Verifying key assumptions: normality, collinearity and functional form. Goodness-of-fit.

Jakub Mućk
SGH Warsaw School of Economics

Least squares estimator

- **Least squares estimator** :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon \tag{1}$$

where
  - y is the (outcome) dependent variable;
  - $x_1, x_2, \ldots, x_K$ is the set of independent variables;
  - ε is the error term.
- The dependent variable is explained with the components that vary with the **the dependent variable** and **the error term**.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_K$ are the coefficients (slopes) on $x_1, x_2, \ldots, x_K$.

- **Least squares estimator** :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon \quad (1)$$

  where
  - y is the (outcome) dependent variable;
  - $x_1, x_2, \ldots, x_K$ is the set of independent variables;
  - ε is the error term.
- The dependent variable is explained with the components that vary with the **the dependent variable** and **the error term**.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_K$ are the coefficients (slopes) on $x_1, x_2, \ldots, x_K$.

$\beta_1, \beta_2, \ldots, \beta_K$ measure the effect of change in $x_1, x_2, \ldots, x_K$ upon the expected value of y (*ceteris paribus*).

- **Assumption #1**: true DGP (data generating process):

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon. \tag{2}$$

- **Assumption #2**: the expected value of the error term is zero:

$$\mathbb{E}\left(\varepsilon\right) = 0, \tag{3}$$

and this implies that $\mathbb{E}\left(y\right) = \mathbf{X}\beta$.

- **Assumption #3**: Spherical variance-covariance error matrix.

$$var(\varepsilon) = \mathbb{E}(\varepsilon\varepsilon') = I\sigma^2 \tag{4}$$

. In particular:

  - the variance of the error term equals σ:

$$var\left(\varepsilon\right) = \sigma^2 = var\left(y\right). \tag{5}$$

  - the covariance between any pair of ε_i and ε_j is zero"

$$cov\left(\varepsilon_i, \varepsilon_j\right) = 0. \tag{6}$$

- **Assumption #4**: **Exogeneity.** The independent variable are **not random** and therefore they are not correlated with the error term.

$$\mathbb{E}(\mathbf{X}\varepsilon) = 0. \tag{7}$$

- **Assumption #5**: the full rank of matrix of explanatory variables (there is no so-called collinearity):

$$rank(\mathbf{X}) = K + 1 \leq N. \tag{8}$$

- **Assumption #6 (optional)**: the normally distributed error term:

$$\varepsilon \sim \mathcal{N}\left(0, \sigma^2\right). \tag{9}$$

Assumptions of the least squares estimators

Under the assumptions A#1-A#5 of the multiple linear regression model, the least squares estimator $\hat{\beta}^{OLS}$ has the smallest variance of all linear and unbiased estimators of β.

$\hat{\beta}^{OLS}$ is the Best Linear Unbiased Estimators (BLUE) of β.

- **The least squares estimator**

$$\hat{\beta}^{OLS} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}. \tag{10}$$

- **The variance of the least square estimator**

$$Var(\hat{\beta}^{OLS}) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} \tag{11}$$

- **If the (optional) assumption about normal distribution of the error term is satisfied** then

$$\beta \sim \mathcal{N}\left(\hat{\beta}^{OLS}, Var(\hat{\beta}^{OLS})\right). \tag{12}$$

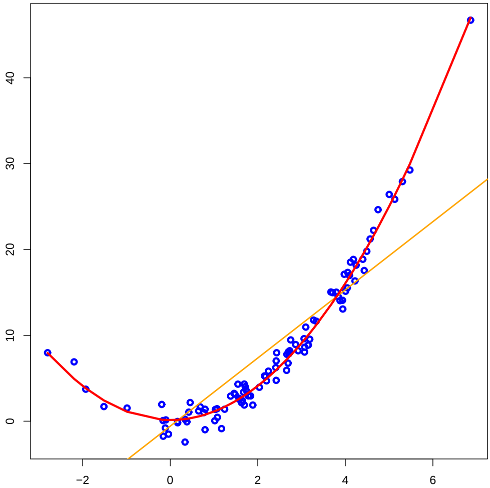
Estimating non-linear relationship

- Economic variables are not always related by straight-line relationships. They display **curvilinear forms**.
- [Example] Wages (w) and experience ($exper$):

$$w = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \varepsilon. \tag{13}$$

