ECON 3740: INTRODUCTION TO ECONOMETRICS

INSTRUCTOR: CHAOYI CHEN
Department of Economics and Finance, University of Guelph

Lecture 17

Lecture outline

Last lecture, we studied the adjusted R^2 and how to choose models between nested/non-nested models. Today, we will

- Predict y when $log(y)$ is the dependent variable
- Study a single dummy variable
  - Motivation to use dummy - incorporate qualitative information
  - Dummy variable Trap
  - Example for using dummy variable

MLR, Further Issue: Adding Regressors to Reduce the Error Variance

- Recall that

$$Var(\widehat{\beta}_j) = \frac{\sigma}{2SST_j(1 - R_j^2)}$$

  - Adding regressors may exacerbate multicollinearity problems ($R_j^2 \uparrow$)
  - On the other hand, adding regressors reduces the error variance ($\sigma^2 \downarrow$)
- Variables that are uncorrelated with other regressors should be added because they reduce error variance ($\sigma^2 \downarrow$) without increasing multicollinearity (R_j^2 remains the same).
- However, such uncorrelated variables may be hard to find.
- Example: Individual Beer Consumption and Beer Prices. If we include individual characteristics in a regression of beer consumption on beer prices leads to more precise estimates of the price elasticity if individual characteristics are uncorrelated with beer prices.

$$log(cons) = \beta_0 + \beta_1 log(price) + indchar + \mu$$

MLR, Further Issue: Predicting y When $log(y)$ is the Dependent Variable

- Let's consider a log-level model

$$log(y) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \mu$$

which implies

$$y = e^{\beta_0+\beta_1 x_1+...+\beta_k x_k+\mu} = e^{\beta_0+\beta_1 x_1+...+\beta_k x_k} e^{\mu} = m(\mathbf{x})e^{\mu}$$

- Under the additional assumption that μ is independent of $(x_1, ..., x_k)$, we have

$$E[y|\mathbf{x}] = e^{\beta_0+\beta_1 x_1+...+\beta_k x_k} E[e^{\mu}|\mathbf{x}] = e^{\beta_0+\beta_1 x_1+...+\beta_k x_k} E[e^{\mu}] = m(\mathbf{x})\alpha_0$$

where the second equality is due to the independence between μ and $\mathbf{x}$, and $\alpha_0 = E[e^{\mu}]$.

- Hence, the predicted y is

$$\widehat{y} = \widehat{m(\mathbf{x})}\widehat{\alpha}_0 = (e^{\widehat{\beta}_0+\widehat{\beta}_1 x_1+...+\widehat{\beta}_k x_k})(\frac{1}{n}\sum_{i=1}^{n} e^{\widehat{\mu}_i})$$

MLR, Further Issue: Predicting y When $log(y)$ is the Dependent Variable Continue

- Recall that $E(\mu) = 0$, therefore, by Jensens Inequality

$$E[e^{\mu}] \geq e^{E(\mu)} = e^0 = 1$$

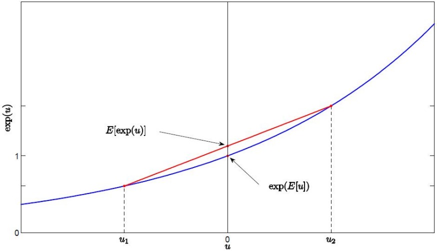

Figure: Illustration of Jensens Inequality

- As a result, $\tilde{y} = \widehat{m(\mathbf{x})} = e^{log(\hat{y})}$ under estimates $E[y|\mathbf{x}]$.

MLR, Further Issue: Predicting y When $log(y)$ is the Dependent Variable Continue

- Hence, we can summarize the following steps to predict y when the dependent variable is $log(y)$
  - 1. Obtain the fitted values, $\widehat{log(y)}$, and residuals, $\widehat{\mu}_i$, from the regression $log(y)$ on $x_1, ..., x_k$
  - 2. Obtain $\widehat{\alpha}_0 = \frac{1}{n}\sum_{i=1}^{n} e^{\widehat{\mu}_i}$
  - 3. Calculate $\widehat{y} = \widehat{\alpha}_0 \widehat{log(y)}$

MLR, Further Issue: Comparing R-Squared of a Logged and an Unlogged Specification

- Recall the CEO salary problem,

$$\widehat{salary} = \underset{(65.23)}{613.43} + \underset{(0.01)}{0.019}sales + \underset{(0.095)}{0.0234}mktval + \underset{(5.61)}{12.7}ceoten$$

where $n = 177$ and $R^2 = 0.201$

- And

$$\widehat{log(salary)} = \underset{(0.257)}{4.504} + \underset{(0.039)}{0.163}\log(sales) + \underset{(0.05)}{0.0109}mktval + \underset{(0.0053)}{0.0117}ceoten$$

where $n = 177$ and $\tilde{R}^2 = 0.318$

- R^2 and $\tilde{R}^2$ are the R-squareds for the predictions of the unlogged salary variable (although the second regression is originally for logged salaries). Both R-squareds can now be directly compared

MLR, Further Issue: Comparing R-Squared of a Logged and an Unlogged Specification, $\tilde{R}^2$

- Recall that

$$R^2 = \widehat{Corr}(y, \widehat{y})^2,$$

  where $\widehat{y}$ is the predicted value of y.

