# Cross-sectional Regression: Binary Dummy Classification

Pengpeng Yue

Version: Fall 2022

# Quantitative Variables vs Qualitative Variables

- ✓ Quantitative Variables: hourly wage rate, years of education, college grade point average
- ✓ Qualitative Variables: gender, race of an individual, the industry of a firm

# Quantitative Variables vs Qualitative Variables

- ✓ Binary variables or Zero-one variables
- ✓ In econometrics, binary variables are most commonly called dummy variables

# Wage1.dta: why dummy

- ✓ Wage1.dta: a study of individual wage determination
- ✓ Female, Married
- ✓ Gender (1 or 2) vs Female (0 or 1)?
- ✓ See Wage1.dta

# How do we incorporate binary information into regression models?

$$\text{wage} = \beta_0 + \delta_0 \text{female } + \beta_1 \text{educ} + \mu \tag{1}$$

$\delta_0$ is the difference in hourly wage between females and males, given the same amount of education.
The coefficient $\delta_0$ determines whether there is discrimination against women: if $\delta_0 < 0$, for the same level of other factors, women earn less than men on average.

# the zero conditional mean assumption

$$\delta_0 = \mathrm{E}(\mathsf{wage}|\mathsf{female} = 1, \mathit{educ}) - \mathrm{E}(\mathsf{wage}|\mathsf{female} = 0, \mathit{educ}) \qquad (2)$$

The key here is that the level of education is the same in both expectations; the difference, $\delta_0$, is due to gender only.

# Hourly wage equation

$$
\begin{aligned}
\widehat{wage} = & -\underset{(.72)}{1.57} - \underset{(.26)}{1.81}\text{female} + \underset{(.049)}{.572}\text{educ} \\
& + \underset{(.012)}{.025}\text{exper} + \underset{(.021)}{.141}\text{tenure} \\
& n = 526, R^2 = .364
\end{aligned} \tag{3}
$$

See Wage1.dta

```
. reg wage female educ exper tenure

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(4, 521)       =     74.40
       Model |  2603.10658         4  650.776644   Prob > F        =    0.0000
    Residual |  4557.30771       521   8.7472317   R-squared       =    0.3635
-------------+----------------------------------   Adj R-squared   =    0.3587
       Total |  7160.41429       525  13.6388844   Root MSE        =    2.9576

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -1.810852   .2648252    -6.84   0.000    -2.331109   -1.290596
        educ |   .5715048   .0493373    11.58   0.000     .4745802    .6684293
       exper |   .0253959   .0115694     2.20   0.029     .0026674    .0481243
      tenure |   .1410051   .0211617     6.66   0.000     .0994323    .1825778
       _cons |  -1.567939   .7245511    -2.16   0.031    -2.991339    -.144538
------------------------------------------------------------------------------
```

# Hourly wage equation

$$\widehat{wage} = 7.10 - 2.51 \text{female}$$
$$(.21) \quad (.30) \tag{4}$$
$$n = 526, R^2 = .116$$

See Wage1.dta

Equation (4) provides a simple way to carry out a comparison-of-means test between the two groups, which in this case are men and women.

# Hourly wage equation: regression in Stata

```
. 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
------------------------------------------------------------------------------
```

# Hourly wage equation

The estimated wage differential between men and women is larger in (3) than in (4) because (4) does not control for differences in education, experience, and tenure, and these are lower, on average, for women than for men in this sample.

# Effects of computer ownership on college GPA

$$colGPA = \beta_0 + \delta_0 PC + \beta_1 hsGPA + \beta_2 ACT + \mu \tag{5}$$

where the dummy variable PC equals one if a student owns a personal computer and zero otherwise. The variables hsGPA (high school GPA) and ACT (achievement test score) are used as controls.
See GPA1.dta

# Reasons PC ownership might have an effect on colGPA

- ✓ work might be of higher quality (Positive)
- ✓ time can be saved by not having to wait at a computer lab (Positive)
- ✓ inclined to play computer games or surf the Internet (Negative)

# Effects of computer ownership on college GPA

$$\widehat{\text{colGPA}} = \underset{(.33)}{1.26} + \underset{(.057)}{.157}PC + \underset{(.094)}{.447}hsGPA + \underset{(.0105)}{.0087}ACT \tag{6}$$
$$n = 141, R^2 = .219$$

# Effects of computer ownership on college GPA: regression in Stata

```
. reg colGPA PC hsGPA ACT

      Source |       SS           df       MS      Number of obs   =       141
-------------+----------------------------------   F(3, 137)       =     12.83
       Model |  4.25741863         3  1.41913954   Prob > F        =    0.0000
    Residual |  15.1486808       137  .110574313   R-squared       =    0.2194
-------------+----------------------------------   Adj R-squared   =    0.2023
       Total |  19.4060994       140  .138614996   Root MSE        =    .33253

------------------------------------------------------------------------------
      colGPA |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          PC |   .1573092   .0572875     2.75   0.007     .0440271    .2705913
       hsGPA |   .4472417   .0936475     4.78   0.000     .2620603     .632423
         ACT |    .008659   .0105342     0.82   0.413    -.0121717    .0294897
       _cons |    1.26352   .3331255     3.79   0.000     .6047871    1.922253
------------------------------------------------------------------------------
```