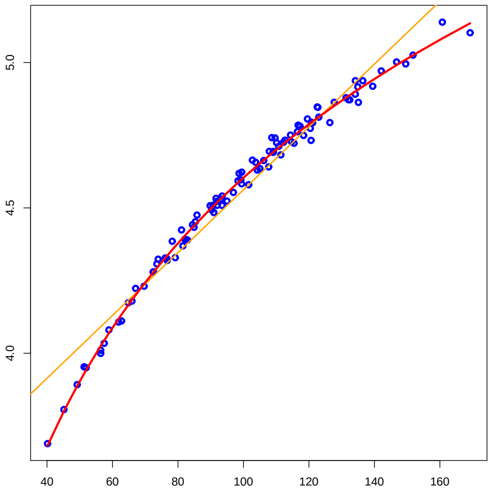
In the above model, the quadratic relationship is assumed. Why?

- In general, the choice of function form is related to:
  1. economic theory,
  2. empirical pattern,
  3. properties of residuals.
- The most popular nonlinear functions:
  - quadratic and cubic relationship,
  - polynomial equations,
  - logs of the dependent and/or independent variable.



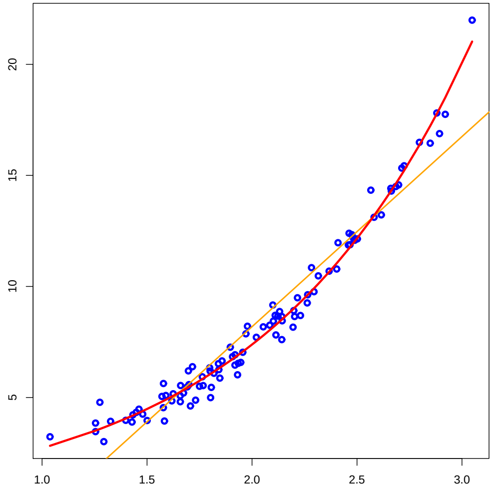
Orange line :
$y = \beta_1 + \beta_2 x$

Red line :
$y = \beta_1 + \beta_2 x^2$



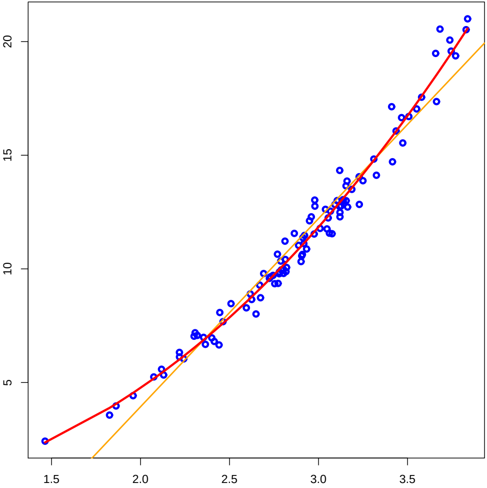
Orange line :
$y = \beta_1 + \beta_2 x$

Red line :
$y = \beta_1 + \beta_2 \ln x$



Orange line :
$y = \beta_1 + \beta_2 x$

Red line :
$\ln y = \beta_1 + \beta_2 x$



Orange line :
$y = \beta_1 + \beta_2 x$

Red line :
$\ln y = \beta_1 + \beta_2 \ln x$

- **Marginal effects** measures expected instantaneous change in the dependent variable (y)in a reaction to change in explanatory variable (x):

$$\text{Marginal effect} = \frac{\partial \mathbb{E}\left(y\right)}{\partial x} \tag{14}$$

In other words, the marginal effects is the slope of the tangent to the curve at a particular point.

