Lecture 1: Regression analysis: a recapitulation of Econometrics and Statistics in Bachelor

Prof. dr. Wolter Hassink
Utrecht University School of Economics
w.h.j.hassink@uu.nl

Position of Empirical Economics in the curriculum

<table>
<tr><th colspan="3">Semester 1</th><th>Semester 2</th></tr>
<tr><th>Period 1</th><th colspan="2">Period 2</th><th>Period 3, 4</th></tr>
<tr><td rowspan="9">EMPIRICAL ECONOMICS</td><td colspan="2">RESEARCH PROJECT</td><td rowspan="9">THESIS</td></tr>
<tr><td>B&F</td><td>Fintech</td></tr>
<tr><td>B&SI</td><td>Frontiers of Business and Social Impact</td></tr>
<tr><td>BD&E</td><td>Frontiers of Entrepreneurship</td></tr>
<tr><td>EP</td><td>Policy Evaluation Skills</td></tr>
<tr><td>IM</td><td>Frontiers of International Management</td></tr>
<tr><td>FM</td><td>Next Generation Finance</td></tr>
<tr><td>SC&R</td><td>SC&R</td></tr>
<tr><td>SF&I</td><td>Sustainability Risk</td></tr>
</table>

Setup of today

- Motivation
- First stage of an empirical project (6 steps)
- Omitted variables
- Exogeneity
- Mechanics of OLS
- Bias and consistency of the OLS estimator
- Quadratic form
- 0-1 indicator variables and categorical variables

NOTE THAT THE SUPPLEMENTARY MATERIAL (SEPARATE SET OF SLIDES) IS ALSO PART OF THE MATERIAL FOR THE TUTORIAL AND THE EXAM

- Goodness of fit
- How to estimate σ^2
- Variance of the OLS estimator
- Heteroskedasticity
- The t-test
- The F-test
- Heteroskedasticity and robust standard errors

MOTIVATION

Multiple linear regression model - motivation

- **Starting point 1:** Theory from economics or business economics.
    - Literature of previous empirical studies.
    - There are multiple competing economic hypotheses.
    - Specific institutional setting on the industry, financial rules, country or law.
- **Starting point 2:** Data set – See some examples on next slide
- **This course:** No controlled experiment
- **Questions that we partially address in this course:**
    - How were the data obtained?
    - Is it a random sample of observations? Time series? Panel data?
    - Validity of the variables? For instance gross or net profits?
    - Measurement error? Administrative data or survey data?
    - Some of these issues will be considered in the period 2 course research project.
- **This course:** Linear regression equation
- **What you need to do?** Formulate a statistical model at the level of the population: give the set of statistical assumptions, given the structure of the data
    - *You should be able to formulate assumptions*
- How to estimate the model? For instance, Ordinary Least Squares?
    - *You should be able to use Stata for empirical analysis*
- How to interpret the estimates
    - *You should be able to interpret regression output*
- **Specific questions of interest:**
    - Is there any causal effect?
    - How large is the economic effect?
    - Can we predict with the statistical model?
    - What is the precision of the estimate?

Type of data sets

Aim: to introduce the different formats of data sets.

- Cross-sectional data: aselect sample of individuals (firm, bank, employee, household, region) (this week)

- Time-series data: an individual (firm, bank, household, region) is followed over time (week 2, 3)

- Panel data: aselect sample of individuals (firm, bank, household, region) is followed over time (week 4, 5)

Key issues in Empirical Economics/ Empirical Research

- Example: let's consider the variables education and salary of individual employees
- What type of empirical questions can we address? Using data, we may want to measure
    - Association/correlation

      $$Corr(salary, exper)$$

    - Causation

      $$\Delta exper \rightarrow \Delta salary$$

    - Forecast
        - What is expected amount of salary at time $t + 1$ given a set of information at time t?

- This week
    - Correlation or causal effect?
    - Main questions
        - What are confounding variables/omitted variables?
        - How does it affect the estimate of the effect?
        - What if we take a different model (quadratic terms or 0-1 indicator variables)?

FIRST STAGE OF AN EMPIRICAL PROJECT: 6 STEPS

Step 1: inspect several summary statistics

- Before we are running a regression, as a first step, we will consider summary statistics of the data set (wage1.dta)

```
. sum salary exper female, detail

                          salary
-------------------------------------------------------------
      Percentiles      Smallest
 1%         1.67            .53
 5%         2.75           1.43
10%         2.92            1.5       Obs                 526
25%         3.33            1.5       Sum of Wgt.         526

50%         4.65                      Mean           5.896103
                        Largest       Std. Dev.      3.693086
75%         6.88          21.86
90%           10           22.2       Variance       13.63888
95%           13          22.86       Skewness       2.007325
99%           20          24.98       Kurtosis       7.970083

                  years potential experience
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              1
 5%            1              1
10%            2              1       Obs                 526
25%            5              1       Sum of Wgt.         526

50%         13.5                      Mean           17.01711
                        Largest       Std. Dev.      13.57216
75%           26             49
90%           38             49       Variance       184.2035
95%           43             50       Skewness       .7068652
99%           49             51       Kurtosis       2.357318

                         =1 if female
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs                 526
25%            0              0       Sum of Wgt.         526

50%            0                      Mean           .4790875
                        Largest       Std. Dev.       .500038
75%            1              1
90%            1              1       Variance        .250038
95%            1              1       Skewness       .0837235
99%            1              1       Kurtosis        1.00701
```

```
. sort female

. by female: sum salary exper

-------------------------------------------------------------------------------
-> female = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      salary |       274    7.099489    4.160858        1.5      24.98
       exper |       274    17.55839    13.49991          1         51

-------------------------------------------------------------------------------
-> female = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      salary |       252    4.587659    2.529363        .53      21.63
       exper |       252    16.42857    13.65274          1         50
```

```
. pwcorr salary exper female, sig

             |   salary    exper   female
-------------+---------------------------
      salary |   1.0000
             |
             |
       exper |   0.1129   1.0000
             |   0.0096
             |
      female |  -0.3401  -0.0416   1.0000
             |   0.0000   0.3407
             |
```

Step 2: estimate the parameters of a linear regression equation

```
. reg salary exper female

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

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


. gen lsalary = log(salary)

. reg lsalary exper female

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  2,   523) =   45.72
       Model |   22.076198     2   11.038099           Prob > F      =  0.0000
    Residual |  126.253553   523  .241402588           R-squared     =  0.1488
-------------+------------------------------           Adj R-squared =  0.1456
       Total |  148.329751   525   .28253286           Root MSE      =  .49133

------------------------------------------------------------------------------
     lsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0037591   .0015813     2.38   0.018     .0006526    .0068656
      female |  -.3929703   .0429205    -9.16   0.000    -.4772881   -.3086526
       _cons |   1.747566   .0406442    43.00   0.000      1.66772    1.827412
------------------------------------------------------------------------------
```

Interpretation of the remaining regression parameters: see Table 2.3 of Wooldridge. log(.): the natural logarithm

Model	**Dependent variable**	**Independent variable**	**Interpretation of β_j**
level-level	y	x	$\Delta y = \beta_j \Delta x$
level-log	y	$\log(x)$	$\Delta y = (\beta_j / 100)\%\Delta x$
log-level	$\log(y)$	x	$\%\Delta y = (100\beta_j)\Delta x$
log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_j \%\Delta x$

Step 3: scrutinize the estimates

- Have a look at the number of observations: missed any data?
- Check the statistical significance: t-statistics, p-values
- Consider estimated parameters and standard errors
- Consider the plausibility of the estimated parameters of the statistically significant parameter estimates (Table 2.3)
- IMPORTANT: Statistically insignificant parameters estimates cannot be interpreted. Even not the sign!
- Note that each regression equation we are running is implicitly based on an economic framework: **we cannot interpret the following regression:**

```
. reg female salary exper

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  2,   523) =   34.21
       Model |  15.1849298     2  7.59246491            Prob > F      =  0.0000
    Residual |  116.085032   523  .221959909           R-squared     =  0.1157
-------------+------------------------------           Adj R-squared =  0.1123
       Total |  131.269962   525  .250038023           Root MSE      =  .47113

------------------------------------------------------------------------------
      female |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      salary |  -.0459987   .0056034    -8.21   0.000    -.0570067   -.0349907
       exper |  -.0001205   .0015247    -0.08   0.937    -.0031158    .0028749
       _cons |   .7523505   .0446449    16.85   0.000     .6646451    .8400559
------------------------------------------------------------------------------
```

Step 4: write down the corresponding population regression equation

We specify a linear regression equation at the level of the population

$$salary = \beta_0 + \beta_1 female + \beta_2 exper + u$$

salary is the payment of the firm to its individual employees.
female is a 0-1 indicator for female (a dummy variable).
experience is measured in number of years the employee has been active on the labour market

All of the other influences on salary are captured by the error term *u*. (for instance: productivity, years of schooling). These variables are NOT included in the data set, but they have an effect on salary.

