Homework 3

STA2101F 2020

Due December 3 2020 11.59 pm

Homework to be submitted through Quercus

Please submit both the .Rmd and the .pdf file for this homework. (It is not possible to annotate .html files in Quercus.) In the .pdf document you should just include properly formatted output, not unedited chunks of ##R output. See the solutions to HW 1 Q 4, for example.

You are welcome to discuss the questions with others, but the solutions and code must be written independently.

1. *Measurement error in regression* Suppose y depends on x in a simple linear regression :

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_j, \quad i = 1, \ldots, n; \\ \epsilon_i &\sim N(0, \sigma_e^2), \quad \text{i.i.d.} \\ x_i &\sim (\mu_x, \sigma_x^2) \quad \text{i.i.d.} \end{aligned}$$

   We assume that x and ϵ are independent. Instead of observing x_i, we are only able to observe a corrupted value $w_i = x_i + u_i$, where u_i is independent of x_i, and $u_i \sim (0, \sigma_u^2)$, also i.i.d.

   (a) Show that the least squares estimator from this regression is

$$\hat{\beta}_1 = \Sigma(y_i - \bar{y})(w_i - \bar{w})/\Sigma(w_i - \bar{w})^2.$$

   (b) Find an expression for the limit in probability of $\hat{\beta}_1$ and thus deduce that it will usually be an under-estimate of the true regression coefficient β_1. In what special circumstance will it be consistent for β_1?

   This result is often relied on to argue that if there is uncertainty in the x's used in a given regression, the association with the response will be attenuated, i.e. less likely to be significantly different from zero.

2. *Faraway, Exercise 11.2*: The dataset `uswages` was drawn as a sample from the Current Population Survey in 1988. The response is `wage`, weekly wages in dollars (adjusted for inflation), and the other variables in the data set are `educ`, years of education, `exper`, years of experience, `race` (1 Black/ 0 White), `smsa` (1 if living in a standard metropolitan statistical area/ 0 if not), a set of dummy variables to indicate region

of employment (`ne, mw, we, so`), and a dummy variable to indicate part-time work (`pt`). Of interest is how years of education are associated with wages, and whether this effect is different in different subgroups determined by the other variables.

(a) Using `wage` as the response, fit one or more non-parametric regression models, with smooth functions of education (and possibly experience), and including relevant dummy variables as appropriate. Choose the method, and model for that method, that you prefer, and summarize the results. Explain why you made the choice that you did.
(b) Fit a parametric model to `wage` or a transformation of `wage` and compare the results to those of the semi-parametric model that you chose in (a).
(c) Compute the square root of the absolute value of the residuals from your preferred non-parametric model fit, and smooth these as a function of `educ`. This was suggested by Faraway – what information about the data is obtained from this?

3. *From Exercise 10.6.1 in SM*: Suppose Y follows a Poisson distribution conditional on an unobserved error ϵ:

$$f(y \mid \epsilon) = (\mu\epsilon)^y \exp(-\mu\epsilon)/y!, \quad y = 0, 1, ...,$$

(a) Assume that ϵ follows a Gamma distribution with expected value 1 and variance $1/\nu$. Derive the expected value and variance of Y, and give an expression for its unconditional distribution. Show that this is in the form of a generalized linear model if ν is known, but not otherwise.
(b) We found evidence for overdispersion in the Galapagos Island data `gala` from `library(faraway)`, and fit the model with `quasi-Poisson` to allow for this. Compare the results from the quasi-Poisson model to those from the negative binomial model.[1]

4. The article "An estimate of the science-wise false discovery rate and application to the top medical literature" by Jager & Leek (*Biostatistics*, 2014), is posted on the course web page and available via the link in (d). In this paper they attempted to estimate the rate of false discoveries in papers published in leading medical journals.

(a) Construct a 2×2 table with "Null hypothesis true/false" as the two column headings, and "Discovery/No Discovery" as the two row headings. Give a definition (algebraic) of the false discovery rate as a function of the entries in your table.
(b) What model did Jager & Leek use for the distribution of p-values?
(c) Their conclusion was that the rate of false discoveries among published results was 14% with an estimated standard error of 1%. How was the standard error estimated?
(d) There were several discussants of this paper, and all the discussions can be found at here. Choose one discussion and summarize in a paragraph the main point of the discussant. Comment briefly on this point, and on the rejoinder by Jager & Leek.

[1]The function `glm.nb` in `library(MASS)` fits the negative binomial model by maximum likelihood.

5. *Bonus Question* The `toxo` (toxoplasmosis) data in the library `SMPracticals` is discussed in SM as Example 10.29 and 10.32. In Example 10.29 the linear predictor is a cubic function of rainfall, with response modelled as an overdispersed Binomial. In Example 10.32 a local likelihood analysis is carried out instead.
   (a) Carry out these two analyses, parametric and non-parametric, and provide summary plots and tables. (You can reproduce the ones in SM, or create your own.)
   (b) Discuss briefly the advantages and disadvantages of each approach.