Econometrics Lecture 4

EC2METRIE

Dr. Anna Salomons

Utrecht School of Economics (U.S.E.)

5 December 2016

This class

- **Functional form**
  - Rescaling independent & dependent variables
  - Variables in logs vs in levels
  - Quadratic terms
  - Dummy variables[1]
  - Interaction effects
  - Chow test
- **Studenmund Ch 7**, excluding section 7.2.5 (note: lagged independent variables are discussed in week 6)

[1] This week, we only consider dummy variables as independent variables- in week 8, the dummy will be the dependent variable.

Rescaling an independent variable

- Original model

$$Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i} + e_i$$

- Now we **rescale the independent variable** X_1, e.g. multiplying it by 10. The original model can be rewritten as;

$$Y_i = \widehat{\beta}_0 + \left(\widehat{\beta}_1 \times 0.1\right)(X_{1i} \times 10) + \widehat{\beta}_2 X_{2i} + e_i$$

Rescaling an independent variable

$$Y_i = \widehat{\beta}_0 + \left(\widehat{\beta}_1 \times 0.1\right)(X_{1i} \times 10) + \widehat{\beta}_2 X_{2i} + e_i$$

- Note the following:
  - The estimated coefficient on X_1 (the rescaled variable) is divided by 10, $\widehat{\beta}_1 \times 0.1$
  - The estimated variance of the coefficient on X_1 is divided by 100: $\widehat{Var}(\widehat{\beta}_1 \times 0.1) = 0.1^2 \times \widehat{Var}(\widehat{\beta}_1)$
  - The standard error of the coefficient on X_1 is divided by 10: $se(\beta_1 \times 0.1) = \sqrt{0.1^2 \times \widehat{Var}(\widehat{\beta}_1)} = 0.1 \times se(\widehat{\beta}_1)$

Rescaling an independent variable

$$Y_i = \widehat{\beta}_0 + \left(\widehat{\beta}_1 \times 0.1\right)(X_{1i} \times 10) + \widehat{\beta}_2 X_{2i} + e_i$$

- Note the following:
  - The t-statistic of the coefficient on X_1 is unaffected: $\frac{\widehat{\beta}_1 \times 0.1}{0.1 \times se(\widehat{\beta}_1)} = \frac{\widehat{\beta}_1}{se(\widehat{\beta}_1)}$
  - The residuals are unaffected, hence so is Root MSE $\widehat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-k-1}}$
  - No effect on R^2

Rescaling an independent variable: some examples

- Annual income in Euros or thousands of Euros
- Working experience in years or in months
- Probabilities (0-1) or percentages (0%-100%)

Example: multiplying the independent variable by 12

```
. descr wage educ age

              storage   display     value
variable name   type    format      label      variable label
-------------------------------------------------------------------------------
wage            double  %10.0g                 earnings per hour
educ            byte    %8.0g                  years of education
age             byte    %8.0g                  age in years

. reg wage educ age

      Source |       SS       df       MS              Number of obs =    4838
-------------+------------------------------           F(  2,  4835) =  645.71
       Model |  160178.442     2  80089.2212           Prob > F      =  0.0000
    Residual |  599696.267  4835   124.03232           R-squared     =  0.2108
-------------+------------------------------           Adj R-squared =  0.2105
       Total |  759874.709  4837  157.096281           Root MSE      =  11.137

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.967164   .0586426    33.54   0.000     1.852197     2.08213
         age |   .1430452   .0127365    11.23   0.000     .1180758    .1680146
       _cons |  -13.20376     .96771   -13.64   0.000    -15.10091   -11.30661
------------------------------------------------------------------------------

. gen educ_months=educ*12

. reg wage educ_months age

      Source |       SS       df       MS              Number of obs =    4838
-------------+------------------------------           F(  2,  4835) =  645.71
       Model |  160178.442     2  80089.2212           Prob > F      =  0.0000
    Residual |  599696.267  4835   124.03232           R-squared     =  0.2108
-------------+------------------------------           Adj R-squared =  0.2105
       Total |  759874.709  4837  157.096281           Root MSE      =  11.137

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 educ_months |   .1639303   .0048869    33.54   0.000     .1543498    .1735108
         age |   .1430452   .0127365    11.23   0.000     .1180758    .1680146
       _cons |  -13.20376     .96771   -13.64   0.000    -15.10091   -11.30661
------------------------------------------------------------------------------
```

Rescaling the dependent variable

- Original model

$$Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i} + e_i$$

- Now we **rescale the dependent variable** Y, e.g. multiplying it by 10. The original model can be rewritten as;

$$Y_i \times 10 = \left(\widehat{\beta}_0 \times 10\right) + \left(\widehat{\beta}_1 \times 10\right) X_{1i} + \left(\widehat{\beta}_2 \times 10\right) X_{2i} + (e_i \times 10)$$

Rescaling the dependent variable

$$Y_i \times 10 = \left(\widehat{\beta}_0 \times 10\right) + \left(\widehat{\beta}_1 \times 10\right) X_{1i} + \left(\widehat{\beta}_2 \times 10\right) X_{2i} + (e_i \times 10)$$

- Note the following:
  - The estimated coefficients are all multiplied by 10, $\widehat{\beta}_j \times 10$
  - The estimated variance of all coefficients is multiplied by 100: $\widehat{Var}(\widehat{\beta}_j \times 10) = 10^2 \times \widehat{Var}(\widehat{\beta}_j)$
  - The standard errors of all estimated coefficients are multiplied by 10: $se(\widehat{\beta}_j \times 10) = \sqrt{10^2 \times \widehat{Var}(\widehat{\beta}_j)} = 10 \times se(\widehat{\beta}_j)$

Rescaling the dependent variable

$$Y_i \times 10 = \left(\widehat{\beta}_0 \times 10\right) + \left(\widehat{\beta}_1 \times 10\right) X_{1i} + \left(\widehat{\beta}_2 \times 10\right) X_{2i} + (e_i \times 10)$$

- Note the following:
  - The t-statistics are unaffected: $\frac{\widehat{\beta}_j \times 10}{se(\widehat{\beta}_j) \times 10} = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)}$
  - The residuals are multiplied by 10, hence so is Root MSE $\sqrt{\frac{\sum (e_i \times 10)^2}{n-k-1}} = \widehat{\sigma} \times 10$
  - No effect on R^2 since $R^2 = 1 - \frac{10^2 \sum e_i^2}{10^2 \sum (y_i - \overline{y})^2} = 1 - \frac{\sum e_i^2}{\sum (y_i - \overline{y})^2}$

Example: multiplying the dependent variable by 40

```
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------
wage            double %10.0g                 earnings per hour

. gen wage_weekly=wage*40

. reg wage educ age

      Source |       SS       df       MS              Number of obs =    4838
-------------+------------------------------           F(  2,  4835) =  645.71
       Model |  160178.442     2  80089.2212           Prob > F      =  0.0000
    Residual |  599696.267  4835   124.03232           R-squared     =  0.2108
-------------+------------------------------           Adj R-squared =  0.2105
       Total |  759874.709  4837  157.096281           Root MSE      =  11.137

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.967164   .0586426    33.54   0.000     1.852197     2.08213
         age |   .1430452   .0127365    11.23   0.000     .1180758    .1680146
       _cons |  -13.20376     .96771   -13.64   0.000    -15.10091   -11.30661
------------------------------------------------------------------------------

. reg wage_weekly educ age

      Source |       SS       df       MS              Number of obs =    4838
-------------+------------------------------           F(  2,  4835) =  645.71
       Model |   256285504     2   128142752           Prob > F      =  0.0000
    Residual |   959514011  4835  198451.709           R-squared     =  0.2108
-------------+------------------------------           Adj R-squared =  0.2105
       Total |  1.2158e+09  4837  251354.045           Root MSE      =  445.48

------------------------------------------------------------------------------
 wage_weekly |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   78.68654   2.345705    33.54   0.000     74.08789    83.28519
         age |   5.721807   .5094615    11.23   0.000     4.723031    6.720583
       _cons |  -528.1504    38.7084   -13.64   0.000    -604.0364   -452.2643
------------------------------------------------------------------------------
```

4 different functional forms

Example: the relationship between smokers' income and cigarette consumption.

1. **Level-level** specification

$$cigs_i = \beta_0 + \beta_1 income_i + \varepsilon_i$$

2. **Log-log** specification (double log)

$$\ln cigs_i = \beta_0 + \beta_1 \ln income_i + \varepsilon_i$$

3. **Log-level** specification (semi log)

$$\ln cigs_i = \beta_0 + \beta_1 income_i + \varepsilon_i$$

4. **Level-log** specification (semi log)

$$cigs_i = \beta_0 + \beta_1 \ln income_i + \varepsilon_i$$

Level-level specification

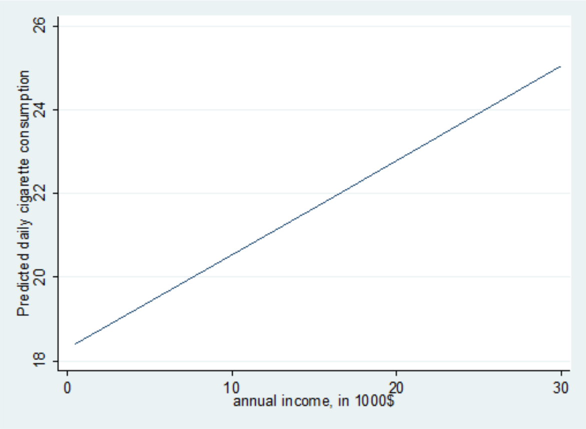
This is what we have seen in previous weeks. Example:

$$cigs_i = \beta_0 + \beta_1 income_i + \varepsilon_i$$

```
. descr cigs income

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------
cigs            byte   %8.0g                  cigs. smoked per day
income          float  %8.0g                  annual income, in 1000$

. sum cigs income

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        cigs |       310     22.6129    13.23543          1         80
      income |       310    19.25645    9.101791         .5         30

. reg cigs income

      Source |       SS       df       MS              Number of obs =     310
-------------+------------------------------           F(  1,   308) =    7.45
       Model |  1278.26682     1  1278.26682           Prob > F      =  0.0067
    Residual |  52851.2816   308   171.59507           R-squared     =  0.0236
-------------+------------------------------           Adj R-squared =  0.0204
       Total |  54129.5484   309  175.176532           Root MSE      =  13.099

------------------------------------------------------------------------------
        cigs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .2234625   .0818741     2.73   0.007     .0623593    .3845658
       _cons |   18.30981   1.743334    10.50   0.000     14.87946    21.74016
------------------------------------------------------------------------------
```

Interpretation: smokers who earn $1000 more per year smoke 0.22 cigarettes more per day.