Step 5: write down the necessary assumptions for a causal interpretation of the estimated parameters

$$salary = \beta_0 + \beta_1 female + \beta_2 exper + u$$

We want to compute the expected value of the dependent variable, given the set of independent variables. It is an expectation of salary for the conditional distribution of salary given female and experience.

$$E(salary \mid female, exper) =$$

$$= E(\beta_0 + \beta_1 female + \beta_2 exper + u \mid female, exper)$$

$$= \beta_0 + \beta_1 female + \beta_2 exper + \underbrace{E(u \mid female, exper)}_{\substack{=0 \text{ if } u \text{ is independent of} \\ female \text{ and } experience}}$$

$$= \beta_0 + \beta_1 female + \beta_2 exper$$

Statistical assumption: $E(u \mid female, exper) = 0$

Zero conditional mean assumption: $E(u \mid female, exper) = 0$

In simple words: all other influences on y are statistically independent of the INCLUDED right hand side variables.

Thus there are no omitted variables on the right-hand side of the equation.

If this assumption is true, we can use regression analysis to predict for specific values of the explanatory variable.

For instance, we compute the expected salary for both *experience* = 11 and *experience* = 10 (note: we do not change the value of *female*, so that we keep *female* constant)

$$E(salary \mid female, exper = 11) = \beta_0 + \beta_1 female + \beta_2 \cdot 11$$
$$\frac{E(salary \mid female, exper = 10) = \beta_0 + \beta_1 female + \beta_2 \cdot 10}{E(\Delta salary \mid \underbrace{\text{no change of } female}_{\Delta female = 0}, \Delta exper = 1) = \beta_2} \_$$

By increasing the experience by one year, expected salary increases by β_2 units, keeping constant female and under the zero conditional mean assumption.

Other example, we compute the expected salary for both *experience* = 0 and the 0-1 variable *female* = 0.