# Summary 1

- ✓ viewed as having relevance for policy analysis
- ✓ gender discrimination in the workforce
- ✓ effect of computer ownership on college performance
- ✓ two groups of subjects: control group vs experimental group or treatment group
- ✓ the choice of the control and treatment groups is not random
- ✓ control for enough other factors and estimate the causal effect

# Control for enough other factors: regression in Stata

```
. reg wage educ exper expersq tenure tenursq nonwhite female married numdep smsa northcen south wes
> t construc ndurman trcommpu trade services profserv profocc clerocc servocc, r
```

Linear regression

| | | |
|---|---|---|
| Number of obs | = | 526 |
| F(22, 503) | = | 17.54 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.4917 |
| Root MSE | = | 2.6899 |

| wage | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| educ | .3264567 | .0662022 | 4.93 | 0.000 | .1963898 | .4565237 |
| exper | .1650197 | .0325703 | 5.07 | 0.000 | .1010291 | .2290104 |
| expersq | -.0034346 | .0006946 | -4.94 | 0.000 | -.0047993 | -.0020699 |
| tenure | .1568149 | .0486385 | 3.22 | 0.001 | .0612553 | .2523746 |
| tenursq | -.0017088 | .0015943 | -1.07 | 0.284 | -.0048412 | .0014235 |
| nonwhite | -.0846544 | .3694596 | -0.23 | 0.819 | -.8105285 | .6412197 |
| female | -1.629549 | .2488436 | -6.55 | 0.000 | -2.11845 | -1.140648 |
| married | .1003878 | .2609132 | 0.38 | 0.701 | -.4122261 | .6130017 |
| numdep | -.0022417 | .0914276 | -0.02 | 0.980 | -.1818688 | .1773853 |
| smsa | .7094254 | .2620937 | 2.71 | 0.007 | .1944922 | 1.224359 |
| northcen | -.5668205 | .351546 | -1.61 | 0.108 | -1.2575 | .1238589 |
| south | -.4482182 | .3203941 | -1.40 | 0.162 | -1.077694 | .1812574 |
| west | .4380886 | .4238091 | 1.03 | 0.302 | -.3945654 | 1.270743 |
| construc | -.5050103 | .7227234 | -0.70 | 0.485 | -1.924939 | .9149182 |
| ndurman | -.8074111 | .4961515 | -1.63 | 0.104 | -1.782196 | .1673734 |
| trcommpu | -1.038444 | .5391083 | -1.93 | 0.055 | -2.097625 | .0207376 |
| trade | -2.03021 | .4388118 | -4.63 | 0.000 | -2.89234 | -1.16808 |
| services | -1.762623 | .4941861 | -3.57 | 0.000 | -2.733546 | -.7916998 |
| profserv | -.9333546 | .5075739 | -1.84 | 0.067 | -1.930581 | .0638714 |
| profocc | 1.890814 | .351843 | 5.37 | 0.000 | 1.199551 | 2.582077 |
| clerocc | .3351192 | .3643302 | 0.92 | 0.358 | -.3806772 | 1.050916 |
| servocc | .0042979 | .3234446 | 0.01 | 0.989 | -.631171 | .6397668 |
| cons | .8033647 | .8829951 | 0.91 | 0.363 | -.931448 | 2.538178 |

# Effects of training grants on hours of training

$$
\begin{aligned}
\widehat{hrsemp} =& 46.67 + 26.25 \text{ grant } - .98 \log(\text{sales}) \\
& (43.41) \quad (5.59) \qquad\quad (3.54) \\
& - 6.07 \log(\text{employ}) \\
& \quad (3.88) \\
n =& 105, R^2 = .237
\end{aligned} \tag{7}
$$

The dependent variable is hours of training per employee, at the firm level. The variable grant is a dummy variable equal to one if the firm received a job training grant for 1988 and zero otherwise. The variables sales and employ represent annual sales and number of employees, respectively.
See JTRAIN.dta

# Effects of training grants on hours of training: regression in Stata

```
. reg hrsemp grant lsales lemploy if year == 1988

      Source |       SS           df       MS      Number of obs   =       105
-------------+----------------------------------   F(3, 101)       =     10.44
       Model |  18622.7268         3  6207.57559   Prob > F        =    0.0000
    Residual |  60031.0921       101  594.367249   R-squared       =    0.2368
-------------+----------------------------------   Adj R-squared   =    0.2141
       Total |  78653.8189       104   756.28672   Root MSE        =     24.38

------------------------------------------------------------------------------
      hrsemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       grant |    26.2545   5.591765     4.70   0.000     15.16194    37.34705
      lsales |  -.9845809   3.539903    -0.28   0.781    -8.006797    6.037635
     lemploy |  -6.069871   3.882893    -1.56   0.121    -13.77249    1.632744
       _cons |   46.66508    43.4121     1.07   0.285    -39.45284     132.783
------------------------------------------------------------------------------
```

# Causal? Or not?

In equation (7), is the difference in training between firms that receive grants and those that do not due to the grant, or is grant receipt simply an indicator of something else?

