# Econometrics Lecture 1

## EC2METRIE

Dr. Anna Salomons

Utrecht School of Economics (U.S.E.)

14 November 2016

# Econometrics

- Econometrics = using data to **measure causal effects**.
- **Economic theory** suggests important relationships (cause-and-effect), often with policy implications.
  - However, there is only so much we can learn from theorizing: we need empirical evidence to test our models.
  - E.g. economic theory almost never suggests quantitative magnitudes of these causal effects.
- Conclusion: we need to **test these relationships in the real world**, using data.
- Note that econometric techniques are used in other social sciences as well: political science, sociology, psychology..

# Examples of questions that econometrics can help you answer:

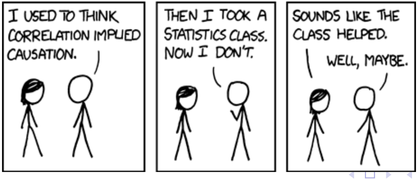
- By how much does a university degree increase your lifetime earnings?
- What is the effect of replacing scholarships with student loans on college attainment rates?
- What is the impact of TV watching on children's cognitive development?
- What is the quantitative effect of reducing class size on student achievement?
- Does increasing the minimum wage cause employment to fall among low-skilled workers?
- What is the impact of voters' trust in government on sympathy for extreme right-wing parties?
- Does foreign aid positively contribute to the growth of an economy?

# Using data to measure causal effects

- Ideally, we would like an **experiment** (randomization) to answer these questions
  - What would be an experiment to estimate the effect of class size on test scores?
- But almost always we only have **observational (=non-experimental) data**.
  - Returns to education
  - Inflation & consumer confidence
  - Foreign aid & growth
  - Minimum wages & unemployment
  - TV watching & cognitive development
  - ...

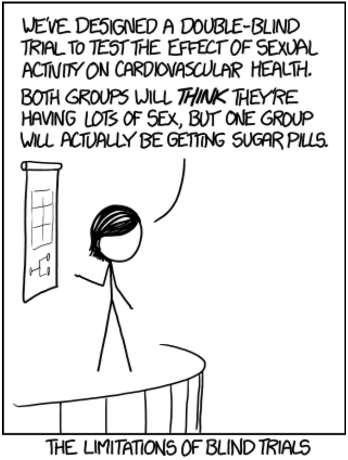
# Aims of this course

- Learn to **apply the basic econometric tools** used for testing economic theories, and **interpret** the estimation results.
- **Focus on applications** – theory is used only as needed to understand the why's of the methods.
- We also deal with **difficulties arising from using observational instead of experimental data** to estimate causal effects
  - confounding effects (omitted variables)
  - "correlation does not imply causation"




(Source: xkcd comics)

# We cannot always perform experiments




(Source: xkcd comics)

# Place in the curriculum

- **Follow-up on Statistics**: further develops regression analysis and testing economic hypotheses
- Further develops **Stata skills** (started in Statistics & Macroeconomics)
- Needed for **Bachelor thesis** and other empirical projects
- **Masters in Economics** requires knowledge of Econometrics

# Course materials

- **Lecture slides**. Note that these may contain additional material not covered in Studenmund.
- **Studenmund**, A.H. (2013), Using Econometrics. A Practical Guide, 6th edition. Pearson Publishing. ISBN 9781292021270 (or ISBN 9780131379985).
- **Course manual** (available on Blackboard).

# Course set-up: meetings

- **Lecture**: weeks 1-8, Mondays, 3.15-5.00 PM, Ruppert Blauw and Rood.
- **Tutorial**: weeks 1-8, Thursdays; individual **schedule available in Osiris**
- **Project group**: meeting in weeks 2, 4, 5 and 7, Tuesdays/Thursdays, meetings of half an hour per group
  - **Schedule posted on Blackboard on Monday 21 November**: first meetings are on Tuesday 22 November for most groups, but there may be a few on Thursday 24 November too.

# Coordinator and tutorial teachers: contact information

- Course coordinator:
  - Sergei Hoxha, econometrics.use@uu.nl
- Tutorial teachers:
  - Tea Elezi, t.elezi@uu.nl
  - Wilfred de Graaf, j.w.degraaf@uu.nl
  - Yolanda Grift, y.grift@uu.nl
  - Sergei Hoxha, s.hoxha@uu.nl
- Project advisors: see course manual

# Project

- Write **empirical research paper** in groups of 5 students, formed during the first tutorial.
  - **Attendance of the first tutorial is mandatory**; only students present in the first tutorial can choose their own project group!
- **Supervised** by a project advisor: your tutorial teacher or additional teaching staff (see course manual).
- Apply the econometric analyses discussed in each week of the course: a **to do list** for the project paper is provided **at the end of each lecture**
- **Datasets** (cross-sectional for weeks 1-5, timeseries for weeks 6-7) are on Blackboard: you can choose your own cross-sectional dataset during the first tutorial, else one is assigned to you. (The time-series dataset can be chosen in week 6.)

# Course set-up: examination

- Note: **no midterm** exam- i.e. classes throughout weeks 1-8
- Individual **written exam**, week 9: **60%** of final grade. See tutorial exercises & sample exams posted on Blackboard.
- **Project paper**, deadline = 20 January 2017: **40%** of final grade
  - Group grade
- See course manual for honor course, repeaters course & exam retake information.

# Course set-up: effort requirement

- Participation in at least 5 out of 8 tutorials.
- Participation in all 4 project group meetings.
- Preliminary draft of project paper (deadline: 23 December 2016) of satisfactory level.
- No effort requirement for course repeaters.

# Course set-up: topics by week

Week 1: Statistics refresher (+overview of regression analysis)

Week 2: Mechanics of the regression model

Week 3: t-test, F-test, specification

Week 4: Functional form and dummy variables

Week 5: Heteroskedasticity and multicollinearity

Week 6: Serial correlation

Week 7: Time-series models

Week 8: Linear probability model (note: this week is not part of the project paper but it is included in the written exam)

# This class

- **Statistics refresher**:
  - Random variables, population, probability distribution
  - Univariate analysis: expected value, variance
  - Bivariate analysis: joint distribution, marginal distribution, conditional distribution, independence, *conditional expectation*, *correlation*, *covariance*
- **Regression analysis**
  - A first introduction to terminology, and causality- more in following weeks.

Topics in *italics* are not discussed in Studenmund.

