POLS2044 WEEK 12
Summing up

Australian National University
School of Politics & International Relations
Dr. Richard Frank

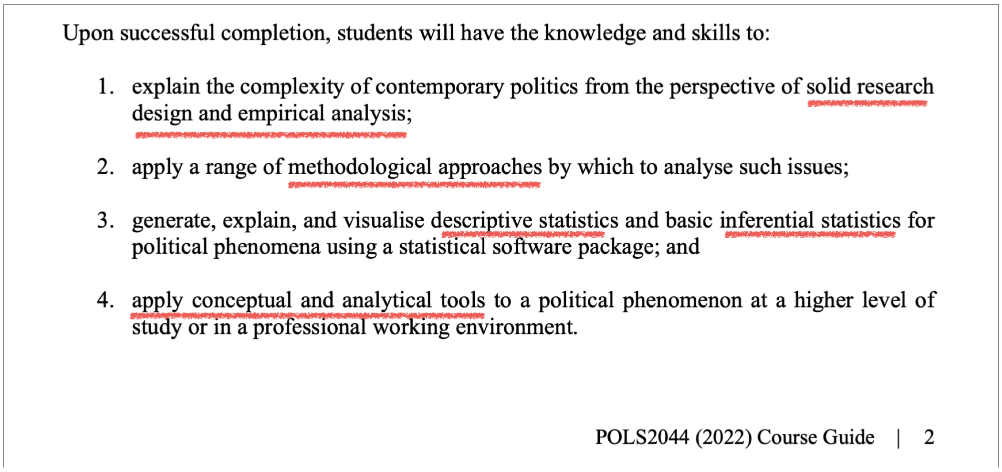
In Week 12 of POLS2044 we will be wrapping up this term by both reflecting on what we have learned this term.

Reading notes and questions

There are no readings for this week. Please use this as an opportunity to read through the textbook and lecture notes again.

LECTURE PART 1: Semester recap

2024 course outline



Upon successful completion, students will have the knowledge and skills to:

1. explain the complexity of contemporary politics from the perspective of solid research design and empirical analysis;
2. apply a range of methodological approaches by which to analyse such issues;
3. generate, explain, and visualise descriptive statistics and basic inferential statistics for political phenomena using a statistical software package; and
4. apply conceptual and analytical tools to a political phenomenon at a higher level of study or in a professional working environment.

POLS2044 (2022) Course Guide | 2

Week 1: Scientific method
Week 2: Causal theorising
Week 3: Research design
Week 4: Concepts and measurement
Week 5: Surveys and sampling
Week 6: Descriptive inference & statistics
Week 7: Probability & statistical inference
Week 8: Bivariate hypothesis testing
Week 9: Bivariate regression
Week 10: Multivariate regression
Week 11: Regression pitfalls

The broad applicability of the scientific method

KKV's (1994) characteristics of scientific research

1. The goal is causal inference.
2. The procedures are public.
3. The conclusions are uncertain.
4. The content is the method not the subject matter.

Goal 1: Help you consume research

Goal 2: help consume information

Goal 3: help you produce research

The scientific method

KKV's (1994) characteristics of scientific research

The goal is causal inference.
The procedures are public.
The conclusions are uncertain.
The content is the method not the subject matter.

Often the scaffolding of intellectual buildings are taken down after being built.

Developing new theoretical arguments

Offer an answer to an interesting research question.
Solve an interesting puzzle.
Identify interesting variation (across time or space)
Move from a specific event to more general theories
Drop the proper nouns
Use a new Y
Use a new X
Add a new Z
Use the literature
Make sure the theory can be disproven.

Developing good ideas

Intellectual taste
Personality
Our interests
Logic
Avoids relabelling
Stands the test of time
Can be described to others clearly and briefly.
Simplifies the world.
Learning from bad ideas

Four hurdles to establishing causality

1. Is there a credible mechanism connecting X and Y?

2. Can we rule out Y causing X (endogeneity)?
3. Is there covariation between X and Y?
4. Have we controlled for potential spuriousness (Z)?

Useful to ask a consistent series of questions of research you come across.

What is the research question or puzzle?
What is the main theory(ies) or argument(s)?
What type of research design is used?
How well does the work surpass the four hurdles?

