Review of OLS Regression
M.Sc. Politics and Policy Analysis

Francesco Mattioli

francesco.mattioli@unibocconi.it

Outline

Regression Analysis

Regression is a statistical tool used to study the relationship between:

- **Dependent variable (y)**: outcome/response variable that we want to predict/explain (also called **regressand** or **left-hand-side variable**)
- **Independent variable(s) (x)**: predictor/explanatory variable(s) used to explain the dependent variable (also called **regressor** or **right-hand-side variable**)

Goals of regression analysis:

- Formalize the relationship between dependent and independent variables (**modelling**)
- Explain the impact of changes of an independent variable on the dependent variable (**inference**)
- Predict the value of the dependent variable based on the value of, at least, one independent variable (**prediction**)

Linear Population Model

(Bivariate) **linear population model**:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

– y is a **linear** function of x (Δy is assumed to be influenced linearly by Δx)

– **Error term**, ε, accounts for other factors that affect y

– β_0 is the **intercept**/constant term, β_1 is the **slope** parameter

– 'True' population parameters β_0 and β_1 are **theoretical**, cannot be observed

– Population parameters can be **estimated** with sample data under specific assumptions

– Estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are **unbiased** thanks to the *mean conditional independence* assumption ($E\left(\varepsilon|x\right) = 0$)

Linear Regression Model

Estimation problem: draw a **random sample** of size n from the population to estimate β_0 and β_1

Linear regression model:

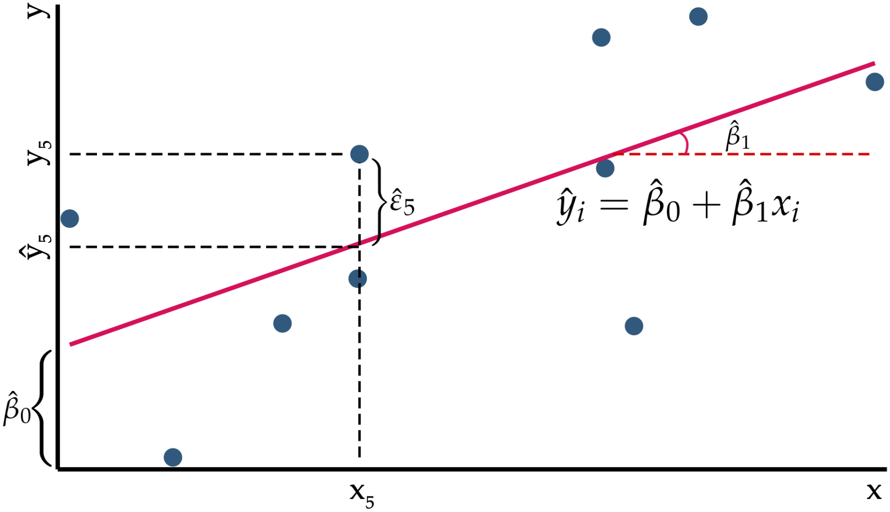
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

After estimation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- $\hat{y}_i$: **predicted** value of y for observation $i = \{1, ..., n\}$
- $\hat{\beta}_0$, $\hat{\beta}_1$: **estimates** of intercept and slope parameters
- x_i: **observed** value of x for observation $i = \{1, ..., n\}$
- $\hat{\beta}_0 + \hat{\beta}_1 x_i$: estimated **regression line**
- $\hat{\varepsilon}_i$: estimate of the error term, i.e. **residual** (it holds $\hat{\varepsilon}_i = y_i - \hat{y}_i$)

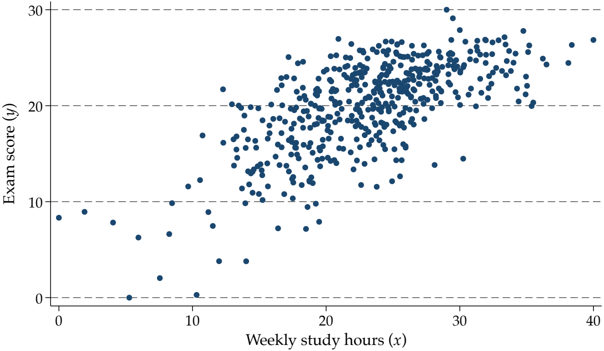
Estimation of regression line – Graphical representation



Example. Estimation of regression line and prediction of y

- We suspect that students' scores depend on how many hours *per* week they spend studying
- Simulated data for a sample of $n = 500$ observations

