Introductory Econometrics I

Multiple Regression: Qualitative Information

Yingjie Feng

School of Economics and Management

Tsinghua University

April 27, 2024

Outline

Describing Qualitative Information

- Describe *binary* qualitative information
  - male or female
  - urban or rural
  - vaccinated or not vaccinated
- It can be captured by defining a **binary variable** (or **dummy variable**, **zero-one** variable)
  - We must decide which outcome is assigned zero, which is one. Choose the variable name to be descriptive.
  - For example, to indicate gender, a variable *female*$= 1$ if the person is female; $female = 0$ if the person is male (better name than *gender* or *sex*)

Describing Qualitative Information

```
list wage educ female married in 1/5
```

	wage	educ	female	married
1.	3.1	11	1	0
2.	3.2	12	1	1
3.	3	11	0	0
4.	6	8	0	1
5.	5.3	12	0	1

- Any two different values would distinguish different types. But "0-1" are convenient for use in regression analysis.
- Define more than two categories with two pieces of qualitative information (say, gender $female$ and marital status $married$)
  - married male ($marrmale$), married female ($marrfem$), single male ($singmale$), and single female ($singfem$)

Outline

A Single Dummy Variable

- How do we interpret a simple regression model with a binary explanatory variable?

$$wage = \beta_0 + \delta_0 female + u$$

- We assume SLR.4:

$$\begin{aligned} & \mathbb{E}[u|female] = 0, \\ \Leftrightarrow \quad & \mathbb{E}[wage|female] = \beta_0 + \delta_0 female \end{aligned}$$

- There are only two values of $female$, 0 and 1. So

$$\mathbb{E}\left[wage|female = 0\right] = \beta_0 + \delta_0 \cdot 0 = \beta_0$$

$$\mathbb{E}\left[wage|female = 1\right] = \beta_0 + \delta_0 \cdot 1 = \beta_0 + \delta_0$$

- The average $wage$ for men is β_0; the average $wage$ for women is $\beta_0 + \delta_0$.

A Single Dummy Variable

- The difference in average *wage* between women and men in **population**:

$$\delta_0 = \mathbb{E}[wage|female = 1] - \mathbb{E}[wage|female = 0]$$

- δ_0 is not really a slope
  - It is just a difference in average outcomes between the two groups
- The population relation is mimicked by the regression estimates.

$$\hat{\beta}_0 = \overline{wage}_m, \qquad \hat{\beta}_0 + \hat{\delta}_0 = \overline{wage}_f, \qquad \hat{\delta}_0 = \overline{wage}_f - \overline{wage}_m$$

  - $\overline{wage}_m$ is the average wage for men in the sample
  - $\overline{wage}_f$ is the average wage for women in the sample
- $\hat{\delta}_0$: the difference in average wage between women and men in the **sample**

A Single Dummy Variable

sum wage female exper

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	526	5.896103	3.693086	.53	24.98
female	526	.4790875	.500038	0	1
exper	526	17.01711	13.57216	1	51

tab female

=1 if female	Freq.	Percent	Cum.
0	274	52.09	52.09
1	252	47.91	100.00
Total	526	100.00	

- The mean (average) of a binary variable is the fraction of ones in the sample
- The fraction of females is 47.91%.

A Single Dummy Variable

```
. reg wage female

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(1, 524)       =     68.54
       Model |  828.220467         1  828.220467   Prob > F        =    0.0000
    Residual |  6332.19382       524  12.0843394   R-squared       =    0.1157
-------------+----------------------------------   Adj R-squared   =    0.1140
       Total |  7160.41429       525  13.6388844   Root MSE        =    3.4763

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -2.51183   .3034092    -8.28   0.000    -3.107878   -1.915782
       _cons |   7.099489   .2100082    33.81   0.000     6.686928     7.51205
------------------------------------------------------------------------------

. tabstat wage, by(female)

Summary for variables: wage
     by categories of: female (=1 if female)

  female |      mean
---------+----------
       0 |  7.099489
       1 |  4.587659
---------+----------
   Total |  5.896103
--------------------

. di 4.588-7.099
-2.511
```

A Single Dummy Variable

- The estimated difference is large. Women earn about \$2.51 less than men per hour, on average ("difference in averages").
- This simple regression allows us to do a **comparison of means test**.
  - The null is

$$H_0 : \mu_f = \mu_m$$

    μ_f: **population** average *wage* for women;

    μ_m: **population** average *wage* for men.
  - The t statistic and confidence interval are directly reported.
- t test leads to a very strong rejection of H_0:

$$t_{female} = -8.28, \qquad p\text{-val} = 0.000$$

- Remember: NO other factors are controlled for! (such as workforce experience and schooling)

A Single Dummy Variable

- If we control for experience, the model written in expected value form is

$$\mathbb{E}[wage|female, exper] = \beta_0 + \delta_0 female + \beta_1 exper$$

  where δ_0 measures the gender difference when we hold fixed *exper*.

- Another way to write δ_0:

$$\delta_0 = \mathbb{E}[wage|female, exper_0] - \mathbb{E}[wage|male, exper_0]$$

  $exper_0$ is any level of experience that is the same for the woman and man.

- The model imposes a **common slope** on *exper* for men and women, β_1. It is only the intercepts that are allowed to differ.