- **Elasticity** measures the percentage change in y in a reaction to percentage change in x:

$$\text{Elasticity} = \frac{\partial \mathbb{E}\left(y\right)}{\partial x}\frac{x}{y}. \tag{15}$$

- **Semi-elasticity** measures the percentage change in y in a reaction to a change in x

$$\text{Semi-Elasticity} = \frac{\partial \mathbb{E}\left(y\right)}{\partial x}\frac{1}{y}. \tag{16}$$

Name	Function	Slope (marginal effects)	Elasticity
Linear	$y = \beta_0 + \beta_1 x$	β_1	$\beta_1 \frac{x}{y}$
Quadratic	$y = \beta_0 + \beta_1 x^2$	$2\beta_1 x$	$2\beta_1 x \frac{x}{y}$
Quadratic (II)	$y = \beta_0 + \beta_1 x + \beta_2 x^2$	$\beta_1 + 2\beta_2 x$	$(\beta_1 + 2\beta_2 x) \frac{x}{y}$
Cubic	$y = \beta_0 + \beta_1 x^3$	$3\beta_1 x^2$	$3\beta_1 x^2 \frac{x}{y}$
Log-Log	$\ln(y) = \beta_0 + \beta_1 \ln(x)$	$\beta_1 \frac{y}{x}$	β_1
Log-Linear	$\ln(y) = \beta_0 + \beta_1 x$	$\beta_1 y$	$\beta_1 x$
a 1 unit change in x leads to (approximately) a 100 β_1% change in y			
Linear-Log	$y = \beta_0 + \beta_1 \ln(x)$	$\beta_1 \frac{1}{x}$	$\beta_1 \frac{1}{y}$
a 1 % change in x leads to (approximately) a $\beta_1/100$ unit change in y			

- **Interaction variable** is the product of (at least) two variable involved in regression and accounts for simultaneous effects of two variables.
- [Example] Wages (w), experience ($exper$) and education ($educ$):

$$w = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 educ + \beta_4 exper \times educ + \varepsilon. \qquad (17)$$

- In this case:

$$\begin{aligned}\text{Marginal effect of education} &= \frac{\partial \mathbb{E}(w)}{\partial educ} = \beta_3 + \beta_4 exper, \\ \text{Marginal effect of experience} &= \frac{\partial \mathbb{E}(w)}{\partial exper} = \beta_2 + 2\beta_3 exper + \beta_4 educ.\end{aligned}$$

Model Specification

A model could be misspecified when

- important explanatory variables are omitted,
- irrelevant explanatory variables are included,
- a wrong functional form is chosen,
- the assumptions of the multiple regression model are not satisfied

Omitted variables I

- Omission of a relevant variable (defined as one whose coefficient is nonzero) might lead to an estimator that is biased. This bias is known as **omitted-variable bias**.
- Let's assume true DGP (data generating process):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon. \tag{18}$$

- Consider the case when we do not have data on x_2.
  Equivalently, we impose the restriction that $\beta_2 = 0$. According to our true DGP this restriction is invalid.
- Then the expected value of the least squares estimator of β_1:

$$\mathbb{E}(\hat{\beta}_1^{LS}) = \beta_1 + \beta_2 \frac{cov(x_1, x_2)}{var(x_2)}, \tag{19}$$

  and the omitted variable bias:

$$bias\left(\hat{\beta}_1^{LS}\right) = \mathbb{E}(\hat{\beta}_1^{LS}) - \beta_1 = \beta_2 \frac{cov(x_1, x_2)}{var(x_2)}. \tag{20}$$

- The omitted bias is larger if:
  - the true slope on omitted variable β_2 is higher,
  - the omitted variable (x_2) is more correlated with the included variable (x_3).
- However, there is no bias when the omitted variable is not correlated with the explanatory variables.

