POLS2044 WEEK 9
Two-variable regression modelling and interpretation

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

In Week 9 of POLS2044 we will be focusing on ways of testing hypotheses related to the relationships between two variables in a regression framework. This week builds directly on the last week's hypothesis tests about difference of means and correlations.

This week I have three main goals. First, I want to (1) introduce you to regression analysis, (2) explain why it is so popular in the social sciences, and (3) explain its assumptions. Second, I want to work through how we might interpret important regression results. Third, I want to give you an opportunity to practice using regression and gauging the statistical significance of the relationship between our independent and dependent variables.

Reading notes

There is one assigned reading, Chapter 9 from the textbook. As in the last few weeks, there are myriad formulae in the readings that may be a bit difficult for some. Remember, I am interested in you being able to understand why we are using different metrics, what elements go into these metrics, how these metrics change as the values of the inputs change, and what their results allow us to conclude about what we care about (answering our research questions).

PART 1: Bivariate regression

Today's motivating questions

Why and how do we run a bivariate regression?

Why and how do you interpret regression results (both yours and others)?

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?

Ceterus paribus assumption

Latin for "other things equal."
Also short for "all other things being equal."

Regression helps us control for other factors to better isolate the effect of the variable we care about.

Estimating the relationship between X and Y

The work horse model: $y = \alpha + \beta x + \varepsilon + \epsilon$

Where:

Y is the outcome you are trying to explain.
X is the main explanatory 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.

Estimating the relationship between X and Y

It can be shown that the least-squares estimators of α and β, which we call $\hat{\alpha}$ and $\hat{\beta}$, are given by

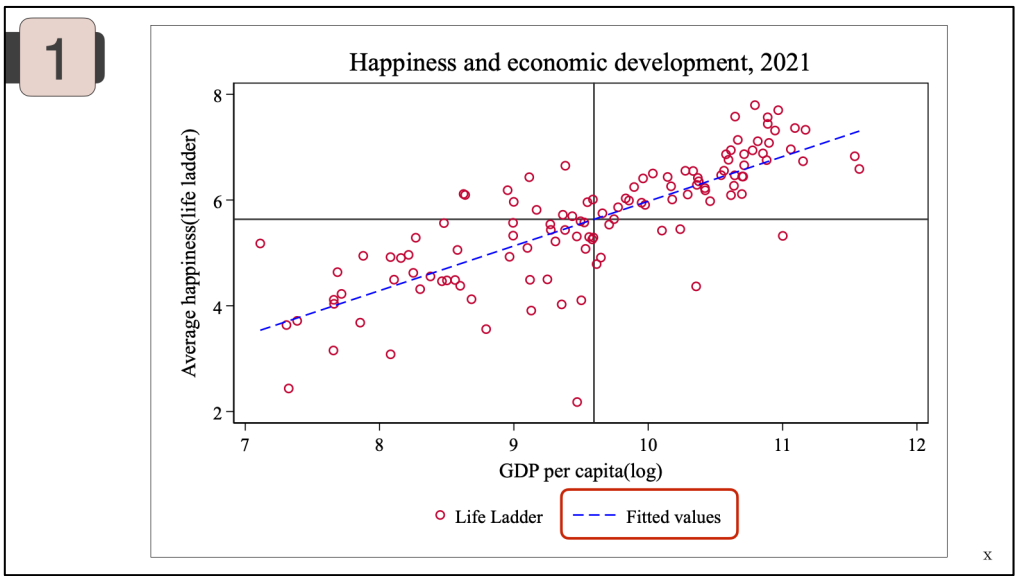
$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

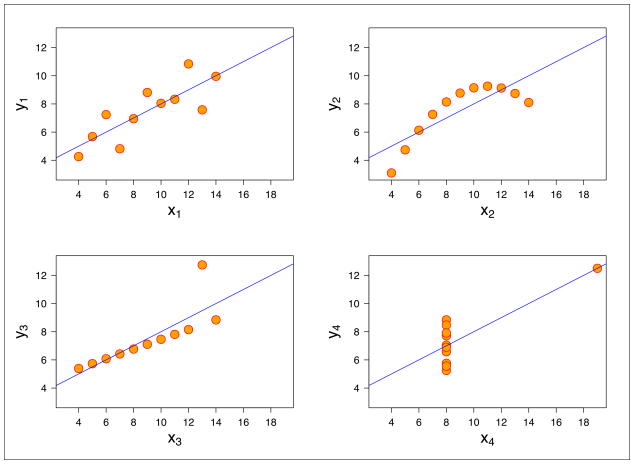
where

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{and} \quad \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

Today's data example is from the World Happiness Report (2023)



Anscombe's quartet



Almost identical descriptive statistics but very different underlying value distributions.