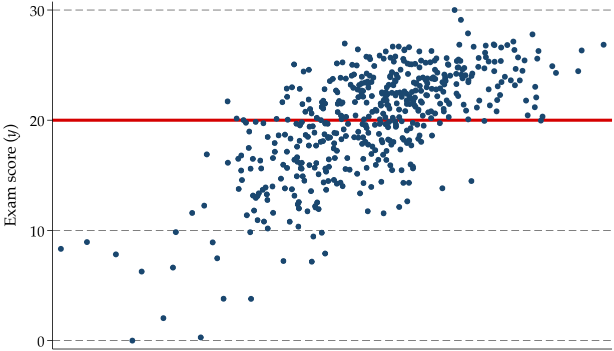
Study (x)	Score (y)
5	13
7	15
15	16
20	20
22	21
35	27
37	28
...	...



- We want to obtain a good prediction of students' scores

Example. Estimation of regression line and prediction of y

- Suppose to ignore information on study hours (x)
- The regression line becomes $\hat{y}_i = \hat{\beta}_0 + \cancel{\hat{\beta}_1 x_i}$ (i.e. slope is 0)

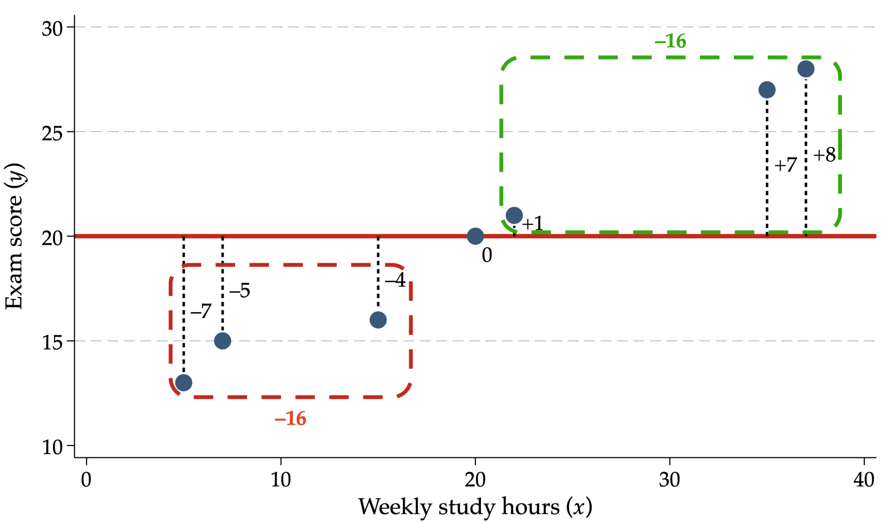


- The best prediction is the mean of y: $\hat{\beta}_0 = 20$
- How good is this prediction? How well does the regression line **fit** the observed data points?

Example. Estimation of regression line and prediction of y

- Observed data points do not fall on the regression line, but they lie either above or below it
- To evaluate how well the line fits the data we can measure the **deviation** from the data points to the line
- Since $\hat{\varepsilon}_i = y_i - \hat{y}_i$, residuals $\hat{\varepsilon}_i$ measure such deviation
- Try to sum all residuals: the smaller the sum of residuals, the better the fit?
- Let's focus on a few data points for illustration

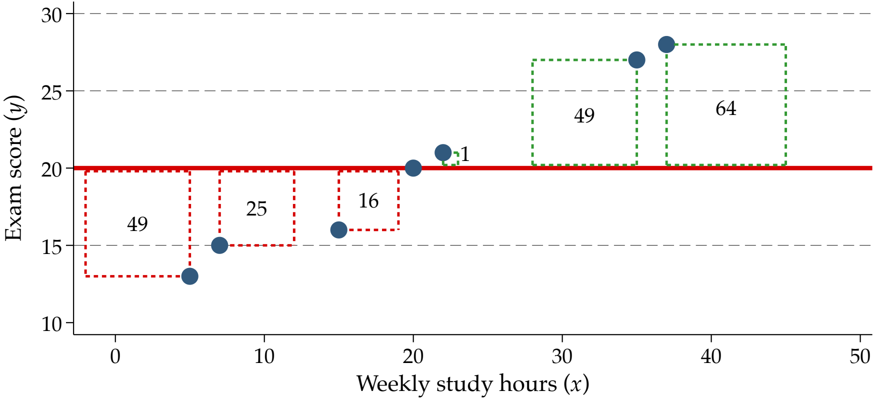
Example. Estimation of regression line and prediction of y



