Lecture 7:

More on functional forms

Jan Zouhar

Introductory Econometrics

What transforms do we use, and when?

- we already know that linear regression can be used to describe non-linear relationships (we've been using logs routinely, after all)
- there is a plethora of functional transforms one can think of, but practitioners mostly restrict themselves to the following four

transform	formula	description
Units change	$x/1000$	Only used as a matter of convenience (to make results easier to read).
Logs	$\log(x)$	Changes interpreted on a relative scale. May help reduce the effect of outliers (CEO salary example).
Squares	x^2	Allows for a u-shaped or inverted-u-shaped relationship (as in age vs wage).
Interactions	$x_1 \cdot x_2$	Effect of x_1 depends on the level of x_2 and vice versa.

More on the use of logarithms

- remember we used the following approximation:

$$\text{change in } \log(y) \approx \text{relative change in } y$$

- relative changes are a bit tricky: if my wage increases by 50% next month, and decreases by 50% the following month, the total effect is a drop of 25%

$$wage \times 1.5 \times 0.5 = 0.75 wage$$

- consider a country where the average wage is 100 for men and 125 for women; then
  - women earn by 25% more than men
  - men earn less by 20% less than women
- in other words, the base category (men or women) matters
- as we know, in regressions it does not (see next slide); is there anything wrong?

```
OLS estimates
Dependent variable: l_wage

                        (1)          (2)

const                0.4317**     0.08352
                    (0.1045)     (0.1011)

educ                0.08584**    0.08584**
                  (0.007183)   (0.007183)

exper              0.009691**   0.009691**
                  (0.001433)   (0.001433)

smsa                 0.1592**     0.1592**
                   (0.04241)    (0.04241)

female              -0.3482**
                   (0.03722)

male                              0.3482**
                                 (0.03722)

        n              526          526
R-squared           0.3696       0.3696
      lnL           -292.1       -292.1
```

Intercept has changed, why?

Coefficients on other variables unaffected by the base category

Different base categories, only the sign has changed

More on the use of logarithms (cont'd)

- β_{female} in model (1) equals $-\beta_{male}$ in model (2)
- interpreting this the usual way,
  - women earn by 35% less than men
  - men earn less by 35% more than women
- but: $0.65 \times 1.35 = 0.88 \neq 1$
- in fact, there is no inconsistence, all of this is due to our approximate interpretation of the logarithm, which only works for small changes (in the log, or small relative changes)
- Exact interpretation: if e.g. $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 female + u$, exponentiating both sides, and writing down for men and women yields

  men: $wage = \exp(\beta_0 + \beta_1 educ + u)$

  women: $wage = \exp(\beta_0 + \beta_1 educ + \beta_2 + u) = \exp(\beta_2) \times \exp(\beta_0 + \beta_1 educ + u)$
- wage for women = $\exp(\beta_2)$ × wage for men

More on the use of logarithms (cont'd)

- interpreting the results in our previous Gretl output:
  - exp(0.35) = 1.42, men earn by 42% more than women
  - exp(−0.35) = 0.70, women earn by 30% less than men
- note that this solves the apparent inconsistency, as 1.42 × 0.7 = 1; or, in general,

$$\begin{aligned}\exp(\beta_{female}) \times \exp(\beta_{male}) &= \exp(\beta_{female} + \beta_{male}) \\ &= \exp(-\beta_{male} + \beta_{male}) \\ &= \exp(0) \\ &= 1\end{aligned}$$

