analysis_6_writeup

May 2, 2019

1 Data Analysis 6

1.1 Dataset Description

We are analyzing a subset of data from the Framingham Heart Study. The dataset contains health data for patients, which has been anonymized for patient confidentiality. The dataset includes records on 1500 participants over a period of time for a set of variables. Data was collected for participants over 4607 days. The features for this dataset are **SEX**, **CURSMOKE**, **DIABETES**, **PREVHYP**, **EDUC**, **AGE**, **TOTCHOL**, **SYSBP**, **DIABP**, **CIGPDAY**, **BMI**, **HEARTRTE**, **GLUCOSE**, **PERIOD**, and **TIME**. **SEX** is a binary categorical variable for the participants biological sex, with 1 for men and 2 for women. **CURSMOKE** is binary categorical variable for whether the participant smokes or not, with 0 for no and 1 for yes. **DIABETES** is binary categorical variable for whether the participant has diabetes or not, with 0 for no and 1 for yes. **PREVHYP** is the response variable and is a binary categorical variable representing whether the participant has hypertension or not. **EDUC** is an ordinal variable for education level with 4 levels; 1 for 0-11 years of school, 2 for high school diploma or GED, 3 for some college or vocational school, and 4 is college degree and more. **AGE** is a continuous variable for age. **TOTCHOL** is a continuous variable for serum total cholesterol. **SYSBP** is a continuous variable for systolic blood pressure. **DIABP** is a continuous variable for diastolic blood pressure. **CIGPDAY** is a continuous variable representing the number of cigarettes the participant smokes per day. **BMI** is a continuous variable representing the patients body mass index. **HEARTRTE** is a continuous variable representing heart rate in beats per minute. Finally, **GLUCOSE** is a continuous variable that represents causal serum glucose. The additional variables in this analysis are **PERIOD** and **TIME**. **PERIOD** is a categorical variable that buckets time into 3 categories. **TIME** is a numeric variable representing the number of days since the baseline exam.

1.2 Data Cleaning

In this case, there was no missing data. Therefore, the only data cleaning to do is to convert categorical variables to factors.

1.3 Exploratory Data Analysis (tests)

Next, a few chi squared tests were run to to test if there is a relationship between categorical variables and the response. **PERIOD** did not seem worth checking, however, **SEX**, **CURSMOKE**, **DIABETES**, and **EDUC**. **SEX** and **EDUC** do not change overtime, therefore, it is irrelevant that this is time series data for the chi squared tests. The tests still apply the same. For **CURSMOKE**

and **DIABETES**, values do change over time for participants, however, I would still argue the test have power and are relevant. If a participant has their **CURSMOKE** or **DIABETES** value change over time, they can also have the response **PREVHYP** change as well. Therefore, these tests still show relationships between these variables and the response. P values ranged from .08 for **SEX** to 1.194e-09 for **CURSMOKE**. All categoricals besides **SEX** had p values below .05. We will consider removing **SEX** from the final model, however, a p value of .08 suggests that there still is a strong case that **SEX** is not independent of **PREVHYP**. The contingency tables for these four categoricals are displayed below.

	free of disease	prevalent disease
Male	366	336
Female	453	345

Table 1: SEX

	free of disease	prevalent disease
Yes	420	456
No	399	225

Table 2: CURSMOKE

	free of disease	prevalent disease
Not Diabetic	811	634
Diabetic	8	47

Table 3: DIABETES

	free of disease	prevalent disease
0-11 years	287	295
High School Diploma or GED	274	212
Some College or Vocational School	153	96
College degree or more	105	78

Table 4: EDUC

1.4 Exploratory Data Analysis (plots)

Next, exploratory plots were generated for the continuous variables. For each, there are overall density plots, seperated by the response value. This is used to see if there are differences in the distributions of each continuous variable for the two classes. These charts do not take into account

time, therefore individual trajectories for each continuous variable for each participant overtime are also plotted. These are also colored by response class to see if there is a difference between the classes. We can see from these charts that **SYSBP** and **CIGPDAY** values seem to be higher for individuals that have hypertension versus those that do not. Therefore, we will consider using these values over others in the models.

