Final Analysis Report

Teamname: Conan the Barbayesian

People: Lars Andersen Bratholm, University of Bristol, School of Chemistry and School of Mathematics, United Kingdom

Conflicts of Interest: None

Contributions: All me

Our final analysis

Description of analysis

Using my definition of depression and computer use, I define the odds ratio with a maximum likelihood estimate and get the confidense interval from an asymptotic normal approximation to the log odds ratio. Any samples missing the *has_dep_diag*, *comp_wend* or *comp_week* columns were ignored.

Analytical choices

1. Outcomes:

   a) I use the *has_dep_diag* column to indicate if the sample (person / data point) is depressed or not.

   b) The percentages with depression using this definition agrees well with other commonly quoted estimates and I don’t think that it’s a huge worry if it’s slightly inaccurate as there’s no reason to believe that there’s a correlation between computer use and the *accuracy* of a depression diagnosis.

2. Exposures:

   a) I classify computer use into three classes using the *comp_wend* and *comp_week* columns.
   1) **Low**, where both *comp_wend* and *comp_wend* is less or equal to ’less than 1 hour', 2) **High**, where one of *comp_wend* and *comp_week* is '3 or more hours' and the other is '1-2 hours' or greater. 3) **Other**, which indicates intermediate values.
   This is illustrated in below table, where the green bins indicate the combinations of *comp_wend* and *comp_week* that corresponds to **low** computer use and the red bins corresponds to **high** computer use.

	0	<1	1-2	3+
0	🟩	🟩		
<1	🟩	🟩		
1-2				🟥
3+			🟥	🟥

   b) This classification of computer use was completely arbitrary and was done before any actual analysis were done.

5. Missing data:

   a) Any sample that is missing an indicator for either depression or computer use is ignored.

   b) This was done since it was easier and faster. The confidence interval could likely be decreased by modelling all the features and integrating out missing features, however this simpler model was all I managed to finish.

7. Statistical models used

   a) Maximum likelihood estimate (MLE)

   b) The log odds ratio is

$$L = \log\left(\frac{p\left(D = Y \mid CU = H\right)/p\left(D = N \mid CU = H\right)}{p\left(D = Y \mid CU = L\right)/p\left(D = N \mid CU = L\right)}\right),$$

where 'Depression' (D), can take values 'Yes' (Y) or 'No' (N), and 'Computer use' (CU), can take values 'High' (H) or 'Low' (L). I assume that the log-odds-ratio is approximately normal and use the maximum likelihood estimates (MLE) of the mean (the MLE of probabilities are the counts divided by the total number) and approximate the variance as (resulting from the asymptotic normal approximation to the log odds ratio)

$$\sigma^2 = \frac{1}{n_{YH}} + \frac{1}{n_{YL}} + \frac{1}{n_{NH}} + \frac{1}{n_{NL}}$$

where n_{YH} is the number of samples with depression and high computer use etc..

c) I found a few places where this approximation were used, including the wikipedia page for the odds ratio. However I didn’t go deep enough to find an explicit derivation or papers that uses this approximation.

Results

Variable definition	Value	95% CI	p-Value (2-sided)
Odds ratio	0.99	0.70-1.41	0.96
AIC	1068.5		

Regarding AIC value, I don't really have any hidden variables so I don't think an information criterion makes sense on it's own, especially since I’m throwing away most of the data. The maximum likelihood used to compute the AIC value uses only the samples that were included in estimating the odds ratio, and was computed from

$$L = \prod_i \left(\frac{n_i}{N}\right)^{n_i},$$

where n_i is the counts in a given bin (e.g. one bin being depression AND high computer use) and N is the total number of samples used. In the AIC, k=3 since there are 4 bins.

Survey

Demographics

1. Lars Andersen Bratholm
2. 35
3. PhD
4. PostDoc
5. 6
6. No
7. No
8. No
9. No
10. It's independent in the sense that I haven't communicated with any teams, so in that regard 7. I expect all the analysis to be highly correlated though, so from that meaning 1. If you mean just how unique my solution is, then probably 2.

Statistical expertise

1. My PhD was related to bio-informatics, so I guess?
2. I've taught statistical mechanics once, but I'm unsure if that would count as a statistics course, so no.
3. No
4. No
5. Yes
    a. 5, but basically monte sampling of posterior likelihood and bayesian linear regression, so I don't know if that counts as statistical models.
6. 3.

Research question expertise

1. No
2. 1.

Subjective beliefs

1. 4.
2. 5.
3. 4.
4. 5.

Authorship

1. Yes, under Lars Andersen Bratholm

Feedback

1. I understand your reasoning behind having a very specific format of the accompaigning code, but I also felt that it was very directed towards R. Maybe an alternative could have been to make some test cases that the code had to pass, so you had more freedom to write

your code in a more modular way? In any case, test cases to automatically check if the code follows the format you want would have been good.

2. I think that there was some edits to some of the files at some point that didn't get accompanied with an email (which you did all the other times). I might be mistaken, but it was confusing at the time.
3. The data dictionary was some weird R format I think, and not csv, since each line didn't correspond to a feature/column, but this is minor
4. I felt like there was missing some details on the structure of the final report, since there was only a preview on OSF, but that is minor.
5. In general, most things have been pretty clear and the minor misunderstandings on my part could easily be cleared up if I didn't always finish things last minute, so thank you for arranging this.