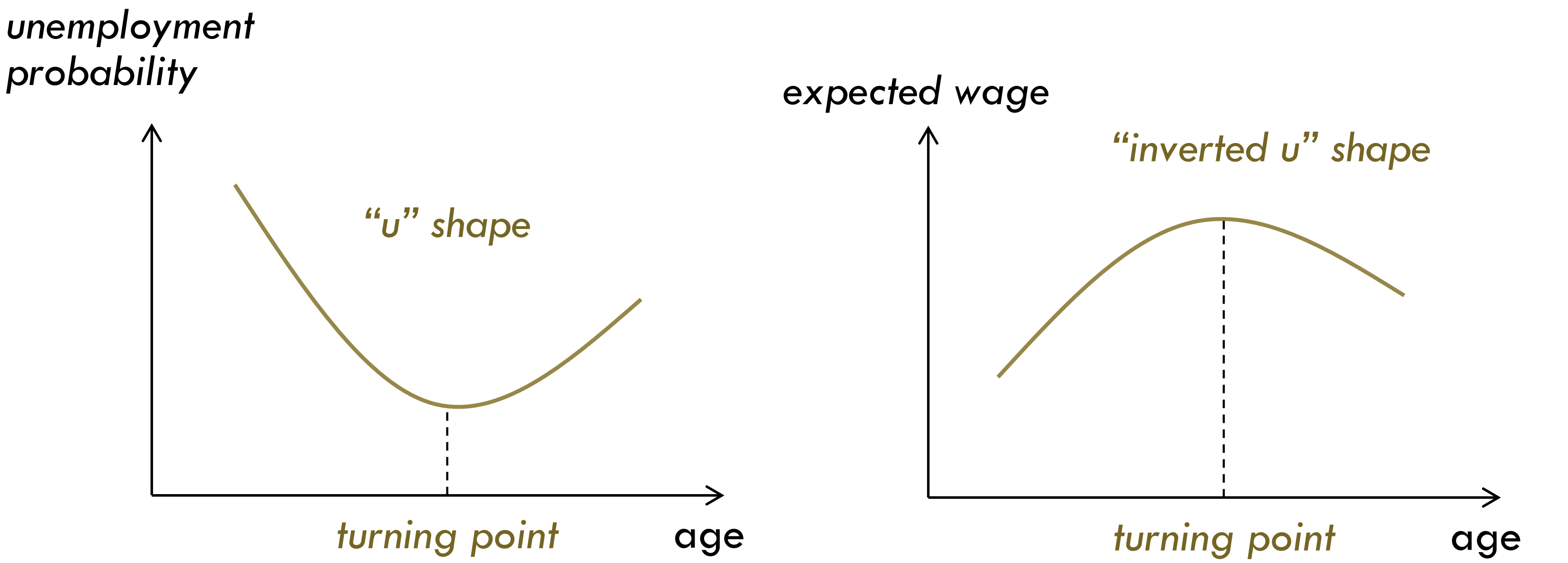
- to conclude, the *exact* relative change in y due to a unit change in x_j is

$$\Delta y / y = \exp(\beta_j) - 1, \quad \text{or}$$

$$\%\Delta y = 100[\exp(\beta_j) - 1]$$

Squares

- allow for a changing sign of the relationship
- note that while logarithms are a non-linear transform, they do not allow the relationship to change sign (log is strictly increasing)
- many nonlinear functions allow this, but the quadratic is the simplest one → hardly ever we use anything beyond that



Squares (cont'd)

Example

- wage vs. work experience
- we estimate

$$wage = \beta_0 + \beta_1 exper + \beta_2 exper^2 + u$$

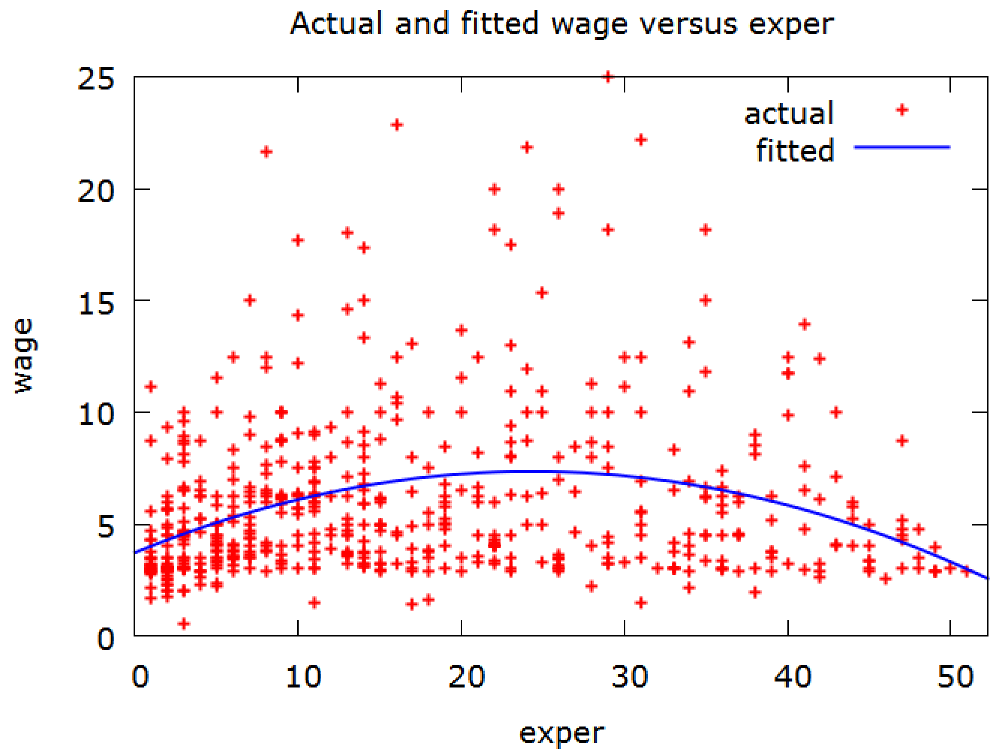
- In Gretl: first we need to create a new variable containing squared experience (Add → Squares of selected variables)
- the estimated equation (using Wooldridge's `wage1` data) is:

```
^wage = 3.73 + 0.298*exper - 0.00613*sq_exper
       (0.346) (0.0410)      (0.000903)

n = 526, R-squared = 0.093
(standard errors in parentheses)
```