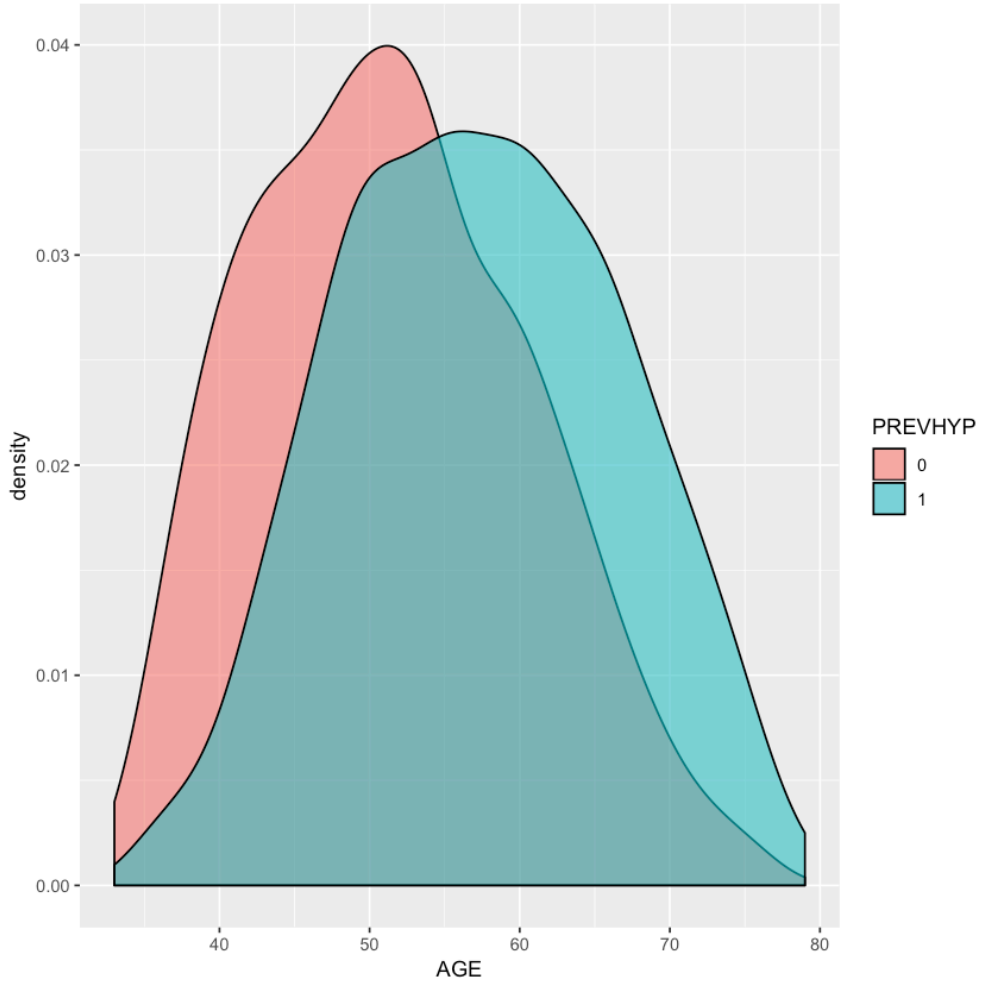

Figure 1: AGE Density

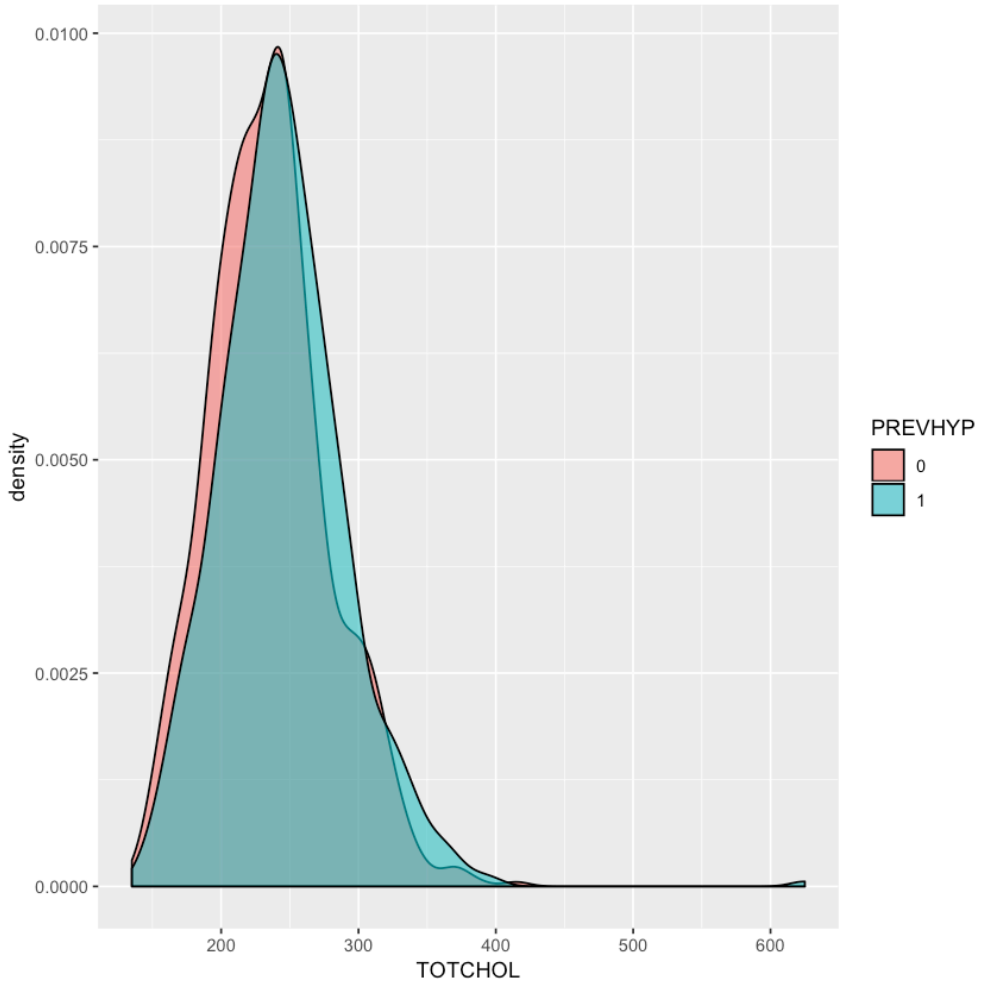

Figure 2: TOTCHOL Density

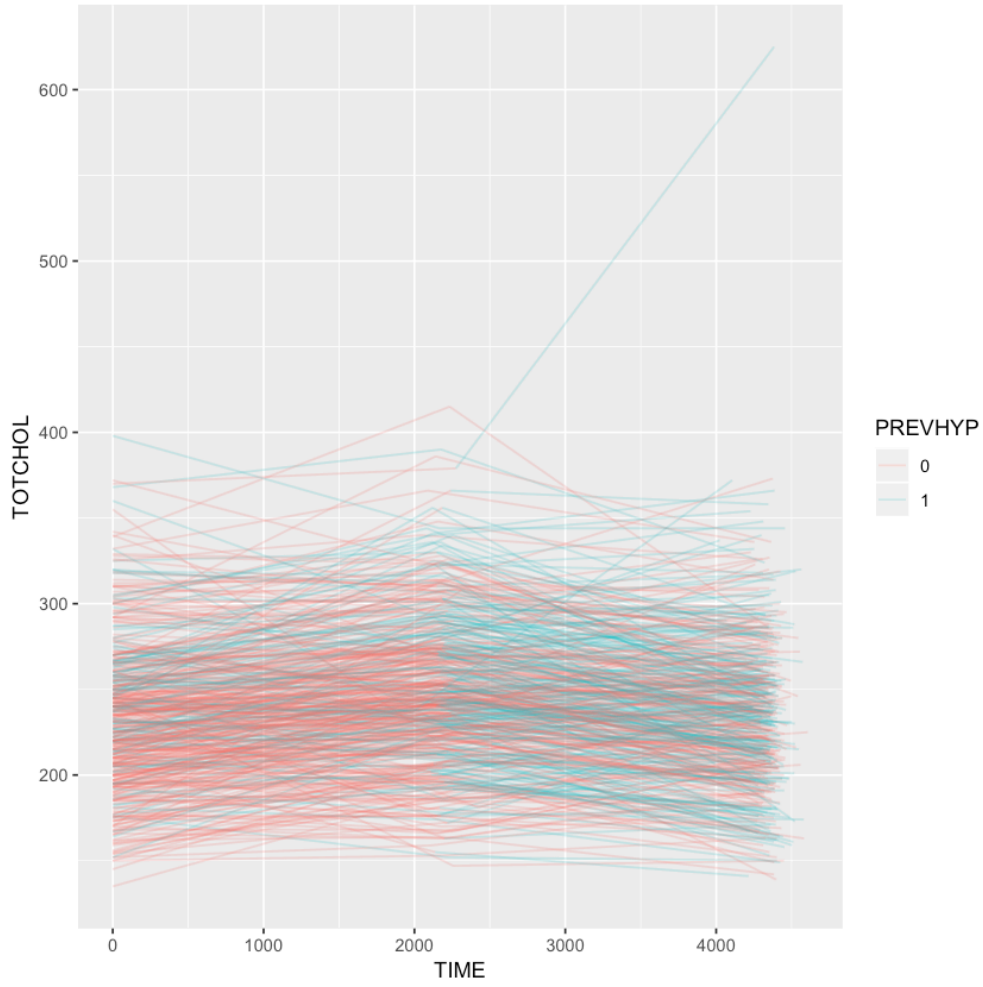

Figure 3: TOTCHOL Trajectories

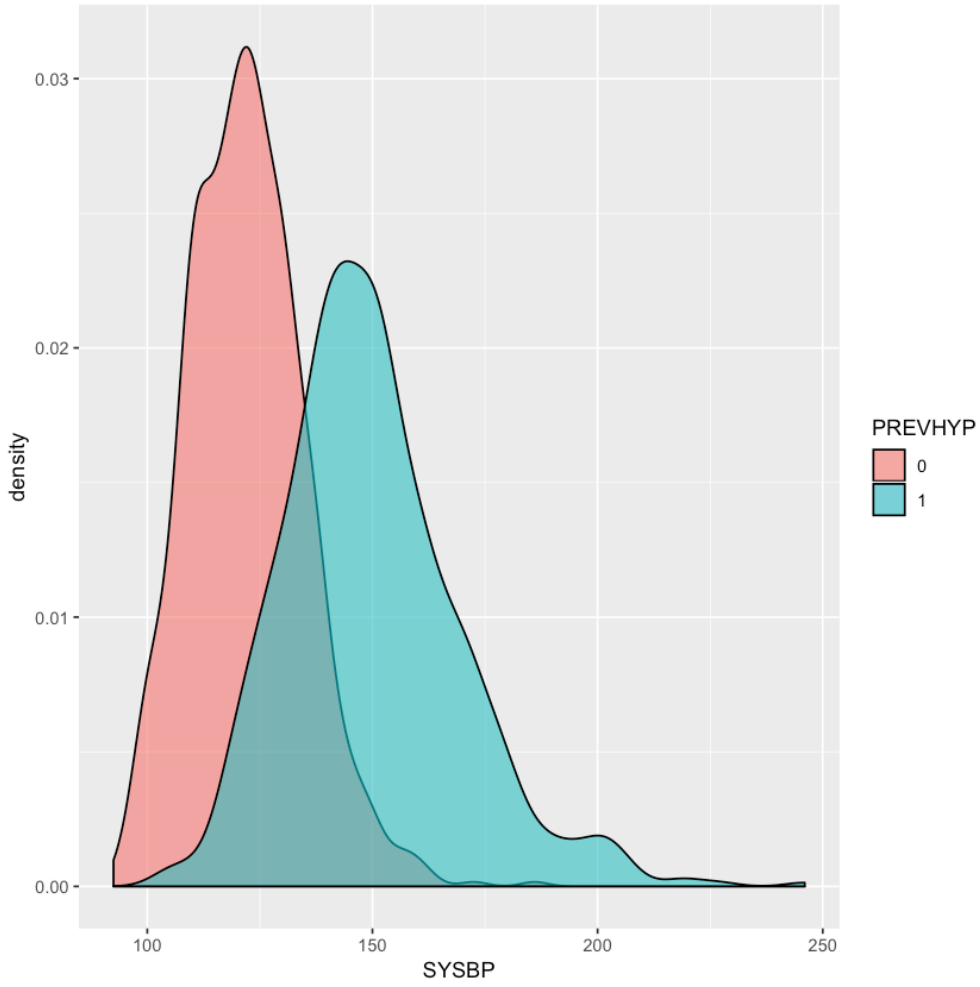

Figure 4: SYSBP Density

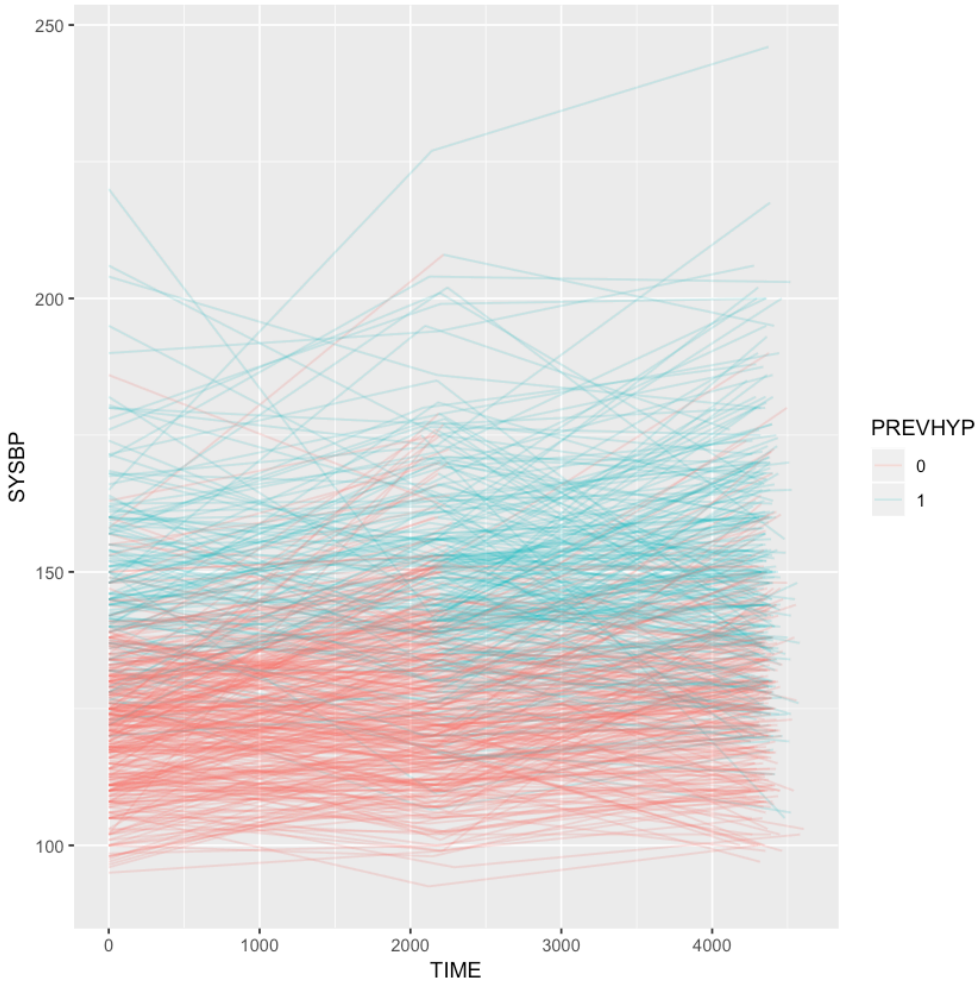

Figure 5: SYSBP Trajectories

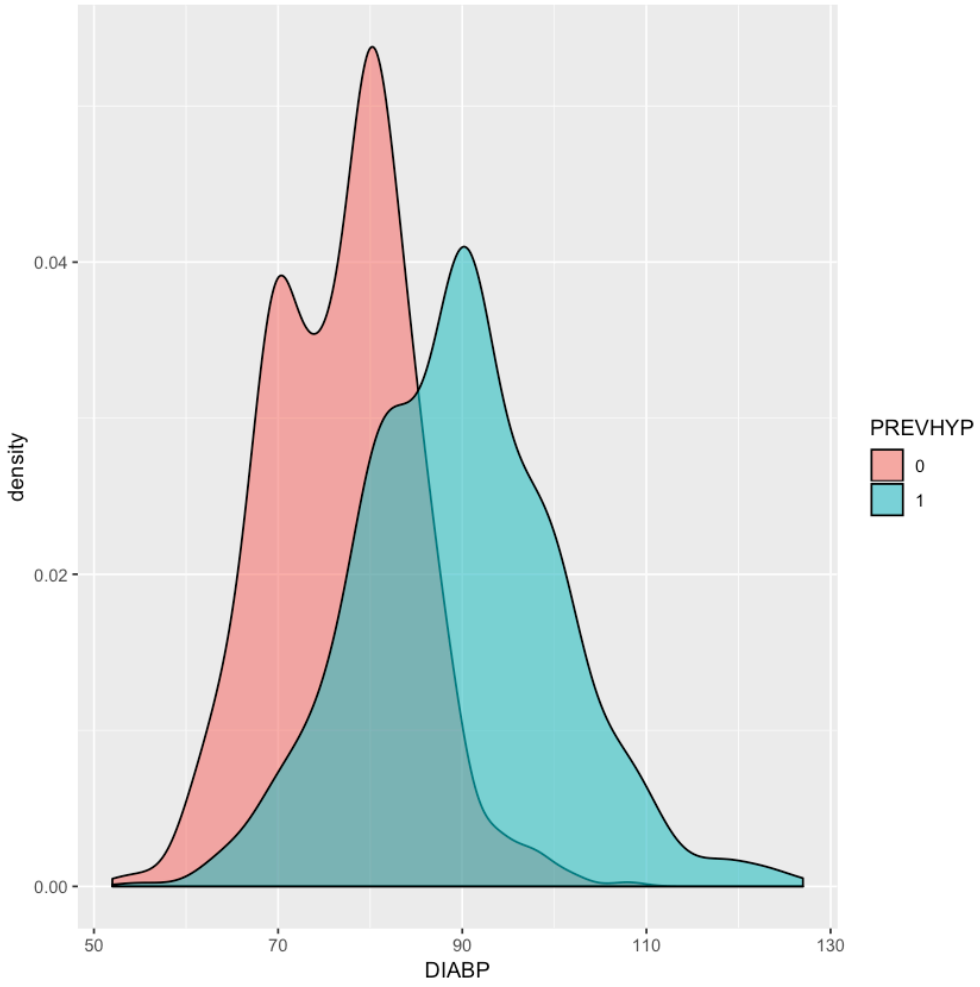

Figure 6: DIABP Density

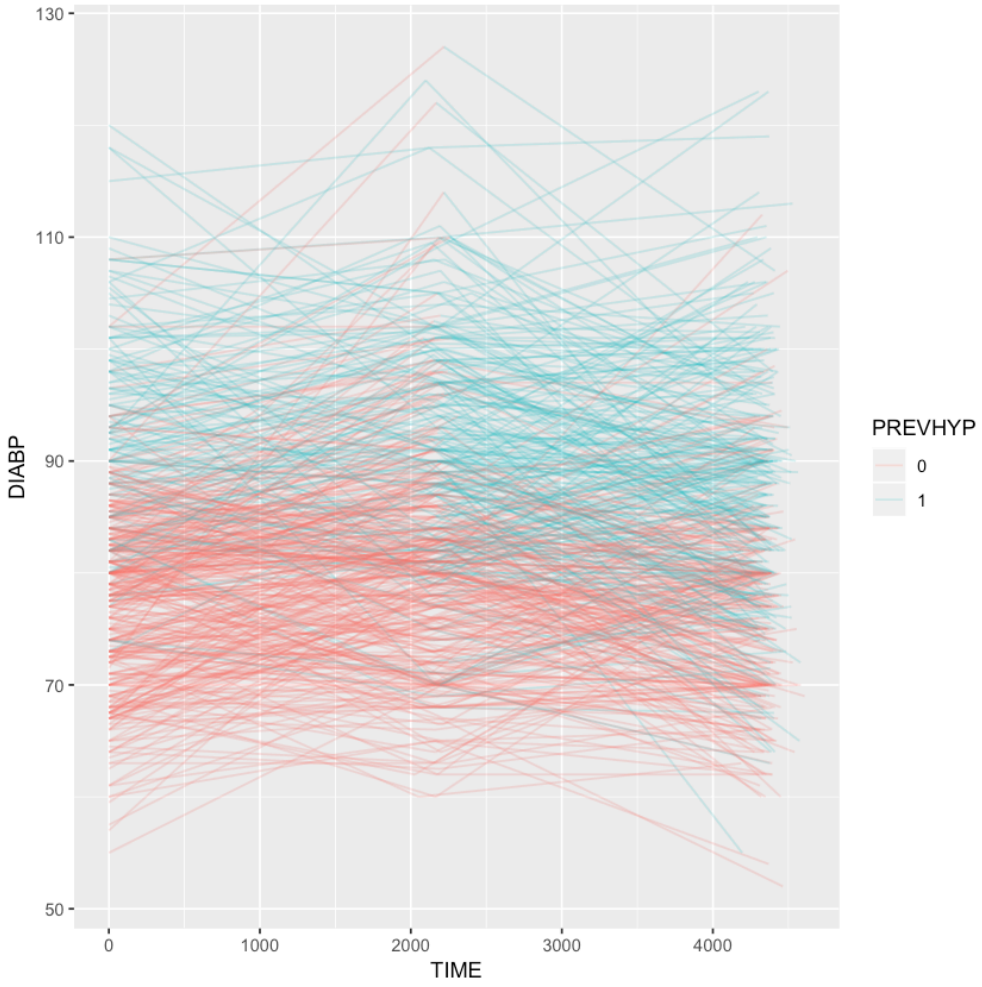

