

# Assignment 1

Last updated on Jan 23, 2022

## Contents

## Introduction

This assignment is to be completed using Rmarkdown. You are to submit two documents

1. The .rmd file containing your code and write up
2. The knitted .html file from the .rmd file in (1)

There are 5 questions in the report.

You can use [this .rmd file](#) as your starting point.

You should submit this assignment by no later than February 4th at 11:59PM eastern time zone.

## Questions 1 (15 points)

You have just had an initial meeting with the head of the Emergency Department at a downtown Toronto hospital. The clinician has provided you with a sample of data from patient visiting the Emergency department over a 5 day period. They are interested in being able to predict how long a patient will remain in the ED (`visit_end_time` - `visit_start_time`)

You'll need to download this CSV file and put it somewhere on your computer

-  [raw\\_ed\\_data.csv](#)

You are going to have a follow up meeting with the clinician in a couple of weeks.

## The Data

The sample data contains every emergency department visit over a 5 day period with the following variables:

- `encounter_id`: a unique ID given to a patient encounter (an encounter is a patient entering and then leaving the Emergency Department)
- `patient_id`: a unique ID for a patient
- `arrival_time`: The date and time of the patient arrival to the ED
- `departure_time`: The date and time that a patient leaves the ED
- `admit_time`: The date and time that a patient is admitted to the hospital
- `admitted`: an indicator for whether or not a patient is admitted to the hospital.
- `los_calc`: A calculated variable measuring the time of a patient arrival to departure in hours
- `presenting_complaint`: The reason for the Emergency Department Visit

All data is manually input into the system by clinician caring for the patient. The LOS and admitted variables are automatically calculated by the system.

You have been given some time to look at the data before your next visit with the clinician.

## (a) 5 points

This messy data is typical of what you might be given from a hospital. Find at least 7 data issues that you will ask the clinician to clarify for you. Simply state the concerns you have. A data issue can be something that seems wrong with the data, or something you would like to have clarified by the clinician.

## (b) 5 points

Write an R function, named `impute_dates()` that will impute values for the missing date variables from the ed data. The function should use tidy evaluation ([see here](#) for more info). The function should take the following arguments

- `data` - a data.frame
- `time_var` - a datetime variable name from the data (e.g. `arrival_time`)
- `type` - takes the character values ("random", "random\_poisson")

You will test this function in part (c) on the variables `arrival_time`, `departure_time`, and `admit_time` from the ed data.

The function will impute based on the following rules:

- if `type == "random"`, the function will pick a random datetime at uniform between "2021-11-01 00:00:00" and "2021-11-05 00:00:00". The random draw will be different for each missing value in the data.
- if `type == "random_poisson"`, the function will take a random draw from a poisson distribution with `lamda = 8` and add this many hours to the `arrival_time`. The random draw will be different for each missing value in the data. That is, if there are 30 missing data points, you will take 30 random draws from `rpois`.

Starter code below

```
# Function to impute missing date variables

# arguments:
# data: a data.frame or tibble
# time_var: the variable from the data frame to impute
impute_dates <- function(data, time_var, type) {

  # code goes here

}

# examples of calling the function

impute_dates(ed_data, admit_time, type = "random")

# or

ed_data %>%
  impute_dates(admit_time, type = "random")
```

## (c) 3 points

Test your function using all 3 date variables.

## (d) 2 points

After reviewing your data issues from part (a), the clinician has his own data analyst clean and provide you with a data set of all emergency department encounters from 2016-2019. They would like you to use this data to build and productionize a model to forecast total arrivals to the emergency department. The model will be implemented in January 2022. Are you comfortable performing this analysis. Why or why not?

## Questions 2 (15 points)

For this question, you will need the following `kidiq` data set

-  `kidiq.csv`

This data contains results of a test taken by school aged children, along with a variable indicating whether their mother graduated high school, their mother's results on an IQ test and their mother's age when she gave birth.