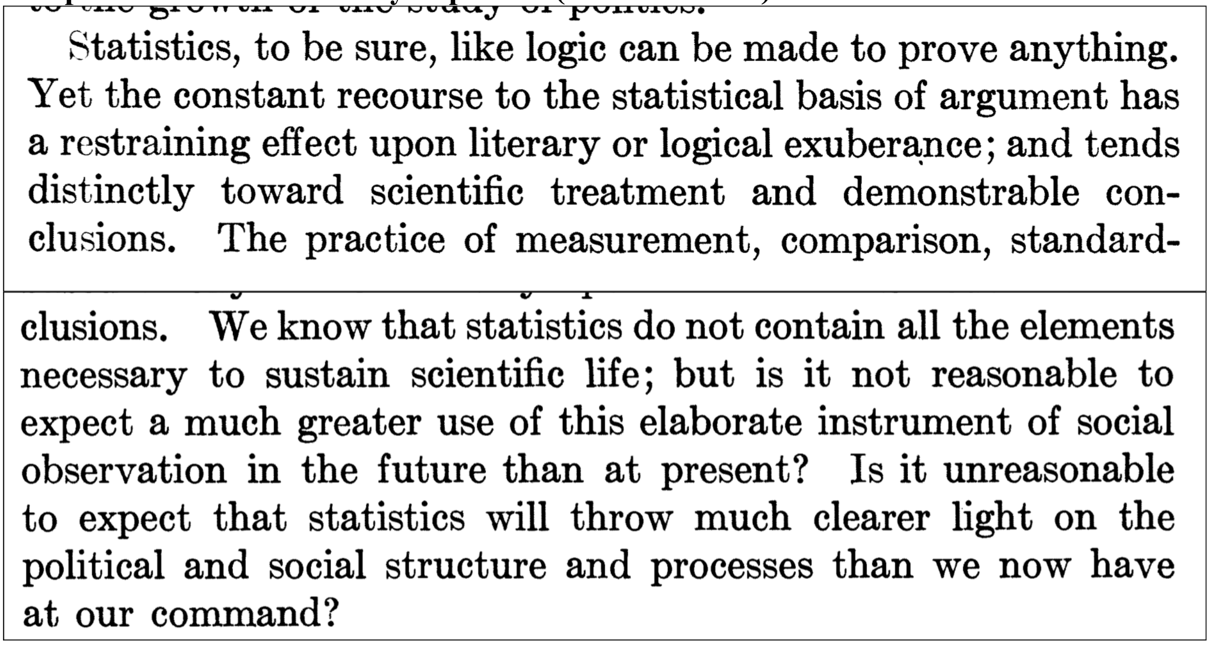
Defining descriptive arguments

"A descriptive argument describes some aspect of the world.
In doing so it aims to answer what questions (e.g. when, whom, out of what, in what manner) about a phenomenon or a set of phenomena."
(Gerring 2012: 722, emphasis added)

Definitions

"A population is any group of people, organisations, objects, or events about which we want to draw conclusions; a case is any member of such a population." (Brians et al. 2011: 132)
"A sample is any subgroup of a population of cases that is identified for analysis." (Brians et al. 2011: 132)
"A representative sample is one in which every major attribute of the larger population from which the sample is drawn is present in roughly the proportion or frequency with which those attributes occur in that larger population." (Brians et al. 2011: 133)

The present state of the study of politics (Merriam 1921)

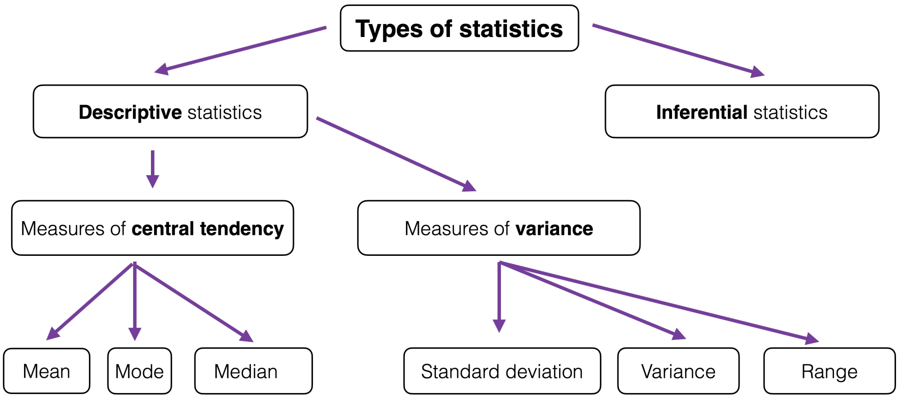


to the growth of the study of politics.

Statistics, to be sure, like logic can be made to prove anything. Yet the constant recourse to the statistical basis of argument has a restraining effect upon literary or logical exuberance; and tends distinctly toward scientific treatment and demonstrable conclusions. The practice of measurement, comparison, standard-

clusions. We know that statistics do not contain all the elements necessary to sustain scientific life; but is it not reasonable to expect a much greater use of this elaborate instrument of social observation in the future than at present? Is it unreasonable to expect that statistics will throw much clearer light on the political and social structure and processes than we now have at our command?

Statistics



Measurement metrics

Label: Employment status of survey respondent
Values: "employed" or "unemployed"
Variable type:
(1) categorical/nominal [unemployed, employed]
(2) ordinal [<5 hours, 5-15 hours, 15-35, >35 hours worked per week]
(3) continuous/interval/ratio [time worked last week]

The challenges of description

Concepts—Economic output, population, democracy
Measurement—GDP, Polity, V-Dem
Why is falsifying descriptive arguments so hard?
Describing a concept: What is democracy and how should we measure it?
Causal argument: Does democracy increase the chance of victory in war?