Example. Estimation of regression line and prediction of y

- Positive and negative residuals offset each other $(16 - 16 = 0)$
- It holds that $E(\hat{\varepsilon}_i) = 0$ if β_0 is included in the regression model
- We need another summary measure of deviation from data points to the regression line to evaluate the fit of the line
- **Squaring** the residuals makes all deviations positive and emphasizes large ones

Example. Estimation of regression line and prediction of y



- The appropriate measure to evaluate how well the line fits the data is the **Residual Sum of Squares (RSS)**

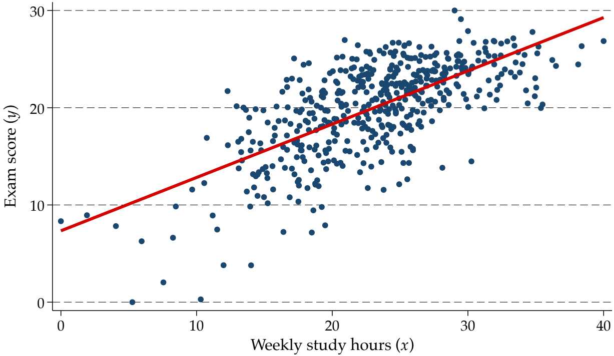
$$\boxed{49} + \boxed{25} + \boxed{16} + \boxed{1} + \boxed{49} + \boxed{64} = 204$$

- The smaller RSS, the better the fit

- The **Ordinary Least Squares (OLS)** method estimates the parameters of the regression line that **minimize RSS**

Ordinary Least Squares (OLS)

- Consider now information on study hours (x)
- The regression line becomes $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ (i.e. slope is $\neq 0$)



- How well does the new regression line fit the observed data points?

Ordinary Least Squares (OLS)

- **Residual Sum of Squares (RSS)**:

$$RSS = \sum_{i=1}^{n} (\hat{\varepsilon}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2$$

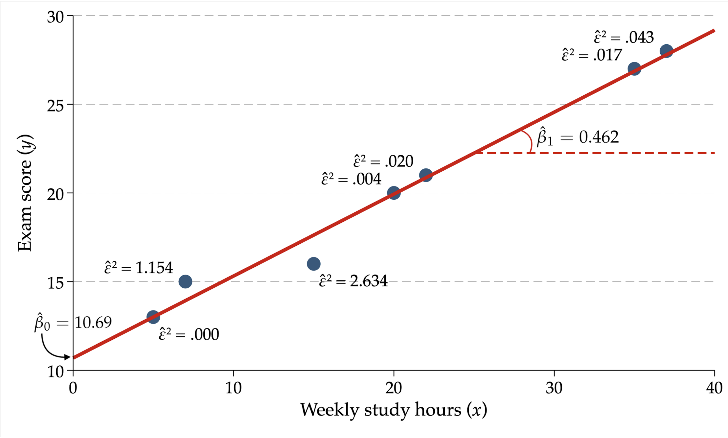
- **OLS** estimators of parameters:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{Cov(x,y)}{Var(x)}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Ordinary Least Squares (OLS)

- Let's focus again on a few data points for illustration: fit the new regression line and compute RSS



$$\hat{\beta}_0 = 10.69 \qquad \hat{\beta}_1 = 0.462 \qquad RSS = 3.874$$

Interpretation of OLS estimates

Intercept

- $\hat{\beta}_0$ is the estimated average value of y when $x = 0$ (e.g. studying 0 hours would lead to an exam score of 10.69/30)
- Misleading/meaningless if 0 is an unlikely/impossible value for x
- Must be included to ensure that (i) $E(\hat{\varepsilon}_i) = 0$, and that (ii) the regression line passes through the point with coordinates $(\overline{y}, \overline{x})$

Slope

- $\hat{\beta}_1$ is the estimated average change in y as a result of a one-unit change in x (e.g. studying one more hour would lead to an increase in the exam score of 0.46 units out of 30)
- Also known as marginal effect of x in a linear model, it is constant across values of x (e.g. same effect when moving from 5 to 6 hours and from 39 to 40 hours)

Goodness of fit

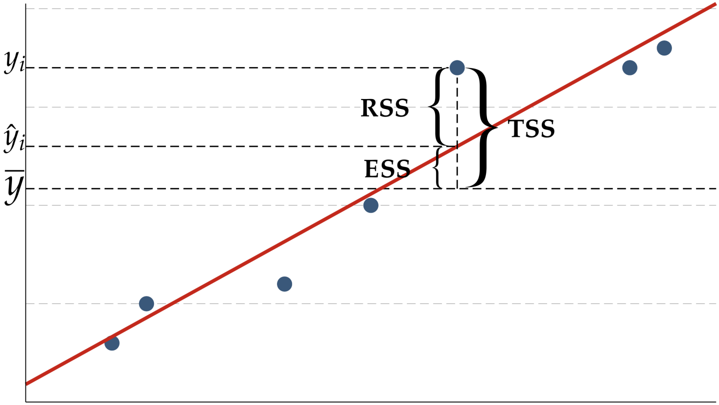
- A good regression line explains most of the sample variation in y. How to measure how much variation the line actually explains?
- Total sample variation in y can be measured through the **Total Sum of Squares** (**TSS**, i.e. a modified version of $Var(y)$):

