# POLS2044 WEEK 10
## Regression modelling and interpreting model results

Australian National University
School of Politics & International Relations
Dr. Richard Frank
https://richardwfrank.com/research_design_2022

In Week 10 of POLS2044 we will be continuing our focus on regression modelling. We have spent time on various ways of (1) describing and developing an understanding of our data—what is the central tendency, how much observed variance is there, what is the most common value, what outliers exist—and (2) looking at relationships between two variables. This week we reinforce Week 9's initial discussion of regression's assumptions and the bivariate regression equation.

The whole reason we are doing all this data analysis, of course, is because we have a theoretically informed expectation about the world that we are trying to evaluate with the best data available. Oh, and because these methods (and probability theory as we discussed last week) are at the heart of much of our daily lives and knowledge of them is increasingly necessary to interpret media summaries of current research (e.g., is red wine and coffee healthy or unhealthy?) as well as increasingly useful for post-undergraduate studies and careers.

This week I have two main goals. First, I want students to understand how and why we move from bivariate to multivariate regression. I also want students to get more comfortable talking about the elements of regression modelling, what they mean substantively, and what assumptions underly them. Second, I want students to get practice interpreting regression results.

---

## Reading notes and questions

There are two readings for this week. Both start with terms we have seen in previous weeks. They then move into multiple regression, which we will be talking about more for the remainder of the semester.

**Wheelan, Charles. 2013. "Chapter 11: Regression Analysis: The Miracle Elixir," in *Naked Statistics: Stripping the Dread from the Data*. London: W.W. Norton: 185-211.**

This chapter is useful to read first as it focuses on the intuition behind both bivariate and multivariate regression. It also connects the various issues we have talked about previously (e.g., sample size, intercepts, dependent and independent variables, sign, size, and significance of coefficients) to several empirical examples (e.g., job stress, height and weight, and gender pay gaps).

Several things to think about while reading this chapter:

1. Can you think of any political outcome that is only correlated with one explanatory factor? Or are there always additional factors that are not of primary interest but have secondary importance?

2. How and why are the sign, size, and significance of coefficients important to understand when analysing regression results?
3. What happens to the tails of the t-distribution as the sample size gets progressively larger?
4. What are confidence intervals (we also discussed them in Week 9) and what are they telling us?
5. What is a useful rule of thumb about statistical significance when the coefficient is at least twice the size of the standard error?
6. What is the $R^2$ and what can it (and cannot) tell us about the model's explained variance?

**Long, Abby. 2016. "10 Things to Know About Reading a Regression Table." Berkeley: EGAP. Available at: [https://egap.org/resource/10-things-to-know-about-reading-a-regression-table/](https://egap.org/resource/10-things-to-know-about-reading-a-regression-table/).**

If the Wheelan chapter focused on the deductive intuition behind regression, this online article takes a more inductive approach—if you are faced with a regression table in an article you are reading, what are the important takeaways to take from it?

One important point this article makes is that there are myriad names given to the same concept (e.g., right-hand side, response, predictor, input, and causal variables).

7. Are there any terms used in this (or other readings) that you do not understand? Can you infer meaning from the text or from an online search? Do you have somewhere where you collect definitions of terms used in this or other courses?
8. What are the main purposes of regression? While this class focuses on causal relationships, do you think the other three purposes Long discusses are also interesting/valuable to your research or understanding?
9. What is "Anscombe's quartet" and how does it connect back to our descriptive statistics week?
10. What is "researcher degrees of freedom"? Have you thought about this risk when deciding whether or not to be convinced by an article's findings?
11. What are the potential biases we should be aware of?
12. What new terms are used in this article? Do you feel confident you could describe them to others?
13. Both readings talk about extensions of regression models that are necessary when the dependent variable is not continuous (e.g., GDP is continuous, but war is not traditionally measured as such). Is the dependent variable you plan on using in your final paper continuous or not?