Figure 7: DIABP Trajectories

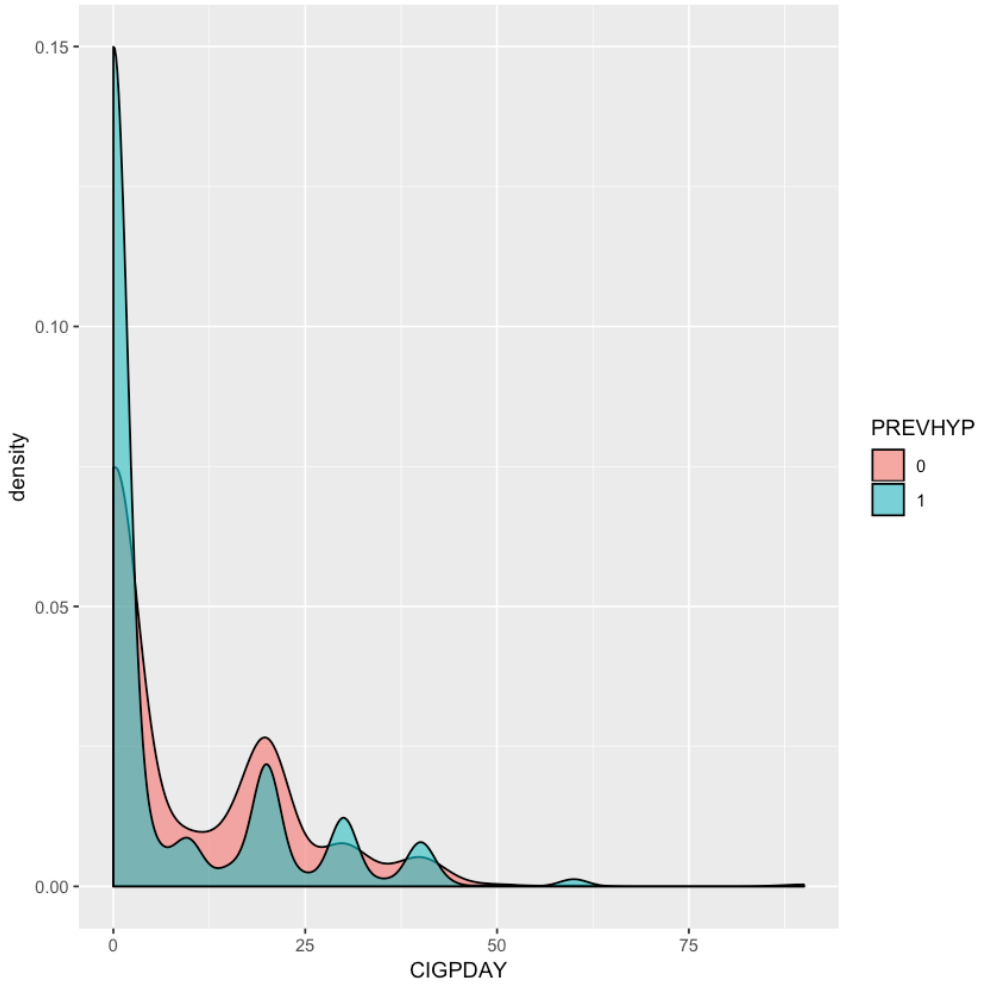

Figure 8: CIGPDAY Density

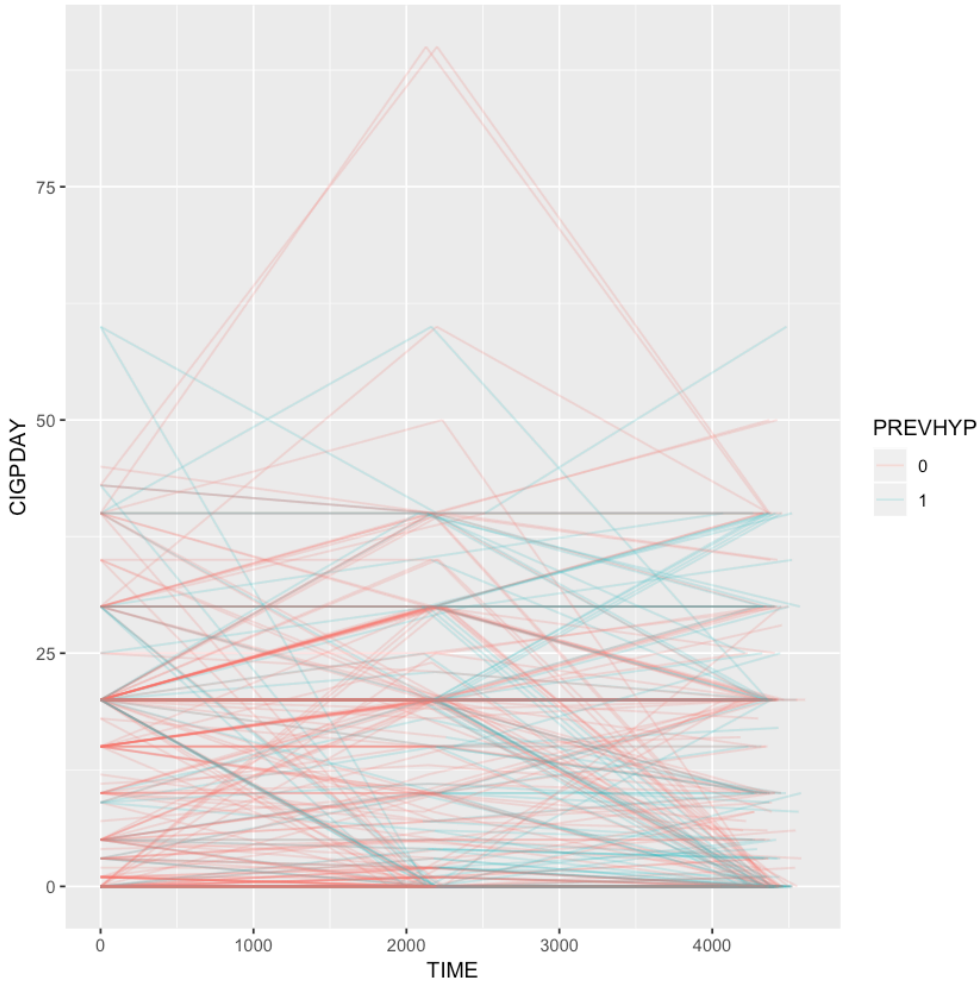

Figure 9: CIGPDAY Trajectories

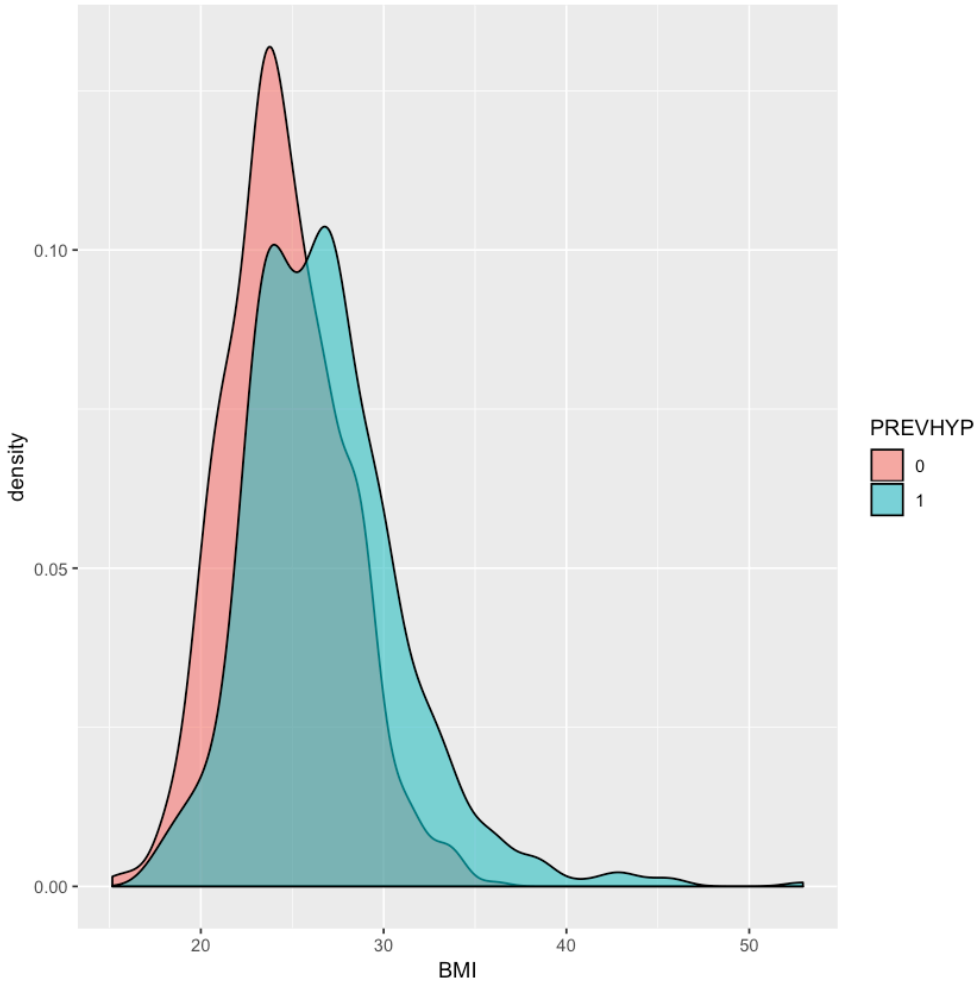

Figure 10: BMI Density

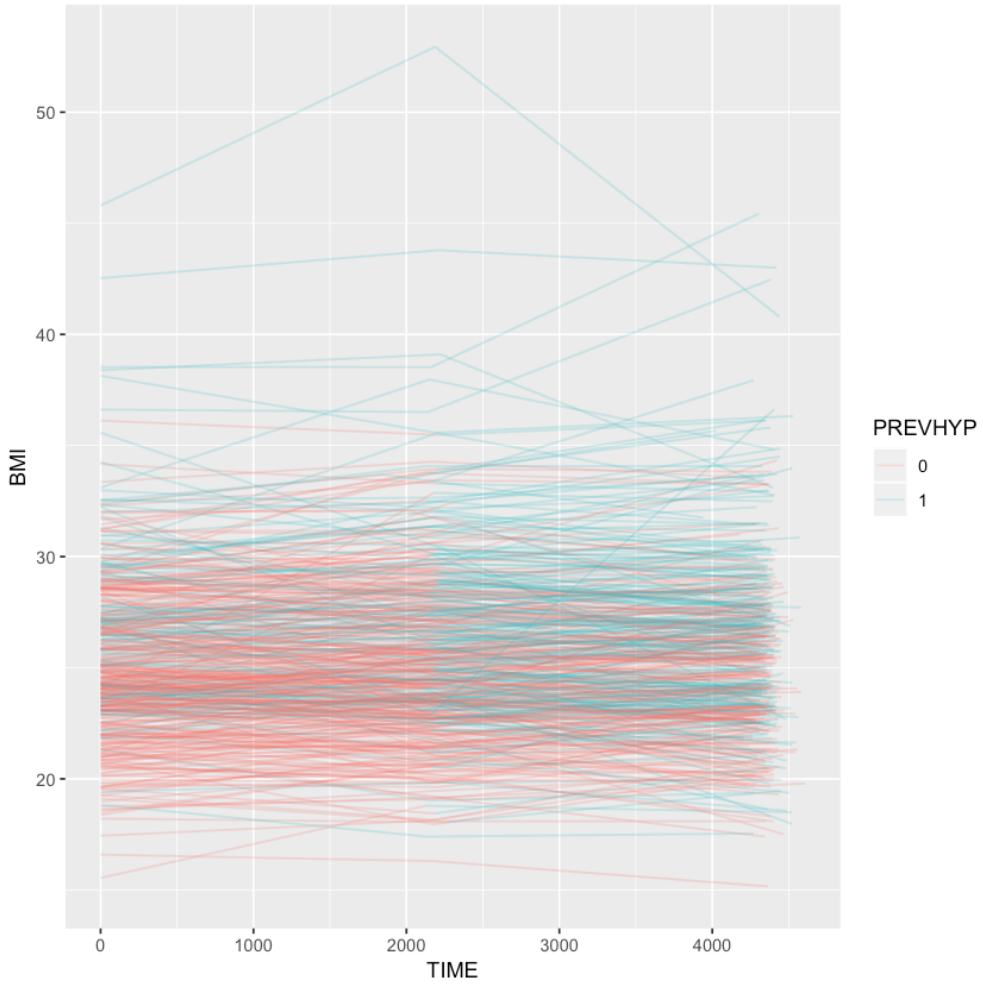

Figure 11: BMI Trajectories

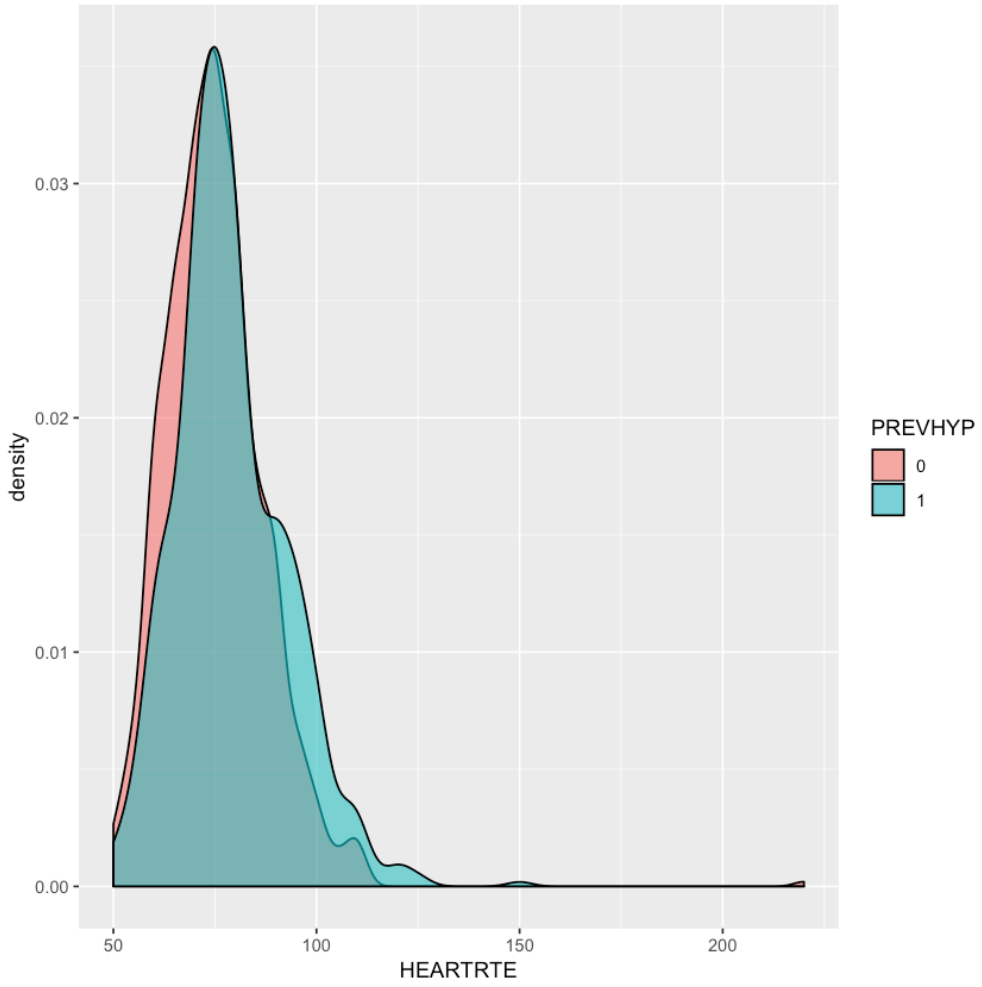

Figure 12: HEARTRTE Density

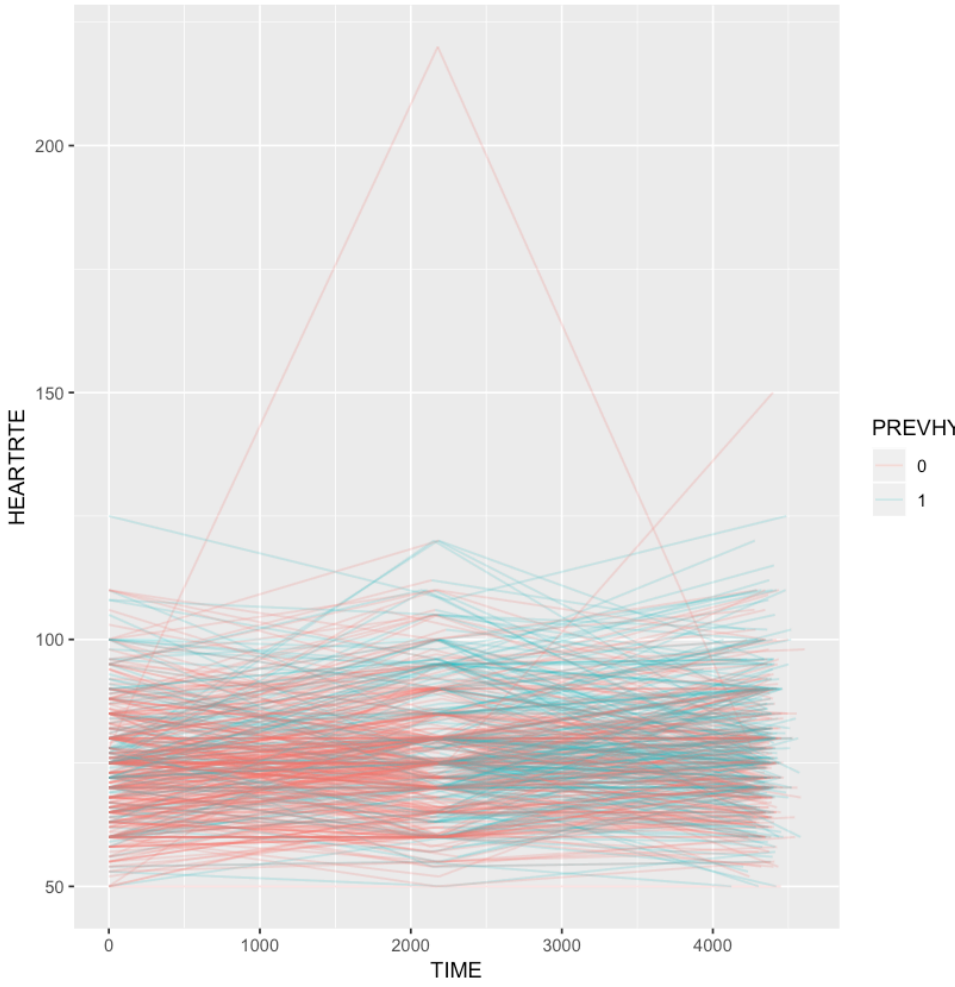

Figure 13: HEARTRTE Trajectories