Result: the intercept is the expected value of the salary for the males with zero years of experience

```
. reg salary exper female

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

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

Step 6: Unbiased or consistent parameter estimates?

- The OLS estimates are: $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ (estimated intercept, parameter on experience, parameter on female)

- It is based on information of all variables: *salary*, *exper*, *female* for all individuals of the sample.

- $\hat{\beta}_1$ is a random variable: it has a distribution, an expected value and a variance.

- Terminology: distribution of $\hat{\beta}_1$ is **the sampling distribution**. All outcomes of $\hat{\beta}_1$ of all random samples (same population) with equal sample size.

- **Definition**: unbiased/consistent parameter of $\hat{\beta}_1$:

  Over many random samples of equal sample size (same n), the average of the estimated parameters is equal to the true unknown parameter of the linear regression model.

- Thus: on average we are right!

- Later today more on this matter.

Omitted variables: theory and examples

- Due to **omitted variables/confounders** (e.g. the variable 'productivity' is not included in the set of regressors), the parameter β_1 does not measure the partial effect of *experience* on *salary*: $E(u \mid female, exper) \neq 0$

- The zero conditional mean assumption is violated if the omitted variable:
    - is contained by the error term, u.
    - is correlated with at least one of the explanatory variables $x_1, ..., x_k$.

- One solution to the omitted variable problem is to include the variable 'productivity' in the model:

$$salary = \beta_0 + \beta_1 female + \beta_2 exper + \beta_3 productivity + u$$

- Problem: '*productivity*' is difficult to observe. It is often not included in datasets.

Example

The purpose of this example is threefold. First, to demonstrate that in all fields of economics, unobservables are important in linear regression equations. Second, that the specific unobservables are based on multiple dimensions, which are hard to measure. Third, it does not make sense to make the claim of the presence of omitted variables, without specifying the reason or source of the confounding variable.

Specification 1. For a dataset of individual employees, the wage depends on years of schooling (*educ*) and a set of additional explanatory variables (= *controls*)

$$wage = \beta_0 + \beta_1 educ + controls + u$$

where u contains the unobserved **ability of the individual**, which may be related to the years of schooling. Thus

$$E(u \mid educ, controls) \neq 0$$

Specification 2. For a dataset of individual firms, the firm profits (*profit*) depend on the firm size (*size*) and a set of additional explanatory variables (= *controls*)

$$profit = \beta_0 + \beta_1 size + controls + u$$

where u contains the unobserved **quality of the management**, which may be related to firm size. Thus

$$E(u \mid size, controls) \neq 0$$

Specification 3. For another dataset of individual firms, a cost regression

$$costs = \beta_0 + \beta_1 size + controls + u_i \qquad (9)$$

where u contains the unobserved **market power of the firm**, which may be related to firm size.

$$E(u \mid size, controls) \neq 0$$

Specification 4. For a dataset of owner-occupied houses that were sold in a specific housing market, the transaction price is explained by the household income of the inhabitants

$$transaction\_price_i = \beta_0 + \beta_1 income_i + controls + u_i \qquad (10)$$

where u contains the unobserved **regional amenities**.

$$E(u \mid income, controls) \neq 0$$

Specification 5. For a dataset of employees who are providing informal care at home, the number of supplied hours in the labour market depends on the informal care

$$hours = \beta_0 + \beta_1 care + controls + u$$

where u contains the unobserved **health** or **care need** of their friends and relatives.

$$E(u \mid care, controls) \neq 0$$

Specification 6. For a dataset of pupils at primary schools

$$Grade = \beta_0 + \beta_1 Class\_size + controls + u$$

where u contains the unobserved **parental motivation**.

$$E(u \mid care, controls) \neq 0$$

Specification 7. For a dataset of employees, sickness absence is lower for highly-educated persons (*education*)

$$sickness = \beta_0 + \beta_1 education + controls + u$$

where u contains the unobservables such as **health**.

How can we reduce the bias of the estimated parameters?

- More observations?
- More explanatory variables?
- Other specification?

EXOGENEITY

Exogeneity

Aim: to introduce the concept of exogeneity.

- How to formulate no relationship between the error term *u* and the explanatory variables (e.g. *experience* and *female*)?
- The key assumption for the multiple regression model is exogeneity of the independent variables. There are two versions applied for aselect samples.

Option 1: Applied to **large datasets**. **Assumption:**

$$E(u) = 0,\ Cov(u, x_1) = 0,\ \ldots,\ Cov(u, x_k) = 0$$

It means:

- The error term u is linearly unrelated to **all** of the explanatory variables (a zero correlation, a zero covariance).
- Any linear combination of *u* is unrelated to any linear combination of the explanatory variable.
  E.g. if $Cov(u, x_1) = 0$ then $Cov(3u + 5, -x_1 + 2) = 0$
  If $Cov(u, exper) = 0$ then $Cov(u, 12 \cdot exper - 1) = 0$
- Note that the **correlation coefficient** between *u* and *x* is defined as

$$Corr(u, x) = \frac{Cov(u, x)}{\sqrt{Var(u)}\sqrt{Var(x)}}$$

Option 2: Applied to **datasets with a limited number of Observations. Assumption:**

Zero-conditional mean assumption

$E(u \mid x_1,...,x_k) = 0$

It means:

- Conditional on the set of the explanatory variables altogether, the expected value of the error term *u* is zero. Formally: the conditional expectation of *u* conditional on all explanatory variable together $x_1,...,x_k$ is equal to 0.
- The error term u is statistically mean independent of all variables $x_1,...,x_k$
- We emphasize that this assumption does not mean that u not statistically mean independent of the explanatory variables separately. E.g., $E(u \mid x_1,...,x_k) = 0$ does not imply that $E(u \mid x_1) = 0$

It can be shown that (you don't need to prove)

$$E(u \mid x) = 0 \quad \Rightarrow \quad \begin{cases} Eu = 0 & \text{(this is not really important)} \\ Cov(u, x) = 0 & \text{(this is very important)} \end{cases}$$

mean independence	zero correlation
=	=
strong assumption	weak assumption

but that

$$E(u \mid x) = 0 \not\Leftarrow \begin{cases} Eu = 0 \\ Cov(u, x) = 0 \end{cases}$$

Note: The implication of exogeneity is that all unobserved factors contained in the error term *u* are exogenous to the regressors.

Question: what is the difference between a **strong** assumption and a **weak** assumption? **We prefer the weakest assumption if it can be applied**. A strong assumption is harder to believe.

Multiple linear regression model - Example

We reconsider our example

$$salary = \beta_0 + \beta_1 female + \beta_2 exper + u$$

Strong assumption

$$E(u \mid female, exper) = 0$$

$\Rightarrow$ Weak assumption

$$\begin{cases} Eu = 0 \\ Cov(u, female) = 0, Cov(u, exper) = 0 \end{cases}$$

Multiple linear regression model - Example

We reconsider our example

$$salary = \beta_0 + \beta_1 female + \beta_2 exper + u$$

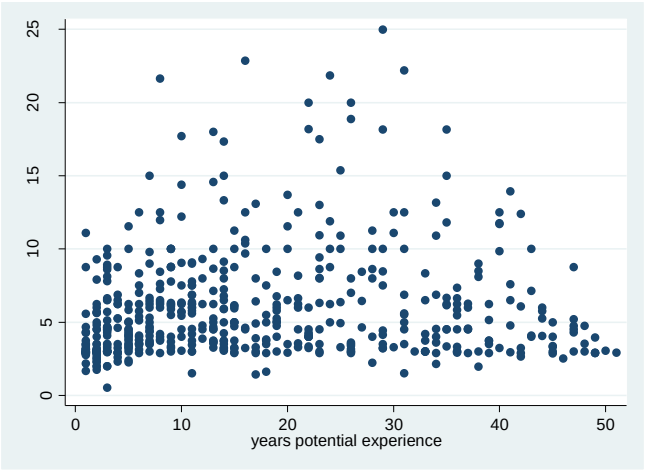
Option 1: $E(u) = 0,\ Cov(u, female) = 0,\ Cov(u, exper) = 0$

Option 2: $E(u \mid female, exper) = 0$

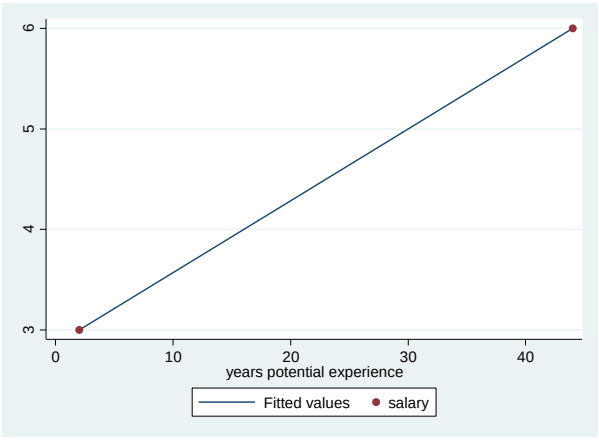
It does not imply $E(u \mid exper) = 0$

THE MECHANICS OF OLS

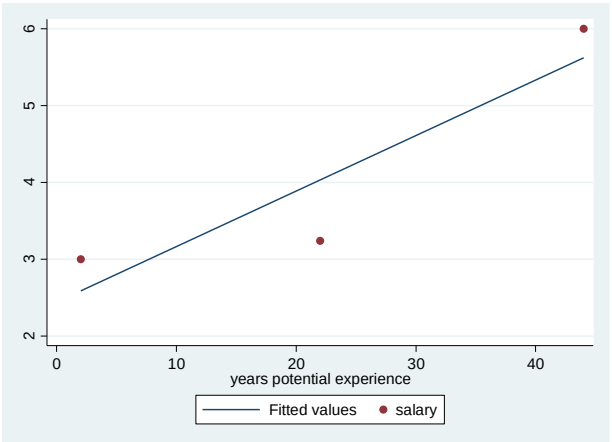
All observations; command: graph twoway scatter salary exper



Two observations; command: graph twoway (lfit salary exper) (scatter salary exper) if n <= 4 & _n > 2



Three observations; command: graph twoway (lfit salary exper) (scatter salary exper) if _n <= 4 & _n > 1



The mechanics of Ordinary Least Squares (OLS) - motivation

Aim: to consider the estimation procedure of OLS.

- Model:

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

- Task of researcher: to estimate the unknown parameters $\beta_0, \beta_1, ..., \beta_k$ using a **random sample of *n* observations**.

- **Estimation method**: **Ordinary Least Squares (OLS)**.

- Without any loss of generality, we explain the OLS method by means of the following **bivariate regression model** (thus *k*=1, see Chapter 2). Note that *x* contains no subscript here

$$y = \beta_0 + \beta_1 x + u \qquad (2.9)$$

- Assumption: there is a random sample of *n* observations $((y_1, x_1), ..., (y_n, x_n))$ (the subscript 1,…,*n* refers to the observation). The model becomes for the *i*-th individual

$$y_i = \beta_0 + \beta_1 x_i + u_i \qquad i = 1, ..., n \qquad (2.9)$$

- Scatter diagram: see illustration above

Example: Stata

Stata: output of reg y x

```
. reg salary exper

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  1,   524) =    6.77
       Model |  91.2751351     1  91.2751351           Prob > F      =  0.0096
    Residual |  7069.13916   524  13.4907236           R-squared     =  0.0127
-------------+------------------------------           Adj R-squared =  0.0109
       Total |  7160.41429   525  13.6388844           Root MSE      =   3.673

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0307219   .0118111     2.60   0.010      .007519    .0539247
       _cons |   5.373305   .2569919    20.91   0.000     4.868444    5.878166
------------------------------------------------------------------------------
```

- We would like to compute the β_0 and β_1 that minimize the following loss function (make the function as small as possible):

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 =$$

$$= \min_{\beta_0,\beta_1} (y_1 - \beta_0 - \beta_1 x_1)^2 + ... + (y_n - \beta_0 - \beta_1 x_n)^2 \qquad (1)$$

- Note that β_0 and β_1 are **unknown parameters**. Furthermore, the dataset $((y_1, x_1),...,(y_n, x_n))$ is observed by the researcher.

- We minimize equation (1) with respect to the unknown β_0 and β_1, treating $((y_1, x_1),...,(y_n, x_n))$ as a constant (in mathematical sense)

There are two first-order conditions:

- Let's take the derivative of (1) with respect to β_0 and set it equal to zero.
- Let's take the derivative of (1) with respect to β_1 and set it equal to zero.

Ordinary Least Squares (OLS) – solution/method 1

The two first-order conditions are referred to as the **normal equations**

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad (2) \qquad (2.14)$$

$$\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad (2.15)$$

(Note: there are two equations with two unknowns, which can be solved mathematically).

The solution to the system of two equations is **the OLS-estimator**: $\hat{\beta}_0$ and $\hat{\beta}_1$. See the output above

```
. reg salary exper

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  1,   524) =    6.77
       Model |  91.2751351     1  91.2751351           Prob > F      =  0.0096
    Residual |  7069.13916   524  13.4907236           R-squared     =  0.0127
-------------+------------------------------           Adj R-squared =  0.0109
       Total |  7160.41429   525  13.6388844           Root MSE      =   3.673

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0307219   .0118111     2.60   0.010      .007519    .0539247
       _cons |   5.373305   .2569919    20.91   0.000     4.868444    5.878166
------------------------------------------------------------------------------
```

Example: Stata