**Studenmund**:

- Chapter 1
- Sections 17.1 and 17.2

# Review of statistical theory

- We are working towards estimating a multivariate (or multiple) population model of the form:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + .. + \beta_k X_{ki} + \varepsilon_i$$

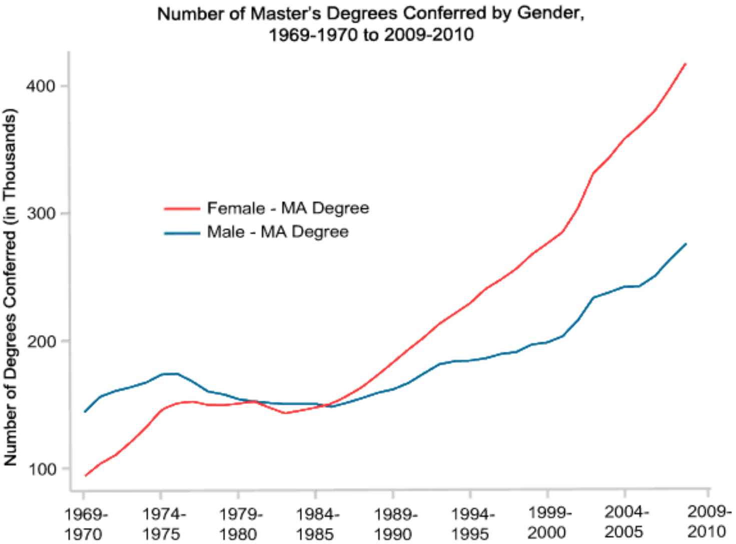
- This is called **regression analysis**- we discuss it in coming weeks
- To do so, we first revisit **univariate analysis** (summarizing one variable) **and bivariate analysis** (summarizing the relationship between 2 variables).
- Most of the material covered this week will be familiar from your Statistics course.

# Motivation for research

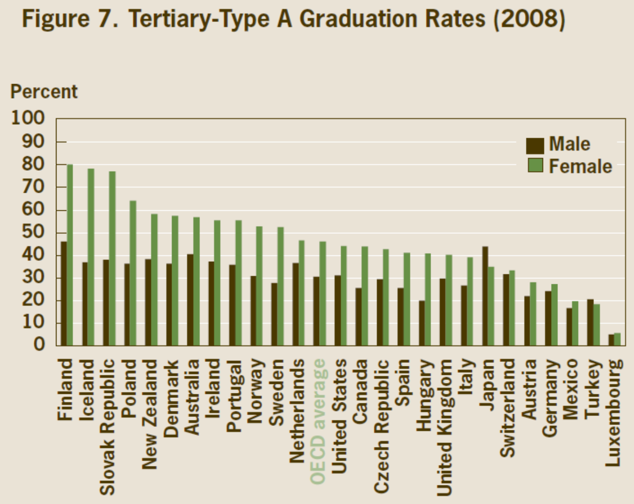
**Dutch newspapers reported** that over the past 30 years:

- The **number of students in tertiary education has increased with 84,000**, but of those, **only 10,000 were male** students!
- The **number of female students has increased 2.5-fold since 1980**.
- Currently, in the Netherlands, there are 6,000 more female students than male students.
- This is in fact an **international phenomenon**. (proof on next slides)

# Evidence for the US



# Evidence for OECD countries



# What about the gender wage gap?

- Documenting the gender wage gap and understanding its causes is still an important research topic among economists
- Two prominent scholars in this field are professors Francine Blau and Lawrence Kahn (both at Cornell University)
- Solution to the gender wage gap proposed by John Oliver (Last Week Tonight): https://www.youtube.com/watch?v=T1p61WrVtEg.

# Example of a research question

- **Do starting salaries of female economics graduates differ from those of male economics graduates?**
  - $H_0$ : there is no difference in starting salaries
  - $H_A$ : there is a difference in starting salaries
- **Note that the alternative hypothesis is based on economic theory**. E.g., we may expect differences in econ graduates' starting salaries by gender due to (a combination of):

# Example of a research question

- **Do starting salaries of female economics graduates differ from those of male economics graduates?**
  - $H_0$ : there is no difference in starting salaries
  - $H_A$ : there is a difference in starting salaries

- **Note that the alternative hypothesis is based on economic theory**. E.g., we may expect differences in econ graduates' starting salaries by gender due to (a combination of):
  - Differences in human capital accumulation;

# Example of a research question

- **Do starting salaries of female economics graduates differ from those of male economics graduates?**
  - $H_0$ : there is no difference in starting salaries
  - $H_A$ : there is a difference in starting salaries
- **Note that the alternative hypothesis is based on economic theory**. E.g., we may expect differences in econ graduates' starting salaries by gender due to (a combination of):
  - Differences in human capital accumulation;
  - Differences in rewards of non-market activities;

# Example of a research question

- **Do starting salaries of female economics graduates differ from those of male economics graduates?**
  - $H_0$ : there is no difference in starting salaries
  - $H_A$ : there is a difference in starting salaries
- **Note that the alternative hypothesis is based on economic theory**. E.g., we may expect differences in econ graduates' starting salaries by gender due to (a combination of):
  - Differences in human capital accumulation;
  - Differences in rewards of non-market activities;
  - Differences in job preferences between men and women;

# Example of a research question

- **Do starting salaries of female economics graduates differ from those of male economics graduates?**
  - $H_0$ : there is no difference in starting salaries
  - $H_A$ : there is a difference in starting salaries

- **Note that the alternative hypothesis is based on economic theory**. E.g., we may expect differences in econ graduates' starting salaries by gender due to (a combination of):
  - Differences in human capital accumulation;
  - Differences in rewards of non-market activities;
  - Differences in job preferences between men and women;
  - Labor market discrimination.

# How to answer this research question

**Do starting salaries of female economics graduates differ from those of male economics graduates?**

- **What is the random variable and the population?**
- Numerically summarize the random variable: univariate analysis
- Analyse the relationship between the random variable and another random variable: bivariate analysis and regression analysis

# Random variable

Note: at this stage we do not have any sample (i.e. dataset) yet!

**Random variable** (r.v.) $X$ = a variable that takes on different values (these are denoted $x_i$) with a given probability for each outcome ($Pr(X = x_i)$).

- In our example, the r.v. is starting salaries of econ grads.
- *Discrete* r.v.: r.v. with a finite number of outcomes ("countable outcomes").
- *Continuous* r.v.: r.v. may take on any numerical value in an interval or collection of intervals ("outcomes from a measuring process").

# Population & probability density function

**Population**: set of all possible outcomes of $X$- we think of populations as infinitely large.

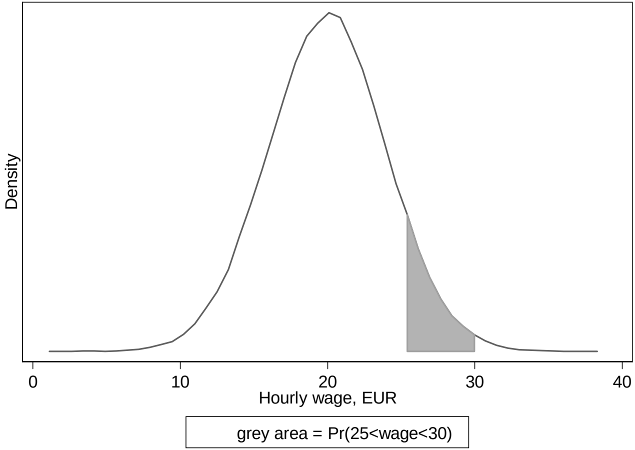
- In our example, the population is all possible starting salaries of economics graduates.

**Probability density function**[1] **(pdf)** function containing the probabilities of different outcomes, denoted $f(x_i) = \Pr(X = x_i)$.

- *Discrete pdf:* pdf for countable outcomes;
- *Continuous pdf:* pdf for non-countable outcomes

[1] Also called the probability distribution function, or the probability function.

