**BIOST 2050: Longitudinal and Clustered Data Analysis**

**Homework Assignment #4**

**Due: Wednesday, October 18, 2023**

A longitudinal study was conducted to investigate the growth in height of infants through their first 2 years since birth. A total of 200 children were included in this study. The height of each child was measured once every 2 months from birth up to the age of 24 months. In the dataset, each child has up to 13 repeated measurements. The primary outcome of interest here is height-to-age z-score (HAZ) and the primary objective of the study is to examine whether social economic status (SES) groups (low vs high) differ in the rate of change in HAZ. At each visit, a HAZ value of less than -2 was defined as stunting (meaning that the child was growing at a reduced than normal rate). As a secondary objective, the study is also interested to examine whether the low SES group has a higher chance of stunting compared to the high SES group.

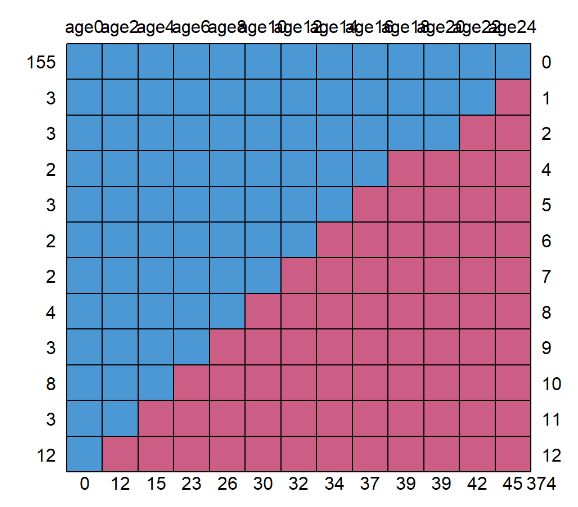
Dataset **hazdata** contains the following variables:

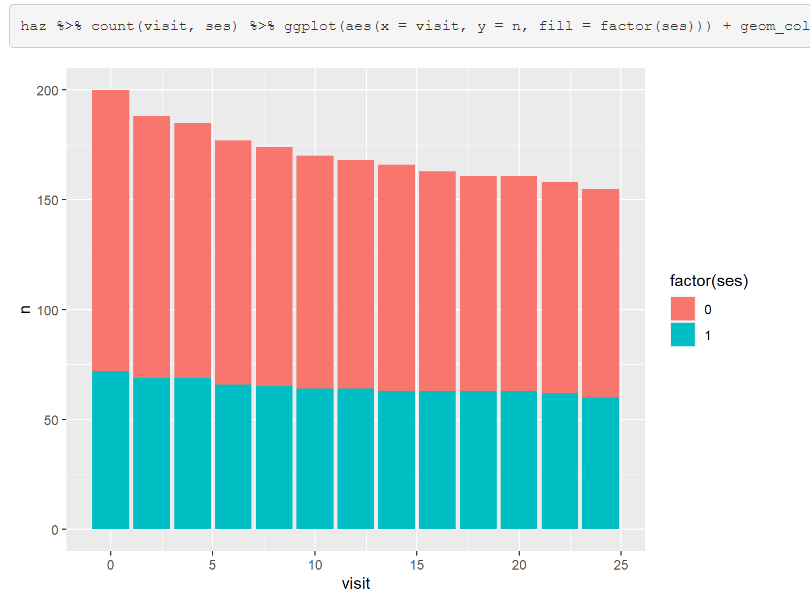
Variable List:

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| **id**  **ses**  **gender**  **visit**  **haz**  **stunt** | Child ID  Social economic status (1 = High, 0 = Low)  Gender (1 = male, 0 = female)  Age (rounded, in months) at each visit/measurement  Height-to-age z-score  Whether stunted or not for each visit (1 = stunted, 0 = not stunted) |

The analysis for this assignment is to test the primary objective and is based on the outcome variable “haz.”

