Quick revision

Correlation

What is correlation?

- Correlation describes the extent to which two features of the world tend to occur together.
- If higher values of one feature are usually seen with higher values of the other, they are positively correlated.
- If the two features show no systematic pattern together, they are uncorrelated.
- If higher values of one feature are usually seen with lower values of the other, they are negatively correlated.

Measuring correlation

- **Covariance**: the average product of deviations of two quantitative variables from the mean:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

- We only interpret the sign, not the magnitude of the association, given that covariance is scale-dependent.

Measuring correlation

- Pearson's r (or *correlation coefficient*): it standardizes average of the product of deviations of two variables from the mean (=standardized covariance)
- We standardize the covariance by dividing by the product of standard deviations of the two variables:

$$r_{xy} = \frac{\text{cov(X,Y)}}{S_x S_y}$$

- Basically, we remove the units and scales from the covariance and get a comparable measure of association.
- It ranges from -1 to 1, with $r = 0$ meaning no correlation.

Inference: Correlation (r vs. ρ)

Inference for ρ

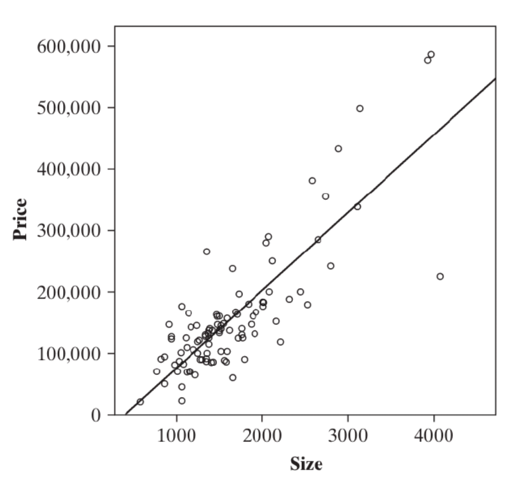
How can we test the statistical significance of correlation?

- Null and alternative hypotheses:
    - H_0: X and Y are not correlated $\Rightarrow \rho_{xy} = 0$
    - H_a: X and Y are correlated $\Rightarrow \rho_{xy} \neq 0$
- Test statistic:

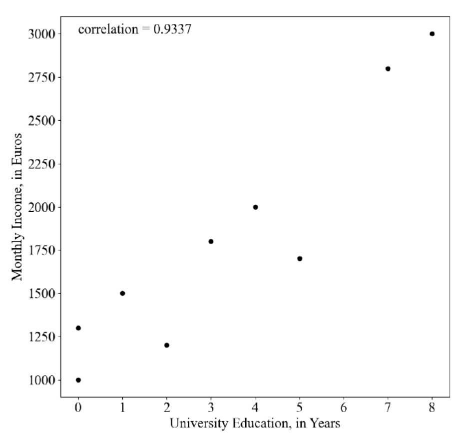
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

An example

- Is there a correlation between house seeling price and house size?
- r = 0.83378
- $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.834\sqrt{98}}{\sqrt{1-0.834^2}} = 14.95$
- How do we interpret this value?
- It tells us how likely we are to observe data in the sample under the assumption that H_0 is true.



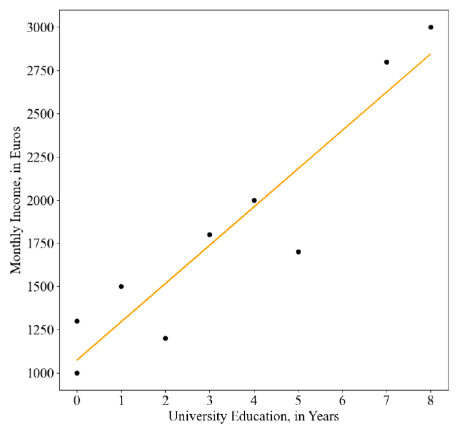
Revision: Bivariate regression



Just by looking at the plot, can you identify the straight line which best describes the joint variation between X and Y?

