Annie Thwing

Advanced Data Analysis

**HAART Treatment and Hard Drug Use**

**Introduction:**

For this project, we are looking at the association of hard drug use at baseline with the effectiveness of highly active antiretroviral treatment (HAART) on a cohort of HIV infected men two years into treatment. This study is made up of HIV positive homosexual and bisexual men from 4 major US cities that were measured before treatment and for 8 consecutive years once treatment began, although we will be focusing on just the baseline measurement and measurement after two years of treatment. We are specifically interested in the difference between hard drug users and non users when it comes to viral load (VLOAD), the number of HIV copies in a mL of blood, CD4+ cell count (LEU3N), a measure associated with immunologic health, aggregate physical quality of life score (AGG\_PHYS), and aggregate mental quality of life score (AGG\_MENT). It is hypothesized that there may be an association of hard drug use with an increased replication of HIV and an inhibited immune system, which would result in higher VLOAD counts and lower CD4+ counts for hard drug users than non hard drug users after two years of treatment, which in turn would lead to lower mental and physical quality of life scores.

**Methods:**

First, we examined and cleaned the dataset in order to gain an understanding of its general structure and any patterns or outliers we needed to be aware of. The dataset provided had a lot of variables, many of which had missing values, or needed to be recoded or transformed, so analyzing and properly adjusting it was a crucial component to this project.

To begin with, every variable was recoded to comply with the groupings suggested by our investigator. Many of the variables had missing variable indicators other than NA, such as -1 or 999 which were subsequently recoded to be NA so that our dataset was consistent throughout. Additionally, many of the categorical variables were recoded so that we only analyzed the specific groupings thought to provide the most insight. Education level for example, was recoded to either represent up to high school level of education, or any education past high school rather than 8 different categories that we were not specifically concerned with. This recoding was also done for race, alcohol use, smoking habits, income level and ART adherence.

Other changes to our dataset that should be noted were done to BMI and VLOAD. Plotting the BMI’s of our subjects revealed a very unlikely and biologically improbable value of 514, which was subsequently removed and coded as missing. Additionally, VLOAD was suggested by our investigator to undergo a log base 10 transformation, as that is the scale this count is normally interpreted on, so we did a log transformation of our VLOAD variable.

In building the model for this dataset, we began by talking to our investigator about which covariates she was interested in including in the model. Since she wanted to see the baseline value of the outcome, age, BMI, race, marijuana use, alcohol use, smoking, income level, education and ART adherence modeled, each of these variables was included as a covariate in our model. Except for adherence level, which was not known until year two, all of the covariates included in the model were the measurements taken at baseline, or year 0 of the study. Additionally, we chose to run a hybrid model, using the difference between the value at year 2 and our baseline value as the outcome, and including our baseline as a covariate in our model. We chose this approach as we believe it best represents what we want to model, as we are not interested in the specific values at year 2 but rather the change in them from year 0.

Finally, since we were asked to run a Bayesian analysis, we needed to take some additional components into account that a frequentist would not have to. We needed to set initial parameter values for each variable in our model, set prior distributions for each parameter in our model, and determine the number of Monte Carlo iterations and burn in iterations for each model. We chose to use conservative priors ~ N(0,1000) as our parameter estimates and standard deviations were small enough for these to make sense, with the exception of our age and BMI variables. Since age and BMI had larger mean values, we chose them to have priors of ~ N(0,10000) after talking to the investigator. Lastly, we set all of our initial parameter values to be 0 through similar logic and after consulting the investigator.

The number of Monte Carlo iterations and burn in iterations were chosen through an informed trial and error. As our diagnostic plots showed slow mixing and poor autocorrelation plots for the standard SAS suggestion of 1,000 burn in values and 10,000 Monte Carlo iterations, we ended up increasing each by a factor of 10 ending up with 10,000 burn in values to be thrown out, and 100,000 Monte Carlo iterations. Additionally we fixed high autocorrelation plots by thinning at a rate of 15, only recording every 15th draw. We found that these parameters produced diagnostic plots we were comfortable with, Geweke diagnostics that did not indicate converge issues, and did not take an excessive amount of time to run.

We then ran crude models including baseline values of the outcome and hard drugs for each outcome and compared them to full models that included all of our desired covariates. We then chose the model with the lowest DIC value as the model from which to interpret our results.

