Ch07. Multiple Regression Analysis with Qualitative Information: Binary (or Dummy) Variables

Ping Yu

HKU Business School
The University of Hong Kong

Describing Qualitative Information

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.
- We consider only **independent dummy** variables.

A Single Dummy Independent Variable

Example: A Simple Wage Equation

- Suppose
$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u,$$
where
$$female = \begin{cases} 1, & \text{if the person is a woman,} \\ 0, & \text{if the person is a man,} \end{cases} \text{ is a dummy variable.}$$
- δ_0 is the wage gain/loss if the person is a woman rather than a man (holding other things fixed).
- Alternative interpretation of δ_0:
$$\begin{aligned} \delta_0 &= E\left[wage|female = 1, educ\right] - E\left[wage|female = 0, educ\right] \\ &= \beta_0 + \delta_0 + \beta_1 educ - \left(\beta_0 + \beta_1 educ\right), \end{aligned}$$
i.e. the difference in mean wage between men and women with the same level of education. [figure here]
- 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.

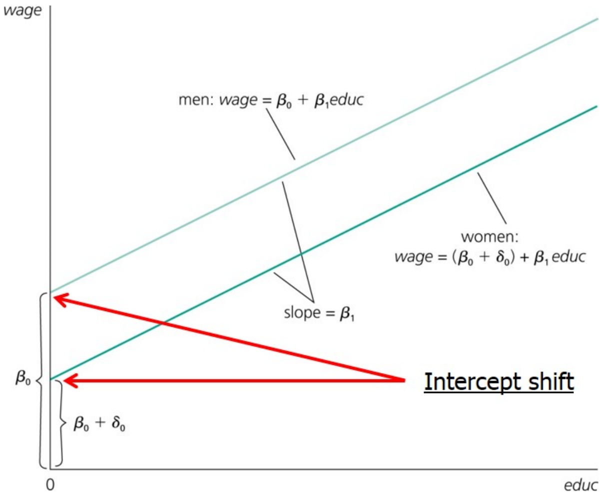


Figure: Graph of $wage = \beta_0 + \delta_0 female + \beta_1 educ$ for $\delta_0 < 0$

Dummy Variable Trap

- The model
$$wage = \beta_0 + \gamma_0 male + \delta_0 female + \beta_1 educ + u$$
cannot be estimated due to perfect collinearity.
- Why? There is an **exact** relationship among the independent variables: $1 = male + female$.
- When using dummy variables, one category always has to be omitted:
$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u,$$
where men is the base group or benchmark group, i.e., the group with the dummy equal to zero/used for comparison, or
$$wage = \beta_0 + \gamma_0 male + \beta_1 educ + u,$$
where women is the base group (or category).
- Alternatively, one could omit the intercept,
$$wage = \gamma_0 male + \delta_0 female + \beta_1 educ + u.$$

Disadvantage Without Intercept

- More difficult to test for differences between the parameters:

$$H_0 : \gamma_0 = \delta_0,$$

  and the t statistic is

$$t = \frac{\widehat{\gamma} - \widehat{\delta}}{se\left(\widehat{\gamma} - \widehat{\delta}\right)},$$

  but $se\left(\widehat{\gamma} - \widehat{\delta}\right)$ is not available from the output of standard econometric softwares (as discussed in chapter 4).

- The R-squared formula is valid only if the regression contains an intercept: Recall that

$$R^2 = 1 - \frac{SSR}{SST} \text{ with } SST = \sum_{i=1}^{n} (y_i - \overline{y})^2,$$

  where $\overline{y}$ is $\widehat{\beta}_0$ in the regression

$$y = \beta_0 + u,$$

  so SST can be treated as the SSR for the restricted regression of $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$ with $\beta_1 = \cdots = \beta_k = 0$.