Two-tailed hypothesis testing of slope coefficient.

Here are my regression results for happiness regressed on GDP: = 0.845; se= 0.060.
My theory's main empirical hypotheses are:
H0 (null hypothesis): = 0
H1 (alternative hypothesis): 0
To test these hypotheses, we do a t-test, in this case we set *null* = 0.
$t = (0.845-0)/0.06 = 14.083$.
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: Bhat plus or minus t * s.e.(Bhat); alpha hat plus or minus t*se(alpha)

My slope's confidence interval is [0.726, 0.963].
My intercept's confidence interval is [-3.627, -1.324].

Crucial regression assumptions: normal population residuals

1. The **population stochastic component** is distributed normally with a mean of 0 and a variance of sigma squared.
This allows us to use the t-table to make probabilistic inferences about population regression given sample regression.
It also assumes that the expected **population errors** are **not biased** one way or another.
The **variance** is assumed to be **the same across values** of our Y and X.

Heteroskedastic standard errors

Crucial regression assumptions: Autocorrelation

2. There is no **autocorrelation** in the population random error terms.

This is important to think about when you have multiple observations of the same units (e.g., people or countries) often through time-series data.

Crucial regression assumptions: **Sample X**

3. Independent variable (X) values measured without error.

This allows us to assume that all deviations from the expected values are due to the population stochastic component (*ui*) rather than measurement error.

Crucial regression assumptions: Model specification

4. No causal variables left out and no non-causal variables included in our model.

5. The relationship (beta) between Y and X stay the same across all values of X.

6. Our independent variable must vary.

7. There must be more cases than parameters (e.g., alpha and beta)

Regression takeaways

Ordinary least squares regression is about fitting a line that minimises the (squared) distance between sample values and the line.

A basic regression provides us with two important estimates:
(1) the slope of the line summarising the relationship between X and Y
(2) the intercept (expected value of Y when X=0).

Multiple regression enables us to control for other factors that might understate or overstate our X-Y relationship if we do not include them.

LECTURE PART 2: Interpreting regression results

Interpreting regression output

Now we have some regression results, what do we do with them?

How do we interpret this table?

Make sure the number of observations makes sense

Remember the dependent variable: Happiness

Find and analyse the independent variable: GDP

Remember the standard limit theorem…

Interpret the estimated coefficient and standard error

Interpret the intercept's results

Interpret the regression statistics: Multiple R

Interpret the regression statistics: R-square

Interpret the regression statistics: Adjusted R square

Interpret the regression statistics: Standard error

Interpret the regression statistics: F statistic

LECTURE PART 3: How do we read a published regression table?

The three S's

Significance
Sign
Size

Long 2016: 10 things to know about reading a regression table

Another example: methods anxiety

> The literature suggests that statistics anxiety negatively affects course performance. (Zeidner 1991; Onwueg- buzie and Seaman 1995; Zanakis and Valenza 1997)

Dependent variable: An anxiety scale

> "AMAS-C is a 49-item self-report measure designed to assess chronic, manifest anxiety in the college student population. The students had to respond to the AMAS-C on a nominal true/false scale. The construct validity of the scale that was obtained through a factor analysis revealed four subscales: worry anxiety, physiological anxiety, test anxiety and social anxiety."
> Papnastasiou and Zembylas (2008: 160)
>
> Note the scale is reversed in some models so that higher values suggest lower anxiety levels.

Independent variables: student survey questions

> 32 questions measured on a seven-point Likert Scale ranging from 1 (strongly disagree) to 7 (strongly agree). These questions are combined into five sub scales:
>
> Usefulness of research to students' profession
> Research anxiety

Positive attitudes to research
Relevance of research to students' personal lives
Research difficulty
Papnastasiou and Zembylas (2008: 160)

Sample

472 students enrolled in an undergraduate methods course for education students at the University of Cyprus from 2002 to 2005.

What population do you think this sample is part of?

Anxiety results

Separate analyses found gender differences in anxiety but not in difficulty.