**(a) 5 points**

Write a function to implement the least squares estimator in matrix form, (called `multi_ols()`). It should take 4 input arguments:

- `data` a data frame containing the data of interest
- `continuous_vars` - a character vector for the continuous input variables to use in your regression.
- `factor_vars` - a character vector for the factor input variables (i.e. categorical data) to use in your regression.
- `y_var` A character string representing your independent variable to use in regression.

Your function should implement the least squares estimator in matrix form

$$\hat{\beta} = (X^{\top} X)^{-1} X^{\top} y$$

and return a named vector of coefficients (the names represent the variable names). Be sure to include the an intercept.

Test your function using the following code

```
# function code goes here
multi_ols <- function(data, continuous_vars, factor_vars, y_var) {

}

# test here
multi_ols(kidiq,
          continuous_vars = c("mom_iq", "mom_age"),
          factor_vars = "mom_work",
          y_var = "kid_score")

# should return same coefficients as
m <- lm(kid_score ~ mom_iq + mom_age + factor(mom_work), data = kidiq)
```

Alter your function to work with these inputs

**(b) 5 points**

- Fit a regression (using `lm()`) of child test scores on mother's age.
- display the fitted model in a table,
- interpret the slope coefficient.
- Based on this analysis, when do you recommend mothers should give birth? What are you assuming in making this recommendation?

**(c) 5 points**

Repeat this for a regression that further includes mother's education, interpreting both slope coefficients in this model. Have your conclusions about the timing of birth changed?

**Question 3 (5 points)**

for each of the questions below, answer True or False

**(a) P-values are calculated based on the assumption that the null is true for the population**

**(b) For a linear regression  $\beta$  coefficient, a P-value is the probability that your coefficient is different from zero.**

**(c) If  $p > .05$  we reject the null hypothesis at the  $\alpha = .05$  level.**

**(d) If  $p < .05$  we reject the null hypothesis at the  $\alpha = .05$  level.**

**(e) After conducting a regression analysis you find that the 95% confidence interval for  $\beta_1 = (0.13, 3.25)$ . This means there is a 95% chance that the true value of  $\beta_1$  is contained in that interval?**

**Question 4 (3 points)**

A regression was fit to data from different countries, predicting the rate of civil conflicts given a set of geographic and political predictors. Here are the estimated coefficients and their z-scores (coefficient divided by standard error), given to three decimal places:

Coefficient	Estimate	z-score
Intercept	-3.814	-20.178
Conflict before 2000	0.020	1.871
Distance to border	0.000	2.450
Distance to capital	0.000	3.629
Population	0.000	2.482
% mountainous	1.641	8.518
% irrigated	-0.027	-1.663
GDP per capital	0.000	-3.589



Why are the coefficients for Distance to border, Distance to capital, GDP per capital, and Population so small?

**Question 5 (15 points)**

This question will use the earnings data from the tutorial. It can be downloaded here

 [Earnings.csv`](#)

In each case fit the following models, print out the model summary, and interpret the coefficients.

**(a) 1 point**

`earnk` as dependent variable and `height` as independent variable

**(b) 1 point**

`log(earnk)` as dependent variable and `height` as independent variable. Filter the data for `earnk > 0`

**(c) 1 point**

`log10(earnk)` as dependent variable and `height` as independent variable. Filter the data for `earnk > 0`

**(d) 2 points**

`log(earnk)` as dependent variable and `height` and `male` as independent variables. Filter the data for `earnk > 0`

**(e) 2 points**

`log(earnk)` as dependent variable and `log(height)` and `male` as independent variables. Filter the data for `earnk > 0`

**(f) 4 points**

`log(earnk)` as dependent variable and `height` and `male` and interaction between `height` and `male` as independent variables. Filter the data for `earnk > 0`

**(g) 4 points**

create a z-score variable for height called `z_height`

log(**earnk**) as dependent variable and **z\_height** and **male** and interaction between **z\_height** and **male** as independent variables. Filter the data for **earnk** > 0