- ✓ It might be that the firms receiving grants would have, on average, trained their workers more even in the absence of a grant
- ✓ must know how the firms receiving grants were determined

# Interpreting Coefficients on Dummy Explanatory Variables When the Dependent Variable Is $log(y)$

The dependent variable appearing in logarithmic form, with one or more dummy variables appearing as independent variables. How do we interpret the dummy variable coefficients in this case?

✓ a percentage interpretation

# Interpreting Coefficients on Dummy Explanatory Variables

$$
\begin{aligned}
\widehat{\log(\text{price})} = & -\underset{(.65)}{1.35} + \underset{(.038)}{.168}\log(\text{lotsize}) + \underset{(.093)}{.707}\log(\text{sqrft}) \\
& + \underset{(.029)}{.027} bdrms + \underset{(.045)}{.054}\ \text{colonial} \\
& n = 88,\ R^2 = .649
\end{aligned} \tag{8}
$$

See HPRICE1.dta

- ✓ Colonial, which is a binary variable equal to one if the house is of the colonial style.
- ✓ a colonial-style house is predicted to sell for about 5.4% more, holding other factors fixed.
- ✓ when $log(y)$ is the dependent variable in a model, the coefficient on a dummy variable, when multiplied by 100, is interpreted as the percentage difference in y

# Interpreting Coefficients on Dummy Explanatory Variables: regression in Stata

```
. reg lprice llotsize lsqrft bdrms colonial

      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(4, 83)        =     38.38
       Model |  5.20397919         4  1.3009948   Prob > F        =    0.0000
    Residual |  2.81362433        83  .033899088   R-squared       =    0.6491
-------------+----------------------------------   Adj R-squared   =    0.6322
       Total |  8.01760352        87  .092156362   Root MSE        =    .18412

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    llotsize |   .1678189   .0381807     4.40   0.000     .0918791    .2437587
      lsqrft |   .7071931    .092802     7.62   0.000     .5226138    .8917725
       bdrms |   .0268305   .0287236     0.93   0.353    -.0302995    .0839605
    colonial |   .0537962   .0447732     1.20   0.233     -.035256    .1428483
       _cons |  -1.349589    .651041    -2.07   0.041    -2.644483   -.0546947
------------------------------------------------------------------------------
```

# Log hourly wage equation

$$
\begin{aligned}
\widehat{\log(wage)} =& .417 - .297 female + .080 educ + .029 exper \\
& (.099)(.036) \qquad (.007) \qquad (.005) \\
& - .00058 exper^2 + .032 tenure - .00059 tenure^2 \qquad (9) \\
& (.00010) \qquad (.007) \qquad (.00023) \\
& n = 526, R^2 = .441
\end{aligned}
$$

the coefficient on female implies that, for the same levels of educ, exper, and tenure, women earn about 100(.297) 29.7% less than men.

```
. reg lwage female educ exper expersq tenure tenursq

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(6, 519)       =     68.18
       Model |  65.3791009         6  10.8965168   Prob > F        =    0.0000
    Residual |  82.9506505       519  .159827843   R-squared       =    0.4408
-------------+----------------------------------   Adj R-squared   =    0.4343
       Total |  148.329751       525  .28253286    Root MSE        =    .39978

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -.296511   .0358055    -8.28   0.000    -.3668524   -.2261696
        educ |   .0801967   .0067573    11.87   0.000     .0669217    .0934716
       exper |   .0294324   .0049752     5.92   0.000     .0196585    .0392063
     expersq |  -.0005827   .0001073    -5.43   0.000    -.0007935   -.0003719
      tenure |   .0317139   .0068452     4.63   0.000     .0182663    .0451616
     tenursq |  -.0005852   .0002347    -2.49   0.013    -.0010463   -.0001241
       _cons |    .416691   .0989279     4.21   0.000     .2223425    .6110394
------------------------------------------------------------------------------
```

# Using Dummy Variables for Multiple Categories

Use several dummy independent variables in the same equation

$$
\begin{aligned}
\widehat{\log(\text{wage})} =& .321 + .213 \text{ marrmale } - .198 \text{ marrfem} \\
& (.100)(.055) \qquad\qquad (.058) \\
& - .110 \text{ singfem } + .079 \text{ educ } + .027 \text{ exper} \\
& \ (.056) \qquad\qquad (.007) \qquad\qquad (.005) \\
& - .00054 \text{ exper}^2 + .029 \text{ tenure } - .00053 \text{ tenure}^2 \\
& \ (.00011) \qquad\qquad (.007) \qquad\qquad (.00023) \\
& n = 526, R^2 = .461
\end{aligned} \tag{10}
$$

# Effects of physical attractiveness on wage

Hamermesh and Biddle (1994)
Each person in the sample was ranked by an interviewer for physical attractiveness, using five categories: homely, quite plain, average, good looking, and strikingly beautiful or handsome
Because there are so few people at the two extremes, the authors put people into one of three groups for the regression analysis: average, below average, and above average, where the base group is average

