Homework 1

STA2101F 2020

Due October 1 2020 11.59 pm

Homework to be submitted through Quercus

You can submit this HW in Word, Latex, or R Markdown, but in future please use R Markdown. If you are using Word or Latex with a R script for the computational work, then this R script should be provided as an Appendix. In the document itself you would just include properly formatted output.

You are welcome to discuss the questions with others, but the solutions and code must be written independently. Any R output that is included in a solution should be formatted as part of the discussion (i.e. not cut and pasted from the Console).

1. **Chooose this question or the next** Find an article about the results of a study, in a scientific journal on a topic of interest to you. The article should discuss a single study, and should provide enough information on the study methods to answer the questions below.
   (a) Give the complete bibliographic reference, as well as a web link, to the published paper.
   (b) Was the study observational or a designed experiment?
   (c) What was the study population? What is the population of interest for the research?
   (d) If the study was observational, was it a prospective, or a retrospective study? If it was an experiment, was it randomized?
   (e) What were the units of analysis?
   (f) What was the primary endpoint and the main analysis of this endpoint?
   (g) What were the main conclusions of the study, in your own words?
2. **Choose this question or the previous** A short article by Professor Rob Hyndman in March describes two approaches to forecasting, time series modelling and agent-based modelling. In the latest release from the Public Health Agency of Canada, there are two forecasts, one on slide 10 and one on slide 12. The government has been criticized for not releasing details of its models, and the latest slide deck does have some references to the literature. In particular on slide 12 they link to this preprint.
   (a) Are these forecasts on slide 10 and slide 12 based on time series modelling or

agent-based modelling?

(b) Did the paper mentioned on slide 12 compare its forecasts to data? What data source(s) did they use?

(c) How many compartments does their SEIR model contain?

(d) They refer on p.15 at the beginning of Section 4 to a *fraction of contact reduction*, and its *critical threshold of the model*. What does this mean, and what is the critical threshold?

(e) How do you think PHAC constructed the forecast on Slide 12?

3. Suppose we have n measurements on a a response y, two explanatory variables x and z, and we assume a model of the form

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \epsilon_i, \quad i = 1, \ldots, n.$$

Further, assume that x is a continuous variate, but $z_i = \pm 1$; for example, x might be age, and z might be gender. We will also make the usual assumptions that the observations are independent, and that $\epsilon_i \sim (0, \sigma^2)$.

(a) Express this model as $y = X\beta + \epsilon$, where y and ϵ are $n \times 1$ vectors, making explicit the entries of the X matrix.

(b) Show that $E(y_i \mid x_i, z_i = -1)$ and $E(y_i \mid x_i, z_i = +1)$ are lines, and give their slopes and intercepts.

(c) What conclusions are implied by the hypothesis $\beta_3 = 0$? What conclusions are implied by the hypothesis $\beta_2 = 0$? What conclusions are implied by the hypothesis $\beta_3 = \beta_2 = 0$?

(d) Give an explicit expression for the least squares estimate of β_2, under the simpler model that assumes $\beta_3 = 0$.

4. The dataset `teengamb` concerns a study on teenage gambling in Britain. You can get more information about the data with `?teengamb`, but you will first need `library(faraway);data(teengamb)`, and possibly `install.packages`("faraway"). The questions below are adapted from FLM.

(a) Using `gamble` as the response, fit a linear regression model to the other four variables and give a table of the estimated coefficients and their estimated standard errors.

(b) With other explanatory variables held fixed, what is the estimated difference in expenditure on gambling for males compared to females?

(c) Fit a model with just `income` as a predictor, and use an F-test to compare it to the full model.

(d) Do the usual linear model assumptions appear to be satisfied for this data? Why or why not? Include at most two plots to support your answer.

(e) Predict the amount that a male with `status`, `income` and `verbal` at the maximum values in this data set would gamble, along with a 95% prediction interval.

(f) Fit a model with `sqrt(gamble)` as the response but with the same explanatory variables. Give a 95% prediction interval for the individual in (e), taking care to convert the interval to the original units of the response.

(g) A model selection strategy that is easily applied here is all possible subsets. Assuming `sex` must be included in all the models, fit all possible combinations of covariates. How stable is the estimated effect of `sex` across these models?

5. (SM Exercise 8.1.1) Which of the following can be written as linear regression models, (i) as they are, (ii) when a single parameter is held fixed, (iii) after transformation? For those that can be so written, give the response variable and the form of the design matrix.

   (a) $y = \beta_0 + \beta_1/x + \beta_2/x^2 + \epsilon$
   (b) $y = \beta_0/(1 + \beta_1 x) + \epsilon$
   (c) $y = 1/(\beta_0 + \beta_1 x + \epsilon)$
   (d) $y = \beta_0 + \beta_1 x^{\beta_2} + \epsilon$
   (e) $y = \beta_0 + \beta_1 x_1^{\beta_2} + \beta_3 x_2^{\beta_4} + \epsilon$

6. **Bonus Question: highly recommended for PhD** Suppose our model for the expected value of Y is nonlinear in β:

$$y_i = \eta_i(\beta) + \epsilon_i, \quad i = 1, \ldots, n; \quad \beta \in \mathbb{R}^p$$

   where we assume $\eta(\cdot)$ is a known function of β (and possibly some covariates), and ϵ_i are i.i.d. $N(0, \sigma^2)$. Show that the least squares estimate of β solves the equation

$$\{y - \eta(\beta)\}^T X(\beta) = 0,$$

   where η is $n \times 1$ and X is $n \times p$. Give an expression for the (i, j)th entry of $X(\beta)$. Suggest an iterative method of solving this equation from some starting point β_0. Try this out on Example 10.1 in SM, using the nonlinear model suggested there:

$$y = \beta_0\{1 - \exp(-x/\beta_1)\} + \epsilon.$$