Estimated shape of the relationship between income and cigarette consumption



Log-log specification

- In a log-log specification, the coefficient gives an **elasticity.**
- **Example:**

$$\ln cigs_i = \beta_0 + \beta_1 \ln income_i + \varepsilon_i$$

- The **income elasticity of cigarette consumption** can then be calculated as:

$$\eta_{income} = \frac{\%\Delta cigs}{\%\Delta income} = \frac{\partial \ln cigs}{\partial \ln income} = \beta_1$$

Log-log specification

$$\eta_{income} = \frac{\partial \ln \mathit{cigs}}{\partial \ln \mathit{income}}$$

Proof: first recognize that

$$\begin{aligned}
\frac{\partial \ln \mathit{cigs}}{\partial \mathit{cigs}} &= \frac{1}{\mathit{cigs}} \Leftrightarrow \partial \ln \mathit{cigs} = \frac{\partial \mathit{cigs}}{\mathit{cigs}} \\
\frac{\partial \ln \mathit{income}}{\partial \mathit{income}} &= \frac{1}{\mathit{income}} \Leftrightarrow \partial \ln \mathit{income} = \frac{\partial \mathit{income}}{\mathit{income}}
\end{aligned}$$

Such that

$$\begin{aligned}
\frac{\partial \ln \mathit{cigs}}{\partial \ln \mathit{income}} &= \frac{\partial \mathit{cigs}}{\mathit{cigs}} \frac{\mathit{income}}{\partial \mathit{income}} \\
&= \frac{\partial \mathit{cigs}}{\partial \mathit{income}} \frac{\mathit{income}}{\mathit{cigs}} \equiv \eta_{income}
\end{aligned}$$

Log-log specification

variable name	storage type	display format	value label	variable label
cigs	byte	%8.0g		cigs. smoked per day
income	float	%8.0g		annual income, in 1000$
lcigs	float	%9.0g		log(cigs)
lincome	float	%9.0g		log(income)

```
. sum cigs income lcigs lincome
```

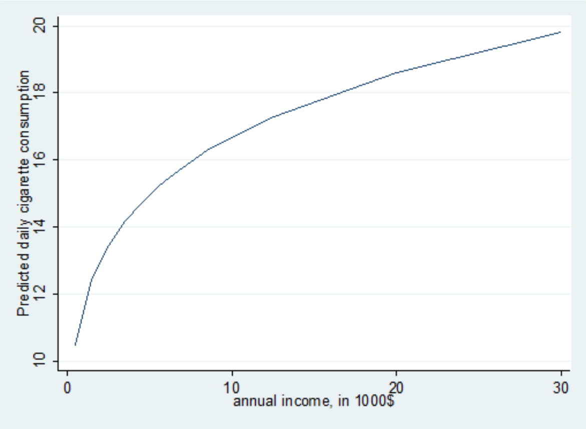
Variable	Obs	Mean	Std. Dev.	Min	Max
cigs	310	22.6129	13.23543	1	80
income	310	19.25645	9.101791	.5	30
lcigs	310	2.890992	.7933564	0	4.382027
lincome	310	2.786507	.6776821	-.6931472	3.401197

Log-log specification

```
. reg lcigs lincome

      Source |       SS       df       MS              Number of obs =     310
-------------+------------------------------           F(  1,   308) =    5.59
       Model |  3.46780403     1  3.46780403           Prob > F      =  0.0187
    Residual |  191.021258   308  .620198889           R-squared     =  0.0178
-------------+------------------------------           Adj R-squared =  0.0146
       Total |  194.489062   309   .62941444           Root MSE      =  .78753

------------------------------------------------------------------------------
       lcigs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lincome |   .1563227   .0661089     2.36   0.019     .0262404     .286405
       _cons |   2.455397   .1895655    12.95   0.000      2.08239    2.828405
------------------------------------------------------------------------------
```

Interpretation: when income increases with 1%, smokers smoke 0.16% more cigarettes per day. I.e., the income elasticity of cigarette consumption is 0.16 for smokers.

Estimated shape of the relationship between income and cigarette consumption



Last week's tutorial exercise

- Compare this to last week's tutorial exercise, calculating the own-price **point elasticity** of chicken consumption from a level-level specification.
  - The point elasticity is different for each point of the demand curve
  - That is, it depends on the price of chicken (PC) and per capita chicken consumption (Y)
- A log-log specification, on the other hand, estimates a **constant elasticity.**

Last week's tutorial exercise: calculating a point elasticity

variable name	storage type	display format	value label	variable label
y	float	%9.0g		per capita chicken consumption
pc	float	%9.0g		price of chicken
pb	float	%9.0g		price of beef
yd	float	%9.0g		disposable income

```
. reg y pc pb yd
```

Source	SS	df	MS
Model	14745.7283	3	4915.24278
Residual	143.072565	36	3.97423792
Total	14888.8009	39	381.764125

Number of obs	40
F(3, 36)	1236.78
Prob > F	0.0000
R-squared	0.9904
Adj R-squared	0.9896
Root MSE	1.9935

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pc	-.60716	.1571203	-3.86	0.000	-.9258147	-.2885054
pb	.0921878	.039883	2.31	0.027	.0113012	.1730743
yd	.2448599	.0110954	22.07	0.000	.2223574	.2673624
_cons	27.59394	1.584457	17.42	0.000	24.38051	30.80737

Last week's tutorial exercise: calculating a point elasticity

```
. sum y pc
```

Variable	Obs	Mean	Std. Dev.	Min	Max
y	40	50.56725	19.53879	23.52	88.87
pc	40	10.24	2.464809	6.5	15.9

- We calculated the average **point elasticity** $\overline{\eta}_{own}$:

$$\eta_{own} = \frac{\partial Y}{\partial PC}\frac{PC}{Y}$$

$$\overline{\eta}_{own} = \frac{\partial Y}{\partial PC}\frac{\overline{PC}}{\overline{Y}} = -0.61 \times \frac{10.24}{50.57} = -0.12$$

Last week's tutorial exercise: estimating a constant elasticity

```
. gen ly=log(y)

. gen lpc=log(pc)

. gen lpb=log(pb)

. gen lyd=log(yd)

. reg ly lpc lpb lyd

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  3,    36) =  799.74
       Model |  5.80622913     3  1.93540971           Prob > F      =  0.0000
    Residual |  .087121785    36   .00242005           R-squared     =  0.9852
-------------+------------------------------           Adj R-squared =  0.9840
       Total |  5.89335091    39  .151111562           Root MSE      =  .04919

------------------------------------------------------------------------------
          ly |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lpc |  -.2206263   .0396401    -5.57   0.000    -.3010202   -.1402324
         lpb |  -.0063099   .0614643    -0.10   0.919    -.1309653    .1183455
         lyd |   .4567236   .0358703    12.73   0.000     .3839753    .5294719
       _cons |   2.405279   .0990111    24.29   0.000     2.204475    2.606083
------------------------------------------------------------------------------
```

The estimated constant own price elasticity is -0.22.

Log-level specification

- In a **log-level specification**, the coefficient×100% gives **the percentage change in the dependent variable, for a one unit increase in the level of the independent variable.**
- **Example:**

$$\ln cigs_i = \beta_0 + \beta_1 income_i + \varepsilon_i$$

Log-level specification

- The coefficient gives

$$\beta_1 = \frac{\partial \ln cigs}{\partial income} \approx \frac{\%\Delta cigs / 100}{\partial income}$$

- Proof:

$$\begin{aligned} \Delta \ln cigs &= \ln(cigs + \Delta cigs) - \ln(cigs) \\ &= \ln\left(\frac{cigs + \Delta cigs}{cigs}\right) \\ &= \ln\left(1 + \frac{\Delta cigs}{cigs}\right) \end{aligned}$$

using the approximation that $\ln(1 + x) \approx x$ for $x \approx 0$:

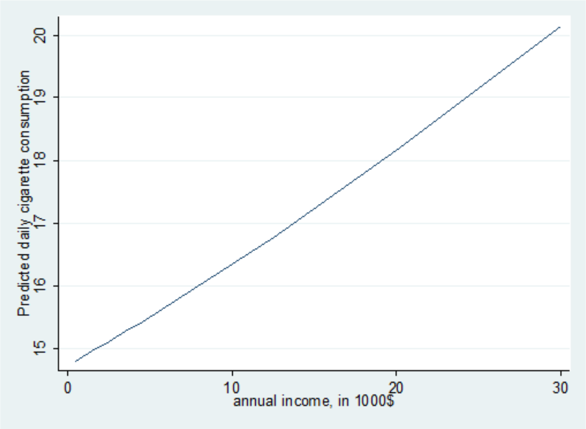
$$\Delta \ln cigs \approx \frac{\Delta cigs}{cigs} = \%\Delta cigs / 100$$

Log-level specification

```
. reg lcigs income

      Source |       SS       df       MS              Number of obs =     310
-------------+------------------------------           F(  1,   308) =    4.51
       Model |  2.80534201     1  2.80534201           Prob > F      =  0.0345
    Residual |   191.68372   308  .62234974            R-squared     =  0.0144
-------------+------------------------------           Adj R-squared =  0.0112
       Total |  194.489062   309  .62941444            Root MSE      =  .78889

------------------------------------------------------------------------------
       lcigs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0104686   .0049307     2.12   0.035     .0007664    .0201707
       _cons |   2.689404   .1049894    25.62   0.000     2.482817    2.895991
------------------------------------------------------------------------------
```

Interpretation: when income increases with $1000, smokers smoke 1.05% ($= 0.0105 \times 100\%$) more cigarettes per day.