# Effects of physical attractiveness on wage for men

Hamermesh and Biddle (1994)

$$\widehat{\log(wage)} = \hat{\beta}_0 - .164 \text{ belavg } + .016 \text{ abvavg } + \text{ other factors} \tag{11}$$
$$\qquad\qquad (.046) \qquad\qquad (.033)$$
$$n = 700, \bar{R}^2 = .403$$

# Effects of physical attractiveness on wage for women

Hamermesh and Biddle (1994)

$$
\begin{aligned}
\widehat{\log(wage)} = \hat{\beta}_0 &- .124 \text{ belavg } + .035 \text{ abvavg } + \text{ other factors} \\
&\quad (.066) \qquad\qquad (.049) \\
n = 409, &\bar{R}^2 = .330
\end{aligned} \tag{12}
$$

Just as variables with quantitative meaning can be interacted in regression models, so can dummy variables.

$$\begin{aligned}\widehat{\log(wage)} =& \underset{(.100)}{.321} - \underset{(.056)}{.110}\text{ female } + \underset{(.055)}{.213}\text{ married} \\ & - \underset{(.072)}{.301}\ \text{female} \times \text{married} + \ldots \end{aligned} \tag{13}$$

# Effects of computer usage on wages

Krueger (1993)

- ✓ defines a dummy variable, which we call compwork, equal to one if an individual uses a computer at work.
- ✓ comphome, equals one if the person uses a computer at home

$$
\begin{aligned}
\widehat{\log(wage)} = & \hat{\beta}_0 + \underset{(.009)}{.177}\ \text{compwork} + \underset{(.019)}{.070}\ \text{comphome} \\
& + \underset{(.023)}{.017}\ \text{compwork} \times \text{comphome} + \text{other factors}
\end{aligned}
\tag{14}
$$

# A Binary Dependent Variable: The Linear Probability Model

- ✓ the properties and applicability of the multiple linear regression model
- ✓ y, takes on only two values: zero and one

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u \tag{15}$$

Assume: $\mathrm{E}\left(u|x_1, \ldots, x_k\right) = 0$

$$\mathrm{E}(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \tag{16}$$

$$\mathrm{P}(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \tag{17}$$

# Linear probability model (LPM)

$$\mathrm{P}(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \tag{18}$$

- ✓ which says that the probability of success
- ✓ $p(\mathbf{x}) = \mathrm{P}(y = 1|\mathbf{x})$ is a linear function of the $x_j$
- ✓ $\mathrm{P}(y = 1|\mathbf{x})$ is called the response probability
- ✓ The multiple linear regression model with a binary dependent variable is called the linear probability model (LPM)

# Linear probability model (LPM)

✓ In the LPM, $\beta_j$ measures the change in the probability of success when $x_j$ changes, holding other factors fixed

$$\Delta \mathrm{P}(y = 1|\mathbf{x}) = \beta_j \Delta x_j \tag{19}$$

# Linear probability model (LPM)

Mroz (1987): *inlf* = 1 if the woman reports working for a wage outside the home at some point during the year

$$
\begin{aligned}
\widehat{\text{inlf}} = & .586 - .0034 \text{ nwifeinc} + .038 \text{ educ} + .039 \text{ exper} \\
& (.154)(.0014) \qquad\qquad (.007) \qquad\qquad (.006) \\
& - .00060 \text{ exper}^2 - .016 \text{ age} - .262 \text{ kidslt } 6 + .013 \text{ kidsge } 6 \qquad (20) \\
& (.00018) \qquad\qquad (.002) \qquad\qquad (.034) \qquad\qquad (.013) \\
n = & 753, R^2 = .264
\end{aligned}
$$

See MROZ.dta.

# Linear probability model (LPM): regression in Stata

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

      Source |       SS           df       MS      Number of obs   =       753
-------------+----------------------------------   F(7, 745)       =     38.22
       Model |  48.8080578         7  6.97257969   Prob > F        =    0.0000
    Residual |  135.919698       745  .182442547   R-squared       =    0.2642
-------------+----------------------------------   Adj R-squared   =    0.2573
       Total |  184.727756       752  .245648611   Root MSE        =    .42713