Example: Hourly Wage Equation with Intercept Shift

- The fitted wage equation is

$$\begin{aligned}\widehat{wage} &= \underset{(.72)}{-1.57} \underset{(.26)}{-1.81} female + \underset{(.049)}{.572} educ + \underset{(.012)}{.025} exper + \underset{(.021)}{.141} tenure \\ n &= 526, R^2 = .364\end{aligned}$$

- 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?
- Not necessarily. Being female may be correlated with other productivity characteristics (e.g., baby birth) that have not been controlled for.

continue

- Let's compare means of subpopulations described by dummies:

$$\begin{aligned}\widehat{wage} &= 7.10 - 2.51 female \\ & \quad (.21) \ (.30) \\ n &= 526, R^2 = .116 \, (< .364 \text{ as expected})\end{aligned}$$

- Not holding other factors constant, women earn \$2.51 per hour less than men, i.e. the difference between the mean wage of men and that of women is \$2.51.
- Discussion:
  - It can easily be tested whether difference in means is significant, $|t| = \left|\frac{-2.51}{.30}\right| = |-8.37| > 1.96$.
  - The wage difference between men and women is larger if no other things are controlled for; i.e. part of the difference is due to differences in education, experience and tenure between men and women.
  - When more factors (such as baby birth) are controlled for, then we expect $\left|\widehat{\delta}_0\right|$ would be even smaller (until insignificance?).

Example: Effects of Training Grants on Hours of Training

- The fitted regression line is

$$\begin{aligned}
\widehat{hrsemp} &= 46.67 + 26.25 grant - .98 \log(sales) - 6.07 \log(employ) \\
& \quad (43.41)(5.59) \qquad (3.54) \qquad\qquad (3.88) \\
n &= 105, R^2 = .237
\end{aligned}$$

  where
  $hrsemp$ = hours training per employee, at the firm level
  $grant$ = dummy indicating whether firm received training grant
  $employ$ = number of employees

- This is an example of program evaluation:
  - treatment group (= grant receivers) vs. control group (= no grant).
  - $t_{grant} = 4.70 > 1.96$, but is the effect of treatment on the outcome of interest causal? The answer depends on whether $E[u|grant] = 0$. It might be that to get grants, some firms give more training to their employees.

a: Using Dummy Explanatory Variables in Equations for $\log(y)$

- Example (Housing Price Regression): The fitted regression line is

$$
\begin{aligned}
\widehat{\log(price)} &= \underset{(.65)}{-1.35} + \underset{(.038)}{.168}\log(lotsize) + \underset{(.093)}{.707}\log(sqrft) \\
&\quad + \underset{(.029)}{.027}bdrms + \underset{(.045)}{.054}colonial \\
n &= 88, R^2 = .649
\end{aligned}
$$

  where
  $colonial =$ dummy for the colonial style [figure here]
- Now,

$$\frac{\partial \log(price)}{\partial colonial} = \frac{\partial price/price}{\partial colonial} = 5.4\%,$$

- As the dummy for colonial style changes from 0 to 1, the house price increases by 5.4 percentage points.

American Colonial Architecture

- American colonial architecture includes several building design styles associated with the colonial period of the United States, including First Period English (late-medieval), French Colonial, Spanish Colonial, Dutch Colonial and Georgian. These styles are associated with the houses, churches and government buildings of the period from about 1600 through the 19th century.

  - From Wiki

Figure: Corwin House, Salem, Massachusetts, built about 1660, First Period English

Using Dummy Variables for Multiple Categories

Using Dummy Variables for Multiple Categories

- 1) Define membership in each category by a dummy variable;
- 2) Leave out one category (which becomes the base category).
- Example (Log Hourly Wage Equation): The fitted regression line is

$$
\begin{aligned}
\widehat{\log(wage)} &= \underset{(.100)}{-.321} + \underset{(.055)}{.213}\,marrmale - \underset{(.0058)}{.198}\,marrfem + \\
&\quad \underset{(.056)}{-.110}\,singfem + \underset{(.007)}{.079}\,educ + \underset{(.005)}{.027}\,exper - \underset{(.00011)}{.00054}\,exper^2 \\
&\quad \underset{(.007)}{+.029}\,tenure - \underset{(.00023)}{.00053}\,tenure^2 \\
n &= 526, R^2 = .461
\end{aligned}
$$

- Holding other things fixed, married women earn 19.8% less than single men (= the base category); similarly, married men earn 21.3% more and single women earn 11.0% ($<$ 19.8%) less than single men. [economic intuition here]

a: Incorporating Ordinal Information by Using Dummy Variables

- Example (City Credit Ratings and Municipal Bond Interest Rates): We can consider two specifications of the regression line.
- The first specification is

$$MBR = \beta_0 + \beta_1 CR + other\ \ factors,$$

where
$MBR =$ municipal bond interest rate
$CR =$ credit rating from $0-4$ ($0 =$ worst, $4 =$ best)

- This specification would probably not be appropriate as the credit rating only contains ordinal information.
- A better way to incorporate this information is to define dummies:

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + other\ \ factors,$$

where $CR_1, \cdots, CR_4$ are dummies indicating whether the particular rating applies, e.g., $CR1 = 1$ if $CR = 1$ and $CR1 = 0$ otherwise.

