INTRODUCTION TO DATA SCIENCE

Introduction to econometrics

Maciej Świtała, PhD

Autumn 2025

UNIVERSITY OF WARSAW | FACULTY OF ECONOMIC SCIENCES

1. Introduction to linear regression model
   - Conceptual framework
   - Ordinary Least Squares (OLS) estimation method
   - Assumptions of the Classical Linear Regression Model (CLRM)
   - Interpretation of parameters in the regression model
   - Testing the statistical significance of the parameters

2. Extensions of the linear regression model
   - Qualitative explanatory variables
   - Modelling interactions and nonlinear relationships
   - Omitted vs. irrelevant variables
   - Model (variables) selection

Introduction to linear regression model

Linear regression

- Linear regression is the simplest econometric model that allows for **explaining** the variability of a selected continuous variable.
- Additionally, linear regression enables the **prediction** of a selected continuous variable, making it a **benchmark** for more advanced predictive algorithms.
- Many far more complex methods can be viewed as **generalisations** or **extensions** of linear regression model.
- Understanding linear regression is the **first step** to learning more advanced methods.

What models are

- Models are a **simplistic** way (i.e., using a small number of estimated parameters) to **explain** the most important relationships between observed variables.
- Econometric models, such as linear regression, are based on a **theoretical model**, meaning that they rely on **assumptions** about the underlying relationships between variables and the structure of the data.
- In contrast, machine-learning algorithms do not involve a theoretical model, rather focus on maximising **prediction accuracy**.

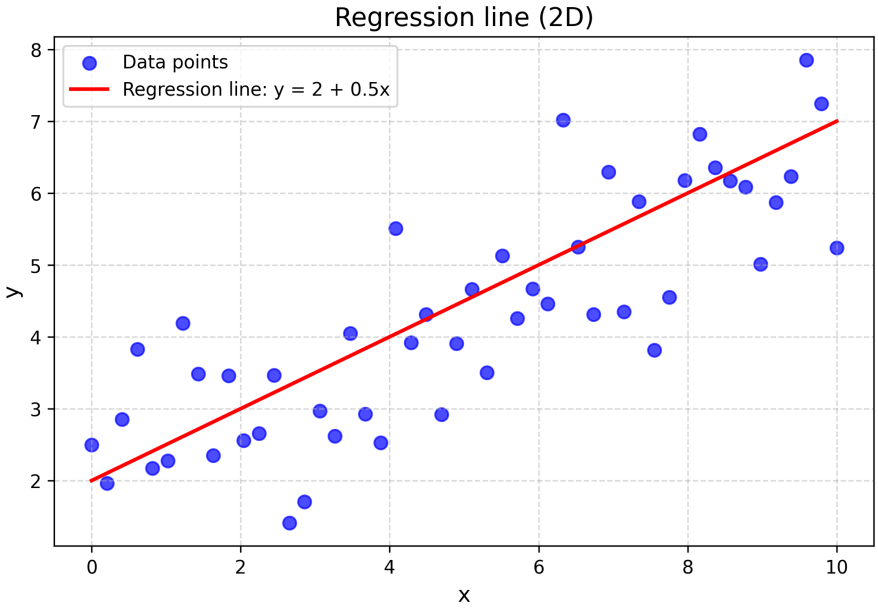
Linear relationship

- When studying the effect of one variable on another, from a mathematical point of view we seek a **functional relationship** between them.
- The simplest form is a linear function:

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

  which has two parameters: β_0 (constant term) and β_1 (slope).
- The **constant term** (intercept) is the point where the line crosses the vertical axis (usually without interpretation).
- The **slope** (regression coefficient) is usually the focus of our analysis. It shows how y changes when x changes.