Regression analysis

Find the line with the best fit: $Y_i = \alpha + \beta X_i + \epsilon_i$



Linear regression model:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

- $\hat{Y}_i$ — Predicted outcome: $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$
- α — Intercept: expected value of Y when $X = 0$
- β — Slope: expected change in Y for a one-unit increase in X
- ϵ_i — Error / residual: difference between the observed and predicted value

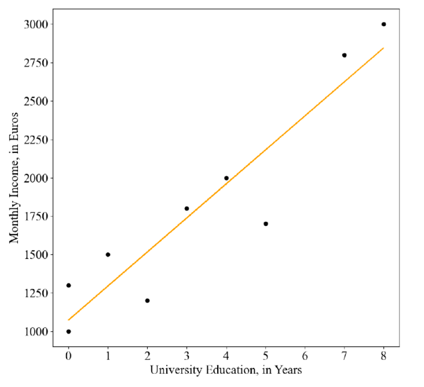
$$\epsilon_i = Y_i - \hat{Y}_i$$

Regression Analysis – Fitted Model

Estimated regression:

$$\widehat{\text{income}} = \hat{\alpha} + \hat{\beta} \cdot \text{education}$$

$$\widehat{\text{income}} = 1072.55 + 221.57 \cdot \text{education}$$



Regression Analysis – Interpreting Coefficients

Intercept $\hat{\alpha} = 1072.549$

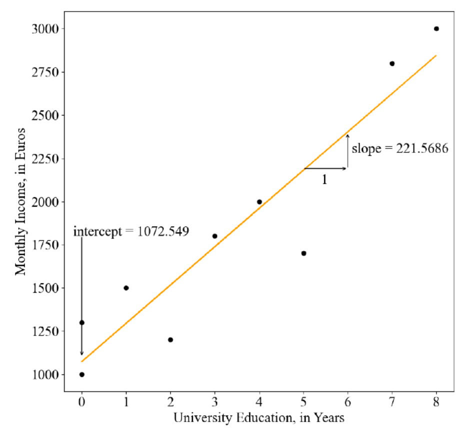
- Expected income when education $= 0$ years:

$$\widehat{\text{income}} = 1072.549 + 221.5685 \cdot 0$$
$$= 1072.549$$

Slope $\hat{\beta} = 221.5685$

- Each additional year of university education increases expected income by **221.57** euros, on average:

$$\widehat{\text{income}} = 1072.549 + 221.5685 \cdot 1$$
$$= 1294.1175$$

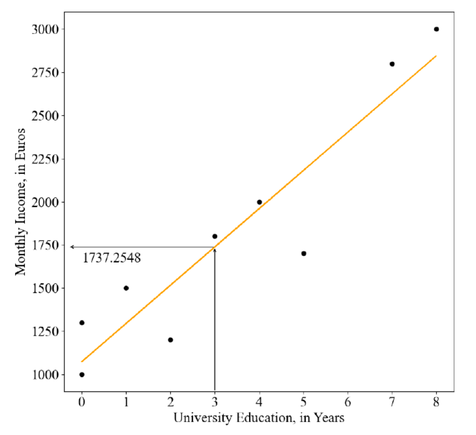


Regression Analysis – Making Predictions

Example: education $= 3$ years

$$\widehat{\text{income}} = 1072.55 + 221.57 \cdot 3 = 1737.25$$

We can plug any X value (years of education) into the model to predict expected income.



Regression Analysis – Residuals

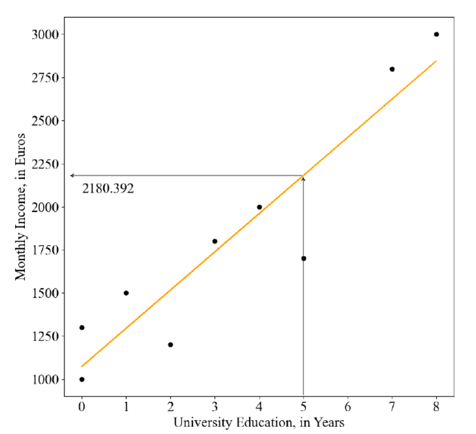
For a person with education = 5 years:

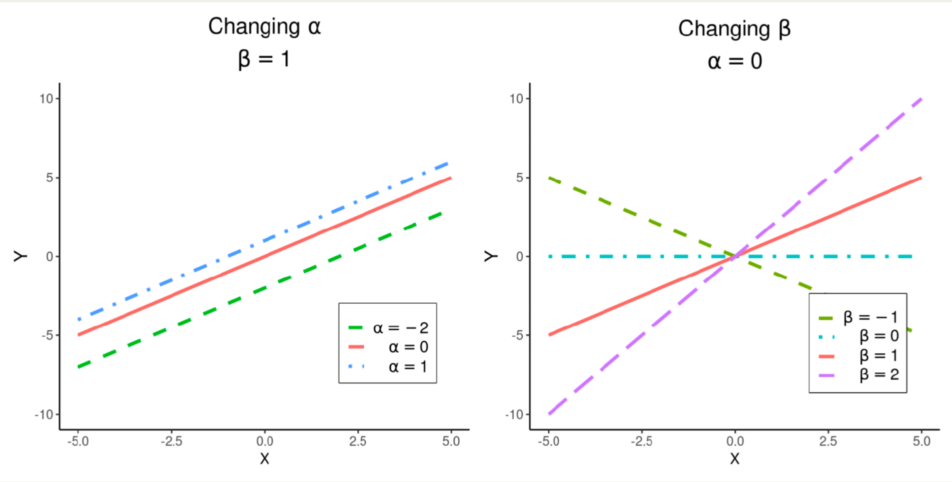
$$\widehat{\text{income}} = 1072.55 + 221.57 \cdot 5 = 2180.39$$

If actual income = 1700:

$$\text{Residual} = Y - \widehat{Y} = 1700 - 2180.39 = -480.39$$

Negative residual → observed income is below predicted.





Varieties of linear relationships

Ordinary Least Squares (OLS)

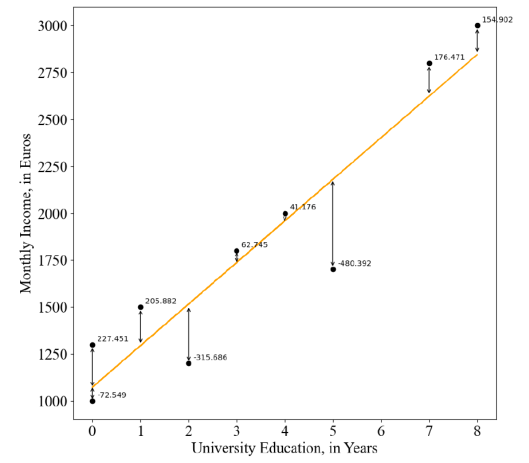
Ordinary Least Squares (OLS)

- OLS is short for "Ordinary Least Squares"
- The best line is the line that minimizes the sum of squared errors (SSE)
- The residuals are the vertical deviations from the line (the observed fitting errors):

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$$

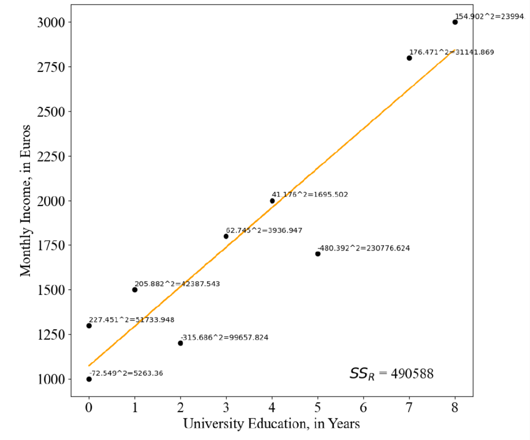
- SSE: the sum of squared differences between the actual and predicted values of Y.
- $$SSE = \sum_{i=1}^{n} (\hat{\epsilon}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (\hat{\alpha} - \hat{\beta} X_i))^2$$
- The goal is to minimize this!

Ordinary Least Squares (OLS)



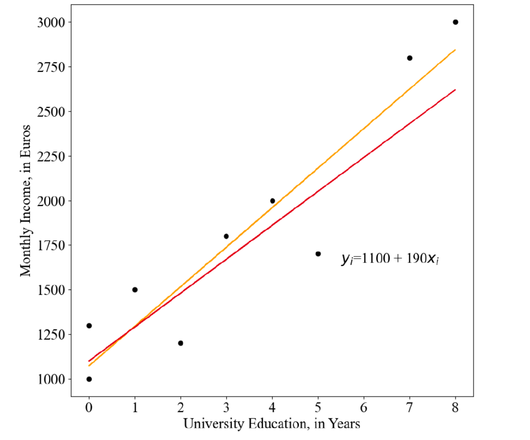
Residuals are the vertical distances between observed points and the regression line.