# Example of a continuous pdf



# Properties of the probability density function

**Properties of discrete pdf** for $N$ possible outcomes for discrete r.v. $X$:

- All outcomes of $X$ (denoted $x_i$) have a non-negative probability of occurring:

$$\Pr(X = x_i) \geq 0 \text{ for } i = 1, 2, .., N$$

- The sum of all probabilities is equal to 1:

$$\sum_{i=1}^{N} \Pr\left(X = x_i\right) = 1 \text{ or } \sum_{i=1}^{N} f\left(x_i\right) = 1$$

# Properties of the probability density function

**Properties of continuous pdf** for all possible outcomes for continuous r.v. $X$:

- All outcomes of $X$ (denoted $x_i$) have a non-zero probability of occurring:
$$\forall x : f(x_i) \geq 0$$
- The sum of all probabilities is equal to 1:
$$\int_{-\infty}^{+\infty} f(x_i)\, dx = 1$$

# How to answer this research question

**Do starting salaries of female economics graduates differ from those of male economics graduates?**

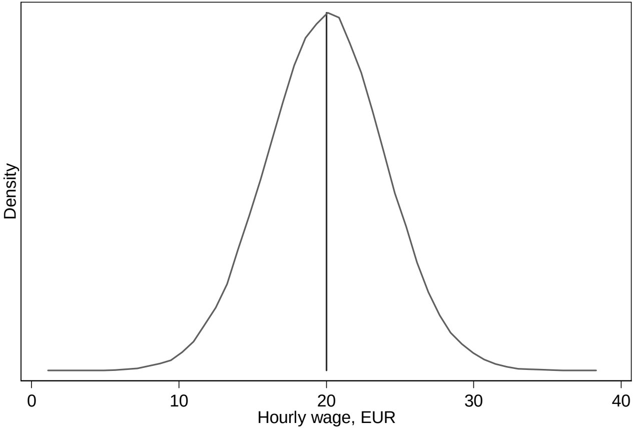
- We have described the random variable and the population, which can be seen from a probability density function.
- We now turn to **numerically summarizing the random variable's population pdf:** this is called **univariate analysis**
  - First moment of the pdf: **expected value** (mean) of $X$
  - Second moment of the pdf: **variance** of $X$

# Expected value

- Consider a discrete random variable $X$, starting salaries of economics graduates.
- The **expected value** of $X$ is its average value (i.e. the mean starting salary) in the population: calculated by weighting each value with the probability that it comes up.
- Hence, to calculate the expected value of $X$:

$$E(X) = EX = \mu_X = \sum_{i=1}^{N} \Pr(X = x_i) * x_i$$

# Expected value: first moment of pdf




E(wage) = 20

# Expected value: numerical example for discrete random variable

The expected value of a roll of a fair die (in a population of infinitely many fair die rolls), $E(X) = 3.5$:

| $x_i$ | $f(x_i) = Pr(X = x_i)$ | $f(x_i) * x_i$ |
|---|---|---|
| 1 | 1/6 | 1/6 |
| 2 | 1/6 | 2/6 |
| 3 | 1/6 | 3/6 |
| 4 | 1/6 | 4/6 |
| 5 | 1/6 | 5/6 |
| 6 | 1/6 | 1 |
| | | SUM = 3.5 |

# Expected value: rules of calculation

1. **When $X$ is a constant with value** $c$, e.g. the expected value of starting salaries when all econ graduates earn exactly the same starting salary:

$$E(c) = c \qquad \text{(rule 1)}$$

2. **When a constant $c$ is added to** $X$, e.g. the expected value of starting salaries when economics graduates all receive the same fixed bonus $c$ on top of a random component $X$:

$$E(X + c) = E(X) + c = \mu_X + c \qquad \text{(rule 2)}$$

# Expected value: rules of calculation

3. **When $X$ is multiplied by constant** $c$: e.g. economics graduates earn $X$ euros or $cX$ dollars, where $c$ is a constant Euro-Dollar exchange rate

$$E(cX) = cE(X) = c\mu_X \qquad \text{(rule 3)}$$

Important: this rule tells us what happens to the expected value of $X$ when the units of measurements of $X$ are changed, e.g. measuring in Euros or thousands of Euros.

# Expected value: rules of calculation

4. **When random variables $X_1$ and $X_2$ are summed**: e.g. the expected value of starting salaries when these are made up of two different income sources (say, baseline wages and overtime pay)

$$E(X_1 + X_2) = E(X_1) + E(X_2) = \mu_{X1} + \mu_{X2} \qquad \text{(rule 4)}$$

$$\text{In general} \quad : \quad E\sum_j X_j = \sum_j EX_j$$

Note that independence between $X_1$ and $X_2$ (discussed later) is not required for this result!

# Expected value: rules of calculation

- Putting the rules together:

$$\begin{aligned} E\left(3X_1 + 2X_2 + 5\right) &= 3E\left(X_1\right) + 2E\left(X_2\right) + 5 \\ &= 3\mu_{X_1} + 2\mu_{X_2} + 5 \end{aligned}$$

- Further practice of these rules in the tutorial.

# Population variance and standard deviation

- The **variance of the random variable** $X$

$$\begin{aligned} Var(X) &= E\left(X - EX\right)^2 = \sum_{i=1}^{N} \left(x_i - \mu_X\right)^2 \Pr\left(X = x_i\right) \\ &= E\left(X^2\right) - \mu_X^2 \text{ (click for proof)} \end{aligned} \tag{var}$$

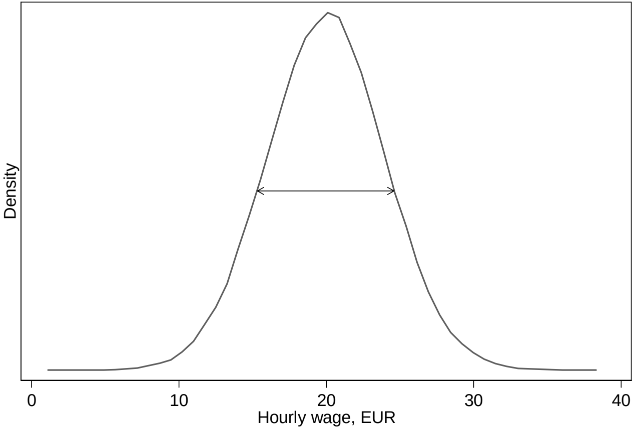
- **Standard deviation of a random variable** $X$

$$Sd(X) = \sqrt{Var(X)} \tag{sd}$$

- Notation:

$$\begin{aligned} Var(X) &= \sigma_X^2 \\ Sd(X) &= \sigma_X \end{aligned}$$

# Variance: second moment of pdf




Variance measures the spread of the pdf

# Population var and sd: numerical example for discrete r.v.

The variance & standard deviation of a roll of a fair die can be calculated as follows (using the previous calculation of $\mu_X = 3.5$):

| $x_i$ | $\mu_X$ | $(x_i - \mu_X)$ | $(x_i - \mu_X)^2$ | $f(x_i)$ | $(x_i - \mu_X)^2 * f(x_i)$ |
|---|---|---|---|---|---|
| 1 | 3.5 | $-2.5$ | 6.25 | 1/6 | $6.25 * 1/6$ |
| 2 | 3.5 | $-1.5$ | 2.25 | 1/6 | $2.25 * 1/6$ |
| 3 | 3.5 | $-0.5$ | 0.25 | 1/6 | $0.25 * 1/6$ |
| 4 | 3.5 | 0.5 | 0.25 | 1/6 | $0.25 * 1/6$ |
| 5 | 3.5 | 1.5 | 2.25 | 1/6 | $2.25 * 1/6$ |
| 6 | 3.5 | 2.5 | 6.25 | 1/6 | $6.25 * 1/6$ |
| | | | | SUM: | $17.5 * \frac{1}{6} = 35/12 \approx 2.92$ |

Hence, $\sigma_X^2 \approx 2.92$ and $\sigma_X \approx \sqrt{2.92} \approx 1.71$.