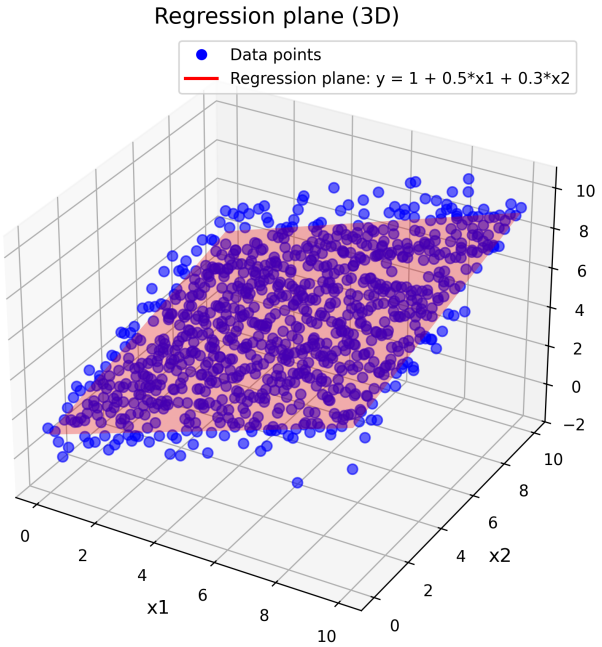
Linear relationship - cont'd



Regression line and hyperplane

- The line of dependence is called the **regression line**.
- We have a line only in a model with a constant and one explanatory variable.
- In the case of two explanatory variables, the result is a **regression plane**.
- With multiple variables, we talk about a **regression hyperplane**.

Regression line and hyperplane - cont'd



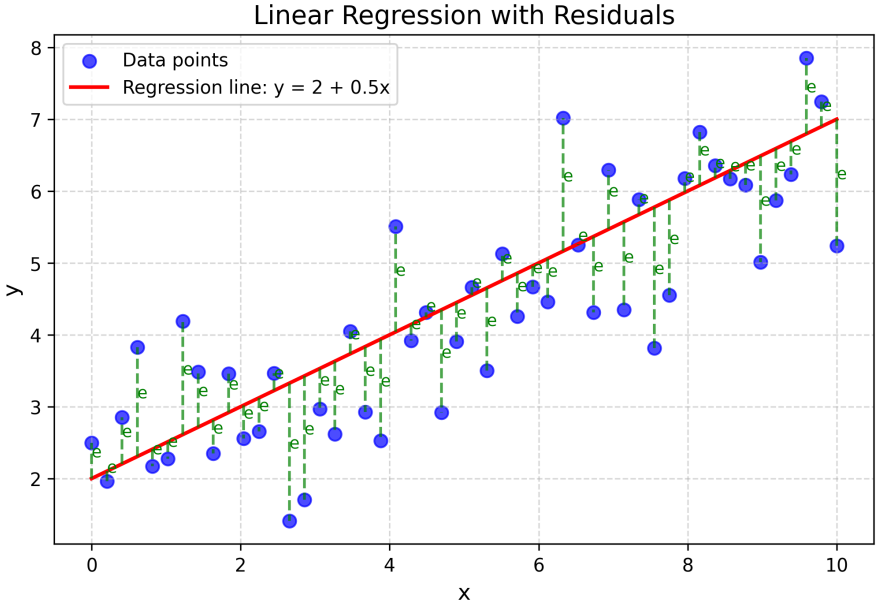
Random disturbance

- In practice, and also as demonstrated before, no line / plane / hyperplane fits the data perfectly!
- Therefore, the linear regression function has the following form:

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

- The last term (ϵ_i) is called the **random error term** or **random disturbance**.
- Its realisation is represented by **residuals** e_i, i.e., the vertical distances between a data point and the regression line.

Random disturbance - cont'd



Where does the random error come from?

- The **random error** ϵ_i describes that part of the variability (variance) of y_i that **cannot be explained** by the explanatory variables used in the regression.
- It cannot be explained because:
  1. the values of other important but **omitted explanatory variables** are unknown,
  2. some important factors may have a **random nature**.

Simple and multiple regression

In simple linear regression:

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

- y_i is called the **dependent variable**, endogenous variable, or regressand,
- x_i is the **independent variable**, explanatory variable, exogenous variable, or regressor.

When multiple explanatory variables are included, the regression function takes the form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

We then speak of the regression of y on $x_1, x_2, \ldots, x_k$.

Estimation of the linear regression model

- Before making predictions, the model parameters must be **estimated** based on a sample.
- The true parameter values $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are unknown.
- They are **estimated** from the observed responses of the dependent variable to changes in explanatory variables.
- Parameter estimates will naturally be **imprecise** (because they are based on a sample, not the entire population).
- Estimates from two different samples will usually differ.
- In linear regression, the estimates are, however, **unbiased** - if we estimated the model on many samples, the average of estimates would equal the true parameters.

