MXN500: Problem Solving Task 2

Lecturer: Kate Saunders

Released: 13/05/2021, Due: 02/06/2021 11:59 pm

Submission Information

Submit your answers, code and datasets to blackboard.

The tutors will need to see your code in order to determine if you deserve the marks for a given question. So make it easy for the tutors to read and navigate. Note, the tutors will not be running your code to get your answers, so you need to provide written answers to all the questions. Three additional marks will be awarded for how readable your code is. Please upload your .csv files so the tutors can reproduce your exact numbers.

Introduction

Large scale climate drivers such as the El Niño Southern Oscillation (ENSO) are known to have an impact on Australian rainfall patterns. ENSO has three phases, El Niño, Neutral and La Niña. Generally, in the La Niña phase of ENSO conditions are cooler and wetter along the Eastern Australian coast. In comparison during the El Niño phase conditions are hotter and drier. In this problem solving task, you will be using your knowledge of regression to explore the relationship between ENSO and Australian rainfall.

Note that it is not possible to measure the strength of ENSO directly, so in regression models the Southern Oscillation Index (SOI) is commonly used to represent the strength of ENSO. The SOI is a climate index that measures the normalised pressure difference between Taihiti and Darwin. You do not need to provide units when displaying SOI on a plot axis as it is an index. ENSO is considered to be in the La Niña phase when there are sustained SOI values above 8, the El Niño phase when there are sustained SOI values below -8, and the Neutral phase otherwise. A csv file containing the monthly SOI values can be obtained from blackboard. More details about ENSO and SOI can be found on the Bureau of Meteorology website (http://www.bom.gov.au/climate/enso).

Section 1: Preprocessing, data wrangling and visualisation

Before we can start fitting models, we need to get a clear picture of our data.

Question 1.1 (1 mark) Download the `soi_data` and convert it to a long format. Show code and display the result by printing the first 3 rows.

Question 1.2 (3 marks) Create a graphical excellent visualisation of the `soi_data`, be sure to include all relevant variables.

Question 1.3 (2 marks) Adapt the code below to add a new categorical variable called `Season` to `soi_data`. Share your coded solution and show the rows corresponding to 2020. (Hint: It may help to read about the dplyr::if_else() function and %in% logical operator)

```
summer_months = c("Dec", "Jan", "Feb")
soi_data <- soi_data %>%
    mutate(Season = NA_character_) %>%
    mutate(Season = if_else(Month %in% summer_months, "Summer", Season))
```

Question 1.4 (2 marks) Create a new data set, `seasonal_soi_data` that gives the mean seasonal SOI value in each year, include named columns in this data set of `Year`, `Season` and `SeasonalSOI`. Show your code and print out the rows corresponding to 2020. (Hint: You can group by more than one variable.)

Question 1.5 (2 marks) Add a new categorical variable called `Phase` to the `seasonal_soi_data`, where the `Phase` variable indicates the phase of ENSO. If the seasonal SOI value is above 8 then the `Phase` value is `LaNina`. If the SOI value is below -8 then the `Phase` value is `ElNino`. Otherwise the value in the `Phase` column is `Neutral`. Show your code and print the updated rows corresponding to 2011. (For reference, in 2011 there was widespread flooding in Brisbane).

Question 1.6 (2 marks) Ensure the relevant variables are converted to factors in `seasonal_soi_data`. Show your code and show the levels.

Section 2: Linear Regression

Data for seasonal rainfall totals that can be used to explore the relationship between Australian rainfall and ENSO is available for download on blackboard in `total_seasonal_rainfall.csv`.

```r
total_seasonal_rainfall <- read_csv("total_seasonal_rainfall.csv") %>%
  mutate(total_seas_prcp = total_seas_prcp/10) %>%
  left_join(seasonal_soi_data)
```

Similarly to the data we used in problem solving task 1, the columns are:

- The GHCN Daily station `id` (character)
- The GHCN Daily station `name` (character)
- The `Year` of the observation (numeric)
- The `Season` of the observation (ordinal, categorical)
- The `total_seas_prcp` is the total amount of rainfall received during that Season in tenths of a mm (numeric). (Converted to mm by the mutate call above.)

