HW Question Week 3

STA2101F 2021

Due October 6 2021 11.59 pm

Homework to be submitted through Quercus

Part 1: Getting ideas for project

Next week you will be asked to submit details on the dataset you will analyse for your final project.

This week we will start collecting sources of data.

Please submit to Piazza a link to a website that either provides an interesting data set in the context of an application, or that provides a library of datasets. (There isn't a mark for this part, but consider it a duty :)

The data set for the project should have a single response variable of interest, and a number of potential covariates related to the response variable. It should have more than thirty observations, and less than about 1000, although exceptions can be made. It will be helpful if there is not a lot of missing data, as we will not cover techniques for missing data in detail. We will also not cover specialized models for time series data nor for data with strong spatial dependence, but data sets with these features can sometimes be analyzed, at least in part, by other means.

The dataset should not be taken from a textbook. Ideally it will not have been analyzed before, but a new analysis of a previously analyzed dataset can be acceptable.

Statistics Canada, and the City of Toronto, and many other government organizations, make data available online. Many journal articles include links to the data as part of the publication.

Part 2: Question on regression

(a) Often a regression analysis is assessed by reporting R^2, which is $SS(REG)/TSS$, or SS_{REG}/TSS, in the notation of Sep 29 slides, or $1 - RSS/TSS$, in the notation of LM §2.9. (Where TSS is corrected for the mean.) Show that R^2 is equal to the square of the correlation between y and $\hat{y} = X\hat{\beta}$, where $\hat{\beta}$ is the usual least squares estimator.

(b) The adjusted R^2, R^2_a is defined (LM-2 §10.3, LM-1 §8.3) as

$$R^2_a = 1 - \frac{RSS/(n-p)}{TSS/(n-1)},$$

and Faraway writes "Adding a predictor will only increase R_a^2 if it has some predictive value". Explain.

(c) For the data `divusa`, treating `divorce` as response, compare the models selected using AIC, R_a^2, and $C_p = RSS_p/\hat{\sigma}^2 + 2p - n$ (LM-2 §10.3, LM-1 §8.3). Comment on any differences.

(d) Do the assumptions of the linear model seem adequate for this data? Include at least two plots as part of your explanation.

(e) *Bonus/PhD (SM Problem 8.21)* Suppose data $(x_1, y_1), \ldots, (x_n, y_n)$ are modelled with a simple linear regression model $y_i = \gamma_0 + \gamma_1(x_i - \bar{x}) + \epsilon_i$. In a *calibration* problem, the value y_+ of a new response is observed, and it is desired to predict x_+, the unknown corresponding value of x. With $\tilde{\sigma}^2 = \Sigma_i\{y_i - \hat{\gamma}_0 - \hat{\gamma}_1(x_i - \bar{x})\}^2/(n-2)$ and $s_x^2 = \Sigma_i(x_i - \bar{x})^2$, argue that

$$T(x_+) = \frac{y_+ - \hat{\gamma}_0 - \hat{\gamma}_1(x_+ - \bar{x})}{[\tilde{\sigma}^2\{1 + n^{-1} + (x_+ - \bar{x})^2/s_x^2\}]^{1/2}}$$

follows a T_{n-2} distribution, so that

$$\chi_{1-2\alpha} = \{x_* : t_{n-2}(\alpha) \leq T(x_*) \leq t_{n-2}(1-\alpha)\}$$

is $(1 - 2\alpha)$ confidence set for x_+. Show that if $\gamma_1 = 0$ that $\chi_{1-2\alpha}$ has infinite length with probability $1 - 2\alpha$.