**Results:**

Our Table 1 below shows a summary of basic descriptive statistics for our dataset, stratified by hard drug use at baseline. We can use it to wrap our heads around the distribution of our data. We can see that our continuous variables, age and BMI, are very evenly distributed between hard drug users at baseline and non hard drug users at baseline. Additionally we can see some very high associations between hard drug use with smoking and education level, and that all of our missing income values were subjects that did not use hard drugs. Since we were asked to include all indicated covariates in our model we do not need to use this table to choose significant covariates, but rather to get an idea of the makeup of our dataset.

After running crude and adjusted models for each of our outcomes, we analyzed DIC to choose the model that best fit our data from which we could interpret and come up with our results. As can be verified in the appendix, all four crude models had drastically higher DIC scores than their corresponding model counterparts adjusting for covariates. These along with good diagnostic plots that didn’t show mixing problems, had smooth posteriors and good autocorrelation plots led us to use the full model for each outcome to interpret our results from.

As can be seen in the figures below, there does not seem to be a significant association between change in VLOAD after 2 years of treatment and hard drug use at baseline with a mean estimated difference between hard drug users and non hard drug users of -0.04, a standard deviation of 0.20 and an HPD interval of (-0.43, 0.35). Additionally, we can see that hard drug use at baseline does not seem to significantly affect the AGG\_MENT score of hard drug users and non hard drug users either, as the mean estimated difference between these two groups after two years of treatment is -0.41 with a standard deviation of 1.78 and an HPD interval of (-3.88, 3.11). Hard drug use at baseline looks like it may trend toward being significantly associated with AGG\_PHYS as those who used hard drugs at baseline had changes an estimated mean of -3.24 less than those who did not use drugs at baseline, with a standard deviation of 1.39 in an HPD interval of (-5.97, -0.51). Finally, it looks like hard drug use at baseline is incredibly associated with CD4+ cell, the white blood cells which fight infection, counts (LEU3N). The average change of those who used hard drugs at baseline, was -83 units from those who did not use hard drugs at baseline, with a standard deviation of 22 and an HPD interval of (-124.4, -38.83).

**Conclusions:**

Conclusions that we can draw from this analysis are that while treatment response 2 years after initiating HAART does not differ significantly for all outcomes studied between hard drug users and non hard drug users at baseline, it does to at least some extent for two. This means that patients already battling an unforgiving virus are at an even worse disadvantage if they were using hard drugs at the start of their treatment. I would argue that the fact that we are only observing the difference at the 2 year time point indicates that the difference may grow even more apparent further into the study. Regardless, I think that more drug rehabilitation programs and information of the interference of drug use with HAART treatments is important information that should be provided to all patients and the general public.

**Figures:**



Table

**Model 1: Outcome LEU3N Full**

**The MCMC Procedure**

| **Posterior Summaries and Intervals** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **N** | **Mean** | **Standard Deviation** | **95% HPD Interval** | |
| **betaInt** | 6667 | 14.7161 | 28.4412 | -42.4795 | 69.1661 |
| **betaBaseline** | 6667 | -0.0478 | 0.0408 | -0.1261 | 0.0325 |
| **betaHASHV** | 6667 | 34.9299 | 14.5932 | 8.1596 | 64.8555 |
| **betaincome** | 6667 | -20.5078 | 12.1020 | -43.9799 | 3.2793 |
| **betaBMI** | 6667 | 5.2428 | 1.6378 | 2.0708 | 8.4726 |
| **betaSMOKE** | 6667 | -9.9613 | 15.5313 | -41.4855 | 18.6217 |
| **betaDKGRP** | 6667 | -5.8570 | 22.4613 | -49.5769 | 37.7652 |
| **betaADH** | 6667 | 37.9615 | 20.1764 | -0.1184 | 77.2804 |
| **betaRACE** | 6667 | -12.6419 | 15.0659 | -43.0112 | 16.4934 |
| **betaEDUCBAS** | 6667 | 22.7270 | 18.2465 | -13.3136 | 57.9402 |
| **betaage** | 6667 | -0.2505 | 0.8963 | -1.9274 | 1.5209 |
| **betahard\_drugs** | 6667 | -83.3148 | 21.8973 | -124.4 | -38.8293 |
| **sigma2** | 6667 | 30563.6 | 2081.1 | 26561.0 | 34572.8 |