A Single Dummy Variable

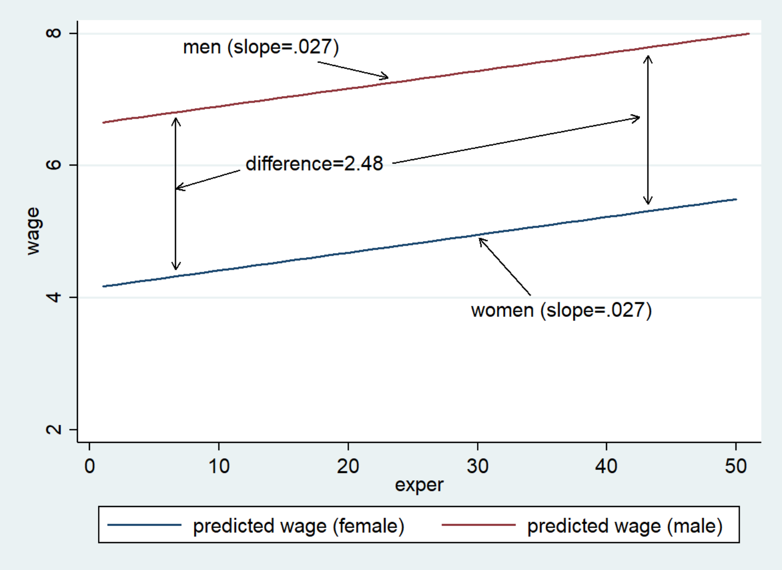
```
reg wage female exper

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(2, 523)       =     37.51
       Model |  898.161983         2  449.080991   Prob > F        =    0.0000
    Residual |  6262.25231       523  11.9737138   R-squared       =    0.1254
-------------+----------------------------------   Adj R-squared   =    0.1221
       Total |  7160.41429       525  13.6388844   Root MSE        =    3.4603

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -2.48142   .3022793    -8.21   0.000     -3.07525   -1.887589
       exper |   .0269163   .0111369     2.42   0.016     .0050379    .0487948
       _cons |   6.626882   .2862475    23.15   0.000     6.064546    7.189218
------------------------------------------------------------------------------
```

- There is still a difference of about \$2.48, slightly smaller than when *exper* is not controlled for.

A Single Dummy Variable



A Single Dummy Variable

- The estimated difference in average wages is the same at all levels of experience: \$2.48.
- Easy to add other variables (but picture is harder to draw). For example, adding years of education shrinks the gap to about \$2.15.

```
reg wage female exper educ

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(3, 522)       =     77.92
       Model |  2214.74206         3  738.247353   Prob > F        =    0.0000
    Residual |  4945.67223       522  9.47446788   R-squared       =    0.3093
-------------+----------------------------------   Adj R-squared   =    0.3053
       Total |  7160.41429       525  13.6388844   Root MSE        =    3.0781

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.155517   .2703055    -7.97   0.000    -2.686537   -1.624497
       exper |   .0642417   .0104003     6.18   0.000     .0438101    .0846734
        educ |   .6025802   .0511174    11.79   0.000     .5021591    .7030012
       _cons |  -1.734481   .7536203    -2.30   0.022    -3.214982   -.2539797
------------------------------------------------------------------------------
```

A Single Dummy Variable

- The previous regressions use **males** as the **base group** (or **benchmark group** or **reference group**).
  - Recall: $female = 0$ refers to males
  - The coefficient -2.15 on $female$ tells us how women do compared with men.
- We will get the same answer if we use women as the base group
  - use a dummy variable for males rather than females
  - $male = 1 - female$
  - The coefficient on $male$ will change sign but must remain the same magnitude.
- The intercept changes because now the base (or reference) group is females.

A Single Dummy Variable

```
gen male=1-female

reg wage male exper educ
```

Source	SS	df	MS
Model	2214.74206	3	738.247353
Residual	4945.67223	522	9.47446788
Total	7160.41429	525	13.6388844

Number of obs	=	526
F(3, 522)	=	77.92
Prob > F	=	0.0000
R-squared	=	0.3093
Adj R-squared	=	0.3053
Root MSE	=	3.0781

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
male	2.155517	.2703055	7.97	0.000	1.624497	2.686537
exper	.0642417	.0104003	6.18	0.000	.0438101	.0846734
educ	.6025802	.0511174	11.79	0.000	.5021591	.7030012
_cons	-3.889998	.7271441	-5.35	0.000	-5.318486	-2.46151

- We get what we had before
  - The intercept for **men** is $-3.890 + 2.156 = -1.734$.
  - The intercept for **women** is -3.890.

Dummy Variable Trap

- Putting $female$ and $male$ both in the equation is redundant.
  - Perfect colinearity: $female + male = 1$
  - We have two groups so need only two intercepts.
  - Stata drops one of the dummies (in this case, the second one listed in the `reg` command).
- This is the simplest example of the so-called **dummy variable trap**
  - Putting in too many dummy variables to represent the given number of groups (two in this case).
  - An intercept is estimated for the base group, we need only **one** dummy variable that distinguishes the two groups.

Dummy Variable Trap

```
. reg wage female male exper educ
note: male omitted because of collinearity

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(3, 522)       =     77.92
       Model |  2214.74206         3  738.247353   Prob > F        =    0.0000
    Residual |  4945.67223       522  9.47446788   R-squared       =    0.3093
-------------+----------------------------------   Adj R-squared   =    0.3053
       Total |  7160.41429       525  13.6388844   Root MSE        =    3.0781

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.155517   .2703055    -7.97   0.000    -2.686537   -1.624497
        male |          0  (omitted)
       exper |   .0642417   .0104003     6.18   0.000     .0438101    .0846734
        educ |   .6025802   .0511174    11.79   0.000     .5021591    .7030012
       _cons |  -1.734481   .7536203    -2.30   0.022    -3.214982   -.2539797
------------------------------------------------------------------------------
```

Dummy Variable: $\log(y)$ as the Dependent Variable

- WAGE1.DTA: $lwage = \log(wage)$ as the dependent variable:

$$\widehat{lwage} = \underset{(.030)}{1.814} - \underset{(.043)}{.397}\, female, \qquad n = 526,\ R^2 = .309$$

- A rough estimate: average wage for women is below that of men by 39.7%.