Estimated shape of the relationship between income and cigarette consumption



Level-log specification

- In a **level-log specification**, the coefficient/100 gives the impact on the level of the dependent variable from a 1% increase in the independent variable.
- **Example:**

$$cigs_i = \beta_0 + \beta_1 \ln income_i + \varepsilon_i$$

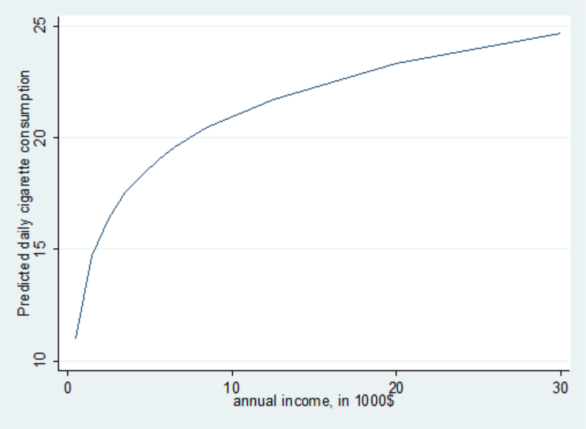
- The reason for this interpretation is that $\Delta \ln income \simeq \frac{\Delta income}{income}$

Level-log specification

```
. reg cigs lincome

      Source |       SS       df       MS              Number of obs =     310
-------------+------------------------------           F(  1,   308) =    9.18
       Model |  1566.29487     1  1566.29487           Prob > F      =  0.0027
    Residual |  52563.2535   308  170.659914           R-squared     =  0.0289
-------------+------------------------------           Adj R-squared =  0.0258
       Total |  54129.5484   309  175.176532           Root MSE      =  13.064

------------------------------------------------------------------------------
        cigs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lincome |   3.322244   1.096631     3.03   0.003     1.164407     5.48008
       _cons |   13.35545   3.144558     4.25   0.000     7.167913    19.54298
------------------------------------------------------------------------------
```

Interpretation: when income increases with 1%, smokers smoke 0.03 ($= 3.32/100$) more cigarettes per day.

Estimated shape of the relationship between income and cigarette consumption



Comparing models: a word of caution

$$\begin{aligned} cigs_i &= \beta_0 + \beta_1 income_i + \varepsilon_i & (1) \\ \ln cigs_i &= \beta_0 + \beta_1 \ln income_i + \varepsilon_i & (2) \\ \ln cigs_i &= \beta_0 + \beta_1 income_i + \varepsilon_i & (3) \\ cigs_i &= \beta_0 + \beta_1 \ln income_i + \varepsilon_i & (4) \end{aligned}$$

- You cannot compare R^2 or $\overline{R}^2$ (or root MSE) across models with different dependent variables! (I.e. we can only compare model 1 with model 4 in this way, as well as model 2 with model 3.)

Comparing models: a word of caution

- This is because R^2, $\overline{R}^2$ and root MSE measure the (un)explained variation in the dependent variable, but different dependent variables (logs or levels) implies that variation is different.
- **Rely on economic reasoning** (and which hypothesis you are interested in testing) to decide which model you prefer.

Linear and quadratic terms

Population regression model:

$$\ln wage_i = \beta_0 + \beta_1 educ_i + \beta_2 age_i + \beta_3 age_i^2 + \varepsilon_i$$

- **Partial effect of education on log wages**= marginal or ceteris paribus effect of one more year of education on the log wage, holding constant age.
- This can be found by taking the first order partial derivative of the equation with respect to *educ*:

$$\frac{\partial \ln wage_i}{\partial \ln educ_i} = \beta_1$$

Linear and quadratic terms

Population regression model:

$$\ln wage_i = \beta_0 + \beta_1 educ_i + \beta_2 age_i + \beta_3 age_i^2 + \varepsilon_i$$

- **Partial effect of age on log wages**= marginal or ceteris paribus effect of one more year of age on the log wage, holding constant education.
- Again, taking the first order derivative to find this:

$$\frac{\partial \ln wage_i}{\partial \ln age_i} = \beta_2 + 2\beta_3 age_i$$

- We see that the marginal effect of age on log wages depends on age.

Linear and quadratic terms

$$\frac{\partial \ln wage_i}{\partial \ln age_i} = \beta_2 + 2\beta_3 age_i$$

- We can describe this effect by **finding the stationary point** and classifying it as a **minimum or maximum**.
- To find the stationary point, set the first derivative equal to zero, and solve for age:

$$\begin{aligned} \frac{\partial \ln wage_i}{\partial \ln age_i} &= \beta_2 + 2\beta_3 age^* = 0 \\ \Leftrightarrow \quad age^* &= -\frac{\beta_2}{2\beta_3} \end{aligned}$$

Linear and quadratic terms

- Classify the stationary point by looking at the sign of the second derivative:

$$\begin{aligned} \frac{\partial^2 \ln wage_i}{\partial \ln age_i^2} &= 2\beta_3 \\ \beta_3 &> 0 \Leftrightarrow \text{min} \\ \beta_3 &< 0 \Leftrightarrow \text{max} \end{aligned}$$

- For an example, see the last exercise of last week's tutorial.

Dummy variable: definition

Dummy variable: can take on two values only, 0 and 1.

- Examples:
  - Dummy for female gender: $female_i = 1$ if the respondent is female; $female_i = 0$ if male
  - Dummy for male gender: $male_i = 1$ if the respondent is male; $male_i = 0$ if female
- Note that
  - For each individual in the sample it holds that $female_i + male_i = 1$
  - $\overline{female}$ = the fraction of women in the sample; $\overline{male}$ = the fraction of men in the sample
  - $\overline{female} + \overline{male} = 1$

Bivariate regression

$$\ln wage_i = \beta_0 + \beta_1 female_i + \varepsilon_i$$

Under OLS assumptions we can write:

- Average log wage for $female_i = 1$:

$$E(\ln wage_i | female_i = 1) = \beta_0 + \beta_1$$

- Average log wage for $female_i = 0$:

$$E(\ln wage_i | female_i = 0) = \beta_0$$

$female_i = 0$ is called the **reference group**, i.e. the group against which the wage comparison is made.

Bivariate regression

To see why:

- Average log wage for $female_i = 1$:

$$E(\ln wage_i | female_i = 1) = E\left[(\beta_0 + \beta_1 female_i + \varepsilon_i) | female_i = 1\right]$$

$$\begin{aligned} &= \beta_0 + E(\beta_1 female_i | female_i = 1) + E(\varepsilon_i | female_i = 1) \\ &= \beta_0 + \beta_1 + E(\varepsilon_i | female_i = 1) = \beta_0 + \beta_1 \end{aligned}$$

- Similarly, average log wage for $female_i = 0$:

$$E(\ln wage_i | female_i = 0) = E\left[(\beta_0 + \beta_1 female_i + \varepsilon_i) | female_i = 0\right]$$

$$\begin{aligned} &= \beta_0 + E(\beta_1 female_i | female_i = 0) + E(\varepsilon_i | female_i = 0) \\ &= \beta_0 + E(\varepsilon_i | female_i = 0) = \beta_0 \end{aligned}$$

[1] OLS assumptions for unbiasedness give that: $E(\varepsilon_i | female_i = 1) = E(\varepsilon_i | female_i = 0) = 0$

Bivariate regression

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      female |      4838    .5049607    .5000271          0          1
        male |      4838    .4950393    .5000271          0          1

. reg lwage female

      Source |       SS       df       MS              Number of obs =    4838
-------------+------------------------------           F(  1,  4836) =  162.18
       Model |  51.7583589     1  51.7583589           Prob > F      =  0.0000
    Residual |  1543.36023  4836  .319139832           R-squared     =  0.0324
-------------+------------------------------           Adj R-squared =  0.0322
       Total |  1595.11859  4837  .329774361           Root MSE      =  .56492