- Quizz: is this a **u** or an **inverted-u** curve? Where is the turning point?

Squares (cont'd)

- a plot may help answer these questions (Graphs → Fitted, Actual plot → Against exper)



- but the turning point will not be guessed accurately from the plot, and the plot looks ugly if we include control variables

Where exactly is the turning point?

- use first-order conditions for a maximum/minimum of a function
- differentiate the equation $wage = \beta_0 + \beta_1 exper + \beta_2 exper^2 + u$ with respect to *exper* and set equal to zero:

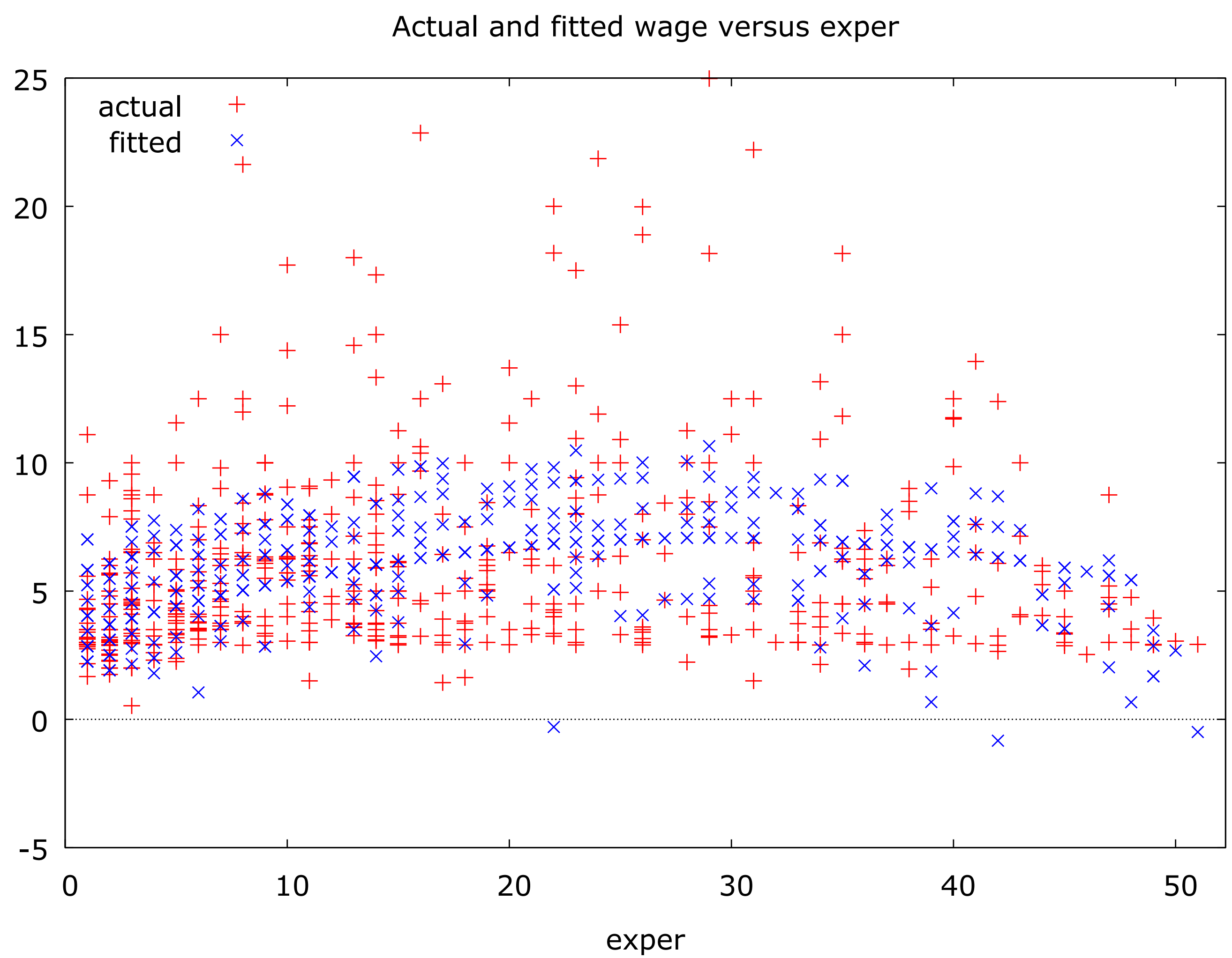
$$\frac{\partial wage}{\partial exper} = \beta_1 + 2\beta_2 exper = 0$$