Describing categorical variables

Usually, we focus on the frequency distribution of categorical variables with a table, pie charts, or bar graphs.
The only central tendency statistic is the mode (the most frequent value).
Quantiles (including percentiles) are also used. They are a measure of position within a distribution.
Categorical variables
We can put cases into categories based on their values, but we cannot rank or order them.

Continuous variables

Sometimes called interval variables or ratio variables (if they have a meaningful 0).
They have equal unit differences.

Describing continuous variables

We are primarily interested in the central tendency and the distribution of values around this central tendency.
We are also interested in outliers.

The midpoint value is the median.
The average value is the mean.
The dispersion around the mean is described by the standard deviation.

Moving from theory to test

```
Independent variable (a concept) ------------Causal theory------- > Outcome (also a concept)
|                                                                                   |
|                                                                                   |
|                                                                                   |
Operationalisation                                                   Operationalisation
|                                                                                   |
|                                                                                   |
|                                                                                   |
Measured proxy-------------------------------- Hypothesis -------->Measured dependent variable
```

Probability's key properties

1. All outcomes have a probability ranging from 0 to 1.
2. The sum of all possible outcomes must be exactly 1.
3. If (and only if) two outcomes are independent, then the probability of those events both occurring is equal to the product of them individually.
4. The chance of either of two outcomes happening is the sum of their probabilities if the options are mutually exclusive.
5. 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

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

Central limit theorem

Sample size has to be large (say greater than 30 observations).

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.

Key point: The sampling distribution is normally shaped even though the underlying frequency distribution is not normally shaped.

The standard normal distribution's properties

It is symmetrical about the mean

The median, mean, and mode are the same.

It has a predictable area under the curve within a specific distance of the mean.

Skewness and kurtosis are zero.

Why conduct hypothesis testing?

Which test should 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

Research design tradeoffs

What do these tests have in common?

They use p-values in their hypothesis tests.

These p-values range from 0 to 1.

They represent the probability that "we would see the observed relationship between the two variables in our sample data if there were truly no relationship between them in the unobserved population." (KW 2018: 164).

They include a null hypothesis.

They assume the selection of a random sample from the underlying population.

They represent a comparison between the actual X-Y sampled relationship to what we expect if there was no X-Y relationship in the underlying population.

The greater the difference between reality and null expectations the more confidence we can be in the X-Y relationship in the underlying population.

What do these tests also have in common? (Limitations)

They do not tell us that the relationship is causal.

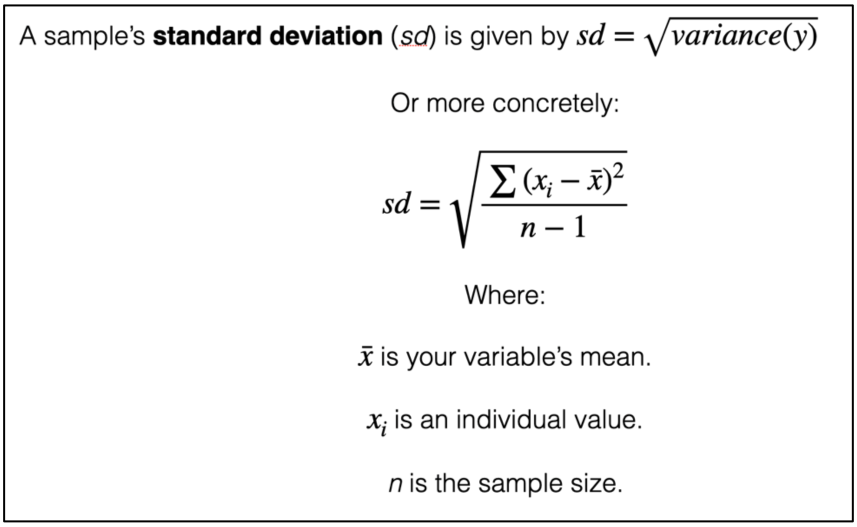
They do not tell us how strong the relationship is.

They do not tell us anything about the quality of our measures.

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 will be largely relevant to us when determining statistical significance.

Finding the **standard deviation**



A sample's **standard deviation** (*sd*) is given by $sd = \sqrt{variance(y)}$

Or more concretely:

$$sd = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

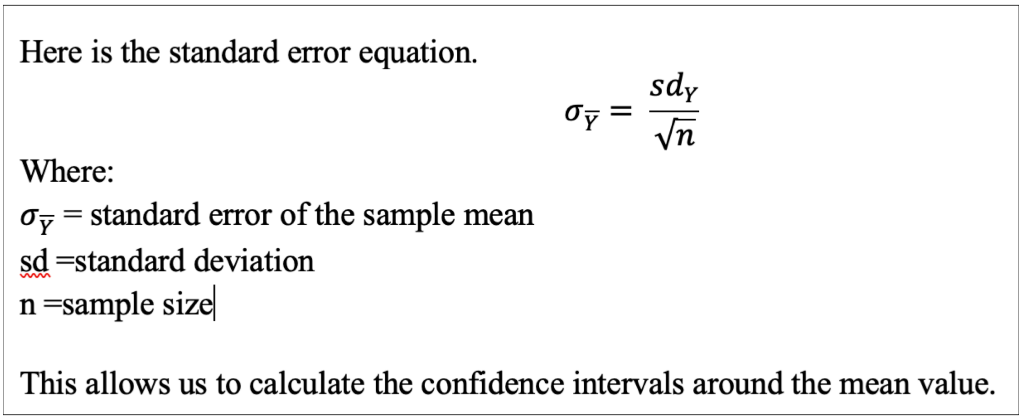
Where:

$\bar{x}$ is your variable's mean.

x_i is an individual value.

n is the sample size.

How are the standard deviation and standard error related?



Here is the standard error equation.

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

Where:
$\sigma_{\overline{Y}}$ = standard error of the sample mean
sd =standard deviation
n =sample size

This allows us to calculate the confidence intervals around the mean value.

How to calculate the margin of error?

Sampling error at 5% significance level= $\boxed{1.96\sqrt{Var/\mathrm{n}}}$

With variance (var)=p(1–p) where is the number of respondents, and p is the proportion favouring an outcome.

From margin of error to confidence intervals

How can we connect our knowledge of probability to better understand polling results?

By converting the polling results into measures of confidence that the population mean is within a certain range around the sample mean.

The margin of error is half the width of the confidence interval.
The confidence interval is thus twice the margin of error centered on the sample mean.

Confidence intervals explained

There is a 95% chance that the confidence interval which extends to two standard errors on either side of the estimate contains the "true value".
This interval is called the 95% confidence interval and is the most commonly used confidence interval. The 95% confidence interval is written as follows:

95% confidence interval for outcome y = [y – [2 * se(y)] , y + [2 * se(y)]]

Confidence intervals

To calculate the confidence interval you need four things:

The number of observations (n)
The mean (X bar)
The standard deviation (s)
The desired confidence level (let's say 95%) you go to the Z table and find the Z(0.95) score, which is 1.96.

Then you plug these values into the following equation

$$\bar{x} \pm Z\frac{s}{\sqrt{n}}$$

Difference of means t-test intuition
The t-statistic basically is a measure of the difference of means over a measure of uncertainty around those means.

Difference of means example
T-statistic: -5.05
Degrees of freedom: 134
P-value: 0.000
Therefore, I conclude that there is less than less than 1 in 1,000 chance that we would see this relationship randomly in our sample if there was no relationship in the underlying population.

Correlation
A correlation is the statistical association between two variables.
It has five important characteristics (nature, direction, sign, strength, statistical significance).
Calculating a correlation coefficient and its statistical significance is straightforward.
Interpreting what it means is a different thing and requires thinking causally.

Pearson's correlation coefficient

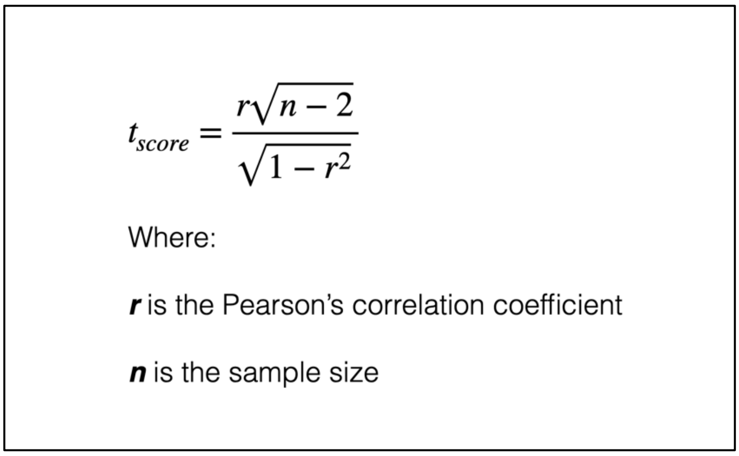
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where:
r is the coefficient of correlation between x and y
x is each individual value (i) of the independent variable
x hat is the average value of x
y is each individual value (i) of the dependent variable
y hat is the average value of y
n is the number of observations

Conducting a significance test

(rho) is the correlation coefficient.
Null hypothesis H0): rho=0 there is not a significant linear correlation between x and y in the sample.
Alternative hypothesis H1 rho is not equals to 0, there is a significant linear correlation between x and y in the sample.
Now we conduct a Student's T-test. What is that?