Ordinary Least Squares (OLS) estimation method

- The **estimates** (or **estimators**) of parameters based on the sample are usually denoted $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ (alternatively $b_0, b_1, b_2, \ldots, b_p$).
- The estimator usually differs from the true value: $\hat{\beta} \neq \beta$, being its approximation based on the analyzed data sample.
- The predicted value of Y resulting from the estimated model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_p x_{pi}$$

or, in matrix form:

$$\hat{Y} = X\hat{\beta}$$

is called the **fitted value** or **theoretical value**.

- The theoretical values result from the model and the explanatory variables in X.

Residuals

- The difference between the empirical (observed) and theoretical value is called the **residual** of the model:

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_p x_{pi}) = y_i - \hat{y}_i$$

  or, in matrix form:

$$e = Y - X\hat{\beta} = Y - \hat{Y}$$

- Residuals represent the **realisation (estimate) of the random error term**.

Ordinary Least Squares (OLS) estimation method - cont'd

- A good model is one for which residuals are small (predicted values of the modeled variable are **close** to the actual ones).
- There are many definitions of **closeness**, but the most common is minimizing the sum of squared residuals.
- Residuals can be positive or negative - so we look at their squares (so they don't cancel each other).
- The smaller the sum of squared residuals, the better the model fits the data.
- We therefore look for parameter estimates $\hat{\beta}$ that **minimize the sum of squared deviations** of empirical from theoretical values:

$$\min_{\hat{\beta}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min_{\hat{\beta}} \sum_{i=1}^{n} e_i^2$$

Form of the OLS estimator

- The quantity minimized in OLS is called the **residual sum of squares**, denoted **RSS**.
- In a model with one explanatory variable, the parameters minimizing RSS can be calculated as:

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

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

where $\bar{x}$ and $\bar{y}$ denote the sample means of variables X and Y, respectively.

General form of the OLS estimator

The general form of the OLS estimator can be written as:

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

where X' denotes the transpose of the matrix, and the exponent -1 denotes the matrix inverse.

It can be estimated if:

- we have at least as many observations as variables,
- no variable is a linear combination of the others (no multicollinearity).

Derivation of the general OLS estimator

1. The OLS estimator minimizes the square of the random error:

$$\underset{\beta}{\operatorname{argmin}}\ \boldsymbol{e}'\boldsymbol{e}$$

2. Knowing that $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}$, we can write the minimized expression as:

$$SSE(\boldsymbol{b}) = \boldsymbol{e}'\boldsymbol{e} = (\boldsymbol{y}-\hat{\boldsymbol{y}})'(\boldsymbol{y}-\hat{\boldsymbol{y}}) = (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}) = \boldsymbol{y}'\boldsymbol{y}-2\boldsymbol{y}'\boldsymbol{X}\boldsymbol{b}+\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}$$

3. Solving this optimization problem yields:

$$\frac{\partial SSE(\boldsymbol{b})}{\partial \boldsymbol{b}} = 0 \Leftrightarrow -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = 0$$

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

Statistical properties of the OLS estimator

Gauss-Markov Theorem

Under certain assumptions, the OLS estimator is the best estimator of the parameter vector among linear and unbiased estimators of that parameter (also known as ***BLUE*** - *Best Linear Unbiased Estimator*).

Properties of the OLS Estimator in the Classical Linear Regression Model

1. **Unbiasedness**: an unbiased estimator has an expected value equal to the estimated parameter: $E(b) = \beta$. Intuitively, the estimate is on average equal to the parameter, so no systematic errors occur.
2. **Efficiency**: an efficient estimator has minimal variance (most precise).
3. **Consistency**: intuitively, as the sample size increases, the estimation error approaches 0, and the estimate converges to the true parameter: $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}, \quad n \longrightarrow \infty$.

Assumptions of the Classical Linear Regression Model (CLRM)

1. There is a **linear relationship** between the dependent variable Y and the explanatory variables X (in practice, this is often a good approximation).
2. The explanatory variables X are **non-random** and do not affect the random error terms.
3. The expected value of the random error is **zero**.
4. The individual error terms are **uncorrelated** (**no autocorrelation** of errors).
5. The variance of the random error is **constant** for all observations (homoskedasticity).