Ordinary Least Squares (OLS)



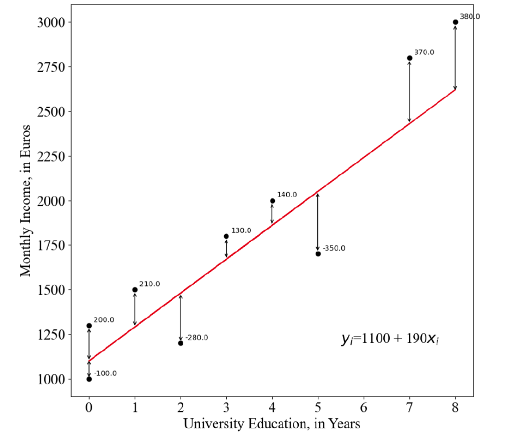
OLS chooses the line that minimizes the squared residuals (errors).

Ordinary Least Squares (OLS)



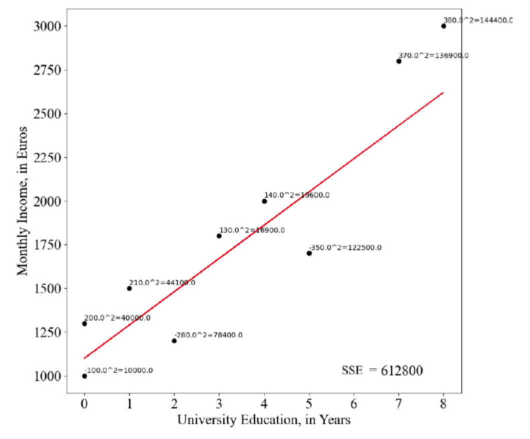
Another possible line — but its total squared residuals (SSE) are larger.

Ordinary Least Squares (OLS)



We are squaring the new residulas to compare lines by their **sum of squared errors (SSE)**.

Ordinary Least Squares (OLS)



$$612{,}800 > 490{,}588 \;\Rightarrow\; \mathrm{SSE}_{\mathrm{red}} > \mathrm{SSE}_{\mathrm{orange}}$$

The orange line has the smaller SSE: it is the better fit.

How to pick the best line?

- How to pick the best line? Get the best slope and best intercept using differential calculus.
- For $Var(x) \neq 0$, the slope coefficient $\hat{\beta}_1$ is given by

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\text{Cov}(x, y)}{\text{Var}(x)}.$$

- The intercept coefficient $\hat{\beta}_0$ is given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{where} \quad \bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}, \quad \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}.$$

- An estimator is unbiased if its expected value is identical to the population value.
- The OLS estimator is best in the sense that it has the lowest variance among all unbiased estimators.

Getting the coefficients

In the lecture you saw how least squares (LS) are point estimates for parameters:

$$\hat{\alpha} = \hat{y} - \hat{\beta}\bar{X} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

These are the best estimates according to the Gauss-Markov theorem.

OLS: Interpretation

- The regression model is: $\hat{Y} = \hat{\alpha} + \hat{\beta}X$
- Interpreting the slope coefficient $\hat{\beta}$:

OLS: Interpretation

- The regression model is: $\hat{Y} = \hat{\alpha} + \hat{\beta}X$
- Interpreting the slope coefficient $\hat{\beta}$: **On average**, a one-unit increase in X leads to a $\hat{\beta}$ unit increase in the predicted value of Y.
- More generally, the marginal effect of an infinitesmall change in X on Y, i.e. $\frac{\partial \hat{Y}}{\partial x} = \hat{\beta}$, is constant (independent of X) in case of OLS.

OLS: Interpretation

- The regression model is: $\hat{Y} = \hat{\alpha} + \hat{\beta}X$
- Interpreting the slope coefficient $\hat{\beta}$: **On average**, a one-unit increase in X leads to a $\hat{\beta}$ unit increase in the predicted value of Y.
- More generally, the marginal effect of an infinitesmall change in X on Y, i.e. $\frac{\partial \hat{Y}}{\partial x} = \hat{\beta}$, is constant (independent of X) in case of OLS.
- Interpreting the intercept coefficient:

OLS: Interpretation

- The regression model is: $\hat{Y} = \hat{\alpha} + \hat{\beta}X$
- Interpreting the slope coefficient $\hat{\beta}$: **On average**, a one-unit increase in X leads to a $\hat{\beta}$ unit increase in the predicted value of Y.
- More generally, the marginal effect of an infinitesmall change in X on Y, i.e. $\frac{\partial \hat{Y}}{\partial x} = \hat{\beta}$, is constant (independent of X) in case of OLS.
- Interpreting the intercept coefficient: When X is zero, the predicted value for $\hat{Y}$ is $\hat{\alpha}$. Note that is may **not** be a meaningful quantity.

