Data Analysis Prelim Exam

Department of Statistics, University of Chicago

Assigned Tuesday Sept 13 2022 at 12pm (noon); Due Thursday Sept 15 2022 at 12pm (noon). The exam is "open book" but **you must not discuss any aspect of this exam with anyone else.** Questions of clarification should be addressed to rina@uchicago.edu, jingshuw@uchicago.edu and mstephens@uchicago.edu

1 Data

The data set for this exam is taken from the paper

> *Immunogenicity of rVSVΔG-ZEBOV-GP Ebola vaccination in exposed and potentially exposed persons in the Democratic Republic of the Congo*, Hoff et al, PNAS 2022, available at `https://www.pnas.org/doi/10.1073/pnas.2118895119`

The data set consists of data from a study examining participants' response to an Ebola vaccine. For the vaccine to be effective, the participant's immune system needs to generate an antibody response after the administration of the vaccine, and for the effectiveness to last over time, this response should persist long after the vaccine was administered. This study measures antibody levels on the day of the vaccine, 3 weeks after the vaccine, and 6 months after the vaccine, to track the antibody response.

The data are contained in the file `pnas.2118895119.sd01.xlsx`. The variable names and descriptions are in the file `pnas.2118895119.sd02.xlsx`. The variables measured in the study are:

- Demographic variables: `age`, `age_cat4` (age categorized into bins), `sex`, `educ` (education level, categorical), `civ_stat` (marital status)
- Other features: `hcw_curr` (indicating whether the participant is a healthcare worker), `BL_ever_cc_ebola` (indicating whether the participant has ever had a close contact for Ebola exposure), `vax_days` (days since receiving the vaccination, at the time of the first measurement)
- In the original study, each participant had their antibody levels measured multiple times at each of the three timepoints (after 0 days, after 21 days, after 6 months), but the publicly available data set only provides the variables `day0_FANG6pt_median`, `day21_FANG6pt_median`, `M6_FANG6pt_median`, which are the median measurements at each of the three timepoints
- `day0_fang6pt`, `day21_fang6pt`, `M6_fang6pt` indicate whether the measurements were successfully obtained or are missing at each of the three timepoints

2 Questions

Your report should include code for all plots and results that you present. The report should be *concise, well-organized, and easy to read.* For example, if you examine plots across a long list of different subsets of the data, you may choose to show only a small selection of representative plots.

1. Please read the original paper and briefly summarize the study. Include a high-level description of the data collection, variables measured, the main trends observed or conclusions drawn by the authors, etc. Highlight some of the main statistical issues, as you see them, that might be present in this data set or in the analysis.

2. a) Figure 1 in the paper uses a boxplot to compare the distributions of Antibody Titer at day 0, day 21 and month 6. Produce a plot that uses histograms instead of boxplots to perform a similar comparison. Indicate the LLOQ and 4LLOQ values as blue and green dashed lines as is done in Figure 1 in the paper.

   b) The paper reports that 87.2% of samples had an antibody response at visit 2 and 95.6% demonstrated an antibody persistence at visit 3. Comment on the criteria used to obtain these numbers. What issues might you bring up for discussion with the authors if you had the opportunity?

3. Use linear regression to investigate how `day21_FANG6pt_median`, the antibody level after 3 weeks, depends on other relevant covariates, and describe your findings. You should explain any choices made throughout the process, including choices about which covariates are reasonable to include in the model. Note that the data provided is not the same as the original data (specifically, the original study had repeated measurements of antibody level for each participant at each time point, but we are only given the median value); due to this, or to choices made in your analysis, your results may agree or may differ from those in the paper. For this problem, you may ignore the issue of missing data (i.e., simply remove data points as needed, without considering the biases that this might potentially introduce).

4. To investigate the relationship between being a health-care worker and age we ran the following logistic regression with response `hcw_curr` and covariates `age` and `age_cat4`:

```
> summary(glm(hcw_curr~age+as.factor(age_cat4),data=data,family=binomial(link = "logit")))

Call:
glm(formula = hcw_curr ~ age + as.factor(age_cat4), family = binomial(link = "logit"),
    data = data)
...
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.89557    0.52900  -1.693 0.090462 .
age                  -0.04157    0.02335  -1.780 0.075073 .
as.factor(age_cat4)2  1.55540    0.42084   3.696 0.000219 ***
as.factor(age_cat4)3  2.10187    0.54316   3.870 0.000109 ***
as.factor(age_cat4)4  2.79035    0.73919   3.775 0.000160 ***
as.factor(age_cat4)5  2.63768    1.07524   2.453 0.014163 *
...
```

   Note that in this fitted model, both the continuous covariate `age` and its categorical version `age_cat4` have significant p-values at level 0.1. Explain why the resulting fitted model is not a natural model, and explain what type of relationship between `hcw_curr` and `age` might lead to the fitted coefficients that we see here. Then propose and run an alternative model that is more natural to capture the relationship between `hcw_curr` and `age`.

5. In this question, we aim to impute missing antibody titer measurements for the 21-day and 6-month visits. To predict the missing antibody titer levels at day 21 and month 6 due to loss to follow-up, we may want to use a mixed-effect model to take into account the fact that each individual can have three measurements (at day 0, day 21, month 6). Let y_{ij} with $j = 0, 1, 2$ be the antibody titer (6pt FANG Median) of individual i at day 0, day 21, and month 6, respectively. Denote X_i as the vector of predictors including age, sex, marital status, health care workers status, vaccination days, education and Ebola close contact or not and consider the following linear mixed-effect model:

$$\log(y_{ij}) = X_i^\top \beta_j + \gamma_j + u_i + \epsilon_{ij}.$$

   (a) Implement the above model to perform imputation of the missing measurements.

   (b) What are the assumptions of this model? Is it a reasonable model to fit for this data?

6. Now we will investigate the missingness pattern of the antibody titers at day 21. Let y_i denote the day 21 antibody level for the ith individual, `day21_FANG6pt_median`, and let X_i denote all features measurable on day 0 for the ith individual (e.g., sex, age, antibody level on day 0, etc). Let $z_i \in \{0, 1\}$ denote whether the day 21 antibody level y_i was missing or not.

   For a response y_i, missing completely at random (MCAR) means that whether the response is missing or not ($z_i \in \{0, 1\}$) are independent from the data (i.e., $(X_i, y_i) \perp\!\!\!\perp z_i$). In contrast, missing at random (MAR) after conditioning on covariates X_i is a weaker condition, and it means that $y_i \perp\!\!\!\perp z_i \mid X_i$. Note that MCAR is a special case of MAR.[1]

   Assuming that the MAR assumption holds, we will now ask whether the stronger MCAR assumption also holds. Please diagnose and evaluate whether the missingness pattern of antibody titers at day 21 is likely missing completely at random, or missing at random only after conditioning on other features of the individuals.

[1] For this question, you can ignore issues of missingness in the X_i's and simply discard points with missing values in the X_i's.