# Variance: rules of calculation

1. **When $X$ is a constant with value** $c$, e.g. the variance of starting salaries when all econ graduates earn exactly the same starting salary:

$$Var(c) = 0 \qquad \text{(rule 1)}$$

2. **When a constant $c$ is added to $X$**, e.g. the variance of starting salaries when econ graduates all receive the same fixed bonus $c$ on top of a random component $X$:

$$Var(X + c) = Var(X) \qquad \text{(rule 2)}$$

Click for proof of rules 1 & 2 (Appendix).

# Variance: rules of calculation

3. **When $X$ is multiplied by a constant** $c$, e.g. the variance of starting salaries when units of measurements are changed, e.g. economics graduates earn $X$ euros or $cX$ dollars, where $c$ is a constant Euro-Dollar exchange rate

$$Var(cX) = c^2 Var(X) \qquad \text{(rule 3)}$$

Click for proof of rule 3 (Appendix).

# Variance: rules of calculation

4. **When (pairwise) independent random variables are summed**:

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2) \qquad \text{(rule 4)}$$

$$\text{In general} \quad : \quad Var\left(\sum_j X_j\right) = \sum_j Var(X_j)$$

# Variance: rules of calculation

5. **When dependent random variables are summed**:

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2) + 2Cov(X_1, X_2) \qquad \text{(rule 5)}$$

In general :

$$Var\left(\sum_j X_j\right) = \sum_j \sum_k Cov(X_j, X_k)$$

Note: when two dependent random variables are differenced:
$Var(X_1 - X_2) = Var(X_1) + Var(X_2) - 2Cov(X_1, X_2)$.

Click for proof of rules 4 & 5 (Appendix).

# Putting some rules of calculation together

- **Rules 3 & 4**: the variance of the sum of $aX_1$ and $bX_2$ (where $a$ and $b$ are constants) if $X_1$ and $X_2$ are independent:

$$Var(aX_1 + bX_2) = a^2 Var(X_2) + b^2 Var(X_2)$$

- **Rules 3 & 5**: the variance of the sum of $aX_1$ and $bX_2$ (where $a$ and $b$ are constants) if $X_1$ and $X_2$ are dependent:

$$Var(aX_1 + bX_2) = a^2 Var(X_1) + b^2 Var(X_2) + 2abCov(X_1, X_2)$$

# Variance: rules of calculation - examples

$$Var(2) = 0$$

$$\begin{aligned} Var(-2X_1 + 6) &= Var(-2X_1) + Var(6) \\ &= 4Var(X_1) + 0 \\ &= 4Var(X_1) \end{aligned}$$

$$\begin{aligned} Var(3X_1 + 2X_2 + 3) &= \text{For independent variables } X_1 \text{ and } X_2, \\ &= Var(3X_1) + Var(2X_2) + Var(3) \\ &= 9Var(X_1) + 4Var(X_2) \end{aligned}$$

More examples in tutorial.

# How to answer this research question

**Do starting salaries of female economics graduates differ from those of male economics graduates?**

- What is the random variable and the population?
- Numerically summarize the random variable: univariate analysis
- **Analyze the relationship between the random variable and another random variable: bivariate analysis and regression analysis**

# Bivariate analysis

- We now start considering the **relationship between the random variable $X$ and another random variable $G$**:
  - $X=$ starting salary of econ grads;
  - $G=$ dummy variable for gender
    - $G = 0$ for male econ grad
    - $G = 1$ for female econ grad.
- To do this, we need the concepts of **joint, marginal and conditional distributions.**

# Joint distribution

- **Joint distribution of two discrete random variables** $X$ and $G$ = the probability that the random variables take on certain values, $x_i$ and $g_j$, simultaneously. It is denoted $f(x_i, g_j)$ or $\Pr(X = x_i, G = g_j)$.
- **Properties** of discrete joint pdf of $X$ and $G$:
  - Any joint event has a non-negative probability of occurring:

$$f(x_i, g_j) = \Pr(X = x_i, G = g_j) \geq 0 \text{ for } i = 1, 2, .., n$$

  - The sum of all joint probabilities is 1:

$$\sum_{i=1}^{N} \sum_{j=1}^{2} f\left(x_i, g_j\right) = \sum_{i=1}^{N} \sum_{j=1}^{2} \Pr(X = x_i; G = g_j) = 1$$

# Marginal distribution

- **Marginal distribution of a random variable** is the same as its probability density function (pdf):

$$\Pr(X = x_i) = \Pr(X = x_i, G = 0) + \Pr(X = x_i, G = 1)$$

same thing, alternative notation;

$$f(x_i) = \sum_{j=1}^{2} f(x_i, g_j)$$

- In bivariate analysis, we call the pdf the marginal distribution to more clearly distinguish it from the joint distribution.

# Joint distribution: numerical example

This table shows the joint income ($X = low$, $medium$ or $high$) distribution of male ($G = 0$) and female ($G = 1$) economics graduates:

| | $G = 0$ | $G = 1$ |
|---|---|---|
| $X = low$ | 0.12 | 0.13 |
| $X = medium$ | 0.27 | 0.23 |
| $X = high$ | 0.18 | 0.07 |

E.g. $Pr(X = low, G = 0) = 0.12$, $Pr(X = high, G = 1) = 0.07$.

Note that the joint probabilities sum to 1.

# Joint and marginal distributions: numerical example

It is straightforward to calculate the two marginal probability distributions by summing the joint probabilities:

| | $G = 0$ | $G = 1$ | $\Pr(X = x_i)$ |
|---|---|---|---|
| $X = low$ | 0.12 | 0.13 | 0.25 |
| $X = medium$ | 0.27 | 0.23 | 0.50 |
| $X = high$ | 0.18 | 0.07 | 0.25 |
| $\Pr(G = g_j)$ | 0.57 | 0.43 | |

The marginal distributions are $f(x_i) = \Pr(X = x_i)$ and $f(g_j) = \Pr(G = g_j)$. Note that $\sum_{i=1}^{3} f(x_i) = \sum_{j=1}^{2} f(g_j) = 1$

