# Analysing the relationship between modernization and democracy

## Social Science Research Methods: University of Sydney

## Introduction

This script is designed to help you analyse the data on development and democracy for the Social Science Research Methods class at the University of Sydney. You can follow along by asking the questions in ChatGPT below and including your own questions and explorations of the data as you like. I’ll guide you through the analysis step-by-step but follow along in ChatGPT.

**Ask questions at any time!**

## Getting to know the data

We always want to get to know our data first. What variables are included in our dataset, over what “units” are the variables measures? Are they measured over spatial units, temporal units, or both? How many observations do we have and where is the missing data. Let’s get to know our dataset a little better.

Questions to answer:

* How many observations are there in your dataset?
* What are these observations of?
* What variables do you have in your dataset?
* How are they measured?
* How many missing observations do you have for:
  + The democracy score
  + GDP per capita
* Overall, how would you describe your dataset? What is this a sample of?

Some prompt tips for ChatGPT:

* Asking the questions above should give you some reasonable answers but here are some additional prompts you might want to try:
  + Show me the top 10 rows of the dataset.
  + What are the variable names in this dataset?
    - Give a brief description of each variable.
  + You can also ask ChatGPT to describe the dataset.
  + Which units are included in this dataset and observed over what period?
  + Produce a plot with year on the x axis, country\_name on the y axis colour the observation black if it is in the dataset?

## Descriptive statistics

Next, it is always good to understand how the variables in our dataset vary (and especially those that are our main independent and dependent variables). This will help us to identify patterns in our dataset and may already prompt us to think about alternative explanations.

Questions to answer in this section:

* Are the countries in this sample mostly democratic or autocratic?
* How have democracy levels changed over time in your sample?
* What are the top 20 cases with the highest democracy scores?
* What are the top 20 cases with the lower democracy scores?
* Which countries had the highest average democracy scores?
* Which countries had the lowest average democracy scores?
* What were the key moments of democratic transition for each country?
* Are countries in this sample mostly rich or poor?
* How has GDP per capita changed in this sample?
* Which countries had the lowest GDP per capita?
* Which countries had the highest GDP per capita?

Prompt tips:

* Has this sample been mostly democratic or autocratic?
* What is the average democracy score?
* Produce a plot with year on the x axis, country\_name on the y axis and the colour the level of v2x\_polyarchy?
* Plot the average v2x\_polyarchy score over time.
* Plot a histogram of v2x\_polyarchy.
* Has this sample been mostly rich or poor?
* What is the average GDP per capita score?
* Plot the average GDP per capita score over time.
* Plot a histogram of GDP per capita.
* For all the countries in this sample, what are the years or periods where there was an important jump towards democracy?
* Which countries had the highest average democracy score over the sample period (round to 3 decimal places)?
* Which countries had the highest average GDP per capita over the sample period (round to 3 decimal places)?

## Bivariate associations

In this section we begin to explore whether GDP per capita and democracy are associated with each other, without controlling for any other factors. This is a “bivariate” analysis (i.e. using only two variables). It’s often good to conduct a bivariate analysis as a “sanity check” to see if there is any basic association between your variables (and in the direction that you expect). But as we learned last week, if you do observe a relationship, this is not the final word, and even if you don’t there could still be a relationship when you start dealing with alternative explanations. Ask me about the development of the literature on peacekeeping for an example.

Questions to answer in this section:

* Are the richest countries also the most democratic?
* Are the poorest countries also the least democratic?
* Are GDP per capita and the democracy score associated with each other?
* There is always randomness in data, how likely is it that this type of association could occur by random chance?
* Try to recall the hurdles to causality and think about which hurdles have not been crossed and what we might do to cross them.

Prompt tips:

* What is the average democracy score for the 20 richest observations and how different is this from the mean?
* What is the average democracy score for the 20 poorest observations and how different is this from the mean?
* Produce a scatterplot of the polyarchy score and GDP per capita.
  + Add a linear trend line.
* What is the correlation between the polyarchy score and GDP per capita?
* Is this correlation coefficient statistically significant?