CLRM assumptions (mathematical form)

1. $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$
2. X is deterministic
3. $E(\varepsilon_i) = 0$ for $i = 1, \ldots, N$
4. $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$
5. $Var(\varepsilon_i) = \sigma^2$ for $i = 1, \ldots, N$ or $Var(\varepsilon) = \sigma^2 \boldsymbol{I}$

Additional assumption - normality of residuals

- For statistical hypothesis testing in CLRM, an additional assumption of normality of residuals is often made:

$$\varepsilon_i \sim NIID(0, \sigma^2)$$

- **Only then** do the statistics used for **hypothesis testing** in the model have the correct theoretical distributions (especially important in small samples).
- If assumptions are violated, regression results should be treated cautiously, more for descriptive than inferential purposes.

Interpretation of parameters in the regression model

- In a model with multiple explanatory variables, the estimated parameter indicates how much the **expected value** of Y changes due to a one-unit change in the explanatory variable X associated with that parameter, holding other variables constant (ceteris paribus):

$$\beta_k = \frac{\Delta E(Y)}{\Delta X_k}$$

- This is called the **partial effect** or **marginal effect**.

(Partial) elasticities

- Often, elasticities derived from parameters are more interesting than the raw estimates.
- **Partial elasticity** is the percentage change in the expected value of Y in response to a 1% change in X_K:

$$e_{Y,X_k} = \frac{\Delta E(y)/E(Y)}{\Delta X_K/X_K}$$

- Its advantage is **dimensionlessness** (independent of units), making comparisons between models easier.

Elasticity in the linear model

- In a linear model, elasticity **depends on the value** of the explanatory variable and the expected value of Y at which it is calculated:

$$e_{Y,X_K} = \frac{\Delta E(y)/E(Y)}{\Delta X_K/X_K} = \frac{\Delta E(Y)}{\Delta X_K}\frac{X_K}{E(Y)} = \beta_K \frac{X_K}{E(Y)}$$

- If needed, it is usually calculated **at the mean values of the explanatory variables**.
- In models where both dependent and explanatory variables are logarithmically transformed, the parameters express **elasticities** rather than partial effects.

Testing the statistical significance of parameters

- After estimating regression coefficients, we can test their **statistical significance**. Statistical significance indicates that the **true value of the parameter is different from zero**.
- We test the **null hypothesis** assuming that the true model coefficient β_i **equals zero**.
- If we **reject** the null hypothesis in favor of the **alternative hypothesis** that the coefficient is **different from zero**, we say that the variable is **significant** in the model.
- Formally: $H_0 : \beta_i = 0$ vs. $H_a : \beta_i \neq 0$.

Testing the statistical significance of parameters - cont'd

- Since we only have a sample of observations, we cannot be certain – we check significance at a chosen **confidence level** (e.g., 95%, 99%).
- While there is no strict rule, at least 20 observations per explanatory variable are recommended.
- In general: more observations increase the likelihood that results are statistically significant.

Testing the statistical significance of parameters - cont'd

- The **test statistic** is:

$$t = \frac{\hat{\beta}_i - 0}{SE(\beta_i)}$$

  where $SE(\beta_i)$ is the standard error of the estimator β_i (reflecting estimation uncertainty from using a sample rather than the entire population).
- The test statistic t follows a Student's t-distribution with $n - 2$ degrees of freedom.
- If the estimation error is **large** and the estimated coefficient **small**, we cannot reject the null hypothesis.
- In this case, a nonzero estimated coefficient may simply result from sampling variability.

Confidence intervals for parameters

- The standard error of a regression parameter is proportional to the measure called the **residual standard error**:

$$RSE = \sqrt{\frac{RSS}{n-p-1}}$$

- Based on the standard error, we can construct **confidence intervals** for the parameters.
- They indicate the range within which the true parameter value lies with high probability $(1-\alpha)$.
- α is called the **significance level**, typically small (5%, 1%, rarely 10%).
- A **larger standard error** results in a wider confidence interval, reflecting **greater uncertainty** about the true parameter value.
- Accuracy of parameter estimates improves with sample size (n).

Testing the significance of the entire model

- In addition to testing individual parameters, we can test the **overall significance** of the model.
- The null hypothesis assumes that all **estimated parameters except the intercept** are zero simultaneously:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

against the alternative hypothesis:

$$H_a : \text{at least one } \beta_i \neq 0.$$

- If only the intercept is statistically significant, the dependent variable does not depend on the independent variables at all.
- In this case, the estimated model cannot explain the variability of the studied phenomenon.

