POLS2044
WEEK 11: Common research design pitfalls

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

In Week 11 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 or more variables. This week we reinforce Week 9 and 10's discussion of ordinary least squares (OLS) regression and highlight common regression pitfalls (and how to avoid them) as well as more general theoretically motivated research pitfalls.

This week I have two main goals. First, I want students to continue developing their understanding of OLS regression—how and why it is useful, what are its assumptions about the data you are using, and how to interpret regression results. Second, I want to highlight fifteen common mistakes made when designing or interpreting empirical models.

Reading notes and questions

There is one reading for this week, Chapter 11 of Kellstedt and Whitten (2018: 246-272). When reading this chapter, I would encourage you to focus most on (1) why dummy variables are different than the sort of continuous variables we have focused on in the last month, (2) when our theoretical arguments lead us towards interactive models, and (3) how influential cases can affect our theoretical and empirical models.

LECTURE PART 1: Theoretical pitfalls

Today's motivating questions

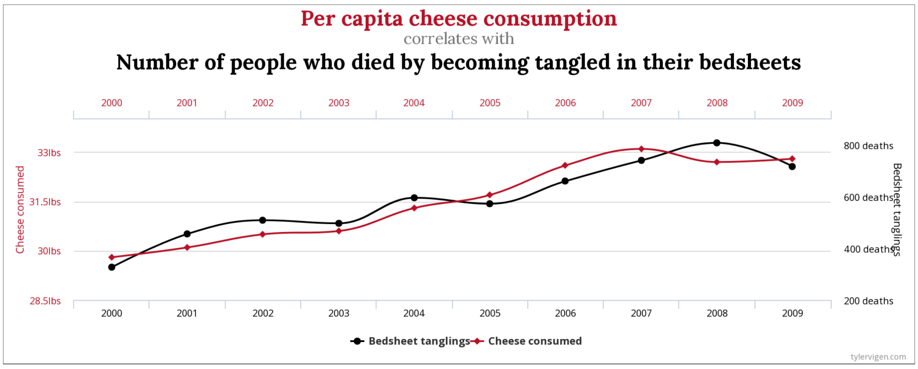
How can we minimise the chance of making mistakes when creating our research design?

What theoretical, empirical, and simple human factors should we be aware of?

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)?

Pitfall #1: Correlation does not equal causation.



Source: https://tylervigen.com/spurious-correlations

Correlation does not equal causation

It is a mistake to think there is a causal link when it could be because of chance or a third factor.

Pitfall #2: Spurious/third variable problem

"A third variable problem occurs when an observed correlation between two variables can actually be explained by a third variable that has not been accounted for."

Sources: https://www.statology.org/third-variable-problem/

X	Y	Z
# fire hydrants	# dogs	# people
Ice cream sales	# shark attacks	Temperature
# volunteers showing up to a natural disaster	Total natural disaster damage	Size of the natural disaster
Trade	Conflict	State capacity

Pitfall #3: Endogeneity

Questions to ask yourself:
Does X cause Y?
Does Y cause X?
Do they both affect each other?

Democracy example

Potential endogeneity between democratic history and individual support for democracy.

Theoretical pitfalls—important takeaways

Before we can even think about running analyses, we need to think theoretically about the myriad possible relationships between the outcome we are trying to explain (Y) and the factors (X's) that could affect it.

Ask yourself the following questions:

Is there a credible mechanism connecting X to Y?
Is there a real risk of endogeneity?
Is there significant covariation between X and Y to explain?
Have we thought about potential spurious factors (Z's)?

LECTURE PART 2: Variable pitfalls

Variable pitfalls

Previously discussed issues:

Links between concepts and proxy measurements
Raw numbers vs. ratio variables
Raw numbers vs. percentages
Raw numbers vs. indices
Mean vs. median vs. mode
Levels of analysis

A few additional pitfalls in this section

Multicollinearity
Logging and squaring variables
Stepwise regression
Data mining/garbage can regressions/overfitting
Dichotomous or categorical dependent variables

Pitfall #4: Multicollinearity

Perfect multicollinearity definition: "when there is an exact linear relationship between any two or more of a regression model's independent variables." (Kellstedt and Whitten 2018: 243)

Multicollinearity is "usually the result of a small number of cases relative to the number of parameters we are estimating, limited independent variable values, or model mis-specification." (Kellstedt and Whitten 2018: 246)

If there are two variables that are perfectly multi-collinear, one will be dropped.