Irrelevant variables I

- Due to omitted-variable bias one might follow strategy to include as many variable as possible.
- However, doing so may also inflate the variance of estimate.
- The inclusion of **irrelevant variables** may reduce the precision of the estimated coefficients for other variables in the equation

- **RESET (REgression Specification Error Test)** is designed to detect omitted variables and incorrect functional form.
- Consider the multiple linear regression:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon. \tag{21}$$

- [Step #1]. Obtain the least square estimates and calculate the fitted values:

$$\hat{y} = \hat{\beta}_0^{LS} + \hat{\beta}_1^{LS} x_1 + \ldots + \hat{\beta}_k^{LS} x_k \tag{22}$$

- [Step #2]. Consider the following auxiliary regressions:

$$\begin{aligned} \text{Model 1}: \quad & y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \gamma_1 \hat{y}^2 + \varepsilon. \\ \text{Model 2}: \quad & y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3 + \varepsilon. \end{aligned}$$

  Obtain the least squares estimators of γ_1 in Model 1 and/or γ_1 and γ_2 in Model 2.
- [Step #3]. Consider the following null:

$$\begin{aligned} \text{Model 1}: \quad & \mathcal{H}_0: \quad \gamma_1 = 0, \\ \text{Model 2}: \quad & \mathcal{H}_0: \quad \gamma_1 = \gamma_2 = 0, \end{aligned}$$

  In both cases the null hypothesis is about **misspecification**.

- The RESET test is very general test allowing for testing functional form. However, if we reject the null we do not know what is the source of misspecification.
- If a number of observations is large one might replace squared and cubic fitted values of outcome variable by squared and cubic of explanatory variables.

Collinearity

- When data are the result of an uncontrolled experiment, many of the economic variables may move together in systematic ways.
- This problem is labeled **collinearity** and explanatory variable are said to be **collinear**.
- Example: multiple regression with two explanatory variable

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon. \tag{23}$$

The variance of the least squares estimator for β_2:

$$var\left(\hat{\beta}_2^{LS}\right) = \frac{\sigma^2}{\left(1 - r_{12}^2\right)\sum_{i=1}^{N}\left(x_{i2} - \bar{x}_2\right)}, \tag{24}$$

where r_{12} is the correlation between x_1 and x_2.
- Extreme case: $r_{23} = 1$ then the x_1 and x_2 are perfectly collinear. In this case the least squares estimator is not defined and we cannot obtain the least squares estimates.
- If r_{12}^2 is large then:
  - the standards errors are large $\implies$ small (in modulus) t statistics. Typically, it leads to the conclusion that parameter estimates are not significantly different from zero,
  - estimates may be very sensitive to the inclusion or exclusion of a few observations,
  - estimates may be very sensitive to the exclusion of insignificant variables.

- **Detecting collinearity:**
  - **pairwise correlation between explanatory variables**,
  - **variance inflation factor (VIF)** which is calculated for each explanatory variable. The VIF is a function of R^2 from auxiliary regression of the selected explanatory variable on the remaining explanatory variables:

$$VIF_i = \frac{1}{1 - R_i^2}. \tag{25}$$

    The values above 10 suggests collinearity.
- **Dealing with collinearity:**
  - Obtaining more infromation.
  - Using non-sample information, i.e., restrictions on parameters.

Normality of the error term

SGH

- The assumption of the error term is crucial to test the hypothesis. However, the error term is random variable and , therefore, is not unobservable.
- The normality of the error term can be justified on the basis of the residuals properties.
- The assessment of this assumption bases on:
  - the residuals histogram,
  - results of the Jarque-Berra test.
- But if the sample is *sufficiently* large then, according to a central limit theorem, the distribution of least squares estimator can be approximated by normal distribution.