1. Descriptive statistical analyses.
2. Provide descriptions on the missing data for the variable of height-to-age z-score. (5 pts)

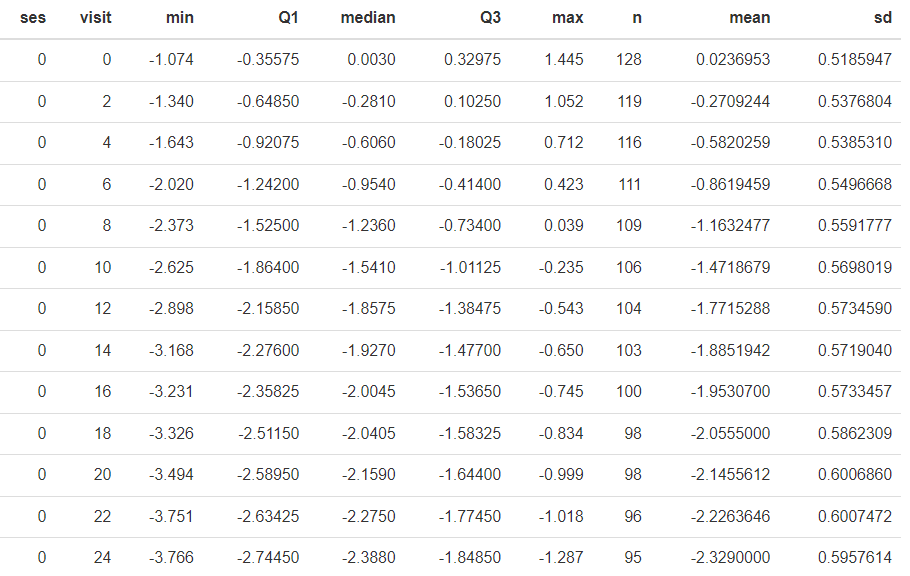




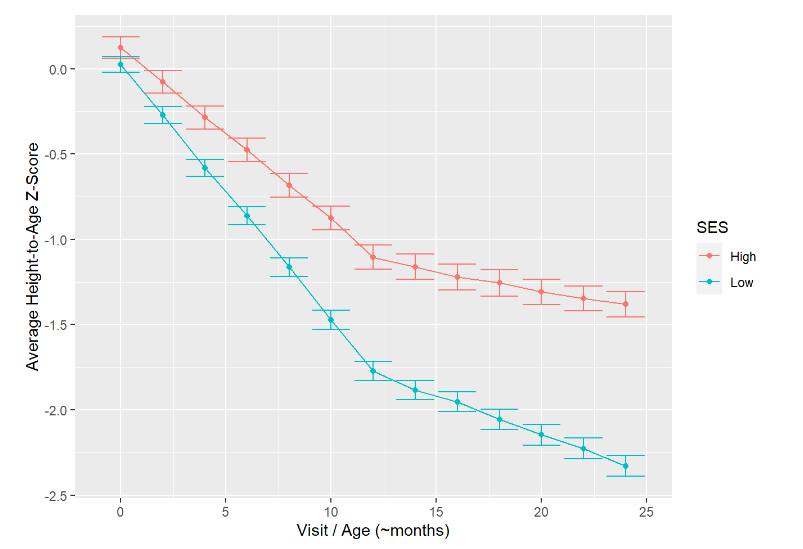
3/4 of the subjects have information on haz at all 13 visits. Anywhere from 2-12 participants have monotonic missingness from the last visit, starting all the way up to the second visit. No participants are missing visit == 0.

It seems that there is a steeper drop out over time of participants from the low SES group compared to the high SES group, particularly at the start of the study. At the same time, there are more participants from the low SES group (and therefore more to dropout) compared to the high SES group.