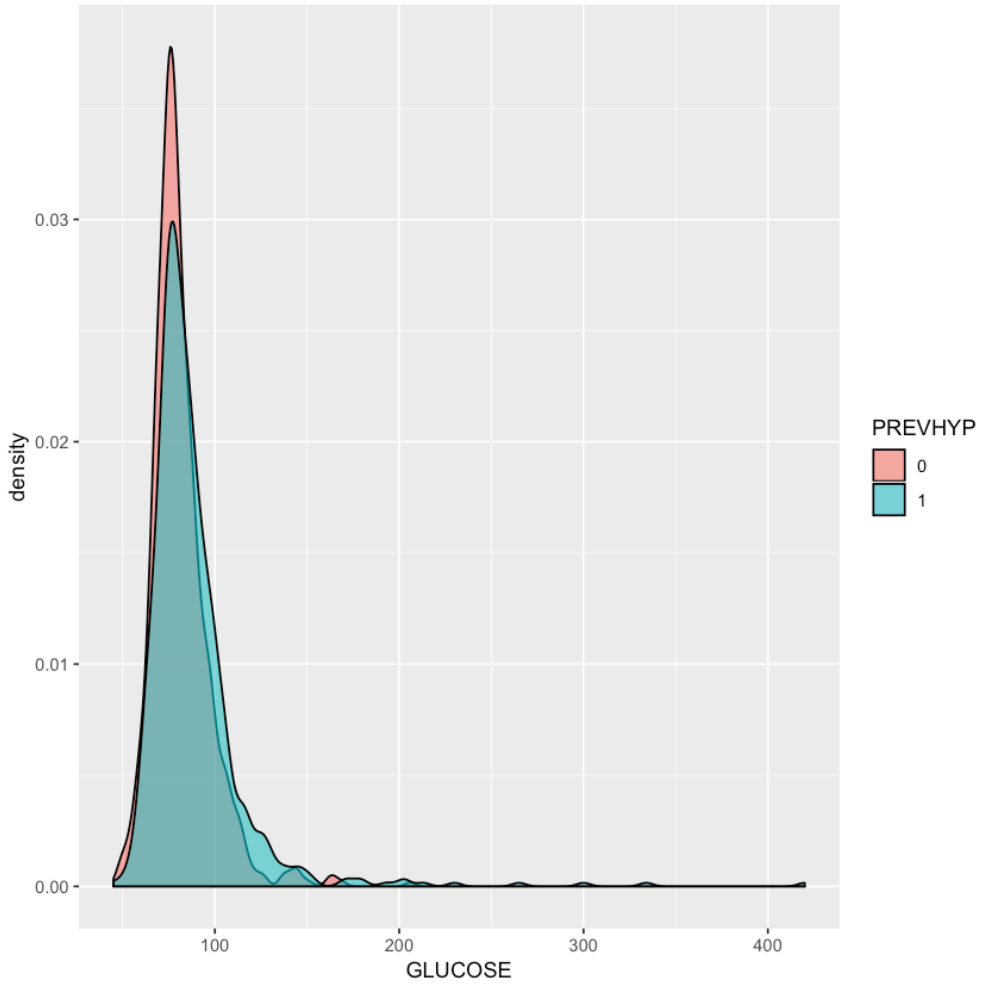

Figure 14: Glucose Density

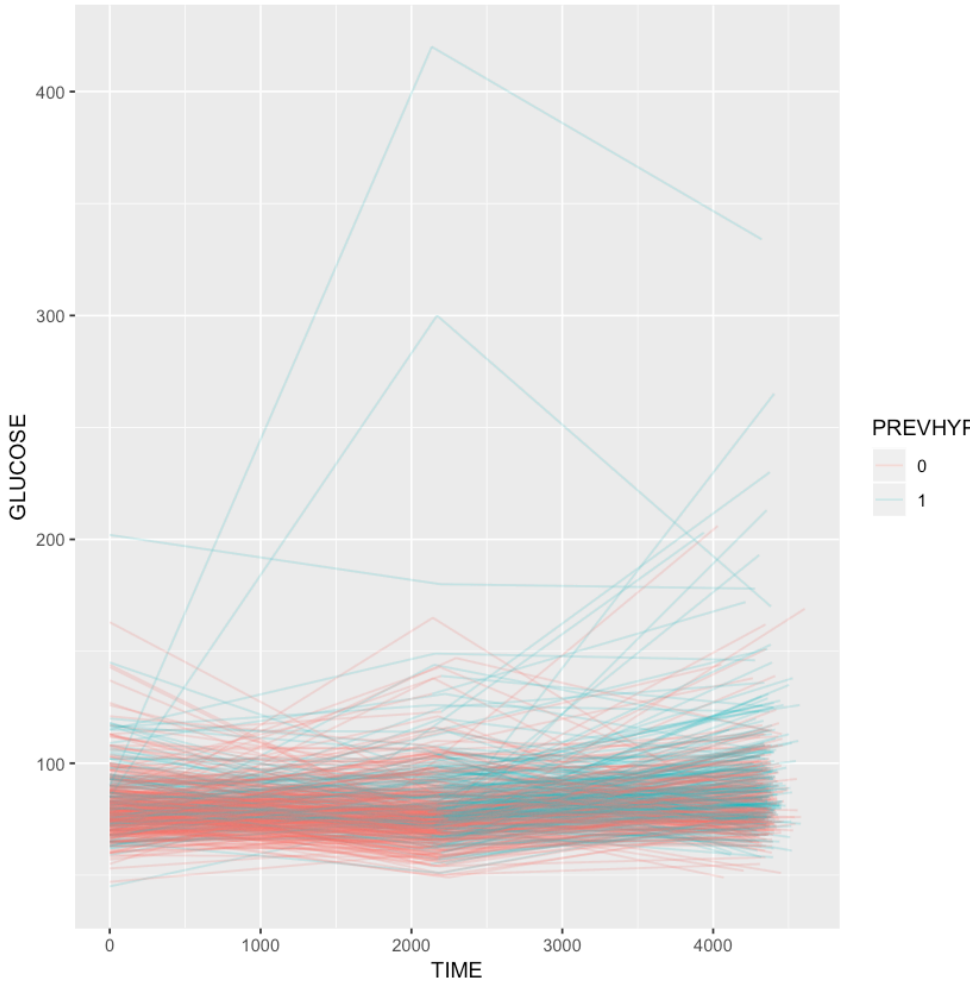

Figure 15: GLUCOSE Trajectories

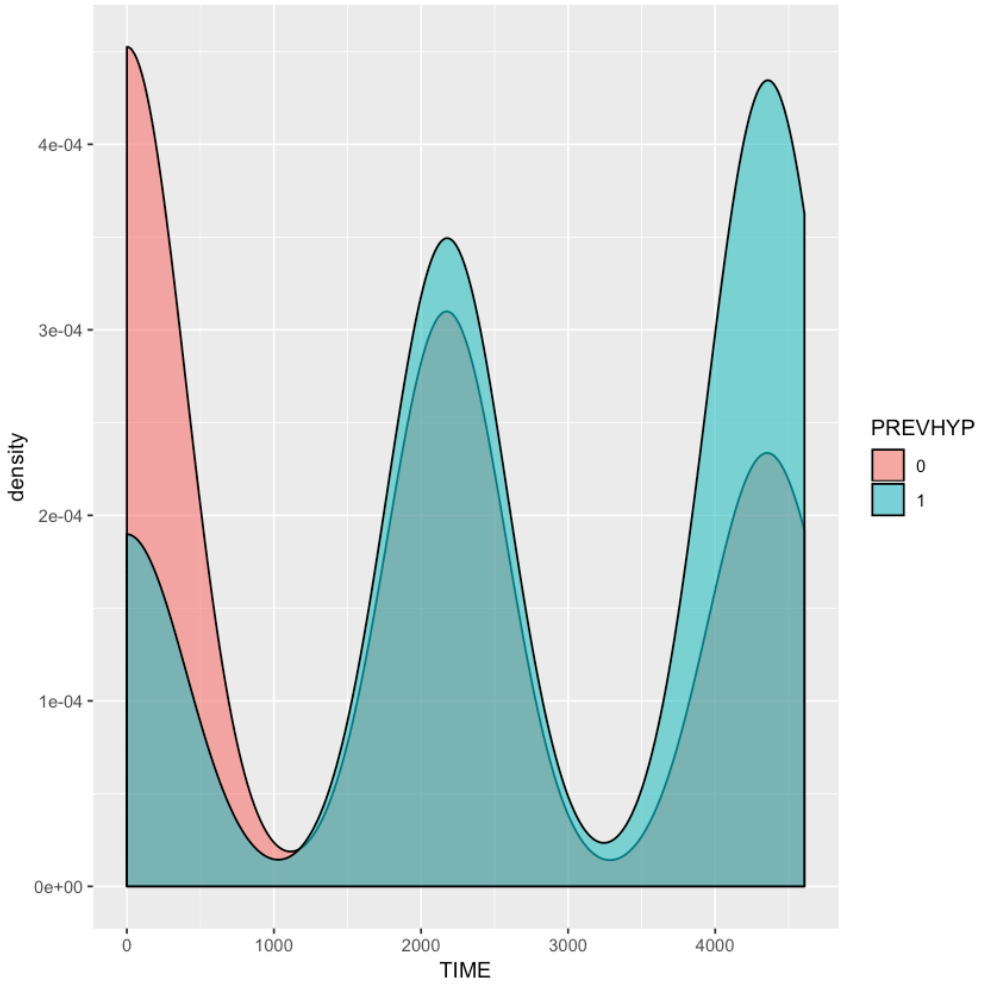

Figure 16: PREVHYP Density

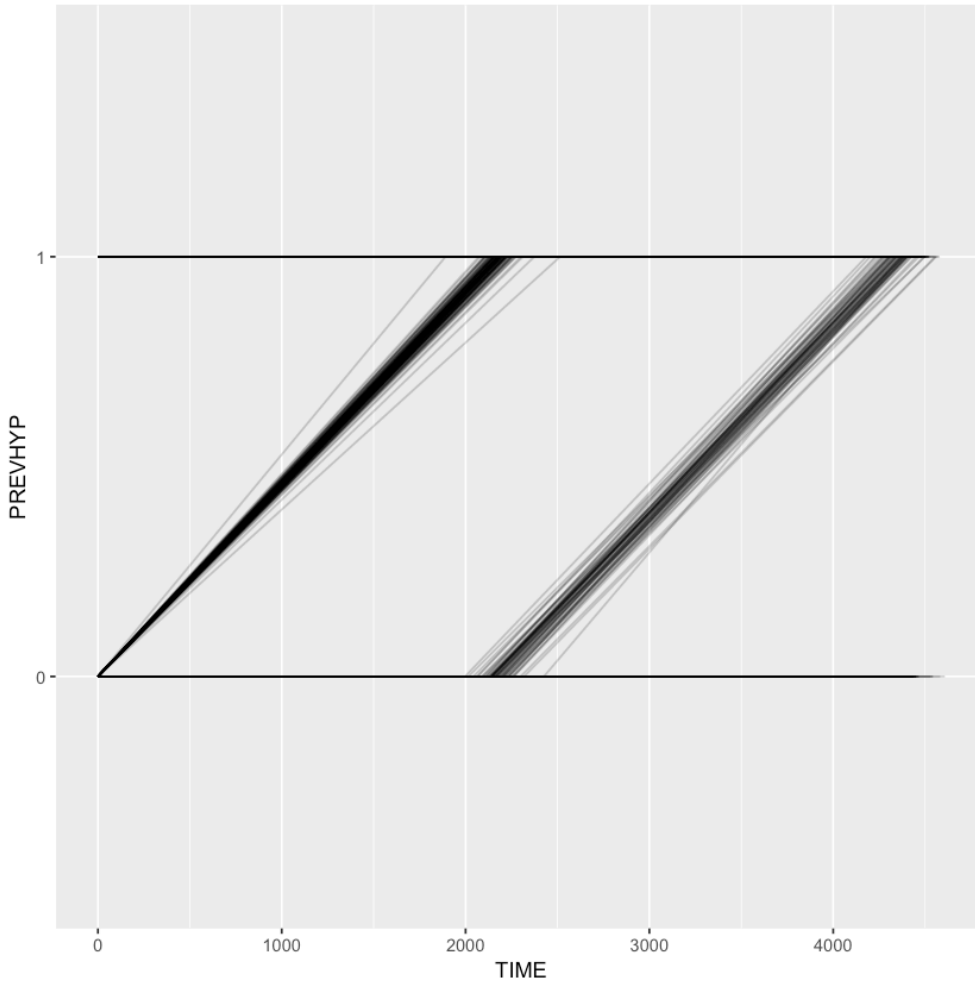

Figure 17: PREVHYP Trajectories

1.5 Model Fitting

Four different models were fit, all of which were Generalized Linear Mixed-Effects models for longitudinal data. As the response is a binary categorical variable, the binomial family is utilized. For every model, I also utilized a random intercept. I decided not to explore interaction terms, for reasons that are apparent later. The first model uses every variable, while the second uses every except for **PERIOD**. I decided to omit **PERIOD** in the second model, as it is just a categorical that bins **TIME** and likely will have a high correlation with **TIME** and not much predictive power. It could be useful in interaction terms, however, I decided not to use any interaction terms. The third model fit utilized the two continuous variables with differing trajectory values, **SYSBP** and **CIGPDAY**, as well as the three categoricals with chi squared tests that had p values below .05, **CURSMOKE**, **DIABETES**, and **EDUC**. The final model fit utilized only the two continuous variables **SYSBP** and **CIGPDAY**. The second model and first model had comperable AIC and BIC as the best models, with values of 952.4222 and 1053.373 for the first model and 950.2637 and 1040.588 for the second model. Given the second model has slightly better values for AIC and BIC and is slightly less complicated, this model is selected for as the final model. The qqplots for all of these models also show slighly fatter tails, however, this is ok for the fit. There is no apparent left or right skew that needs to be adjusted for.