- so the turning point is: $exper = -\dfrac{\beta_1}{2\beta_2}$
- our estimate of the turning point (based on the estimated equation) is

$$\text{estimated turning point} = -\frac{\text{coefficient on the linear term}}{2 \times \text{coefficient on the squared term}}$$

- in our example, this is $exper = -\dfrac{0.298}{2(-0.00613)} = 24.3$ years

```
^wage = -3.96 + 0.268*exper - 0.00461*sq_exper + 0.595*educ
       (0.752) (0.0369)      (0.000822)           (0.0530)
```



More on squares

- u or inverted-u shape? Determined by the **sign** of the coefficient on the **squared term** (positive → u; negative → inverted u)
- **partial effect** of experience:

$$\frac{\Delta wage}{\Delta exper} \approx \frac{\partial wage}{\partial exper} = \beta_1 + 2\beta_2 exper, \quad \text{so} \quad \Delta wage \approx \left(\beta_1 + 2\beta_2 exper\right)\Delta exper$$

- in particular, the change in wage brought about by a unit increase in experience ($\Delta exper = 1$) is $\beta_1 + 2\beta_2 exper$
- now wait, we used to log the wage in most regressions
- fortunately, log is an increasing function, log(*wage*) increases whenever *wage* does, so our turning point formulas work even for

$$\log(wage) = \beta_0 + \beta_1 exper + \beta_2 exper^2 + u$$

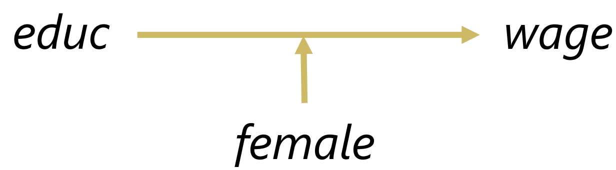
- partial effect: $\Delta \log(wage) \approx \left(\beta_1 + 2\beta_2 exper\right)\Delta exper, \quad \text{so}$

$$\%\Delta wage \approx 100\left(\beta_1 + 2\beta_2 exper\right)\Delta exper$$

Interactions

Example: Do returns to schooling differ for men and women?

- Or: is the effect of education on the wage **moderated** by gender?



- What do you think is the case in your country? Any objective reasons why women should be rewarded more/less for their education than men?
- How do we formulate a model that allows the effect of education to vary with gender?