```
. reg salary exper

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  1,   524) =    6.77
       Model |  91.2751351     1  91.2751351           Prob > F      =  0.0096
    Residual |  7069.13916   524  13.4907236           R-squared     =  0.0127
-------------+------------------------------           Adj R-squared =  0.0109
       Total |  7160.41429   525  13.6388844           Root MSE      =   3.673

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0307219   .0118111     2.60   0.010      .007519    .0539247
       _cons |   5.373305   .2569919    20.91   0.000     4.868444    5.878166
------------------------------------------------------------------------------
```

```
. corr salary exper, cov
(obs=526)

             |   salary    exper
-------------+------------------
      salary |  13.6389
       exper |  5.65908   184.204
```

```
. display 5.65908/184.204
.0307218
```

- The denominator (variance of experience) is NOT allowed to be equal to zero: NO PERFECT MULTICOLLINEARITY
- Variance of experience is zero if all of the 526 observations of the sample have equal experience.

Ordinary Least Squares (OLS) – solution/method 2

The system of **normal equations** (2) may be solved (2 linear equations with two unknowns β_0 and β_1. This gives the solutions, which is OLS-estimator

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (2.19)$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (2.17)$$

Notes

- If the denominator of (2.19) is zero: **perfect multicollinearity**.
  - Requirement to the data set: there must be sampling variation for each of the independent variables of the regression equations. Thus, the **independent variables are non-constant**.
- Alternative formulation to (2.19):

$$\hat{\beta}_1 = \frac{Cov(x, y)}{Var(x)}$$

where

$$Cov(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$Var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

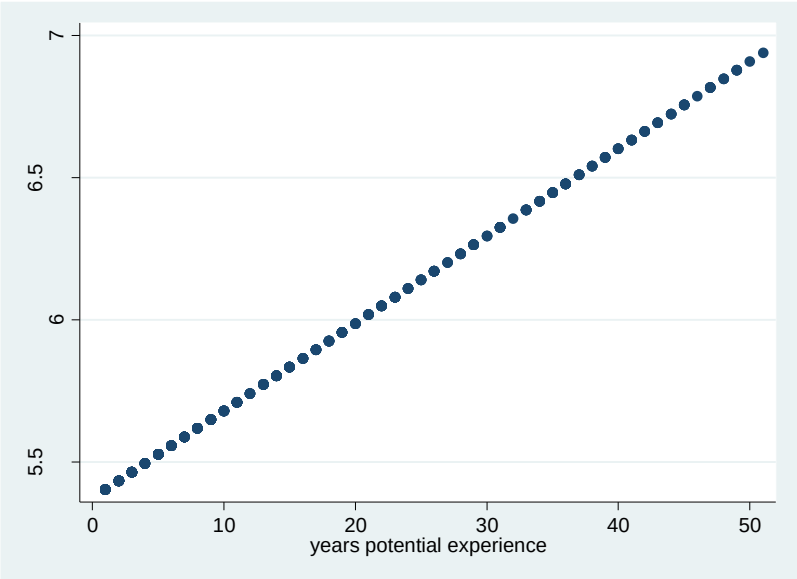
Example: Stata

```
. reg salary exper

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  1,   524) =    6.77
       Model |  91.2751351     1  91.2751351           Prob > F      =  0.0096
    Residual |  7069.13916   524  13.4907236           R-squared     =  0.0127
-------------+------------------------------           Adj R-squared =  0.0109
       Total |  7160.41429   525  13.6388844           Root MSE      =   3.673

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0307219   .0118111     2.60   0.010      .007519    .0539247
       _cons |   5.373305   .2569919    20.91   0.000     4.868444    5.878166
------------------------------------------------------------------------------

. predict uhat, resid

. sum uhat

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        uhat |       526   -2.15e-09    3.669472  -4.935471   18.71576

. corr uhat exper
(obs=526)

