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

**Homework Assignment #2**

**Due: Friday, September 29, 2023**

Answer the questions and justify your answers. Note that you will receive a major deduction if you answer a question by giving only the software output without justifying your answer.

An observational cohort study was conducted for individuals who underwent bariatric surgery at 10 US hospitals in 6 geographically diverse clinical sites. The study aimed to investigate if the surgery improves kidney functioning and if the improvement differs between Roux-en-Y gastric bypass (RYGB) surgery and laparoscopic adjustable gastric band (LAGB) surgery. The study was also intended to identify factors associated with improvement in kidney functioning in the first 3 years after bariatric surgery.

Participants were recruited between February 21, 2005, and February 17, 2009. Those who underwent first-time bariatric surgical procedures between March 14, 2006, and April 24, 2009 were followed for 3 years (through October 24, 2012). Assessments were conducted before the procedure and annually thereafter.

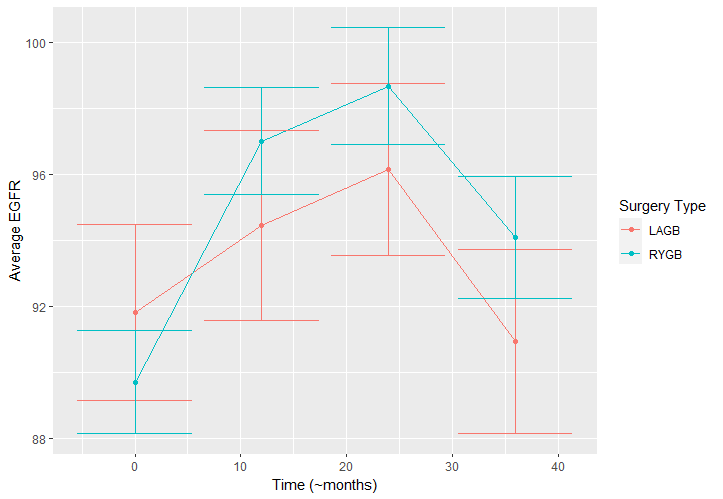
Estimated glomerular filtration rate (eGFR) is a measurement for kidney function and is used to determine the stage of kidney disease. The lower the eGFR the worse the kidney function is. An eGFR score of 90 or above suggests a healthy kidney; any value less than that would indicate some loss of kidney function.

Dataset egfr contains a subset of the original study dataset that contains variables specific to addressing the above aims. The following are the variables of interest.

|  |  |
| --- | --- |
| Variable | Description |
| site | Clinical site ID |
| subject | Subject ID |
| surgery | Bariatric Surgery Type  **1** = "RYGB", **2** = "LAGB" |
| time | Visit time in months  **0** = Baseline, **12** = 12-Month, **24** = 24-Month, **36** = 36-Month |
| age | Participant's age (years) at baseline |
| income3 | Household income at each visit  **0**= Less than $25,000, **1**= $25,000-$99,999, **2**= $100,000 or greater |
| egfr | Estimated glomerular filtration rate (eGFR) at each visit |

Investigators are interested in evaluating whether one type of surgery resulted in greater improvement in eGFR than the other type of surgery, and whether eGFR values were consistently better over time for one type of surgery as compared to the other, assuming that the effects of adjusting covariates on eGFR were equal in different surgery types. Adjusting covariates include age and income level. Note that income level is a time-varying measurement.

* 1. **Provide the mean trajectory plot of eGFR for each surgery type. (15 pts)**



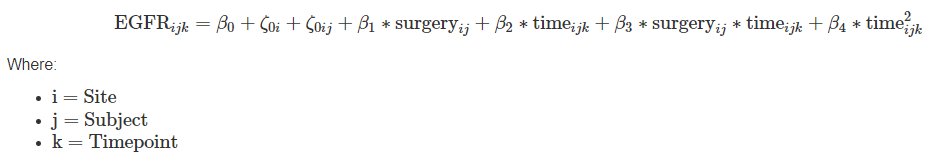
Plotted with standard error bars for the mean, over time, by surgery type. There appears to be a quadratic pattern (inverted U) in mean EGFR over time. The average EGFR at baseline between groups appears insignificant. There appears to be a significantly higher EGFR after RYGB compared to LAGB at timepoint 12, while at timepoint 36, there seems to be significantly lower EGFR after RYGB compared to LAGB.