Student's t-test



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

Where:

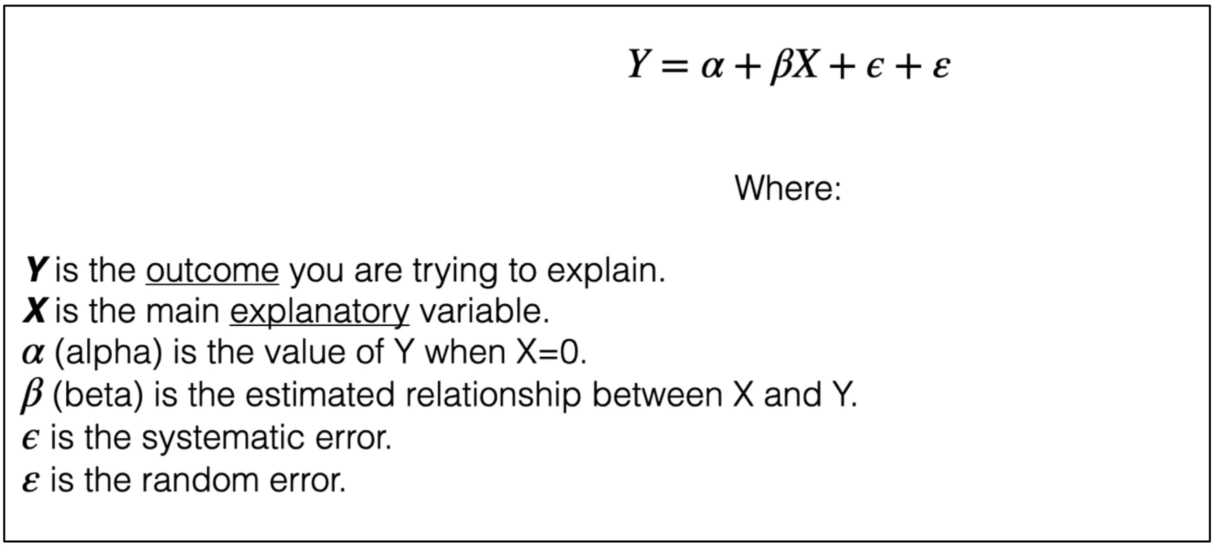
r is the Pearson's correlation coefficient

n is the sample size

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) or whether two variables are correlated?
Rather we want a more complex understanding of the directionality and significance in the relationship between an X and Y?
Or perhaps we want to predict our outcome as we vary values of our independent variable?

The bivariate regression model

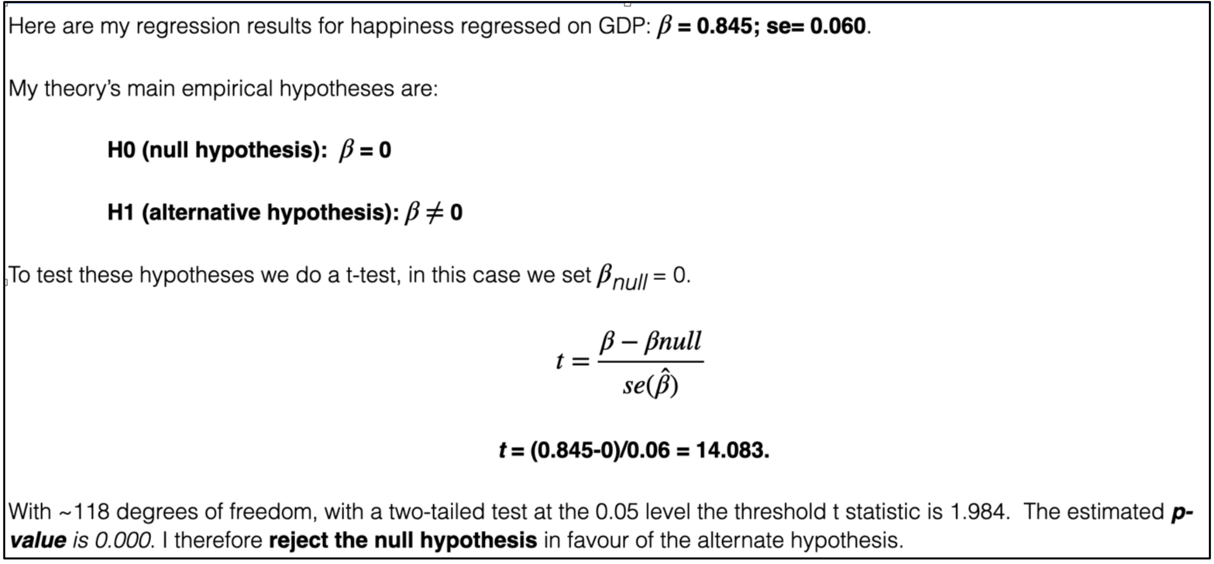


$$Y = \alpha + \beta X + \epsilon + \varepsilon$$

Where:

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

Two-tailed hypothesis testing of slope coefficient



Here are my regression results for happiness regressed on GDP: ***β* = 0.845; se= 0.060**.