It is fine to use OLS regression in this class even if your dependent variable is not continuous, but your results would be easier to interpret if they are.

## LECTURE PART 1: Introduction

### Probability's key properties

All outcomes have a probability ranging from 0 to 1.
The sum of all possible outcomes must be exactly 1.
If (and only if) two outcomes are independent, then the probability of those events both occurring is equal to the product of them individually.
The chance of either of two outcomes happening is the sum of their probabilities if the options are mutually exclusive.
If the events are not mutually exclusive, the probability of getting A or B consists of the sum of their individual probabilities minus the probability of both events happening.

### Probability pitfalls

Assuming events are independent when they are not (e.g., rain today and tomorrow).
Assuming events are not independent when they are (e.g., hot streaks).
Clusters do happen (e.g. getting struck by lightening).
There is often reversion to the mean (e.g. doing well on an exam).
Moving from aggregate statistics to predicting individual behaviour (e.g., profiling).
Garbage in, garbage out (e.g., data quality).
Analytical tools are moving faster than our knowledge of what to do with results (e.g. predictive AI).

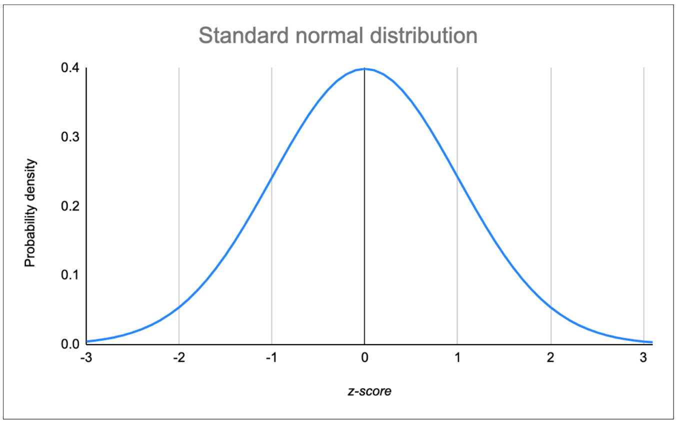
### Central limit theorem

Sample size has to be large (say >30).

The sample mean will be distributed roughly as a normal distribution around the population mean.

The sample standard deviation will equal the population standard deviation over the square root of the number of sample observations.

### The standard normal distribution




Source: https://www.scribbr.com/statistics/standard-normal-distribution/

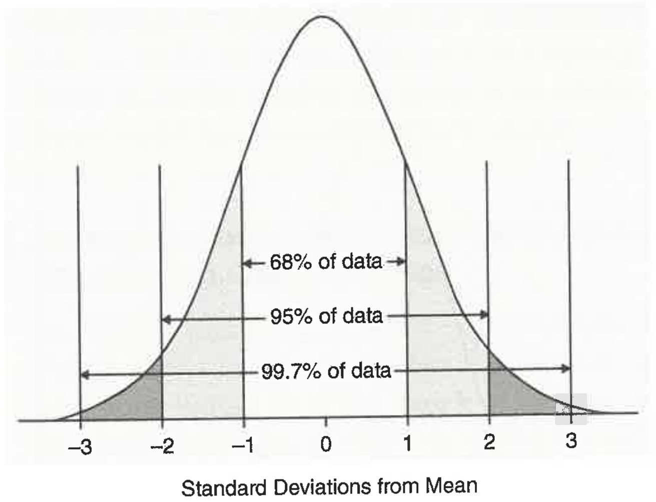



**Figure 7.2** The 68–95–99 rule

Source: Kellstedt and Whitten (2018: 150)

**Important to note!**

A distribution of sample values is the frequency distribution, which does not have to be normal.

However, because of the central limit theorem, the repeated sampling distribution mean will be naturally distributed, even if the underlying frequency is not.

**Probability takeaways**

Probabilities involve uncertainty.
Political scientists need estimates of uncertainty as we have sample data instead of population data.
Probability theory comes with important assumptions, strengths, and weaknesses.
It is largely relevant to us when determining statistical significance.