## “Statistical Significance”

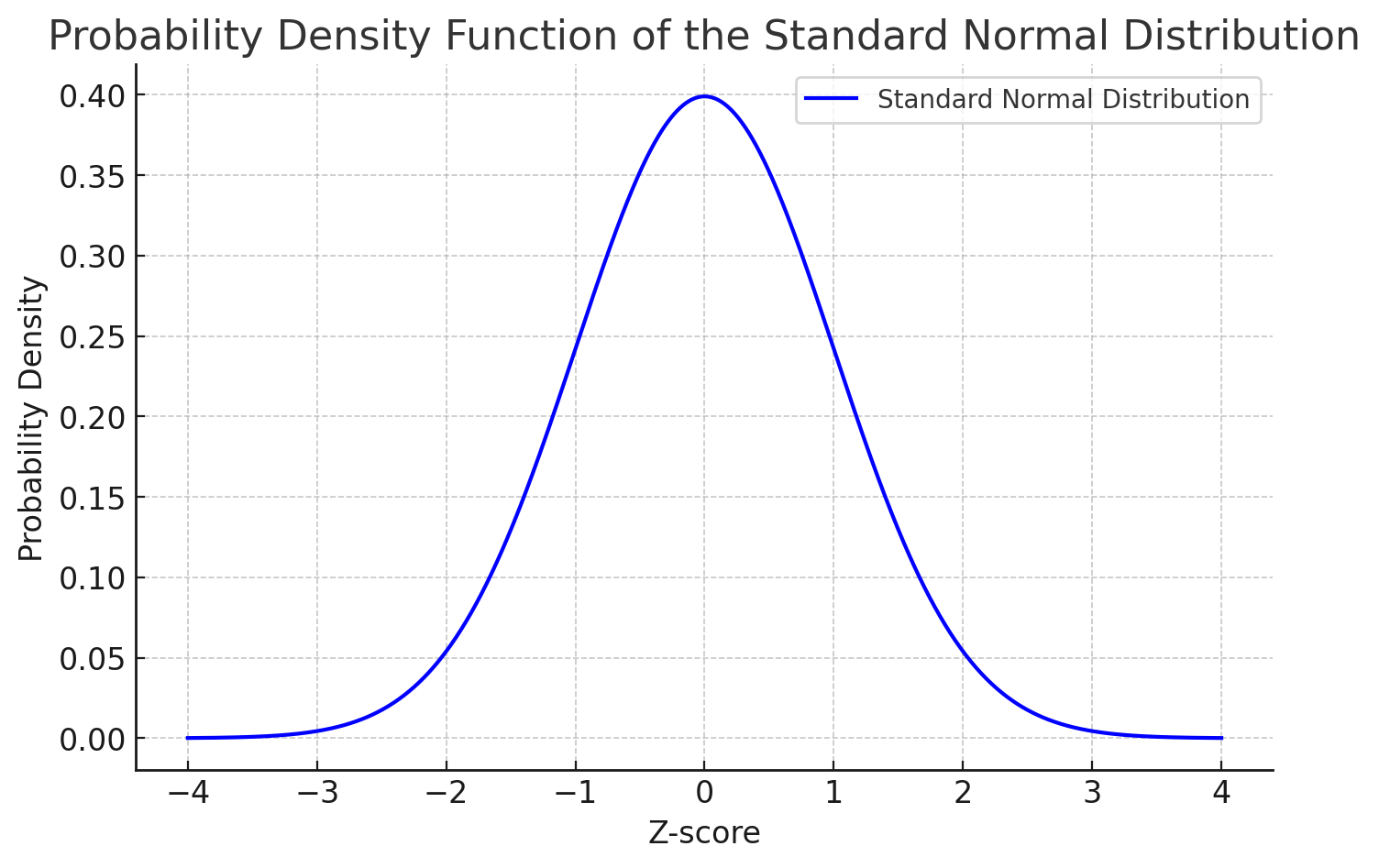
Let’s take a brief aside to define statistical significance. Your correlation analysis above was made up of a coefficient and a p-value. The coefficient tells you how “strong” the association between the two variables is (in the specific case above, how much of the variance in democracy we can explain with GDP per capita). The p-value is a number reflecting the probability that the “true” value of this coefficient is zero, given your data and model. If this number is very low, that probability is low. If it is high, that probability is high. A threshold of p<0.05 is usually the threshold for statistical “significance” but is an arbitrary number and a (increasingly unpopular) convention among social scientists.

P-values might seem like a bit of magic, but they are based on an elegant principle called the central limit theorem. This is the idea that when we take multiple, random draws from any distribution, and then plot the mean of those draws (called the sample mean) the distribution will converge to a normal distribution. Let’s quickly look at it in ChatGPT.