------------------------------------------------------------------------------
        inlf |      Coef.   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
------------------------------------------------------------------------------
```

# Linear probability model (LPM)

$$\begin{aligned}\widehat{\text{inlf}} =& .586 - .0034 \text{ nwifeinc} + .038 \text{ educ} + .039 \text{ exper} \\ & (.154)(.0014) \qquad\qquad (.007) \qquad\qquad (.006) \\ & - .00060 \text{ exper}^2 - .016 \text{ age} - .262 \text{ kidslt } 6 + .013 \text{ kidsge } 6 \\ & (.00018) \qquad\qquad (.002) \qquad\qquad (.034) \qquad\qquad (.013) \\ n =& 753, R^2 = .264\end{aligned} \tag{21}$$

- ✓ The coefficient on nwifeinc implies that, an increase of \$10,00, the probability that a woman is in the labor force falls by .0034
- ✓ The coefficient on educ means that, everything else in (20) held fixed, another year of education increases the probability of labor force participation by .038

# Summary

- ✓ How to use qualitative information in regression analysis
- ✓ A dummy variable could be defined to distinguish between two groups, and the coefficient estimate on the dummy variable estimates the ceteris paribus difference between the two groups
- ✓ Dummy variables are also useful for incorporating ordinal information, such as a credit or a beauty rating
- ✓ Dummy variables can be interacted with quantitative variables to allow slope differences across different groups
- ✓ The linear probability model, which is simply estimated by OLS, allows us to explain a binary response using regression analysis
- ✓ The LPM does have some drawbacks: it can produce predicted probabilities that are less than zero or greater than one

Beauty and the Labor Market
Hamermesh and Biddle

# About this paper

- ✓ develop a theory of sorting across occupations based on looks
- ✓ derive its implications for testing for the source of earnings differentials related to looks
- ✓ Data: 1977 Quality of Employment, 1971 Quality of American Life, 1981 Canadian Quality of Life Survey (all contain interviewers' ratings of the respondents' physical apperance)

# Findings

- ✓ Plain people earn less than people of average looks, who earn less than the good-looking
- ✓ The penalty for plainness is 5 to 10 percent, slightly larger than the premium for beauty
- ✓ the effects are slightly larger for men than women

# Motivation

- ✓ Discrimination in the labor market has generated immense amounts of research by economists
- ✓ Blacks, Hispanics, women, linguistic minorities, physically handicapped workers...
- ✓ the first study of the economics of discrimination in the labor market against yet another group: the ugly and its obverse, possible favoritism for the beautiful

# Related literature

- ✓ Quinn (1978) finds correlations of interviewers' ratings of the looks of respondents with their incomes
- ✓ Roszell et al. (1989) find faster income growth among better-looking respondents
- ✓ Frieze et al. (1991) find ratings of beauty based on photographs of the students are correlated positively with both starting and subsequent salaries for males.
- ✓ Hatfield and Sprecher (1986) find men beauty enhanced the worker's likelihood of being chosen for both clerical and professional/managerial jobs.

# Methodology

- ✓ determine whether standard earnings equations yield a looks differential
- ✓ determine whether the differential differs across occupations in ways the model suggests
- ✓ look for evidence of the sorting implied by the productivity model
- ✓ check whether more attractive workers tend to be concentrated in those occupations

# Date and Variable

- ✓ Two broad household surveys for the U.S. and one for Canada
- ✓ 1977 Quality of Employment, 1515 workers
- ✓ 1971 Quality of American Life, 2164 wokrers
- ✓ 1981 Canadian Quality of Life Survey, 3415 workers

# Date and Variable

✓ Beauty: the surveys asked the interviewer to "rate the respondent's physical appearance" on the five-point scale, homely below average (plain), average, above average (good looking), strikingly handsome.

```
. 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 | |

# Beauty: male vs female

```
. tab looks female
```

| from 1 to 5 | =1 if female 0 | 1 | Total |
|---|---|---|---|
| 1 | 8 | 5 | 13 |
| 2 | 88 | 54 | 142 |
| 3 | 489 | 233 | 722 |
| 4 | 228 | 136 | 364 |
| 5 | 11 | 8 | 19 |
| Total | 824 | 436 | 1,260 |

# Summary Statistics

```
. sum
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| wage | 1,260 | 6.30669 | 4.660639 | 1.02 | 77.72 |
| lwage | 1,260 | 1.6588 | .5945075 | .0198026 | 4.353113 |
| belavg | 1,260 | .1230159 | .3285858 | 0 | 1 |
| abvavg | 1,260 | .3039683 | .4601517 | 0 | 1 |
| exper | 1,260 | 18.20635 | 11.96349 | 0 | 48 |
| looks | 1,260 | 3.185714 | .6848774 | 1 | 5 |
| union | 1,260 | .2722222 | .4452804 | 0 | 1 |
| goodhlth | 1,260 | .9333333 | .2495429 | 0 | 1 |
| black | 1,260 | .0738095 | .2615645 | 0 | 1 |
| female | 1,260 | .3460317 | .4758923 | 0 | 1 |
| married | 1,260 | .6912698 | .462153 | 0 | 1 |
| south | 1,260 | .1746032 | .3797781 | 0 | 1 |
| bigcity | 1,260 | .2190476 | .4137652 | 0 | 1 |
| smllcity | 1,260 | .4666667 | .4990857 | 0 | 1 |
| service | 1,260 | .2738095 | .4460895 | 0 | 1 |
| expersq | 1,260 | 474.4825 | 534.6454 | 0 | 2304 |
| educ | 1,260 | 12.56349 | 2.624489 | 5 | 17 |

# Summary Statistics

```
. sum looks belavg abvavg
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| looks | 1,260 | 3.185714 | .6848774 | 1 | 5 |
| belavg | 1,260 | .1230159 | .3285858 | 0 | 1 |
| abvavg | 1,260 | .3039683 | .4601517 | 0 | 1 |

# Model

$$log(wage) = \alpha + \beta_0 \text{belave} + \beta_1 \text{abvavg} + \text{other factors} \tag{22}$$