**Why conduct hypothesis testing?**

It forces us to clearly link our theory to its real world implications.
It forces us to think about the null hypothesis.
It forces us to frame our implications in a falsifiable manner.

**Which hypothesis test do we choose?**

| | | **Independent variable type** | |
|---|---|---|---|
| | | ***Categorical*** | ***Continuous*** |
| **Dependent variable type** | ***Categorical*** | Tabular (goodness of fit) analysis | Logit/probit |
| | ***Continuous*** | Difference of means test or regression | Pearson's correlation coefficient or regression |

**What do these tests have in common?**

They use p-values in their hypothesis tests.
These p-values range from 0 to 1.
They include a null hypothesis.

They do not tell us that the relationship is causal.
They do not tell us how strong the relationship is.
They do not tell us anything about the quality of our measures.

**Hypothesis testing important takeaways**

There are different types of hypothesis tests for different types of data and hypotheses.
They all involve some form of significance test.
These significance tests rely on probabilities related to the distribution of the sample means.

**Today's motivating questions**

Why and how do we compare different samples?
Why and how do we run a multiple regression?
Why and how do you interpret regression results?

---

**LECTURE PART 2: Difference of means test**

**Motivating question**

Are the average values of two different measures significantly different from each other?
Are men statistically significantly taller than women?
Do plumbers earn more than lawyers?
Do democracies have a higher average GDP than non-democracies?

**Two-sample t-test for difference of means**

**Null hypothesis**—the average values (means) are drawn from the same underlying distribution.
**Alternate hypothesis**—the average values (means) are not drawn from the same underlying distribution.

The test statistic for this difference of means test is the **t-statistic** because it follows the t-distribution.

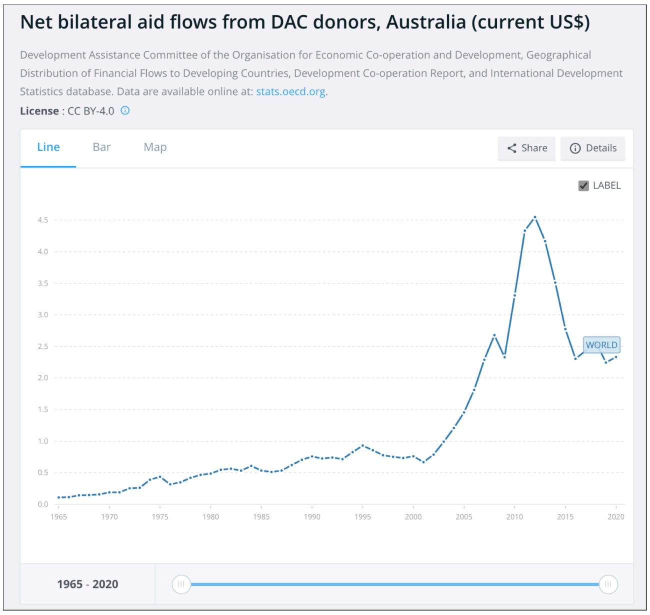
$$t = \frac{\overline{Y_1} - \overline{Y_2}}{se(\overline{Y_1} - \overline{Y_2})}$$

Where   is the mean value of variable y1 and se is the standard error of the difference of the two means.

$$se(\overline{Y}_1 - \overline{Y}_2) = \sqrt{\left(\frac{(n_1-1)s_1^2 + (n_1-1)s_2^2}{n_1+n_2-2}\right)} + \sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