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2068753   .0162446   -12.74   0.000    -.2387221   -.1750285
       _cons |   2.941271   .0115435   254.80   0.000      2.91864    2.963901
------------------------------------------------------------------------------
```

Interpretation: women earn 20.7% lower wages than men. (Note that we can interpret this coefficient since it's statistically significant.)

Multivariate regression

$$\begin{aligned}
\ln wage_i &= \beta_0 + \beta_1 female_i + \beta_2 educ_i + \varepsilon_i \\
\text{Under OLS assumptions} &: \\
E(\ln wage_i | educ, female_i) &= \beta_0 + \beta_1 female_i + \beta_2 educ_i
\end{aligned}$$

- Average log wage for $female_i = 1$:

$$E(\ln wage_i | female_i = 1) = \beta_0 + \beta_1 + \beta_2 educ_i$$

- Average log wage for $female_i = 0$:

$$E(\ln wage_i | female_i = 0) = \beta_0 + \beta_2 educ_i$$

- Average log wage difference between women and men:

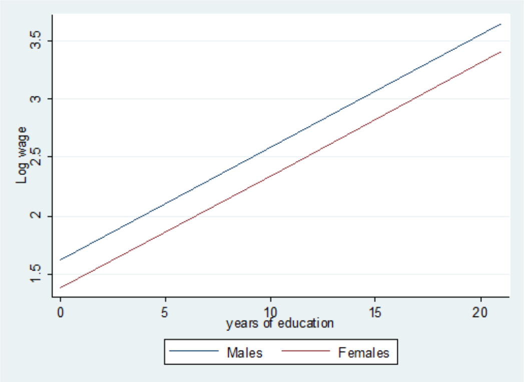
$$E(\ln wage_i | female_i = 1) - E(\ln wage_i | female_i = 0) = \beta_1$$

Multivariate regression

```
. reg lwage female educ

      Source |       SS           df       MS      Number of obs   =      4838
-------------+----------------------------------   F(  2,  4835)   =    768.30
       Model |  384.683739         2  192.34187   Prob > F        =    0.0000
    Residual |  1210.43485      4835  .250348469   R-squared       =    0.2412
-------------+----------------------------------   Adj R-squared   =    0.2408
       Total |  1595.11859      4837  .329774361   Root MSE        =    .50035

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2398371    .014416   -16.64   0.000    -.2680991    -.211575
        educ |   .0961508   .0026366    36.47   0.000     .0909817    .1013198
       _cons |   1.626193   .0374833    43.38   0.000     1.552709    1.699678
------------------------------------------------------------------------------
```

Interpretation: women earn 24% less than men, holding constant the education level. (Note that we can interpret this coefficient since it's statistically significant.)

Dummy changes the intercept



Men and women have a difference log wage intercept (but identical slopes, given by the coefficient on educ).

Dummy variable trap and perfect collinearity

Why is the following model incorrect:

$$\ln wage_i = \beta_0 + \beta_1 female_i + \beta_2 male_i + \varepsilon_i$$

Perfectly collinearity:

- $female_i$ and $male_i$ are perfect linear functions of each other, in particular, for each individual observation $female_i + male_i = 1$.
- Hence, one of the OLS assumptions for unbiasedness is violated.
- This is known as the **dummy variable trap:** cannot include a full set of dummies, we need an **omitted category** which serves as the reference category.

Dummy variable trap and perfect collinearity

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

      Source |       SS       df       MS              Number of obs =    4838
-------------+------------------------------           F(  2,  4835) =  768.30
       Model |  384.683739     2  192.34187           Prob > F      =  0.0000
    Residual |  1210.43485  4835  .250348469           R-squared     =  0.2412
-------------+------------------------------           Adj R-squared =  0.2408
       Total |  1595.11859  4837  .329774361           Root MSE      =  .50035

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2398371    .014416   -16.64   0.000    -.2680991    -.211575
        male |  (omitted)
        educ |   .0961508   .0026366    36.47   0.000     .0909817    .1013198
       _cons |   1.626193   .0374833    43.38   0.000     1.552709    1.699678
------------------------------------------------------------------------------
```

If you make this mistake, Stata will automatically omit one of the categories for you. The next two slides show that it does not matter which category you decide to exclude.

Omitted category (=reference group): men

```
. reg lwage female educ

      Source |       SS       df       MS              Number of obs =    4838
-------------+------------------------------           F(  2,  4835) =  768.30
       Model |  384.683739     2  192.34187           Prob > F      =  0.0000
    Residual |  1210.43485  4835  .250348469          R-squared     =  0.2412
-------------+------------------------------           Adj R-squared =  0.2408
       Total |  1595.11859  4837  .329774361          Root MSE      =  .50035

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2398371    .014416   -16.64   0.000    -.2680991   -.211575
        educ |   .0961508   .0026366    36.47   0.000     .0909817    .1013198
       _cons |   1.626193   .0374833    43.38   0.000     1.552709    1.699678
------------------------------------------------------------------------------
```

Interpretation: women earn 24% lower wages than men, cet. par. on education.

Omitted category (=reference group): women

```
. reg lwage male educ

      Source |       SS       df       MS              Number of obs =    4838
-------------+------------------------------           F(  2,  4835) =  768.30
       Model |  384.683739     2  192.34187            Prob > F      =  0.0000
    Residual |  1210.43485  4835 .250348469            R-squared     =  0.2412
-------------+------------------------------           Adj R-squared =  0.2408
       Total |  1595.11859  4837 .329774361            Root MSE      =  .50035

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .2398371    .014416    16.64   0.000      .211575    .2680991
        educ |   .0961508   .0026366    36.47   0.000     .0909817    .1013198
       _cons |   1.386356   .0383269    36.17   0.000     1.311218    1.461495
------------------------------------------------------------------------------
```

Interpretation: men earn 24% higher wages than women, cet. par. on education.

Different types of dummy variables

- Dummy variables for **2 groups**: e.g. gender (male or female), married (married or not married).
- Dummy variables for **>2 groups**
  - Dummies for **categorical variables**: cannot be ranked (e.g. race, industry or region)
  - Dummies for **ordinal variables**: can be ranked (e.g. degree of religiousness, degree of happiness,...)

Dummy variables for multiple groups

Examples:

- values of *region*:
  - 1 for north
  - 2 for east
  - 3 for west
  - 4 for south
- values of *industry*:
  - 1 for manufacturing
  - 2 for services
  - 3 for utilities
  - 4 for public sector
- values of *race*:
  - 1 for black
  - 2 for asian
  - 3 for white
- *Region*, *industry*, and *race* are all **categorical variables**: their **values cannot be ranked**

Dummy variables for multiple groups

Examples:

- values of *happiness*:
  - 1 for very unhappy
  - 2 for somewhat unhappy
  - 3 for somewhat happy
  - 4 for very happy
- values of *agegroup*:
  - 1 for 18<age<34
  - 2 for 34<age<54
  - 3 for 55<age<65
- values of *religiousness*:
  - 1 for not religious
  - 2 for somewhat religious
  - 3 for very religious
- *Happiness*, *agegroup*, and *religiousness* are all **ordinal variables**: their **values can be ranked**

Inclusion of categorical variables

Do not include categorical variables directly into the regression equation.

- **Wrong specification**:

$$\ln wage_i = \beta_0 + \beta_1 female_i + \beta_2 educ_i + \beta_3 race_i + \varepsilon_i$$

  The coefficient on race **cannot be interpreted** ("when race increases with 1" makes no sense since the categories of race cannot be ranked).

- **Correct procedure:**
  - Create a separate dummy variable for each of the k categories of race
  - Include $k-1$ of those dummies into the equation (not k, due to dummy variable trap!)

Inclusion of categorical variables

- **Race has three values**: 1 for black, 2 for asian, 3 for white.
- Create **three dummies**:
  - *gen black=1 if race==1*
  - *replace black=0 if race!=1*
  - *gen asian=1 if race==2*
  - *replace asian=0 if race!=2*
  - *gen white=1 if race==3*
  - *replace white=0 if race!=3*
- In Stata, we can also create all dummies in one go by using *tab race, gen(drace)*
  This creates 3 separate dummy variables, drace1 drace2 and drace3.

Inclusion of categorical variables

$$\ln wage_i = \beta_0 + \beta_1 female_i + \beta_2 educ_i + \beta_3 black_i + \beta_4 asian_i + \varepsilon_i$$

```
. reg lwage female educ black asian

      Source |       SS       df       MS              Number of obs =    4838
-------------+------------------------------           F(  4,  4833) =  395.29
       Model |  393.211516     4  98.302879           Prob > F      =  0.0000
    Residual |  1201.90707  4833  .248687579           R-squared     =  0.2465
-------------+------------------------------           Adj R-squared =  0.2459
       Total |  1595.11859  4837  .329774361           Root MSE      =  .49869

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2361326   .0143826   -16.42   0.000    -.2643291   -.2079361
        educ |   .0953349   .0026343    36.19   0.000     .0901704    .1004994
       black |  -.1394592   .0238211    -5.85   0.000    -.1861594   -.0927589
       asian |  -.0187773   .0338864    -0.55   0.580      -.08521    .0476554
       _cons |   1.650723   .0375929    43.91   0.000     1.577023    1.724422
------------------------------------------------------------------------------
```

Interpretation of categorical variables

$$\ln wage_i = \beta_0 + \beta_1 female_i + \beta_2 educ_i + \beta_3 black_i + \beta_4 asian_i + \varepsilon_i$$

lwage	Coef.	Std. Err.	t	P>|t|
female	-.2361326	.0143826	-16.42	0.000
educ	.0953349	.0026343	36.19	0.000
black	-.1394592	.0238211	-5.85	0.000
asian	-.0187773	.0338864	-0.55	0.580

The omitted category is white: all **interpretations are relative to this omitted category**

- $\widehat{\beta}_3$: Black workers earn on average 14% less than white workers, cet. par. on education and gender.
- $\widehat{\beta}_4$: There is no statistically significant wage difference between asian and white workers, cet. par. on education and gender.

Testing the statistical significance of multiple dummies

$$\ln wage_i = \beta_0 + \beta_1 female_i + \beta_2 educ_i + \beta_3 black_i + \beta_4 asian_i + \varepsilon_i$$

- **Does race have a significant impact on wages**, controlling for education and gender?

$$\begin{aligned} H_0 &: \beta_3 = \beta_4 = 0 \\ H_A &: H_0 \text{ not true} \end{aligned}$$

- Multiple population parameters in H_0, hence need an **F-test**!