Prompt tips:

* Plot a uniform probability distribution with mean 1.
* Take 100 draws from this distribution and calculate the mean. Do this 1000 times and keep the mean values in a table.
* Plot a histogram of the sample means.
* Plot a Poisson probability distribution with mean 1.
* Take 100 draws from this distribution and calculate the mean. Do this 1000 times and keep the mean values in a table.
* Plot a histogram of the sample means.

Those histograms should look very alike! The central limit theorem allows us to ask how likely is it that we would get this “strong” an association between two variables if the “true” association (i.e. the mean of the “true” probability distribution that produced our association is 0). We can do this by comparing our association with what is called a standard normal distribution. If it is a long way from the center (mean) of the distribution for the “null hypothesis” (the postulation that the true association is zero) then the probability of getting this “strong” an association is small, if it is close to the center, the probability is higher.



This is the basic reasoning behind p-values and “statistical significance”. There are many other ways to assess hypotheses, but p-values do give us a feel for how likely or unlikely it would be to find such an association – given our model and data – if the true association is zero.

For now, let’s interpret p-values to represent the probability that the true relationship in our data is 0, given our model and data. Another way to think about it is the probability that this strong an association could emerge randomly, if the true result is zero.

P-values do not overcome all the hurdles to causal inference (and have their own pitfalls). A low p-value does not mean your model is good, your, data are measured well, that you have no omitted variables (alternative explanations), that reverse causality is not an issue, or that you can generalize to a broader population. A low p-value also does not mean that the association is strong (with “big” data it’s easier to find weak but significant relationships). Please consult the Kellstedt and Whitten book for more information about p-values.

Let’s start dealing with alternative explanations below.

## Multivariate associations

Now we start using “regression” models. These models are often used in the social sciences because they allow researchers to include other variables and “control for” the associations between those alternative variables and the dependent variable. We will start with a bivariate model, which gives the same results as our correlation analysis. We will then start adding control variables.

You should be aware that there are many diagnostics that can be used to assess whether the model is “robust”. These include checking for “heteroskedasticity” (which is the case the variance of the residuals is not constant, often within countries or over time), “autocorrelation” (where residuals are correlated with each other), and “multicollinearity” (where independent variables are correlated with each other). We are not going to talk or assess these here. To my mind the main “robustness” tests that we want to run are thinking about whether these models help us to cross hurdles to causality, but these tests are important things to consider when using regression in your own research.

Questions to think about as you examine the results:

* Are GDP per capita and polyarchy positively or negatively associated with each other?
* What happens to the association between GDP per capita and polyarchy when we add oil wealth as a control variable?
* What happens to the association when we include other control variables?
* What other control variables might you want to include in this analysis (if any)
* Are there remaining issues that still present “hurdles” to causality that we have not crossed?

Prompt tips:

* Run a linear regression model of the polyarchy score on GDP per capita and output the results as a nice table with stars for levels of statistical significance. Call this model 1.
  + From this model, what is the expected polyarchy score for a country with a GDP per capita of $500?
  + From this model, what is the expected polyarchy score for a country with a GDP per capita of $10,000?
  + From this model, what is the expected polyarchy score for a country with a GDP per capita of $20,000?
  + Plot how the expected democracy score changes for levels of GDP per capita from this model.
  + Produce a marginal effects plot for all variables but not the intercept for this model.
* Add oil wealth as a control variable to the first model and output the results as a nice table with stars for levels of statistical significance. Call this model 2.
  + Plot how the expected democracy score changes for levels of GDP per capita from this model and add the line from model 1.
  + Produce a marginal effects plot for all variables but not the intercept for this model.
* Add population and year as control variables and output a table in the same format. Call this model 3
  + Produce a marginal effects plot for all variables but not the intercept for this model.
  + Plot how the expected democracy score changes for levels of GDP per capita from this model and add the lines from model 1 and model 2.

## Other aspects to explore

You might still have a number of concerns about your findings. Some of those concerns might be related to our hurdles to causality. For example, you might wonder if reverse causality is an issue because democracy and GDP per capita are measured in the same year. In this case we don’t know if changes in GDP per capita occurred before, or after, changes in democracy. One way to do this is with a *lagged* independent variable. In this case, instead of using the GDP per capita variable measured in the same year as the democracy score, we can use the GDP per capita measure in the year before the democracy score. This way we can be more confident (although not 100% sure) that changes in our independent variable occur before changes in the dependent variable. Let’s do this.