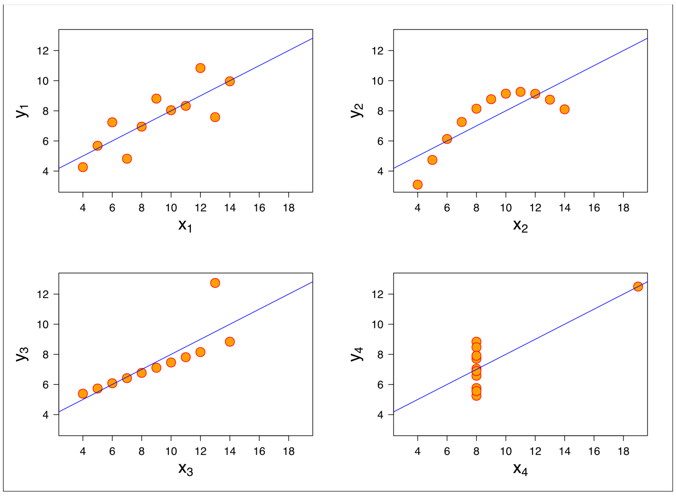
Where: n1 is the sample size of Ybar 1  and n2 is the sample size of  Ybar 2 and s1 and s2   are the sample variances.

**Example: Does Australia give more aid to Asia-Pacific countries?**

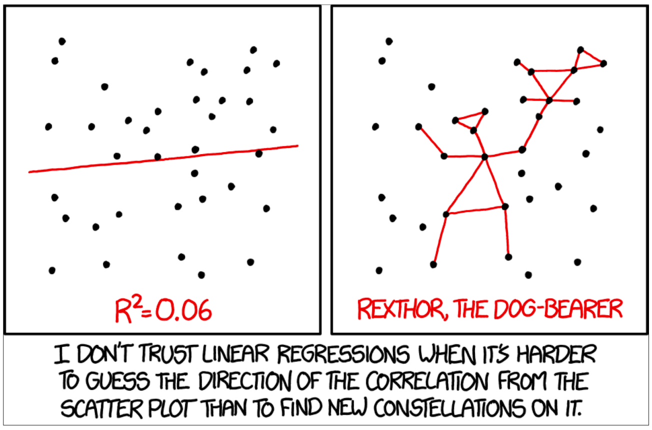



2020 data from the World Bank
([https://data.worldbank.org/indicator/DC.DAC.AUSL.CD](https://data.worldbank.org/indicator/DC.DAC.AUSL.CD))

**Anscombe's quartet**




Almost identical descriptive statistics but very different underlying value distributions.




Source: https://xkcd.com/1725/

**Exercise—calculating difference of means in Excel.**

**T-distribution critical values**

**Important takeaways**

This type of hypothesis testing is useful when you have two (or more) identifiable groups that may systematically differ in their mean values of the outcome we are interested in.
The math is less important than the intuition.
There are a number of other test statistics that you can use depending on the nature and distribution of your variables as well as the hypothesis you are interested in testing.

---

## LECTURE PART 3: Why run a regression?

**Why run a regression?**

What if we are interested not just if there is a statistically significant difference in a sample (goodness of fit) or pairs of samples (difference of means test)?
Rather we want a more complex understanding of the directionality and significance in the relationship between an X and Y than a simple correlation can tell us?
Or perhaps we want to predict our outcome as we vary values of our independent variables?

**Ceterus paribus assumption**

Latin for "other things equal."
Also short for "all other things being equal."
Regression helps us control for other factors to better isolate the effect of the variable we care about.

**Taken to the extreme**

Political Behavior
https://doi.org/10.1007/s11109-021-09720-y

ORIGINAL PAPER

## Correlates of Voter Turnout

Richard W. Frank[1] · Ferran Martínez i Coma[2]

Accepted: 11 May 2021


**Abstract**
Despite decades of research, there is no consensus as to the core correlates of national-level voter turnout. We argue that this is, in part, due to the lack of comprehensive, systematic empirical analysis. This paper conducts such an analysis. We identify 44 articles on turnout from 1986 to 2017. These articles include over 127 potential predictors of voter turnout, and we collect data on seventy of these variables. Using extreme bounds analysis, we run over 15 million regressions to determine which of these 70 variables are robustly associated with voter turnout in 579 elections in 80 democracies from 1945 to 2014. Overall, 22 variables are robustly associated with voter turnout, including compulsory voting, concurrent elections, competitive elections, inflation, previous turnout, and economic globalization.

