


Applied Statistical Analysis I

Contingency tables, correlation & bivariate regression

Elena Karagianni, PhD Candidate

karagiae@tcd.ie

 October 8, 2025

Department of Political Science, Trinity College Dublin

Joint and conditional probability distributions

Joint and conditional probability distributions

- What is a contingency table?

Joint and conditional probability distributions

- What is a contingency table?
- What is a joint probability distribution?

Joint and conditional probability distributions

- What is a contingency table?
- What is a joint probability distribution?
- What is a conditional probability distribution?

Contingency Tables

- A contingency (or 'cross-tabulation') table displays the number of subjects observed at all combinations of possible outcomes for the two variables.
- Used for the analysis of **categorical** variables.

Gender	Democrat	Independent	Republican	Total
Females	573	516	422	1511
Males	386	475	399	1260
Total	959	991	821	2771

Contingency Tables

- A contingency (or 'cross-tabulation') table displays the number of subjects observed at all combinations of possible outcomes for the two variables.
- Used for the analysis of **categorical** variables.

Gender	Democrat	Independent	Republican	Total
Females	573	516	422	1511
Males	386	475	399	1260
Total	959	991	821	2771

The row totals and the column totals are called the **marginal distributions**. For example, the sample marginal distribution for party identification is 959, 991, 821 for each category.

Contingency Tables

Gender	Democrat	Independent	Republican	Total
Females	573	516	422	1511
Males	386	475	399	1260
Total	959	991	821	2771

- Is there an association between gender and party affiliation?
- Does party affiliation depend on gender?

Joint distribution

- The **joint distribution** describes the probability that two variables (e.g., X and Y) simultaneously take some values.
- What is the probability of the gender being 'female' and the party id being 'Democrat'?

$$\frac{573}{2771} = 0.2067846 = 21\%$$

Gender	Democrat	Independent	Republican	Total
Females	573	516	422	1511
Males	386	475	399	1260
Total	959	991	821	2771

Conditional distribution

- The **conditional distribution** describes the probability of one variable Y taking different values, conditional on another variable X having a specific value.
- What is the probability of the party id being 'Democrat' conditional on the gender being 'female'?

$$\frac{573}{1511} = 0.3792191 = 38\%$$

Gender	Democrat	Independent	Republican	Total
Females	573	516	422	1511
Males	386	475	399	1260
Total	959	991	821	2771

Why is this important?

- To answer questions such as “Is party identification associated with gender?” we refer to the concepts of **statistical independence and dependence**.
- “Two categorical variables are statistically independent if the population conditional distributions on one of them are identical at each category of the other”.
- “The variables are statistically dependent if the conditional distributions are not identical”

(Agresti and Finlay 2009, 223)

Independence and Dependence

Why is this important?

To answer questions such as “Is party identification associated with gender?” we refer to the concepts of **statistical independence and dependence**.

Independence and Dependence

Why is this important?

To answer questions such as “Is party identification associated with gender?” we refer to the concepts of **statistical independence and dependence**.

- **Statistical independence:**

Two categorical variables are independent if the **population conditional distributions** on one variable are *identical for each category* of the other.

- **Statistical dependence:**

The variables are dependent if these conditional distributions *differ across categories*.

(Agresti & Finlay, 2009, p. 223)

Independence and Dependence

Ethnic Group	Democrat	Independent	Republican	Total
White	440 (44%)	140 (14%)	420 (42%)	1000 (100%)
Black	44 (44%)	14 (14%)	42 (42%)	100 (100%)
Hispanic	110 (44%)	35 (14%)	105 (42%)	250 (100%)

→ The conditional distribution is the same in each row.

→ What does that mean?

χ^2 Test of Independence

χ^2 Test of Independence

- What is the χ^2 Test of Independence?
- What is the χ^2 -distribution?

χ^2 Test of Independence

- We want to test if two categorical variables are independent $\rightarrow \chi^2$ test
- Formulate the hypotheses:

χ^2 Test of Independence

- We want to test if two categorical variables are independent $\rightarrow \chi^2$ test
- Formulate the hypotheses:
 - H_0 : The variables are statistically independent.
 - H_A : The variables are statistically dependent.
- Conditions for test: randomization and large sample Agresti and Finlay, 2009, 224

χ^2 Test of Independence

- This test compares the observed with the expected frequencies (f_o vs. f_e)
- Test statistic:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

- **Expected frequency:** This is the count expected in a cell if the variables were independent (as per H_0).

$$f_{e,ij} = \frac{(\text{row total}) \cdot (\text{column total})}{N}$$

Expected frequencies (f_o)

Gender	Democrat	Independent	Republican	Total
Female	573 (522.9)	516 (540.4)	422 (447.7)	1511
Male	386 (436.1)	475 (450.6)	399 (373.3)	1260
Total	959	991	821	2771

$$f_e = \frac{959 \times 1511}{2771} = 522.9$$

- Observed f_o for Female Democrats = 573, but expected $f_e = 522.9$.
- That's a bit higher than what would be expected under independence.