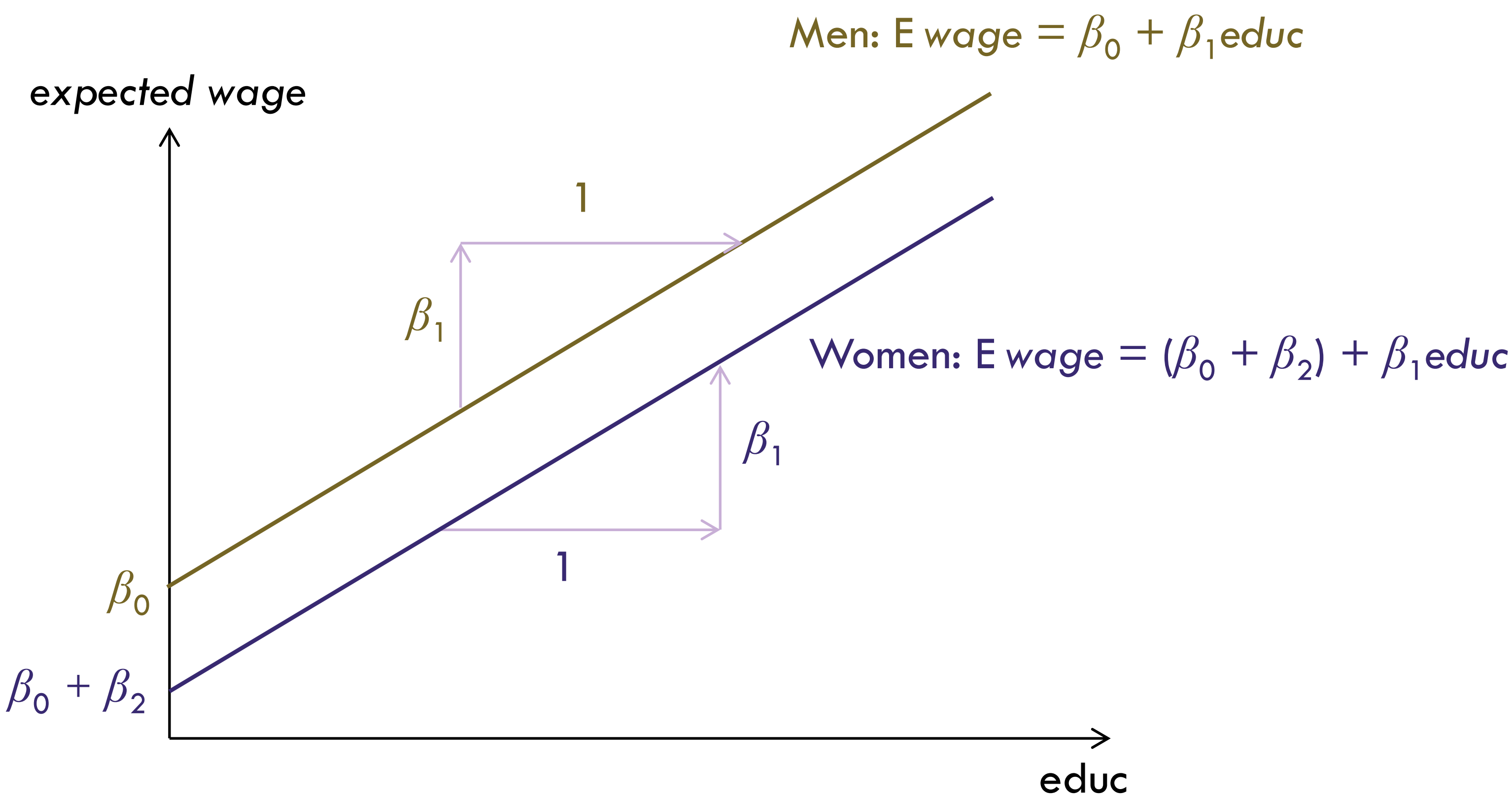
$$wage = \beta_0 + \beta_1 educ + \beta_2 female + u \qquad (1)$$

$$wage = \beta_0 + \beta_1 educ + \beta_2 female + \beta_3 female \cdot educ + u \qquad (2)$$

- It is easily seen that the effect of additional year of education, $\dfrac{\Delta wage}{\Delta educ}$, is
  - β_1 in equation (1)
  - $\beta_1 + \beta_3 female$ in equation (2)