1. Provide adequate descriptive statistics (numerical and graphical) to show the mean trajectories of the height-to-age z score for each of the SES groups. Comment on your findings. Discuss how you will model the time variable. (10 pts)

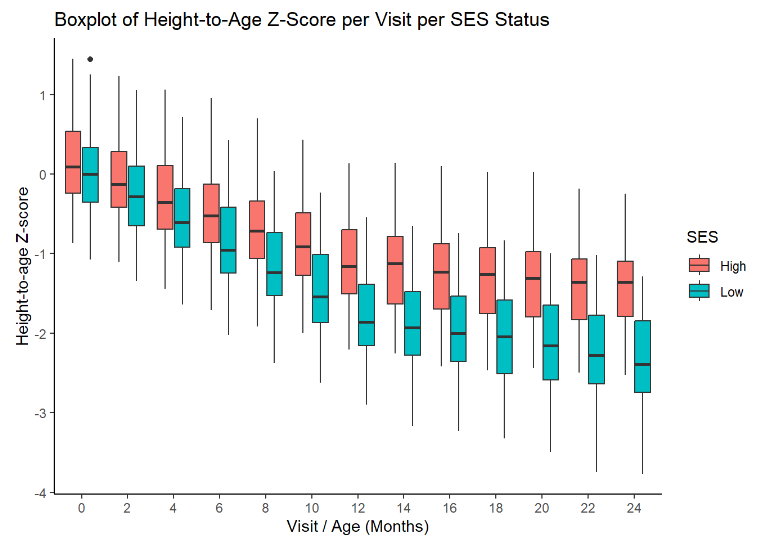


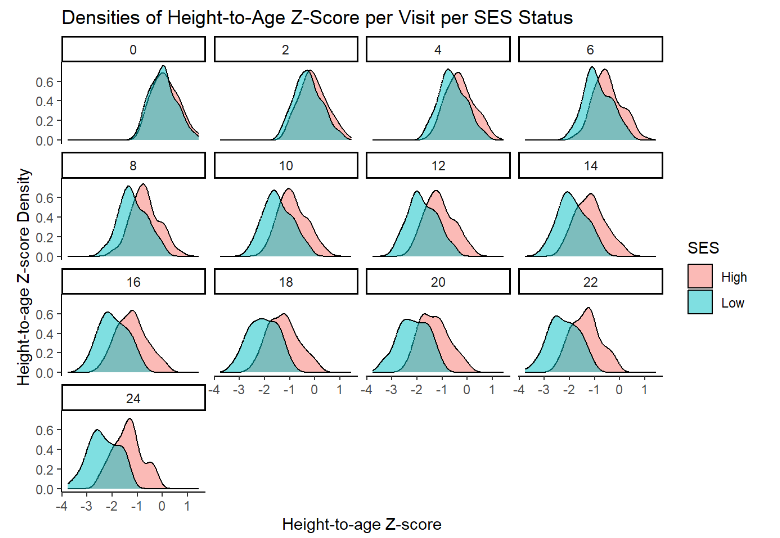


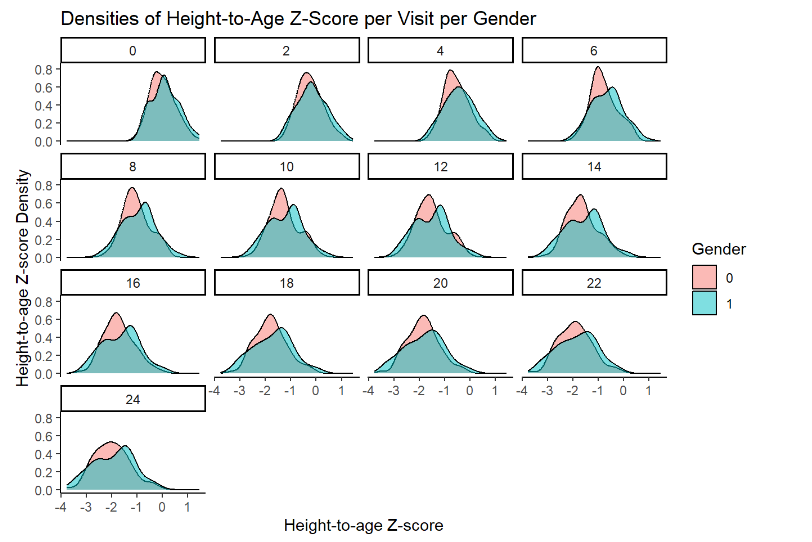


It is clear that the mean Height-to-Age Z-Score for children in the Low SES group decreases more rapidly in the first 24 months than for children in the High SES group. Around the Visit/Age of 12 months, there is a clear bend in the originally negative linear trajectory of the mean scores by time. This signals to me to include a quadratic term for time, in an attempt to capture this curvature.