# The impact of looks on employee's earnings, QES 1977

```
. reg lwage belavg abvavg educ exper expersq union-service, r
```

Linear regression

| | | | |
|---|---|---|---|
| Number of obs | = | 1,260 |
| F(14, 1245) | = | 73.98 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.4131 |
| Root MSE | = | .45802 |

| lwage | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| belavg | -.132568 | .0395233 | -3.35 | 0.001 | -.2101076 | -.0550285 |
| abvavg | .0121734 | .0307301 | 0.40 | 0.692 | -.0481151 | .0724618 |
| educ | .0673383 | .0056074 | 12.01 | 0.000 | .0563373 | .0783394 |
| exper | .0407251 | .0041554 | 9.80 | 0.000 | .0325728 | .0488774 |
| expersq | -.0006345 | .0000921 | -6.89 | 0.000 | -.0008151 | -.0004538 |
| union | .1601863 | .027663 | 5.79 | 0.000 | .105915 | .2144576 |
| goodhlth | .0688725 | .0633036 | 1.09 | 0.277 | -.055321 | .1930661 |
| black | -.0848692 | .0587534 | -1.44 | 0.149 | -.2001358 | .0303973 |
| female | -.3804478 | .0310708 | -12.24 | 0.000 | -.4414046 | -.3194909 |
| married | .0363606 | .0305869 | 1.19 | 0.235 | -.023647 | .0963682 |
| south | .0682183 | .0320746 | 2.13 | 0.034 | .0052921 | .1311445 |
| bigcity | .2391346 | .037477 | 6.38 | 0.000 | .1656096 | .3126596 |
| smllcity | .0902632 | .0307589 | 2.93 | 0.003 | .0299181 | .1506083 |
| service | -.1445479 | .0335156 | -4.31 | 0.000 | -.2103013 | -.0787945 |
| _cons | .323038 | .1028998 | 3.14 | 0.002 | .1211618 | .5249141 |

# The impact of looks on employee's earnings, QES 1977: men

```
. reg lwage belavg abvavg educ exper expersq union-service if !female , r
note: female omitted because of collinearity

Linear regression                               Number of obs     =        824
                                                F(13, 810)        =      27.43
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3084
                                                Root MSE          =     .45282
```

| lwage | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| belavg | -.1433863 | .0502696 | -2.85 | 0.004 | -.2420603 | -.0447122 |
| abvavg | -.0010065 | .0375645 | -0.03 | 0.979 | -.0747418 | .0727288 |
| educ | .0603151 | .0071227 | 8.47 | 0.000 | .046334 | .0742963 |
| exper | .0494652 | .0053137 | 9.31 | 0.000 | .0390348 | .0598955 |
| expersq | -.0007947 | .0001117 | -7.11 | 0.000 | -.0010139 | -.0005754 |
| union | .109175 | .0311586 | 3.50 | 0.000 | .0480138 | .1703362 |
| goodhlth | .001204 | .0858561 | 0.01 | 0.989 | -.1673227 | .1697307 |
| black | -.2771892 | .0706494 | -3.92 | 0.000 | -.4158667 | -.1385117 |
| female | 0 | (omitted) | | | | |
| married | .0824294 | .0397771 | 2.07 | 0.039 | .0043511 | .1605077 |
| south | .1037158 | .0387396 | 2.68 | 0.008 | .027674 | .1797576 |
| bigcity | .2734916 | .0457833 | 5.97 | 0.000 | .1836238 | .3633595 |
| smllcity | .1346177 | .0384882 | 3.50 | 0.000 | .0590693 | .2101661 |
| service | -.2089609 | .0471125 | -4.44 | 0.000 | -.3014379 | -.1164839 |
| _cons | .3580113 | .133867 | 2.67 | 0.008 | .0952441 | .6207785 |

# The impact of looks on employee's earnings, QES 1977: women

```
. reg lwage belavg abvavg educ exper expersq union-service if female , r
note: female omitted because of collinearity

Linear regression                               Number of obs     =        436
                                                F(13, 422)        =      16.36
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3003
                                                Root MSE          =     .44534
```

| lwage | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| belavg | -.1151564 | .0591999 | -1.95 | 0.052 | -.2315198 | .001207 |
| abvavg | .0575209 | .052332 | 1.10 | 0.272 | -.045343 | .1603848 |
| educ | .0769358 | .0090836 | 8.47 | 0.000 | .0590809 | .0947906 |
| exper | .0300475 | .0073818 | 4.07 | 0.000 | .0155377 | .0445572 |
| expersq | -.0005099 | .0001989 | -2.56 | 0.011 | -.000901 | -.0001189 |
| union | .2843611 | .0569626 | 4.99 | 0.000 | .1723954 | .3963268 |
| goodhlth | .1279672 | .0806454 | 1.59 | 0.113 | -.0305496 | .2864839 |
| black | .1058475 | .083257 | 1.27 | 0.204 | -.0578026 | .2694976 |
| female | 0 | (omitted) | | | | |
| married | -.0549752 | .046386 | -1.19 | 0.237 | -.1461516 | .0362011 |
| south | -.0044875 | .0573383 | -0.08 | 0.938 | -.1171918 | .1082167 |
| bigcity | .172293 | .0635147 | 2.71 | 0.007 | .0474484 | .2971376 |
| smllcity | .0130385 | .0500073 | 0.26 | 0.794 | -.0852559 | .1113329 |
| service | -.0907494 | .0462749 | -1.96 | 0.051 | -.1817073 | .0002086 |
| _cons | -.1027681 | .1357105 | -0.76 | 0.449 | -.3695208 | .1639846 |

Any new idea?