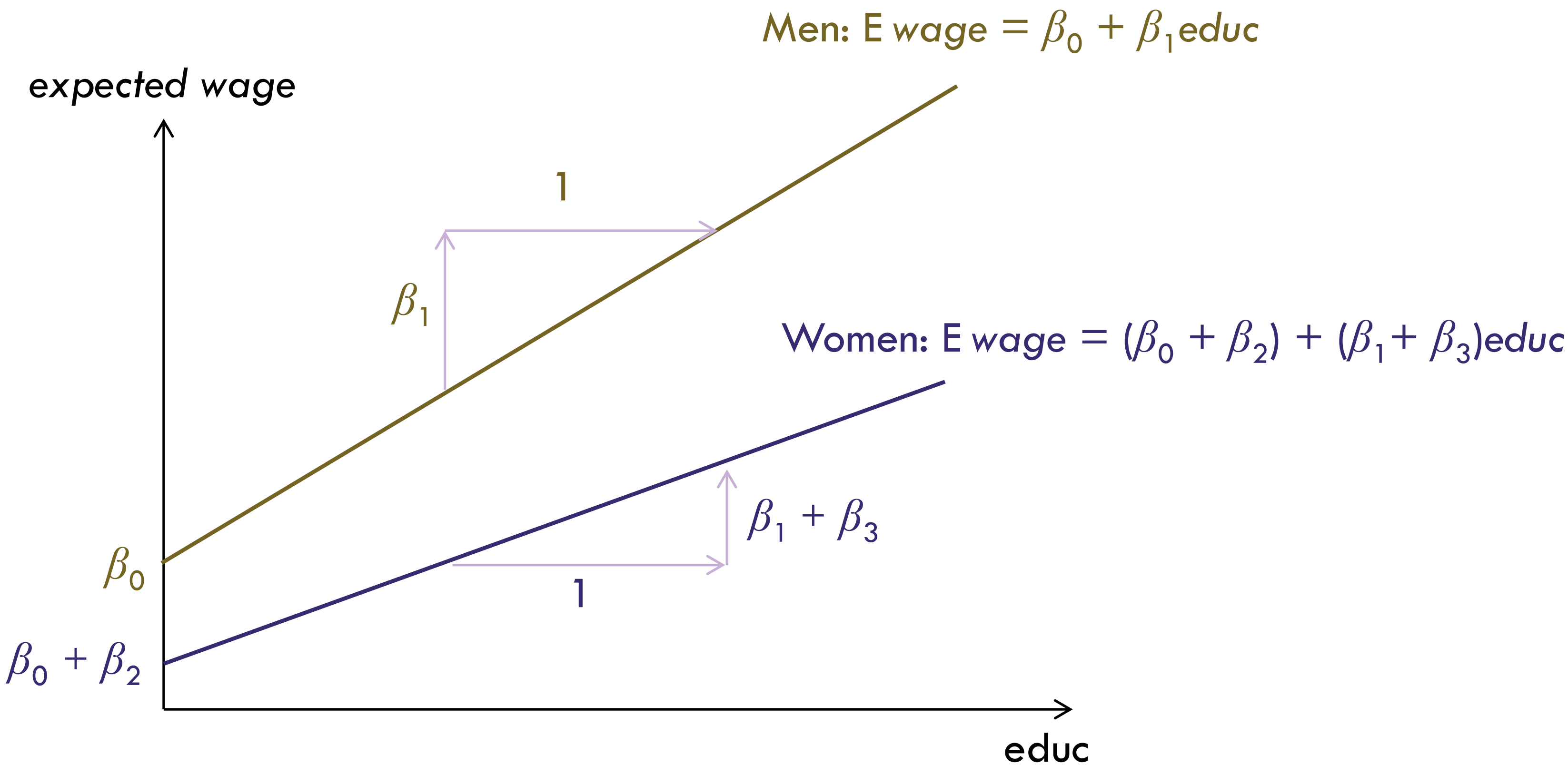
Interactions (cont'd)

$$wage = \beta_0 + \beta_1 educ + \beta_2 female + u$$



Interactions (cont'd)

$$wage = \beta_0 + \beta_1 educ + \beta_2 female + \beta_3 female \cdot educ + u$$



```
Model 1: OLS, using observations 1-526
Dependent variable: wage

               coefficient   std. error   t-ratio    p-value
  ----------------------------------------------------------
  const          0.200496     0.843562      0.2377   0.8122
  educ           0.539476     0.0642229     8.400    4.24e-016 ***
  female        -1.19852      1.32504      -0.9045   0.3661
  femaleXeduc   -0.0859990    0.103639     -0.8298   0.4070

Mean dependent var   5.896103   S.D. dependent var   3.693086
Sum squared resid    5300.170   S.E. of regression   3.186469
R-squared            0.259796   Adjusted R-squared   0.255542
F(3, 522)            61.07022   P-value(F)           7.44e-34
Log-likelihood      -1353.942   Akaike criterion     2715.885
Schwarz criterion    2732.946   Hannan-Quinn         2722.565
```

- What is the interpretation of the intercept?
- What is the interpretation of the β_{educ}?
- What is the interpretation of the β_{female}?
- What is the effect of an additional year of education on a woman's wage?
- Do returns to schooling differ for men and women?