**Keywords** Elections · Turnout · Extreme bounds analysis · Meta-analysis

**Introduction**

A common challenge in the study of comparative politics is balancing theoretical and empirical comprehensiveness with substantive importance. Consider voter turnout. If we ask what the most statistically significant and substantively important predictors of national-level voter turnout in democratic elections are, even after more than 50 years of comparative voter turnout research, there are few certainties beyond the fact that compulsory voting increases turnout. For example, several studies including Radcliff and Davis (2000) find larger district magnitudes increase turnout while others like Tavits (2008) find either no significant relationship or even a negative one (Fumagalli & Narciso, 2012).

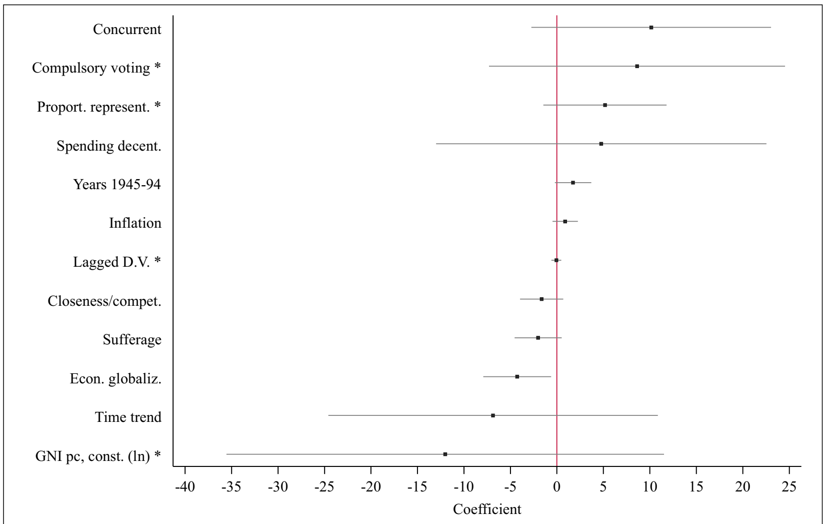
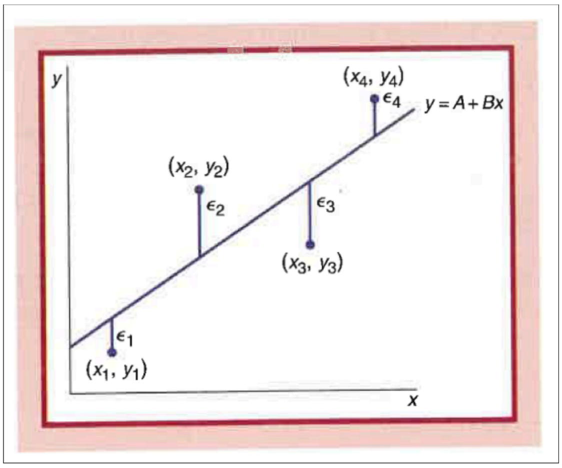



**Fig. 2** Robust predictors of voter turnout, mean coefficients and extreme bounds. *Note* Results from fixed effects models reported in Table 5. Table A7 includes complete results. *Identifies core variables in all fixed effects models. The only core variable found to be (Sala-i-Martin) robust is proportional representation. Population is a core variable but not included in figure due to the disproportionate size of its extreme bounds

## Estimating the relationship between X and Y



**The work horse model**

Y=alpha + Beta*X + error terms

Where:

*Y* is the <u>outcome</u> you are trying to explain.
*X* is the main <u>explanatory</u> variable.
(alpha) is the value of Y when X=0.
(beta) is the estimated relationship between X and Y.
is the systematic error.
is the random error.