Grade results

Side note: Why standardise coefficients?

Most of the time our independent variables use different measurement units.
This makes direct comparison of regression coefficients difficult.
Standardising the coefficients puts the coefficients on the same scale, which aids comparability.
This comes at a cost of easily understanding one unit change in the independent variable.

Side note: You can also standardise the variable instead of the coefficient.

Today's motivating questions

Why and how do we run a regression?
Why and how do you interpret regression results?

Important Week 9 terms

Directional and non-directional hypotheses
Parameters
Parameter estimate
Population error term
Residual
Model standard error
R-squared
Stochastic
T-ratio
T-test
Slope
Intercept
Ceterus paribus
Autocorrelation

WEEK 9 WORKSHOP

This week's workshop is geared towards (1) thinking about potential causes of happiness, (2) explore how we might conduct hypothesis tests on a particular cause of happiness' slope coefficient, and (3) practice interpreting the statistical results.

When you are done with all parts of this week's workshop activities, please submit **both your spreadsheet and your answers** to the ***questions*** below to the Workshop submission link.

I am not including explicit steps in this week's workshop document as they did tend some students to not read the entire text and skip to the steps in previous weeks. The steps did not really make sense unless you read the rest of the text.

WORKSHOP Part I. What causes happiness?

Today we will all be analysing the same dataset on happiness from lecture. If you are curious, you can find the report and the statistical appendix online using the information in the footnote.[1] This report includes the following variables, which are also described in the spreadsheet in the "data description" tab:

Variable	Description
country	Country name
happiness	This is a happiness score (or subjective wellbeing) from a Gallop World Poll question: "Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?"
gdp	GDP per capita in purchasing power parity (PPP) at constant 2017 international dollar prices are from World Development Indicators (WDI, version 17, metadata last updated on Jan 22, 2023).
socialsupport	Social support (or having someone to count on in times of trouble) is the national average of the binary responses (either 0 or 1) to the GWP question "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"
life expectancy	Healthy life expectancies at birth are based on the data extracted from the World Health Organization's (WHO) Global 1 Health Observatory data repository (Last updated: 2020-12-04).
freedom	Freedom to make life choices is the national average of responses to the GWP question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"
generosity	Generosity is the residual of regressing national average of response to the GWP question "Have you donated money to a charity in the past month?" on GDP per capita.
corruption	Corruption Perception: The measure is the national average of the survey responses to two questions in the GWP: "Is corruption widespread throughout the government or not" and "Is corruption widespread within businesses or not?" The overall perception is just the average of the two 0-or-1 responses. In case the perception of government corruption is missing, we use the perception of business corruption as the overall perception. The corruption perception at the national level is just the average response of the overall perception at the individual level.

[1] Helliwell, J. F., Layard, R., Sachs, J. D., Aknin, L. B., De Neve, J.-E., & Wang, S. (Eds.). (2023). World Happiness Report 2023 (11th ed.). Sustainable Development Solutions Network. < https://worldhappiness.report/ed/2023/#appendices-and-data > Data descriptions are from this report's Statistical Appendix < https://happiness-report.s3.amazonaws.com/2023/WHR+23_Statistical_Appendix.pdf >

positiveaffect	Positive affect is defined as the average of three positive affect measures in GWP: laugh, enjoyment and doing interesting things in the Gallup World Poll waves 3-7. These measures are the responses to the following three questions, respectively: "Did you smile or laugh a lot yesterday?", and "Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Enjoyment?", "Did you learn or do something interesting yesterday?
negativeaffect	Negative affect is defined as the average of three negative affect measures in GWP. They are worry, sadness and anger, respectively the responses to "Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Worry?", "Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Sadness?", and "Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Anger?"

As you can see there are eight potential explanatory variables. I analysed *GDP* in class, so there are seven remaining potential explanatory factors.

Each student in your group will be choosing a different potential explanatory variable. Please have every student in your group roll the die and then return it to me. The number you get is the independent variable you will be exploring.

Die side	Variable
1	*socialsupport*
2	*life_expectancy*
3	*freedom*
4	*generosity*
5	*corruption*
6	*negativeaffect*

Now that you have your variable, please look at its description either above or in the data-description sheet. Think about a research question and what causal mechanisms may lead your variable to affect a person's overall happiness.