My theory's main empirical hypotheses are:

H0 (null hypothesis): $\beta = 0$

H1 (alternative hypothesis): $\beta \neq 0$

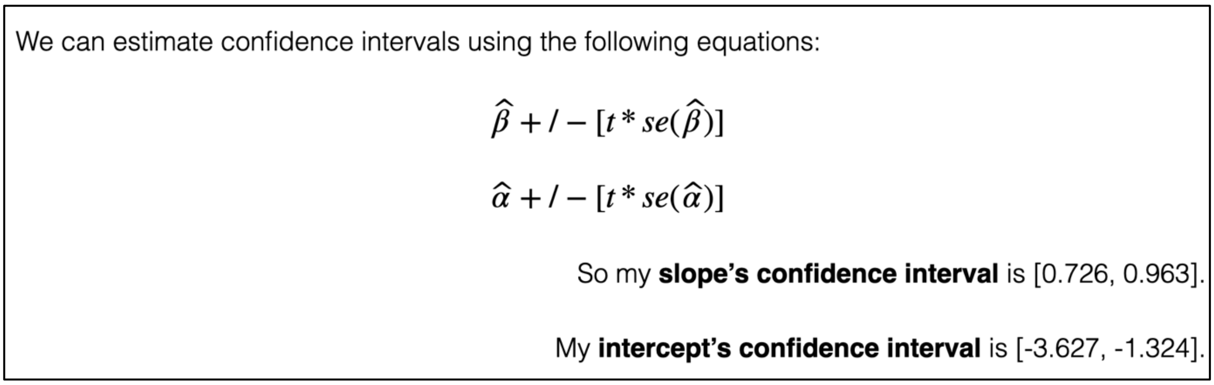
To test these hypotheses we do a t-test, in this case we set $\beta_{null} = 0$.

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

***t* = (0.845-0)/0.06 = 14.083.**

With ~118 degrees of freedom, with a two-tailed test at the 0.05 level the threshold t statistic is 1.984. The estimated ***p-value*** *is 0.000.* I therefore **reject the null hypothesis** in favour of the alternate hypothesis.

Confidence intervals



We can estimate confidence intervals using the following equations:

$$\hat{\beta} + / - [t * se(\hat{\beta})]$$

$$\hat{\alpha} + / - [t * se(\hat{\alpha})]$$

So my **slope's confidence interval** is [0.726, 0.963].

My **intercept's confidence interval** is [-3.627, -1.324].

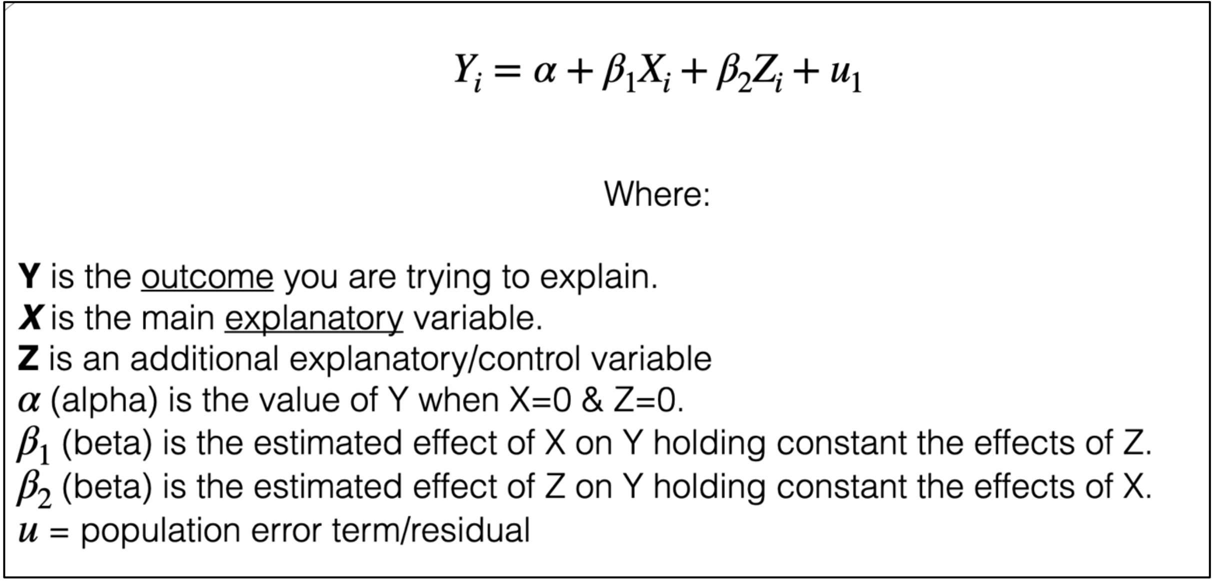
Interpret the estimated coefficient and standard error