1. Give an exploratory assessment on whether the high SES group has a higher value of the height-to-age z score (over time) as compared to the low SES group. (5 pts)







It was clear from the figure in 1b) that mean haz decreases more quickly over time in the Low SES group than in the High SES group. It seems that this pattern holds when displayed in terms of boxplots (median, IQR, +/- 1.5\*IQR). This means it is not just a few outliers in the low SES group bringing the mean haz downwards more than in the high SES group. The boxplots also demonstrates that there is some stunting in the high SES group, but not as much as in the low SES group.

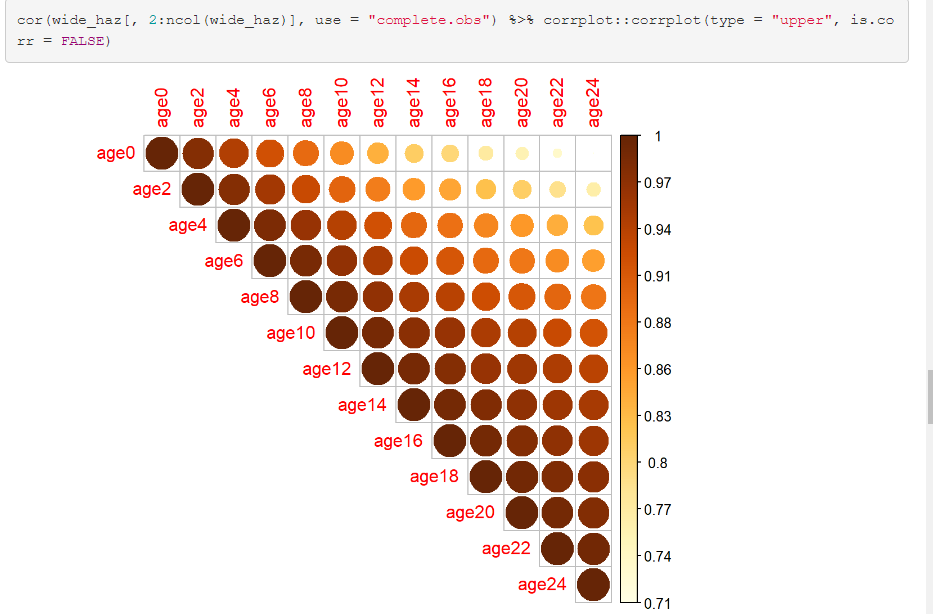
We can observe from the densities that haz, the z-score, is indeed normalized.

All of these figures motivates the usage of an interaction term between SES and time and main effects of SES/Time (i.e., the main/secondary objectives of the study). The densities by gender don’t indicate much of an effect of gender, but this will still be controlled for.

1. The objective is to assess whether social economic status (SES) groups differ in the rate of change in HAZ, adjusting for SES and gender.
2. Fit an appropriate generalized linear model with the generalized estimating equation (GEE) that you will use to address the primary study question. Explicitly write out your model by specifying the model of the mean structure and the working variance-covariance matrix. State all the assumptions necessary for your model. Additionally, comment on your decision regarding how to treat the time variable and how to handle the within-subject correlations. (24 pts)

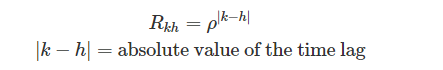
It is clear the outcome (a z-score) is most close to normally distributed. Therefore, the link function is the identity function, and the mean structure/function is from the Gaussian family of distributions.

(= \eta (the linear predictor))



It is clear that there is a positive and diminishing autocorrelation across time, from (~1 to ~.71), as the time lag increases. Therefore, I will attempt to use an AR(1) working correlation matrix.