- In general, **the Jarque-Berra test** allows to investigate whether sample data have the skewness and kurtosis that match to normal distribution.
- The skewness ($\mathcal{S}$) and kurtosis ($\mathcal{K}$) of residuals ($\hat{e}_i$)

$$\mathcal{S} = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{e}_i - \hat{\bar{e}}\right)^3}{\left(\frac{1}{N}\sum_{i=1}^{N}\left(\hat{e}_i - \hat{\bar{e}}\right)^2\right)^{\frac{3}{2}}} \quad \text{and} \quad \mathcal{K} = \frac{\frac{1}{N}\sum_{i=1}^{N}(e_i - \bar{e})^4}{\left(\frac{1}{N}\sum_{i=1}^{N}\left(\hat{e}_i - \hat{\bar{e}}\right)^2\right)^2} - 3$$

- The test statistics:

$$\mathcal{JB} = \frac{N}{6}\left(\mathcal{S}^2 + \frac{1}{4}(\mathcal{K} - 3)^2\right) \sim \chi^2_{(2)}. \tag{26}$$

Goodness-of -fit

- The observed values (y_i) of dependent variable can be decomposed into the fitted values ($\hat{y}_i$) and the residuals ($\hat{e}_i$):

$$y_i = \hat{y}_i + \hat{e}_i, \tag{27}$$

- subtracting the sample mean ($\bar{y}$) from both sides:

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + \hat{e}_i. \tag{28}$$

- Squaring and summing both sides of above equation:

$$\sum_{i=1}^{N} (y_i - \bar{y})^2 = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{N} \hat{e}_i^2, \tag{29}$$

  In the above expression we use assumption that $\sum_{i=1}^{N} (\hat{y}_i - \bar{y}) \hat{e}_i = 0$ since the $x_1, \ldots, x_K$ are not random.

- The decomposition of total variation in dependent variable:

$$SST = SSR + SSE, \tag{30}$$

  where
  - SST is the sum of squares and $SST = \sum_{i=1}^{N} (y_i - \bar{y})^2$,

Goodness-of -fit II

SGH

- SSR is the sum of squares due to regression and $SSR = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$,
- SSE is the sum of squares due to regression and $SSE = \sum_{i=1}^{N} \hat{e}_i^2$.

Coefficient of determination R^2 is the proportion of variation that can be explained by independent variables:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}, \tag{31}$$

$R^2 \in < 0, 1 >$.

- The correlation coefficient ρ_{xy} between x and y is defined by:

$$\rho_{xy} = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \tag{32}$$

and the sample correlation coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y}, \tag{33}$$

takes the values between -1 and 1.

- **In simple linear regression**: the relationship between R^2 and r_{xy} is as follows:

$$R^2 = r_{xy}^2, \tag{34}$$

and, therefore, the R^2 can also be computed as the square of the sample correlation coefficient between y_i and $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Adjusted R^2

- The coefficient of determination R^2 is always higher if we include additional explanatory variable even if the added variable is not justified/ relevant.
- **The adjusted coefficient of determination $\bar{R}^2$:**

$$\bar{R}^2 = 1 - \frac{SSE}{SST}\frac{(N-1)}{(N-K)}, \quad (35)$$

where SSE is the sum of squared errors and SST is the sum of squares.

- With the adjusted coefficient of determination we account for a decrease in degree of freedoms: $(N-1)/(N-K)$.
- However, it has no convenient interpretation.

- Information criteria are alternative measures of goodness-of-fit. They have no interpretation but, like adjusted R^2, account for a decrease in degrees of freedom.
- **The Akaike information criterion (AIC)**:

$$AIC = \ln\left(\frac{SSE}{N}\right) + \frac{2K}{N}. \tag{36}$$

- **The Bayesian information criterion (SIC)**:

$$SIC = \ln\left(\frac{SSE}{N}\right) + \frac{K\ln(K)}{N}. \tag{37}$$

- Using the above criteria, the lower values of AIC/BIC signals better fit to data.