It can be shown that the least-squares estimators of $\alpha$ and $\beta$, which we call $\hat{\alpha}$ and $\hat{\beta}$, are given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\,\overline{x}$$

where

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{and} \quad \overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

**Let us revisit our industrial strikes example and run a multiple regression in Excel.**

**Regression takeaways**

Ordinary least squares regression is about fitting a line that minimises the (squared) distance between sample values and the line.
A basic regression provides us with two important estimates:
(1) the slope of the line summarising the relationship between x and y
(2) the expected value of y when x=0
Multiple regression enables us to control for other factors that might understate or overstate our X-Y relationship if we do not include them.
The intuition helps us understand what it can and cannot tell us.

**The three S's**

Size, sign, and significance

**Example of regression output from Long (2016)**

**Confidence intervals**

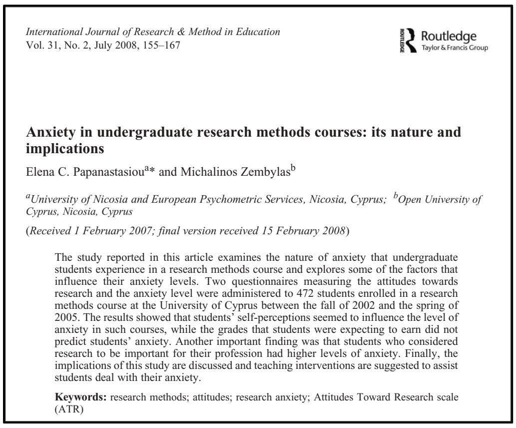
$$\sigma_{\bar{Y}} = \frac{sd_Y}{\sqrt{n}}$$

The lower bound of the confidence interval is the mean minus two standard errors of the mean,

The upper bound is the mean plus two standard errors of the mean.

What would happen to the confidence intervals if the sample was cut in half?

**Another example: methods anxiety**


*International Journal of Research & Method in Education*
Vol. 31, No. 2, July 2008, 155–167

**Anxiety in undergraduate research methods courses: its nature and implications**

Elena C. Papanastasiou[a]* and Michalinos Zembylas[b]

[a]*University of Nicosia and European Psychometric Services, Nicosia, Cyprus;* [b]*Open University of Cyprus, Nicosia, Cyprus*

(*Received 1 February 2007; final version received 15 February 2008*)

The study reported in this article examines the nature of anxiety that undergraduate students experience in a research methods course and explores some of the factors that influence their anxiety levels. Two questionnaires measuring the attitudes towards research and the anxiety level were administered to 472 students enrolled in a research methods course at the University of Cyprus between the fall of 2002 and the spring of 2005. The results showed that students' self-perceptions seemed to influence the level of anxiety in such courses, while the grades that students were expecting to earn did not predict students' anxiety. Another important finding was that students who considered research to be important for their profession had higher levels of anxiety. Finally, the implications of this study are discussed and teaching interventions are suggested to assist students deal with their anxiety.

**Keywords:** research methods; attitudes; research anxiety; Attitudes Toward Research scale (ATR)

The literature suggests that statistics anxiety negatively affects course performance. (Zeidner 1991; Onwueg- buzie and Seaman 1995; Zanakis and Valenza 1997)

> Finally, there are several implications from previous research on statistics anxiety in relation to teaching and learning strategies that can alleviate anxiety. For example, Gal and Ginsburg (1994) emphasized that in order to make statistics less threatening and more effective, attention should be focused on students' beliefs and attitudes. Other researchers report specific strategies that help reduce students' anxiety levels; these strategies include: encouraging students, using humour, teaching gimmicks, helping students to understand the course objectives, administering open book exams, using performance assessments, using effective teaching style, provide extensive feedback, addressing ways to relieve anxiety, applying statistics to real world examples and assigning students to work in groups (Onwuegbuzie and Wilson 2003).

