# INFO370 Problem Set: Linear Regression

April 30, 2022

## Instructions

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

1. Estimate a simple regression model and interpret the results.
2. Interpret a multiple regression model.

- Please write clearly! Answer questions in a way, that if the code chunks are removed from your document, the result is still readable!

## 1 When will we see BOE? (45pt)

"Blue Ocean Event" (BOE) is a colloqial name for the event where there is no sea ice in the northern hemisphere. If the current trends continue, we are expected to see the first BOE around 2050 or so. Your task is to predict when will the average sea ice extent go to zero. We will be looking at two months: September and March. Northern sea ice reaches its yearly minimum in September and maximum in March.

The dataset *ice-extent.csv* can be downloaded from NSDIC webpage but I recommend to get the version on canvas (the website has the data split into several files). The data is based on satellite imagery (not necessarily in visible light). The main variables are

**year** 1978-2022
**mo** month, 1-12
**data-type** data source (satellite)
**region** "N" for northern, "S" for southern hemisphere
**extent** in M km$^2$. Extent is the sea area that is covered by ice by at least 15%.

**area** in M km$^2$. Sea ice area. Area is considered a less reliable figure than extent because of measurement issues—satellites have hard time distinguishing between open water, and water on ice.

You can play with interactive daily chart at NSIDC webpage if you are interested for more.

Your tasks are the following:

1. Load the data. Only keep the northern hemisphere data. Remove all missing extents–we will only work with extent below.

   Note: make sure you understand how missings are coded! There *are* missings in data.

2. Make a plot where you show the September, December and March ice extent over years.

   Note: you may consider *seaborn* library in addition to matplotlib if you want ready-made functionality for data plotting. Python notes will give you a crash course. But you can also just use plt.plot.

3. Estimate a linear regression model for September data where you describe the sea ice extent as a function of years.

   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.

4. Interpret the coefficients. What does the slope mean? Is it statistically significant?

5. 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?

6. Create a plot (like what you did in the lab with global temperature) where you show the regression line through the September data points.

   Comment the plot. Do you see a trend? Do you see any acceleration/deceleration (i.e. the trend is getting faster/slower)?

7. Based on your linear prediction: when will your trend line hit 0 and we will have BOE?

Note: such predictions are highly speculative as we ignore the physical reality of the polar regions.

Note 2: your figure tells when will *the trend* reach zero. The first BOE will probably happen 10-20 years earlier as the yearly extent jumps up and down around the trend.

8. Repeat these calculations for March—the month of the yearly maximum. When will the Arctic be completely ice free (if the current trends continue)?

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

2. 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. 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. Analyze the game score *GmSc*. What is its distribution? Which distribution does the histogram resemble? Range? Mean? Standard deviation?

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

   $$\mathit{GmSc}_\mathsf{g} = \beta_0 + \beta_1 \cdot \mathit{FGA}_\mathsf{g} + \epsilon_\mathsf{g} \tag{1}$$

   where $\mathsf{g}$ indexes games.

   Display the results and answer the following questions:

   (a) What is the interpretation of *Intercept* ($\beta_0$)?

   (b) What is the interpretation of *FGA* ($\beta_1$)? Is it statistically significant?

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

   $$\mathit{GmSc}_\mathsf{g} = \beta_0 + \beta_1 \cdot \mathit{FG}_\mathsf{g} + \beta_2 \cdot \mathit{FGA}_\mathsf{g} + \epsilon_\mathsf{g}. \tag{2}$$

   If done correctly, you should see results approximately 6.9, 3.4, -0.7.

   Answer the following questions:

   (a) What is the interpretation of $FG$? Is it statistically significant?

   (b) What is the interpretation of $FGA$ ($\beta_2$)? 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?

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

7. 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. 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 (10pt)

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

$$\mathit{PHQ9}_i = \beta_0 + \beta_{\mathit{Age}} \cdot \mathit{Age}_i + \beta_{\mathit{Gender}} \cdot \mathit{Gender}_i + \\ + \beta_{\mathit{Education}} \cdot \mathit{Education}_i + \cdots + \epsilon_i.$$

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

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

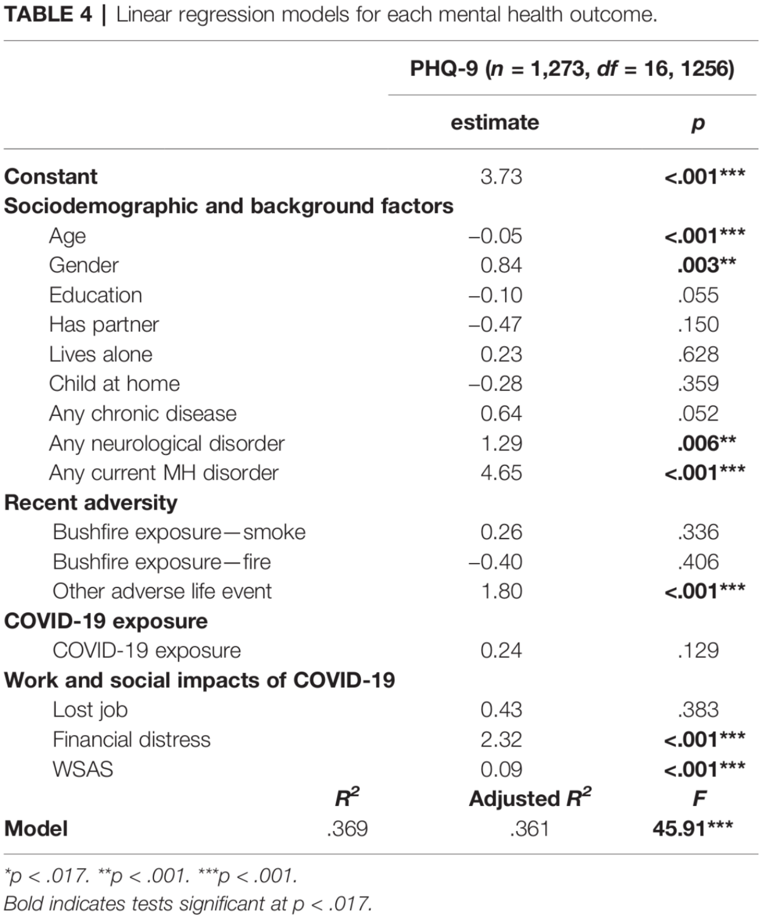

**TABLE 4 |** Linear regression models for each mental health outcome.

| | | **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.*

## 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.