# Conditional distribution

**Conditional distribution** is the distribution of a random variable $X$ conditional on a specific value of another random variable $G$. It is defined as the ratio of the joint distribution over the marginal distribution:

$$f(X|G) = \frac{f(x_i, g_j)}{f(g_j)}$$

- For instance, the **conditional income distribution for male econ grads** is:

$$\Pr(X = x_i | G = 0) = \frac{\Pr(X = x_i, G = 0)}{\Pr(G = 0)}$$

- Similarly, the **conditional income distribution for female econ grads** is:

$$\Pr(X = x_i | G = 1) = \frac{\Pr(X = x_i, G = 1)}{\Pr(G = 1)}$$

# Conditional distribution: numerical example

We can calculate the **conditional income distributions for male and female econ graduates** in our example:

| | $G = 0$ | $G = 1$ | $\Pr(X = x_i)$ |
|---|---|---|---|
| $X = low$ | 0.12 | 0.13 | 0.25 |
| $X = medium$ | 0.27 | 0.23 | 0.50 |
| $X = high$ | 0.18 | 0.07 | 0.25 |
| $\Pr(G = g_j)$ | 0.57 | 0.43 | |

| | $\Pr(X = x_i \mid G = 0)$ | $\Pr(X = x_i \mid G = 1)$ |
|---|---|---|
| $X = low$ | $\frac{0.12}{0.57} \approx 0.21$ | $\frac{0.13}{0.43} \approx 0.30$ |
| $X = medium$ | $\frac{0.27}{0.57} \approx 0.47$ | $\frac{0.23}{0.43} \approx 0.53$ |
| $X = high$ | $\frac{0.18}{0.57} \approx 0.32$ | $\frac{0.07}{0.43} \approx 0.16$ |

# Independence

Two r.v.'s ($X$ and $G$) are **independent** if the distribution of each variable is unaffected by any particular outcome the other variable takes on[2]: the **joint distribution is equal to product of marginal distributions**

$$\Pr(X = x_i, G = g_j) = \Pr(X = x_i) * \Pr(G = g_j)$$

**Consequences** of independence:

- The conditional distribution is equal to marginal distribution

$$\Pr(X = x_i | G = g_j) = \Pr(X = x_i)$$

- The covariance and correlation between the random variables is zero: $Cov(X, G) = Corr(X, G) = 0$.

[2] Intuitively, "independent" means that the occurrence of one event makes it neither more nor less probable that the other event occurs- in the case of random variables, this has to hold for all events (=outcomes) captured by the random variables.

# Independence condition: example

$X$ and $G$ independent if joint distribution = product of marginal distributions:

| | $G=0$ | $G=1$ | $\Pr(X=x_i)$ |
|---|---|---|---|
| $X = low$ | 0.12 | 0.13 | 0.25 |
| $X = medium$ | 0.27 | 0.23 | 0.50 |
| $X = high$ | 0.18 | 0.07 | 0.25 |
| $\Pr(G=g_j)$ | 0.57 | 0.43 | |

$$\Pr(X = low, G = 0) \stackrel{?}{=} \Pr(X = low) * \Pr(G = 0)$$

$$0.12 \stackrel{?}{=} 0.25 * 0.57$$

$$0.12 \neq 0.14$$

Hence $X$ and $G$ are **not independent.**

# Independence consequence: example

Consequence: conditional distribution $\neq$ marginal distribution since $X$ and $G$ are dependent:

| | $G=0$ | $G=1$ | $\Pr(X=x_i)$ | $\Pr(X=x_i \vert G=0)$ |
|---|---|---|---|---|
| $X = low$ | 0.12 | 0.13 | 0.25 | $\frac{0.12}{0.57} \approx 0.21$ |
| $X = medium$ | 0.27 | 0.23 | 0.50 | $\frac{0.27}{0.57} \approx 0.47$ |
| $X = high$ | 0.18 | 0.07 | 0.25 | $\frac{0.18}{0.57} \approx 0.32$ |
| $\Pr(G=g_j)$ | 0.57 | 0.43 | | |

$$\text{e.g. } \Pr(X = low \vert G = 0) \stackrel{?}{=} \Pr(X = low)$$

$$0.21 \neq 0.25$$

# Conditional expectations

Just like we can summarize the pdf by its expectation $E(X)$, we can summarize the conditional distribution by the conditional expectation $E(X|G)$.

Recall the general formula of the **expectation**:

$$E(X) = \sum_{i=1}^{N} \Pr(X = x_i) * x_i$$

The **conditional expectation** $E(X|G)$ is the same, but using the **conditional instead of the marginal distribution**:

$$E(X|G) = \sum_{i=1}^{N} \Pr(X = x_i | G = g_j) * x_i$$

# Conditional expectations

The **conditional expectation of income for males**:

$$E(X|G=0) = \sum_{i=1}^{N} \Pr(X = x_i | G = 0) * x_i$$

The **conditional expectation of income for females**:

$$E(X|G=1) = \sum_{i=1}^{N} \Pr(X = x_i | G = 1) * x_i$$

If $X$ and $G$ are independent, it follows that the conditional expectations equal the (unconditional) expectation.

$$E(X|G=0) = E(X|G=1) = E(X)$$

# Conditional expectation: numerical example

Assume the low hourly wage is EUR 10, the medium is EUR 15 and the high is EUR 20: calculate the expected income conditional on G=0 (i.e. for males):

| $x_i$ | $\Pr(X = x_i \mid G = 0)$ | $\Pr(X = x_i \mid G = 0) * x_i$ |
|---|---|---|
| $X = 10$ | $\frac{0.12}{0.57} \approx 0.21$ | $0.21 * 10 = 2.10$ |
| $X = 15$ | $\frac{0.27}{0.57} \approx 0.47$ | $0.47 * 15 = 7.05$ |
| $X = 20$ | $\frac{0.18}{0.57} \approx 0.32$ | $0.32 * 20 = 6.40$ |
| | SUM | 15.55 |

Hence $E(X \mid G = 0) = 15.55$.

# Conditional expectations: rules of calculation

In addition to the already discussed rules for expectations, we have:

1. **When $X$ is multiplied by a constant with value** $c$, the conditional expectation is:

$$\begin{aligned} E(cX|G &= 0) = cE(X|G=0) \\ E(cX|G &= 1) = cE(X|G=1) \end{aligned} \qquad \text{(rule 1)}$$

   Click for proof

2. **When $X$ is multiplied by a function of $G$,** $h(G)$, the conditional expectation is (where $h(0)$ denotes the function $h(G)$ evaluated at $G=0$; and $h(1)$ evaluated at $G=1$):

$$\begin{aligned} E(h(G)X|G &= 0) = h(0)E(X|G=0) \\ E(h(G)X|G &= 1) = h(1)E(X|G=1) \end{aligned} \qquad \text{(rule 2)}$$

# Conditional expectations: examples of rules of calculation

$$
\begin{aligned}
E\left[5X|X=3\right] &= \\
&= 5E\left[X|X=3\right] \\
&= 5*3=15
\end{aligned}
$$

$$E\left[2Y+4XY+5X|X=2\right]=$$

$$
\begin{aligned}
&= E\left[2Y|X=2\right]+E\left[4XY|X=2\right]+E\left[5X|X=2\right] \\
&= 2E\left[Y|X=2\right]+4E\left[XY|X=2\right]+5E\left[X|X=2\right] \\
&= 2E\left[Y|X=2\right]+4*2E\left[Y|X=2\right]+10 \\
&= 10E\left[Y|X=2\right]+10
\end{aligned}
$$

# Covariance

The **covariance between X and G** is a measure of linear association between $X$ and $G$:

$$
\begin{aligned}
Cov(X,G) &= \sigma_{XG} = E\left(X - EX\right)\left(G - EG\right) \\
&= E(XG) - E(X)E(G) \text{ (for proof: see stats course)}
\end{aligned}
$$

- $Cov(X,G) > 0$ : X and G are positively linearly associated
- $Cov(X,G) < 0$ : X and G are negatively linearly associated
- $Cov(X,G) = 0$ : X and G are not *linearly* associated
- **When two variables are independently distributed, their covariance is 0**. But **the converse need not be true**: a covariance of 0 does not necessarily imply independence as the association may be non-linear!