- Recall: the precise formula to calculate percentage change is $100 \cdot (\exp(\beta_j) - 1)$

- More precise estimate: $\exp(-.397) - 1 \approx -.328$, or 32.8% lower for women.

- If we use women as the base group, $\exp(.397) - 1 \approx .487$, so men earn 48.7% more than women on average.

Outline

Dummy Variables for Multiple Categories

- In the wage example we have two qualitative variables, gender ($female$) and marital status ($married$).
- Define four exhaustive and mutually exclusive groups.
  - Married males ($marrmale$), married females ($marrfem$), single males ($singmale$), and single females ($singfem$)
- We can define each of these dummy variables by interactions of $female$ and $married$:

$$\begin{aligned}
marrmale &= married \cdot (1 - female) \\
marrfem &= married \cdot female \\
singmale &= (1 - married) \cdot (1 - female) \\
singfem &= (1 - married) \cdot female
\end{aligned}$$

Dummy Variables for Multiple Categories

$$
\begin{aligned}
marrmale &= married \cdot (1 - female) \\
marrfem &= married \cdot female \\
singmale &= (1 - married) \cdot (1 - female) \\
singfem &= (1 - married) \cdot female
\end{aligned}
$$

- We can allow each of the four groups to have a different intercept by choosing a base group (omit the dummy for the base group) and then including dummies for the other three groups.
- For example, choose single males as the base group
  - Do not add $singmale$ in the regression
  - include $marrmale$, $marrfem$, and $singfem$
  - The coefficients on these dummies are difference compared with single men.

Dummy Variables for Multiple Categories

Use WAGE1.DTA

```
gen marrmale=male*married

gen marrfem=female*married

gen singfem=female*(1-married)

gen exper2=exper^2

gen tenure2=tenure^2

reg lwage marrmale marrfem singfem educ exper exper2 tenure tenure2
```

Source	SS	df	MS
Model	68.3617623	8	8.54522029
Residual	79.9679891	517	.154676961
Total	148.329751	525	.28253286

Number of obs	=	526
F(8, 517)	=	55.25
Prob > F	=	0.0000
R-squared	=	0.4609
Adj R-squared	=	0.4525
Root MSE	=	.39329

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
marrmale	.2126757	.0553572	3.84	0.000	.103923	.3214284
marrfem	-.1982676	.0578355	-3.43	0.001	-.311889	-.0846462
singfem	-.1103502	.0557421	-1.98	0.048	-.219859	-.0008414
educ	.0789103	.0066945	11.79	0.000	.0657585	.092062
exper	.0268006	.0052428	5.11	0.000	.0165007	.0371005
exper2	-.0005352	.0001104	-4.85	0.000	-.0007522	-.0003183
tenure	.0290875	.006762	4.30	0.000	.0158031	.0423719
tenure2	-.0005331	.0002312	-2.31	0.022	-.0009874	-.0000789
_cons	.3213781	.100009	3.21	0.001	.1249041	.5178521

Dummy Variables for Multiple Categories

- Control for education, experience, and tenure:

$$\begin{aligned}\widehat{lwage} \;=\; & \underset{(.100)}{.321} + \underset{(.055)}{.213}\, marrmale - \underset{(.058)}{.198}\, marrfem - \underset{(.056)}{.110}\, singfem \\ & + \underset{(.007)}{.079}\, educ + \underset{(.005)}{.027}\, exper - \underset{(.00011)}{.00054} exper^2 \\ & + \underset{(.007)}{.029}\, tenure - \underset{(.00023)}{.00053} tenure^2, \qquad n = 526,\ R^2 = .461\end{aligned}$$

- Interpretation: holding education, experience and tenure fixed,
  - A married man earns about 21.3% more than a single man on average.
  - Remember this compares two men with the same level of schooling, experience, and tenure with the current employer

Dummy Variables for Multiple Categories

- Control for education, experience, and tenure:

$$
\begin{aligned}
\widehat{lwage} = \; & \underset{(.100)}{.321} + \underset{(.055)}{.213}\, marrmale - \underset{(.058)}{.198}\, marrfem - \underset{(.056)}{.110}\, singfem \\
& + \underset{(.007)}{.079}\, educ + \underset{(.005)}{.027}\, exper - \underset{(.00011)}{.00054} exper^2 \\
& + \underset{(.007)}{.029}\, tenure - \underset{(.00023)}{.00053} tenure^2, \qquad n = 526,\; R^2 = .461
\end{aligned}
$$

- Interpretation: holding education, experience and tenure fixed,
  - *Marriage premium* for men has long been noted by labor economists.
    - Does marriage make men more productive?
    - Is being married a signal to employers (say, of stability and reliability)?
    - Is there a selection issue in that more productive men are likely to be married, on average?
  - The regression cannot tell us which explanation is correct.

