Problem Set 2: Macro Analysis

1 Purpose

The purpose of this problem set is to assess your understanding of one key method of quantitative public opinion research: cross-sectional and panel regression analysis.

2 Overview

You are asked to watch a short slidecast covering relevant material and then complete the following tasks. Questions regarding this material should be raised on the Moodle discussion board (`https://moodle.lse.ac.uk/mod/forum/view.php?id=521541`) or in instructor office hours.

3 Your Task

1. Thinking of different ways of designing public opinion research, define "cross-sectional design," "repeated cross-sectional design," and "panel design". Identify the advantages and disadvantages of each design.

   Solution:

   (a) A *cross-sectional design* is a survey design involving observation of a sample of units from a population at one point in time

   (b) A *repeated cross-sectional design* is a survey design involving observation of multiple, independent samples of units from a population at different points at time (typically using the same sampling design each time)

   (c) A *panel design* is a survey design involving observation of a single sample of units at multiple points in time, or equivalently multiple dependent samples at multiple points in time.

   The advantages and disadvantages of each approach are numerous. Cross-sections are simple and relatively cheap and repeated cross-sections increase costs only on a per-interview basis (since the sampling frame and design are already established). Panels either reduce or increase costs relative to that approach depending on the relative costs of recruiting a new respondent versus retaining a former respondent. Panel attrition, however, is inevitable over long periods of time. Panel designs are the only approach that enables the assessment of within-subjects comparisons (though "pseudo-panel" techniques may provide reasonable estimates of this using matching algorithms to compare units across repeated cross-sections).

2. Consider the following hypothetical data on attitudes toward the European Union (measured as 0=negative and 1=positive) in the UK and in Germany collected at two points in time:

   - Germany, time 1: .70
   - Germany, time 2: .80
   - United Kingdom, time 1: .60
   - United Kingdom, time 2: .40

(a) Assuming repeated cross-sectional surveys of n = 500 in each country at each point in time, what is the standard error of each estimate?

Solution:
Here we use the standard formulae for variance of a proportion. Given the large population, we can ignore the finite population correction.

$$Var(p) = (1-f)\frac{p(1-p)}{n-1} = \frac{p(1-p)}{n-1}$$
$$Var(p_{1,\text{Germany}}) = \frac{0.7 * 0.3}{499}$$
$$SE_1 = 0.021$$
$$Var(p_{2,\text{Germany}}) = \frac{0.8 * 0.2}{499}$$
$$SE_2 = 0.018$$

$$Var(p_{1,\text{UK}}) = \frac{0.6 * 0.4}{499}$$
$$SE_1 = 0.022$$
$$Var(p_{2,\text{UK}}) = \frac{0.4 * 0.6}{499}$$
$$SE_2 = 0.022$$

We rely here on the observed proportions to calculate sampling variance and standard error rather than the maximum possible variance because the sample element variance, s^2 is treated as an estimate of the population element variance, σ^2, which is what actually determines sampling variance.

(b) How much farther apart are the UK and Germany at time 2 than they were at time 1?

Solution:
Simple algebra:

$$(0.80 - 0.40) - (0.70 - 0.60) = 0.30 \quad (1)$$

Germany and the UK are 30 percentage points further apart at time 2 than they were at time 1.

(c) Are the differences between the UK and Germany at each time point (i.e., separately for time 1 and time 2) statistically significant?

Solution:
The calculation here require simply taking the difference in proportions over the standard error of the of the proportion for the combined samples. Because the sample sizes are the same, the calculation of the pooled standard error is pretty simple using the average level of support across the two countries at each point in time:s

$$SE_1 = sqrt(\frac{.65 * .35}{1000}) = 0.015$$
$$SE_2 = sqrt(\frac{.60 * .40}{1000}) = 0.0155$$

Time 1:

$$z = \frac{0.7 - 0.6}{0.015} = 6.67 \quad (2)$$

Time 2:

$$z = \frac{0.8 - 0.4}{0.0155} = 25.81 \tag{3}$$

Both of these z-statistics are extremely large such that $p < 0.05$ so the between-country differences are statistically distinguishable from zero.

(d) If the data were instead collected from a panel (i.e., the same respondents interviewed twice), what additional information could we learn about opinion dynamics in the two countries? What would this tell us about the individual-level change in opinion (above and beyond the macro-level changes)?

Solution:
If we collected a panel, we could additionally ascertain within-subjects changes over time. This would allow us to see whether individuals changed positions between time 1 and time 2. For example, between time 1 and time 2, the support in Germany changed from 0.7 to 0.8. This could be from 5% of the population changing positions all in one direction or many more changing in either direction.

3. Consider the following table reporting the results of a hypothetical regression of support for same-sex marriage rights (measured 1=strongly oppose to 7=strongly support) as a function of left-right political ideology (measured from 0=left to 1=right):

	Coef.	SE
Ideology	2.5	0.7
Intercept	4.1	0.6
R-squared	.32	
n	1500	

(a) In plain language, what relationship does ideology appear to have with support for same-sex marriage rights? Is this relationship substantively large or small?

