Quiz 3 Solution

Third Term, 2020-2021

The purpose of this quiz is to assess your knowledge of the course materials covered during weeks 5 and 6 of class and covered in Problem Set 3.

Instructions:

- This is an open book quiz; you may consult your course notes and handouts.
- You should not discuss this quiz with any other student during on Friday March 12th through Sunday March 14th.
- This quiz is designed to be completed in 20-30 minutes.
- You may provide your solution by annotating the pdf version of this quiz or writing your solution on paper and submitting a picture of your solution.

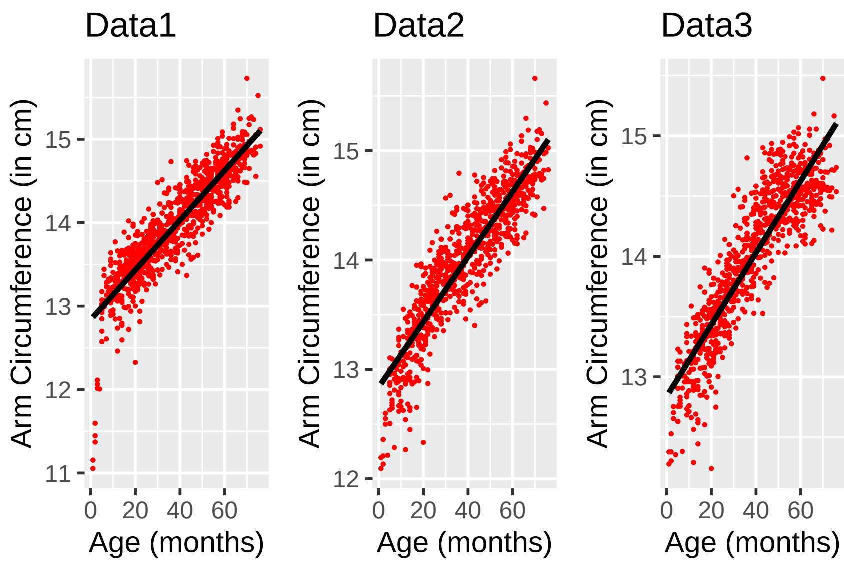
By signing my name, I enter agree to abide by the instructions above and the Johns Hopkins University School of Public Health Academic Code:

Name (Print): ____

Signature: ____

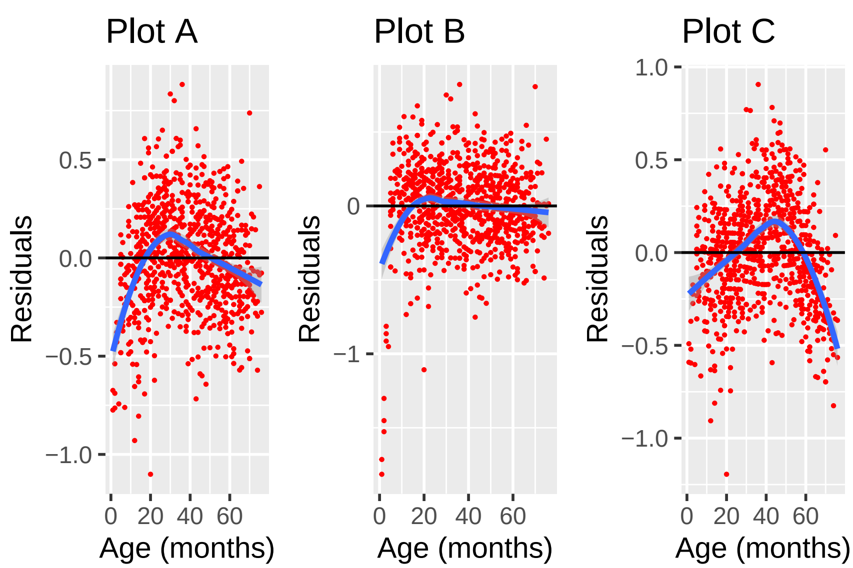
Question 1:

Below you will find scatterplots of three datasets (Data1, Data2 and Data3). A simple linear regression of arm circumference vs. age was fit to each dataset (see the fitted line on the each scatterplot).



