Final HW

STA2101F 2020

Due December 20 2020 11.59 pm on Quercus

Please work alone

Please submit both the .Rmd and the .pdf file for this homework. 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.

Fifteen percent of your mark will be based on the quality of your presentation, so please take time with formatting, explanations, and so on.

1. In the linear model

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

it may sometimes be appropriate to assume $\epsilon_i \sim (0, \sigma^2 v_i)$, where $v_1, \ldots, v_n$ are known positive quantities, attached to each observation. In matrix form we would write

$$y = X\beta + \epsilon; \quad \epsilon \sim (0, \sigma^2 V),$$

where V is an $n \times n$ diagonal matrix with v_i on the diagonal.

(a) Derive expressions for the *weighted least squares* estimator of β, defined by

$$\hat{\beta}_{WLS} = \arg\min_{\beta} \sum_{i=1}^{n} \frac{1}{v_i}(y_i - x_i^T \beta)^2 = \arg\min_{\beta} (y - X\beta)^T V^{-1} (y - X\beta).$$

(b) Give an expression for the residual sum of squares after fitting this model, and derive an unbiased estimator of σ^2. *Hint*: since the weights are known, model (1) can be linearly transformed to an ordinary least squares model.

(c) If $v_i = v_i(\alpha)$ depends on a parameter α that is unknown, suggest a method for computing the estimator in (a).

(d) The dataset `gammaray` in `library(faraway)` has measurements on the X-ray decay light curve of a Gamma ray burst. The column `error` gives an estimate of the standard error of each flux measurement. (See `help(gammaray)` for more details.) Fit a weighted least squares model with `flux` as the response and `time` as the explanatory variable, using weights derived from `error`. Compare the coefficients and their estimated standard errors from this model and an unweighted

least squares model. Compare as well the plots of the residuals against the fitted values from each model.

(e) Does a linear model seem appropriate for this data? Explain why or why not. If not, suggest an alternate model that seems more appropriate.

2. The data `Boston` in `library(MASS)` contains data on the value of homes in the suburbs of Boston, along with several explanatory variables.
   (a) Create a binary variable `crim2` that is 1 if `crim` exceeds the median in the dataset, and 0 otherwise.
   (b) Fit a logistic regression model with `crim2` as the response, and all the other variables (except `crim`) as covariates. Provide a summary table of the estimated coefficients and their estimated standard errors.
   (c) Describe in words the interpretation of the estimated coefficient of `nox`.
   (d) Create a simpler logistic regression model by omitting covariates that do not appear to be needed in the model. Explain why you chose the model you did.
   (e) Using your simpler model, create a confusion matrix, or classification table, showing the actual values of `crim2` along with the predicted values, using the classification rule that assigns 1 if the fitted value is positive and 0 if the fitted value is negative.

3. The density function for a response that follows a generalized linear model is, as given in class,

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y;\phi)\right\}.$$

The chi-squared density with known degrees of freedom q (assumed to be an integer ≥ 1) and unknown parameter σ has density function

$$\frac{y^{q/2-1}}{2^{q/2}\sigma^q\Gamma(q/2)}\exp\left(-\frac{y}{2\sigma^2}\right), \quad y>0, \sigma>0.$$

   (a) Show that the χ^2_q density can be written in generalized linear model form, and identify θ and ϕ.
   (b) The yield of an industrial process was measured r_i times, independently, at m different temperatures t_i. The resulting yields $Z_{ij}, i = 1,\ldots,m; j = 1,\ldots,r_i$ at t_i may be assumed to follow a normal distribution with $\mathbb{E}(Z_{ij}) = \alpha_i$ and $\text{var}(Z_{ij}) = \tau_i$, where α_i and τ_i depend on t_i. Create a *derived response*

$$Y_i = \sum_{j=1}^{r_i}(Z_{ij} - \bar{Z}_i)^2,$$

   where $\bar{Z}_i = r_i^{-1}\sum_{j=1}^{r_i} Z_{ij}$. What is the distribution of Y_i?
   (c) Explain how a generalized linear model with $Y_1,\ldots,Y_n$ as a response could be used to assess the dependence of τ_i on temperature t_i.
   (d) Briefly discuss the advantages and disadvantages of the canonical link function of your model.

4. This question refers to the paper "The Coronavirus Exponential: A Preliminary Investigation Into the Public's Understanding", published in the *Harvard Data Science Review* by Podkul et al., in Special Issue 1, posted May 14 2020. The paper is posted on my web page, and on Quercus under Assignments/Final homework.

   (a) The paper describes two survey experiments that were conducted. In survey experiment 1, described in §2.2, half the subjects were asked a question related to linear growth and half were asked a question related to exponential growth. The authors say survey respondents were randomly assigned to one question or the other. What are the advantages of random assignment in this context?

   (b) In survey experiment 2, described in §2.3, they assessed understanding of exponential growth from a different point of view. How is it different, and why did they undertake this experiment?

   (c) In §2.1, the authors state "We emphasize that our web-panel-based survey is not a probabilistic sample". Why do they say this? What effect do you think this might have on their conclusions?

   (d) In the same paragraph they note that "we have done our best to mitigate and assuage concerns". What steps do you think they took to do this?

   (e) What information is presented in Figure 1? How does this relate to survey experiment 1?

   (f) A further analysis described in §3 investigated the relationship between "worry about the epidemic" and the answers to survey experiment 2. Write out an equation that describes the analysis they did to study this relationship.

For holiday fun: the dataset is available at https://github.com/optimus-forecasting-and-polling/HDSR-Paper-Materials/blob/master/README.md.