STAT 961: Homework 4

Name

Due Friday, November 19 at 11:59pm

1 Instructions

Setup. Pull the latest version of this assignment from Github and set your working directory to `stat-961-fall-2021/homework/homework-4`. Consult the getting started guide if you need to brush up on `R`, `LaTeX`, or `Git`.

Collaboration. The collaboration policy is as stated on the Syllabus:

> "Students are permitted to work together on homework assignments, but solutions must be written up and submitted individually. Students must disclose any sources of assistance they received; furthermore, they are prohibited from verbatim copying from any source and from consulting solutions to problems that may be available online and/or from past iterations of the course."

In accordance with this policy,

Please list anyone you discussed this homework with:

Please list what external references you consulted (e.g. articles, books, or websites):

Writeup. Use this document as a starting point for your writeup, adding your solutions between `\begin{sol}` and `\end{sol}`. See the preparing reports guide for guidance on compilation, creation of figures and tables, and presentation quality. Show all the code you wrote to produce your numerical results, and include complete derivations typeset in LaTeX for the mathematical questions.

Programming. The `tidyverse` paradigm for data manipulation (`dplyr`) and plotting (`ggplot2`) are strongly encouraged, but points will not be deducted for using base `R`.

```
library(tidyverse)
```

Grading. Each sub-part of each problem will be worth 3 points: 0 points for no solution or completely wrong solution; 1 point for some progress; 2 points for a mostly correct solution; 3 points for a complete and correct solution modulo small flaws. The presentation quality of the solution for each problem (as exemplified by the guidelines in Section 3 of the preparing reports guide) will be evaluated out of an additional 3 points.

Submission. Compile your writeup to PDF and submit to Gradescope.

Problem 1. Inverting the Wald, likelihood ratio, and score tests for a Poisson GLM.

You have two email accounts: your personal one and your academic one. Last month, you received y_1 and y_2 emails in your personal and academic inboxes, respectively. Interested in the extent to which you receive more (or less) email in your academic inbox, you set up the following Poisson regression model:

$$y_i \overset{\text{ind}}{\sim} \text{Poi}(\mu_i); \quad \log \mu_i = \beta_0 + \beta_1 x_i; \quad i \in \{1, 2\},$$

where $x_i \in \{0, 1\}$ is an indicator for your academic inbox. Your goal is to build a level-α confidence interval for e^{β_1} (the factor by which the expected number of emails in your academic inbox exceeds that in your personal inbox), and to this end you will invert the Wald, likelihood ratio, and score tests.

(a) What is the unrestricted maximum likelihood estimate $(\widehat{\beta}_0, \widehat{\beta}_1)$? What are the corresponding fitted means $(\widehat{\mu}_1, \widehat{\mu}_2)$? What is the maximum likelihood estimate for β_0 if β_1 is fixed at some value $\beta_1^0 \in \mathbb{R}$? What are the corresponding fitted means? What do the fitted means reduce to when $\beta_1^0 = 0$, and why does this make sense?

(b) What is the large-sample normal approximation to the sampling distribution of $\widehat{\boldsymbol{\beta}}$? What is the resulting level-α Wald confidence interval for e^{β_1} (defined by transforming the endpoints of the Wald confidence interval for β_1)? Express your answer explicitly.

(c) Given some $\beta_1^0 \in \mathbb{R}$, what is the likelihood ratio test statistic for $H_0 : \beta_1 = \beta_1^0$? What is the level-$\alpha$ confidence interval for e^{β_1} that results from inverting this test? The endpoints of your interval may be specified as solutions to a nonlinear equation.

(d) Formulate the test $H_0 : \beta_1 = \beta_1^0$ as a goodness of fit test. What is the corresponding score test statistic? What is the level-α confidence interval for e^{β_1} that results from inverting this test? Express your answer explicitly.

Solution 1.

Problem 2. Comparing the three confidence interval constructions from Problem 1.

Let's use a numerical simulation to compare the three confidence interval constructions from Problem 1 in finite samples.

(a) Write functions called `get_ci_wald`, `get_ci_lrt`, and `get_ci_score` that take as arguments (`y_1`, `y_2`, `alpha`) and return the corresponding confidence intervals for e^{β_1}. If the confidence interval is undefined for a given pair (y_1, y_2), your function should return $(-\infty, \infty)$.

(b) To get a first sense of how the three intervals compare, compute level $\alpha = 0.05$ intervals for $(y_1, y_2) = (10^1, 10^1), (10^{1.5}, 10^{1.5}), \ldots, (10^5, 10^5)$. Plot the lower and upper endpoints of these intervals as functions of y_1 (you should arrive at a plot containing six curves, corresponding to the lower and upper endpoints of the three methods). Add a dashed horizontal line at the MLE for e^{β_1} (which is the same for each given pair (y_1, y_2)). How do the interval widths compare, both across methods and across (y_1, y_2) values?

(c) Next, calculate the average length and coverage of the three level $\alpha = 0.05$ confidence intervals for e^{β_1} in the following simulation setting. Set $(\mu_1, \mu_2) = (10^1, 10^1), (10^{1.5}, 10^{1.5}), \ldots, (10^5, 10^5)$. For each pair (μ_1, μ_2), generate 5000 realizations of (y_1, y_2) and compute the three confidence intervals for each realization. Plot the average length and coverage for each of the three interval constructions as a function of μ_1 (please omit the undefined/infinite-length intervals from the calculations of length and coverage). Compare and contrast the average lengths and coverages of the three constructions, both across methods and across (μ_1, μ_2) values.

(d) Last month you received 60 emails in your personal inbox and 90 in your academic inbox. Pick one of the three confidence interval constructions above that you feel has good coverage and small width. According to this construction, what is the confidence interval for e^{β_1}? Can you reject the null hypothesis that the two inboxes receive emails at the same rate?

Solution 2.

Problem 3. Case study: Child development.

Children were asked to build towers as high as they could out of cubical and cylindrical blocks.[1] The number of blocks used and the time taken were recorded (see `blocks_data` below). In this problem, only consider the number of blocks used and the age of the child.

```
blocks_data = read_tsv("../../data/blocks.tsv")
print(blocks_data, n = 5)

## # A tibble: 100 x 6
##   Child Number  Time Trial Shape   Age
##   <chr>  <dbl> <dbl> <dbl> <chr> <dbl>
## 1 A         11  30       1 Cube   4.67
## 2 B          9  19       1 Cube   5
## 3 C          8  18.6     1 Cube   4.42
## 4 D          9  23       1 Cube   4.33
## 5 E         10  29       1 Cube   4.33
## # ... with 95 more rows
```

(a) Create a scatter plot of blocks used versus age; since there are exact duplicates of `(Number, Age)` in the data, use `geom_count()` instead of `geom_point()`. Propose a GLM to model the number of blocks used as a function of age.

(b) Fit this GLM using R, and write down the fitted model. Determine the standard error for each regression parameter, and find the 95% Wald confidence intervals for the regression coefficients.

(c) Use Wald, score, and likelihood ratio tests to determine if age seems necessary in the model. Compare the results and comment.

(d) Plot the number of blocks used against age as in part (a), adding the relationship described by the fitted model as well as lines indicating the lower and upper 95% confidence intervals for these fitted values.

Acknowledgment: This problem was drawn from "Generalized Linear Models With Examples in R" (Dunn and Smyth, 2018).

Solution 3.

[1] Johnson, B., Courtney, D.M.: Tower building. Child Development 2(2), 161–162 (1931).