F-test for overall model significance

- The **F-test** is used for this purpose, with test statistic:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

  where TSS is the total sum of squares (Total Sum of Squares), i.e., the sum of squared deviations of Y from its mean: $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$.
- Sometimes, the **overall model is significant** even if all individual variables are **insignificant**.

Testing the insignificance of selected variables

- Similarly, one can test the **joint insignificance** of several (e.g., q) selected variables:
$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \ldots = \beta_p = 0$$
- Estimate a second model excluding these q variables.
- If the residual sum of squares for this model is RSS_0, the appropriate test statistic is:
$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

Extensions of the linear regression model

Qualitative explanatory variables

- Often, among explanatory variables, there are **qualitative** (discrete) variables.
- These are variables that take a finite number of values, usually descriptive in nature.
- A special case is a **binary** variable (dummy variable), which takes only two possible values.

Qualitative explanatory variables - cont'd

- If a variable is binary, its values must be **encoded as** 0 and 1 for modeling purposes.
- For clarity, denote a binary variable in the model as D:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p D_i + \epsilon_i$$

- For example, D could represent gender: $D = 0$ for men, $D = 1$ for women.

Qualitative explanatory variables - cont'd

- We can effectively write two separate models (with slight re-grouping of β_0):
  - for $D_i = 1$: $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_0 + \beta_p + \epsilon_i$
  - for $D_i = 0$: $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_0 + \epsilon_i$
- β_p shows **how much the intercept differs** for women ($D = 1$) compared to men ($D = 0$).
- In other words, it measures the **difference in the average value of** Y for women compared to men with **similar values of other characteristics**.
- $D = 0$ is called the **baseline level** – the choice of which level to encode as 0 is arbitrary.

Qualitative explanatory variables - cont'd

- If a qualitative variable has **more than two levels**, a single dummy variable cannot represent all levels.
- For example, consider education level encoded as: 1 =none, 2 =primary, 3 =secondary, 4 =higher.
- For **each level**, we create a separate dummy variable that is 1 if the original variable is at that level, and 0 otherwise.
- For the example above, we would have:
  - $edu_1 = 1$ if education level $= 1$, 0 otherwise
  - $edu_2 = 1$ if education level $= 2$, 0 otherwise
  - $edu_3 = 1$ if education level $= 3$, 0 otherwise
  - $edu_4 = 1$ if education level $= 4$, 0 otherwise

Qualitative explanatory variables - cont'd

- We cannot include all new variables in the model, because they are **linearly dependent**.
- Their **sum is always 1**, since for each observation the variable must take one of the possible levels.
- This leads to **multicollinearity** in the model.
- Therefore, we choose one level as the **baseline**, and the dummy for that level is **not included in the model**.

Qualitative explanatory variables - cont'd

- Consider a model with no-education as the baseline (variable edu_1 excluded):

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p edu_2 + \beta_{p+1} edu_3 + \beta_{p+2} edu_4 + \epsilon_i$$

- Coefficients of the discrete variable indicate the difference in the mean y_i **relative to the baseline level**.
- For example, β_{p+1} shows how much the expected value of Y differs for someone with **secondary education compared to someone with no education**, keeping other characteristics constant.

Interactions

- Standard linear models assume that the effect of variables on the dependent variable is **additive**.
- We can also include **cross-effects** (interaction effects), when the effect of one variable depends on the level of another.
- This effect occurs when the influence of one independent variable changes depending on the value of another independent variable.
- This is called an **interaction**.

Interactions - cont'd

- Interactions are included by adding variables that are **products** of existing explanatory variables.
- Technically, one can add interactions of variables not in the model, but the model will then not be **hierarchically well formulated**.
- Adding many interactions can cause **multicollinearity**.
- This problem can be mitigated by **centering variables** before forming interactions.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i$$

Modeling nonlinear relationships

- Nonlinearity arises when we try to fit a straight line to data that does not follow a "linear shape".
- Often solved by **logarithmic transformation** of regressors and/or the dependent variable.
- Another approach is to add **higher powers** of explanatory variables (polynomial model).
- Quadratic polynomial model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

- Cubic polynomial model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i$$

Omitted vs. irrelevant variables

- Consider two models:
  - (A) $y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i$
  - (B) $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$
- Omitted variables: true model is (B), we estimate (A).
- Irrelevant variables: true model is (A), we estimate (B).

Omitted vs. irrelevant variables - cont'd

- Omitting variables has worse consequences than including irrelevant variables.
- With irrelevant variables, estimator b is still **unbiased**, but **inefficient** (higher variance).
- With omitted variables, estimator b is **usually biased**, leading to systematic errors.