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 = ESS + RSS$$

- TSS can be decomposed in a part of variation that is explained by the model (**Explained/Regression/Model Sum of Squares**, **ESS**) and part attributable to factors other than x, captured by the residuals (**Residual Sum of Squares**, **RSS**):

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Decomposition of variation (intuition for one value of y)



Goodness of fit

- **Coefficient of determination**: measures the fraction of total variation in y explained by the model:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

- If β_0 is included in the regression model: $0 \leq R^2 \leq 1$

- If only one indep. variable x is included in the regression model: $R^2 = Corr\,(y,x)^2$

- If k indep. variables x are included in the regression model, R^2 increases by construction without necessarily improving explanation, and should be **adjusted**:

$$\overline{R}^2 = 1 - \frac{\frac{RSS}{n-k-1}}{\frac{TSS}{n-1}} \leq R^2$$

Multiple regression model

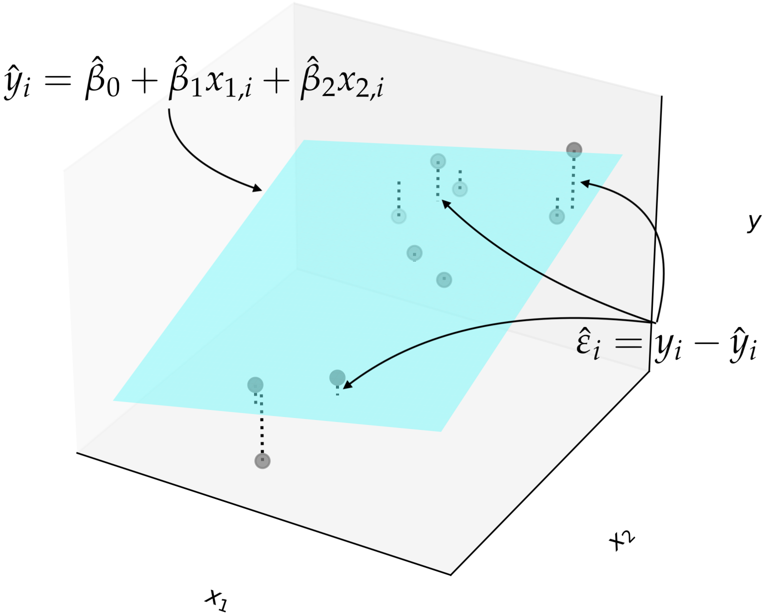
- A single independent variable x might not be sufficient to explain y (e.g. the exam score might depend on study hours, parental education, gender, etc.)

- The regression model can be improved by including other k independent variables, getting the following form:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_k x_{k,i} + \varepsilon_i$$

- $\hat{\beta}_1$ is the estimated average change in y as a result of a one-unit change in x_1 holding constant the characteristics x_2 to x_k

Estimation of a regression plane – Graphical representation



OLS estimators

$\hat{\beta}_k$ is an **estimator** of the true population parameter β_k. As seen in Module 1, this implies:

- **Sampling distribution**: $\hat{\beta}_k \approx N\left(\beta_k, \sigma^2_{\beta_k}\right)$ as $n \to \infty$
- **Hypothesis testing**: $H_0 : \beta_k = 0 \quad \text{vs} \quad H_1 : \beta_k \neq 0$
- **T-statistic**: $t = \frac{\hat{\beta}_k - 0}{SE\left(\hat{\beta}_k\right)} = \frac{\hat{\beta}_k - 0}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_k}}} \sim N(0,1)$
- **Confidence intervals**: $CI_{0.95} = \hat{\beta}_k \pm 1.96 \cdot SE\left(\hat{\beta}_k\right)$