Below you will find the three plots of residuals vs. age that correspond to the simple linear regression models fit above. Match (one to one) the residual vs. age plot to each of the datasets.

- The residual plot for Data1 is **Plot B**__________.
- The residual plot for Data2 is **Plot A**__________.
- The residual plot for Data3 is **Plot C**__________.



Question 2:

I fit the following mean model to the longitudinal Nepali Anthropometry dataset:

$$WT_{ij} = \beta_0 + \beta_1 \times (age_{ij} - 6) + \beta_2 \times (age_{ij} - 6)^+ + \epsilon_{ij}$$

Where WT_{ij} is the weight for child i at follow-up j, $j = 1, 2, 3, 4, 5$, $(age_{ij} - 6)$ and $(age_{ij} - 6)^+$ defines a linear spline model with knot at 6 months.

I fit this model using ordinary least squares and then computed $Corr(\hat{\epsilon}_{ij}, \hat{\epsilon}_{ik})$ for each j and k in $1, 2, 3, 4, 5$.

The correlation matrix is displayed below.

```
##        visit1 visit2 visit3 visit4 visit5
## visit1  1.000  0.936  0.924  0.913  0.866
## visit2  0.936  1.000  0.939  0.908  0.873
## visit3  0.924  0.939  1.000  0.950  0.927
## visit4  0.913  0.908  0.950  1.000  0.937
## visit5  0.866  0.873  0.927  0.937  1.000
```

Interpret the value in the 2nd row, 3rd column of the correlation matrix.

Answer The value 0.939 represents the correlation between residual information in weight after removing age information at follow-up visit 2 and 3.

Question 3:

Suppose I suggest that instead of the ordinary least squares model, we should fit a weighted least squares where we assume an *AR1* correlation model for the residuals for child i, i.e. $Corr(\epsilon_{ij}, \epsilon_{ik}) = \rho^{|j-k|}$.

Use the empirical correlation matrix above to argue for or against this recommendation.

Answer If we assumed an *AR1* model for the correlation structure, then our estimate of ρ would be 0.94 (the average of the lag correlation estimates, 0.936, 0.937, 0.939, 0.950). This would imply that the lag 2 correlation should be roughly $\rho^2 = 0.94^2 = 0.88$, the lag 3 correlation should be roughly $\rho^3 = 0.94^3 = 0.83$ and the lag 4 correlation should be $\rho^4 = 0.94^4 = 0.78$.

Based on the empirical correlation matrix above, the *AR1* would underestimate the correlations at lag 2, 3, and 4. Therefore, I would argue that the *AR1* model is not a reasonable model for the observed correlation.

Question 4:

I fit two weighted least squares models to the data using the *gee* package in R. I assumed a working exchangeable correlation model and a working AR1 correlation model.

Below you are presented with a table containing the coefficient estimates and in parentheses the naive standard error followed by the robust standard error. Which of the two correlation models do you think fits the data "best" and why?

```
##                        Exchangeable                AR1
## Intercept      4.142(0.329, 0.322)  4.057(0.39, 0.249)
## (age_ij - 6)   0.484(0.051, 0.052)   0.493(0.06, 0.04)
## (age_ij - 6)+ -0.354(0.051, 0.052) -0.36(0.061, 0.041)
```

Answer: We notice two things: 1) there is very little difference between the standard error estimates computed assuming an exchangeable working correlation model and the robust standard error estimates (computed using an exchangeable working correlation model), 2) the standard error estimates computed assuming an AR1 working correlation model are considerably different than the robust standard error estimates (computed using an AR1 working correlation model).

Given that when we assume exchangeable, the model based and robust standard error estimates are more similar; then the data better support the exchangeable correlation model compared to the AR1 model.