This is defined as (for all subjects):



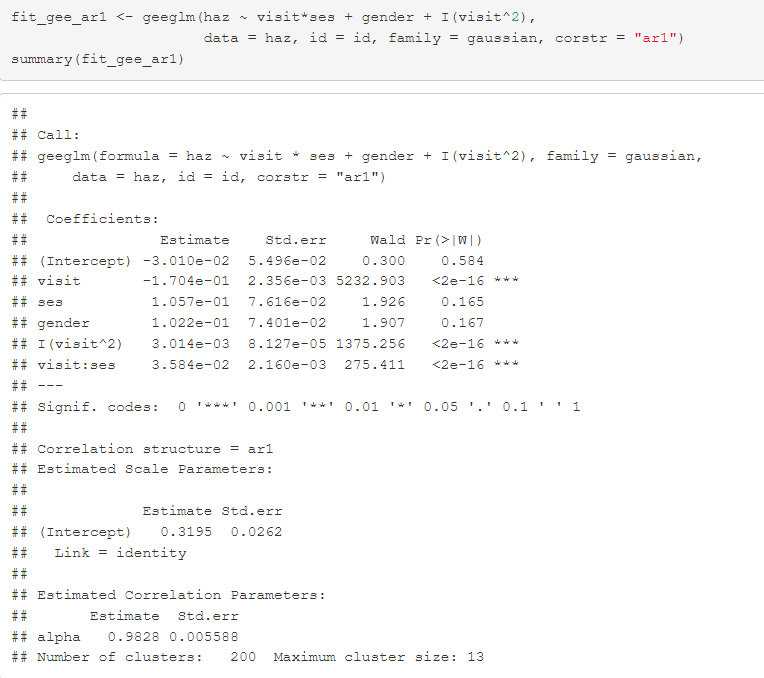
rho is the only parameter for the working correlation matrix. Exponentiation by the absolute difference between timepoints indicates that the larger the absolute time lag is the smaller the autocorrelation is.

Necessary assumptions for my model are:

* That the autocorrelations do not drop to zero above any threshold value of the absolute time lag, as in the Toeplitz structure.
* There are autocorrelations over time and that they decay exponentially as a function of the absolute time lag
* The outcome is normally distributed and independent between different children.
* Data is missing completely at random (MCAR)
* Panel data (non-continuous time)

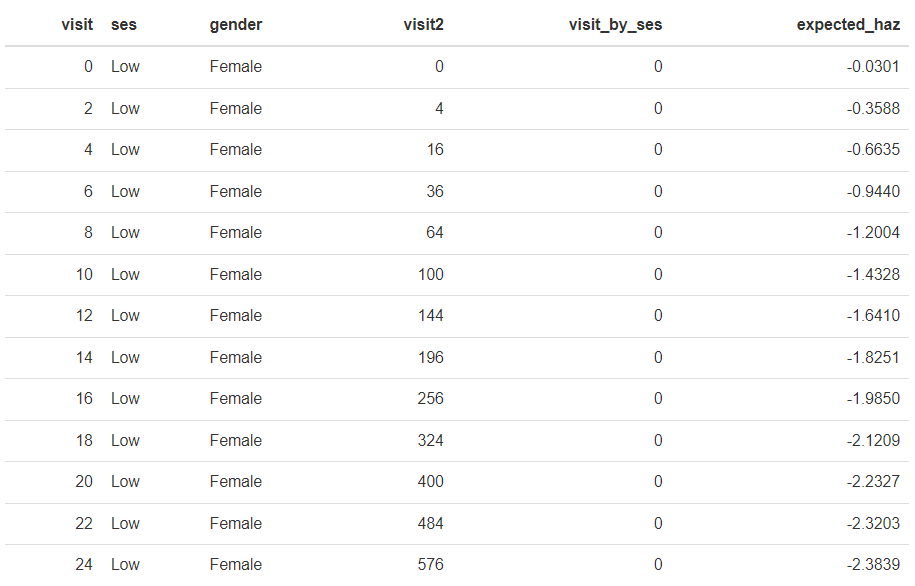
I treat time as a main effect, an interaction effect with ses, and a quadratic effect. I include the main effect because of the clear trend in the figures in part 1, I include the interaction because of the clear difference in trajectories between SES groups (and the objective of the study). I include the quadratic effect of time because of the non-linear curvature in haz trajectory.