1. ***What is a research question linking your causal factor to the outcome?***

For example, for GDP I might write "Do more developed countries have happier citizens than less developed countries?" Now try and outline a plausible causal mechanism to get over the first causal hurdle. For example, for GDP I might argue that more developed countries have more limited work weeks and higher pay, which gives people more time to pursue their hobbies. I am not saying that this is the most plausible causal link I am just saying this might be one I think of off the top of my head.

2. ***What is a plausible causal mechanism linking your causal factor to happiness? Why do you find it plausible?***

Next, we need to think about the observable implications of our argument. If I think that more GDP gives people higher average wages and more regular work hours, then as GDP increases, all else being equal, average happiness should also increase. My alternate hypothesis would be "As a country's GDP increases, its average citizen happiness also increases." My null hypothesis would be: "A country's GDP has no effect on its average citizen happiness."

3. ***What is your null hypothesis and alternate hypothesis? Why did you word these hypotheses the way you did?***

Great! Now that we have a clear idea of what our research questions are, what plausible causal mechanism we think leads our independent variable to have on our dependent variable, and our testable hypotheses, we can productively analyse the data and see if our intuition is supported by the evidence.

WORKSHOP Part II: Running a bivariate regression

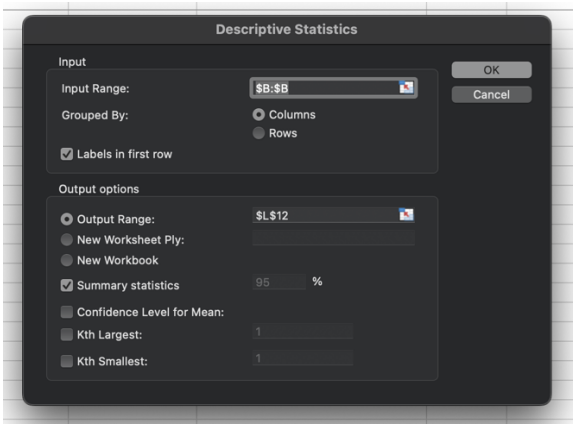
Now comes the part where we must turn to the spreadsheet. You have a dependent variable and a potentially significant independent variable. What are we going to find? Let's see! Open the spreadsheet and go to the "data" sheet. Here you will see data on 118 countries in 2021. I would have used 2022 data, but there were fewer countries measured in 2022 and I wanted to get as large a sample as possible.

Find your independent variable and the dependent variable, happiness. I showed you summary statistics in lecture for GDP and happiness, but when you are faced with a new variable, I always recommend learning a bit more about how it is distributed before turning to regression.

Here is the descriptive statistics for happiness. The mode is unavailable because no two countries have the same average happiness score.

happiness	
Mean	5.629586258
Standard Error	0.10602944
Median	5.7812784
Mode	#N/A
Standard Deviation	1.151774537
Sample Variance	1.326584585
Kurtosis	0.015021831
Skewness	-0.513999589
Range	5.6155684
Minimum	2.1788094
Maximum	7.7943778
Sum	664.2911784
Count	118

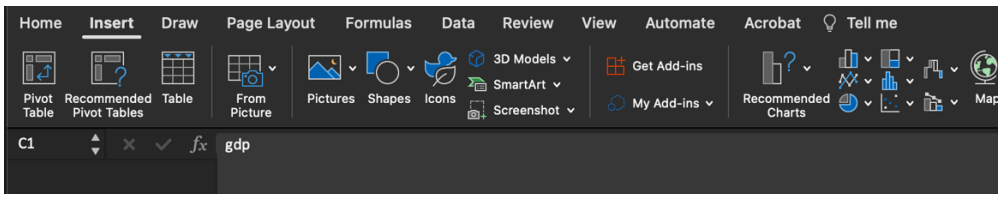
Here are two that we will focus on. First, generate the descriptive statistics using the ToolPak addon. Here are the settings for the happiness variable.

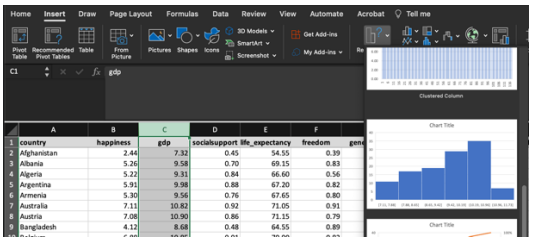


Here is my output for GDP.