- All effects are measured in comparison to the worst rating (= base category).

Difference Between These Two Specifications

- Specification 1:

$$\begin{aligned} CR &= 0 \Longrightarrow MBR = \beta_0, \\ CR &= 1 \Longrightarrow MBR = \beta_0 + \beta_1, \\ CR &= 2 \Longrightarrow MBR = \beta_0 + 2\beta_1, \\ CR &= 3 \Longrightarrow MBR = \beta_0 + 3\beta_1, \\ CR &= 4 \Longrightarrow MBR = \beta_0 + 4\beta_1, \end{aligned}$$

  where the increase in *MBR* for each rating improvement is the same - β_1.

- Specification 2:

$$\begin{aligned} CR &= 0 \Longrightarrow MBR = \beta_0, \\ CR &= 1 \Longrightarrow MBR = \beta_0 + \delta_1, \\ CR &= 2 \Longrightarrow MBR = \beta_0 + \delta_2, \\ CR &= 3 \Longrightarrow MBR = \beta_0 + \delta_3, \\ CR &= 4 \Longrightarrow MBR = \beta_0 + \delta_4, \end{aligned}$$

  where the increase in *MBR* for each rating improvement can be different due to the arbitrariness of $\delta_1, \cdots, \delta_4$.

Interactions Involving Dummy Variables

a: Interactions among Dummy Variables

- Reconsider the female and marital status effect on log(wage) by adding the *female* $\cdot$ *married* interaction term:

$$\begin{aligned}\log\widehat{(wage)} &= \underset{(.100)}{-.321} - \underset{(.056)}{.110} female + \underset{(.055)}{.213} married - \underset{(.072)}{.301} female \cdot married + \cdots\end{aligned}$$

- These two specifications are equivalent: four categories are generated. [2 $\times$ 2 table in the next slide]
- *marrmale*: setting *married* $=1$ and *female* $=0$, we get $\widehat{\delta}_2 = .213$ as before.
- *marrfem*: setting *married* $=1$ and *female* $=1$, we get $\widehat{\delta}_1 + \widehat{\delta}_2 + \widehat{\delta}_3 = -.110 + .213 - .301 = -.198$ as before.
- *singfem*: setting *married* $=0$ and *female* $=1$, we get $\widehat{\delta}_1 = -.110$ as before.

(*) What is the Meaning of the Coefficient of $female \cdot married$, $\widehat{\delta}_3$?

- Four Categories:

	Female	Male
Single	δ_1	0
Married	$\delta_1 + \delta_2 + \delta_3$	δ_2

- **DID**:

$$
\begin{aligned}
& \left(E\left[\log(wage)\,|female = 1, married = 1\right] - E\left[\log(wage)\,|female = 1, married = 0\right]\right) \\
& - \left(E\left[\log(wage)\,|female = 0, married = 1\right] - E\left[\log(wage)\,|female = 0, married = 0\right]\right) \\
= & \left[(\delta_1 + \delta_2 + \delta_3) - \delta_1\right] - \left[\delta_2 - 0\right] \\
= & \text{ difference (in gender) in difference (in marriage)} \\
= & \left[(\delta_1 + \delta_2 + \delta_3) - \delta_2\right] - \left[\delta_1 - 0\right] \\
= & \text{ difference (in marriage) in difference (in gender)}
\end{aligned}
$$

- Analog:

$$\frac{\partial^2 y}{\partial x_1 \partial x_2}$$

b: Allowing for Different Slopes

- Consider the model

$$\log(wage) = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female \cdot educ + u.$$

where

$$\begin{aligned} \beta_0 &= \text{intercept of men, } \beta_1 = \text{slope of men,} \\ \beta_0 + \delta_0 &= \text{intercept of women, } \beta_1 + \delta_1 = \text{slope of women.} \end{aligned}$$

- Interacting both the intercept and the slope with the female dummy enables one to model completely independent wage equations for men and women. [figure here]
- Interested Hypotheses:

$$H_0 : \delta_1 = 0,$$

i.e., the return to education is the same for men and women, and

$$H_0 : \delta_0 = \delta_1 = 0,$$

i.e., the whole wage equation is the same for men and women.