             |     uhat    exper
-------------+------------------
        uhat |   1.0000
       exper |  -0.0000   1.0000
```

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad i = 1,\ldots,n$$

predict yhat

graph twoway scatter yhat exper



Ordinary Least Squares (OLS) – solution/method 3

The **fitted (predicted) value** of the dependent variable for the *i*-th observation, $\hat{y}_i$ is defined as follows:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad i = 1,...,n \tag{2.20}$$

The **residual** $\hat{u}_i$ is defined as the difference between the observed y_i and its fitted $\hat{y}_i$:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \qquad i = 1,...,n \tag{2.21}$$

The **Sum of Squared Residuals** is defined as

$$SSR = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{2.22}$$

Notes:

- OLS: $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen in such a way that the residual sum of squares (*SSR*) is minimized (see Figure 2.7).
- The residual $\hat{u}_i$ can be interpreted as an estimate of the unknown error term u_i. However, the words "residual" and "error term" are no synonyms of each other!!!

Properties of the residuals:

$$\bar{\hat{u}} = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad \text{(3a)} \quad (2.14)$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i \hat{u}_i = \frac{1}{n}\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad \text{(3b)} \quad (2.15)$$

- Interpretation of (3a): OLS produces an average residual of value 0.

- Interpretation of (3b): OLS ensures a zero covariance (or correlation) between residual and each of the explanatory variables, given that the average of the residuals is zero.

Conclusions - The mechanics of Ordinary Least Squares (OLS)

1) By means of ordinary least squares for $\hat{\beta}_0$ and $\hat{\beta}_1$ all information of salary and experience is used.

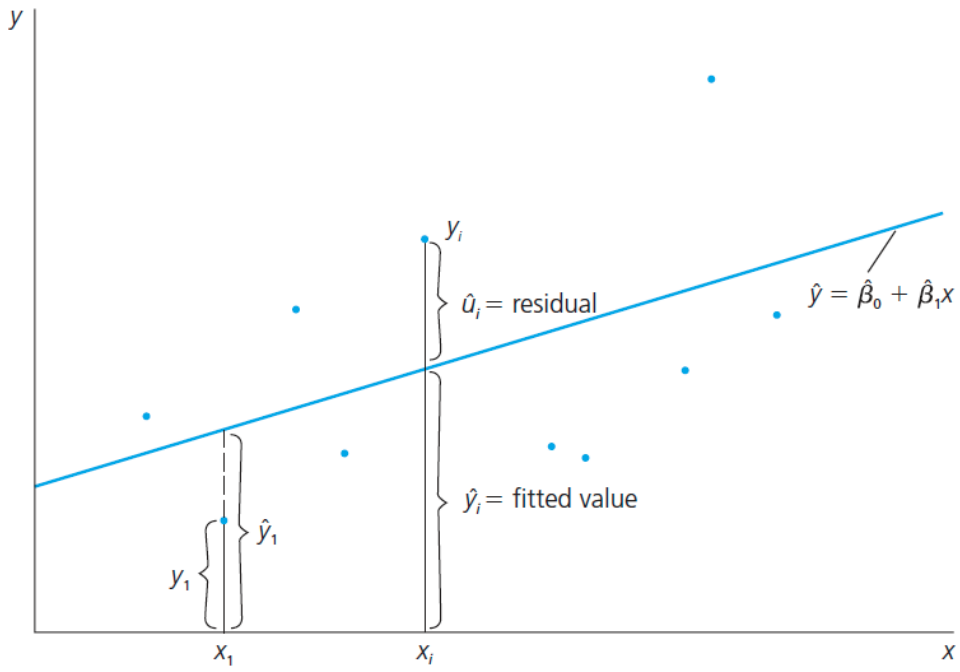
2) The residual is on average equal to zero (over all observations).

3) The residual is uncorrelated to the right-hand side variables.

4) The fitted observations of all observations (used to estimate the equation) are at the linear regression equation.

5) We can make use of $\hat{\beta}_0 + \hat{\beta}_1 x_i$ to predict the dependent variable.

FIGURE 2.4 Fitted values and residuals.



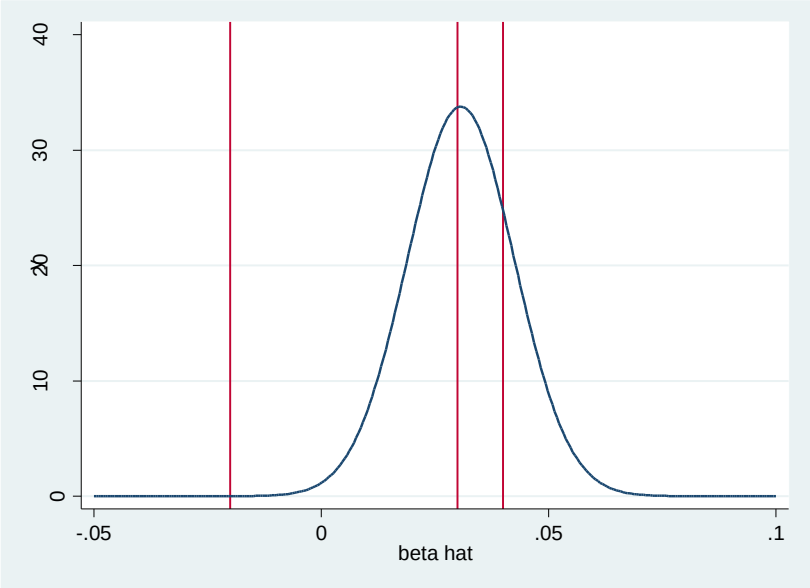
UNBIASED AND CONSISTENT ESTIMATORS

The sampling distribution of beta hat

A sampling distribution of $\hat{\beta}_1$ is sktched below.

$\hat{\beta}_1$ will be an **underestimate** if $E(\hat{\beta}) < \beta$ (the vertical line on the right-hand side of the graph).

$\hat{\beta}_1$ will be an **overestimate** if $E(\hat{\beta}) > \beta$ (the vertical line on the left-hand side of the graph).



Distribution of estimated regression parameter

Aim: introduction of the sampling distribution of $\hat{\beta}_j$.

Estimator: Algorithm (accounting rule/ formula) which provides the rule to estimate the unknown parameters. This rule can be formulated before the sample is drawn. The estimator is a random variable (because *y* is also a random variable) **with a probability distribution.**

Estimate: After a sample is drawn, one can apply the algorithm, which gives the observed values of $((y_1, x_1),...,(y_n, x_n))$. The estimate is NOT a random variable (it is a number)! Different samples will yield different OLS estimates.

- Consequence: $\hat{\beta}_1$ is a random variable that has a statistical distribution: it is based on random variables $((y_1, x_1),...,(y_n, x_n))$
- We can construct a **sampling distribution** of the OLS estimator of $\hat{\beta}_1$, which gives the statistical distribution of all possible outcomes of $\hat{\beta}_1$ for a very large number of samples (from the same population) of equal sample size *n*. (see graph on the next page)
- Note that the area below the sampling distribution of $\hat{\beta}_1$ is exactly equal to 1.
- We are interested in two features of the sampling distribution:
    - **Location of** $\hat{\beta}_1$**:**
        - Unbiased estimator (using a **small** sample)
        - Consistent estimator (using a **large** sample)
        - Question: Under which assumption is OLS an unbiased/consistent estimator?
    - **Variation of** $\hat{\beta}_1$:
        - The sampling distribution has a variance: $Var(\hat{\beta}_1)$
        - Question: Formula for $Var(\hat{\beta}_1)$?

Unbiased OLS-estimator of regression parameter

Aim: introduction of unbiased estimator of $\hat{\beta}_j$, included the four MLR-assumptions.

- **Definition:** the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ are **unbiased** if their expected values are equal to the unknown population parameters $\beta_0, \beta_1, ..., \beta_k$: $E(\hat{\beta}_j) = \beta_j$ for all j=0,1,…,k.
- **Intuition:** the distribution of the random variable $\hat{\beta}_j$ is centrered around the true unknown parameter β_j. Hence, although it is likely that the estimate $\hat{\beta}_j$ does not exactly equal β_j, on average - across many samples (of same size n) – it the estimates are exactly equal to β_j. Thus, on average $\hat{\beta}_j$ is equal to β_j.
- It is irrespective of the size of the sample: also valid for **small datasets.**
- Theorem 3.1 Wooldridge: The OLS estimators $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ are **unbiased** under the following four assumptions about the population regression model ((3.37)

MLR.1) The population regression model is linear in its parameters β_j.

(note: this assumption is about the regression parameters and not about the explanatory variables; MLR: Multivariate Linear Regression)

MLR 2) Random sample $\{(x_{1i}, x_{2i}, ..., x_{ki} y_i), i = 1, ..., n\}$ from a population whose behaviour can be described by the regression equation.

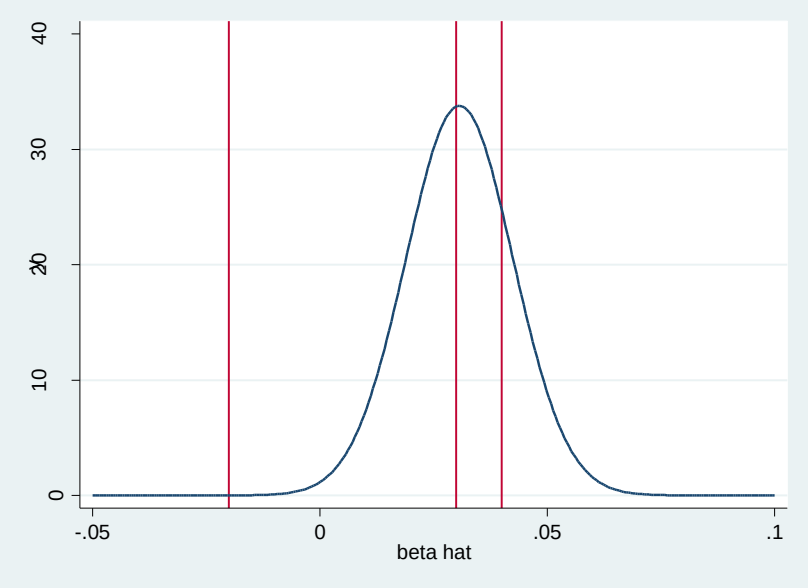
MLR 3) No perfect collinearity: 1) No perfect linear relation between the explanatory variables. 2) Sample variation in each explanatory variable.

MLR 4) $E(u \mid x_1, ..., x_k) = 0$ **The expected value of the error term *u* does not depend on the set of the right-hand side (rhs) variables.**

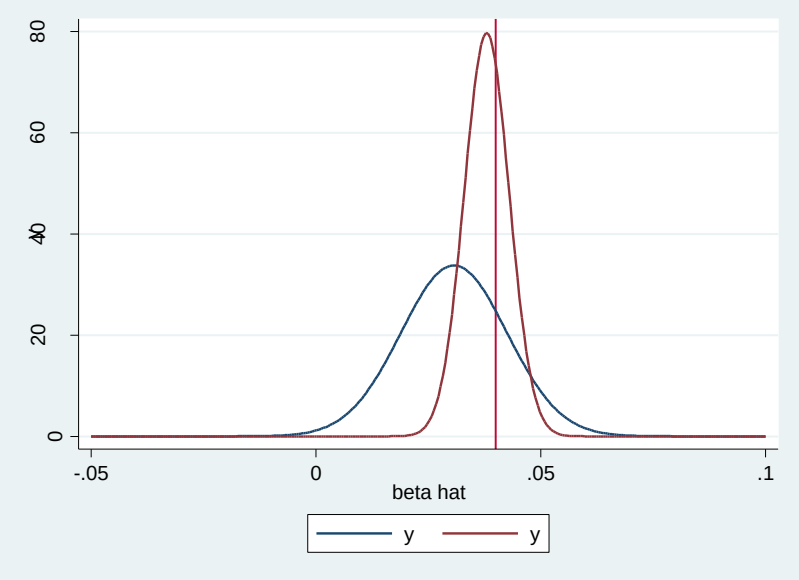
Example: consistency to the OLS estimator

```
. reg salary exper female if _n <=50

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =    7.78
       Model |  236.942957     2  118.471478           Prob > F      =  0.0012
    Residual |  716.131507    47  15.2368406           R-squared     =  0.2486
-------------+------------------------------           Adj R-squared =  0.2166
       Total |  953.074464    49  19.4504993           Root MSE      =  3.9034