An example

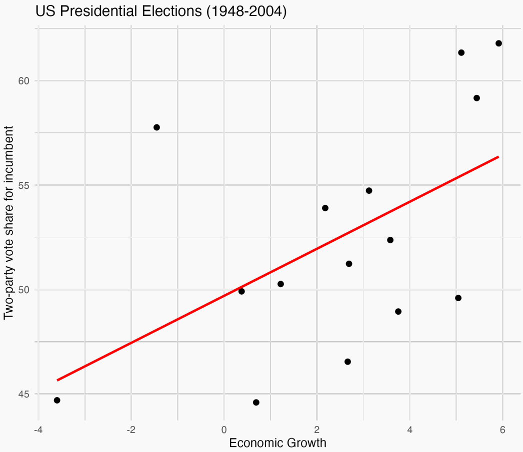
Year	VoteShare	Growth	$(y_i - \bar{y})$	$(x_i - \bar{x})$
1948	52.37	3.579	-0.088	1.131
1952	44.595	.691	-7.863	-1.757
1956	57.764	-1.451	5.306	-3.899
1960	49.913	.377	-2.545	-2.071
1964	61.344	5.109	8.886	2.661
1968	49.596	5.043	-2.862	2.595
1972	61.789	5.914	9.331	3.466
1976	48.948	3.751	-3.510	1.303
1980	44.697	-3.597	-7.761	-6.045
1984	59.17	5.440	6.712	2.992
1988	53.902	2.178	1.444	-0.270
1992	46.545	2.662	-5.913	0.214
1996	54.736	3.121	2.278	0.673
2000	50.265	1.219	-2.193	-1.229
2004	51.233	2.690	-1.225	0.242
	$\bar{y} = 52.4578$	$\bar{x} = 2.4484$		

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{111.559}{99.0181} = 1.127$$

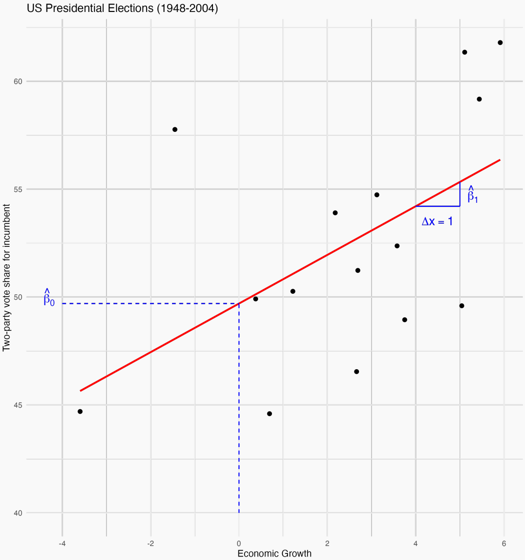
$$\hat{\alpha} = \bar{y} - \hat{\beta}_1\bar{x} = 52.4578 - 1.127 \times 2.4484 = 49.699$$

An example

- OLS model estimation: $\widehat{\text{VoteShare}}_i = 49.699 + 1.127 \times \text{Growth}_i$
- SSE is minimized at $\Sigma_{i=1}^{n} e_i^2 = 311.1486$



An example


US Presidential Elections (1948-2004)
Two-party vote share for incumbent
Economic Growth
60
55
50
45
40
-4
-2
0
2
4
6
$\hat{\beta}_0$
$\hat{\beta}_1$
$\Delta x = 1$

