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.

These 15 points were assigned approximately as follows: 5 if I could knit your .Rmd file; 5 if a .pdf file was submitted (although I was generous about .html submissions), and 5 points for style – no, or very little, raw `R` output, clear formatting of tables, clarity of your explanations, etc.

1. [25 points: 5 for each part.] 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

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

This can be derived using either by differentiating $SS(\beta) = (y - X\beta)V^{-1}(y - X\beta)$ and setting it equal to zero, or making the transformation $V^{-1/2}y$ and similarly with X and invoking the expression for the OLS estimator. The answer is $\hat{\beta}_{WLS} = (X^TV^{-1}X)^{-1}X^TV^{-1}Y$.

(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. It wasn't

completely clear whether or not the residual sum of squares should be weighted. That is, I had in mind using $SS(\hat{\beta})$ as the residual sum of squres, with $SS(\beta)$ defined above. In this case an unbiased estimate of σ^2 is

$$\tilde{\sigma}^2 = SS(\hat{\beta}_{WLS})/(n-p),$$

where p is the length of β and n is the length of y. Some people used instead $(y - X\hat{\beta}_{WLS})^T(y - X\hat{\beta}_{WLS})$, which gives a messy expression, but I didn't deduct marks for that.

(c) If $v_i = v_i(\alpha)$ depends on a parameter α that is unknown, suggest a method for computing the estimator in (a). I was looking for a suggestion of iterating; this can be done most naturally by assuming ϵ has a normal distribution and using likelihood methods. But iteratively re-weighted LS should work too. The quantity to be minimized over α and β would be $SS(\alpha, \beta) = (y - X\beta)V(\alpha)^{-1}(y - X\beta)$, and the solution in β for a given α is as in (a), so the next step would be to minimize $SS(\alpha, \hat{\beta}_\alpha)$ with respect to α, and put this new estimate into $\hat{\beta}_\alpha$ etc. Some people suggested using the OLS residuals to estimate v_i, as is also suggested in SM, Ch.8, but this is a non-parametric estimate and doesn't exploit the simplicity of a V which just depends on one parameter. I didn't take off very many marks for this proposal if it was clearly set out. Some people suggested a regression of, e.g. $\log r_i^2$ against some covariate, in order to estimate α. This would also be okay, depending on the problem.

(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.
This data set is very weird, and both ordinary and weighted least squares give unusual answers and strange residual plots. Following the logic of (a), the weights would be $1/$`gammaray$error`^{2}; many people used $1/$`gammaray$error`, and I didn't deduct marks for this. Although both fits are awful, you still got full marks if you fit both models and gave the requested residual plots. Fitting `flux` to $1/$`time` (which is a rate), is much better, and fitting `log(flux)` against `log(time)` is also not bad. Kudos to Peter Collins who looked up the original paper and noted that the measurements occurring before and after 2000 seconds were made by different methods. This explains the breakpoint in the plot of `flux` against `1/time`.

2. [15 points total] 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. Most people did quite well on this (easy) question. Although it was not intended, it turned out that the coefficient of `nox` was ridiculously large (44 in the reduced model and 48 in the original model.) However, an increase of 1 part per million (which is the units of `nox`), is huge, the range of `nox` in the data is much smaller, so it would make more sense to assess the effect of an increase of, for example, 0.1 parts per million or at most 0.5 parts per million, the range of the data.

3. [20 points; 5 per part] *SM Problem 10.10.8* 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 this density can be written in generalized linear model form, and identify θ and ϕ. The solution is not unique, but the simplest is to identify θ with $-1/(2\sigma^2)$, in which case ϕ is just 1. I was pretty flexible.
(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? The chi-squared density of (a), with $q=r_i-1$ and $\sigma^2=\tau_i$.
(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.
The linear predictor would be, most likely, $\beta_0 + \beta_1 t_i$, and then if we use the canonical link we'd be modelling $1/\tau_i$ as linear. This makes the calculations simpler, but doesn't respect the fact that τ_i must be positive. We could fit this with `glm` with `family(gamma)`, as the χ^2 distribution is in the gamma family; the shape parameter is related to $r_i - 1$, and is known; the scale parameter (or mean parameter) is related to τ_i.

4. [25 points total]} 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? Random assignment of subjects to question should eliminate, on average, any differences between groups that might otherwise be confounding variables (variables that affect both the treatment assignment and the response). This applies only to the subjects in the sample, of course. Their representativeness of the US population is unclear, as noted below.
   (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? Kind of obvious – they gave the subjects the easier task of identifying the visual difference between linear and exponential growth. They also used a different question, although it's not clear why. You have to dig into the github repo to discover that they used the same subjects for both experiments, so perhaps they thought there would be less of a learning effect if they changed the question.
   (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? The sample was not selected according to a probability mechansim. (Many people said "each person didn't have an equal chance to be selected", i.e. it was not a *simple random sample*, but there are many other probability samples; stratified, cluster samples etc.) The key point is that with probability sampling you know the probability that each subject is selected, and you can use this probability distribution to estimate the population quantity of interest, **and** to estimate variance of your estimate. Without this, any margin of error relies on heavy regression modelling, and if the model is incorrect so are the estimates.
   (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? Lots of

modelling, and weighting to mimic the population of interest, at least as far as a few key demographic variables go. This is pretty common with web-based panel surveys, it's a (slightly desperate) attempt to try to ensure that the conclusions apply to the population of interest, rather than the group of survey respondents only. Note that the authors say "Although the present project simply seeks to report the findings for the survey's respondents, these figures are weighted to be representative the all U.S. adults". This is much harder than it sounds. For example, their quota sampling would have prescribed the number of, e.g., adults over 65 to be contacted. This group then become proxies for all adults in the US over 65, and their answers are weighted accordingly. But it is effectively impossible to ensure, or even to check, if this group is more or less likely than the general population of adults over 65 to understand exponential vs linear growth.

(e) What information is presented in Figure 1? How does this relate to survey experiment 1? Shows the distribution of responses to the question.

(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. This is a probit model: let y_i be the response of subject i, coded as '1' for "Very worried" or "Somewhat worried" and 0 otherwise. The linear predictor for subject i is $x_i^T\beta$ in the usual way, where the vector x_i records values for various covariates. The responses y_i are assumed to be Bernoulli, with $p_i = \Phi(x_i^T\beta)$, or $\Phi^{-1}(p_i) = x_i^T\beta$. Here $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. The difference between this and logistic regression, where p_i is modelled using the cdf of the logistic distribution, is numerically slight unless the p_i are very close to 0 or 1. You can fit this model in `R` using `family = binomial(link = "probit")`.

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