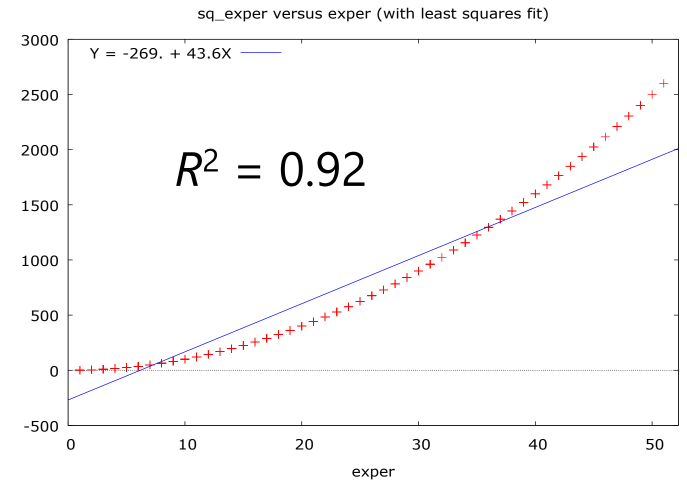
Variable centering

- Sample median of *educ* is 12
- Create new variable *educ_12* = *educ* − 12; new interpretation?

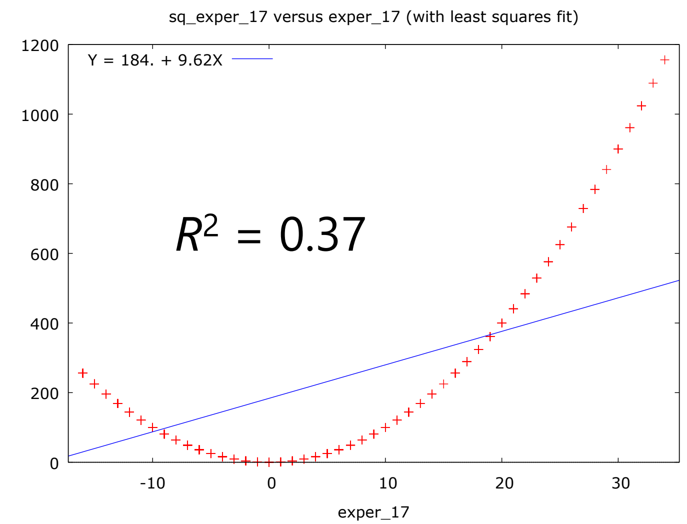
```
Model 3: OLS, using observations 1-526
Dependent variable: l_wage

                    coefficient   std. error   t-ratio    p-value
  -------------------------------------------------------------
  const              1.46091      0.0493213    29.62     1.27e-113 ***
  educ_12            0.0876179    0.00902612    9.707    1.39e-020 ***
  female            -0.345893     0.0379530    -9.114    1.73e-018 ***
  femaleXeduc_12    -0.00481837   0.0138472    -0.3480   0.7280
  exper              0.00970891   0.00143735    6.755    3.85e-011 ***
  smsa               0.159559     0.0424996     3.754    0.0002    ***
  nonwhite          -0.00966693   0.0613298    -0.1576   0.8748

Mean dependent var   1.623268   S.D. dependent var   0.531538
Sum squared resid    93.47959   S.E. of regression   0.424399
R-squared            0.369785   Adjusted R-squared   0.362500
F(6, 519)            50.75480   P-value(F)           4.38e-49
Log-likelihood      -292.0139   Akaike criterion     598.0278
Schwarz criterion    627.8849   Hannan-Quinn         609.7182
```

Multicollinearity vs. squares & interactions

```
Variance Inflation Factors
Minimum possible value = 1.0
Values > 10.0 may indicate a collinearity
problem

       exper   13.216
    sq_exper   13.493
        educ    1.867
      female   22.899
 femaleXeduc   22.869
    nonwhite    1.013
        smsa    1.059
```




```
Variance Inflation Factors
Minimum possible value = 1.0
Values > 10.0 may indicate a collinearity
problem

      exper_17    1.639
   sq_exper_17    1.639
       educ_12    1.867
        female    1.050
femaleXeduc_12    1.650
      nonwhite    1.013
          smsa    1.059
```



How do we decide about the functional form?

- even if we restrict ourselves to squares, logs, and interactions, there's many different functional forms we can produce with given variables; how do we choose?
- lecture 2 revisited:

> **Why use simple models:**
>
> Simple models are:
>
> - easier to estimate.
> - easier to interpret (e.g., $\beta_1 = \Delta\text{wage}/\Delta\text{educ}$ etc.).
> - easier to analyze from the statistical standpoint.
> - safe: they serve as a good approximation to the real relationship, the functional nature of which might be unknown and/or complicated. Things can't go too wrong when using a simple model.
>
> Further reading: Angrist and Pischke (2008): *Mostly Harmless Econometrics: An Empiricist's Companion*.