OLS estimation of: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$

```
. regress trstprt eduyrs agea if cntry=="IT"

      Source |       SS           df       MS      Number of obs   =     2,483
-------------+----------------------------------   F(2, 2480)      =     14.29
       Model |  144.747683         2  72.3738417   Prob > F        =    0.0000
    Residual |  12563.2878     2,480  5.06584184   R-squared       =    0.0114
-------------+----------------------------------   Adj R-squared   =    0.0106
       Total |  12708.0354     2,482  5.12007874   Root MSE        =    2.2507

------------------------------------------------------------------------------
     trstprt | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      eduyrs |   .0473376    .011471     4.13   0.000     .0248439    .0698312
        agea |  -.0041704   .0026174    -1.59   0.111     -.009303    .0009622
       _cons |   2.712842   .2349016    11.55   0.000     2.252218    3.173465
------------------------------------------------------------------------------
```

$$R^2 = \frac{ESS}{TSS} = \frac{144.75}{12708} = 0.0114$$

$$\overline{R}^2 = 1 - \frac{RSS}{n-k-1} \Big/ \frac{TSS}{n-1} = 1 - \frac{12563}{2483-2-1} \Big/ \frac{12708}{2483-1} = 0.0106$$

OLS estimation of: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$

```
. regress trstprt eduyrs agea if cntry=="IT"

      Source |       SS           df       MS      Number of obs   =     2,483
-------------+----------------------------------   F(2, 2480)      =     14.29
       Model |  144.747683         2  72.3738417   Prob > F        =    0.0000
    Residual |  12563.2878     2,480  5.06584184   R-squared       =    0.0114
-------------+----------------------------------   Adj R-squared   =    0.0106
       Total |  12708.0354     2,482  5.12007874   Root MSE        =    2.2507

------------------------------------------------------------------------------
     trstprt | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      eduyrs |   .0473376    .011471     4.13   0.000     .0248439    .0698312
        agea |  -.0041704   .0026174    -1.59   0.111     -.009303    .0009622
       _cons |   2.712842   .2349016    11.55   0.000     2.252218    3.173465
------------------------------------------------------------------------------
```

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - 0}{SE\left(\hat{\beta}_1\right)} = \frac{0.047}{0.0115} = 4.13 > 1.96 \Rightarrow \textbf{Reject } H_0$$

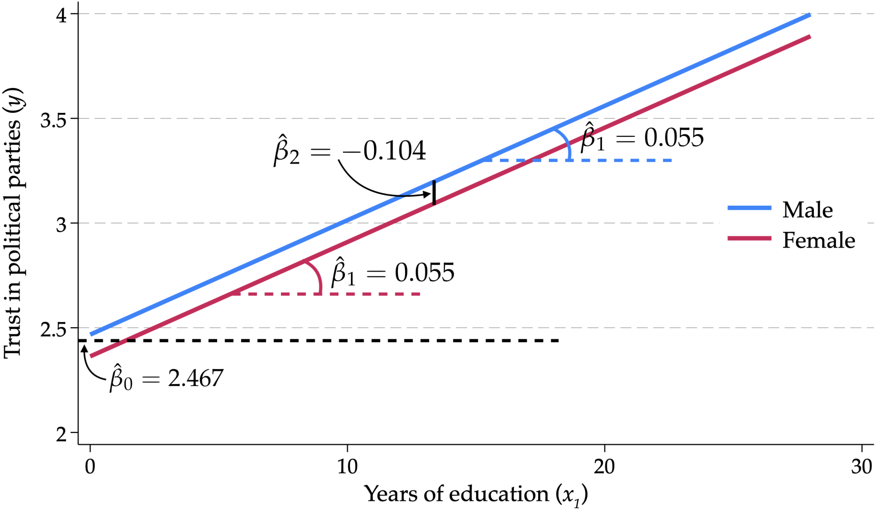
$$CI_{0.95,\hat{\beta}_1} = \hat{\beta}_1 \pm 1.96 \cdot SE\left(\hat{\beta}_1\right) = 0.047 \pm 1.96 \cdot 0.0115 = [0.0248; 0.0698]$$

Dummy variables

- A **dummy variable** is a binary variable taking value 0 or 1, used to indicate the absence/presence of an individual characteristic (e.g. female/male, employed/unemployed)
- Suppose that in the model $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$, $x_{2,i}$ is a gender dummy variable equal to 0 if i is male or to 1 if i is female
- $\hat{\beta}_2$ is the estimated average change in y as a result of the characteristic x_2 being present, compared to when x is absent
- The estimated regression model becomes:
  - $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \cancel{\hat{\beta}_2}$ if i is male
  - $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2$ if i is female
- We can visualize two separate regression lines with same slope but different intercept, the difference corresponding to the value $\hat{\beta}_2$