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0503378   .0472735     1.06   0.292    -.0447642    .1454397
      female |  -4.069598   1.143777    -3.56   0.001     -6.37058   -1.768616
       _cons |     8.6927   1.240624     7.01   0.000     6.196886    11.18851
------------------------------------------------------------------------------

. reg salary exper female if _n <=250

      Source |       SS       df       MS              Number of obs =     250
-------------+------------------------------           F(  2,   247) =   21.85
       Model |  627.795528     2  313.897764           Prob > F      =  0.0000
    Residual |  3548.38907   247  14.3659477           R-squared     =  0.1503
-------------+------------------------------           Adj R-squared =  0.1434
       Total |   4176.1846   249  16.7718257           Root MSE      =  3.7902

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0552235   .0185188     2.98   0.003     .0187486    .0916984
      female |  -2.717782   .4810281    -5.65   0.000    -3.665222   -1.770342
       _cons |    6.83937   .4649701    14.71   0.000     5.923558    7.755181
------------------------------------------------------------------------------

. reg salary exper female if _n <=400

      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F(  2,   397) =   28.43
       Model |  714.751777     2  357.375889           Prob > F      =  0.0000
    Residual |  4989.99459   397  12.5692559           R-squared     =  0.1253
-------------+------------------------------           Adj R-squared =  0.1209
       Total |  5704.74637   399  14.2976099           Root MSE      =  3.5453

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0390148   .0132004     2.96   0.003     .0130634    .0649662
      female |  -2.381591   .3555952    -6.70   0.000    -3.080676   -1.682506
       _cons |   6.495541   .3387759    19.17   0.000     5.829522     7.16156
------------------------------------------------------------------------------

. reg salary exper female

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

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

Above: sampling distribution of beta hat (estimated regression parameter) for many random samples with equal sample size *n*.



Below: sampling distribution of beta hat (estimated regression parameter) for an increasing sample size *n* (red curve is based on a bigger number of observations)



Definition of consistency to the OLS estimator

Aim: to give an intuitive definition of consistency.

Intuition: we consider the distribution of the random variable $\hat{\beta}_j$

Definition: the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ are consistent if

1) their expected values become close to the unknown population parameters $\beta_0, \beta_1, ..., \beta_k$:

$$E(\hat{\beta}_j) \to \beta_j \quad \text{as} \quad n \to \infty$$

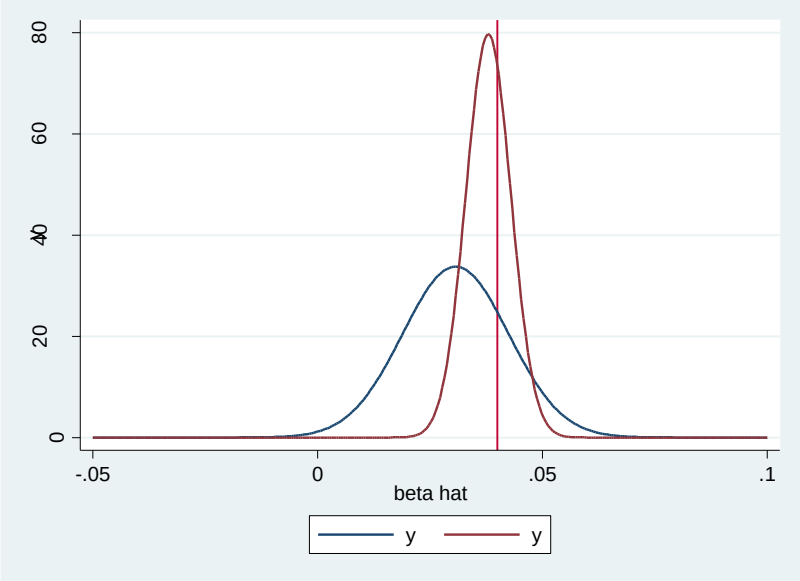
An increase of the sample size *n* leads to a better approximation to the true unknown regression parameter.

2) variance of the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ becomes smaller for a larger sample size *n*:

$$Var(\hat{\beta}_j) \to 0 \quad \text{as} \quad n \to \infty$$

Notation: $\text{plim}\hat{\beta}_j = \beta_j$

- Implication: it can be used only for **large datasets.**



Application of consistency to the OLS estimator

Aim: under which assumptions is there a consistent OLS estimator? These assumptions are weaker than those for an unbiased estimator.

Multivariate regression equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u \qquad (5.1)$$

Remember:

$$\underbrace{E(u \mid x) = 0}_{\text{Assumption MLR.4 for unbiased estimator}} \Rightarrow \underbrace{\begin{cases} Eu = 0 \\ Cov(u, x) = 0 \end{cases}}_{\text{Assumption MLR.4' for consistent estimator}} \qquad E(u \mid x) = 0 \not\Leftarrow \begin{cases} Eu = 0 \\ Cov(u, x) = 0 \end{cases}$$

strong assumption	weak assumption

- Unbiased estimator: in particular useful for small samples, but also for large samples.
- Consistent estimator: not useful for small samples, only applicable to large data sets.

Under assumptions MLR.1-MLR.4, the OLS-estimator $\hat{\beta}_j$ is **consistent** for β_j, for all j=0,1,…,k

MLR.1) The population regression model is linear in its parameters β_j.

MLR 2) Observations $\{(x_{1i}, x_{2i}, ..., x_{ki} y_i),\ i = 1, ..., n\}$ are randomly drawn from a population whose behaviour can be described by the regression equation.

MLR 3) No perfect collinearity: 1) No perfect linear relation between the explanatory variables. 2) Sample variation in each explanatory variable

MLR.4') Zero mean and zero correlation $E(u) = 0$ **and** $Cov(x_j, u) = 0$ **for *j*=0,1,…,*k***

Conclusion – unbiased or consistent estimator

- After a regression, we consider whether the four MLR assumptions have been fulfilled.

- For an unbiased estimator: small sample property. It is based on the following. The true unknown parameter is not equal to the estimated parameter. However, the average of the estimated regression parameter over many random samples (with same size *n*) is equal to the unknown parameter.

- For a consistent estimator: large sample property. As the sample size increases, a) the estimated parameter tends toward the unknown regression parameter; b) the variance of the estimated parameter becomes smaller.

QUESTION 1: OMITTED VARIABLES? ANSWER: IS OFTEN A PROBLEM

QUESTION 2: HETEROSKEDASTICITY OF WHAT?

$$salary = \beta_0 + \beta_1 female + \beta_2 exper + u$$

- We can test for heteroskedasticy using the Breusch Pagan test (see additional material and the tutorial)
- Solution: We can compute the robust standard errors

```
. reg  female salary exper, robust

Linear regression                                      Number of obs =     526
                                                       F(  2,   523) =   37.69
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1157
                                                       Root MSE      =  .47113

------------------------------------------------------------------------------
             |               Robust
      female |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      salary |  -.0459987   .0054495    -8.44   0.000    -.0567043   -.0352932
       exper |  -.0001205   .0015547    -0.08   0.938    -.0031747    .0029338
       _cons |   .7523505   .0436496    17.24   0.000     .6666004    .8381006
------------------------------------------------------------------------------
```

QUESTION 3: ASSUMPTION OF NORMALITY OF WHAT?

- Small sample ($n < 50$): assume normality of the error term u of the population regression model: use the t-test and F-test
- Large samples: apply Central Limit Theorem. No assumption of normality of error term u is needed.

QUADRATIC FORM

Quadratic form

Aim: to introduce a quadratic form.

$$salary = \beta_0 + \beta_1 female + \beta_2 exper + \beta_3 exper^2 + u$$

The shape of the parabola depends on the following:

- $\beta_2 > 0$ and $\beta_3 < 0$: inverse U-shape
- $\beta_2 < 0$ and $\beta_3 > 0$: U-shape
-

- WHY QUADRATIC FORM? Answer: Non-linear relationship between experience and salary

The difference in the marginal effects of these two specifications

Aim: to interpret the parameters of a quadratic form.