**Model 1: Outcome VLOAD Full**

**The MCMC Procedure**

| **Posterior Summaries and Intervals** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **N** | **Mean** | **Standard Deviation** | **95% HPD Interval** | |
| **betaInt** | 6667 | 1.0242 | 0.6268 | -0.1300 | 2.3422 |
| **betaBaseline** | 6667 | -0.5333 | 0.0608 | -0.6514 | -0.4178 |
| **betaHASHV** | 6667 | -0.1694 | 0.1090 | -0.3868 | 0.0429 |
| **betaincome** | 6667 | -0.2355 | 0.0919 | -0.4134 | -0.0570 |
| **betaBMI** | 6667 | -0.0223 | 0.0121 | -0.0458 | 0.00242 |
| **betaSMOKE** | 6667 | -0.0977 | 0.1183 | -0.3264 | 0.1326 |
| **betaDKGRP** | 6667 | 0.1311 | 0.1968 | -0.2578 | 0.5107 |
| **betaADH** | 6667 | -0.4430 | 0.1758 | -0.7984 | -0.1068 |
| **betaRACE** | 6667 | 0.1253 | 0.1258 | -0.1260 | 0.3667 |
| **betaEDUCBAS** | 6667 | 0.00410 | 0.1477 | -0.2782 | 0.2948 |
| **betaage** | 6667 | -0.00024 | 0.00620 | -0.0121 | 0.0122 |
| **betahard\_drugs** | 6667 | -0.0358 | 0.2005 | -0.4245 | 0.3489 |
| **sigma2** | 6667 | 1.2426 | 0.0828 | 1.0833 | 1.4040 |

**Model 1: Outcome AGG\_MENT Full**

**The MCMC Procedure**

| **Posterior Summaries and Intervals** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **N** | **Mean** | **Standard Deviation** | **95% HPD Interval** | |
| **betaInt** | 6667 | 15.0854 | 4.5774 | 6.1383 | 24.0028 |
| **betaBaseline** | 6667 | -0.5237 | 0.0336 | -0.5887 | -0.4554 |
| **betaHASHV** | 6667 | 1.1791 | 0.9629 | -0.6795 | 3.0433 |
| **betaincome** | 6667 | 1.5794 | 0.7731 | 0.0139 | 3.0598 |
| **betaBMI** | 6667 | 0.0427 | 0.1071 | -0.1507 | 0.2565 |
| **betaSMOKE** | 6667 | 1.8205 | 1.0575 | -0.1613 | 3.9475 |
| **betaDKGRP** | 6667 | -0.0717 | 1.7933 | -3.5839 | 3.3784 |
| **betaADH** | 6667 | 2.3123 | 1.5761 | -0.7699 | 5.3234 |
| **betaRACE** | 6667 | 0.3547 | 1.1151 | -1.8202 | 2.4807 |
| **betaEDUCBAS** | 6667 | 0.7812 | 1.2937 | -1.6815 | 3.3645 |
| **betaage** | 6667 | 0.0607 | 0.0534 | -0.0435 | 0.1626 |
| **betahard\_drugs** | 6667 | -0.4129 | 1.7785 | -3.8754 | 3.1079 |
| **sigma2** | 6667 | 96.7696 | 6.3198 | 84.8747 | 109.3 |

**Model 1: Outcome AGG\_PHYS Full**

**The MCMC Procedure**

| **Posterior Summaries and Intervals** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **N** | **Mean** | **Standard Deviation** | **95% HPD Interval** | |
| **betaInt** | 6667 | 11.8720 | 4.3223 | 3.6872 | 20.7491 |
| **betaBaseline** | 6667 | -0.3137 | 0.0455 | -0.3991 | -0.2209 |
| **betaHASHV** | 6667 | 0.3583 | 0.7479 | -1.0885 | 1.8139 |
| **betaincome** | 6667 | 1.2267 | 0.6118 | 0.0599 | 2.4411 |
| **betaBMI** | 6667 | 0.0573 | 0.0841 | -0.1064 | 0.2207 |
| **betaSMOKE** | 6667 | -0.7675 | 0.8273 | -2.3297 | 0.9131 |
| **betaDKGRP** | 6667 | -0.8176 | 1.4074 | -3.6785 | 1.8413 |
| **betaADH** | 6667 | 1.6945 | 1.2670 | -0.7716 | 4.1584 |
| **betaRACE** | 6667 | 1.2948 | 0.8535 | -0.4461 | 2.9076 |
| **betaEDUCBAS** | 6667 | 1.3326 | 1.0246 | -0.6951 | 3.3153 |
| **betaage** | 6667 | -0.1071 | 0.0433 | -0.1941 | -0.0226 |
| **betahard\_drugs** | 6667 | -3.2426 | 1.3923 | -5.9715 | -0.5149 |
| **sigma2** | 6667 | 60.2560 | 3.9991 | 52.8538 | 68.4168 |

**Reproducible Research Information:**

<https://github.com/BIOS6623-UCD/bios6623-athwing/tree/master/Project1>