Testing the statistical significance of race

```
. reg lwage female educ black asian

      Source |       SS       df       MS              Number of obs =    4838
-------------+------------------------------           F(  4,  4833) =  395.29
       Model |  393.211516     4  98.302879            Prob > F      =  0.0000
    Residual |  1201.90707  4833  .248687579           R-squared     =  0.2465
-------------+------------------------------           Adj R-squared =  0.2459
       Total |  1595.11859  4837  .329774361           Root MSE      =  .49869

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2361326   .0143826   -16.42   0.000    -.2643291   -.2079361
        educ |   .0953349   .0026343    36.19   0.000     .0901704    .1004994
       black |  -.1394592   .0238211    -5.85   0.000    -.1861594   -.0927589
       asian |  -.0187773   .0338864    -0.55   0.580      -.08521    .0476554
       _cons |   1.650723   .0375929    43.91   0.000     1.577023    1.724422
------------------------------------------------------------------------------

. test black asian

 ( 1)  black = 0
 ( 2)  asian = 0

       F(  2,  4833) =   17.15
            Prob > F =    0.0000
```

Conclusion: reject H_0, race has a statistically significant impact on wages, ceteris paribus.

Choice of reference category

- As before, the **choice of reference category is not important.**
  - That is, we can omit *white*, *black* or *asian*.
- This can be shown more formally: parameters for equations with different reference categories can all be found from an equation with any single reference group.

Changing the reference category from white to black

$$\ln wage_i = \beta_0 + \beta_1 fem_i + \beta_2 educ_i + \beta_3 black_i + \beta_4 asian_i + \varepsilon_i$$

Using that
$white_i + black_i + asian_i = 1 \Leftrightarrow black_i = 1 - white_i - asian_i$

$$\begin{aligned}
\ln wage_i &= \left\{ \begin{array}{c} \beta_0 + \beta_1 fem_i + \beta_2 educ_i + \beta_3 (1 - white_i - asian_i) \\ + \beta_4 asian_i + \varepsilon_i \end{array} \right\} \\
\ln wage_i &= \left\{ \begin{array}{c} \beta_0 + \beta_1 fem_i + \beta_2 educ_i + \beta_3 - \beta_3 white_i - \beta_3 asian_i \\ + \beta_4 asian_i + \varepsilon_i \end{array} \right\} \\
\ln wage_i &= \left\{ \begin{array}{c} (\beta_0 + \beta_3) + \beta_1 fem_i + \beta_2 educ_i - \beta_3 white_i \\ + (\beta_4 - \beta_3) asian_i + \varepsilon_i \end{array} \right\}
\end{aligned}$$

Reference category: white vs black

White omitted:

$$\widehat{\ln wage_i} = \widehat{\beta}_0 + \widehat{\beta}_1 fem_i + \widehat{\beta}_2 educ_i + \widehat{\beta}_3 black_i + \widehat{\beta}_4 asian_i$$

$$\widehat{\ln wage_i} = 1.65 - 0.24 fem_i + 0.095 educ_i - 0.14 black_i - 0.02 asian_i$$

Black omitted:

$$\widehat{\ln wage_i} = \left\{ \begin{array}{c} \left(\widehat{\beta}_0 + \widehat{\beta}_3\right) + \widehat{\beta}_1 fem_i + \widehat{\beta}_2 educ_i - \widehat{\beta}_3 white_i \\ + \left(\widehat{\beta}_4 - \widehat{\beta}_3\right) asian_i \end{array} \right\}$$

$$\widehat{\ln wage_i} = 1.51 - 0.24 fem_i + 0.095 educ_i + 0.14 white_i + 0.12 asian_i$$

where

$$\left(\widehat{\beta}_0 + \widehat{\beta}_3\right) = 1.65 - 0.14 = 1.51$$

$$-\widehat{\beta}_3 = 0.14$$

$$\left(\widehat{\beta}_4 - \widehat{\beta}_3\right) = -0.02 + 0.14 = 0.12$$

Choice of reference category: black

```
. reg lwage female educ asian white

      Source |       SS           df       MS      Number of obs   =      4838
-------------+----------------------------------   F(  4,   4833)  =    395.29
       Model |  393.211516         4  98.302879    Prob > F        =    0.0000
    Residual |  1201.90707      4833  .248687579   R-squared       =    0.2465
-------------+----------------------------------   Adj R-squared   =    0.2459
       Total |  1595.11859      4837  .329774361   Root MSE        =    .49869

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2361326   .0143826   -16.42   0.000    -.2643291   -.2079361
        educ |   .0953349   .0026343    36.19   0.000     .0901704    .1004994
       asian |   .1206818    .039982     3.02   0.003      .042299    .1990647
       white |   .1394592   .0238211     5.85   0.000     .0927589    .1861594
       _cons |   1.511263    .042203    35.81   0.000     1.428526    1.594001
------------------------------------------------------------------------------
```

$$\widehat{\ln wage_i} = 1.51 - 0.24 fem_i + 0.095 educ_i + 0.14 white_i + 0.12 asian_i$$

Changing the reference category from white to asian

$$\ln wage = \beta_0 + \beta_1 fem_i + \beta_2 educ_i + \beta_3 black_i + \beta_4 asian_i$$

Using that

$white_i + black_i + asian_i = 1 \Leftrightarrow asian_i = 1 - white_i - black_i$

$$\ln wage_i = \left\{ \begin{array}{c} \beta_0 + \beta_1 fem_i + \beta_2 educ_i + \beta_3 black_i \\ +\beta_4 \left(1 - white_i - black_i\right) + \varepsilon_i \end{array} \right\}$$

$$\ln wage_i = \left\{ \begin{array}{c} \beta_0 + \beta_1 fem_i + \beta_2 educ_i + \beta_3 black_i + \beta_4 - \beta_4 white_i \\ -\beta_4 black_i + \varepsilon_i \end{array} \right\}$$

$$\ln wage_i = \left\{ \begin{array}{c} \left(\beta_0 + \beta_4\right) + \beta_1 fem_i + \beta_2 educ_i + \left(\beta_3 - \beta_4\right) black_i \\ -\beta_4 white_i + \varepsilon_i \end{array} \right\}$$

Reference category: white vs asian

White omitted:

$$\begin{aligned}\widehat{\ln wage_i} &= \widehat{\beta}_0 + \widehat{\beta}_1 fem_i + \widehat{\beta}_2 educ_i + \widehat{\beta}_3 black_i + \widehat{\beta}_4 asian_i \\ \widehat{\ln wage_i} &= 1.65 - 0.24 fem_i + 0.095 educ_i - 0.14 black_i - 0.02 asian_i\end{aligned}$$

Asian omitted:

$$\begin{aligned}\widehat{\ln wage_i} &= \left\{ \begin{array}{c} \left(\widehat{\beta}_0 + \widehat{\beta}_4\right) + \widehat{\beta}_1 fem_i + \widehat{\beta}_2 educ_i + \left(\widehat{\beta}_3 - \widehat{\beta}_4\right) black_i \\ -\widehat{\beta}_4 white_i \end{array} \right\} \\ \widehat{\ln wage_i} &= 1.63 - 0.24 fem_i + 0.095 educ_i - 0.12 black_i + 0.02 white_i\end{aligned}$$

where

$$\begin{aligned}\left(\widehat{\beta}_0 + \widehat{\beta}_4\right) &= 1.65 - -0.02 = 1.63 \\ \left(\widehat{\beta}_3 - \widehat{\beta}_4\right) &= -0.14 - -0.02 = -0.12 \\ -\widehat{\beta}_4 &= 0.02\end{aligned}$$

Choice of reference category: asian

```
. reg lwage female educ black white

      Source |       SS       df       MS              Number of obs =    4838
-------------+------------------------------           F(  4,  4833) =  395.29
       Model |  393.211516     4  98.302879            Prob > F      =  0.0000
    Residual |  1201.90707  4833  .248687579           R-squared     =  0.2465
-------------+------------------------------           Adj R-squared =  0.2459
       Total |  1595.11859  4837  .329774361           Root MSE      =  .49869

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2361326   .0143826   -16.42   0.000    -.2643291   -.2079361
        educ |   .0953349   .0026343    36.19   0.000     .0901704    .1004994
       black |  -.1206818    .039982    -3.02   0.003    -.1990647    -.042299
       white |   .0187773   .0338864     0.55   0.580    -.0476554      .08521
       _cons |   1.631945   .0503872    32.39   0.000     1.533164    1.730727
------------------------------------------------------------------------------
```

$$\widehat{\ln wage_i} = 1.63 - 0.24 fem_i + 0.095 educ_i - 0.12 black_i + 0.02 white_i$$


Official Selection 2015
sundance
film festival
Dreamcatcher

Brenda Meyers-Powell of the Dreamcatcher foundation

Another example of categorical dummies

- The **study of illegal markets**: intersection of many fields (law, economics, sociology, psychology..)
- Markets are illegal when either the product itself (e.g. heroine), the exchange of it for money (e.g. prostitution, human organs), or the way in which it is produced or sold (e.g. counterfeit Rolexes, or production using child labor) violates legal stipulations.
- Example: **prostitution in Mexico**.