Think theoretically if both variables are capturing the same underlying trait of the sample you are using.

Pitfall #5: transforming (or leaving) variables

Scholars often transform their variables for theoretical or practical reasons. **Why**?

Pitfall #6: Stepwise regression

A regression approach in which you automatically specify a final model through trial and error of adding or subtracting independent variables according to some model fit criterion.

Stepwise regression critiques

Stepwise regression can lead to overfitting.
It will explain the current data but may not do well with new data.
It can inflate accuracy estimates and statistical significance.

Pitfall #7: Data mining/garbage-can regressions/overfitting

If we include 20 variables in a model, then on average we will find one statistically significant relationship.
Most variables include missing data. The more variables you include, the smaller your sample becomes.
Some variables may do well with prediction but have only tenuous theoretical links.
Humans can only conceptualise a small number of moving parts at the same time.

Chris Achen's critique of garbage-can regressions

Pitfall #8: Dichotomous or categorical dependent variables

Example using GDP and democracy

Addressing limited dependent variables

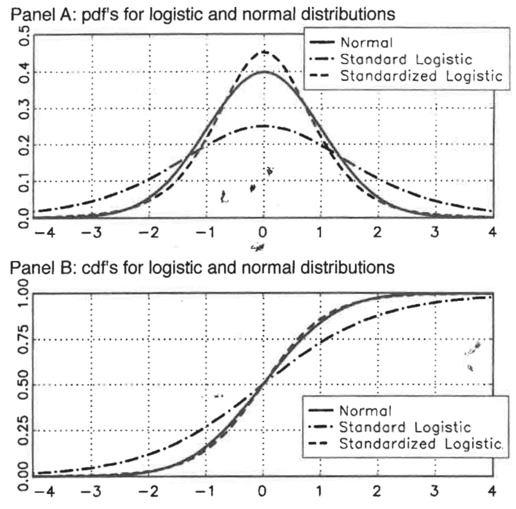


Figure 3.3. Normal and Logistic Distributions

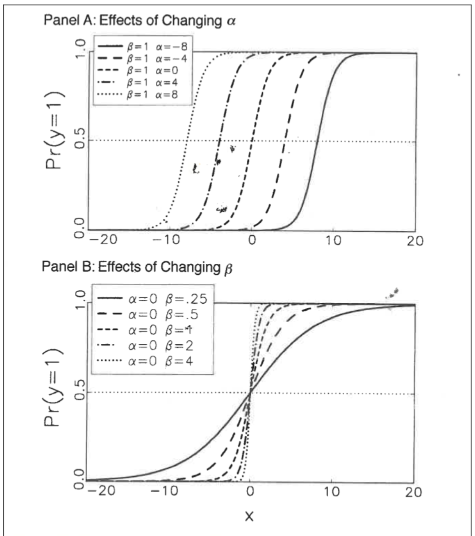


Figure 3.8. Effects of Changing the Slope and Intercept on the Binary Response Model: $\Pr(y = 1 \mid x) = F(\alpha + \beta x)$

Source: Long (1997: 43, 63)

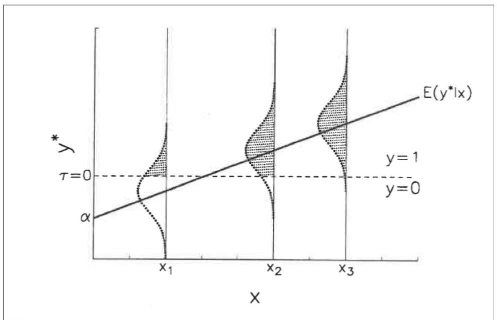


Figure 3.2. The Distribution of y^* Given x in the Binary Response Model

$$y^* = \mathbf{X}\beta + u$$

Where

$$y_i = \begin{cases} 1, & if\ \ y_i^* > \kappa \\ 0, & if\ \ y_i^* \leq \kappa \end{cases}$$

Source: Long (1997: 41, 63)

Limited dependent variables regression functions

Logit and Probit.