I choose to handle within-subject correlations as singular autocorrelation that decays with the lag between measurements due to the observed autocorrelation matrix for haz showing autocorrelation that decays with larger lags.

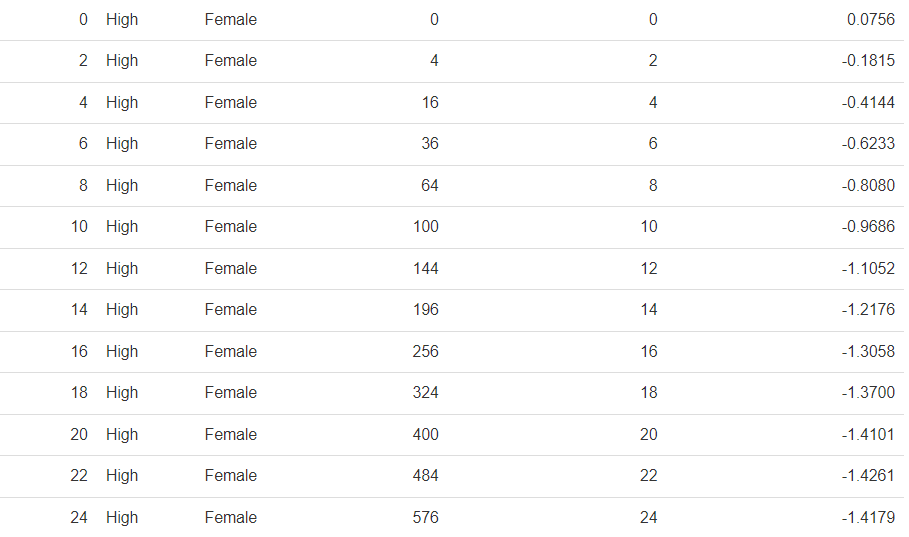


1. Based on the fitted model in part (a), how does HAZ change over the ages? Does this change differ for the different SES groups? Justify your answer. (20 pts)

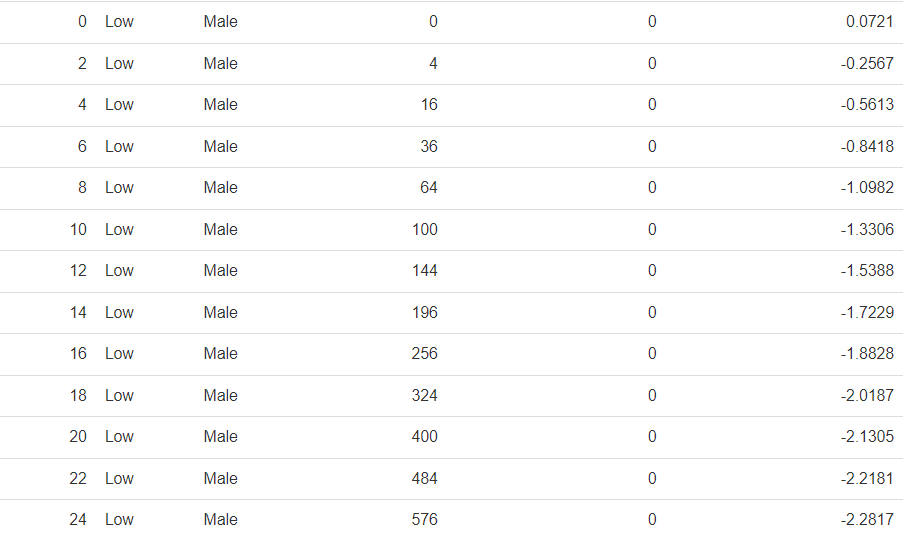
By the linear age main effect and ses:age interaction effect alone, all individuals show a decrease in haz for each unit (month) increase in age, but those with a high SES show a less steep decrease in haz for each unit(month) increase in age compared to those with a low SES. Including the effect of quadratic age appears to “dampen” the steepness of the linear/interaction effect of age for each unit increase in quadratic age. This can be observed by looking at expected haz for all possible combinations of our variables:



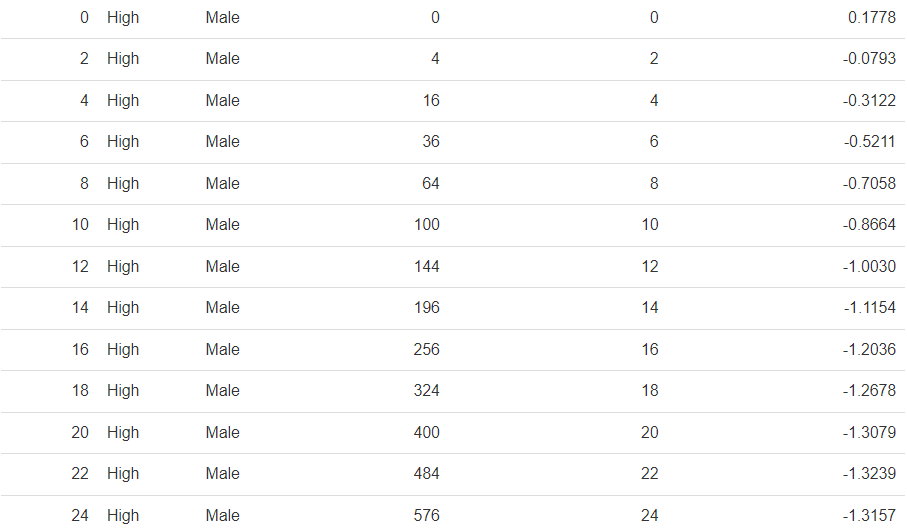




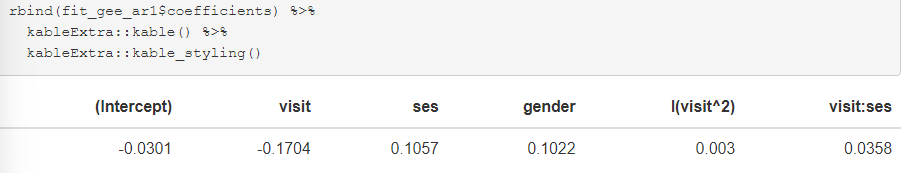








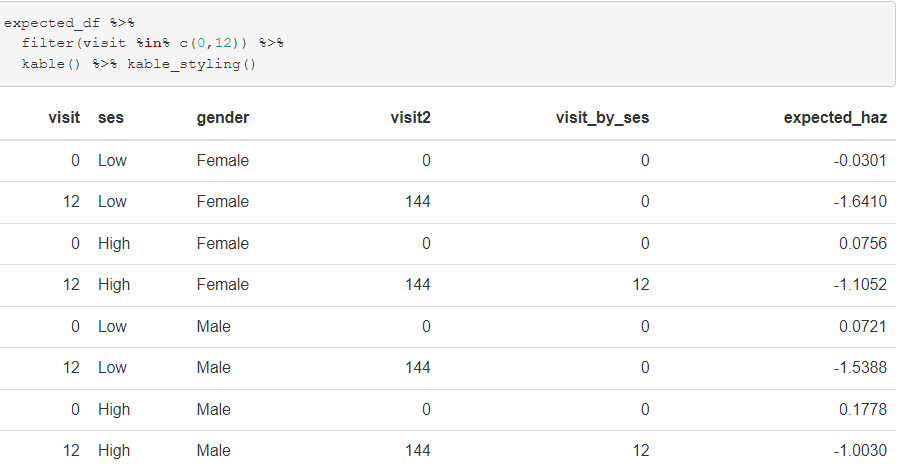
Where each row of each column is multiplied by its corresponding coefficients + the intercept below to obtain the expected hazard ratio as a function of time within each SES status and gender:



As can be seen from the expected values, holding gender constant, there is a steep decreases in HAZ at first, which is “dampened” as visit^2 grows in magnitude over time. The stee decrease is steeper for low SES participants, and dampened less for low SES participants, compared to high SES participants.