Source: Papnastasiou and Zembylas (2008: 158)

**Research questions**

**Research methods**

The research questions that are examined in this study are the following:

(1) What are the levels of anxiety experienced by undergraduate students enrolled in a research methods class?
(2) What is the relationship between research methods anxiety and other anxiety types and attitudes towards research?
(3) What variables can explain and predict the anxiety levels of these students?
(4) How does the students' anxiety affect their achievement in the course?

**Dependent variable: An anxiety scale**

"AMAS-C is a 49-item self-report measure designed to assess chronic, manifest anxiety in the college student population. The students had to respond to the AMAS-C on a nominal true/false scale. The construct validity of the scale that was obtained through a factor analysis revealed four subscales: worry anxiety, physiological anxiety, test anxiety and social anxiety." -Papnastasiou and Zembylas (2008: 160)

Note the scale is reversed in some models so that higher values suggest lower anxiety levels.

**Independent variables: student survey questions**

32 questions measured on a seven-point Likert Scale ranging from 1 (strongly disagree) to 7 (strongly agree). These questions are combined into five sub scales:

Usefulness of research to students' profession
Research anxiety
Positive attitudes to research
Relevance of research to students' personal lives
Research difficulty

Papnastasiou and Zembylas (2008: 160)

**Sample**

472 students enrolled in an undergraduate methods course for education students at the University of Cyprus from 2002 to 2005.

What population do you think this sample is part of?

**Results**

Table 3. Predicting anxiety from the other ATR scales.

| Subscales | Unstandardized coefficients β | Std. error | Standardized coefficients β | t | Sig. |
|---|---|---|---|---|---|
| Constant | 0.033 | 0.374 | | 0.087 | 0.931 |
| Usefulness for the profession | −0.333 | 0.089 | −0.255 | −3.755 | 0.000 |
| Positive attitudes | 0.570 | 0.073 | 0.472 | 7.805 | 0.000 |
| Relevance to life | 0.118 | 0.081 | 0.086 | 1.465 | 0.144 |
| Research difficulty | 0.477 | 0.054 | 0.419 | 8.852 | 0.000 |

F = 62.258, p = 0.000.
R2=.452

Separate analyses found gender differences in anxiety but not in difficulty.

Source: Papanastasiou and Zembylas (2008: 162)

**Grade results**

Table 6. Predicting the research methods course grade.

| | Unstandardized coefficients β | Std. error | Standardized coefficients β | t | Sig. |
|---|---|---|---|---|---|
| Constant | 6.840 | 0.451 | | 15.183 | 0.000 |
| Usefulness for the profession | 0.387 | 0.109 | 0.318 | 3.546 | 0.000 |
| Anxiety | −0.177 | 0.070 | −0.190 | −2.520 | 0.012 |
| Positive attitudes | 0.075 | 0.096 | 0.066 | 0.774 | 0.439 |
| Relevance to life | −0.128 | 0.097 | −0.101 | −1.319 | 0.188 |
| Research difficulty | 0.066 | 0.073 | 0.062 | 0.901 | 0.368 |

F = 6.56, p = 0.000
R2 = 0.102.

Source: Papanastasiou and Zembylas (2008: 163)

**Why standardise coefficients?**

Most of the time our independent variables use different measurement units.
This makes direct comparison of regression coefficients difficult.
Standardising the coefficients puts the coefficients on the same scale, which aids comparability.
This comes at a cost of easily understanding one unit change in the independent variable.

**Standardising coefficients**

$$\beta_1^* = \hat{\beta}_1 \left( \frac{sd_{x_1}}{sd_y} \right)$$

Where:

$\hat{\beta}_1$ equals the unstandardised coefficient for variable $x_1$.

$sd_{x_1}$ equals the sample standard deviation for variable $x_1$.

$sd_y$ equals the sample standard deviation for the outcome variable y.

**Side note**: You can also standardise the variable instead of the coefficient.

Source: Gujarati, Damodar N. 2003. *Basic Econometrics*. 4th edition. New York: McGraw Hill: 175.