Prompt tips:

* Create a new variable called gdp\_per\_capita\_lagged\_1, which is the value of GDP per capita in the year before the observation year. Do this treating each country as a separate group.
* Show the output for one case.
* Re-run model 3, but used the lagged GDP per capita variable instead of the GDP per capita variable. Call this model 4.
* Output the results as a table.
* Plot the expected democracy score for values of GDP per capita. Add the lines for GDP per capita from models 1, 2, and 3.

We might also be concerned that there is something special (that we can’t observe) about individual countries that helps explain both their levels of development and their levels of democracy. For example, some deep historical factors might explain why South Korea is a democracy today, while North Korea is not. This is the argument of a series of articles by the Nobel Prize winning economists Daron Acemoglu and James Robinson.

We can use a method called “fixed effects” to account for this concern. Very basically, a fixed effect involves including a binary (i.e. 1/0) variable for each unit in our sample. For us, this means including a variable that indicates whether the country is Brazil (Yes/No), Argentina (Yes/No), Peru (Yes/No) which “controls” for features specific to that country. Put differently, it removes the variation across countries and leave behind only the variation in our independent variables that occurs within countries. This is quite handy for causal inference, because we are often interested in the over time variation within units, holding things constant (fixed effects models have costs too, please ask if you are interested). You can use a fixed effect for anything (region, location, year) but it’s most common to use fixed effects to control for “unit” level features. In our case the unit is a country.

Prompt tips:

* Re-run model 4 including fixed effects for country\_name. Call this model 5.
* Output the results as a table
* Plot the expected democracy score over values of GDP per capita. Add the lines from models 1, 2, 3, and 4 to this plot.

Questions:

* What do you notice about the relationship when we include fixed effects? What might this mean?

## More advanced analysis

### Binary dependent variables

Quantitative methods are very flexible. The above models are very basic (but often very useful) in the sense that they assume that our dependent variable can be measured on a continuous scale and that the relationship between the independent and dependent variable(s) is/are linear (i.e. can be summarized by a line).

Maybe our dependent variable is not a continuous variable. Maybe it is a binary variable and what we want to know is how does the probability of an outcome changes with changes in an independent variable. In our example, maybe we want to measure democracy as something that a country is or is not. In this case we can use a “logit” or “probit” model to estimate how changes in an independent variable associate with changes in the odds (or probability of something happening). Instead of modelling the relationship as a line, logit models allow us to model them as a logistic curve.

Some prompt tips for a logit regression:

* Create a new variable called democracy\_binary, which is 1 where the democracy score is above 0.5 and 0 if it is below.
* Run a logistic regression model with democracy\_binary as the dependent variable, and GDP per capita and oil wealth as the independent variables.
* Output this as a table.
* Plot how the probability of democracy changes for levels of GDP per capita from this model.
* Produce a marginal effects plot for GDP per capita and oil wealth.

### Non-linear relationships

All of the models that we have been working with assume that the relationship between two variables can be explained by a line, or, more specifically, a linear equation in the form:

But maybe this is not our hypothesized relationship. Maybe (as many of you did in class) you expect that the relationship between two variables is best explained as a curve. Happily, there are quadratic and polynomial equations that we can use to test for nonlinear relationships. These are in the form of or higher order polynomials. To test for these in our data, all we need to do is add a squared term for our variable of interest (or a squared and cubed term for a more complex relationship). Let’s try that now for GDP per capita and democracy (the continuous measure).

Prompt tips:

* Create a new variable called “gdp\_per\_capita\_squared”, which is the square of the GDP per capita variable.
* Run model 3 again but add GDP per capita squared.
* Output the results as a table.
* Plot how the expected democracy score changes for values of GDP per capita, add the lines from model 1, 2, and 3 to the plot.

## Interaction terms

In other cases, we might expect that the impact of our dependent variable depends on the value of another variable. For example, we might expect that GDP per capita only causes democracy at low levels of oil wealth. When oil wealth is low, we get democratizing effects of development, when it is high, increased development has a lesser impact, or even a negative impact.

To test this kind of hypothesis we can use an interaction term. This is getting a bit more complex now, but to test whether the impact of our independent variable depends on the value of another, we can multiply two variables together and include them all in the regression equation (the GDP per capita term, the oil wealth term, and the GDP per capita X Oil wealth term).

Prompt tips:

* Create a new variable called gdp\_oil\_interaction which is GDP per capita multiplied by oil wealth. This is an interaction term.
* Include this interaction term in model 3.
* Plot two lines, one how the expected democracy score changes with GDP per capita when oil wealth is low and one where it is high. Add the lines from models 1, 2, and 3, to this plot.