See that the functions include the probability of y=1 and y=0

Variable pitfalls—Important takeaways

Scholars engage in a daily balancing act when deciding: which variables to include; in what form should we include them; how to estimate our models; and which model is appropriate for the distribution of our Y.

LECTURE PART 3: Sample pitfalls

Sample pitfalls

Time series versus cross-sectional samples
Simpson's paradox
Leave-one-out cross-validation
Extrapolating beyond the data you have
Using regression on a non-linear relationship

Pitfall #9: Time series vs. cross-sectional sample?

Example of Polity2 score of South Africa over time and Africa cross-sectionally in 2018.

Pitfall #10: Simpson's Paradox

It appears that there is an "apparent trend in the data that can be eliminated or reversed by splitting the data into natural groups."
(Reinhart 2015:4)

Example using QoG data on unemployment by region

Pitfall #10: Overlooking cross-validation

A way to evaluate regressions is to run them a number of times, each time leaving out a different observation and using the results to predict this observation (leave-one-out cross-validation).

Pitfall #11: Extrapolating beyond the data you have

Along a similar vein to Simpson's paradox is the danger of thinking your results apply to a population that may or not be like the sample you used.

Pitfall #12: Using a regression on a non-linear relationship

Assuming linearity can either lead to null results or understating true relationship (Type 2 errors).

Using a regression on a non-linear relationship

Example from Gleditsch, Nils Petter, Kathryn Furlong, Håvard Hegre, Bethany Lacina, and Taylor Owen. 2006. "Conflicts Over Shared Rivers: Resource Scarcity or Fuzzy Boundaries?" *Political Geography* 25: 361-382.

Sample pitfalls—important takeaways

It is easy to get results either counter to your expectations or null effects if your theories are not well matched to your data sample.

Think about whether your theory is more about change within units (e.g. countries or people) over time or between units.
Think about whether the relationship is linear or non-linear.
Make sure to evaluate the robustness of your findings.

LECTURE PART 4: Researcher pitfalls

Pitfall #13: Publication bias

Example from Gerber and Malhotra (2008) and Breznau et al. (2022)

Pitfall #14: Theoretical biases

Researchers are human, and they often have a tendency of using a particular perspective that favours particular populations, opinions, and research questions.

There are also risks of:

Confirmation bias—interpret incoming information in light of what you already believe
Interpretation bias—e.g., hostile attribution bias
Fundamental attribution error—attribute outcomes as coming more from people's preferences rather than the situation or the structural environment.

Pitfall #15: Empirical biases

Researchers tend to use the same:

methods (e.g. OLS or probit),
data (e.g. Polity IV), and
interpretation (coefficient significance)

across papers and (often) research sub-fields.

Research is what other people do.
I think I can formulate a falsifiable hypothesis.
All I need to do is run a regression.
The best research involves storytelling & indicative data/case analysis.

Today’s motivating questions

How can we minimise the chance of making mistakes when creating our research design?

What theoretical, empirical, and simple human factors should we be aware of?

Important Week 11 terms

Confirmation bias
Data mining
Dummy variable
Extrapolation/interpolation
Fundamental attribution error
Index/indices
Interactive effect
Interactive model
Interpretation bias
Leave-one-out cross-validation
Limited dependent variable
Multicollinearity
Publication bias
Stepwise regression
Transformed variable

WORKSHOP ACTIVITIES

Welcome to our final workshop! Today we are going to be reinforcing your knowledge of, and comfort with, the subject matter of weeks 6-10 before adding in one last thing...

For this workshop, we will be using the Quality of Governance Indicators data from Week 8. It is available in "Wattle/Week 11/Workshop/" The dataset is called *week_11_workshop_data.xlsx.*

We will be focusing on running the statistics we have run in previous workshops. All necessary Excel commands are available in the relevant spreadsheets from previous weeks. This process is geared to give you an opportunity to revisit the material and reinforce your knowledge of these techniques.

1. Descriptive statistics
2. A difference of means test
3. A correlation coefficient
4. Bivariate regression
5. Multivariate regression

Finally, we will estimate the expected value of our dependent variable given our multivariate results.