χ^2 Test of Independence

Let's do the χ^2 -test:

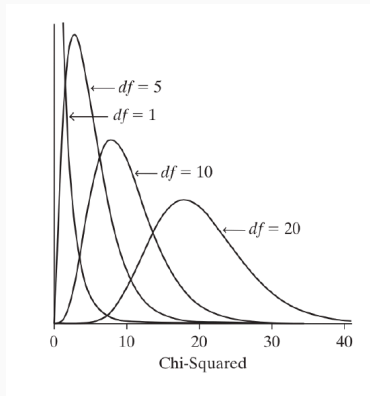
$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(573 - 522.9)^2}{522.9} + \frac{(516 - 540.4)^2}{540.4} + \dots + \frac{(399 - 373.3)^2}{373.3} = 16.2$$

- How do we interpret this value?
- How likely are we to observe data in sample (this test statistic), under the assumption that H_0 is true?
- → We need to look at the probability distribution!

χ^2 Distribution

Characteristics:

- Always **non-negative** (sums of squared differences divided by positive expected frequencies).
- **Right-skewed**.
- Shape depends on the **degrees of freedom (df)**:
 - Mean $\mu = df$
 - Standard deviation $\sigma = \sqrt{2 df}$
 - Spreads out for larger df
 - As df increases, skew decreases and the curve becomes more bell-shaped



χ^2 Distribution — Degrees of Freedom

- For a contingency table with r rows and c columns:

$$\text{df} = (r - 1)(c - 1)$$

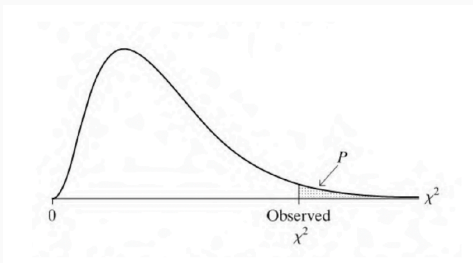
- Example: 2×3 table

$$\text{df} = (2 - 1)(3 - 1) = 1 \times 2 = 2$$

The number of degrees of freedom determines the exact shape of the χ^2 distribution.

χ^2 Test – p-value

- The **larger** the χ^2 value, the **stronger** the evidence against H_0 (independence).
- The **p-value** is the **right-tail probability**:
- It measures how likely it is, under H_0 , to obtain a test statistic at least as extreme as the observed one.



Decision rule: If $p < \alpha$ (e.g., 0.05) \Rightarrow **reject H_0** .

Correlation

Scatterplot

A plot, showing two continuous variables (e.g., X and Y) alongside each other → for each observation, the value on X is plotted against the value on Y.

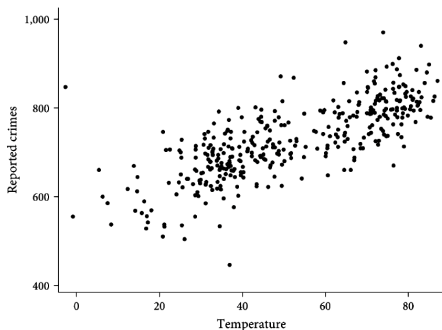


Figure 2.1. Crime and temperature (in degrees Fahrenheit) in Chicago across days in 2018.

- Each point corresponds to an observation in our data.
- In this example, each point is a day in Chicago in 2018.
- The location of each point shows the average temperature (X) and the amount of crime (Y) on that day.

- What is correlation?
- How can we measure it?

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**.

Correlation

How can we measure correlation?