## Covariance: rules of calculation

1. **Covariance between X and a constant c:**

$$Cov(X, c) = 0 \qquad \text{(rule 1)}$$

2. **Covariance between $aX$ and $bG$ where $a$ and $b$ are constants:**

$$Cov(aX, bG) = abCov(X, G) \qquad \text{(rule 2)}$$

3. **The covariance between $X$ and $X$:**

$$Cov(X, X) = E\,(X - EX)^2 = Var(X) \qquad \text{(rule 3)}$$

Click for proof of rules 1 & 2.

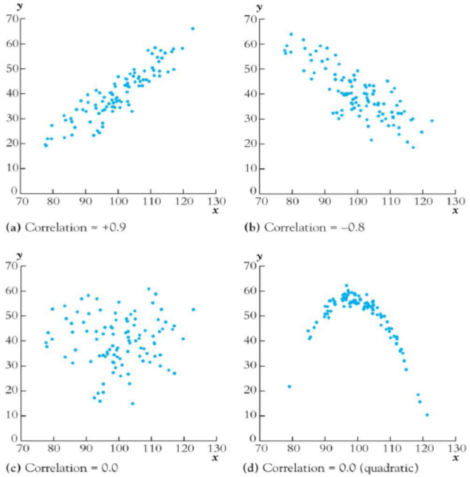
# Correlation

The **correlation between X and G** is a scale-invariant measure of linear association between $X$ and $G$:

$$Corr(X,G) = \rho_{XG} = \frac{Cov(X,G)}{\sqrt{Var(X)}\sqrt{Var(G)}} = \frac{Cov(X,G)}{sd(X)sd(G)}$$

- $-1 \leq Corr(X,G) \leq 1$
- $Corr(X,G) = 1$ : perfect linear positive relationship between X and G
- $Corr(X,G) = -1$ : perfect linear negative relationship between X and G
- $Corr(X,G) = 0$ : no linear relationship between X and G

# Correlation or covariance of 0 does not imply independence



# Rules of calculation: examples

$$
\begin{aligned}
Cov(-2X_1 + 3, 2X_2 + 4) &= \\
&= Cov(-2X_1, 2X_2) \\
&= -4Cov(X_1, X_2)
\end{aligned}
$$

$$Var(-2X_1 + 2X_2 + 4) =$$

$$
\begin{aligned}
&= Var(-2X_1 + 2X_2) \\
&= Var(-2X_1) + Var(2X_2) + 2Cov(-2X_1 + 2X_2) \\
&= 4VarX_1 + 4Var(X_2) - 8Cov(X_1 + X_2)
\end{aligned}
$$

# Rules of calculation: examples

$$
\begin{aligned}
Var(2X_1 + 4) &= \\
&= Var(2X_1) \\
&= 4Var(X_1)
\end{aligned}
$$

$$
\begin{aligned}
Corr(X_1, 2X_1 + 4) &= \\
&= Corr(X_1, 2X_1) \\
&= \frac{Cov(X_1, 2X_1)}{\sqrt{Var(X_1)}\sqrt{Var(2X_1)}} \\
&= \frac{2Cov(X_1, X_1)}{\sqrt{Var(X_1)}\sqrt{4Var(X_1)}} \\
&= \frac{2Var(X_1)}{\sqrt{Var(X_1)}2\sqrt{Var(X_1)}} = 1
\end{aligned}
$$

# How to answer this research question

**Do starting salaries of female economics graduates differ from those of male economics graduates?**

- We have summarized the population of r.v. X, as well as its relationship with r.v. G
- However, the **population** is always **unobserved**
- Hence we need to **use a sample to infer about the population**
  - List sample estimators of the population parameters (mean, variance, covariance, correlation) discussed so far
  - For proofs of the unbiasedness of these estimators: see Statistics course.
  - More on sampling distributions & inference next week.

# Random sampling

- We will assume a **random sample**: randomly selected subset of the population.
- As such, $x_1, x_2, x_3 ... x_n$ denotes a random sample of size $n$ from the population (which has population mean $\mu_X$ and variance $\sigma_X^2$), where $x_i$ =value of $x$ for the $i^{th}$ individual (or other entity, e.g. firm, household) sampled.
- We can then **use the sample to infer about the population** by using **estimators** = procedures for constructing an estimate of a population parameter.

## Sample estimators of population parameters

- The **sample mean** $\overline{x}$ is an unbiased estimator of the population mean $\mu_X$

$$\begin{aligned} \overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ E(\overline{x}) &= \mu_X \end{aligned}$$

The sample mean has a variance (across different samples: the sampling distribution) which is decreasing in the sample size $n$

$$Var(\overline{x}) = \frac{\sigma_X^2}{n}$$

# Sample estimators of population parameters

- The **sample variance** $s_X^2$ is an unbiased estimator of the population variance $\sigma_X^2$

$$\begin{aligned} s_X^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ E(s_X^2) &= \sigma_X^2 \end{aligned}$$

# Sample estimators of population parameters

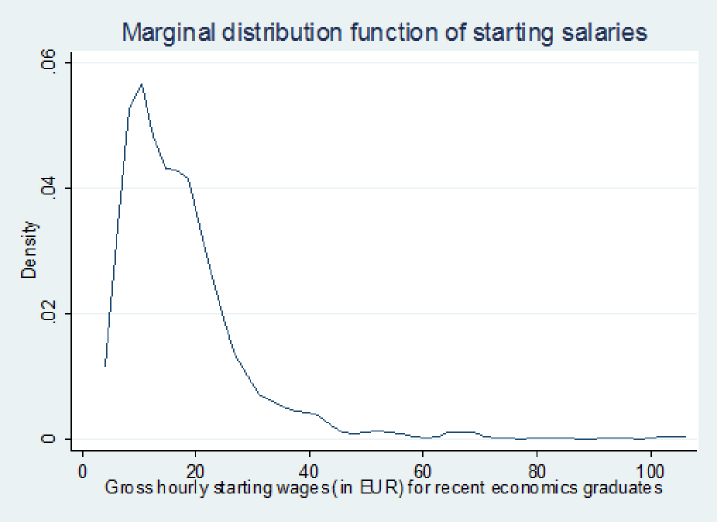
- **Sample covariance** between $X$ and $G$, $s_{XG}$, is unbiased estimator of population covariance $\sigma_{XG}$

$$
\begin{aligned}
s_{XG} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(g_j-\overline{g}) \\
E(s_{XG}) &= \sigma_{XG}
\end{aligned}
$$

- **Sample correlation coefficient** between $X$ and $G$, $r_{XG}$, is unbiased estimator of population correlation coefficient $\rho_{XG}$

$$
\begin{aligned}
r_{XG} &= \frac{s_{XG}}{\sqrt{s_X^2}\sqrt{s_G^2}} \\
E(r_{XG}) &= \rho_{XG}
\end{aligned}
$$