Dummy variables – Graphical representation

$$\hat{y}_i = 2.467 + 0.055 \cdot x_{1,i} - 0.104 \cdot x_{2,i}$$



Dummy variable trap

- Suppose we want to evaluate whether, in Italy, trust in political parties differs across residents in the North (N), Center (C), or South (S)
- We create 3 dummy variables, each taking value 1 if i lives in the respective area N, C, or S and 0 otherwise (i.e. x_N,x_C,x_S), and estimate the model: $y_i = \beta_0 + \beta_N x_{N,i} + \beta_C x_{C,i} + \beta_S x_{S,i} + \varepsilon_i$
- **Dummy variable trap**: including all 3 dummies at the same time determines **perfect multicollinearity**; the information provided by any two dummies is sufficient to determine the value of the third dummy, which becomes redundant

Dummy variable trap

- Suppose we want to evaluate whether, in Italy, trust in political parties differs across residents in the North (N), Center (C), or South (S)
- We create 3 dummy variables, each taking value 1 if i lives in the respective area N, C, or S and 0 otherwise (i.e. x_N,x_C,x_S), and estimate the model: $y_i = \beta_0 + \beta_N x_{N,i} + \beta_C x_{C,i} + \cancel{\beta_S x_{S,i}} + \varepsilon_i$
- **Dummy variable trap**: including all 3 dummies at the same time determines **perfect multicollinearity**; the information provided by any two dummies is sufficient to determine the value of the third dummy, which becomes redundant
- If i doesn't live neither in the North ($x_{N,i} = 0$) nor in the Center ($x_{C,i} = 0$), then i lives necessarily in the South ($x_{S,i} = 1$) and there is no need to include x_S in the model
- The redundant dummy excluded from estimation represents the reference category to interpret the parameter estimates of the included dummies

Dummy variable trap – OLS estimation

```
. regress trstprt eduyrs north center south if cntry=="IT"
note: south omitted because of collinearity.

      Source |       SS           df       MS      Number of obs   =     2,516
-------------+----------------------------------   F(3, 2512)      =     12.00
       Model |  182.209659         3  60.7365531   Prob > F        =    0.0000
    Residual |  12710.0272     2,512  5.05972421   R-squared       =    0.0141
-------------+----------------------------------   Adj R-squared   =    0.0130
       Total |  12892.2369     2,515  5.12613793   Root MSE        =    2.2494

------------------------------------------------------------------------------
     trstprt | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      eduyrs |    .056178   .0105909     5.30   0.000     .0354102    .0769458
       north |   .2937215   .0996418     2.95   0.003      .098333    .4891099
      center |   .1344463   .1332165     1.01   0.313     -.126779    .3956716
       south |          0  (omitted)
       _cons |   2.228934   .1532395    14.55   0.000     1.928446    2.529423
------------------------------------------------------------------------------
```

- Although we tried to include x_S, STATA excluded it to avoid the dummy variable trap: South is the reference category
- $\hat{\beta}_N$ is the difference in political trust between the North and the reference, i.e. the South, holding education constant; $\hat{\beta}_C$ is the difference between the Center and the South

Interactions

- The relationship between x_k and y might differ depending on the level of another independent variable x_j where $k \neq j$: in such cases, **interactions** must be added to the model

- Interacting means **multiplying** two or more variables, thus generating additional parameters whose estimates require careful interpretation

- The regression model with an interaction between x_1 and x_2 is:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 \left(x_{1,i} \times x_{2,i}\right) + \varepsilon_i$$

- When including an **interaction term**, $x_1 \times x_2$, it's important to include also the **main terms** separately, x_1 and x_2

- Three forms of interaction are possible, depending on the variables involved (continuous or categorical)

Interaction #1: continuous $\times$ continuous

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 Age_i + \hat{\beta}_2 Education_i + \hat{\beta}_3 \left(Age_i \times Education_i \right)$$

```
. regress polintr agea eduyrs c.agea#c.eduyrs
```

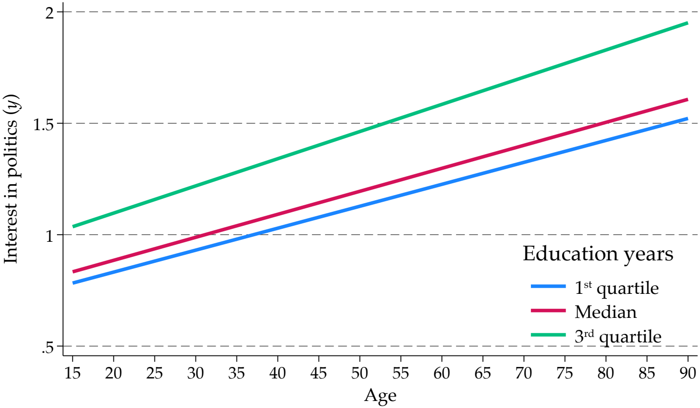
Source	SS	df	MS
Model	3423.22592	3	1141.07531
Residual	27576.7646	36,609	.753278282
Total	30999.9905	36,612	.846716665