Dummy Variables for Multiple Categories

- Control for education, experience, and tenure:

$$\begin{aligned}
\widehat{lwage} = \; & \underset{(.100)}{.321} + \underset{(.055)}{.213}\, marrmale - \underset{(.058)}{.198}\, marrfem - \underset{(.056)}{.110}\, singfem \\
& + \underset{(.007)}{.079}\, educ + \underset{(.005)}{.027}\, exper - \underset{(.00011)}{.00054} exper^2 \\
& + \underset{(.007)}{.029}\, tenure - \underset{(.00023)}{.00053} tenure^2, \qquad n = 526,\ R^2 = .461
\end{aligned}$$

- Interpretation: holding education, experience and tenure fixed,
  - A married woman earns about 19.8% *less* than a single man.
  - A single woman earns about 11.0% less than a comparable single man.
  - Marriage "penalty" for women?

Dummy Variables for Multiple Categories

- Control for education, experience, and tenure:

$$\begin{aligned}\widehat{lwage} \quad = \quad & \underset{(.100)}{.321} + \underset{(.055)}{.213}\, marrmale - \underset{(.058)}{.198}\, marrfem - \underset{(.056)}{.110}\, singfem \\ & + \underset{(.007)}{.079}\, educ + \underset{(.005)}{.027}\, exper - \underset{(.00011)}{.00054} exper^2 \\ & + \underset{(.007)}{.029}\, tenure - \underset{(.00023)}{.00053} tenure^2, \qquad n = 526,\ R^2 = .461\end{aligned}$$

- Interpretation: holding education, experience and tenure fixed,
  - Compare married women and single women?

$$\begin{aligned}\text{intercept for married women} \quad &= \quad .321 - .198 \\ \text{intercept for single women} \quad &= \quad .321 - .110 \\ \text{difference} \quad &= \quad -.198 - (-.110) = -.088\end{aligned}$$

  - Married women earn about 8.8% less than single women.

Dummy Variables for Multiple Categories

- Control for education, experience, and tenure:

$$\begin{aligned}\widehat{lwage} \quad = \quad & \underset{(.100)}{.321} + \underset{(.055)}{.213}\, marrmale - \underset{(.058)}{.198}\, marrfem - \underset{(.056)}{.110}\, singfem \\ & + \underset{(.007)}{.079}\, educ + \underset{(.005)}{.027}\, exper - \underset{(.00011)}{.00054} exper^2 \\ & + \underset{(.007)}{.029}\, tenure - \underset{(.00023)}{.00053} tenure^2, \qquad n = 526,\ R^2 = .461\end{aligned}$$

- Interpretation: holding education, experience and tenure fixed,
  - Compare married women and single women?
    - Is this difference statistically significant?
    - Two approaches: (1) `lincom` command in Stata. (2) Choose, say, married females as the base group and re-estimate the model (including the dummies *marrmale*, *singmale*, and *singfem*).
    - The t statistic for the estimated difference $-.088$ is -1.68, which is significant at the 10% level (but not much lower than that).

Dummy Variables for Multiple Categories

lincom marrfem-singfem

(1) **marrfem - singfem = 0**

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	-.0879174	.0523481	-1.68	0.094	-.1907586	.0149238

reg lwage singmale marrmale singfem educ exper exper2 tenure tenure2

Source	SS	df	MS
Model	68.3617623	8	8.54522029
Residual	79.9679891	517	.154676961
Total	148.329751	525	.28253286

Number of obs	=	526
F(8, 517)	=	55.25
Prob > F	=	0.0000
R-squared	=	0.4609
Adj R-squared	=	0.4525
Root MSE	=	.39329

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
singmale	.1982676	.0578355	3.43	0.001	.0846462	.311889
marrmale	.4109433	.0457709	8.98	0.000	.3210234	.5008631
singfem	.0879174	.0523481	1.68	0.094	-.0149238	.1907586
educ	.0789103	.0066945	11.79	0.000	.0657585	.092062
exper	.0268006	.0052428	5.11	0.000	.0165007	.0371005
exper2	-.0005352	.0001104	-4.85	0.000	-.0007522	-.0003183
tenure	.0290875	.006762	4.30	0.000	.0158031	.0423719
tenure2	-.0005331	.0002312	-2.31	0.022	-.0009874	-.0000789
_cons	.1231105	.1057937	1.16	0.245	-.0847279	.3309488

Dummy Variables: Incorporate Ordinal Information

- BEAUTY.DTA includes a ranking of physical attractiveness of each man or woman, on a scale of 1 to 5 (5 indicates "strikingly beautiful or handsome")
  - This is a subset of the data used in Hamermesh and Biddle (1994, *American Economic Review*).
- As we move up the scale from 1 to 5, why should a one-unit increase mean the same amount of "beauty"?
- Such variables are called **ordinal variables**:
  - Order of outcomes conveys information (5 is better than 4; 2 is better than 1)
  - We do not know that the difference between 5 and 4 is the same as 2 and 1.
  - Very few people are at the extreme values 1 and 5. It makes sense to combine into three categories: below average ($belavg$), average, and above average ($abvavg$).

