POLS2044 WEEK 11
Common research design pitfalls

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

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 are two readings for this week. Instead of the originally assigned Angrist and Pischke (2009) chapter, please read chapter 12 of Wheelan (2013).

Both the Reinhart (2015) and Wheelan (2013) chapters discuss some pitfalls and challenges when conducting and interpreting multiple regression. The goal is to try and avoid Type I (false positive) and Type 2 (false negative) errors.

Reinhart, Alex. 2015. "Chapter 8: Model Abuse," in *Statistics Done Wrong: The Woefully Complete Guide*. San Francisco: No Starch Press: 79-88.

This chapter focuses on several ways of analysing data that increases the risk of bias towards either Type I or II errors.

1. What is overfitting?

The watermelon example is a clear example of overfitting, 16,00 independent variables regressed on the ripeness of only 43 melons! However, think about your research and to what extent that this might also be the case in your research area.

2. What is the connection between overfitting and the concept of degrees of freedom we have covered in earlier weeks?
3. What is stepwise regression? Why is Reinhart critical of it? How are forward selection and backward elimination both raise the risk of overfitting and may make the resulting model useless for out-of-sample modelling and prediction?
4. What is leave-one-out cross-validation?

5. Do you think leave-one-out cross-validation might be an interesting approach of use for your paper (if it had been available in Excel)?
6. Relatedly, are there any observations in your data that may be overly influential in your model?

We have talked extensively in this class about the differences between correlation and causation. What is interesting in this chapter is the links to the *ceterus paribus* assumption.

7. Why does Reinhart (2015) think that it may not be possible to hold all other variables constant in practice?
8. What is Simpson's Paradox?
9. Can you think of any political science examples of Simpson's Paradox?

Wheelan, Charles. 2013. "Chapter 12: Common Regression Mistakes," in *Naked Statistics: Stripping the Dread from the Data*. London: W.W. Norton: 212-224.

"What could possibly go wrong? All kinds of things." (Wheelan 2013: 213)

10. What are the seven common regression mistakes at the heart of this chapter?
11. How are these mistakes similar or different from those in Reinhart (2015)?
12. What are the two key lessons he concludes with?

LECTURE PART 1: Theoretical pitfalls

Pitfall commercial

(https://youtu.be/DA4V-n8Ft3g).

Do you recognise the first child actor?

Today's motivating questions

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

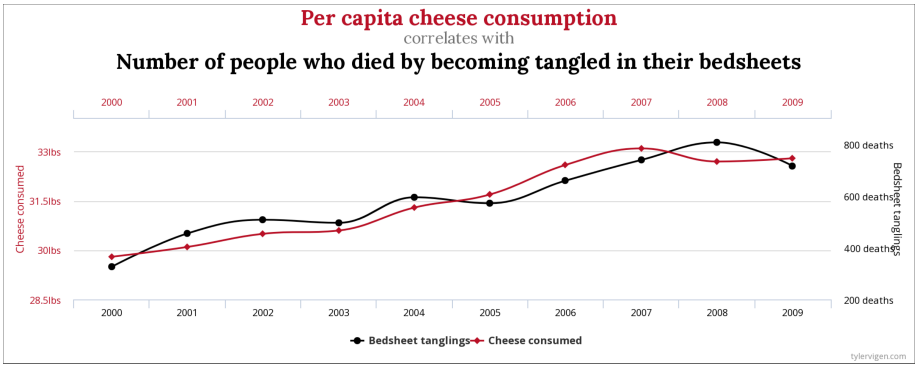
What theoretical, empirical, and simple human factors should we be aware of?

Puzzling

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
Race	Educational attainment	Racism
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.