Regression Diagnostics

Regression Diagnostics: Residuals

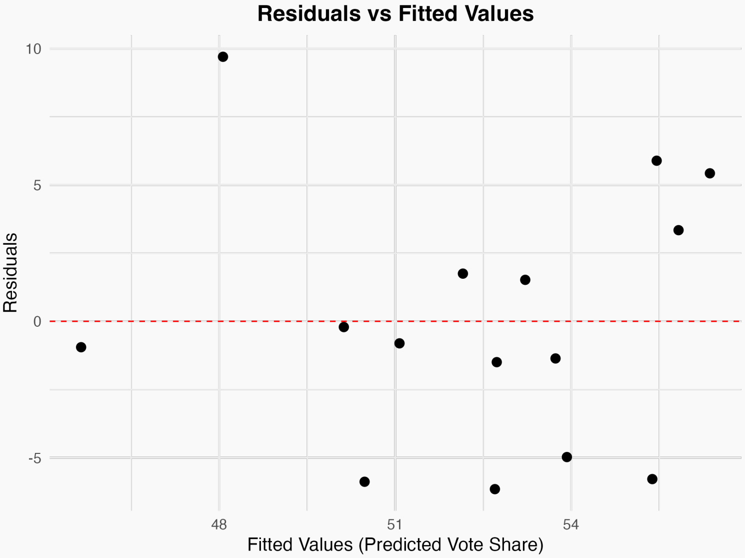
- Patterns in residuals signal that systematic influences on Y still have not been captured by our model, or that our model misrepresents the data, or that errors do not have a constant variance.
- The residual plot is a diagnostic plot as it helps us to detect patterns in the residuals.

Regression Diagnostics: Residuals

- Patterns in residuals signal that systematic influences on Y still have not been captured by our model, or that our model misrepresents the data, or that errors do not have a constant variance.
- The residual plot is a diagnostic plot as it helps us to detect patterns in the residuals.
- Residual plot: a *scatterplot of the regression residuals* against the explanatory variable X or the predicted values $\hat{Y}$.
- Ideally, residual plots should looks as if the pattern was generated by pure chance.
- By construction, OLS residuals sum to 0:

$$\Sigma_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta_1 x_i}) = \Sigma_{i=1}^{n}(y_i - \hat{y}_i) = \Sigma_{i=1}^{n} e_i = 0$$

Regression Diagnostics: Residuals



Assumptions of Linear Regression

OLS Assumptions: Why they matter

OLS gives us the **Best Linear Unbiased Estimators (BLUE)** under a set of assumptions known as the **Gauss–Markov conditions**.

OLS Assumptions: Why they matter

OLS gives us the **Best Linear Unbiased Estimators (BLUE)** under a set of assumptions known as the **Gauss–Markov conditions**.

- If these assumptions hold, OLS estimates of α and β are:
  - Unbiased: $\mathbb{E}[\hat{\beta}] = \beta$
  - Efficient: they have the smallest variance among all linear unbiased estimators
  - Consistent: as $n \to \infty$, $\hat{\beta} \to \beta$

OLS Assumptions: Why they matter

OLS gives us the **Best Linear Unbiased Estimators (BLUE)** under a set of assumptions known as the **Gauss–Markov conditions**.

- If these assumptions hold, OLS estimates of α and β are:
  - Unbiased: $\mathbb{E}[\hat{\beta}] = \beta$
  - Efficient: they have the smallest variance among all linear unbiased estimators
  - Consistent: as $n \to \infty$, $\hat{\beta} \to \beta$
- If assumptions are violated, inferences (t-tests, confidence intervals, p-values) can be **misleading**.

OLS Assumptions about the error term

Model:

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad \text{with } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Assumptions:

1. **Linearity in parameters:** the model is correctly specified as linear in α and β.

OLS Assumptions about the error term

Model:

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad \text{with } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Assumptions:

1. **Linearity in parameters:** the model is correctly specified as linear in α and β.
2. **Zero conditional mean:** $\mathbb{E}[\epsilon_i | X_i] = 0$. $\Rightarrow$ X contains no omitted variables that systematically affect Y.

OLS Assumptions about the error term

Model:

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad \text{with } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Assumptions:

1. **Linearity in parameters:** the model is correctly specified as linear in α and β.
2. **Zero conditional mean:** $\mathbb{E}[\epsilon_i|X_i] = 0$. $\Rightarrow$ X contains no omitted variables that systematically affect Y.
3. **Homoscedasticity:** $\text{Var}(\epsilon_i|X_i) = \sigma^2$ for all i. $\Rightarrow$ the spread of residuals is constant across all X.

OLS Assumptions about the error term

Model:

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad \text{with } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Assumptions:

1. **Linearity in parameters:** the model is correctly specified as linear in α and β.
2. **Zero conditional mean:** $\mathbb{E}[\epsilon_i | X_i] = 0$. $\Rightarrow$ X contains no omitted variables that systematically affect Y.
3. **Homoscedasticity:** $\text{Var}(\epsilon_i | X_i) = \sigma^2$ for all i. $\Rightarrow$ the spread of residuals is constant across all X.
4. **No autocorrelation:** $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. $\Rightarrow$ errors across observations are independent.