First, we consider

$$salary = \beta_0 + \beta_1 female + \beta_2 exper + u$$

The marginal effect: $\frac{\partial salary}{\partial exper} = \beta_2$ (=first derivative of salary with respect to experience)

$$salary = \beta_0 + \beta_1 female + \beta_2 exper + \beta_3 exper^2 + u$$

In the second model the derivative is: $\frac{\partial salary}{\partial exper} = \beta_2 + 2\beta_3 exper$

- The marginal effect of experience on the salary changes for each year of experience
- At how many years of experience should we calculate the marginal effect? For instance, we can calculate the marginal effect of *experience* on *salary* at the average level of experience across the sample

$$\frac{\partial salary}{\partial exper} = \beta_2 + 2\beta_3 \overline{exper}$$

Example: How to calculate the marginal effect in Stata?

Aim: Stata provides the estimated parameter of the marginal effect as well as its standard error, which allows the application of statistical tests.

- Apply the stata command **nlcom** (abbreviation of: nonlinear combination of estimated parameters).
- _b[*varname*] refers to the estimated parameter of the last regression equation for the variable *varname*.

```
. sum exper, detail

                 years potential experience
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              1
 5%            1              1
10%            2              1       Obs                 526
25%            5              1       Sum of Wgt.         526

50%         13.5                      Mean           17.01711
                        Largest       Std. Dev.      13.57216
75%           26             49
90%           38             49       Variance       184.2035
95%           43             50       Skewness       .7068652
99%           49             51       Kurtosis       2.357318


. reg salary female exper exper2

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  3,   522) =   42.81
       Model |  1413.97372     3  471.324575           Prob > F      =  0.0000
    Residual |  5746.44098   522  11.0085076           R-squared     =  0.1975
-------------+------------------------------           Adj R-squared =  0.1929
       Total |  7160.41471   525  13.6388852           Root MSE      =  3.3179

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.394192   .2901199    -8.25   0.000    -2.964138   -1.824246
       exper |   .2809817   .0386219     7.28   0.000     .2051083    .3568551
      exper2 |  -.0058216   .0008505    -6.85   0.000    -.0074924   -.0041508
       _cons |    5.01779   .3613739    13.89   0.000     4.307864    5.727716
------------------------------------------------------------------------------
```

```
. nlcom _b[exper]+2*_b[exper2]*13.5

       _nl_1:  _b[exper]+2*_b[exper2]*13.5

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   .1237982   .0177299     6.98   0.000     .0890482    .1585482
------------------------------------------------------------------------------
```

- The marginal effect at the median of experience (13.5) is 0.0124.
- The standard error of the marginal effect is 0.0177.
- The *t*-value is 6.98 (which is larger than the critical value for *t*-tests at the 5% level, 1.96).
- Conclusion, the marginal effect of experience on salary, calculated at the median value of experience, is statistically different from zero.

```
. nlcom _b[exper]+2*_b[exper2]*17

       _nl_1:  _b[exper]+2*_b[exper2]*17

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   .0830469   .0134638     6.17   0.000     .0566584    .1094354
------------------------------------------------------------------------------
```

- The marginal effect at the average of experience (17) is 0.0830.
- The standard error of the marginal effect is 0.0135.
- The *t*-value is 6.17 (which is larger than the critical value for *t*-tests at the 5% level, 1.96).
- Conclusion, the marginal effect of experience on salary, calculated at the average level of experience, is statistically different from zero.

DUMMY VARIABLES AND INTERACTION TERMS

Interaction terms

Aim: to show that by using interaction terms, we can have different effects for different groups. We must apply marginal effects.

- The next example adds an interaction term (cross-term) between *female* and *experience*.

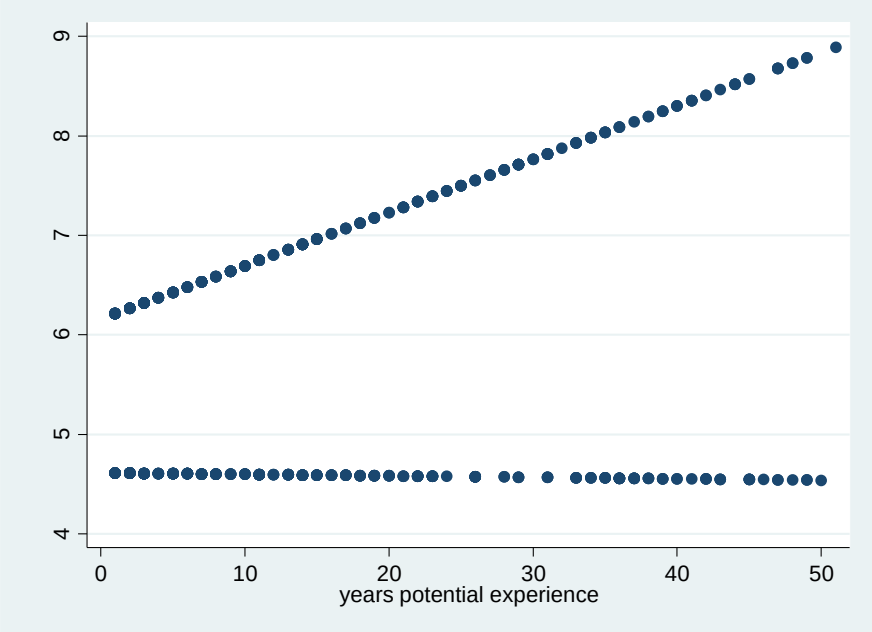
$$salary = \beta_0 + \beta_1 female + \beta_2 exper + \beta_3 female \cdot exper + u$$

```
. gen female_exper = female*exper
. reg salary female exper female_exper

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  3,   522) =   27.31
       Model |  971.286375     3  323.762125            Prob > F      =  0.0000
    Residual |  6189.12833   522  11.8565677           R-squared     =  0.1356
-------------+------------------------------           Adj R-squared =  0.1307
       Total |  7160.41471   525  13.6388852           Root MSE      =  3.4433

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -1.546547   .4818603    -3.21   0.001     -2.49317    -.599923
       exper |   .0536048   .0154372     3.47   0.001     .0232782    .0839314
female_exper |  -.0550699    .022175    -2.48   0.013     -.098633   -.0115068
       _cons |   6.158276   .3416741    18.02   0.000      5.48705    6.829501
------------------------------------------------------------------------------
```

- The interaction-term is statistically significant at the 5% level (α=0.05) with a t-statistic of -2.48. The *p*-value is equal to 0.013.

. predict salaryp
. twoway scatter salaryp exper

The flat line are the fitted values of salary for the females. The other, increasing line are the fitted values for the males.



The marginal effect of experience on *salary*

$$salary = \beta_0 + \beta_1 female + \beta_2 exper + \beta_3 female \cdot exper + u$$

$$\frac{\partial salary}{\partial exper} = \beta_2 + \beta_3 female$$ (derivative of salary with respect to experience)

- The estimated marginal effect of experience is

for females (*female*=1):

$$\hat{\beta}_2 + \hat{\beta}_3 \cdot 1 = 0.0536048 - 0.0550699 = -0.0014651$$

for males (*female* = 0)

$$\hat{\beta}_2 + \hat{\beta}_3 \cdot 0 = 0.0536048$$

Example

- In Stata; for females (*female*=1):

```
. nlcom _b[exper]+_b[female_exper]

       _nl_1:  _b[exper]+_b[female_exper]

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   -.0014651   .0159193    -0.09   0.927    -.0326663     .029736
------------------------------------------------------------------------------
```

Conclusion: returns to experience (the marginal effect of *experience* on *salary*) are larger for men than for women.

Categorical variables

Categorical dummy variables

Aim: to introduce multiple dummy variables as an explanatory variable.

- It is useful to separate the sample in to various classes of experience.

