Homework 2

STA2101F 2020

Due November 5 2020 11.59 pm

Amended on Oct 26

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. *Two-way layout with replication.* If we have a balanced two-factor experiment with replication, a suitable linear model might be

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \tag{1}$$

where $i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K$. In this model we separate the main effects of factors A and B from the interaction effect. If the observations were set out in a two-way layout, there would be K observations in each cell.

   (a) If we assume that ϵ_{ijk} are all independent and identically distributed $N(0, \sigma^2)$, argue that the variation in a single cell, $\Sigma_k(y_{ijk} - \bar{y}_{ij.})^2$ estimates σ^2 with $K - 1$ degrees of freedom. From this conclude that $\Sigma_{ij}\Sigma_k(y_{ijk} - \bar{y}_{ij.})^2$ estimates σ^2 with $IJ(K - 1)$ degrees of freedom. Suggest and construct a plot that might be used to check on the assumption of constant variance.

   (b) The data set `butterfat` in `library(faraway)` is a two-way layout with 10 observations per cell. Present a two-way table of cell means, showing as well the row and column means. Is there evidence of interaction between the two factors? Explain. I mistakenly called the dataset ‘buttermilk‘ in the Oct 22 version.

   (c) Fit the linear model of equation (1) to this data and give the analysis of variance table. Give an algebraic expression for the sum of squares due to interaction in the analysis of variance table. What is the estimate of σ^2? Does it have $IJ(K - 1)$ degrees of freedom?[1]

[1] if not you've made a mistake somewhere.

(d) Provide at most three plots of the data and/or fitted model. What information is available from these plots?

(e) Is the best breed in terms of butterfat content clearly superior to the second best breed? Why or why not? Explain your calculations.

2. (*FELM Exercise 2.2*) The dataset `wbca` in the library `faraway` comes from a study of breast cancer in Wisconsin (see `?wbca`). There are 681 cases of potentially cancerous tumours, of which 238 are actually malignant. Malignancy of a tumor is traditionally determined by an invasive surgical procedure. The purpose of this study was to determine whether a new procedure called fine needle aspiration, which draws only a small sample of tissue, could be effective in determining tumor status.

   (a) Fit a binomial regression with `Class` as the response and the other nine variables as predictors. Report the residual deviance and associated degrees of freedom. Can this information be used to determine if this model fits the data? Explain.

   (b) Use the `step` function to choose final model.

   (c) For a patient with the following measured values of the predictors: `Adhes` = 1, `BNucl` = 1, `Chrom` = 3, `Epith` = 2, `Mitos` = 1, `NNucl` = 1, `Thick` = 4, `UShap` = 1, `USize` = 1, predict the `Class` of the tumor, using the full model and the final model. Give confidence intervals for your predictions, and explain how you obtained them.

   (d) Suppose that a cancer is classified as benign (`Class` = 1) if $p(\hat{\beta}) > 0.5$, and otherwise is classified as malignant. Give a table showing the misclassification errors if this method is applied to the current data, using the final model from (b). Give a similar table but using the cutoff 0.9 instead.

   (e) The error estimates in (d) are overly optimistic, because they are computed on the same data that was used to fit the model. A better strategy is to fit the model on training data and assess the errors on test data. Choose a random 1/3 of the data to hold out as a test set, and find the best fitting model on the remainder. Construct the tables of misclassification errors for the cutoffs 0.5 and 0.9, and compare to the results in (d). *Note: If you have issues with convergence on the randomly selected training set, you could try following Faraway's suggestion of choosing each 3rd case to construct the test set.*

3. *SM, Exercise 10.4.1* Suppose $y_1, \ldots, y_n$ are independent Binomial(n_i, p_i) random variables, with $p_i = \exp(x_i^T\beta)/\{1 + \exp(x_i^T\beta)\}, i = 1, \ldots, n$.

   (a) Derive expressions for the log-likelihood function, the equations defining the maximum likelihood estimator of β, and the residual deviance.

   (b) Use the approximation $\log(1 + x) \simeq x - x^2/2$ to show that the residual deviance is approximately

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1 - \hat{p}_i)}.$$

   It is enough to prove this result with denominator $n_ip_i(1-p_i)$, which is a bit easier.

   (c) Show that if $n_i \equiv 1$, that the deviance is a function only of $\hat{\beta}$, and not otherwise a function of y. Show that if $n_i \equiv 1$ and $p_i \equiv p$, that $X^2 = n$. *It is worth looking at*

Davison's version of this question in SM p.497.

(d) Explain why the results in (c) mean that neither the residual deviance nor Pearson's X^2 can be used to assess goodness of fit with binary data.

4. This question refers to the article "Clinical characteristics and predictors of outcomes of hospitalized patients with Coronavirus Disease 2019 in a multiethnic London National Health Service Trust: a retrospective cohort study" by Perez-Guzman et al. (2020) in *Clinical Infectious Diseases*. The article and supplementary material are posted on my web page for this class.

   (a) What is a "retrospective cohort study", and why is this study described that way?

   (b) What is the "Elixhauser comorbidity score"?

   (c) What was the primary outcome of the study?

   (d) The methods section refers to "unadjusted" and "adjusted" logistic regression. What is the difference between these?

   (e) There are 9 models reported in Table S4 of the Supplementary Material. Which of these models ended up being summarized in Table 2 of the main paper?

   (f) The first 22 or 23 lines of Table 2 show one column of estimates called "unadjusted", and a second column "adjusted for age." As the authors note, "when accounting for age, only diabetes and chronic kidney disease remained statistically significant" (p.6). Why do you think this is?

   (g) On p.7 the authors state that they performed "unadjusted and adjusted logistic regression analyses to assess the odds of death of BAME groups compared to white patients". In your own words, what did these analyses show?

   (h) In the discussion section, the authors say "our findings .. sugest that important biological differences may be potentially driving differences in COVID-19 hospitalization outcomes by ethnicity, which cannot solely be explained by socioeconomic differences". What is the basis for this claim?

   (i) Summarize in your own words what you think are the two most important findings in this paper.

5. *Bonus Question* The article by Roozenbeek et al. (2020) *RSOS* was discussed in class on October 15. The authors carried out several statistical analyses, as described in Section 2.3: Pearson's correlation coefficent; one-way analysis of variance; ordingary least squares regression; two logistic regressions, and an OLS linear regression.

   (a) Choose two of these analyses and try to reproduce them using the authors' data.[2]

   (b) Try to reproduce one of the plots in the paper or the supplementary document.

   (c) In the supplementary information, Figure S1 shows residual plots from OLS regression. Why do you think there is an apparent linear boundary on the residuals vs fitted plot?

   (d) In the paper (Footnote 10 on p.8), the authors refer to a "robust standard error regression". What method did they use for this? (It might be obvious from their code what `R` package they used, but I am looking for some detail on what statistical method they used.)

[2] You are welcome to use their code or your own; I found it necessary to read the data using `read.csv`, for example, not `read.xls` as they recommend.

(e) The headline from this paper stated “Poor numerical literacy linked to great susceptibility to Covid-19 fake news”. Do you think this headline is supported by the analyses in the paper? Why or why not?