For each season, we aim to fit a linear regression model to estimate the relationship between total seasonal precipitation and the mean seasonal SOI value using

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

where the $\varepsilon_i \sim N(0, \sigma^2)$ are the errors/residuals, distributed normally with a mean of 0 and standard deviation σ. As there are 7 cities and 4 seasons, this is a total of 28 different linear regression models.

Question 2.1 (1 mark) Consider the linear model for the `BRISBANE REGIONAL OFFICE` station and the `Spring` season. Complete the following sentence so that it refers to the terms in the regression model above.

For the Brisbane Station in Spring, a linear model was specified to model how the total seasonal precipitation, ..., is related to the mean seasonal SOI value, ... The parameter ... describes the rate of change in the total seasonal precipitation with an increase in mean seasonal SOI value. The parameter ... represents the total seasonal precipitation when the mean seasonal SOI value is 0.

Question 2.2 (2 marks) For the `BRISBANE REGIONAL OFFICE` and `Spring`, fit a linear model as described above using `R`. Based on your parameter estimates, write down your linear model substituting the parameter values into the equation.

Question 2.3 (1 mark) How much variability in the data is explained by this model?

Question 2.4 (2 marks) Provide a 95% confidence interval for the parameter estimates.

Question 2.5 (3 marks) Visualise the fitted values compared with the residuals, and visualise the standardised quantiles of the residuals compared with the theoretical quantiles. Discuss the validity of the underlying assumptions of linear regression.

Question 2.6 (2 marks) Interpret the results of the linear regression model and explain the physical meaning.

Question 2.7 (5 marks) Create a visualisation for each city and for each season that compares the proposed linear model to the null model. Explain your visualisation with a caption. (Hint: use `geom_smooth()`)

Question 2.8 (2 marks) Based on your visualisation for which cities and seasons would you expect there to be a significant linear relationship between total seasonal precipitation and the mean seasonal SOI value?

Question 2.9 (2 marks) Interpret your visualisation, what does it convey about differences between locations, and differences in how total seasonal rainfall responds to ENSO?

Section 3: Polynomial Lines of Best Fit

Many climate scientists hypothesise that when it comes to rainfall, that wet can get wetter, but dry can't get drier. In otherwords, how Australian rainfall responds to the different phases of ENSO may not be equal in both La Niña and El Niño phases. For this reason, one might want to check if polynomial regression better suits the data.

Question 3.1 (2 marks) For `BRISBANE REGIONAL OFFICE` in `Spring` fit a linear regression using polynomial explanatory variables of up to order 2 and the `SeasonalSOI`. Write down the equation.

Question 3.2 (3 marks) Print out a summary of your fitted model, interpret the results and explain the related the physical meaning.

Question 3.3 (2 marks) Create a prediction interval for a mean seasonal SOI value of 25 and a mean seasonal SOI value of -25. Comment on the result.

Question 3.4 (1 mark) Decide whether a linear or polynomial regression is preferred using a statistical test.

Section 4: Linear Regression with Categorical Explanatory Variables

Within the analysis so far the role of ENSO phases has been modelled solely using the SOI. If the influence of ENSO on Australian rainfall is different in different ENSO phases if follows that the Phases should interact with strength of ENSO in the linear regression.

Question 4.1 (3 marks) Consider again `BRISBANE REGIONAL OFFICE` in `Spring` and fit regression to explore how the phases influence the mean of the total seasonal precipitation. Use `ElNino` as the reference level in your categorical regression. (Do not include `SeasonalSOI`). Write down the equation.

Question 4.2 (3 marks) Print out a summary of your fitted model, interpret the results and explain the related the physical meaning.

Question 4.3 (2 marks) Given what we know, how might one expand the simple linear regression equation to include the asymmetric influence of `Phase` as an explanatory variable along with `SeasonalSOI`? Write down a new equation and explain in words what the equation means?

Readability and clarity of code (3 marks) The tutors will award marks based on how clear and readible your code is. To help the tutors with this, please make sure to comment your code for each of the different questions.