Interpret the regression statistics: R-square

Interpret the regression statistics: Standard error

Interpret the regression statistics: F statistic

Multivariate regression model

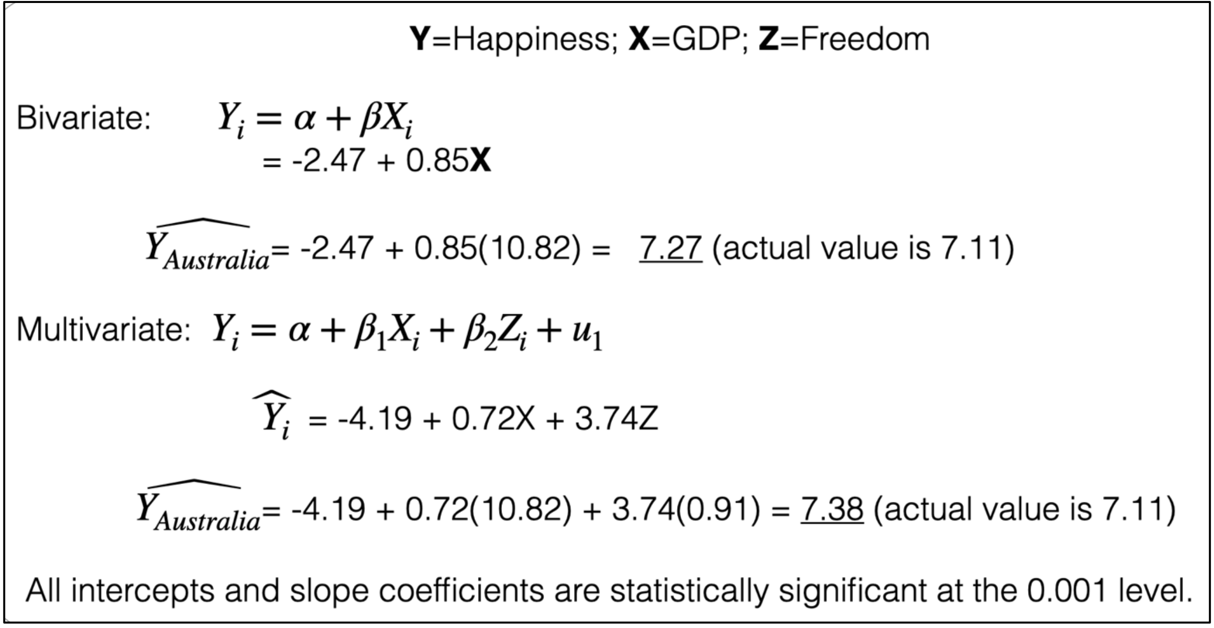


$$Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + u_1$$

Where:

Y is the outcome you are trying to explain.
X is the main explanatory variable.
Z is an additional explanatory/control variable
α (alpha) is the value of Y when X=0 & Z=0.
β_1 (beta) is the estimated effect of X on Y holding constant the effects of Z.
β_2 (beta) is the estimated effect of Z on Y holding constant the effects of X.
u = population error term/residual

Estimating the relationship between X and Y, controlling for Z



Y=Happiness; **X**=GDP; **Z**=Freedom

Bivariate: $Y_i = \alpha + \beta X_i$
= -2.47 + 0.85**X**

$\widehat{Y_{Australia}}$= -2.47 + 0.85(10.82) = 7.27 (actual value is 7.11)

Multivariate: $Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + u_1$

$\widehat{Y_i}$ = -4.19 + 0.72X + 3.74Z

$\widehat{Y_{Australia}}$= -4.19 + 0.72(10.82) + 3.74(0.91) = 7.38 (actual value is 7.11)

All intercepts and slope coefficients are statistically significant at the 0.001 level.

Interpreting a regression table

Tell your readers in words what you want them to take away from your table.
Often focus is on both statistical and substantive significance.
Connect results back to your theory and hypotheses.

Research pitfalls

#1: Correlation does not equal causation.
#2: Spurious/omitted variable problem
#3: Endogeneity

#4: Multicollinearity
#5: Transforming (or leaving) variables
#6: Stepwise regression
#7: Data mining/garbage-can regressions/overfitting
#8: Dichotomous or categorical dependent variables
#9: Time series vs. cross-sectional sample?
#10: Simpson's Paradox (trend can be eliminated/ reversed by splitting data into groups.
#11: Overlooking cross-validation
#12: Extrapolating beyond the data you have
#13: Using a regression on a non-linear relationship
#14: Publication bias
#15: Theoretical biases (Confirmation bias, interpretation bias, fundamental attribution error)
#16: Empirical biases

LECTURE PART 2: IMPORTANT TERMS

Learning terms this semester—a modest proposal

Write down the definition.
Describe why scholars think this term is useful.
Think of an example that resonates with you.