1. Based on the fitted model in part (a), for each SES group, what is the estimated HAZ between age one year versus birth? (6 pts)

We can filter the expected df to see the estimated HAZ between age one year and birth for each SES group and for each gender, since we must make a choice about gender since it affects estimated HAZ. It is coded so it is simple enough to display both, but I could answer this question just holding one constant.



We can see that, regardless of gender, the decrease in estimated HAZ from birth to one year is larger for low SES compared to high SES.

1. Fit an appropriate covariance pattern model that you will use to address the primary study question. Explicitly write out your model by specifying the model of the mean structure and the model of the variance-covariance matrix. State all the assumptions necessary for your model. (24 pts)

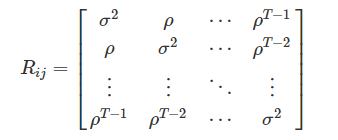
The covariance pattern model for each occasion j for subject i is:



The mean structure for each occasion j for subject i is:



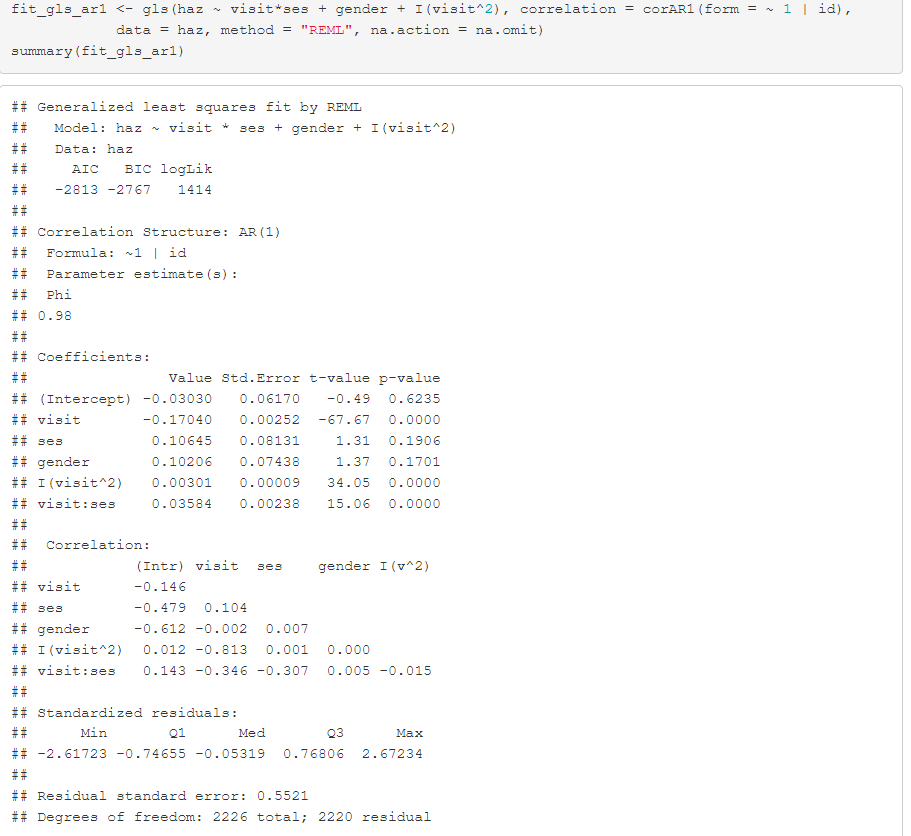
The variance-covariance structure AR(1) is:



Where the rows and columns refer to subject i for occasion j and the rho changes as an exponential function of time.

Necessary assumptions for my model are:

* That the autocorrelations do not drop to zero above any threshold value of the absolute time lag, as in the Toeplitz structure.
* There are autocorrelations over time and that they decay exponentially as a function of the absolute time lag
* The outcome is normally distributed and independent between different children.
* If data is not MCAR, then the covariance structure must be correct to obtain consistent beta estimates
* Panel data (non-continuous time)



1. Do you the findings of the primary objective change when using the covariance pattern model compared to GEE? (6 pts)

No, they don’t, because the coefficients are all virtually the same.