OLS Assumptions about the error term

Model:

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad \text{with } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Assumptions:

1. **Linearity in parameters:** the model is correctly specified as linear in α and β.
2. **Zero conditional mean:** $\mathbb{E}[\epsilon_i | X_i] = 0$. $\Rightarrow$ X contains no omitted variables that systematically affect Y.
3. **Homoscedasticity:** $\text{Var}(\epsilon_i | X_i) = \sigma^2$ for all i. $\Rightarrow$ the spread of residuals is constant across all X.
4. **No autocorrelation:** $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. $\Rightarrow$ errors across observations are independent.
5. **Normality:** ϵ_i is normally distributed (needed for valid t and F tests in small samples).

Intuition: why these assumptions matter

- If $\mathbb{E}[\epsilon_i|X_i] \neq 0$, $\hat{\beta}$ is biased: we systematically over- or underestimate the true relationship.

Intuition: why these assumptions matter

- If $\mathbb{E}[\epsilon_i|X_i] \neq 0$, $\hat{\beta}$ is biased: we systematically over- or underestimate the true relationship.
- If $\text{Var}(\epsilon_i|X_i)$ is not constant (heteroscedasticity), $\hat{\beta}$ remains unbiased but inefficient: standard errors and p-values are unreliable.

Intuition: why these assumptions matter

- If $\mathbb{E}[\epsilon_i|X_i] \neq 0$, $\hat{\beta}$ is biased: we systematically over- or underestimate the true relationship.
- If $\text{Var}(\epsilon_i|X_i)$ is not constant (heteroscedasticity), $\hat{\beta}$ remains unbiased but inefficient: standard errors and p-values are unreliable.
- If errors are autocorrelated (common in time series), inference is again unreliable: $\text{se}(\hat{\beta})$ is underestimated.

Intuition: why these assumptions matter

- If $\mathbb{E}[\epsilon_i|X_i] \neq 0$, $\hat{\beta}$ is biased: we systematically over- or underestimate the true relationship.
- If $\text{Var}(\epsilon_i|X_i)$ is not constant (heteroscedasticity), $\hat{\beta}$ remains unbiased but inefficient: standard errors and p-values are unreliable.
- If errors are autocorrelated (common in time series), inference is again unreliable: $\text{se}(\hat{\beta})$ is underestimated.
- If normality is violated, large samples still allow inference via the Central Limit Theorem, but small-sample t-tests become invalid.

Intuition: why these assumptions matter

- If $\mathbb{E}[\epsilon_i|X_i] \neq 0$, $\hat{\beta}$ is biased: we systematically over- or underestimate the true relationship.
- If $\text{Var}(\epsilon_i|X_i)$ is not constant (heteroscedasticity), $\hat{\beta}$ remains unbiased but inefficient: standard errors and p-values are unreliable.
- If errors are autocorrelated (common in time series), inference is again unreliable: $\text{se}(\hat{\beta})$ is underestimated.
- If normality is violated, large samples still allow inference via the Central Limit Theorem, but small-sample t-tests become invalid.

OLS is remarkably robust – unbiasedness mainly requires the **zero conditional mean assumption**. Other violations primarily affect the reliability of inference (standard errors and significance tests).

OLS conditions

- **Variation in X:** $Var(X) \neq 0$ — otherwise $\hat{\beta}$ is undefined.
- **Sample size:** $n > k$ (number of parameters) — we need enough observations.
- **No perfect multicollinearity:** in multiple regression, no independent variable can be an exact linear combination of others.

These are mechanical conditions ensuring that the OLS equations have a unique, computable solution.

Regression inference

Inference about β: the logic

OLS gives us an estimate $\hat{\beta}$, but it's based on a sample. We want to know: does the evidence suggest that the true slope β differs from zero?

Steps:

1. Formulate hypotheses:

Inference about β: the logic

OLS gives us an estimate $\hat{\beta}$, but it's based on a sample. We want to know: does the evidence suggest that the true slope β differs from zero?

Steps:

1. Formulate hypotheses:

$$H_0 : \beta = 0 \quad \text{vs.} \quad H_A : \beta \neq 0$$

Inference about β: the logic

OLS gives us an estimate $\hat{\beta}$, but it's based on a sample. We want to know: does the evidence suggest that the true slope β differs from zero?