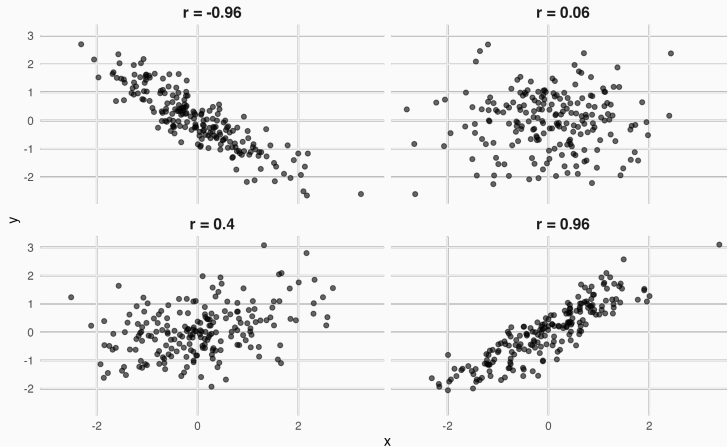
- The most common measure is the **Pearson correlation coefficient** r .
- It is the standardized covariance between two variables X and Y :

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

where \bar{x}, \bar{y} are sample means and s_x, s_y are sample standard deviations.

- Range: $r \in [-1, 1]$
 - $r = 0$: no linear association
 - $r > 0$: positive association
 - $r < 0$: negative association
 - The closer $|r|$ is to 1, the stronger the association.

Correlation

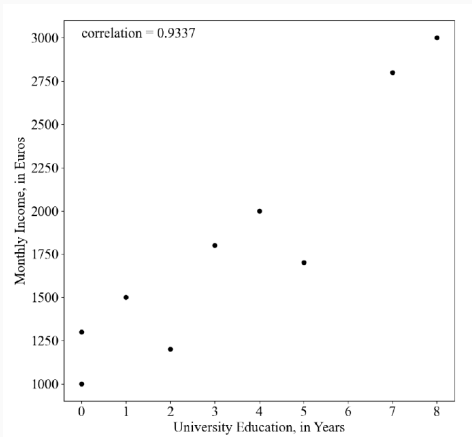


Bivariate regression

Linear regression

- What is a linear regression model?
- What interpretations can we make?

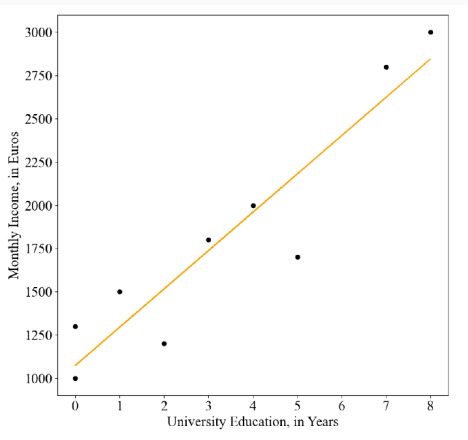
Linear regression



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

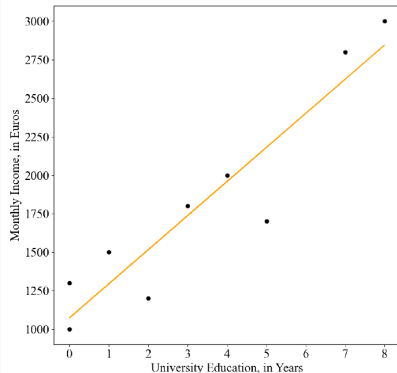
$$\epsilon_i = Y_i - \hat{Y}_i$$

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 $\hat{\alpha} = 1072.549$

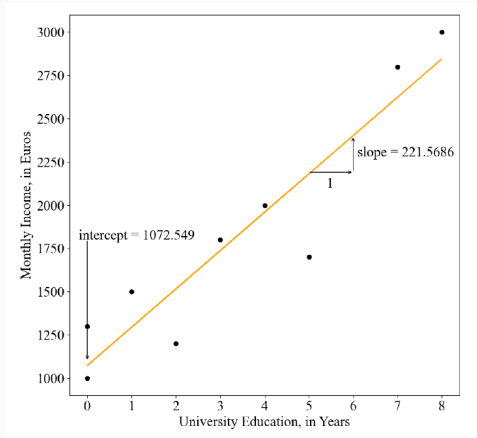
- Expected income when education = 0 years:

$$\begin{aligned}\widehat{\text{income}} &= 1072.549 + 221.5685 \cdot 0 \\ &= 1072.549\end{aligned}$$

Slope $\hat{\beta} = 221.5685$

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

$$\begin{aligned}\widehat{\text{income}} &= 1072.549 + 221.5685 \cdot 1 \\ &= 1294.1175\end{aligned}$$