Dataset on an illegal market

Obtained from Manisha Shah and Stefano Bertozzi, "Risky Business: The Market for Unprotected Sex", Journal of Political Economy (2005), 113, pp. 518-550.

variable name	variable label
price	Price of the transaction in Mexican pesos
lnprice	log(price) of transaction
attractive	1 if the sex worker is attractive; 0 otherwise
school	1 if sex worker has completed secondary school or higher; 0 otherwise
age	age of sex worker in years
rich	1 if client is rich; 0 otherwise
alcohol	1 if client consumed alcohol prior to the transaction
bar	1 if transaction originated in a bar; 0 otherwise
street	1 if transaction originated in a street; 0 otherwise
othersite	1 if transaction originated in another site; 0 otherwise

Dataset on an illegal market

Variable	Obs	Mean	Std. Dev.	Min	Max
price	3016	449.5823	400.4593	9.999999	5799.999
lnprice	3016	5.839489	.7155389	2.302585	8.665613
attractive	3016	.137931	.3448848	0	1
school	3016	.3169761	.4653752	0	1
age	3016	27.40981	7.729452	12	54
rich	3016	.8428382	.3640136	0	1
alcohol	3016	.846817	.3602236	0	1
bar	3016	.8047082	.3964909	0	1
street	3016	.1747347	.379803	0	1
othersite	3016	.020557	.1419194	0	1

A pricing equation for illegal transactions

Model the transaction price as a function of:

- characteristics of the sex worker (schooling, age, attractiveness);
- characteristics of the customer (rich, alcohol); and
- characteristics of the transaction (transaction place of origin: bar, street or other)

$$\ln price_i = \left\{ \begin{array}{c} \beta_0 + \beta_1 attractive_i + \beta_2 school_i + \beta_3 age_i + \\ \beta_4 rich_i + \beta_5 alcohol_i + \\ \beta_6 bar_i + \beta_7 othersite_i + \varepsilon_i \end{array} \right\}$$

Note that **street is the omitted category**

Estimates

```
. reg lnprice attractive school age rich alcohol bar other
```

Source	SS	df	MS
Model	501.703241	7	71.6718916
Residual	1041.9644	3008	.34639774
Total	1543.66764	3015	.511995901

Number of obs	=	3016
F(7, 3008)	=	206.91
Prob > F	=	0.0000
R-squared	=	0.3250
Adj R-squared	=	0.3234
Root MSE	=	.58856

lnprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attractive	.2394121	.0315921	7.58	0.000	.1774678	.3013563
school	.1637754	.0238151	6.88	0.000	.11708	.2104709
age	-.0210136	.0014531	-14.46	0.000	-.0238627	-.0181645
rich	.2924201	.0304404	9.61	0.000	.232734	.3521061
alcohol	.2403329	.0358481	6.70	0.000	.1700436	.3106222
bar	.4781665	.0348945	13.70	0.000	.409747	.5465861
othersite	.2621039	.0793876	3.30	0.001	.1064444	.4177633
_cons	5.49038	.0612582	89.63	0.000	5.370268	5.610492

Estimated equation

$$\ln price_i = \left\{ \begin{array}{c} 5.49 + 0.24 attractive_i + 0.16 school_i - 0.02 age_i + \\ 0.29 rich_i + 0.24 alcohol_i + \\ 0.48 bar_i + 0.26 othersite_i + \varepsilon_i \end{array} \right\}$$

Interpretations:

- $\widehat{\beta}_{bar}$: transactions that originated in bars had a 48% higher price than those that originated in the street, all else equal;
- $\widehat{\beta}_{othersite}$: transactions that originated in other sites (i.e. not in bars or on the street) had a 26% higher price than those that originated in the street, all else equal.

Note that these interpretations are relative to the omitted category "street"; both effects are individually significant which means the prices are significantly different from that of the omitted category.

Location, location, location?

Does location matter for the transaction price, after controlling for characteristics of the sex worker and the customer?

$$\ln price_i = \left\{ \begin{array}{c} \beta_0 + \beta_1 attractive_i + \beta_2 school_i + \beta_3 age_i + \\ \beta_4 rich_i + \beta_5 alcohol_i + \\ \beta_6 bar_i + \beta_7 othersite_i + \varepsilon_i \end{array} \right\}$$

$$\begin{aligned} H_0 &: \beta_6 = \beta_7 = 0 \\ H_A &: H_0 \text{ not true} \end{aligned}$$

Location matters

lnprice	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval
attractive	.2394121	.0315921	7.58	0.000	.1774678 .301356
school	.1637754	.0238151	6.88	0.000	.11708 .210470
age	-.0210136	.0014531	-14.46	0.000	-.0238627 -.018164
rich	.2924201	.0304404	9.61	0.000	.232734 .352106
alcohol	.2403329	.0358481	6.70	0.000	.1700436 .310622
bar	.4781665	.0348945	13.70	0.000	.409747 .546586
othersite	.2621039	.0793876	3.30	0.001	.1064444 .417763
_cons	5.49038	.0612582	89.63	0.000	5.370268 5.61049

```
. test bar other

 ( 1)  bar = 0
 ( 2)  othersite = 0

       F(  2,  3008) =   93.89
            Prob > F =    0.0000
```

H_0 rejected: location of the transaction matters for the price, cet. par.

Inclusion of ordinal variables

- Unlike categorical variables, **ordinal variables can be included directly into the regression equation**. From our affairs example (last week):

$$naffairs_i = \beta_0 + \beta_1 yrsmarried_i + \beta_2 religion_i + \varepsilon_i$$

- But we **can also include separate dummies for all but one group**, i.e.

$$naffairs_i = \left\{ \begin{array}{c} \beta_0 + \beta_1 yrsmarried_i + \beta_2 veryrelig_i + \beta_3 somerelig_i \\ + \beta_4 slightrelig_i + \beta_5 notrelig_i + \varepsilon_i \end{array} \right\}$$

This is a more flexible specification since it allows for a different intercept for each value of the religion variable.

Inclusion of ordinal variable: directly

```
              storage  display     value
variable name   type   format      label      variable label
-----------------------------------------------------------------------------------
relig           byte   %9.0g                  5 = very relig., 4 = somewhat, 3 = slightly, 2 = not
                                                at all, 1 = anti

. reg naffairs yrsmarr relig

      Source |       SS       df       MS              Number of obs =     601
-------------+------------------------------           F(  2,   598) =   22.84
       Model |  463.279031     2  231.639515           Prob > F      =  0.0000
    Residual |  6065.8025    598  10.1434824           R-squared     =  0.0710
-------------+------------------------------           Adj R-squared =  0.0678
       Total |  6529.08153   600  10.8818026           Root MSE      =  3.1849

------------------------------------------------------------------------------
    naffairs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     yrsmarr |   .1357706   .0239144     5.68   0.000     .0888041     .182737
       relig |  -.5496927   .1141186    -4.82   0.000    -.7738147   -.3255708
       _cons |   2.058719   .3889054     5.29   0.000     1.294932    2.822505
------------------------------------------------------------------------------
```

Use a **t-test** to test for the importance of religion on the number of affairs, cet. par. on years of marriage.

Inclusion of separate ordinal dummies (omitted group = anti-religious)

```
. reg naffairs yrsmarr  vryrel smerel slghtrel notrel

      Source |       SS       df       MS              Number of obs =     601
-------------+------------------------------           F(  5,   595) =   10.29
       Model |   519.84179     5  103.968358           Prob > F      =  0.0000
    Residual |  6009.23974   595  10.0995626           R-squared     =  0.0796
-------------+------------------------------           Adj R-squared =  0.0719
       Total |  6529.08153   600  10.8818026           Root MSE      =   3.178

------------------------------------------------------------------------------
    naffairs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     yrsmarr |   .1359918   .0239071     5.69   0.000     .0890393    .1829443
      vryrel |   -2.1628    .6011444    -3.60   0.000    -3.343423   -.982177
      smerel |  -2.038553   .5163265    -3.95   0.000    -3.052597   -1.024509
    slghtrel |  -.7286147    .537735    -1.35   0.176    -1.784704    .3274748
      notrel |  -.9102627   .5215395    -1.75   0.081    -1.934545    .1140194
       _cons |   1.644965   .4874632     3.37   0.001     .6876068    2.602322
------------------------------------------------------------------------------
```

Use an **F-test** to test for the importance of religion on the number of affairs, cet. par. on years of marriage.

Inclusion of ordinal dummies

- **How to decide** whether to include an ordinal variable directly, or as separate dummy variables?
- Can **compare adjusted R^2 across two models**: the one with the highest adjusted R^2 is preferable.
- More practice in the tutorial.

Interaction terms involving a dummy variable

We now consider models with an **interaction term** in X_1 and X_2

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$$

where either X_1 or X_2 is a dummy variable.
Note that whenever we have an interaction term, we should also always include the variables which make up the interaction (here, X_1 and X_2) individually.

Let's look at an example.

A Mincer model with an interaction term

- Consider the following **Mincer model**:

$$\ln wage_i = \beta_0 + \beta_1 age_i + \beta_2 educ_i + \beta_3 fem_i + \varepsilon_i$$

- We can include an **interaction term between gender and education:**

$$\ln wage_i = \beta_0 + \beta_1 age_i + \beta_2 educ_i + \beta_3 fem_i + \beta_4 educ_i \times fem_i + \varepsilon_i$$

  This allows for two different reasons that women earn different wages:

  - Direct effect: **different intercept** (β_3)
  - Indirect effect through education: **different slope** (β_4)

A Mincer model with an interaction term

To see how an interaction term gives both intercept and slope differences, rewrite the model for men and women separately:

- **Model for men** (i.e. filling in $fem_i = 0$):

$$\ln wage_i = \beta_0 + \beta_1 age_i + \beta_2 educ_i + \varepsilon_i$$

- **Model for women** (i.e. filling in $fem_i = 1$):

$$\begin{aligned} \ln wage_i &= \beta_0 + \beta_1 age_i + \beta_2 educ_i + \beta_3 + \beta_4 educ_i + \varepsilon_i \\ &= (\beta_0 + \beta_3) + \beta_1 age_i + (\beta_2 + \beta_4)\, educ_i + \varepsilon_i \end{aligned}$$

- This shows that β_3 reflects the intercept difference and β_4 reflects the slope difference for education, between men and women.