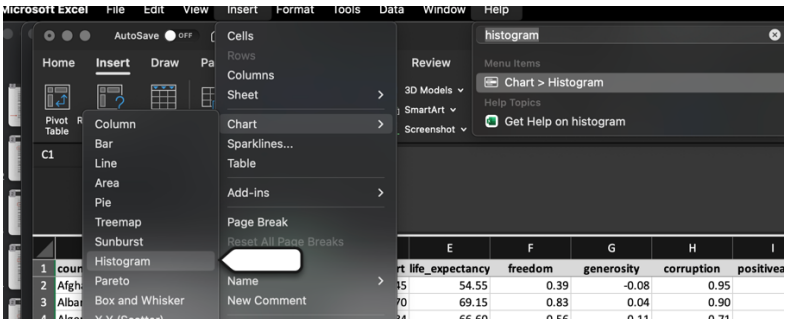
gdp	
Mean	9.58634335
Standard Error	0.099023675
Median	9.60410835
Mode	#N/A
Standard Deviation	1.075672441
Sample Variance	1.1570712
Kurtosis	-0.750888966
Skewness	-0.384689455
Range	4.4593607
Minimum	7.1121373
Maximum	11.571498
Sum	1131.188515
Count	118

Now, let's generate a histogram to see how your independent variable is distributed. This is one instance, where I have found the built-in graph to be easier to use than the ToolPak version. First highlight your variable's column, go to the Excel "Insert" tab, recommended charts, then choose histogram.





Your version of Excel may vary a bit in how its menus are laid out. I have found that if you click the Help menu then enter "histogram" it will show you where you generate a histogram.



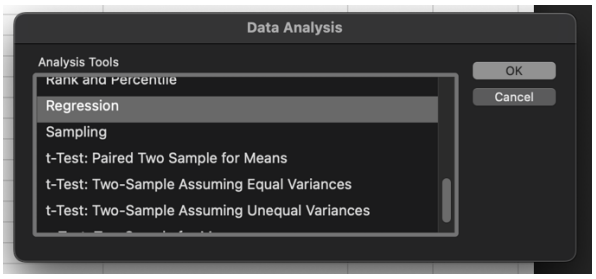
Now that you have a histogram, you can change the number of bins, labels, etc to make it as clean and easy to interpret as possible. If you are interested in a video walkthrough, I have found the following video potentially useful for you: https://youtu.be/yh5ihdHwmTk?si=ziL-x_UAaS4Isuzq.

4. ***Place your summary statistics and histogram in the "analysis" sheet of this week's spreadsheet.***

I am asking students to place their descriptive statistics, histograms, and regression results into the "analysis" sheet to make it easier for you to analyse the results and describe them below. It should also help us to see your results more clearly.

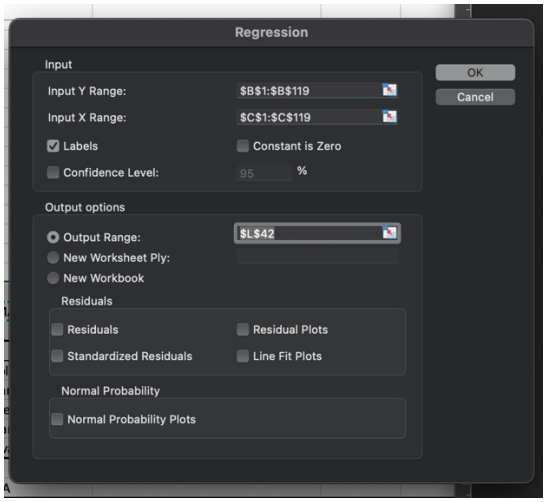
5. ***Did you find anything interesting about the distribution of your independent variable? If so, what was it?***

Hopefully, now you have a clearer idea of what your raw data looks like. Now it is time to run a regression! Open the ToolPak and select regression.



Then select the happiness column data for the "Input Y Range" and your independent variable column data for your "Input X Range". Make sure to include the variable name in column 1 by clicking the Labels box. This shows Excel what the variable name is to put in the regression results.

Here is a screenshot for my selections to regress happiness on GDP.



6. ***Please cut and paste your regression results into the "analysis" sheet.***

Now let us turn to interpreting the results.

WORKSHOP Part III: Interpreting the results

The hard part of trying to get the ToolPak stuff to work is thankfully behind us. In this section, we are going to focus on interpreting our results. I am not going to try and cover every cell of the results because we really care about a few things in particular. I also have not been able to find out why we are getting two sets of lower and upper 95% confidence intervals, so let us just set that aside for now…

Remember that in lecture I told you to focus on significance, sign, and size. Let us take each one in turn.