Steps:

1. Formulate hypotheses:

$$H_0 : \beta = 0 \quad \text{vs.} \quad H_A : \beta \neq 0$$

2. Compute the **test statistic**:

$$t = \frac{\hat{\beta} - 0}{\text{se}(\hat{\beta})}$$

Inference about β: the logic

OLS gives us an estimate $\hat{\beta}$, but it's based on a sample. We want to know: does the evidence suggest that the true slope β differs from zero?

Steps:

1. Formulate hypotheses:

$$H_0 : \beta = 0 \quad \text{vs.} \quad H_A : \beta \neq 0$$

2. Compute the **test statistic**:

$$t = \frac{\hat{\beta} - 0}{\mathrm{se}(\hat{\beta})}$$

3. Compare $|t|$ to the critical value $t_{n-2,1-\alpha/2}$ or compute a p-value.

If $|t|$ is large (small p-value), we reject H_0 and conclude that X has a statistically significant linear association with Y.

Standard errors and sampling uncertainty

- The **standard error of** $\hat{\beta}$ measures how much $\hat{\beta}$ would vary across repeated random samples.

$$\mathrm{se}(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

- The residual standard deviation $\hat{\sigma}$ is:

$$\hat{\sigma} = \sqrt{\frac{\mathrm{SSE}}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

- Smaller $\mathrm{se}(\hat{\beta})$ means more precise estimation of β.

Standard errors and sampling uncertainty

- The **standard error of** $\hat{\beta}$ measures how much $\hat{\beta}$ would vary across repeated random samples.

$$\text{se}(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}$$

- The residual standard deviation $\hat{\sigma}$ is:

$$\hat{\sigma} = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$

- Smaller $\text{se}(\hat{\beta})$ means more precise estimation of β.

Precision improves when: (1) residuals are smaller (better fit), (2) n is larger, and (3) X has greater variation.

Confidence interval for β

We can express statistical uncertainty using a confidence interval:

$$\hat{\beta} \pm t_{n-2,1-\alpha/2} \cdot \mathrm{se}(\hat{\beta})$$

- 95% confidence interval $\Rightarrow$ we are 95% confident the true β lies in this range.
- If the CI does **not include 0**, the effect is statistically significant at the 5% level.

Example: $\hat{\beta} = 221.6$, $\mathrm{se}(\hat{\beta}) = 35.0 \Rightarrow$ CI = $[152.2, 291.0] \Rightarrow$ Each additional year of education increases expected income by 152–291 euros.

Interpreting significance and magnitude

- **Statistical significance:** whether the relationship is distinguishable from zero given sampling uncertainty.
- **Substantive magnitude:** whether the size of $\hat{\beta}$ is meaningful in context.
- In large samples, even small effects can be statistically significant.
- In small samples, large but noisy effects may fail to reach significance.

Always report:

- point estimate $\hat{\beta}$
- standard error
- p-value or confidence interval

and interpret them in the context of the research question.

Model Fit and Goodness of Fit

Explained vs. unexplained variation

$$Y_i = \hat{Y}_i + e_i$$

- **Total variation:** $SST = \sum(Y_i - \bar{Y})^2$
- **Explained variation:** $SSR = \sum(\hat{Y}_i - \bar{Y})^2$
- **Unexplained variation:** $SSE = \sum(Y_i - \hat{Y}_i)^2$

Relationship:

$$SST = SSR + SSE$$

R^2: Coefficient of determination

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

- Measures the proportion of variance in Y explained by X.
- $R^2 = 0$ means the model explains none of the variation.
- $R^2 = 1$ means perfect fit (rarely observable in real data).
- A higher R^2 indicates better fit, but not necessarily a better or causal model.

In political science and social data, R^2 values around 0.3–0.6 are typical — relationships are often probabilistic, not deterministic.

Adjusted R^2 (for multiple regression)

When we include more variables:

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

- Penalizes adding variables that don't improve model fit much.
- Only increases if new variable improves explanatory power beyond chance.

Adjusted R^2 is preferred when comparing models with different numbers of predictors.

Summary

Putting it all together

1. Estimate coefficients $(\hat{\alpha}, \hat{\beta})$ by OLS.
2. Check assumptions and residual plots.
3. Compute standard errors.
4. Perform hypothesis test:
$$t = \frac{\hat{\beta}}{se(\hat{\beta})}$$
5. Compute and interpret confidence intervals.
6. Evaluate overall model fit (R^2, adjusted R^2).
7. Translate results into substantive conclusions.