**Today's motivating questions**

Why and how do we compare different samples?
Why and how do we run a multiple regression?
Why and how do you interpret regression results?

## TUTORIAL ACTIVITIES

Over the last twelve weeks we have covered almost all the steps in the scientific method we discussed back in week 1. The last few weeks have built your familiarity with several descriptive and inferential statistical methods culminating with bivariate regression last week.

In this week's tutorial, you are going to have the opportunity to apply the two techniques discussed and demonstrated in this week's lecture—a difference of means test and a multiple regression. The goal is to understand when each technique is of use, how it is used, and what it can and cannot help us with when evaluating causal arguments and their observable hypotheses.

Please divide yourselves into groups of ~4. The goal is for every student to run these statistical methods themselves on their own computers; however, you should work together as teams to help each other work through the tasks below and discuss your results and answer any logistical questions. Of course, the tutor is always there if you have any more specific questions or are not sure how to start to answer the prompts below.

The Excel spreadsheet (week_10_tutorial_data.xlsx) that you need for this week's activities has been uploaded to Wattle under POLS2044/Week 10/tutorial. The first sheet in this spreadsheet ("data description") provides information about the dataset we are using and a link to the codebook for this dataset. I recommend that you download the codebook and spend a minute or two looking at the variables' names and descriptions.

**Part 1: A difference of means test**

For this first activity, please go to the second tab ("difference of means") of the tutorial dataset. You will see in Column A the names of 156 countries. Columns B-K include the variables described in the "data description" tab with two notable additions. Columns D and E break up the GDP per capita data in Column C into GDP values for EU (Column D) and non-EU (Column E) countries. I am curious whether EU countries have on average a higher GDP per capita than non-EU countries.

1. Please write (a) a suitable hypothesis and (b) a null hypothesis for this difference of means test.

*Hint*—look at the H0 and H1 outlined in the in-class Task 2 and see if they can be reworked for your purposes here.

Next, run a difference of means test for Columns D and E. A screenshot on the right of the table (starting in Column L) shows the type of t-test you want to run as well as suggestions for the values to enter. Does the hypothesized mean difference make sense? What would it mean to change the alpha value?

2. Do your results support rejecting the null hypothesis in favour of your hypothesis or do you fail to reject the null hypothesis? How do you reach this conclusion?

**Part 2: A multiple regression**

So far, I have asked you to follow some rather clear procedures. In this part of the tutorial, you will have the opportunity to (1) choose <u>your</u> own outcome of interest, (2) choose an explanatory variable that may affect your outcome, (3) choose some potential control variables, (4) run a regression, and (5) explore the results.

Please go to the tab labelled "multiple regression." You will see the eight variables you learned about above.

3. Choose one variable as your outcome variable (Y).
4. Choose one variable as your explanatory variable (X).
5. Choose three variables as your control variables (the vector **Z**).
6. Write a hypothesis and a null hypothesis for the expected relationship between your X and Y.

There are no right or wrong choices here. Some students in your group may choose different variables than you. That is great! The goal here is to see the list of variables and see if you can start to tell yourself a halfway decent story about how one factor affects another. If you really are stuck for a way to start, do ask a tutor.

Once you have your variables chosen, run a multiple regression as I did in class. There is a screenshot of a regression option (that makes no sense) to the right of the table that shows you the different fields that need entering. Remember the readings for this week and the lecture. You should initially focus on statistical significance, sign, and size of the coefficients.

7. Evaluate your #6 hypotheses. Do you reject or fail to reject the null? Why or why not?

8. Try and explain as much of your regression results as you can to the other members of your group (or to the rest of your tutorial).
9. What is the easiest to explain and what is the hardest?
10. Do you think you would get similar results if you had data for these countries in 2021 or 1991 (or 1881)? Why or why not?
11. What do you think would happen to your results if you ran your model on only the first 30 observations in this dataset? Or if you deleted clear outliers/extreme values in your data?