- When $log(salary)$ is the dependent variable, the predicted value of y is $\widehat{m(\mathbf{x})}\widehat{\alpha}_0 = \widehat{\alpha}_0 \tilde{y}$.

- Since $\widehat{\alpha}_0 > 0$,

$$\widehat{Corr}(y, \widehat{y}) = \widehat{Corr}(y, \widehat{\alpha}_0 \tilde{y}) = \widehat{Corr}(y, \widehat{y})$$

  which is invariant to $\widehat{\alpha}_0$. Why? For any $a > 0$,

$$Corr(X, aY) = \frac{Cov(X, aY)}{\sqrt{Var(X)Var(aY)}} = \frac{aCov(X, Y)}{\sqrt{a^2 Var(X)Var(Y)}}$$

$$= \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = Corr(X, Y)$$

- Hence, $\tilde{R}^2 = \widehat{Corr}(y, \tilde{y})^2$

MLR, Further Issue: Quantitative and Qualitative Information

- Quantitative Variables: hourly wage, years of education, college GPA, amount of air pollution, firm sales, number of arrests, etc., where the magnitude of variable conveys useful information.
- Qualitative Variable: gender, race, industry (manufacturing, retail, finance, etc.), region (South, North, West, etc.), rating grade (A, B, C, D, F, etc), etc.
- A way to incorporate qualitative information is to use dummy variables.
- A dummy variable is also called a binary variable or a zero-one variable.
- Dummy variables may appear as the dependent or as independent variables. In the latter discussion, we consider only **independent** dummy variables.

MLR, Further Issue: A Single Dummy Independent Variable- An Example

- Let's consider following population regression model

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \mu$$

where

$$\text{female} = \begin{cases} 1, \text{if the person is a woman,} \\ 0, \text{if the person is a man,} \end{cases} \quad \text{is a dummy variable.}$$

- Intuitively, δ_0 measures the wage gain/loss if the person is a woman rather than a man (holding other things fixed).
- Alternative interpretation of δ_0

$$\begin{aligned} \delta_0 = E[wage|female = 1, educ] - E[wage|female = 0, educ] \\ = \beta_0 + \delta_0 female + \beta_1 educ - (\beta_0 + \beta_1 educ) \end{aligned}$$

which gives the difference in mean wage between men and women with the same level of education

MLR, Further Issue: A Single Dummy Independent Variable- An Example Continue

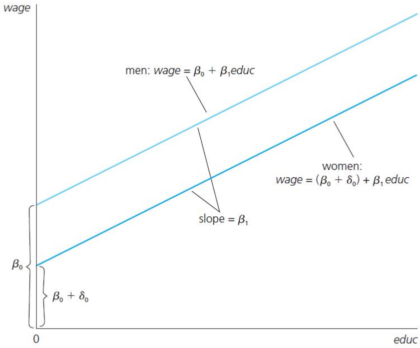

Figure: Graph of wage $= \beta_0 + \delta_0 female + \beta_1 educ$ for $\delta_0 < 0$

- Note that the mean wage difference is the same at all levels of education, i.e., the mean wage equations for men and women are parallel.

MLR, Further Issue: Dummy Variable Trap

- To investigate previous problem, one may consider to propose a population regression model as follows

$$wage = \beta_0 + \gamma_0 male + \delta_0 female + \beta_1 educ + \mu$$

However, this model **cannot** be estimated due to perfect collinearity.

- Why? There is an exact relationship among the independent variables: $1 = male + female$.
- As a result, when using dummy variables, one category always has to be omitted. Take the previous example as an example

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \mu$$

where men is the base group or benchmark group, i.e., the group with the dummy equal to zero/used for comparison, or

$$wage = \beta_0 + \gamma_0 male + \beta_1 educ + \mu$$

where women is the base group (or category).

- One would like to investigate the effect of gender on wage. Hence, after estimation, we have the following fitted regression line

$$\begin{aligned}\widehat{wage} = \; & \underset{(0.72)}{-1.57} \quad \underset{(0.26)}{-1.81 female} \quad \underset{(0.049)}{+0.572 educ} \\ & \underset{(0.012)}{+0.025 exper} \quad \underset{(0.021)}{+0.141 tenure}\end{aligned}$$

where $n = 526$, $R^2 = 0.364$

- Holding education, experience, and tenure fixed, women earn $\widehat{\delta}_0 = \$1.81$ less per hour than men.
- Does that mean that women are discriminated against?
- May be not necessarily. Being female may be correlated with other productivity characteristics (e.g., baby birth) that have not been controlled for.