  $dume1 = 1$ if $0 \leq exper < 5$ and 0 otherwise
  $dume2 = 1$ if $5 \leq exper < 10$ and 0 otherwise
  $dume3 = 1$ if $10 \leq exper < 15$ and 0 otherwise
  $dume4 = 1$ if $15 \leq exper < 20$ and 0 otherwise
  $dume5 = 1$ if $20 \leq exper$ and 0 otherwise

- In stata:

```
gen dume1 = 0
gen dume2 = 0
gen dume3 = 0
gen dume4 = 0
gen dume5 = 0
. replace dume1 = 1 if exper >=0 & exper <5
(112 real changes made)
gen dume2 = 0
. replace dume2 = 1 if exper >=5 & exper <10
(94 real changes made)
. replace dume3 = 1 if exper >=10 & exper <15
(77 real changes made)
replace dume4 = 1 if exper >=15 & exper <20
. replace dume5 = 1 if exper >=20
(191 real changes made)
```

$$salary = \beta_0 + \beta_1 female + \beta_2 dume1 + \beta_3 dume2 + \beta_4 dume3 + \beta_5 dume4 + u$$

The regression parameters can be estimated by OLS:

```
. reg salary female dume1-dume4
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  5,   520) =   21.00
       Model |  1203.05757     5  240.611513            Prob > F      =  0.0000
    Residual |  5957.35714   520   11.456456           R-squared     =  0.1680
-------------+------------------------------           Adj R-squared =  0.1600
       Total |  7160.41471   525  13.6388852           Root MSE      =  3.3847
------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.427586    .296175    -8.20   0.000    -3.009433    -1.84574
       dume1 |  -2.145217    .403492    -5.32   0.000    -2.937891   -1.352542
       dume2 |  -.6860064   .4267015    -1.61   0.109    -1.524277    .1522643
       dume3 |   .0748075   .4569837     0.16   0.870    -.8229536    .9725687
       dume4 |  -.3286454   .5302376    -0.62   0.536    -1.370317    .7130258
       _cons |   7.660038   .2773947    27.61   0.000     7.115086     8.20499
------------------------------------------------------------------------------
```

F-test $H_0 : \beta_2 = 0, \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$; $H_1 : H_0$ is not true

```
. test dume1 dume2 dume3 dume4
 ( 1)  dume1 = 0
 ( 2)  dume2 = 0
 ( 3)  dume3 = 0
 ( 4)  dume4 = 0

       F(  4,   520) =    8.18
            Prob > F =    0.0000
```

```
. testparm dume*
 ( 1)  dume1 = 0
 ( 2)  dume2 = 0
 ( 3)  dume3 = 0
 ( 4)  dume4 = 0

       F(  4,   520) =    8.18
            Prob > F =    0.0000
```

F-test $H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$; $H_1 : H_0$ is not true

```
. test dume2 dume3 dume4
 ( 1)  dume2 = 0
 ( 2)  dume3 = 0
 ( 3)  dume4 = 0

       F(  3,   520) =    1.06
            Prob > F =    0.3667
```

Conclusions of *F*-tests:

1) Experience has a significant effect on salary.
2) It seems that only the first category (*dume1*) has an effect on *salary*

Aim: to interpret the estimated parameters.

$$salary = \beta_0 + \beta_1 female + \beta_2 dume1 + \beta_3 dume2 + \beta_4 dume3 + \beta_5 dume4 + u$$

- Explanation coefficient β_2: workers with between 0 and 5 years of experience have a β_2 higher salary, RELATIVE TO THE REFERENCE CATEGORY *dume5* (20 years and older) and ceteris paribus on female.
- t-value of $\hat{\beta}_2$: -2.14. At α=0.05, the salary of persons in this category is statistically significant different from the reference category (*dume5*).
- t-value of $\hat{\beta}_3$: -0.69. At α=0.05, the salary of persons in this category is statistically insignificant different from the reference category (*dume5*).
- t-value of $\hat{\beta}_4$: 0.07. At α=0.05, the salary of persons in this category is statistically insignificant different from the reference category (*dume5*).
- t-value of $\hat{\beta}_5$: -0.33. At α=0.05, the salary of persons in this category is statistically insignificant different from the reference category (*dume5*).

- Conclusions:
  - Experience has a significant effect on salary (*F*-test)
  - Only the lowest experience category has an individually significant effect on *salary*.
  - The salary is about $ 2.15 lower in this category than in the reference category.
  - The salary in higher categories (2,3, and 4) cannot be distinguished from the reference category (5).

Alternative model specification, reference group = 1
Aim: to show that the choice of the reference category is irrelevant.

The specification is different from above, but it delivers the same results:

$$salary = \beta_0 + \beta_1 female + \beta_2 dume2 + \beta_3 dume3 + \beta_4 dume4 + \beta_5 dume5 + u$$

Note: this specification should yield the same conclusions as before when the reference group = 5.

```
. reg salary female dume2-dume5

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  5,   520) =   21.00
       Model |  1203.05757     5  240.611513            Prob > F      =  0.0000
    Residual |  5957.35714   520   11.456456           R-squared     =  0.1680
-------------+------------------------------           Adj R-squared =  0.1600
       Total |  7160.41471   525  13.6388852           Root MSE      =  3.3847

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.427586    .296175    -8.20   0.000    -3.009433    -1.84574
       dume2 |    1.45921    .473538     3.08   0.002     .5289275    2.389493
       dume3 |   2.220024   .5012954     4.43   0.000     1.235211    3.204837
       dume4 |   1.816571   .5680172     3.20   0.001     .7006807    2.932462
       dume5 |   2.145217    .403492     5.32   0.000     1.352542    2.937891
       _cons |   5.514821   .3547028    15.55   0.000     4.817995    6.211648
------------------------------------------------------------------------------

. testparm dume*

 ( 1)  dume2 = 0
 ( 2)  dume3 = 0
 ( 3)  dume4 = 0
 ( 4)  dume5 = 0

       F(  4,   520) =    8.18
            Prob > F =    0.0000
```

- Note that the parameter estimate on *dume5* is the negative of the parameter estimate on *dume1* above.
- Note that the value of the *F*-statistic (8.18) is equal to the value of the *F*-statistic above.
- The regression results for the dummy variables are one to one and the conclusions will not differ.

Prove that this alternative specification with reference group = 1 yields the same conclusions as when the reference group is 5.

(1) Use the fact that the dummy variables all add up to 1.

(2) Rewrite the first model

$$\begin{aligned} salary &= \beta_0 + \beta_1 female + \beta_2 dume2 + \beta_3 dume3 + \beta_4 dume4 + \beta_5 dume5 + u \\ &= \beta_0 + \beta_1 female + \beta_2 dume2 + \beta_3 dume3 + \beta_4 dume4 \\ &\quad + \beta_5((1 - dume1 - dume2 - dume3 - dume4) + u \\ &= (\beta_0 + \beta_5) + \beta_1 female - \beta_5 dume1 + (\beta_2 - \beta_5) dume2 \\ &\quad + (\beta_3 - \beta_5) dume3 + (\beta_4 - \beta_5) dume4 + u \end{aligned}$$

And this can be written as:

$$salary = \gamma_1 + \beta_1 female + \gamma_2 dume1 + \gamma_3 dume2 + \gamma_4 dume3 + \gamma_5 dume4 + u$$

With, for instance, $\gamma_3 = \beta_2 - \beta_5$

- Conclusions:
  - *Experience* has an effect on *salary* (*F*-test)
  - The effects of *experience* on *salary* in all categories are different from the reference category.
  - Higher *experience* categories (2,3, and 4) cannot be distinguished from category 5 (see the 3 tests below): all *p*-values are above 0.05.

```
. test dume2 = dume5

 ( 1)  dume2 - dume5 = 0

       F(  1,   520) =    0.44
            Prob > F =    0.5057

. test dume3 = dume5

 ( 1)  dume3 - dume5 = 0

       F(  1,   520) =    1.15
            Prob > F =    0.2833

. test dume4 = dume5

 ( 1)  dume4 - dume5 = 0

       F(  1,   520) =    0.05
            Prob > F =    0.8158
```

Conclusions

- How to start an empirical research project?
- See steps 1 – 5: statistics, linear regression, etc.
- Major question: unbiased estimator?
- Reason: omitted variables/ confounder
- Solution: include additional explanatory variables
- Consider the assumption of normality of error term: small sample/ large sample
- Compute robust standard errors of the estimated regression parameters