First, look at the coefficient and standard error of your independent variable. Hopefully, you will see that the coefficient is more than twice the size of the standard error. This is a quick and dirty way to see in others results whether the coefficients are statistically significantly different from zero. Then look at the t-statistic and the associated p-value for your variable.

7. ***Given your t-statistic and p-value, do you conclude that you can reject the null hypothesis in favour of the alternate hypothesis? Why or why not?***

In the interests of transparency, there are five variables that are statistically significant and one that is not. Do not worry if you are not able to reject the null. Just think about whether switching to a one-tailed test with a directional hypothesis would enable you to reject the null.

Second, we are going to look at the sign of the coefficient.

8. ***Is your coefficient's sign positive or negative? What does this tell you about how happiness changes as your independent variable changes?***

Third, we can look at the size of the coefficient. This is a bit out of scope for today as we do not have other results to compare it to. However, think about what this coefficient means. For instance, GDP's coefficient is 0.86 (s.e. 0.059). What this means is that the slope of the regression line goes up by 0.85 units of happiness for every one unit change in logged GDP

per capita (PPP). These are not intuitive metrics, but if you standardise the coefficients or the values or have more meaningful scales of your IV and DV, this is worth looking at. We are going to spend more time on predicted values in later weeks.

Next, we want to look at the intercept.

9. ***What is the value of your intercept? Is it statistically significant? What do you think the coefficient means substantially?***

For instance, for the GDP regression, my intercept coefficient is -2.635 (s.e. -0.569; t-stat -4.634; p. 0.000). I conclude that (1) the alpha is statistically significant and that (2) the expected value of happiness when GDP is 0 is -2.635. Now this is impossible in the real world as no country has no economic activity (GDP) within its borders nor can people have negative values on the happiness scale. What it tells us is where the regression line intersects the Y-axis.

Finally, let us look at the overall regression statistics.

10. ***How many observations are in your results? Do you think your results would change if this number increased to either (1) every country in the world or (2) only a sample of 30 countries? Why or why not?***

11. ***What is your R-squared? What does this value tell us substantially about how much of the variation in happiness is explained by your independent variable? How might it compare to the R^2 for my GDP equation (0.65)?***

12. ***What is your F-statistic? Is it significant? What does this conclusion mean about the explanatory power of your model?***

Since this model only has one predictor, it is interesting to look at the Multiple R, given our lecture discussion.

13. ***What is your Multiple R? Why is it identical to another statistic we calculated last week, and what is that statistic?***

Wow, we have covered a lot of territory over the last few weeks. Hopefully, you have become a bit more comfortable with hypothesis testing and conducting hypothesis tests on regression coefficients. Additionally, I hope you can see the number of different elements we have covered so far come together in ordinary least squares regression and how our theoretical interests and research questions drive us towards ways of modelling the world.

When you are done with all parts of this week's workshop activities, please submit **both your spreadsheet and your answers** to the ***questions*** above to the Workshop submission link **(item 9.1**).

Part IV: For the data ninjas (very optional)

So, you still have time left before the end of workshop and are up for another challenge? This section is for you.

First, it is important to note that the exercises above were minimal efforts to analyse existing data.

The *World Happiness Report, 2023* (https://happiness-report.s3.amazonaws.com/2023/WHR+23.pdf) does run and include its own regression results.

For instance, look at Table 2.1 (page 38). Look at the estimated coefficient of your independent variable in the contest of the last model on the right labelled "Cantril Ladder (0-10)". As you can see all the variables we looked at today are in this model.

14. ***Are their results for your IV comparable (in significance, sign, and size) to yours? Why or why not?***

15. ***How does their sample differ from yours?***

16. ***What regression statistics do they include in Table 2.1, and can you interpret them? How do they differ from your regression statistics?***

Still here? Why not try and run your own multiple regression on the 2021 data.

17. ***How (and why) does your multiple regression results differ from those in Table 2.1?***

Glutton for punishment? There are myriad other works looking at subjective well-being and happiness.

18. ***Find an empirical article on happiness that includes a survey. Analyse their results and compare it to your results and the World Happiness Report. What do they do similarly or differently to the World Happiness Report?***