Tests for functional form misspecification

- even though some statistical tests have been developed to detect functional form misspecification, we should use them sparingly: they can lead to overspecified (= overly complicated) models that do not interpret easily
- the most important criteria are: (i) our research question and the underlying economic theory, and (ii) the desired interpretation of the parameters (see Slide 2 of this presentation)

Using F-tests for joint significance

- it is straightforward to check for the omission of squares and interactions in a particular model using an F-test
- just add squares and/or interactions of the regressors and use the F-test for joint significance
- Gretl uses this for logarithms as well

Ramsey's RESET test

- a popular test for general functional form misspecification
- procedure:
  1. First, use OLS to estimate your equation, say

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

  2. Save the fitted values, $\hat{y}$.
  3. Estimate the equation

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + u$$

  and use the F-test for joint significance of $\hat{y}^2$ and $\hat{y}^3$.
- note that $\hat{y}^2$ and $\hat{y}^3$ are themselves functions of cubes, squares, and interactions of the x's, but using $\hat{y}^2$ and $\hat{y}^3$ instead of all possible interactions and squares saves up on degrees of freedom dramatically

```
Auxiliary regression for RESET specification test
OLS, using observations 1-328
Dependent variable: l_price

             coefficient   std. error    t-ratio   p-value
  -------------------------------------------------------
  const      -778.711      214.096       -3.637    0.0003  ***
  km1000        0.138152     0.0372679    3.707    0.0002  ***
  age          10.2993       2.78202      3.702    0.0003  ***
  combi        -8.39722      2.26483     -3.708    0.0002  ***
  diesel      -15.3748       4.14411     -3.710    0.0002  ***
  LPG          -4.84540      1.31218     -3.693    0.0003  ***
  octavia     -52.6445      14.2247      -3.701    0.0003  ***
  superb     -100.411       27.0420      -3.713    0.0002  ***
  yhat^2        7.51842      2.06297      3.644    0.0003  ***
  yhat^3       -0.199197     0.0561879   -3.545    0.0005  ***

Warning: data matrix close to singularity!

Test statistic: F = 24.093873,
with p-value = P(F(2,318) > 24.0939) = 1.81e-010
```

- Numerical instability!
- In this case, the version with a squared term only is preferred

```
Auxiliary regression for RESET specification test
OLS, using observations 1-328
Dependent variable: l_price

             coefficient    std. error   t-ratio   p-value
  --------------------------------------------------------------
  const      −19.9472       5.54465      −3.598    0.0004   ***
  km1000       0.00611032   0.00131867    4.634    5.24e-06 ***
  age          0.442007     0.0944437     4.680    4.24e-06 ***
  combi       −0.373065     0.0820537    −4.547    7.75e-06 ***
  diesel      −0.692139     0.147900     −4.680    4.25e-06 ***
  LPG         −0.200290     0.0722966    −2.770    0.0059   ***
  octavia     −2.24280      0.479250     −4.680    4.25e-06 ***
  superb      −4.60119      0.969330     −4.747    3.13e-06 ***
  yhat^2       0.205809     0.0351040     5.863    1.14e-08 ***

Test statistic: F = 34.372892,
with p-value = P(F(1,319) > 34.3729) = 1.14e-008
```

Price or log(*price*)?

price

```
Non-linearity test (squares)
  Test statistic: LM = 87.3563
  with p-value = P(Chi-square(2) > 87.3563) = 1.07352e-019

Non-linearity test (logs) -
  Test statistic: LM = 52.1271
  with p-value = P(Chi-square(2) > 52.1271) = 4.79459e-012

RESET test for specification
  Test statistic: F(2, 318) = 82.1404
  with p-value = P(F(2, 318) > 82.1404) = 1.7427e-029
```

log(*price*)

```
Non-linearity test (squares) -
  Test statistic: LM = 37.1925
  with p-value = P(Chi-square(2) > 37.1925) = 8.38964e-009

Non-linearity test (logs) -
Test statistic: LM = 11.4947
  with p-value = P(Chi-square(2) > 11.4947) = 0.00319124

RESET test for specification -
Test statistic: F(2, 318) = 24.0939
  with p-value = P(F(2, 318) > 24.0939) = 1.8072e-010
```

Lecture 7:
More on functional forms

Jan Zouhar

Introductory Econometrics