How computers have changed the wage structure- evidence from microdata
Krueger

# About this paper

- ✓ Aim: examines whether employees who use a computer at work earn a higher wage rate than otherwise similar workers who do not use a computer at work.
- ✓ Data: the Current Population Survey and High School and Beyond Survey
- ✓ Methods: models to correct for unobserved variables (that might be correlated with both job-related computer use and earnings)

# Findings

- ✓ workers who use computers on their job earn roughly a 10 to 15 percent higher wage rate
- ✓ the expansion in computer use in the 1990s can account for between one-third and one-half of the observed increase in the rate of return to education
- ✓ occupations that experienced greater growth in computer use between 1984 and 1989 also experienced above average wage growth.

# Motivation

- ✓ significant changes in the structure of wages took place in the United States in the 1980s
- ✓ two leading hypotheses that have emerged to explain the rapid changes in the wage structure
  - ✓ increased international competition in several industries has hurt the economic position of low-skilled and less-educated workers in the U.S. (Murphy and Welch, 1991)
  - ✓ repaid, skilled-biased technological change in the 1980s caused profound changes in the relative productivity of various types of workers (Bound and Johnson, 1989; Mincer, 1991; Allen, 1991)
- ✓ this paper explores the impact of the computer revolution on the wage structure using three microdata sets.

# Effects of computer usage on wages

Krueger (1993)

- ✓ defines a dummy variable, which we call compwork, equal to one if an individual uses a computer at work.
- ✓ comphome, equals one if the person uses a computer at home

$$
\begin{aligned}
\widehat{\log(wage)} = & \hat{\beta}_0 + \underset{(.009)}{.177}\ \text{compwork} + \underset{(.019)}{.070}\ \text{comphome} \\
& + \underset{(.023)}{.017}\ \text{compwork} \times \text{comphome} + \text{other factors}
\end{aligned} \tag{23}
$$

Any new idea?

The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions

Mroz

# About this paper

- ✓ Aim: undertakes a systematic analysis of several theoretic and statistical assumptions used in many empirical models of female labor supply
- ✓ Data: PSID 1975 labor supply data
- ✓ Two Assumptions: (1) the Tobit assumption used to control for self-selection into the labor force and (2) exogeneity assumptions on the wife's wage rate and her labor market experience.

# Motivation

- ✓ the studies not sufficient to reach any firm conclusion
- ✓ questions relating to the consequences of measurement error, sample selection bias, and the inclusion of taxes
- ✓ this study attempts a systematic analysis of many of the theoretical and statistical issues raised in previous studies of female labor supply

# Methodology

✓ examine three methodological considerations: exogeneity assumptions, statistical control for self-selection into the labor force, and the impact of controlling for taxes.

# Date and Variable

- ✓ the data coms from the University of Michigan Panel Study of Income Dynamics (PSID) for the year 1975 (interview year 1976)
- ✓ the sample consists of 753 married white women between the ages of 30 and 60 in 1975, with 428 working at some time during the year

