takes values of 0 or 1)

Let's say we run our traditional linear regression with a dummy variable as our y:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

where

$$y \in \{0, 1\}$$

We call this type of analysis a "linear probability model" because we interpret results as affecting the probability that $y = 1$.

Linear Probability Model: Interpretation

What changes with a dummy variable as our y?

- We **can't** interpret β_j as the unit change in y given a one-unit increase in x_j holding all other factors fixed
- y either changes from $0 \to 1$ or $1 \to 0$ or doesn't change
- Instead, β_j measures the change in the **probability** of y taking the value 1 when x_j changes by one unit holding all other factors fixed
- $\hat{\beta}_j$ measures the **predicted** change in this probability

Linear Probability Model: Example

Let's say we wanted to study the probability of women working in the labor force (*inlf*). We run a regression with a binary variable as the y indicating whether a woman is in the labor force on several explanatory variables:

$$\widehat{inlf} = 0.586 - 0.0034nwifeinc + 0.038educ + 0.039exper$$
$$- 0.0060exper^2 - 0.017age - 0.262kindslt6 + 0.0130kidsge6$$

How do we interpret the coefficient on education?

Linear Probability Model: Example

Let's say we wanted to study the probability of women working outside the home. We run a regression with a binary variables as the Y indicating a woman was working outside the home on several explanatory variables:

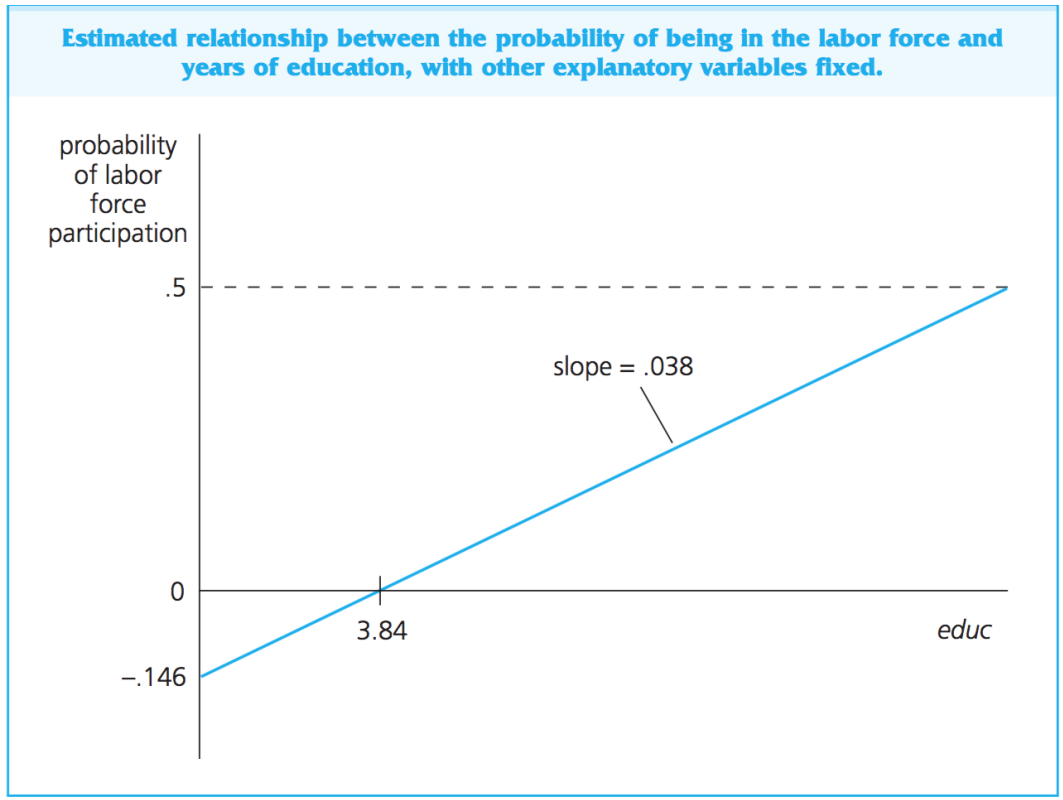
$$\widehat{inlf} = 0.586 - 0.0034nwifeinc + 0.038educ + 0.039exper$$
$$- 0.0060exper^2 - 0.017age - 0.262kindslt6 + 0.0130kidsge6$$

How do we interpret the coefficient on education?

Another year of education increases the predicted probability of labor force participation by 0.038 holding all else constant

Linear Probability Model: Example

We fix the values of the other variables and graph the relationship:



- It is possible to get negative probabilities (yikes)
- The marginal effect of an additional year of education on the probability of participation is constant (at 0.038)

Drawbacks of Linear Prob. Model

1. Predicted probabilities from regression aren't bounded between zero and one
2. There must be heteroskedasticity in the linear probability model, since the variance of y—based on the probability $y = 1$—is now a function of our x variables. This violates our assumption of homoskedasticity:

$$Var(u|x) = Var(u) = \sigma^2$$

Therefore, our standard error calculations are more difficult

Other models (logit, probit, tobit) help deal with these limitations, but LPM usually still useful for intuition