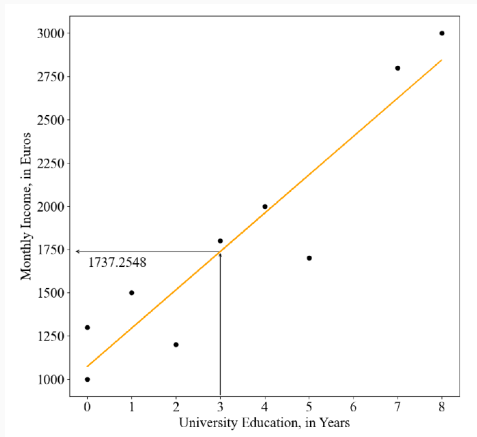


Regression Analysis – Making Predictions

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

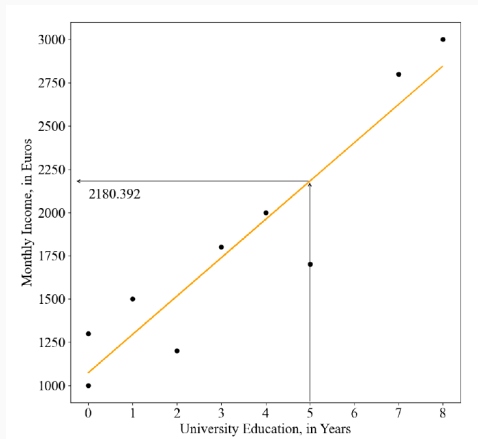
For a person with education = 5 years:

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

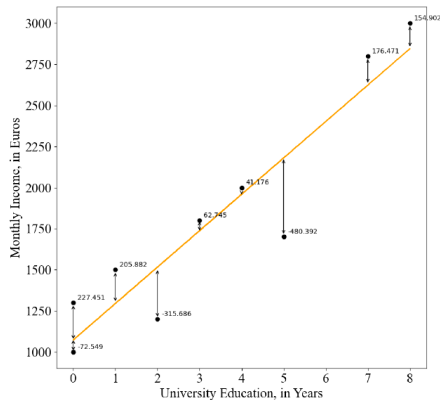
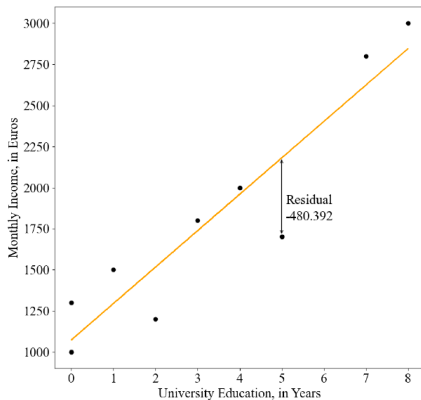
If actual income = 1700:

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

Negative residual \rightarrow observed income is below predicted.



Regression Analysis — Checking Residuals



Residual plots help check if the model fits well

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):

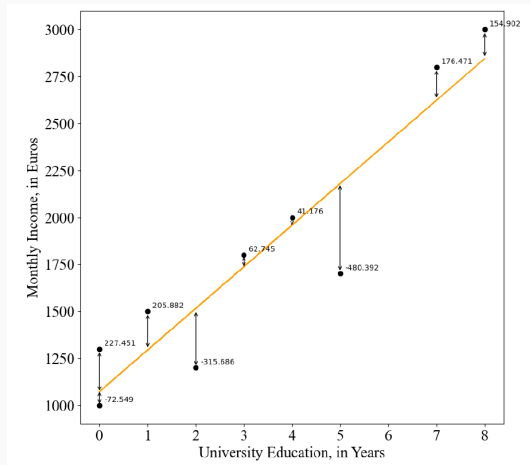
$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$$

- **SSE**: the sum of squared differences between the actual and predicted values of Y.

$$SSE = \sum_{i=1}^n (\hat{e}_i)^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - (\hat{\alpha} - \hat{\beta} X_i))^2$$

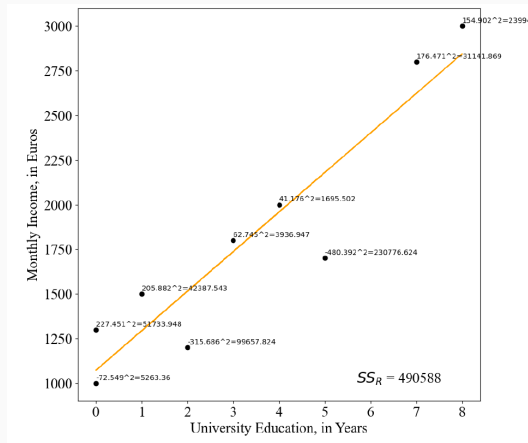
- The goal is to minimize this!