# Marginal distribution function



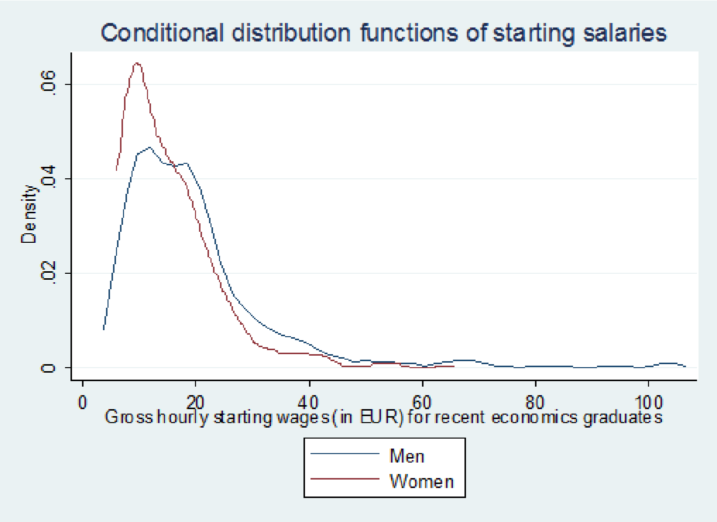
# Sample means

```
. describe starting_wage female

variable name   variable label
--------------------------------------------------------------------------------
starting_wage   Gross hourly starting wages (in EUR) for recent economics graduates
female          Gender dummy, 0 for males and 1 for females

. sum starting_wage female

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
starting_w~e |        629    17.73411    12.04007   6.041697   104.0769
      female |        629    .4737679    .4997088          0          1
```

$$\overline{x} = 17.73 \qquad \overline{g} = 0.47$$

Mean starting wage in this sample is EUR 17.73, and 47% of the sampled graduates are female.

# Conditional distribution functions (conditional on gender)



# Conditional sample means (for men and women)

```
. sum starting_wage if female==0
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| starting_w~e | 331 | 19.86311 | 14.07225 | 6.088898 | 104.0769 |

```
. sum starting_wage if female==1
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| starting_w~e | 298 | 15.36936 | 8.707941 | 6.041697 | 65.84978 |

Average hourly salary for a male econ graduate in this sample is EUR 19.86, versus EUR 15.37 for a female econ graduate.

# Sample (co)variances & sample correlation

```
. corr starting_wage female, cov
(obs=629)

             |  starti~e   female
-------------+------------------
starting_w~e |   144.963
      female |  -1.12213  .249709
```

```
. corr starting_wage female
(obs=629)

             |  starti~e   female
-------------+------------------
starting_w~e |    1.0000
      female |   -0.1865   1.0000
```

$s_X^2 = 144.96 \qquad s_{XG} = -1.12 \qquad s_G^2 = 0.25 \qquad r_{XG} = -0.19$

# Hypothesis test about sample correlation

```
. pwcorr starting_wage female, sig
```

| | starti~e | female |
|---|---|---|
| starting_w~e | 1.0000 | |
| female | -0.1865 | 1.0000 |
| | 0.0000 | |

$$
\begin{aligned}
H_0 &: \rho_{XG} = 0 \\
H_A &: \rho_{XG} \neq 0
\end{aligned}
$$

P-value is $0.00 < 0.05$, hence reject $H_0$ : starting wages are significantly correlated with gender.

# How to answer economic research questions

- This week, we just introduce the **terminology of the population regression model**
- In coming weeks, we will see how we can use a sample to infer on the population parameters from such multiple regression models.

# Terminology of the multiple regression model

**Population regression model**, in general form:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i$$

- $Y$: **dependent variable** (or explained variable, endogenous variable, regressand)
- $X_1, X_2, ..X_k$: **independent variables** (or explanatory variables, exogenous variables, regressors)
- $\varepsilon$: **error term** (or disturbance)- captures all other influences on $Y$ (i.e. other than the influences of $X_1, X_2, ..X_k$)

# Terminology of the multiple regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i$$

- Regression of $Y$ on $X$
- Models the effect of the $X's$ (independent) on $Y$ (dependent)
- The regression equation is linear in parameters $\beta_0, \beta_1, \beta_2..\beta_k$
- Subscript $i$ indexes individual population observations (1 through $N$):

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + ... + \beta_k X_{k1} + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + ... + \beta_k X_{k2} + \varepsilon_2 \\ & ... \\ Y_N &= \beta_0 + \beta_1 X_{1N} + \beta_2 X_{2N} + ... + \beta_k X_{kN} + \varepsilon_N \end{aligned}$$

# Terminology of the multiple regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i$$

- $\beta_0, \beta_1, \beta_2 .. \beta_k$ : the **population parameters** (or coefficients)
- $\beta_0$ : **intercept** (or constant) - expected value of Y when $X_1 = X_2 = ... = X_k = 0$.
- $\beta_1, \beta_2 .. \beta_k$ : **slope** parameters
- $\beta_1$ : expected effect of a one-unit change in $X_1$ on $Y$, holding constant $X_2, X_3, ..X_k$.
  - When can $\beta_1$ be interpreted as **a causal effect of $X_1$ on $Y$**, ceteris paribus? As we will now show in an example, when $E(\varepsilon_i | X_1, X_2 .. X_k) = 0$.

# Causality in the multiple regression model: example

$$W_i = \beta_0 + \beta_1 G_i + \beta_2 A_i + \varepsilon$$

This model informs us about how starting wages ($W$) relate to the gender ($G$) and age ($A$) of graduates.

We can then write the expected wage for a female graduate aged 25 as:

$$E\left(W|G=1, A=25\right) = E\left[\left(\beta_0 + \beta_1 G_i + \beta_2 A_i + \varepsilon\right)|\left(G=1, A=25\right)\right]$$

# Causality in the multiple regression model: example

Expected wage for a female graduate aged 25:

$$E\left(W|G=1,A=25\right)=E\left[\left(\beta_0+\beta_1 G_i+\beta_2 A_i+\varepsilon\right)|\left(G=1,A=25\right)\right]$$

We can work out the expectation (using rules of calculation):

$$\begin{aligned}
&= E\left[\left(\beta_0+\beta_1 G_i+\beta_2 A_i+\varepsilon\right)|\left(G=1,A=25\right)\right] \\
&= \left\{\begin{array}{c} E[\beta_0|(G=1,A=25)]+E[\beta_1 G_i|(G=1,A=25)] \\ +E[\beta_2 A_i|(G=1,A=25)]+E[\varepsilon|(G=1,A=25)] \end{array}\right\} \\
&= \beta_0+\beta_1 * 1+\beta_2 * 25+E[\varepsilon|(G=1,A=25)]
\end{aligned}$$

# Causality in the multiple regression model: example

Expected wage for female graduate aged 25:

$$E\left(W|G=1,A=25\right)=\beta_0+\beta_1+25\beta_2+E[\varepsilon|(G=1,A=25)]$$

Similarly, expected wage for female graduate aged 26:

$$E\left(W|G=1,A=26\right)=\beta_0+\beta_1+26\beta_2+E[\varepsilon|(G=1,A=26)]$$

The difference between the two is the **causal impact of being one year older on wages, ceteris paribus on gender** (since gender is held constant)-this is **given by** $\beta_2$ **ONLY IF** $E[\varepsilon|(G,A)]=0$:

$$\begin{aligned}\beta_2 &= E\left(W|G=1,A=25\right)-E\left(W|G=1,A=26\right)\\ \textit{iff}\;\; E[\varepsilon|(G,A)] &= 0\end{aligned}$$

In general, $E[\varepsilon|X_1,X_2..X_k]=0$ is known as the **conditional mean assumption** and we will get back to it in later weeks.

