INFO370 Problem Set: Linear Regression

February 5, 2023

Instructions

This problem set revolves around linear regression, in particular interpretation of linear regression results. It contains three parts:

1. Estimate a simple regression model and compute a few simple figures based on the results
2. Estimate and understand a multiple regression model
3. Interpret regression results in the literature

- Please write clearly! Answer questions in a way that if the code chunks are removed from your document, the result is still readable!
  - All substantial questions need explanations. You do not have to explain the simple things like "how many rows are there in data", but if you make a plot of global temperature, you should explain what do you see there.
  - Write explanations in markdown cells using markdown. *Do not* write those as code comments! Do not print them using `print()` command.
- Do not make code lines too long. 80-100 characters is a good choice (python standard is 80). Your reader may not be able to follow all the code if the line is too long–most of us are using small laptop screens! You can split code line using the backslash escape symbol as the last character on line.

1 When will we hit 2C of global warming? (46pt)

In this question you are working with HadCRUT dataset (Hadley Centre/Climatic Research Unit Temperature). It is one of the most widely used datasets to asses global temperature trends. It is maintained by Met Office (UK Government). The data source includes both annual and montly time series. Copyrighted under Open Government License. It can be downloaded from Metoffice download page. Here we use the annual version (one temperature record for year), you can also get it from canvas. It contains variables:
Time year, 1850-2022
Anomaly (deg C) Global temperature anomaly (deg C) wrt 1961-1990 average
Lower confidence limit (2.5%)
Upper confidence limit (97.5%)

(As you see, the variable names are long and complex.)

1. (1pt) Load data. Ensure it looks good.
2. (2pt) Make a plot of the temperature anomaly for all years in the dataset. Describe what you see–how has the global temperature behaved through the last 170 years?

   Hint: you may want to rename the data variables to something simpler.

In 2015, the countries joined the Paris Agreement to limit global warming to 1.5C above the pre-industrial levels (1850-1900). It is believed that staying within 1.5C limits will avoid the worst effects of warming. However, the second-best will be to stay withing 2C limits. Your next task is to find when will we pass these thresholds if the current trends continue.

1. (2pt) Compute the baseline anomaly (for 1850-1900).

   Hint 1: about -0.35C.

   Hint 2: "Baseline anomaly" is the mean anomaly for 1850–1900.

If you made the plot correctly, you see that the global temperature shows both ups and downs through the time period. But the last ~ 60 years seem to follow a fairly constant upward trend. So our next step is to look at the last 60 years of data only (1963-2022), compute the trend, and see when will the trend hit +1.5C and +2C above the pre-industrial baseline. In the questions below, use only the last 60 years of data!

2. (3pt) Make a plot that shows the last 60 years of data, overlied with the regression line. In your opinion–does the regression line describe the trend well?

   Hint: you *can* add regression line to the plot with matplotlib, but I recommend you to check out `sns.regplot` instead.

3. (2pt) But we have a problem with these data–the anomaly figures reported there are computed w.r.t. the 1961-1990 period. Create a new variable that describes the temperature anomaly *above the pre-industrial baseline* that you computed above in Q1.1.

   Hint: it should be -0.06 in year 1850.

4. (4pt) Estimate a linear regression model where you describe the anomaly (measured above baseline in q3) as a function of years. How large is R^2? What does this number tell you?

   Note: in terms of interpeting the coefficients below, I recommend to use years relative to 2000, i.e. $1999 = -1$, $2000 = 0$, $2001 = 1$ etc.

5. (4pt) Interpret the coefficients. What does the slope mean? Is it statistically significant?
6. (6pt) What does the intercept mean if

   (a) you use years as is (1999, 2000, 2001, ...)?

   (b) you use years relative to 2000 (-1, 0, 1, ...)?

   Is it statistically significant?

7. (4pt) Use the intercept and slope, and compute how much above the pre-industrial baseline is the world now, in 2022.

8. (9pt) Now use the slope and intercept to find which year will the trend hit 1.5C, and when will it hit 2C.

   Hint: we should hit 1.5 degrees ~ 2038, and 2C around the time of your retirement.

9. (4pt) Now think about these two threshold years you computed. What do you think, might the thresholds be actually crossed earlier or later than what you computed, or maybe not at all? What does it depend on? Can you improve your model to make your predictions better?

---

2 How Is Basketball Game Score Calculated? (45pt)

James Harden playing for Rockets in 2017. Keith Allison from Hanover, MD, USA, CC BY-SA 2.0 https://creativecommons.org/licenses/by-sa/2.0, via Wikimedia Commons.

In this section you will work with basketball data. Basketball is a big business, and there is a lot of analytics collected about high-profile games. Game score is one of the popular measures of player's performance in game. But how is it calculated?

Here we look at one particular dataset about James Harden's (see photo) 2021-2022 season, downloaded from basketball-reference.com. We recommend you to be familiarize yourself with the basics of basketball, including what are *field goals*, *turnovers*, and *personal fouls* (wikipedia is a good source).

The dataset contains 30 variables, including field goals, field goal attempts, 3-point field goals, rebounds and personal fouls. See my data repo readme for reference.

The central variable in current context is *GmSc*, the game score. It is a summary performance score for the player (given he played in the game).

Here are the tasks:

1. (1pt) Load data (*harden-21-22.csv*). Do basic checks.

2. (3pt) These data also include games where he did not play. Find how many games did James Harden actually play in this season.

   Hint: there are no general ways how to do this. But just look at the data and figure it out based on what do you see there. It can be coded in different ways, but first you have to see how the relevant data looks like.