Live demonstration of multicollinearity examples

Using Quality of Government data

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

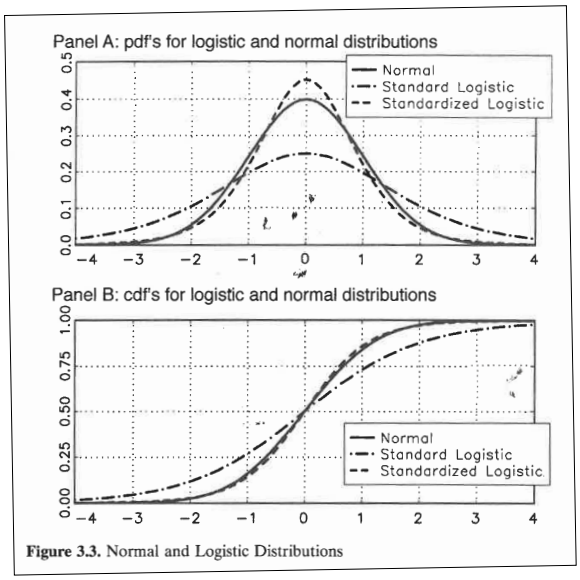


Figure 3.3. Normal and Logistic Distributions

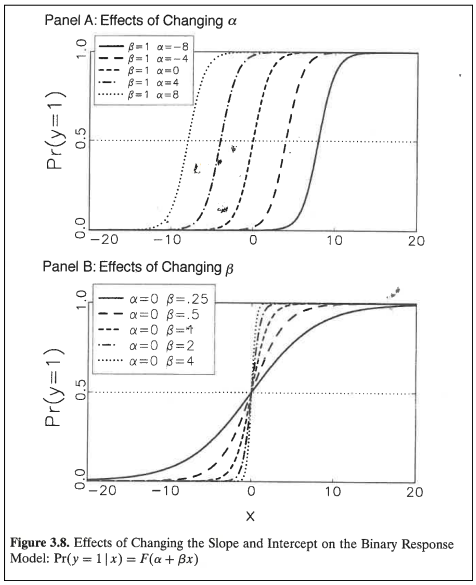


Figure 3.8. Effects of Changing the Slope and Intercept on the Binary Response Model: $\Pr(y = 1 \mid x) = F(\alpha + \beta x)$

Source: Long (1997: 43, 63)

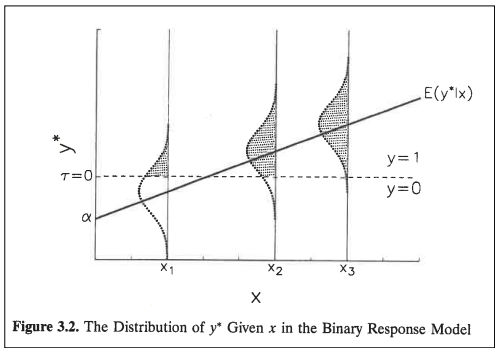


Figure 3.2. The Distribution of y^* Given x in the Binary Response Model

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

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: 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 similar to 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)

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 have a tendency 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.

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?

TUTORIAL ACTIVITIES

The final essay is now a few short weeks from being due. This week's tutorial is geared towards discussing (1) your research, (2) how it connects to the course material, and (3) how you can use these new techniques in your final papers, and (4) how you can avoid the pitfalls I discussed in lecture.

Today's tutorial is broken up into two parts—one-part small group, one-part the entire tutorial.

Part 1: Small group discussion of research projects (~4 students, 20 minutes)

So, you have our feedback on your research proposals. The final paper is due in a few weeks. Now what?

Please take a few minutes to discuss with the members of your group (1) what you feel confident about in your final paper plans, (2) what you feel less confidence about, (3) and what questions you have about the final paper project.

1. Are there any similarities in your group in your responses to the three elements above? If so, what are they?

In this week's lecture, I highlight several potential pitfalls as we design, conduct research, and write up our findings. To help jog your memory, I have summarized my lecture in Table 1 below.

Table 1. This week's research design pitfalls

Pitfall	Description
Theoretical pitfalls	
1	Correlation does not equal causation
2	Spurious/third variable problem
3	Endogeneity
Variable pitfalls	
4	Multicollinearity
5	transforming (or leaving) variables
6	Stepwise regression
7	Data mining/garbage-can regressions/overfitting
8	Dichotomous or categorical dependent variables
Sample pitfalls	
9	Time series vs. cross-sectional sample?
10	Simpson's Paradox
11	Extrapolating beyond the data you have
12	Using a regression on a non-linear relationship
Researcher pitfalls	
13	Publication bias

14	Theoretical biases
15	Empirical biases

As a group, please discuss which pitfalls you are the most concerned with in your own research.

2. Which pitfalls were the most frequent topic of conversation?

3. As a group can you help each other think of ways of avoiding these research challenges? If so, what ways did you come up with?

Part 2: Whole-tutorial discussion (remainder of tutorial)

In lecture and tutorial, we have been introduced to myriad ways of (1) asking a research question, (2) building on the literature when describing your own causal mechanism, (3) writing explicit, falsifiable, and clear hypotheses (and null hypotheses), (4) designing and executing an empirical test of your hypothesis, (5) and interpreting and discussing your results.

This is a lot of (often difficult) ground to cover. In this section, then, I want to make sure that you can all come together and discuss the main highlights of your small group work and ask questions of the tutor as well as the entire group.

4. What did you find the most interesting, compelling, boring, unclear, or challenging topics/issues/methods we have covered so far?