# Things to do for your project paper this week (I)

- **Look at the different cross-sectional datasets** available and see which one(s) you find interesting.
- When you form a paper group in the first tutorial, also **indicate your preferred dataset** to your tutorial teacher.
  - If you do not indicate a preference in the first tutorial, a dataset will be assigned to you.
- Find out **what information your dataset contains**:
  - Examine what variables are in the dataset and what the main unit of observation is (individual, firm, country, ...).
  - Construct summary statistics for a number of different variables that interest you (**univariate analysis**): means, standard deviations, minimum & maximum.

# Things to do for your project paper this week (II)

- Choose a **dependent variable** for your regression analysis: you want to understand variation in this variable (e.g. wage, economic growth, sales, crime rate, student test score, human trafficking, sympathy for extreme right-wing parties, ..).
- Choose the **main independent variable** on which the analysis is focused: you want to examine the economic effect of this variable on the dependent variable.
- **Formulate research question** related to these two variables (e.g. what is the impact of workers' physical attractiveness on their wages?)
- Then, start exploring the **bivariate relationship** between these two variables: correlation + simple OLS.
- See Appendix 1 of course manual for details on the final lay-out of the paper.

# A note on appendix slides

Note: you are only very exceptionally required to perform proofs on the exam for this course, meaning it is not necessary for passing or even for obtaining a good grade. These proofs are included here to make it easier to understand where the various rules of calculation come from.

# Expected value: Caution!

In general,

$$E\left[g(X)\right] \neq g\left[E(X)\right]$$

For example, take $g(X) = X^2$:

$$EX^2 = E(X^2) \neq \left[E(X)\right]^2$$

# Proof of alternative variance formula

$$
\begin{aligned}
Var(X) &= E\left(X - EX\right)^2 && \text{(proof)} \\
&= E\left(X - \mu_X\right)^2 \\
&= E\left[\left(X - \mu_X\right)\left(X - \mu_X\right)\right] \\
&= E(X^2 - 2X\mu_X - \mu_X^2) \\
&= E(X^2) - 2E(X\mu_X) + E(\mu_X^2) \\
&= E(X^2) - 2\mu_X E(X) + \mu_X^2 \\
&= E(X^2) - 2\mu_X\mu_X + \mu_X^2 \\
&= E(X^2) - 2\mu_X^2 + \mu_X^2 \\
&= E\left(X^2\right) - \mu_X^2
\end{aligned}
$$

Click here to go back to the main text.

# Proof of variance rules 1 & 2

**Proof of rule 1**:

$$Var(c) = E\left(c^2\right) - c^2 = c^2 - c^2 = 0 \qquad \text{(proof)}$$

**Proof of rule 2**:

$$\begin{aligned} Var(X+c) &= E\left(X + c - E[X+c]\right)^2 && \text{(proof)} \\ &= E\left(X + c - E[X] - c\right)^2 \\ &= E\left(X - E[X]\right)^2 \\ &= Var(X) \end{aligned}$$

Click here to go back to the main text.

# Proof of variance rule 3

$$
\begin{aligned}
Var(cX) &= E\left((cX)^2\right) - (c\mu_X)^2 &\qquad \text{(proof)}\\
&= E\left(c^2X^2\right) - c^2\mu_X^2 \\
&= c^2E\left(X^2\right) - c^2\mu_X^2 \\
&= c^2\left[E\left(X^2\right) - \mu_X^2\right] \\
&= c^2 Var(X)
\end{aligned}
$$

Click here to go back to the main text.

# Proof of variance rules 4 & 5

**Proof of rule 5**:

$$
\begin{aligned}
Var(X_1 + X_2) &= E\left([X_1 + X_2] - E[X_1 + X_2]\right)^2 \qquad \text{(proof)} \\
&= E\left([X_1 - EX_1] + [X_2 - EX_2]\right)^2 \\
&= E\left([X_1 - EX_1] + [X_2 - EX_2]\right)^2 \\
&= E\begin{pmatrix} [X_1 - EX_1]^2 + [X_2 - EX_2]^2 \\ +2[X_1 - EX_1][X_2 - EX_2] \end{pmatrix} \\
&= E[X_1 - EX_1]^2 + E[X_2 - EX_2]^2 \\
&\quad +2E[X_1 - EX_1][X_2 - EX_2] \\
&= Var(X_1) + Var(X_2) + 2Cov(X_1, X_2)
\end{aligned}
$$

If $X_1$ and $X_2$ are independent, then $Cov(X_1, X_2) = 0$ and the formula simplifies to rule 4.

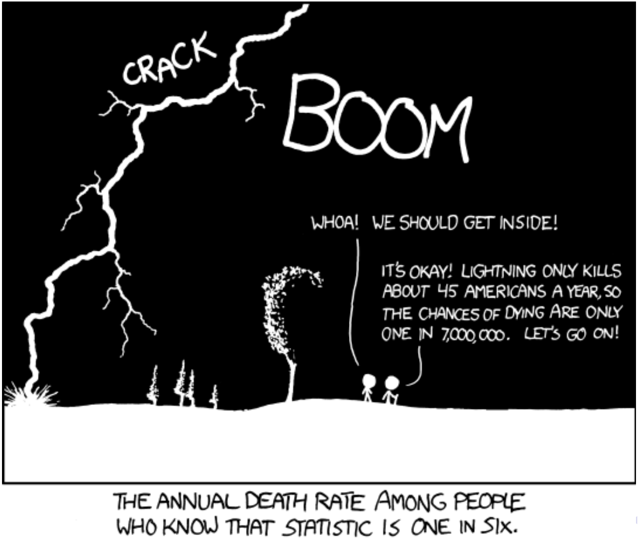
Click here to go back to the main text.

# Proof of conditional expectation rule 1

$$
\begin{aligned}
E(cX|G &= 0) = \sum_{i=1}^{N} \Pr(X = x_i|G = 0) * cx_i \quad \text{(proof)} \\
&= c \sum_{i=1}^{N} \Pr(X = x_i|G = 0) * x_i \\
&= cE(X|G = 0)
\end{aligned}
$$

Click here to go back to the main text.

# An example of a conditional expected value (xkcd.com)



# Proof of covariance rules 1 & 2

**Rule 1**:

$$
\begin{aligned}
Cov(X, c) &= E(X - \mu_X)(c - \mu_c) && \text{(proof)} \\
&= E(X - \mu_X)(c - c) \\
&= 0
\end{aligned}
$$

**Rule 2:**

$$
\begin{aligned}
Cov(aX, bG) &= E(aX - E(aX))(bG - E(bG)) && \text{(proof)} \\
&= E(aX - aE(X))(bG - bE(G)) \\
&= E(aX - a\mu_X)(bG - b\mu_G) \\
&= abE(X - \mu_X)(G - \mu_G) \\
&= abCov(X, G)
\end{aligned}
$$

Click here to go back to the main text.