Assessed coursework 1

ST346 Generalized Linear Models for Regression and Classification

Deadline: 22 Oct 2024, 1 pm

Please read these instructions carefully!

This assignment counts for **10%** of your final module mark. The maximum score for this coursework is **25 marks**. Numbers in brackets indicate the points available for each question.

Your solutions should be submitted electronically in the form of a **machine-readable** PDF document using the submission portal on the ST346 moodle page, see the guidance on online submissions.

Penalties for late submission are listed in the assessment handbook.

Make sure you read the questions carefully and provide full answers (in full sentences!). You must show your working if you want credit for your answers. For most questions this will take the form of `R` code embedded in your document. You should also include a text explanation of what you are doing. Your report should be of a professional standard.

If you have any queries about the coursework please post them on the ST346 forum, but do not post any part of your solutions.

To access the data needed for this assignment, download the file `CourseworkData1.rda` from the ST346 moodle web page and read it into `R` using the function `load()`. This will create two data frames in your workspace:

- `cleaners` and
- `mushrooms`.

Please be aware that your work will be submitted to TurnItIn, a piece of plagiarism-detection software. Cases of suspected collusion or plagiarism will be followed up as outlined in the course guide. Note that detailed discussions of the assignment or comparisons of numerical/graphical results or computer code are **not permitted.** Furthermore the use of AI such as ChatGPT or other generative artificial intelligence tools are not permitted.

Good luck with the assignment!

Question 1 - Weighted regression

For planning purposes, a cleaning company would like to assess how many offices (on average) a team of cleaning staff is able to clean in a working day. For this purpose, they have collected past data on how many offices the various cleaning teams cleaned. The variables are

- `Staff`: the number of staff members on the cleaning team;
- `Offices`: the number of offices cleaned in a working day.

The aim is to use the size of the cleaning team to predict the number of offices cleaned.

(a) [**2 marks**] Produce a **suitable** plot of the data. Discuss whether is it reasonable to assume a linear relationship between the size of the cleaning team and the number of offices cleaned.

(b) [**1 mark**] Fit a normal linear model that predicts the number of offices cleaned from the number of cleaning staff on the team. Give an interpretation of the slope coefficient of the fitted model.

(c) [**2 marks**] Produce a residuals versus fitted values plot for the model fitted in (b). Critically evaluate the plot.

(d) [**2 marks**] Note that there are only a small number of distinct values for the size of teams. For each team size x there are multiple measurements for the response variable Y where Y = number of offices cleaned. The table below presents, for each size of team, the sample variance for the response variable.

Team size x	estimate of Var(Y | x)
2	9.00
4	24.67
6	22.00
8	44.12
10	62.84
12	53.14
16	144.01

Develop `R` code that reproduces the estimates in the table above. Verify that your code produces the same (rounded) estimates as above.

Hint: To produce the data for the table, you might use, for example, the function `aggregate(x, by, FUN)`. The `R` help function has some examples on how to use `aggregate()`.

(e) [**4 marks**] Consider the weighted normal linear model

$$Y_i \quad \sim \quad \mathcal{N}\Big(\beta_0 + \beta_1 x_i,\ \phi/w_i\Big), \qquad i \in \{1, \ldots, 53\}.$$

where x_i is the size of the ith team and Y_i is the number of offices cleaned by the team.

The weighted normal linear model assumes that $w_1, \ldots, w_n$ are known. Unfortunately these are not known in this example. However, in part (d) we were able to estimate the variances $\mathbb{V}ar(Y_i \mid x_i)$ for $i = 1, \ldots, 53$. This means that we can produce estimates of the weights.

Assume that $w_i = 1$ for $x_i = 2$ and that $\mathbb{V}ar(Y_i \mid x_i) = \mathbb{V}ar(Y_j \mid x_j)$ if $x_i = x_j$.

- Derive (with explanation/justification) an expression for the weights $w_1, \ldots, w_{53}$ in terms of the response variances $\mathbb{V}ar(Y|x)$ where $x \in \{2, 4, 6, 8, 10, 12, 16\}$. (This expression should not depend on ϕ.)
- Use the derived expression for w_i and the information from (d) to compute estimates of the weights $w_1, \ldots, w_{53}$.
- Report the estimated weights for $x_i \in \{2, 4, 6, 8, 10, 12, 16\}$.
- Create an additional column in the data frame `cleaners` that records the estimated weight for the corresponding datapoint.

(f) [**3 marks**] Using the weights computed in (e) fit the weighted normal linear model and present the model summary. Produce a plot of the weighted residuals against the fitted values. Critically evaluate the plot.

(g) [**2 marks**] The weighted normal linear model corresponds to an (unweighted) normal linear model on a transformed scale. Derive the model equation for the model on the transformed scale and then fit the model. Report the model summary of the fitted model.

(h) [**1 mark**] Verify numerically that the weighted residuals of the model in (f) and the response residuals for the model in (g) are identical.

Question 2 - Logistic regression

For this question you will explore a small subset of data from a dataset that was simulated to resemble real-life data, see Wagner, Heider and Hattab (2021).[1]

The aim is to classify mushrooms as either edible or poisonous. Data for 2000 (hypothetical) mushrooms is provided with the following variables:

- `Class`: binary type of mushroom (edible or poisonous).
- `Diameter`: diameter of the mushroom's cap in cm.

(a) [**2 marks**] Consider a logistic regression model with `Class` as response variable and `Diameter` as predictor variable. Thus the linear predictor is defined as

$$\eta_i \quad = \quad \alpha + \beta \, \texttt{Diameter}_i \qquad \text{for } i = 1, \ldots, 2000.$$

Based on the above model, derive an algebraic expression for the probability of a mushroom being poisonous as a function of its cap diameter.

(b) [**1 mark**] Estimate the odds ratio of being poisonous for a mushroom with a 5 cm cap diameter versus a mushroom with a 10 cm cap diameter.

(c) [**2 marks**] Produce a suitable plot showing the empirical relationship between the predictor `Diameter` and the response `Class`.

(**Hint:** This should be a plot of the data and not involve any fitted model!)

Using evidence from the plot to justify your answer, decide whether it is reasonable to assume the functional form in (a) to describe relationship between the cap diameter of a mushroom and the probability of it being poisonous?

(d) [**2 marks**] Fit a new regression model in which the linear predictor is a quadratic polynomial in `Diameter`. Then produce a plot showing the predicted probability of a mushroom being poisonous based on its cap diameter. Briefly comment on the plot.

(e) [**1 mark**] To the nearest half centimeter, for which cap diameters does the predicted probability of a mushroom being poisonous fall below 50%?

[1] Wagner, D., Heider, D., & Hattab, G. (2021). Mushroom data creation, curation, and simulation to support classification tasks. Scientific Reports, 11, 8134. 10.1038/s41598-021-87602-3.