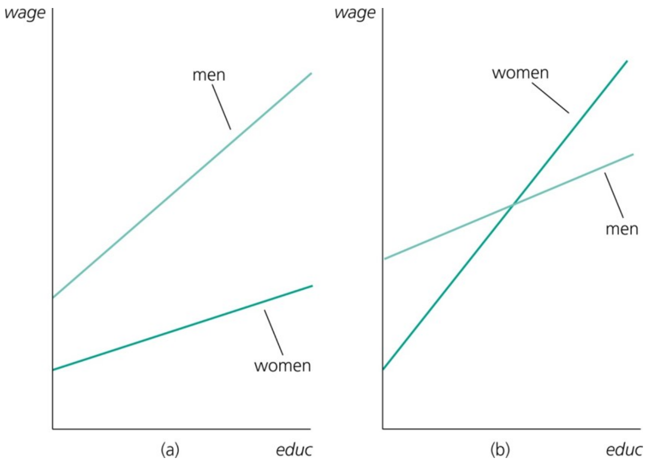


Figure: (a) $\delta_0 < 0, \delta_1 < 0$; (b) $\delta_0 < 0, \delta_1 > 0$

Example: Log Hourly Wage Equation

- The fitted regression line is

$$
\begin{aligned}
\widehat{\log(wage)} &= .389 - .227 female + .082 educ \\
&\quad\; (.119)\;(.168) \qquad\quad (.008) \\
&\quad - .0056 female \cdot educ + .029 exper - .00058 exper^2 \\
&\quad\; (.0131) \qquad\qquad\qquad (.005) \qquad\quad (.00011) \\
&\quad + .032 tenure - .00059 tenure^2 \\
&\quad\; (.007) \qquad\quad (.00024) \\
n &= 526, R^2 = .441
\end{aligned}
$$

- $|t_{female \cdot educ}| = \left|\frac{-.0056}{.0131}\right| = |-.43| < 1.96$: No evidence against hypothesis that the return to education is the same for men and women.
- $|t_{female}| = \left|\frac{-.227}{.168}\right| = |-1.35| < 1.96$: Does this mean that there is no significant evidence of lower pay for women at the same levels of *educ*, *exper*, and *tenure*? No: this is only the effect for $educ = 0$ since

$$\frac{\partial \log(wage)}{\partial female} = -.227 - .0056 educ.$$

- To answer the question one has to recenter the interaction term, e.g., around $educ = 12.5$ (= average education) to have $female \cdot (educ - 12.5)$:
$\frac{\partial \log(wage)}{\partial female} = \widehat{\delta}_0 + \widehat{\delta}_1 (educ - 12.5)$ with new $\widehat{\delta}_0 = -.297$

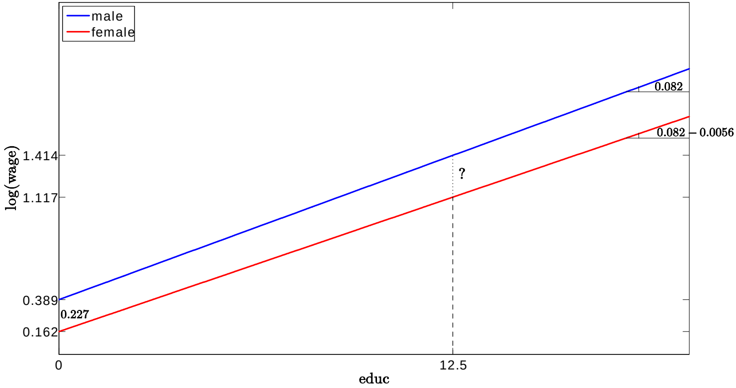


Figure: The New new $\widehat{\delta}_0 = -.227 + 12.5 \times (-.0056) = -.297 < -.227$

c: Testing for Differences in Regression Functions across Groups

- This is a **special** F test with the unrestricted model containing full set of interactions,

$$\begin{aligned} cumgpa &= \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female \cdot sat + \beta_2 hsperc \\ &\quad + \delta_2 female \cdot hsperc + \beta_3 tothrs + \delta_3 female \cdot tothrs + u \end{aligned}$$

  and the restricted model with same regression for both groups,

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u,$$

  where
  $cumpga =$ college GPA
  $sat =$ standardized aptitude test score
  $hsperc =$ high school rank percentile
  $tothrs =$ total hours spent in college courses

- The null hypothesis is

$$H_0 : \delta_0 = \delta_1 = \delta_2 = \delta_3 = 0.$$

- All interaction effects are zero, i.e., the same regression coefficients apply to both men and women.