Remember to submit your own work (not anyone else's) to "Wattle/Week 11/Workshop/Item 11.1" at the end of workshop. **Please only submit one Word document with the numbers of the questions and your responses, not this entire document with your answers in another colour.**

Part 1: Descriptive inference

We will begin by learning a bit about our variables, how they are distributed, and what their central tendencies are.

Step 1: Choose any three variables from this dataset as long as they are not `EU_gdp` or `non_EU_gdp`.

Step 2: Find their descriptive statistics using the Analysis TookPak's descriptive statistics option.

Step 3: Fill in a table summarising your descriptive statistics using the template in Table 1. These will be needed for the last workshop question below.

Question 1: Show your completed descriptive statistics table below.

Table 1. Summary statistics and interpretation

Variable name	Mean	Median	Mode	Standard deviation	Minimum	Maximum

Question #2: Describe any descriptive findings you find particularly interesting about the distribution of your variable. Any substantive descriptive conclusions you can reach given these descriptive statistics?

Step 4: Create a scatterplot or histogram of one (or two) of your three variables, preferably your dependent variable.

Question #3: What variable(s) did you choose and why did you choose them? Did the graph conform to your expectations? Why/why not? If you created a histogram, does the distribution approximate a bell curve or is it skewed or distributed in any other notable way? If you created a scatter plot, can you discern any clear pattern in the relationship between your variables?

Part 2: Hypothesis testing—A difference of means test

Now, please run a difference of means test of the average values of GDP in European Union and non-European Union countries. Before doing so, you need to specify your observable expectations.

Step 5: Please write (a) a suitable hypothesis and (b) a null hypothesis for your difference of means test.

Question #4: What are your two hypotheses?

Step 6: run your difference of means test.

Once you have run your difference of means test, please answer the following questions.

Question #5: Is the difference between the average GDP of European Union countries (`EU_gdp`) and non-European countries (`non_EU_gdp`) in the sample statistically significantly different from each other?

Question #6: Do your results support (a) rejecting the null hypothesis in favour of your alternate hypothesis or (b) do you fail to reject the null hypothesis? How do you reach this conclusion?

Part 3: Hypothesis testing—Correlation coefficient

Step 7: Correlate two of the three variables you chose above for your descriptive statistics table in Part 1 and run a Student's t-test.

Given our sample size, the threshold T-statistic for more than 120 degrees of freedom is 1.96. Remember to calculate the T-score, the equation is:

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

where ***r*** is Pearson's correlation coefficient and ***n*** is the number of observations.

> **Question #7:** What is the value of the Pearson's correlation coefficient you calculated? Is it as high (or low) as you were expecting? What is your t-score? Is your t-score greater than the threshold value? What does this tell you about the relationship between these two variables?

Part 4: Hypothesis testing—Bivariate regression

Now that you have a bit of background about several variables and their relations to each other, it is time to run a bivariate regression. Make sure to keep the results output, as you will be using it again in a later section.

Step 8: From the three variables in Part 1, choose a variable to be your outcome variable (dependent variable, Y) and one as your explanatory variable (independent variable, X).

> **Question #8:** What is a potential research question about these variables?
>
> **Question #9:** What is a plausible (the most plausible you can think of) causal mechanism linking your Y and X?
>
> **Question #10:** What is an observable null hypothesis and alternate hypothesis?

Step 9: Run your bivariate regression.

> **Question #11**: Do your results allow you to reject your null hypothesis or not? How do you reach this conclusion?
>
> **Question #12:** What is your R-squared and your F-statistic and explain whether your F-statistic is statistically significant.

Part 5: Hypothesis testing—Multivariate regression

Hopefully, it has occurred to you that there are other potential explanatory factors that affect your dependent variable. In this section you are going to add in one additional variable (a control variable from Part 1) to your model.

> **Question #13:** What is your control variable? Why do you think it is worth controlling for this factor.

Step 10: Run your multivariate regression. ***Remember that Excel will only allow multiple columns as X variables if they are next to each other, so you will likely need to copy and paste columns next to each other.***

Step 11: Make a regression results table akin to what you made last week. Include two columns, one for each regression you run. A template table is included below as Table 2.