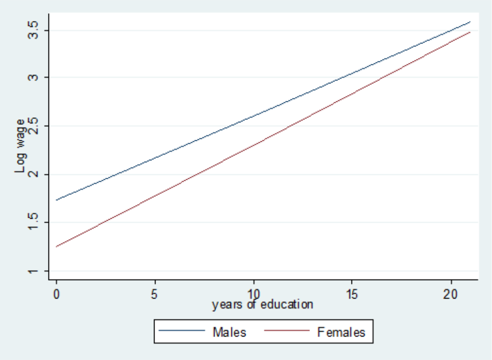
Estimates

```
. gen educ_fem=educ*female

. reg lwage age educ female educ_fem

      Source |       SS       df       MS              Number of obs =    4838
-------------+------------------------------           F(  4,  4833) =  435.48
       Model |  422.597779     4  105.649445           Prob > F      =  0.0000
    Residual |  1172.52081  4833  .242607243           R-squared     =  0.2649
-------------+------------------------------           Adj R-squared =  0.2643
       Total |  1595.11859  4837  .329774361           Root MSE      =  .49255

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0068006   .0005659    12.02   0.000     .0056913    .0079099
        educ |     .08455   .0034573    24.46   0.000      .077772    .0913279
      female |  -.5725595   .0743861    -7.70   0.000    -.7183901   -.4267289
    educ_fem |   .0235023   .0052602     4.47   0.000       .01319    .0338146
       _cons |   1.498829   .0519311    28.86   0.000     1.397021    1.600638
------------------------------------------------------------------------------
```

Interpretation of estimates

- $\widehat{\beta_2} = 0.085$: men earn 8.5% higher wages for each additional year of education, cet. par.
- $\widehat{\beta_3} = -0.57$: women have a 57% lower wage intercept than men (i.e. earn 57% lower wages than men at an age and education of 0)
- $\widehat{\beta_4} = 0.024$: compared to men, women earn 2.4 percentage points higher wages for each additional year of education, cet. par.
- $\widehat{\beta_2} + \widehat{\beta_4} = 0.085 + 0.024 = 0.109$: women earn 10.9% higher wages for each additional year of education, cet. par.

Intercept and slope dummies visualized



$$\widehat{\ln wage_i} = 1.50 + 0.01 age + 0.08 educ_i - 0.57 fem_i + 0.02 educ_i \times fem_i$$

Intercept difference is -0.57; slope difference is 0.02.

A Mincer model with an interaction term

$$\ln wage_i = \beta_0 + \beta_1 age_i + \beta_2 educ_i + \beta_3 fem_i + \beta_4 educ_i \times fem_i + \varepsilon_i$$

Rewriting:

$$\ln wage_i = (\beta_0 + \beta_3 fem_i) + \beta_1 age_i + (\beta_2 + \beta_4 fem_i)\, educ_i + \varepsilon_i$$

- If $\beta_3 = 0$, the intercept is the same for men and women
- If $\beta_4 = 0$, the slope (i.e. return to education) is the same for men and women

A Mincer model with an interaction term

$$\ln wage_i = (\beta_0 + \beta_3 fem_i) + \beta_1 age_i + (\beta_2 + \beta_4 fem_i)\, educ_i + \varepsilon_i$$

Possible hypothesis tests:

- Same intercept (different slopes are allowed): $H_0 : \beta_3 = 0$, $H_A : \beta_3 \neq 0$
- Same slope (different intercepts are allowed): $H_0 : \beta_4 = 0$, $H_A : \beta_4 \neq 0$
- Same intercept *and* same slope: $H_0 : \beta_3 = \beta_4 = 0$, $H_A : H_0$ not true

Hypothesis tests

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0068006	.0005659	12.02	0.000	.0056913	.0079099
educ	.08455	.0034573	24.46	0.000	.077772	.0913279
female	-.5725595	.0743861	-7.70	0.000	-.7183901	-.4267289
educ_fem	.0235023	.0052602	4.47	0.000	.01319	.0338146
_cons	1.498829	.0519311	28.86	0.000	1.397021	1.600638

```
. test female educ_fem

 ( 1)  female = 0
 ( 2)  educ_fem = 0

       F(  2,  4833) =  160.38
            Prob > F =    0.0000
```

$$H_0 : \beta_3 = 0 \rightarrow H_0 \text{ rejected}$$
$$H_0 : \beta_4 = 0 \rightarrow H_0 \text{ rejected}$$
$$H_0 : \beta_3 = \beta_4 = 0 \rightarrow H_0 \text{ rejected}$$

Note: I include this discussion of interaction terms between 2 continuous variables in case you need this for your paper / BSc thesis; I won't ask it on the exam.

Interaction between continuous variables

Model with an **interaction term** in X_1 and X_2 (and neither is a dummy variable)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$$

Marginal effect of X_1 on Y is the first partial derivative of Y wrt X_1

$$\frac{\partial Y_i}{\partial X_{1i}} = \beta_1 + \beta_3 X_{2i}$$

So for each value of X_2 there is a different marginal effect of X_1 on Y : we can show all these different marginal effects in a histogram; or calculate value for the average value of X_2.

Interaction between continuous variables: example

$$wage_i = \beta_0 + \beta_1 age_i + \beta_2 nrkids_i + \beta_3 age_i \times nrkids_i + \varepsilon_i$$

The effect of the number of children someone has on their wages is given by

$$\frac{\partial wage_i}{\partial nrkids_i} = \beta_2 + \beta_3 age_i$$

The average effect is

$$\beta_2 + \beta_3 \overline{age}$$

Interaction between continuous variables: example

variable name	storage type	display format	value label	variable label
wage	double	%10.0g		earnings per hour
age	byte	%8.0g		age in years
nkids	byte	%8.0g		number of children living with

```
. gen age_nkids=age*nkids

. reg wage age nkids age_nkids
```

Source	SS	df	MS
Model	29722.1111	3	9907.37038
Residual	730152.598	4834	151.045221
Total	759874.709	4837	157.096281

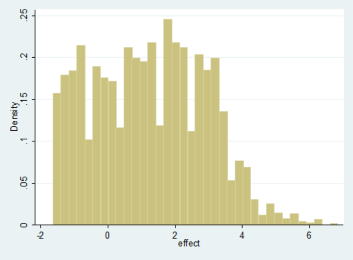
Number of obs	=	4838
F(3, 4834)	=	65.59
Prob > F	=	0.0000
R-squared	=	0.0391
Adj R-squared	=	0.0385
Root MSE	=	12.29

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.1386198	.0154784	8.96	0.000	.1082752	.1689644
nkids	-4.41183	.7807822	-5.65	0.000	-5.942519	-2.881142
age_nkids	.1323203	.0197556	6.70	0.000	.0935904	.1710502
_cons	13.79807	.7164421	19.26	0.000	12.39352	15.20262

Interaction between continuous variables: example

```
. gen effect=-4.41183+.1323203 *age

. sum effect
```

Variable	Obs	Mean	Std. Dev.	Min	Max
effect	4838	1.224227	1.665623	-1.633104	6.835395



Interaction between continuous variables: example

- **Interpretation**: the negative effect of having children on wages decreases with age

$$\frac{\partial wage_i}{\partial nrkids_i} = -4.41 + 0.13 age_i$$

```
. sum age
```

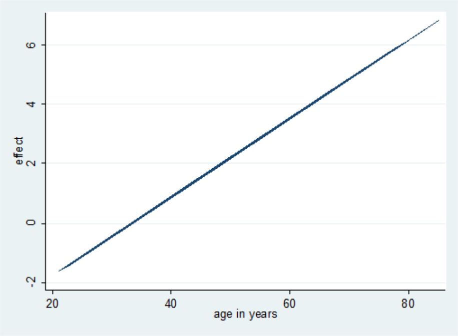
Variable	Obs	Mean	Std. Dev.	Min	Max
age	4838	42.59405	12.58781	21	85

- The effect at the average age (42.6) is positive: having one more child increases hourly wages with \$1.22.
- The effect becomes positive at an age of 33.92 (before that age, having children decreases wages)

$$\begin{aligned} -4.41 + 0.13 age^* &= 0 \\ \Leftrightarrow \quad age^* &= \frac{4.41}{0.13} = 33.92 \end{aligned}$$

Interaction between continuous variables: example

The predicted effect of having one more child on wages for different ages can also be graphed:



Differences across groups

We have seen:

- How including a **dummy variable allows for a different intercept** between two groups (example: men and women)

$$\ln wage_i = \beta_0 + \beta_1 age_i + \beta_2 educ_i + \beta_3 fem_i + \varepsilon_i$$

- How additionally including an **interaction term** allows for a **different intercept and 1 different slope** between two groups:

$$\ln wage_i = \left\{ \begin{array}{c} \beta_0 + \beta_1 age_i + \beta_2 educ_i + \beta_3 fem_i \\ +\beta_4 educ_i \times fem_i + \varepsilon_i \end{array} \right\}$$

- Sometimes, we want to an even more flexible specification: we want to allow two groups to have completely **different regression equations**, i.e. allowing the intercept and **all** slopes to differ.