Estimation of the Unrestricted Model

- The estimated unrestricted model is

$$
\begin{aligned}
\widehat{cumgpa} &= \underset{(.21)}{1.48} \underset{(.411)}{-.353} female + \underset{(.0002)}{.0011} sat + \underset{(.00039)}{.00075} female \cdot sat \\
&\quad \underset{(.0014)}{-.0085} hsperc \underset{(.00316)}{-.00055} female \cdot hsperc + \underset{(.0009)}{.0023} tothrs \\
&\quad \underset{(.00163)}{-.00012} female \cdot tothrs \\
n &= 366, R^2 = .406, \overline{R}^2 = .394
\end{aligned}
$$

- It can be shown that [proof not required]

$$SSR_{ur} = SSR_{male} + SSR_{female},$$

where SSR_{male} is the SSR in the regression

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u,$$

using only the data of male, and SSR_{female} is the SSR using only the data of female.

Testing Results

- Tested individually, the hypothesis that the interaction effects are zero cannot be rejected.
- Tested jointly, the F statistic is

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(85.515 - 78.355)/4}{78.355/(366-7-1)} \approx 8.18,$$

  and the null is rejected.
- $SSR_{ur} = SSR_{male} + SSR_{female} = 58.752 + 19.603 = 78.355$, $n_{male} = 276$, $n_{female} = 90$ and $n = 366$.
- This relationship is true only if **all** interaction terms are included in the unrestricted model.

Chow Test

- In general, if the test is computed in this way, i.e.,

$$F=\frac{\left[SSR_P-(SSR_1+SSR_2)\right]/(k+1)}{(SSR_1+SSR_2)/\left[n-2(k+1)\right]},$$

  it is called the Chow-Test. [photo here]
  - SSR_P is the SSR for the pooled (restricted) regression; [see below]
  - SSR_1 and SSR_2 are the SSRs for the two separate regressions; [see below]
  - The number of restrictions is $k+1$ with k being the number of nonconstant regressors in the **pooled** regression; [see below]
  - The total number of parameters in the unrestricted model is $2(k+1)$. [see below]
- SSR_P: $y_i=\beta_0+\beta_1x_{i1}+\cdots+\beta_kx_{ik}+u_i$, $i=1,\cdots,n$.
- SSR_{ur}: $y_i=\beta_0+\beta_1x_{i1}+\cdots+\beta_kx_{ik}+\delta_0D_i+\delta_1D_ix_{i1}+\cdots+\delta_kD_ix_{ik}+u_i$, $i=1,\cdots,n$.
- SSR_1: $y_i=\beta_0+\beta_1x_{i1}+\cdots+\beta_kx_{ik}+u_i$, $i=1,\cdots,n_1$.
- SSR_2: $y_i=\beta_0+\beta_1x_{i1}+\cdots+\beta_kx_{ik}+u_i$, $i=n_1+1,\cdots,n$.
- H_0: $\delta_0=\delta_1=\cdots=\delta_k=0$.

Gregory C. Chow (1929-),
Princeton, 1955ChicagoPhD

- Chow, G.C., 1960, Tests of Equality Between Sets of Coefficients in Two Linear Regressions, *Econometrica*, 28, 591-605.
- Caution: Chow-Test assumes a constant error variance across groups as assumed in the F test.
  - $E[u_i|D_i] = \sigma^2$ for $D_i = 0$ and 1.

Applications of the Chow-Test

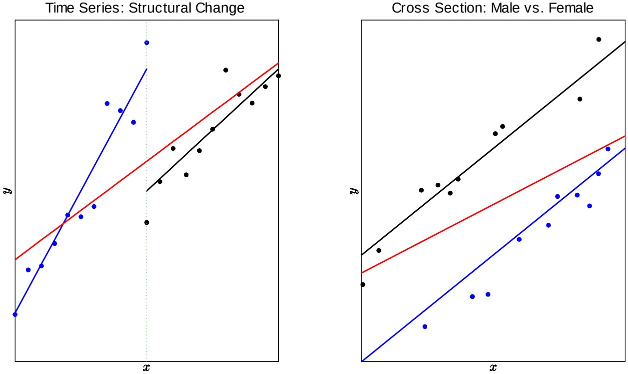


Figure: Restricted and Unrestricted Models in the Chow Test

- You must pass the Chow test to pass this course!