Ordinary Least Squares (OLS)



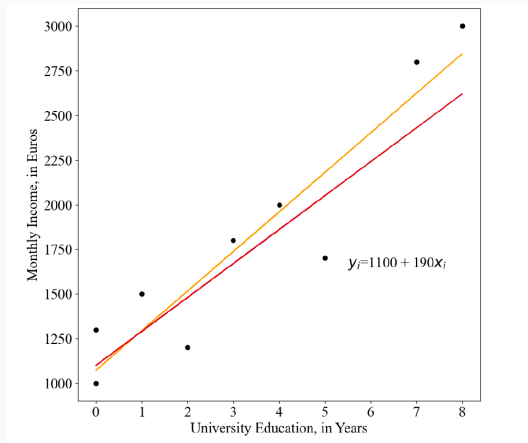
Residuals are the vertical distances between observed points and the regression line.

Ordinary Least Squares (OLS)



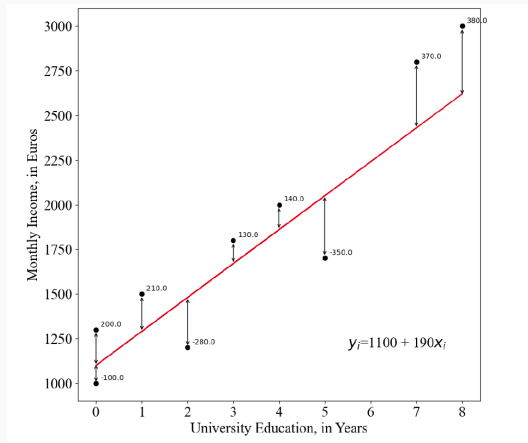
OLS chooses the line that minimizes the squared residuals (errors).

Ordinary Least Squares (OLS)



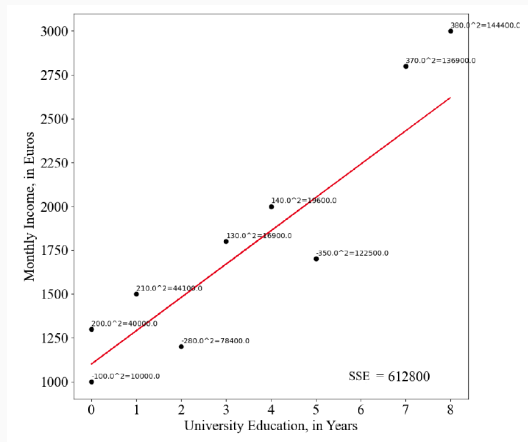
Another possible line — but its total squared residuals (SSE) are larger.

Ordinary Least Squares (OLS)



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

Ordinary Least Squares (OLS)



$$612,800 > 490,588 \Rightarrow SSE_{\text{red}} > SSE_{\text{orange}}$$

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

Assumptions

What are the assumptions of linear regression?

OLS Assumptions

Suppose we have the following bivariate linear model:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

We need **two** assumptions to derive **unbiased** regression coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$:

A1: An almost trivial assumption is that coefficients (i.e., parameters) are **linear**.

A2: We make a **zero conditional mean** assumption:

$$E(\epsilon_i | X) = 0$$

These assumptions are sufficient to estimate **unbiased coefficients** $\hat{\beta}_0$ and $\hat{\beta}_1$ with OLS.

OLS Assumptions

To also estimate the **variance** of the coefficients, we need to make additional assumptions:

A3: We assume constant variance (**homoskedasticity**), regardless of the values of X :

$$\text{Var}(\epsilon_i | X) = \sigma^2$$

A4: We assume **no correlation** among any pair of error terms:

$$\text{Cov}(\epsilon_i, \epsilon_j | X_i, X_j) = 0 \quad \forall i \neq j$$

A5: We assume **normality** of the error term:

$$\epsilon_i | X \sim \mathcal{N}(0, \sigma^2)$$

Standard Errors for Regression Coefficients

- If A3, the **zero conditional mean assumption**, $E(\epsilon_i | X) = 0$, holds, we get:

$$E(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad E(\hat{\beta}_1) = \beta_1$$

- Assuming A5, **normally distributed errors**, $\epsilon | X \sim \mathcal{N}(0, \sigma^2)$, the OLS coefficients themselves are normally distributed:

$$\hat{\beta}_0 \sim \mathcal{N}(\beta_0, \text{Var}(\hat{\beta}_0)) \quad \hat{\beta}_1 \sim \mathcal{N}(\beta_1, \text{Var}(\hat{\beta}_1))$$

- This allows us to calculate **standard errors** based on the normal approximation.