Table 2. Causes of my outcome variable, 2021

Independent variables	**Model 1: bivariate**	**Model 2: multivariate**
Variable 1		
	( )	( )
Variable 2		
	( )	( )
Intercept		
	( )	( )
Observations		
R^2		
F-statistic		

Note: * = p<0.05, two-tailed test. Standard errors in parentheses.

Question #14: Place your completed Table 2 here.

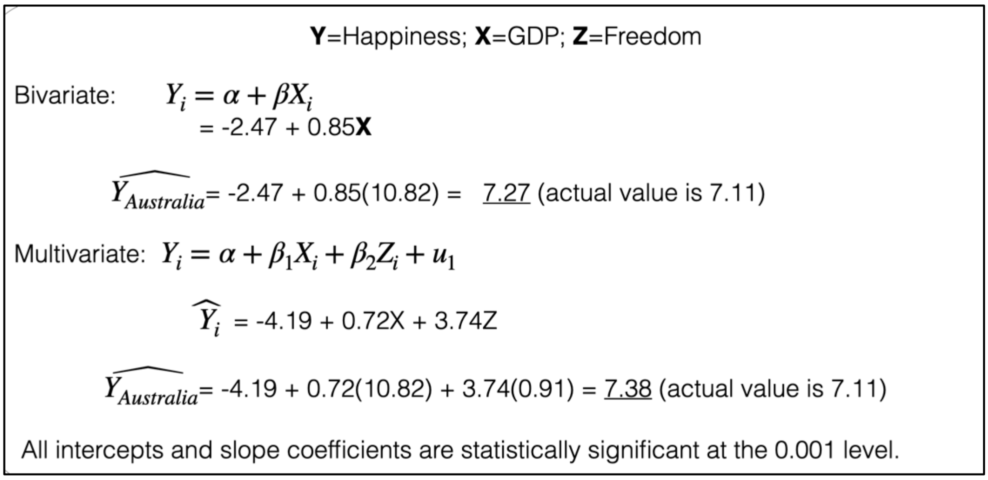
Question #15: What do you conclude given your regression results? In other words, what do you want readers to take away from your table?

Question #16: Would your conclusions change if you shifted from a non-directional to a directional hypothesis (and thus use a one-tailed test instead of a two-tailed test)?

Part 6: Expected values

Finally, I want to push just a bit further and use the regression results. I am curious to see what the expected values of your outcome variable would be for specific values of your independent variables.

Please harken back to week 11 and the slide where I looked at Australia's expected happiness given my results for `GDP` and `Freedom` regressed on `Happiness`.



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

Bivariate: $Y_i = \alpha + \beta X_i$

$= -2.47 + 0.85\mathbf{X}$

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

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

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

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

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

The equation you want to use is $\widehat{\boldsymbol{Y}_{\boldsymbol{\iota}}} = \boldsymbol{\alpha} + \boldsymbol{\beta}_{\mathbf{1}}\boldsymbol{X}_{\mathbf{1}} + \boldsymbol{\beta}_{\mathbf{2}}\boldsymbol{X}_{\mathbf{2}}$ where *alpha* is the intercept, *beta$_1$* is the slope for your independent variable, X_1 is the independent variable value (calculated below in step 12a), *beta$_2$* is the slope for your control variable, and *X_2* is the control variable value (calculated below in step 12b below).

Step 12: Please calculate the expected value of your dependent variable given the simultaneous values of (a) your independent variable that is one standard deviation **above** the mean and (b) your control variable at one standard deviation **below** the mean. To find the values to plug in, you will need to run descriptive statistics on your independent and control variables like what you did in Part 1 above.

> **Question #17:** What is the estimated value of your dependent variable given these values? How does it compare to the average value of your dependent variable?
>
> **Question #18:** What do you think would happen if you set the values of your independent and control variables to their lowest values? Their highest?

Final thoughts

Hopefully, you found this an interesting exercise in both (1) thinking theoretically, (2) how you run a variety of statistics to evaluate your theories, and (3) how you describe the results to others.

Do let me know if you have any questions about any part of this process. Thanks again for all your hard work and discussions in workshops this term. I hope to see you in lecture next week!

-Richard