Differences in regression equations across groups

Consider the following **Mincer model**:

$$\ln wage_i = \beta_0 + \beta_1 age_i + \beta_2 educ_i + \varepsilon_i$$

To allow both the intercept & all coefficients of this Mincer model to differ between men and women, we can write **equations separately for men and women**:

$$\begin{aligned} \ln wage_i &= \beta_0^M + \beta_1^M age_i + \beta_2^M educ_i + \varepsilon_i \quad \text{for males} \\ \ln wage_i &= \beta_0^F + \beta_1^F age_i + \beta_2^F educ_i + \varepsilon_i \quad \text{for females} \end{aligned}$$

Testing differences in regression equations across groups: Chow test

Chow test: tests whether groups have different regression functions.

Restricted model:

$$\ln wage_i = \beta_0 + \beta_1 age_i + \beta_2 educ_i + \varepsilon_i$$

Unrestricted models:

$$\begin{aligned} \ln wage_i &= \beta_0^M + \beta_1^M age_i + \beta_2^M educ_i + \varepsilon_i \text{ for males} \quad (1) \\ \ln wage_i &= \beta_0^F + \beta_1^F age_i + \beta_2^F educ_i + \varepsilon_i \text{ for females} \quad (2) \end{aligned}$$

We want to compare the restricted to the unrestricted models: use the Chow test (which is a particular type of F-test).

Chow test

1. Write restricted and unrestricted models; define corresponding null and alternative hypotheses.
2. Choose a significance level α
3. Estimate the restricted and unrestricted models
4. Calculate the Chow test statistic, which is an F-statistic comparing the RSS between the restricted and unrestricted models
5. Find the critical F-statistic, F_c
6. Reject H_0 if $F > F_c$

Chow test: steps 1 & 2

Restricted model:

$$\ln wage_i = \beta_0 + \beta_1 age_i + \beta_2 educ_i + \varepsilon_i$$

Unrestricted models:

$$\begin{aligned} \ln wage_i &= \beta_0^M + \beta_1^M age_i + \beta_2^M educ_i + \varepsilon_i \quad \text{for males} && (1) \\ \ln wage_i &= \beta_0^F + \beta_1^F age_i + \beta_2^F educ_i + \varepsilon_i \quad \text{for females} && (2) \end{aligned}$$

Hypotheses:

$$\begin{aligned} H_0 &: \beta_0^M = \beta_0^F, \beta_1^M = \beta_1^F, \beta_2^M = \beta_2^F \\ H_A &: H_0 \text{ not true} \\ \alpha &= 0.05 \end{aligned}$$

Chow test: step 3 - estimate of restricted model

```
. reg lwage age educ

      Source |       SS       df       MS              Number of obs =    4838
-------------+------------------------------           F(  2,  4835) =  666.62
       Model |  344.778477     2  172.389238           Prob > F      =  0.0000
    Residual |  1250.34011  4835  .258601884           R-squared     =  0.2161
-------------+------------------------------           Adj R-squared =  0.2158
       Total |  1595.11859  4837  .329774361           Root MSE      =  .50853

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0061996   .0005816    10.66   0.000     .0050594    .0073397
        educ |   .0920017   .0026777    34.36   0.000     .0867521    .0972512
       _cons |   1.298487   .0441869    29.39   0.000      1.21186    1.385113
------------------------------------------------------------------------------
```

Chow test: step 3 - estimates of unrestricted models

```
. reg lwage age educ if female==0

      Source |       SS       df       MS              Number of obs =    2395
-------------+------------------------------           F(  2,  2392) =  368.64
       Model |  184.315041     2  92.1575206           Prob > F      =  0.0000
    Residual |  597.984653  2392  .249993584           R-squared     =  0.2356
-------------+------------------------------           Adj R-squared =  0.2350
       Total |  782.299694  2394  .326775144           Root MSE      =  .49999

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0081278   .0008216     9.89   0.000     .0065166    .0097389
        educ |   .0838045    .003525    23.77   0.000     .0768921     .090717
       _cons |   1.453204   .0564518    25.74   0.000     1.342504    1.563904
------------------------------------------------------------------------------

. reg lwage age educ if female==1

      Source |       SS       df       MS              Number of obs =    2443
-------------+------------------------------           F(  2,  2440) =  399.67
       Model |  187.800312     2  93.9001559           Prob > F      =  0.0000
    Residual |  573.260221  2440  .234942714           R-squared     =  0.2468
-------------+------------------------------           Adj R-squared =  0.2461
       Total |  761.060533  2442    .3116546           Root MSE      =  .48471

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0055318   .0007788     7.10   0.000     .0040046    .0070589
        educ |   .1077437   .0038926    27.68   0.000     .1001106    .1153769
       _cons |   .9853033   .0661849    14.89   0.000      .855519    1.115088
------------------------------------------------------------------------------
```

Chow test: step 4 - calculate the F-stat

$$F = \frac{\left(RSS_M - RSS_1 - RSS_2\right)/(k+1)}{\left(RSS_1 + RSS_2\right)/(n_1 + n_2 - 2(k+1))} \backsim F_{[k+1],[n_1+n_2-2(k+1)]}$$

where

RSS_M	:	RSS from restricted model
RSS_1	:	RSS from unrestricted model 1
RSS_2	:	RSS from unrestricted model 2
n_1	:	nr of obs from unrestricted model 1
n_2	:	nr of obs from unrestricted model 2
k	:	nr of parameters (all models have same k)

Chow test: step 4 - calculate the F-stat

$$
\begin{aligned}
F &= \frac{(RSS_M - RSS_1 - RSS_2)\,/\,(k+1)}{(RSS_1 - RSS_2)\,/\,(n_1 + n_2 - 2(k+1))} \\
&= \frac{(1250.34 - 597.98 - 573.26)\;/\;(2+1)}{(597.98 + 573.26)\;/\;(2395 + 2443 - 2\,(2+1))} \\
&= \frac{(1250.34 - 597.98 - 573.26)\,/3}{(597.98 + 573.26)\,/4832} \\
&= 108.8
\end{aligned}
$$

Chow test: steps 5 & 6 - compare to critical F-stat

$$F_c = F_{3,4832,0.05} = 2.60$$

$$F > F_c \quad \text{since } 108.8 > 2.60$$

Hence reject H_0 : this means we **reject the restricted model in favor of the unrestricted models**.

The conclusion is therefore that men and women have **significantly different regression equations**.

An alternative procedure to the Chow test

- To allow both the intercept & all coefficients of this Mincer model to differ between men and women, we can write **equations separately for men and women, as done in the Chow test**:

$$
\begin{aligned}
\ln wage_i &= \beta_0^M + \beta_1^M age_i + \beta_2^M educ_i + \varepsilon_i \quad \text{for males} \\
\ln wage_i &= \beta_0^F + \beta_1^F age_i + \beta_2^F educ_i + \varepsilon_i \quad \text{for females}
\end{aligned}
$$

- Or, we could **include a dummy for gender and create interactions of all slopes with this gender dummy**:

$$
\ln wage_i = \left\{ \begin{array}{c} \beta_0 + \beta_1 age_i + \beta_2 educ_i + \beta_3 fem_i \\ +\beta_4 educ_i \times fem_i + \beta_5 age_i \times fem_i + \varepsilon_i \end{array} \right\}
$$

This combines the two unrestricted models into one model!

An alternative procedure to the Chow test

Restricted model:

$$\ln wage_i = \beta_0 + \beta_1 age_i + \beta_2 educ_i + \varepsilon_i$$

Unrestricted model:

$$\ln wage_i = \left\{ \begin{array}{c} \beta_0 + \beta_1 age_i + \beta_2 educ_i + \beta_3 fem_i \\ +\beta_4 educ_i \times fem_i + \beta_5 age_i \times fem_i + \varepsilon_i \end{array} \right\}$$

We can now perform a **regular F-test** to see which of these two models is better:

$$\begin{aligned} H_0 &: \beta_3 = \beta_4 = \beta_5 = 0 \\ H_A &: H_0 \text{ not true} \end{aligned}$$

If we reject H_0, the unrestricted model is better, and we conclude that men and women have different regression equations. We will get **exactly the same F-stat value** as with the Chow test procedure!

An alternative procedure to the Chow test

```
. gen age_fem=age*female

. gen educ_fem=educ*female

. reg lwage age educ female age_fem educ_fem

      Source |       SS       df       MS              Number of obs =    4838
-------------+------------------------------           F(  5,  4832) =  349.74
       Model |  423.873712     5  84.7747424           Prob > F      =  0.0000
    Residual |  1171.24487  4832  .242393393           R-squared     =  0.2657
-------------+------------------------------           Adj R-squared =  0.2650
       Total |  1595.11859  4837  .329774361           Root MSE      =  .49233

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0081278    .000809    10.05   0.000     .0065417    .0097138
        educ |   .0838045    .003471    24.14   0.000     .0769997    .0906094
      female |  -.4679007   .0872312    -5.36   0.000    -.6389135   -.2968879
     age_fem |   -.002596   .0011315    -2.29   0.022    -.0048142   -.0003778
    educ_fem |   .0239392   .0052613     4.55   0.000     .0136247    .0342537
       _cons |   1.453204   .0555871    26.14   0.000     1.344228     1.56218
------------------------------------------------------------------------------

. test female age_fem educ_fem

 ( 1)  female = 0
 ( 2)  age_fem = 0
 ( 3)  educ_fem = 0

       F(  3,  4832) =  108.77
            Prob > F =    0.0000
```

Project paper

Reconsider your specification, improving the functional form by considering the following modifications (i.e. trying them out, but only retaining them when appropriate):

- (Re)scaling of variables?
- Dependent and independent variables in logs or levels?
- Quadratic terms on the right-hand side of the equation?
- Inclusion of dummy variables- categorical or ordinal? Add interpretation of these variables.
- Include an interaction term & its interpretation.
- Perform a Chow test.