Dummy Variables: Incorporate Ordinal Information

- 12.3% of people are "below average," 30.4% are "above average," and everyone else (57.3%) has $looks = 3$ (labeled "average").

```
. tab looks

from 1 to 5 |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         13        1.03        1.03
          2 |        142       11.27       12.30
          3 |        722       57.30       69.60
          4 |        364       28.89       98.49
          5 |         19        1.51      100.00
------------+-----------------------------------
      Total |      1,260      100.00

. tab abvavg

=1 if looks |
        >=4 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        877       69.60       69.60
          1 |        383       30.40      100.00
------------+-----------------------------------
      Total |      1,260      100.00

. tab belavg

=1 if looks |
       <= 2 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,105       87.70       87.70
          1 |        155       12.30      100.00
------------+-----------------------------------
      Total |      1,260      100.00
```

Dummy Variables: Incorporate Ordinal Information

```
reg lwage belavg abvavg

      Source |       SS           df       MS      Number of obs   =     1,260
-------------+----------------------------------   F(2, 1257)      =      7.98
       Model |  5.58214128         2  2.79107064   Prob > F        =    0.0004
    Residual |  439.397831     1,257  .349560725   R-squared       =    0.0125
-------------+----------------------------------   Adj R-squared   =    0.0110
       Total |  444.979972     1,259  .353439215   Root MSE        =    .59124

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      belavg |  -.2087896   .0523391    -3.99   0.000    -.3114713   -.1061078
      abvavg |  -.0454376   .0373744    -1.22   0.224    -.1187607    .0278855
       _cons |   1.698296   .0220035    77.18   0.000     1.655128    1.741463
------------------------------------------------------------------------------
```

- Base group: people with "average" looks
- Those with "below average" looks earn about 20.9% less than those with average looks. The t statistic is very significant.
- Those with above average looks are estimated to earn about 4.5% less than those with average looks, but the p-value is .266 (little evidence for a non-zero effect)

Dummy Variables: Incorporate Ordinal Information

- Now control for some other factors, including gender and education.

```
reg lwage belavg abvavg educ exper expersq female
```

Source	SS	df	MS
Model	160.094314	6	26.6823857
Residual	284.885658	1,253	.227362856
Total	444.979972	1,259	.353439215

Number of obs	=	1,260
F(6, 1253)	=	117.36
Prob > F	=	0.0000
R-squared	=	0.3598
Adj R-squared	=	0.3567
Root MSE	=	.47683

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
belavg	-.1542032	.0423296	-3.64	0.000	-.2372479	-.0711585
abvavg	-.0066465	.0306562	-0.22	0.828	-.0667896	.0534966
educ	.0663221	.0053094	12.49	0.000	.0559058	.0767384
exper	.0408305	.0044034	9.27	0.000	.0321916	.0494694
expersq	-.0006301	.0000985	-6.40	0.000	-.0008233	-.0004368
female	-.4532832	.029217	-15.51	0.000	-.5106029	-.3959636
_cons	.558981	.0795603	7.03	0.000	.4028949	.7150671

- The effect of having "below average" looks is now -15%. The effect of "above average looks" is still insignificant and gets smaller in magnitude.

Dummy Variables: Incorporate Ordinal Information

- One shortcoming in the previous analysis:
  - It ignores occupation. Maybe we should allow people to sort into occupation (perhaps partly based on looks) and see if there is a "looks premium" in a given occupation. Biddle and Hamermesh (1998, *Journal of Labor Economics*) study lawyers' looks and earnings.
- Variables such as credit ratings, or any variables asked on a scale, are ordered variables. For example,
  - A credit rating on a scale from 1 to 7
  - Rate one's "happiness" on a scale of 1 to 5

Outline

Outline

Interactions Among Dummy Variables

- Example: *lwage* equation
  - Gender and marital status define four different groups: married male, married female, single male, single female
  - We can achieve the same thing by interacting between *female* and *married*.
- Regress *lwage* on *female*, *married*, *female* · *married* (and others)

$$\widehat{lwage} = \underset{(.100)}{.321} - \underset{(.056)}{.110}\, female + \underset{(.055)}{.213}\, married - \underset{(.072)}{.301}\, female \cdot married + ...$$

- The intercept for
  - single male: $female = 0$, $married = 0$, so 0.321
  - single female: $female = 1$, $married = 0$, so $0.321 - 0.110$
  - married male: $female = 0$, $married = 1$, so $0.321 + 0.213$
  - married female: $female = 1$, $married = 1$, so $0.321 - 0.110 + 0.213 - 0.301$

Interactions Among Dummy Variables

- Example: *lwage* equation
  - Gender and marital status define four different groups: married male, married female, single male, single female
  - We can achieve the same thing by interacting between $female$ and $married$.
- Regress $lwage$ on $female$, $married$, $female \cdot married$ (and others)

$$\widehat{lwage} = \underset{(.100)}{.321} - \underset{(.056)}{.110}\, female + \underset{(.055)}{.213}\, married - \underset{(.072)}{.301}\, female \cdot married + ...$$

- One advantage with interaction: we can read off the difference in marriage premium between women and men.
  - For males: slope on $married$, $+0.213$
  - For female: slope on $married$+slope on $female \cdot married$, $0.213 - 0.301 = -0.088$
  - Slope on the interaction -0.301: gender difference in marriage premium

Outline

Allowing for Different Slopes

- Regression models with different slopes and different intercepts.
- Interact dummy variables with quantitative variables
- Recall

  $$lwage = \beta_0 + \delta_0 female + \beta_1 exper + u,$$

  Intercept for men is β_0 and that for women is $\beta_0 + \delta_0$. The slope on *exper*, β_1, is common across men and women.
- An extended model is

  $$lwage = (\beta_0 + \delta_0 female) + (\beta_1 + \delta_1 female) \cdot exper + u$$

  The slope for men is β_1 and the slope for women is $\beta_1 + \delta_1$

Allowing for Different Slopes

- An extended model is

$$lwage = (\beta_0 + \delta_0 female) + (\beta_1 + \delta_1 female) \cdot exper + u$$