3. (5pt) Clean the data and ensure the relevant variables are of numeric type so we can use those in the regression models. It is your task to find what is wrong with the data in its present form (it is downloaded directly from basketball-reference.com), and fix these issues.

   Hint: a good way to transform text to number is `pd.to_numeric`.

   Hint 2: you do not have to convert variables you are not using.

4. (6pt) Analyze the game score *GmSc*. What is its range? Mean? Standard deviation? Which distribution does the histogram resemble?

5. (7pt) First, let's run a simple regression model explaining game score *GmSc* by field goal attempts *FGA*:

$$GmSc_{\mathtt{g}} = \beta_0 + \beta_1 \cdot FGA_{\mathtt{g}} + \epsilon_{\mathtt{g}} \tag{1}$$

   where $\mathtt{g}$ indexes games.

   Display the results and answer the following questions:

   (a) What is the interpretation of *Intercept* (β_0)?
   (b) What is the interpretation of *FGA* (β_1)? Is it statistically significant?

6. (10pt) Next, let's analyse how is game score related to field goals (*FG*) and field goal attempts (*FGA*). Estimate the model

$$GmSc_{\mathtt{g}} = \beta_0 + \beta_1 \cdot FG_{\mathtt{g}} + \beta_2 \cdot FGA_{\mathtt{g}} + \epsilon_{\mathtt{g}}. \tag{2}$$

   If done correctly, you should see results approximately 9.0, 3.0, -0.6 here.

   Answer the following questions:

   (a) What is the interpretation of *FG*? Is it statistically significant?
   (b) What is the interpretation of *FGA* (β_2)? Is it statistically significant?
   (c) How do you explain the fact that model 1 shows positive and model 2 shows a negative estimate for *FGA*? There is a very easy an intuitive explanation that everyone will understand, including those who have no clue about stats. Can you phrase it in that way?

   Hint: try to understand what exactly is the difference between interpreting slope for simple regression and multiple regression.

   (d) What is the $\mathtt{R}^2$ of the model? How does it compare to the model 1? What do you conclude from this comparison?

7. (7pt) Now include all the independent numerical variables, i.e. *FG*, *FGA*, *3P*, *3PA*, *FT*, *FTA*, *ORB*, *DRB*, *AST*, *STL*, *BLK*, *TOV*, *PF* into the model. Estimate it, and discuss the results.

   Answer the following questions:

(a) How do standard errors and `t`-values look like in this model?

(b) What is R^2 of this model? What does it tell you about how game score is calculated?

(c) What do the results tell about turnover (TOV)? Is it good or bad for the team?

Suggestion: check out patsy `Q()` quoting to include non-valid variable names.

8. (6pt) Finally, consult the game score explanation at https://www.nbastuffer.com/analytics101/game-score/. Did you recover the same formula?

---

3 Interpret regression results in the literature (9pt)

The final task involves just interpretation, no separate analysis is needed.

Table 1 displays linear regression results from Dawel *et al.* (2020). You do not have to read the paper in order to answer these questions, but it is uploaded on canvas (files/readings/-dawel+2020FiP.pdf) in case you want understand more.

The authors estimate a model

$$PHQ9_i = \beta_0 + \beta_{Age} \cdot Age_i + \beta_{Gender} \cdot Gender_i + \\ + \beta_{Education} \cdot Education_i + \cdots + \epsilon_i.$$

Variable explanations:

PHQ-9 Patient Health Questionnaire, used to detect depression. Larger number mean worse mental health.

WSAS Work and Social Adjustment Scale, measures COVID caused disruptions in life.

Age in years

Education in years

Other variables should be easy to understand, except I cannot find what exactly does "gender" mean.

Answer the following questions:

1. (3pt) Do those who have a partner have better mental health (as measured by PHQ-9)? Is the effect statistically significant?
2. (3pt) What is the effect of COVID exposure? Is it improving or worsening mental health? Is the effect statistically significant?
3. (3pt) How is Financial distress related to mental health? Is the effect statistically significant?

Finally tell us how many hours did you spend on this PS.

References

Dawel, A., Shou, Y., Smithson, M., Cherbuin, N., Banfield, M., Calear, A. L., Farrer, L. M., Gray, D., Gulliver, A., Housen, T., McCallum, S. M., Morse, A. R., Murray, K., Newman, E., Rodney Harris, R. M. and Batterham, P. J. (2020) The effect of covid-19 on mental health and wellbeing in a representative sample of australian adults, *Frontiers in Psychiatry*, **11**, 1026.

Figure 1: Table 4 from Dawel *et al.* (2020).

TABLE 4 | Linear regression models for each mental health outcome.

	PHQ-9 (*n* = 1,273, *df* = 16, 1256)	
	estimate	***p***
Constant	3.73	**<.001*** **
Sociodemographic and background factors		
Age	–0.05	**<.001*** **
Gender	0.84	**.003****
Education	–0.10	.055
Has partner	–0.47	.150
Lives alone	0.23	.628
Child at home	–0.28	.359
Any chronic disease	0.64	.052
Any neurological disorder	1.29	**.006****
Any current MH disorder	4.65	**<.001*** **
Recent adversity		
Bushfire exposure—smoke	0.26	.336
Bushfire exposure—fire	–0.40	.406
Other adverse life event	1.80	**<.001*** **
COVID-19 exposure		
COVID-19 exposure	0.24	.129
Work and social impacts of COVID-19		
Lost job	0.43	.383
Financial distress	2.32	**<.001*** **
WSAS	0.09	**<.001*** **
	R^2	**Adjusted R^2**	***F***
Model	.369	.361	**45.91*** **

**p < .017. **p < .001. ***p < .001.*

Bold indicates tests significant at p < .017.