# Summary Statistics

```
. sum
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| inlf | 753 | .5683931 | .4956295 | 0 | 1 |
| hours | 753 | 740.5764 | 871.3142 | 0 | 4950 |
| kidslt6 | 753 | .2377158 | .523959 | 0 | 3 |
| kidsge6 | 753 | 1.353254 | 1.319874 | 0 | 8 |
| age | 753 | 42.53785 | 8.072574 | 30 | 60 |
| educ | 753 | 12.28685 | 2.280246 | 5 | 17 |
| wage | 428 | 4.177682 | 3.310282 | .1282 | 25 |
| repwage | 753 | 1.849734 | 2.419887 | 0 | 9.98 |
| hushrs | 753 | 2267.271 | 595.5666 | 175 | 5010 |
| husage | 753 | 45.12085 | 8.058793 | 30 | 60 |
| huseduc | 753 | 12.49137 | 3.020804 | 3 | 17 |
| huswage | 753 | 7.482179 | 4.230559 | .4121 | 40.509 |
| faminc | 753 | 23080.59 | 12190.2 | 1500 | 96000 |
| mtr | 753 | .6788632 | .0834955 | .4415 | .9415 |
| motheduc | 753 | 9.250996 | 3.367468 | 0 | 17 |
| fatheduc | 753 | 8.808765 | 3.57229 | 0 | 17 |
| unem | 753 | 8.623506 | 3.114934 | 3 | 14 |
| city | 753 | .6427623 | .4795042 | 0 | 1 |
| exper | 753 | 10.63081 | 8.06913 | 0 | 45 |
| nwifeinc | 753 | 20.12896 | 11.6348 | -.0290575 | 96 |
| lwage | 428 | 1.190173 | .7231978 | -2.054164 | 3.218876 |
| expersq | 753 | 178.0385 | 249.6308 | 0 | 2025 |

# Linear probability model (LPM)

Mroz (1987): $inlf = 1$ if the woman reports working for a wage outside the home at some point during the year

$$
\begin{aligned}
\widehat{\text{inlf}} =& .586 - .0034 \text{ nwifeinc } + .038 \text{ educ } + .039 \text{ exper} \\
& (.154)(.0014) \qquad\qquad (.007) \qquad\qquad (.006) \\
& - .00060 \text{ exper }^2 - .016 \text{ age } - .262 \text{ kidslt } 6 + .013 \text{ kidsge } 6 \qquad (24) \\
& (.00018) \qquad\qquad (.002) \qquad\qquad (.034) \qquad\qquad (.013) \\
n =& 753, R^2 = .264
\end{aligned}
$$

See MROZ.dta.

# Linear probability model (LPM): regression in Stata

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

| Source | SS | df | MS | | | |
|---|---|---|---|---|---|---|
| Model | 48.8080578 | 7 | 6.97257969 | Number of obs | = | 753 |
| Residual | 135.919698 | 745 | .182442547 | F(7, 745) | = | 38.22 |
| | | | | Prob > F | = | 0.0000 |
| | | | | R-squared | = | 0.2642 |
| | | | | Adj R-squared | = | 0.2573 |
| Total | 184.727756 | 752 | .245648611 | Root MSE | = | .42713 |

| inlf | Coef. | 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 |

Any new idea?

# Take Away

- ✓ A dummy variable could be defined to distinguish between two groups, and the coefficient estimate on the dummy variable estimates the ceteris paribus difference between the two groups
- ✓ Dummy variables are also useful for incorporating ordinal information, such as a credit or a beauty rating
- ✓ Dummy variables can be interacted with quantitative variables to allow slope differences across different groups

One more thing

# China Household Survey Data (CHFS)

- ✓ Part One: Demographic Characteristics
- ✓ Part Two: Assets and Debts
- ✓ Part Three: Insurance and Security
- ✓ Part Four: Expenditures and Income
- ✓ Part Five: Financial Knowledge, Local Governance and Subject Evaluation

# Part One: Demographic Characteristics

- ✓ Basic Family Member Information: age, edu, martial status, family size,...
- ✓ Work & Income Information of Household Member

# Part Two: Assets and Debts

- ✓ Non-Financial Assets: Production and Operation; Housing and Land; Vehicles
- ✓ Financial Assets: Demand Deposits; Deposits; Stocks; Funds; Financial Products; Bonds; Derivatives; Non-RMB Assets; Precious Metal; Other Financial Assets; Cash; Lent-out Money
- ✓ Debts

# Part Three: Insurance and Security

- ✓ Social Security
- ✓ Commercial Insurance

# Part Four: Expenditures and Income

- ✓ Nonproductive Expenditures
- ✓ Transfer Expenditures
- ✓ Other Expenditures
- ✓ Transfer Income
- ✓ Other Income

# Part Five: Financial Knowledge, Local Governance and Subject Evaluation

- ✓ Financial Knowledge
- ✓ Local Governance
- ✓ Environment Protection
- ✓ Tax
- ✓ Birth
- ✓ Exposition to Financial Crime
- ✓ Voluntary Service

# A Real-time Survey in China about Covid-19

- the South China Morning Post: the first confirmed case in China can be traced back to November 17, 2019
- The survey used in our paper started on February 12, 2020, about six weeks after the new coronavirus was identified

# A Real-time Survey in China about Covid-19

- The survey conducted by the Survey and Research Center for China Household Finance
- Detailed information on Chinese households on topics including household demographics, the impact of COVID-19 on salaried employees and business owners, household investment in financial markets, household income and consumption, and household perception and expectations of the economy
- Merge this new data with the latest wave of the China Household Finance Survey conducted in 2019

# A Real-time Survey in China about Covid-19

The survey consists of multiple sections that include detailed information about Chinese households.

- Section A: demographic information.
- Section B: the impact of COVID-19 on salaried employees.
- Section C: the impact of COVID-19 on business owners.
- Section D: household investment in financial markets.
- Section E: household income and consumption.
- Section F: household perception and expectations of the economy.

# A Real-time Survey in China about Covid-19

- The survey was conducted in two consecutive periods with different households.
- The first questionnaire was completed between February 12, 2020 and March 11, 2020, that is Period 1.
- Then, a revised version of the initial questionnaire was completed between March 12, 2020 and March 22, 2020, that is Period 2.

# A Real-time Survey in China about Covid-19

- Period 1: 2,367 responses
- Period 2: 1,186 responses
- the Total: 3,553
- 88% of this total sample consist of people who were surveyed in the last wave of the China Household Finance Survey (CHFS) in 2019
- To see the impact of COVID-19 on household-owned businesses, our final sample includes ONLY the households that have their own businesses.
- This final dataset includes 304 observations which corresponds to 8.6% of the households who responded to the survey.

Thank You!