- For men, $female = 0$. For women, $female = 1$. Then,

	Intercept	Slope
Male	β_0	β_1
Female	$\beta_0 + \delta_0$	$\beta_1 + \delta_1$
Difference (Female – Male)	δ_0	δ_1

- Here we use Greek letter “delta” to emphasize that δ_0 and δ_1 are differences
- Estimation: write the model as

$$lwage = \beta_0 + \delta_0 female + \beta_1 exper + \delta_1 female \cdot exper + u$$

Allowing for Different Slopes

Use WAGE1.DTA

```
gen femexper=female * exper

reg lwage female exper femexper
```

Source	SS	df	MS
Model	22.9051677	3	7.63505589
Residual	125.424584	522	.24027698
Total	148.329751	525	.28253286

Number of obs	=	526
F(3, 522)	=	31.78
Prob > F	=	0.0000
R-squared	=	0.1544
Adj R-squared	=	0.1496
Root MSE	=	.49018

lwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
female	-.2934319	.0685958	-4.28	0.000	-.4281897	-.158674
exper	.0066007	.0021976	3.00	0.003	.0022835	.0109179
femexper	-.0058634	.0031567	-1.86	0.064	-.0120649	.000338
_cons	1.697672	.0486394	34.90	0.000	1.602119	1.793225

Allowing for Different Slopes

- Use WAGE1.DTA:

$$\begin{aligned}\widehat{lwage} &= \underset{(.049)}{1.698} - \underset{(.069)}{.293}\, female + \underset{(.002)}{.007}\, exper - \underset{(.003)}{.006}\, female \cdot exper \\ n &= 526,\ R^2 = .154\end{aligned}$$

  - The intercept for men is 1.698 and the slope is .007 – about 0.7% for each year of experience.
  - The intercept for women is $1.698 - .293 = 1.405$ and the slope is $.007 - .006 = .001$ – about 0.1% for each year of experience.
  - The interaction term is marginally statistically significant, with p-value = .064. (So at the 10% level but not the 5%.)

Allowing for Different Slopes

- Use WAGE1.DTA:

$$\begin{aligned}\widehat{lwage} &= \underset{(.049)}{1.698} - \underset{(.069)}{.293}\,female + \underset{(.002)}{.007}\,exper - \underset{(.003)}{.006}\,female \cdot exper \\ n &= 526,\ R^2 = .154\end{aligned}$$

  - At any level of experience, the predicted difference in $lwage$ between females and males is

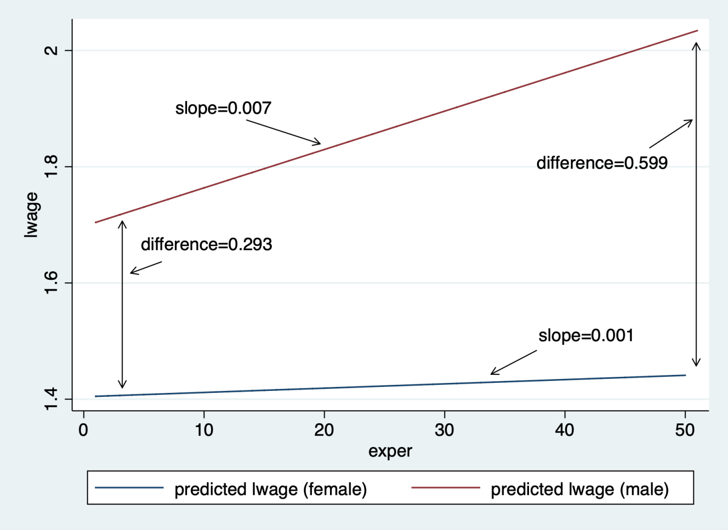
$$-.293 - .006\ exper$$

  - The coefficient $-.293$ on $female$: is the predicted difference in $lwage$ between a woman and a man when $exper = 0$ (not a very interesting subpopulation)
  - A more interesting quantity may be the gap at around the mean $\overline{exper} = 17$:

$$-.293 - .006 \cdot 17 = -.395$$

    or about 39.5% less for women.
  - The gap increases in $exper$.

Allowing for Different Slopes



Allowing for Different Slopes

- Use the same centering scheme as before to directly get the gap estimate at $exper = 17$

```
gen femexper_17=female*(exper-17)

reg lwage female exper femexper_17
```

Source	SS	df	MS
Model	22.9051677	3	7.63505589
Residual	125.424584	522	.24027698
Total	148.329751	525	.28253286

Number of obs	=	526
F(3, 522)	=	31.78
Prob > F	=	0.0000
R-squared	=	0.1544
Adj R-squared	=	0.1496
Root MSE	=	.49018

lwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
female	-.3931104	.0428204	-9.18	0.000	-.4772318	-.3089889
exper	.0066007	.0021976	3.00	0.003	.0022835	.0109179
femexper_17	-.0058634	.0031567	-1.86	0.064	-.0120649	.000338
_cons	1.697672	.0486394	34.90	0.000	1.602119	1.793225

Allowing for Different Slopes

- Add *educ* and the interaction term $female \cdot educ$

```
gen femeduc=female*educ

reg lwage female educ femeduc exper femexper
```

Source	SS	df	MS
Model	54.8532145	5	10.9706429
Residual	93.4765369	520	.179762571
Total	148.329751	525	.28253286