Number of obs	=	36,613
F(3, 36609)	=	1514.81
Prob > F	=	0.0000
R-squared	=	0.1104
Adj R-squared	=	0.1104
Root MSE	=	.86792

polintr	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
agea	.0046829	.0008267	5.66	0.000	.0030625	.0063034
eduyrs	.0435676	.0036624	11.90	0.000	.0363891	.0507461
c.agea#c.eduyrs	.0004698	.0000635	7.40	0.000	.0003454	.0005942
_cons	.1558196	.0496646	3.14	0.002	.0584756	.2531635

- $\hat{\beta}_1$: 1 more year of age increases interest in politics by 0.004 when education $= 0$
- $\hat{\beta}_2$: 1 more year of education increases interest in politics by 0.043 when age $= 0$
- $\hat{\beta}_3$: 1 more year of age increases interest in politics by 0.0004 for every additional year of education (or viceversa)

Interaction #1: continuous × continuous – Graph



- Plot a line showing the relationship between y and one of the interacted continuous variable for each chosen level of the other interacted continuous variable
- Interest in politics increases with age at any level of education, but the increase is larger for people with more education years

Interaction #2: categorical $\times$ continuous

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 Female_i + \hat{\beta}_2 Education_i + \hat{\beta}_3 \left(Female_i \times Education_i\right)$$

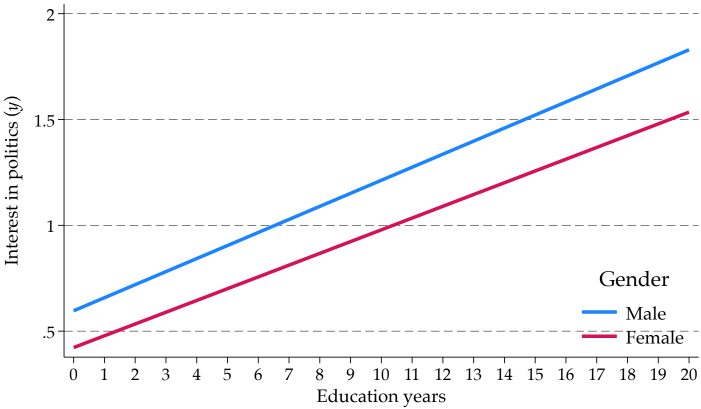
```
. regress polintr i.gndr eduyrs i.gndr#c.eduyrs

      Source |       SS           df       MS      Number of obs   =    36,864
-------------+----------------------------------   F(3, 36860)     =   1154.50
       Model |  2683.43116         3  894.477055   Prob > F        =    0.0000
    Residual |   28558.126    36,860  .774772817   R-squared       =    0.0859
-------------+----------------------------------   Adj R-squared   =    0.0858
       Total |  31241.5572    36,863  .847504467   Root MSE        =    .88021

------------------------------------------------------------------------------
     polintr | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        gndr |
      Female |    -.17373   .0309238    -5.62   0.000    -.2343415   -.1131186
      eduyrs |   .0616628   .0016784    36.74   0.000      .058373    .0649526
             |
gndr#c.eduyrs|
      Female |  -.0060462   .0022606    -2.67   0.007     -.010477   -.0016153
             |
       _cons |    .596249   .0229519    25.98   0.000     .5512627    .6412353
------------------------------------------------------------------------------
```

- $\hat{\beta}_1$: females have lower interest in politics by 0.173 when education $= 0$
- $\hat{\beta}_2$: 1 more year of education increases interest in politics by 0.061 for males
- $\hat{\beta}_3$: females have lower interest in politics by 0.006 for every additional year of education (or viceversa)

Interaction #2: categorical × continuous – Graph



- Plot a line showing the relationship between y and the interacted continuous variable for each level of the interacted categorical variable
- Interest in politics increases with education for both sexes, but the increase is smaller for females

Interaction #3: categorical × categorical

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 Female_i + \hat{\beta}_2 Non\text{-}native_i + \hat{\beta}_3 \left(Female_i \times Non\text{-}native_i\right)$$