* 1. **Describe a 3-level hierarchical clustering model that you choose to address the primary questions and the associated assumptions. Explicitly specify the fixed effects and the random effects, if any, and justify your choices. (25 pts)**

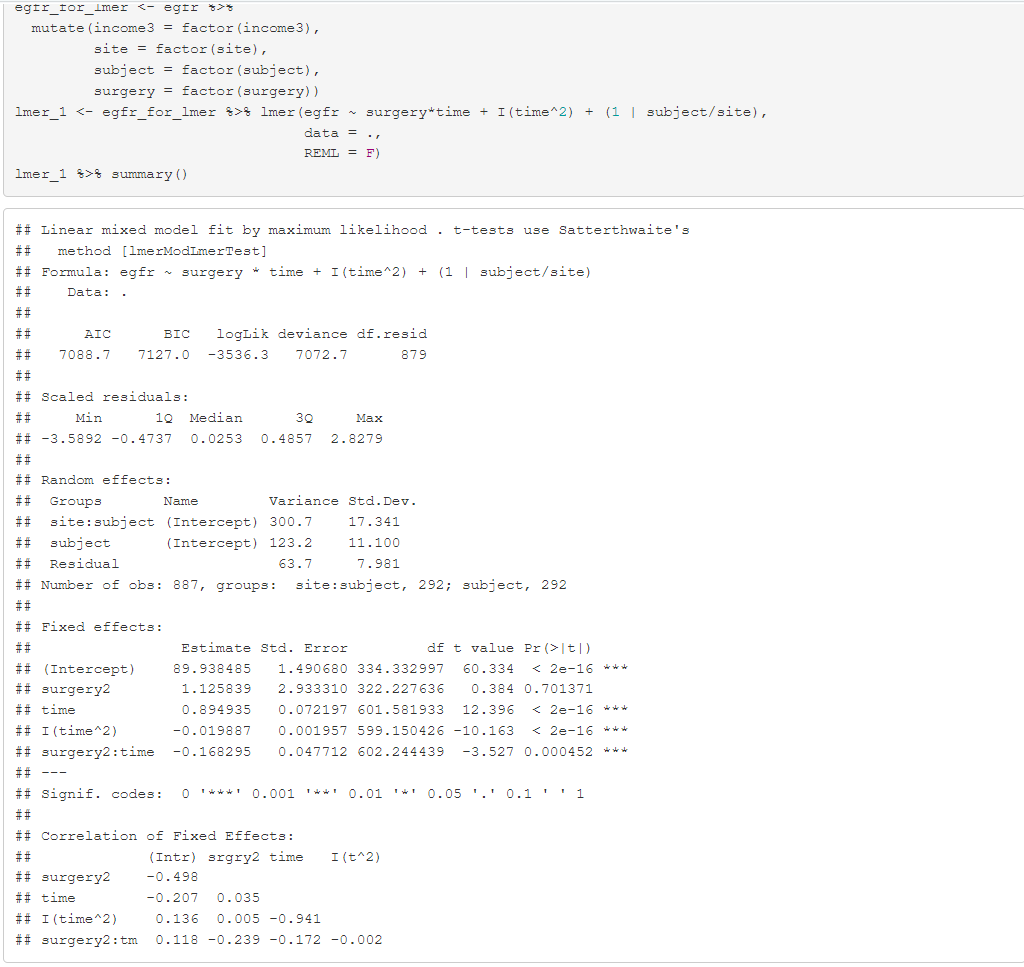
The aim of the study is to identify if one surgery is better than the other for EGFR levels. Therefore, surgery is included as a covariate. It is also to identify if one surgery has a consistently better trajectory over time than another, so time and its interaction with surgery are included as covariates. There is also a clear polynomial (~quadratic) pattern over time in EGFR levels, so time2 is included as a covariate. In the dataset, we also have adjusting covariates: age (time-independent) and income level (time-dependent). Therefore, these are also included as covariates.

Between-site heterogeneity in EGFR can be modeled by a site-level random intercept and between-subject heterogeneity in EGFR can be modeled by a subject-level random intercept. With this alone, then, the model will have an intercept, 6 random intercepts for site, 300 random intercepts for subject, 2 estimates for random intercept variance, and 1 estimate for the residual variance. This is 310 parameters for at most 1200 observations across 300 subjects, not taking into account missingness. Therefore, it seems unwise to include additional random coefficients.