Number of obs	=	526
F(5, 520)	=	61.03
Prob > F	=	0.0000
R-squared	=	0.3698
Adj R-squared	=	0.3637
Root MSE	=	.42398

lwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
female	.0850418	.2035344	0.42	0.676	-.3148089	.4848925
educ	.1018128	.0091857	11.08	0.000	.0837672	.1198583
femeduc	-.0194987	.0144173	-1.35	0.177	-.047822	.0088246
exper	.0149081	.0020432	7.30	0.000	.0108941	.0189221
femexper	-.0107923	.0028679	-3.76	0.000	-.0164264	-.0051581
_cons	.2497944	.1372369	1.82	0.069	-.0198125	.5194014

Allowing for Different Slopes

$$lwage = \beta_0 + \delta_0 female + \beta_1 exper + \delta_1 female \cdot exper + \beta_2 educ + \delta_2 female \cdot educ + u$$

- We can test

H_0: no difference between women and men

by testing the three variables $female$, $femeduc$ and $femexper$ jointly, i.e.,

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

```
test female femeduc femexper

 ( 1)  female = 0
 ( 2)  femeduc = 0
 ( 3)  femexper = 0

       F(  3,   520) =   33.13
            Prob > F =    0.0000
```

Allowing for Different Slopes

- A final **warning**: it is hard to justify omitting the level of a variable but including an interaction that involves that variable.
- Suppose we drop $educ$ but include $female \cdot educ$.

$$lwage = \beta_0 + \delta_0 female + \beta_1 female \cdot educ + ...$$

- The model imposes a zero return to education for men, which we cannot justify
- The coefficient on $female \cdot educ$ is now a direct estimate of the return to education for women, rather than being the difference in the returns between women and men

Outline

Testing for Differences Across Groups

- **Chow test**: Does the regression function change across groups?
  - Allow all parameters to vary across groups
- In the general k variable case, we can define a dummy variable, w, indicating the two groups. Then

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \\ &\quad + \delta_0 w + \delta_1 w \cdot x_1 + \delta_2 w \cdot x_2 + ... + \delta_k w \cdot x_k + u \\ H_0 &: \delta_0 = 0, \delta_1 = 0, \delta_2 = 0, ..., \delta_k = 0 \end{aligned}$$

  for $k+1$ restrictions.
- We can use a standard F test of the $k+1$ exclusion restrictions.
- Reject if F statistic is greater than the critical value from $\mathcal{F}_{k+1,n-2(k+1)}$

Testing for Differences Across Groups

- Rather than construct all of the interactions and run the regression with interactions, there is an equivalent way to implement this test (without using the interactions).

  1. Pool the data and estimate a single regression. This is the **restricted** model, and produces the restricted SSR. Call this the *pooled* SSR, SSR_P.

  2. Estimate the regressions on the two groups (say, 1 and 2) separately. Get SSR_1 and SSR_2. The **unrestricted** SSR is $SSR_1 + SSR_2$ (this is the same as the regression that includes the full set of interactions).

- The F statistic is

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

and, under H_0, has the $\mathcal{F}_{k+1, n-2(k+1)}$ distribution under H_0

Outline

The Linear Probability Model

- We have studied many ways that binary (dummy) explanatory variables can appear in regression analysis.
- **Question**: What if the *dependent* variable is a dummy variable?
  - We want to explain the outcome of a yes/no or zero/one event.
- For example, we want to study the question of married women's labor force participation. Or, we want to know whether a young man is arrested for a crime during a certain period of time.
- In these cases, the variable y we want to explain is a binary or dummy variable.

The Linear Probability Model

- How do we interpret the population model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

  when y is binary?

- y can only change from 0 to 1 or 1 to 0.

- Suppose $\beta_1 = .035$ and $x_1 = educ$. What does it mean for a one year increase in *educ* to increase y by .035?

- Same problem arises for other discrete variables, such as $y = number\ of\ arrests$ or $y = number\ of\ children$ (cannot have a fraction more of a child)

The Linear Probability Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

- We rely on the expected value formulation of linear regression.

  Recall under MLR.4,

$$\mathbb{E}[y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k,$$

  where $\mathbf{x}$ is shorthand for $(x_1, x_2, ..., x_k)$.

- Remember, we can interpret β_j as

$$\Delta\mathbb{E}[y|\mathbf{x}] = \beta_j \Delta x_j, \text{ holding other explanatory variables fixed}$$

- Key relationship when y is binary:

$$\mathbb{E}[y|\mathbf{x}] = \mathbb{P}[y = 1|\mathbf{x}]$$

The Linear Probability Model

- We call $\mathbb{P}[y=1|\mathbf{x}]$ the **response probability**.
- When we apply the linear model to binary y, we are really saying

$$\mathbb{P}[y=1|\mathbf{x}]=\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k$$

and we call the model the **linear probability model (LPM)**. (Other **binary response models** are covered in Chapter 17 of Wooldridge)

  - **Important**: all partial effects are effects on the probability that $y=1$

$$\Delta\mathbb{P}[y=1|\mathbf{x}]=\beta_j\Delta x_j,\ \text{holding other explanatory variables fixed}$$

  - Since $\mathbb{P}[y=0|\mathbf{x}]=1-\mathbb{P}[y=1|\mathbf{x}]$, it is the only probability we need.
  - The sample analog holds as well. When we have the OLS regression line

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+...+\hat{\beta}_k x_k,$$

    $\hat{y}$ is now the predicted probability.

