Homework 3

PubH 7406: Biostatistical Inference II – Jared D. Huling

Due: Thurs, March 2, 2023 at the beginning of class

Logistic Regression - 38 points

Instructions

Turn in the homework in the form of a PDF. It is fine to use existing functions to answer questions. However, note that simply providing output from statistical software is not sufficient and will not receive full points. Any output/results must be interpreted in the context of the problem and accompanying explanations of the results are necessary. Use clearly-defined and explained statistical notation to accompany results. For example, any statistical tests should be accompanied by a formal statement of the null and alternative hypotheses with additional context explaining what the hypotheses mean in the context of the problem. Please follow the instructions on homeworks in the syllabus in order to receive full credit.

If you have any questions, please ask in the course Q&A on Canvas so that others can see any responses.

The Questions

1. **(8 points)** Problem 5.19, pg 202 of CDA.

2. **(8 points)** Problem 6.5, pg 244 of CDA. Data can be downloaded as:

```
read.csv("http://jaredhuling.org/data/pubh7406/table_6_3_data.csv")
```

3. **(22 points):** This problem focuses on the number of deaths from Leukemia and other cancers among survivors of Hiroshima during the period 1950–1959).

The following table tabulates the number of deaths from Leukemia and other cancers among survivors of Hiroshima during the period 1950-1959. The subjects were aged between 25 and 60 in the year 1950, and are classified by the dose of radiation they received in radons.

radiation	midpoint	leukemia	other cancer	total cancers
0	0	13	378	391
1 to 9	5	5	200	205
10 to 49	29.5	5	151	156
50 to 99	74.5	3	47	50
100 to 199	149.5	4	31	35
200 +	249.5	18	33	51

In this table we converted the radiation level to a continuous predictor, coding the value as the midpoint of the range (e.g., for a radiation of 1 to 9 we use the value $(1+9)/2 = 5$). For the largest radiation value we choose a value of 249.5 (assuming the radiation goes from 200 to 299). The data can be read as

```
read.csv("https://jaredhuling.org/data/pubh7406/hiroshima.csv")
```

1. Fit a model that estimates the logit of the probability of having leukemia in terms of the radiation factor variable. Interpret this statistical model on the odds scale, including confidence intervals for any quantities that you estimate.

2. For the fitted model from the previous part, determine which individual variable effects are significant after correction for multiple testing. Correct for multiple tests using a procedure that controls the FWER and also correct using an approach that controls the FDR but not FWER. Discuss the findings.

3. Compare the factor model from the previous part with the simple logistic regression model that uses the midpoint of radiation as a continuous predictor. Compare the models using deviances and residuals plots. Which model do you prefer? Why?

4. Without re-fitting your statistical model, calculate the ratio of the odds of having leukemia comparing a radiation level of '100 to 199' and a radiation level of '50 to 99'. Show all your calculations.

   Hint: You should be able to write this ratio as a function of $\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}$, for some vector $\boldsymbol{a}$.

5. Using the fact that $\text{var}(\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}) = \boldsymbol{a}^T\mathbf{V}\boldsymbol{a}$, where $\mathbf{V}$ is the estimated covariance matrix for $\widehat{\boldsymbol{\beta}}$, calculate a 95% confidence interval for the odds ratio you calculated in the previous part. Show all your calculations.