Solution:
Conservative/right-wing ideology appears to be positively correlated with support for same-sex marriage rights. A full-domain shift from 0 to 1 on this scale is associated with an increase in support of 2.5 scale points on the 1 to 7 response scale. This is probably a substantively large increase given that support among those in the far left-wing is already 4.1 on that scale, making support about far-right ideologues 6.6.

(b) Is the relationship between ideology and same-sex marriage support statistically distinguishable from zero? (Show your work.)

Solution:
To determine statistical significance, calculate the t-statistic:

$$t = \frac{\beta}{SE_\beta} = \frac{2.5}{0.7} = 3.57 \tag{4}$$

This is a large t-statistic and, given the sample size, represents a very small p-value ($p < 0.05$).

(c) For what reasons might we hesitate to draw a causal inference from these data?

Solution:
Two reasons come to mind. One is substantive: the relationship appears to be the opposite of what one would a priori expect (which is that left-wing ideologues would be more support of same sex marriage). The more technical reason is that this relationship is likely to be confounded by many factors, none of which are conditioned upon.

4. Imagine we are interested in knowing the effect of economic performance on support for the government, conditional on whether one's own party is in government. To test this a researcher collects monthly data on percentage GDP growth and, from a repeated cross-sectional survey over a ten-year period, an average feeling thermometer rating of government support (ranging from 0 to 100) separately by party (assume two parties: government and opposition).

   (a) Describe the structure of the dataset. What is the unit of analysis? How many observations are there? What are the variables?

   Solution:
   This is an aggregated dataset, so the unit of analysis is a party-month. There are $12 * 10 = 120$ periods and 2 observations per period for a total $n = 240$.

   (b) Assuming economic performance and party identification are exogeneous, describe a regression model that would provide an estimate of the effect of economic performance on the feeling thermometer for each group. How would you interpret the regression coefficients in each case?

   Solution:
   To estimate the influence of party identification and economic performance, the regression model is simply:

   $$\text{Thermometer}_t = \beta_0 + \beta_1 \text{Government} + \beta_2 \text{GDP}_t + \beta_3 \text{Government} * \text{GDP}_t \quad (5)$$

   The interpretation here would be that β_0 is the mean level of opposition support for the government when GDP growth is zero. β_1 is the mean level of government party support for the government when GDP growth is zero. β_2 is the effect of a one-percentage change in GDP on average thermometer ratings of opposition party supporters and β_3 is the corresponding effect for government party supporters.

5. Consider a four-wave panel survey that begins with 2000 respondents. The goal of the study is to assess the relationship between economic uncertainty (measured as "high" or "low") and support for economic redistribution (measured on a 0-1 scale), measured at each panel wave.

   (a) Assuming no attrition, if all respondents to all waves were pooled and analyzed together in a single regression model of opinion on economic uncertainty, what are we ignoring in our analysis? Can we consider the coefficient on economic uncertainty to be a causal effect?

   Solution:
   We are ignoring the fact that we have dependent observations, thereby ignoring unobserved heterogeneity (person-specific variations along unobserved dimensions).
   We should not consider this coefficient to be a causal effect.

   (b) In the scenario in (a), what consequence does this approach have for the uncertainty surrounding our estimates?

Solution:
By pooling observations, we effectively are inflating our sample size by 400% and thus decreasing our sampling variances (and standard errors). We only have 2000 observations but we are basically pretending that we have $n = 8000$.

(c) Another way to estimate the effect of uncertainty on support for redistribution is a fixed effects regression. In what ways would this approach improve our inference? In what ways would it limit us?

Solution:
The fixed effects approach respects the panel structure and allows us to make within-subjects comparisons. Because fixed effects regression involves estimation of a person-specific "effect" (or deviation of the unit-specific over-time mean from the population global mean), we can account for all unobserved time-invariant variables without even needing to observe for them or estimate their influence. The challenge, then, is that we cannot estimate the influence of those time-invariant factors.

(d) If 15% of remaining respondents attrit (leave the panel) after each panel wave, how many respondents are left in wave 4? Briefly describe how you could (or would) deal with this attrition in practical or statistical ways.

Solution:
Calculating this requires simple algebra in the style of compounding interest:

$$n_{t1} = 2000 \tag{6}$$
$$n_{t2} = 0.85 * 2000 = 1700 \tag{7}$$
$$n_{t3} = 0.85 * 1700 = 1445 \tag{8}$$
$$n_{t4} = 0.85 * 1445 = 1228 \tag{9}$$
$$\tag{10}$$

Addressing this practical ways would involve attempting to retain some or all those who have left the panel. One could do this strategically by focusing on those most likely to leave the panel (e.g., if the lowest income groups leave at higher rates, selectively expending resources to retain these individuals compared to higher-income individuals who leave the panel). One could also try to recruit more committed panelists at $t1$ at the expense of representativeness. Another strategy would be to recruit replenishment samples or attempt to replace individuals in the panel with similar new recruits.

Statistically, it might be possible to deal with attrition by either reweighting the remaining respondents (i.e., post-stratification) to match the original sample or the population, or to use data imputation techniques to impute the missing values.

4 Submission Instructions

You should submit your problem set as a Word (.docx) document or PDF via Moodle.

5 Feedback

You will receive feedback within two weeks.