The Linear Probability Model

- **EXAMPLE:** Married Women's Labor Force Participation (MROZ.DTA)
  - $inlf = 1$ if a woman worked for a wage during a certain year, and 0 if not.
  - We estimate a linear probability model to see the effects of variables on the probability of being in the labor force.

```
. des inlf nwifeinc educ exper age kidslt6 kidsge6

Variable      Storage   Display    Value
    name         type    format    label      Variable label
------------------------------------------------------------------------
inlf            byte    %9.0g                 =1 if in lab frce, 1975
nwifeinc        float   %9.0g                 (faminc - wage*hours)/1000
educ            byte    %9.0g                 years of schooling
exper           byte    %9.0g                 actual labor mkt exper
age             byte    %9.0g                 woman's age in yrs
kidslt6         byte    %9.0g                 # kids < 6 years
kidsge6         byte    %9.0g                 # kids 6-18
```

The Linear Probability Model

```
reg inlf nwifeinc educ exper expersq age kidslt6 kidsge6
```

Source	SS	df	MS
Model	48.8080578	7	6.97257969
Residual	135.919698	745	.182442547
Total	184.727756	752	.245648611

Number of obs	=	753
F(7, 745)	=	38.22
Prob > F	=	0.0000
R-squared	=	0.2642
Adj R-squared	=	0.2573
Root MSE	=	.42713

inlf	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
nwifeinc	-.0034052	.0014485	-2.35	0.019	-.0062488	-.0005616
educ	.0379953	.007376	5.15	0.000	.023515	.0524756
exper	.0394924	.0056727	6.96	0.000	.0283561	.0506287
expersq	-.0005963	.0001848	-3.23	0.001	-.0009591	-.0002335
age	-.0160908	.0024847	-6.48	0.000	-.0209686	-.011213
kidslt6	-.2618105	.0335058	-7.81	0.000	-.3275875	-.1960335
kidsge6	.0130122	.013196	0.99	0.324	-.0128935	.0389179
_cons	.5855192	.154178	3.80	0.000	.2828442	.8881943

The Linear Probability Model

- The estimated equation is

$$\begin{aligned}\widehat{inlf} &= \underset{(.154)}{.586} - \underset{(.0014)}{.0034} nwifeinc + \underset{(.007)}{.038}\, educ + \underset{(.006)}{.039}\, exper - \underset{(.00018)}{.00060} exper^2 \\ &\quad - \underset{(.002)}{.016}\, age - \underset{(.034)}{.262}\, kidslt6 + \underset{(.013)}{.013}\, kidsge6 \\ n &= 753,\ R^2 = .264\end{aligned}$$

  - The coefficient on $nwifeinc$ (other sources of income) shows a modest effect: if it increases by 20 ($\$20,000$, about one standard deviation), the probability of being in the labor force falls by .068, or 6.8 percentage points. The t statistic shows it is statistically significant at the 2% level.
  - Each year of education increases the probability by an estimated .038, or 3.8 percentage points.

The Linear Probability Model

- The estimated equation is

$$
\begin{aligned}
\widehat{inlf} &= \underset{(.154)}{.586} - \underset{(.0014)}{.0034} nwifeinc + \underset{(.007)}{.038}\, educ + \underset{(.006)}{.039}\, exper - \underset{(.00018)}{.00060} exper^2 \\
&\quad - \underset{(.002)}{.016}\, age - \underset{(.034)}{.262}\, kidslt6 + \underset{(.013)}{.013}\, kidsge6 \\
n &= 753,\ R^2 = .264
\end{aligned}
$$

  - Past workforce experience has a positive but diminishing effect. The effect of the first year is about .039, and this diminishes to zero at *exper* $= .039/(2 \cdot .0006) = 32.5$. (Only 13 women have *exper* > 32.)
  - Having young children has a very large negative effect: being in the labor force falls by .262 for each young child.
  - It is unwise to extrapolate to extreme values when using any linear model

Shortcomings of the LPM

- Using a linear model for a binary outcome is convenient because estimation is easy and so is interpretation.

- But the LPM does have some shortcomings.

  1. The fitted values from an OLS regression are never guaranteed to be between zero and one

     - Slightly embarrassing but is rarely a big deal. We usually use the LPM to estimate partial effects, not to make predictions.

Shortcomings of the LPM

- Using a linear model for a binary outcome is convenient because estimation is easy and so is interpretation.
- But the LPM does have some shortcomings.

  2. The estimated partial effects are constant throughout the range of explanatory variables (possibly silly estimated effects for large changes)

     - ⋆ This is more of a problem because we know, say, a variable with a positive effect on $\mathbb{P}[y = 1|\mathbf{x}]$ must eventually have a diminishing effect. But the linear model implies a constant effect (when the variable appears by itself).
     - ⋆ But LPM does a good job of approximating partial effects if we do not look at extreme values of the explanatory variables.

Shortcomings of the LPM

- Using a linear model for a binary outcome is convenient because estimation is easy and so is interpretation.
- But the LPM does have some shortcomings.
  3. Because y is binary, the LPM must exhibit **heteroskedasticity** except in the one case where no x_j affects $P(y = 1|\mathbf{x})$:

$$\mathbb{V}[y|\mathbf{x}] = p(\mathbf{x})[1 - p(\mathbf{x})]$$

  where $p(\mathbf{x}) = \mathbb{P}[y = 1|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$

    - This is a case where we *know* MLR.5 must fail, and we know how. So, currently, we treat the usual t and F tests with suspicion, and the confidence intervals.
    - We will relax MLR.5 later