Concordantly, models that include random coefficients fail to converge. A model including income also fails to converge, which I feel OK leaving out, because a grouped boxplot of egfr by income does not appear to demonstrate any sort of linear relationship with egfr. A model including age fails to converge when using REML to estimate the random effects, despite age appearing to have a negative relationship with EGFR. This is likely due to the extremely high >-.95 correlation between the effect of age and the intercept, as such, not much is lost by leaving age out. This all makes my final model:



* 1. **Fit the model you specified in part (a) and summarize your findings. (15 pts)**



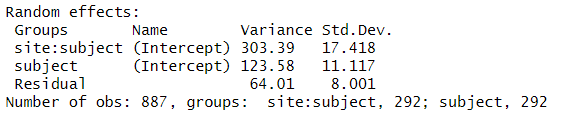
The average overall EGFR, when all covariates are 0 (i.e., when surgery == 1, “RYGB” surgery), is significantly greater than zero. The simple effect of “LAGB” surgery on EGFR is non-significant (i.e., there is no significant difference between LAGB and RYGB). The effects of time are significant. For instance, for RYGB surgeries, at month 12, there is a 0.89 – 0.02\*2\*12 = +0.41 unit difference in average EGFR. At month 36, there is a -0.55 unit difference in average EGFR. For LAGB surgeries, at month 12, there is a 0.89 – 0.02\*2\*12 – 0.168 = +0.242 unit difference in average EGFR. At month 36, there is a -0.718 unit difference in average EGFR.

For the purposes of answering the study aim. It appears people after RYGB surgery appear to do better at timepoint 12 than those after LAGB surgeries. However, at timepoints 24 and 36, people after RYGB surgery appear to have worse EGFR than people after LAGB surgeries. This highlights the importance of estimating random effects, because my original visualization in part a) has this pattern the other way around.

* 1. **Based on the fitted model in part (b), how much of the total variation in eGFR is attributable to site variability, and how much is attributable to subject variability (i.e., not due to repeated measurements)? Of the total between-subject variation in eGFR, how much is attributable to between sties? (18 pts)**

There is level-3 between site variability. There is level-2 between subject variability. There is also level-1 between occasion variability. The sum of these variances is the total variation in EGFR. The proportion of this attributable to site variability is the site variability / total variation. The proportion of this attributable to subject variability is the (site variability + subject variability) / total variation. The proportion of between-subject variation attributable to sites is site variability / (site variability + subject variability).

The random effects when fitting my model with REML are



Total variation = 303.39 + 123.58 + 64.01 = 490.48

Variation attributable to Site = 303.39 / 490.48 = 0.6185573

Variation attributable to Subject = 303.39 + 123.58 / 490.48 = 0.8705146

Between-subject variation attributable to Site = 303.39 / (303.39 + 123.58) = 0.7105651

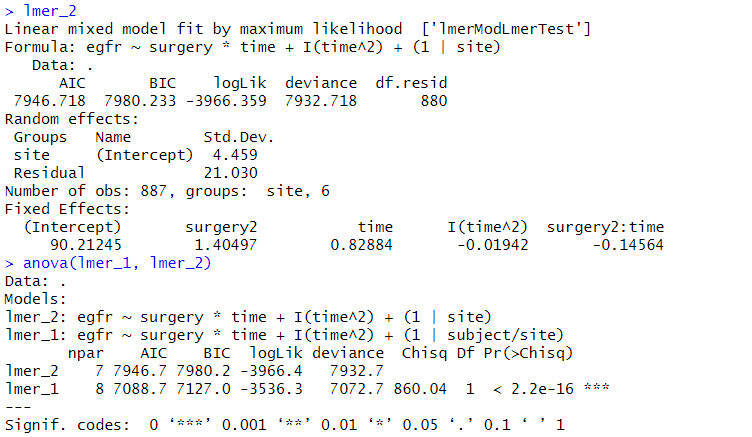
Over half of the total variation is attributable to site, 87% of the total variation is attributable to subject and site, and over half (almost ¾) of the variance between-subjects can be attributed to Site.

e) **Based on the fitted model in part (b), is the between-subject within-site variance significantly different from zero? Write down the hypotheses and perform the test. (15 pts)**

Null hypothesis: the between-subject within-site variance is zero

Alternative hypothesis: the between-subject within-site variance is not zero.

This can be tested with a likelihood ratio test comparing a model that includes a random-effect for site only with my original model



From this LRT, I have statistical reason to reject the null hypothesis that between-subject, within-site variance is 0, because the model that captures between-subject, within-site variance performs better than the model that only captures between-site variance.

**F) Based on your findings in part (c) and part (d), describe how you will modify the model that you specified in part (a), provide justification. (12 pts)**

Part c) and d) indicated that there is about 29% of the variance between subjects that is not accounted for by site variance and indicated that about 26% of the total variation can be accounted for by considering between-subject variance in addition to between-site variance. Part e) indicated that there is statistical reason to *not* believe that this unique, between-subject variance is 0. Therefore, I desire to keep my nested random effects and make no changes to my original model.