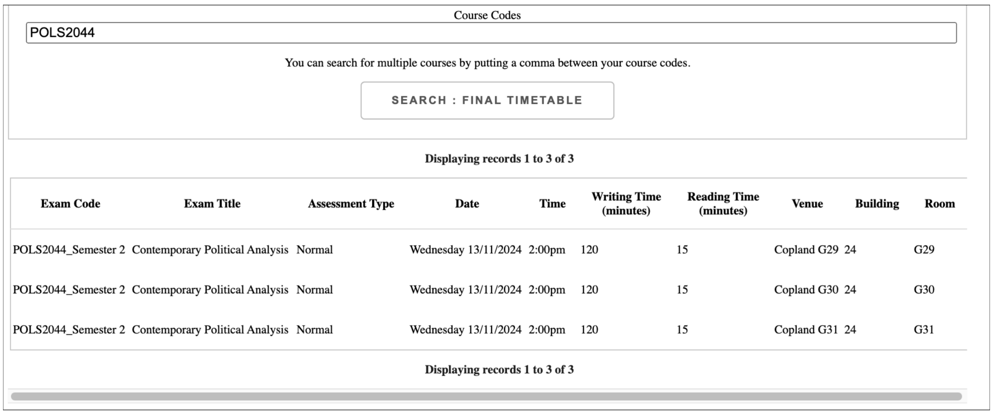
Term	Definition	Use	Example
Autocorrelation			
Bias			
Bivariate vs multivariate regression			
Categorical/nominal variable			
Causality			
Central tendency			
Ceterus paribus			
Chi-squared test			
Cluster or multistage sampling			
Complete & incomplete information			
Conceptual clarity			
Confidence interval			
Confirmation bias			
Confounding variable			
Construct validity			
Content validity			
Continuous/interval/ratio variable			
Convenience sample			
Correlation			
Correlation coefficient			
Covariance			

Cross-sectional and time-series data			
Data mining			
Datum and data			
Degrees of freedom			
Dependent variable			
Deterministic vs. probabilistic relationship			
Difference of means test			
Directional and non-directional hypotheses			
Dummy variable			
Endogeneity			
Equal unit difference			
Expected utility			
Experimental research design			
Extrapolation/interpolation			
Face validity			
Field experiment			
Formal theory			
Fundamental attribution error			
Generality			
Hypothesis testing			
Hypothesis/null hypothesis			
Independent variable			
Index/indices			
Interactive effect			
Interactive model			
Intercept			
Internal and external validity			
Interpretation bias			
Leave-one-out cross-validation			
Limited dependent variable			
List experiment			
Lower & upper bounds			
Matrix			
Mean			
Measure			
Measurement bias			
Median			
Mode			
Model standard error			
Multicollinearity			
Natural experiment			
Nonprobability sample			

Observational research design			
Omitted variable bias			
Ordinal variable			
Outliers			
P-value			
Parameter estimate			
Parameters			
Parsimony			
Perfect multicollinearity			
Placebo			
Population			
Population error term			
Preference ordering			
Publication bias			
Purposive sample			
Quantiles			
R-squared			
Random sample			
Rational choice			
Rational utility maximisers			
Reliability			
Replication			
Representative sample			
Residual			
Sample			
Sample mean			
Sampling error			
Selection bias			
Slope			
Snowball sample			
Spatial & temporal dimensions			
Spuriousness			
Standard deviation			
Standard error of the mean			
Stepwise regression			
Stochastic			
Stratified random sample			
Substantive significance			
Survey experiment			
Systematic random sample			
T-ratio			
T-statistic			

T-test			
Tabular analysis			
Theory			
Transformed variable			
Transitive & intransitive preferences			
Treatment and control groups			
Utility			
Variable			
Variable label			
Variable types			
Variable values			
Variance			
Vector			
Volunteer sample			

LECTURE PART 3: FINAL EXAM



Course Codes: POLS2044

You can search for multiple courses by putting a comma between your course codes.

Displaying records 1 to 3 of 3

Exam Code	Exam Title	Assessment Type	Date	Time	Writing Time (minutes)	Reading Time (minutes)	Venue	Building	Room
POLS2044_Semester 2	Contemporary Political Analysis	Normal	Wednesday 13/11/2024	2:00pm	120	15	Copland G29	24	G29
POLS2044_Semester 2	Contemporary Political Analysis	Normal	Wednesday 13/11/2024	2:00pm	120	15	Copland G30	24	G30
POLS2044_Semester 2	Contemporary Political Analysis	Normal	Wednesday 13/11/2024	2:00pm	120	15	Copland G31	24	G31

Displaying records 1 to 3 of 3

MULTIPLE CHOICE
(20% total, each question is worth 2% of your final mark)

Please answer the following ten multiple-choice questions. When reading the questions please be sure to read them carefully and answer the question asked.

SHORT(ER) ANSWER
(50% total, each question is worth 5% of your final mark)

In this section, please answer the following ten questions, making sure to answer all parts of the question. These questions are designed to take two to five sentences to answer adequately.

LONG(ER) ANSWER
(30%, each question is worth 15% of your final mark)

Please answer the following two questions using full sentences, complete paragraphs, and clearly structured answers.

Thank you for a great semester!

WORKSHOP ACTIVITIES

There is no workshop this week. I will be in both workshop spaces in the normal time to answer student questions. Thanks, from the POLS2044 team for being part of this class!