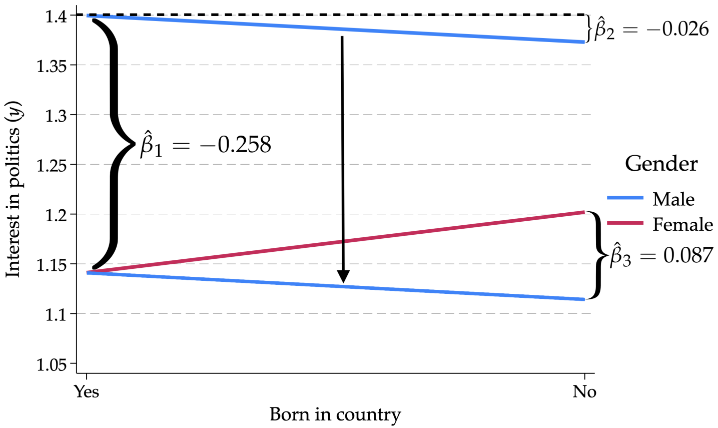
```
. regress polintr i.gndr i.brncntr i.gndr#i.brncntr

      Source |       SS           df       MS      Number of obs   =    37,486
-------------+----------------------------------   F(3, 37482)     =    238.58
       Model |  595.383619         3  198.461206   Prob > F        =    0.0000
    Residual |  31179.3323    37,482  .831848148   R-squared       =    0.0187
-------------+----------------------------------   Adj R-squared   =    0.0187
       Total |  31774.7159    37,485  .847664823   Root MSE        =    .91206

------------------------------------------------------------------------------
     polintr | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        gndr |
      Female |  -.2584995   .0098587   -26.22   0.000    -.2778228   -.2391762
             |
     brncntr |
          No |  -.0267188   .0252232    -1.06   0.289     -.076157    .0227193
             |
gndr#brncntr |
   Female#No |   .0875409   .0344135     2.54   0.011     .0200895    .1549923
             |
       _cons |   1.399612   .0072143   194.01   0.000     1.385472    1.413752
------------------------------------------------------------------------------
```

- $\hat{\beta}_1$: female natives have lower interest in politics than male natives by 0.258
- $\hat{\beta}_2$: male non-natives have lower interest in politics than male natives by 0.026
- $\hat{\beta}_3$: females have higher interest in politics than males by 0.087 among non-natives compared to natives

Interaction #3: categorical × categorical – Graph



- Plot a line showing the relationship between y and one of the interacted categorical variables for each level of the other interacted categorical variable
- Interest in politics is larger for males than females, but the gap is smaller when comparing non-natives to natives

Regressions with non-linear terms

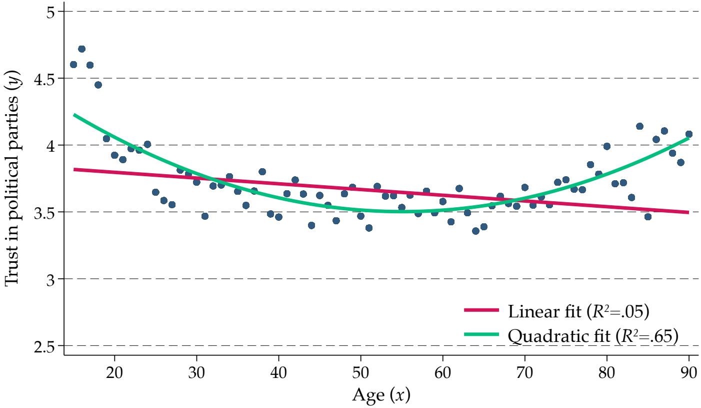
- In **linear** models the marginal effect of x is **constant** across values of x, i.e. y changes by the same amount when a one-unit change occurs at low or high levels of x
- Good for intuition and interpretation, less good for accurate modelling and prediction
- Solution: x can also be included in the model **non-linearly** by including higher order polynomials of x (e.g. quadratic, x^2, cubic, x^3, quartic, x^5, etc.)
- The marginal effect of x becomes **not constant** across values of x, i.e. y changes by a different amount when a one-unit change occurs at low or high levels of x

Regressions with non-linear terms – Example

Example. Let's model trust in political parties, y, as function of age, x

Estimating $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ produces a poor linear fit

Estimating $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_1 x_i^2$ produces a nice quadratic fit



OLS assumptions: summary

- Population model linear in parameters: $y = \beta_0 + \beta_1 x + \varepsilon$
- Random sampling: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Sample variation in the explanatory variable: $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- Mean conditional independence: $E(\varepsilon|x) = 0$
- Constant variance (homoskedasticity): $Var(\varepsilon|x) = \sigma^2$