(**) A Binary Dependent Variable: The Linear Probability Model

Linear Regression When the Dependent Variable is Binary

- Suppose
$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u, \tag{1}$$
with $E[u|\mathbf{x}] = 0$; then
$$E[y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$
- If the dependent variable y only takes on the values 1 and 0, then
$$E[y|\mathbf{x}] = 1 \cdot P(y = 1|\mathbf{x}) + 0 \cdot P(y = 0|\mathbf{x}) = P(y = 1|\mathbf{x}),$$
i.e.,
$$P(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$
where $P(y = 1|\mathbf{x})$ is called the response probability.
- Since $P(y = 1|\mathbf{x})$ is linear in $\mathbf{x}$, the model (1) is called the linear probability model (LPM).
- In the LPM,
$$\beta_j = \partial P(y = 1|\mathbf{x}) / \partial x_j$$
describes the effect of the explanatory variable x_j on the probability that $y = 1$.

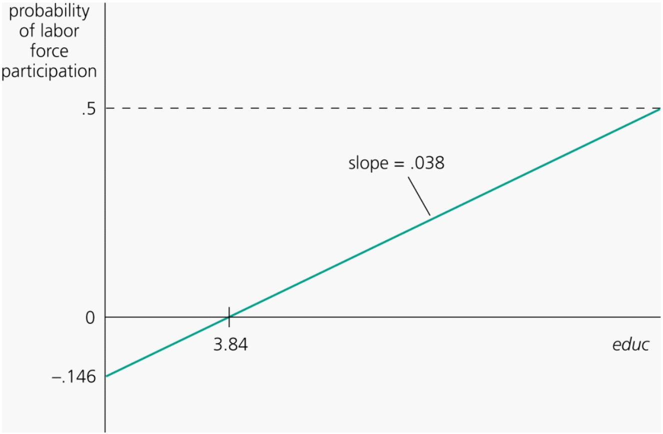
Example: Labor Force Participation of Married Women

- The fitted regression line is

$$\begin{aligned} \widehat{infl} &= .596 - .0034 nwifeinc + .038 educ + .039 exper \\ &\quad (.154)(.0014) \qquad\qquad (.007) \qquad\quad (.006) \\ &\quad - .00060 exper^2 - .016 age - .262 kidslt6 + .0130 kidsge6 \\ &\quad (.00018) \qquad\qquad (.002) \qquad (.0034) \qquad\qquad (.0132) \\ n &= 753, R^2 = .264 \end{aligned}$$

  where
  $infl$ = dummy for "in the labor force" of a married women
  $nwifeinc$ = husband's earnings (in thousands of dollars)
  $kidslt6$ = number of children less than six years old
  $kidsge6$ = number of kids between 6 and 18 years of age
- All variables except kidsge6 are statistically significant, and all of the significant variables have correct signs.
- If the number of kids under six years increases by one, the probability that the woman works falls by 26.2%.

Graph for nwifeinc=50, exper=5, age=30, kindslt6=1, kidsge6=0



- The maximum level of education in the sample is $educ = 17$. For the given case, this leads to a predicted probability to be in the labor force of about 50%.
- Negative predicted probability but no problem because no woman in the sample has $educ < 5$.

Disadvantages and Advantages of the LPM

Disadvantages:

- Predicted probabilities may be larger than one or smaller than zero.
- Marginal probability effects sometimes logically impossible, e.g., the first small child would reduce the probability by a large amount, but subsequent children would have a smaller marginal effect; going from zero to four young children reduces the probability of working by $\Delta\widehat{infl} = .262 \times 4 = 1.048$!
- The LPM is necessarily heteroskedastic:

$$Var\left(y|\mathbf{x}\right) = P\left(y = 1|\mathbf{x}\right)\left(1 - P\left(y = 1|\mathbf{x}\right)\right)$$

by the variance formula of a Bernoulli random variable.
- Heterosceasticity consistent standard errors need to be computed.

Advantages:

- Easy estimation and interpretation.
- Estimated effects and predictions often reasonably good in practice.