Irrelevant variables

- The estimator is **inefficient**, i.e., has higher variance.
- Higher variance leads to larger standard errors for parameter estimates.
- Larger standard errors result in smaller t statistics for testing variable significance.
- This may lead to incorrect conclusions about variable significance $\Longrightarrow$ the null hypothesis of irrelevance may be rejected too often.

Consequences of omitting relevant variables

Omitting relevant variables may lead to incorrect conclusions:

- quantitative errors: incorrect coefficient estimates,
- qualitative errors: a variable may be incorrectly deemed significant.

Choosing potential models (variables)

Potential models can be chosen based on:

- goodness of fit (R^2, adjusted R^2),
- information criteria (AIC, BIC, HQC),
- variable significance (are all variables significant?),
- diagnostic tests (does the model satisfy assumptions?),
- predictive performance (RMSE, MAPE),
- other criteria.

Decomposition of Total Sum of Squares

- Total squared deviations from the mean measure the variability of a variable:

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

  i.e., total variability (TSS) = explained variability (ESS) + unexplained variability (RSS)
- Only possible in a model with an intercept!

Measures of model fit

- RSS provides an **absolute** measure of model misfit.
- Since it is measured in Y units, it is hard to judge a "good" value.
- R^2 is a better relative measure:

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

- TSS measures **total variability** of Y around the mean; RSS measures the part **unexplained** by the model.
- Thus $TSS - RSS$ is the variability **explained** by the linear regression.

Measures of model fit - cont'd

- R^2 measures the proportion of Y variability explained by X.
- It always ranges between 0 and 1.
- It is scale-independent.
- Higher R^2 generally indicates a better model.
- A value close to 0 may indicate that:
  - the linear model is inappropriate, or
  - the estimation error is large (e.g., due to omitted relevant variables).
- There is no strict rule for how high R^2 should be—it depends on the application.

Measures of model fit - cont'd

- **Removing a variable** from the model usually **reduces** R^2—rarely it remains unchanged.
- **Adding a regressor** (even an irrelevant one) usually **increases** R^2—rarely it remains unchanged.
- This does not mean variables should be added indiscriminately.
- For models with many variables, use **adjusted** R^2:

$$adjR^2 = 1 - \frac{n-1}{n-p}(1-R^2)$$

- Adjusted R^2 can be interpreted as a measure of fit for statistically significant variables.

Regression notes

- Treat R^2 as a descriptive statistic; do not rely on it solely for model comparison.
- There is no clear threshold for "satisfactory" R^2—depends on the type of study.
- Removing a regressor usually worsens R^2 (rarely no effect).
- Adding a regressor (even irrelevant) usually improves R^2 (rarely no effect). This does not mean adding indiscriminately.
- Adjusted R^2 measures fit for statistically significant variables.

Model with or without intercept?

- Models without an intercept should be estimated only with clear theoretical justification.
- R^2 does not have the same interpretation as in models with an intercept.

Information criteria

- Lower information criterion values indicate a better model (can be negative).
- Comparable only for models estimated on the same sample.
- Useful also for non-OLS estimation (e.g., WLS), where R^2 cannot be computed.
- In all criteria, the first term describes fit quality, and the second penalizes model complexity.

Information criteria - cont'd

- Akaike Information Criterion (AIC):

$$\mathrm{AIC} = \ln\left(\frac{\boldsymbol{e}'\boldsymbol{e}}{2}\right) + \frac{2K}{n}$$

- Bayesian Information Criterion (BIC, SBC, SIC, SC):

$$\mathrm{BIC} = \ln\left(\frac{\boldsymbol{e}'\boldsymbol{e}}{2}\right) + \frac{K \ln n}{n}$$

- Hannan-Quinn Criterion (HQC):

$$\mathrm{HQC} = n \ \ln\left(\frac{\boldsymbol{e}'\boldsymbol{e}}{n}\right) + 2k \ln(\ln n)$$

Prediction error measures

- Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

- Measures average absolute deviation, ignoring direction.
- All observations are equally weighted.

Prediction error measures - cont'd

- Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- Measures magnitude of error, ignoring direction.
- Observations with large errors receive higher weight.

Thank you for your attention!

Maciej Świtała, PhD
ms.switala@uw.edu.pl