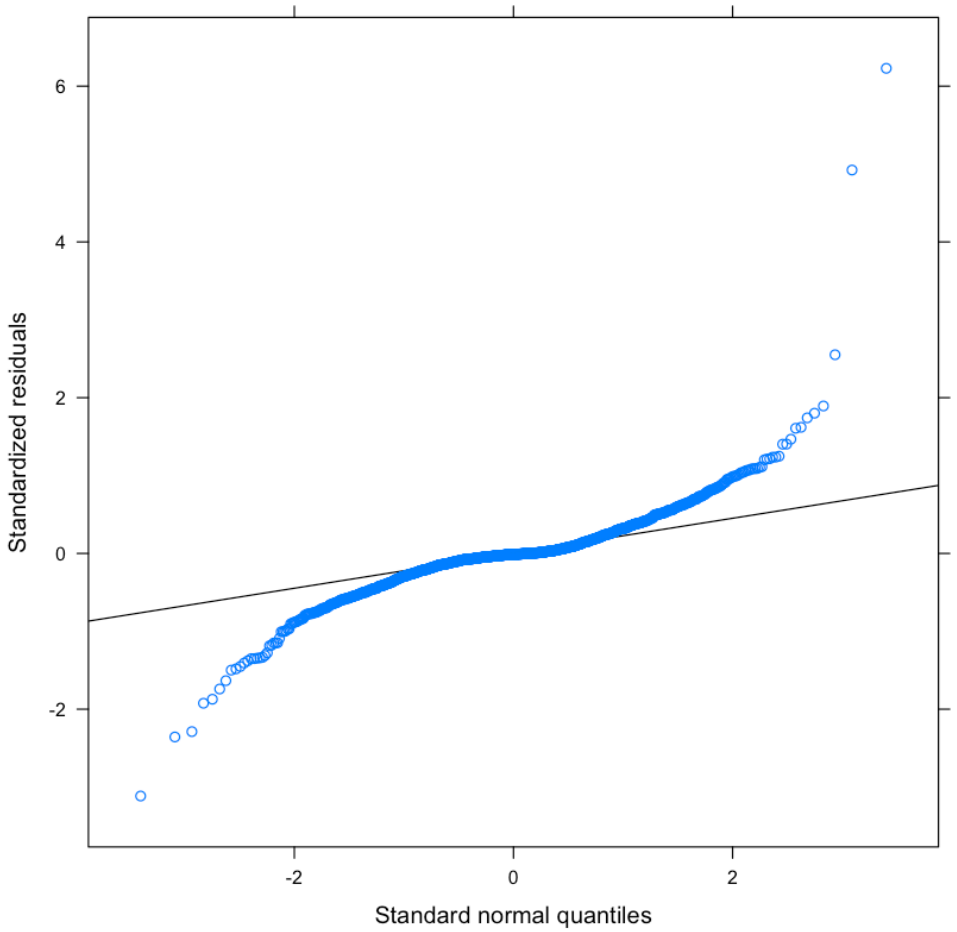

Figure 18: Model 1 QQplot

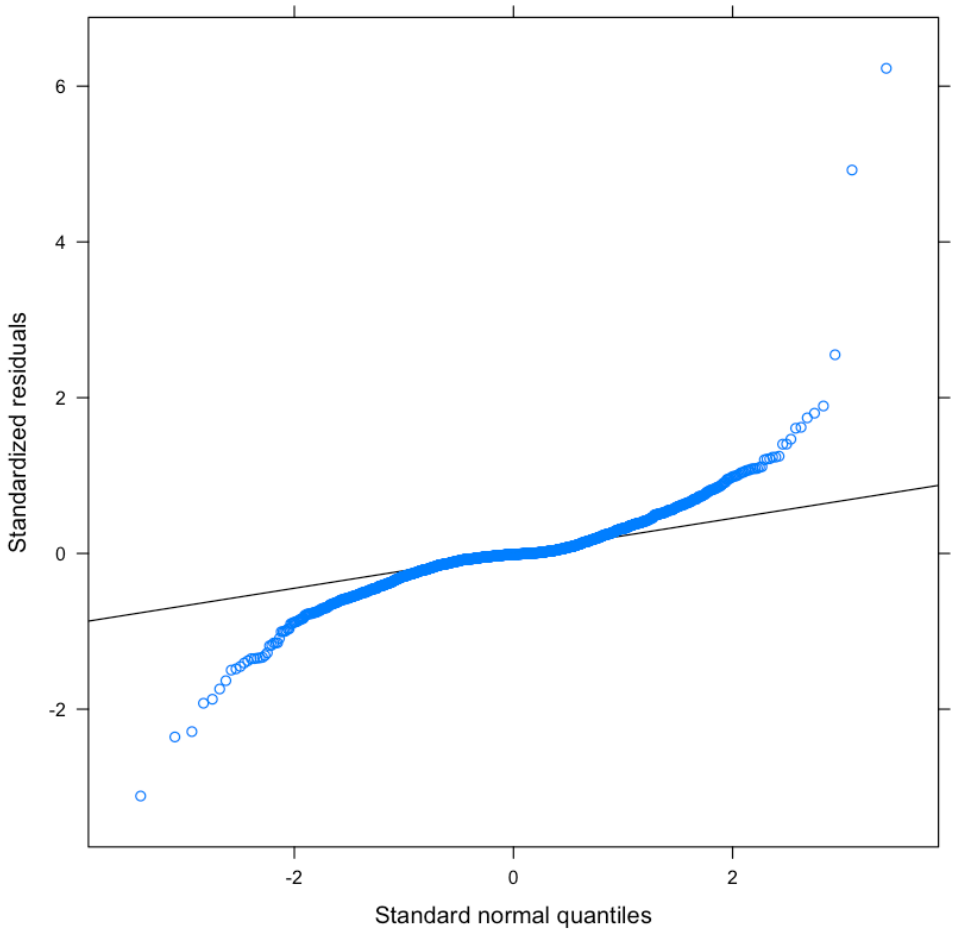

Figure 19: Model 2 QQplot

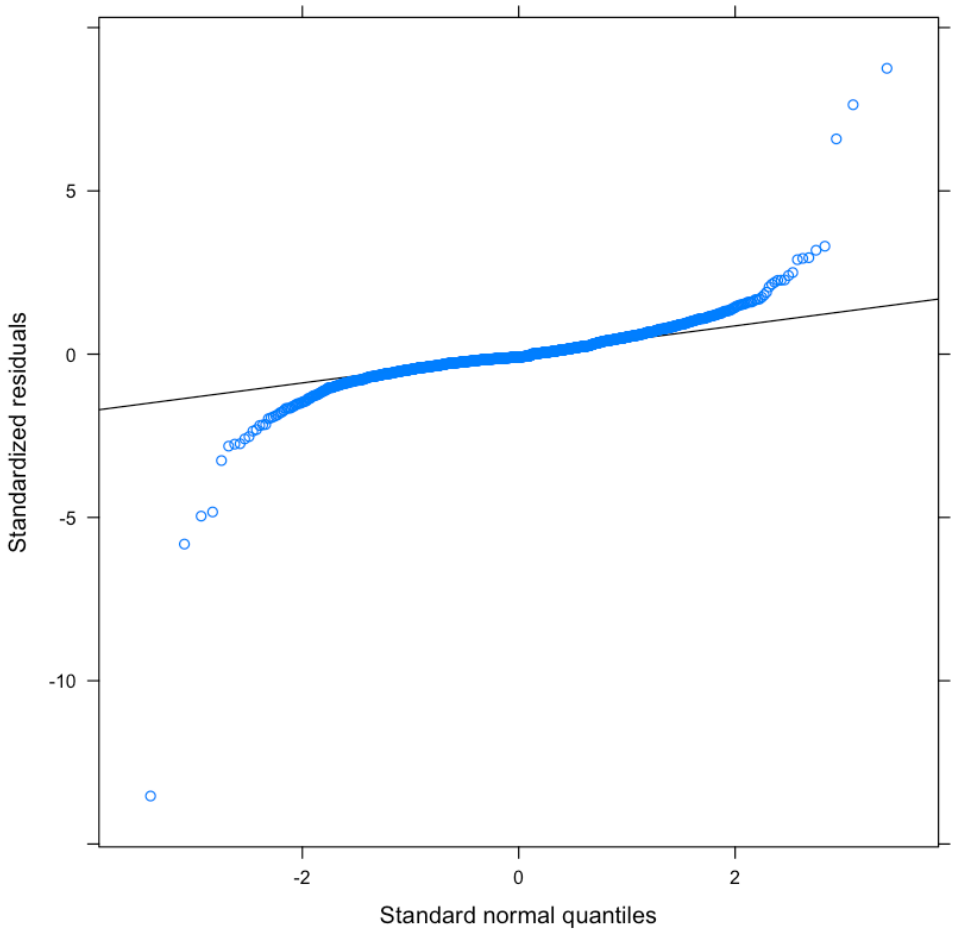

Figure 20: Model 3 QQplot

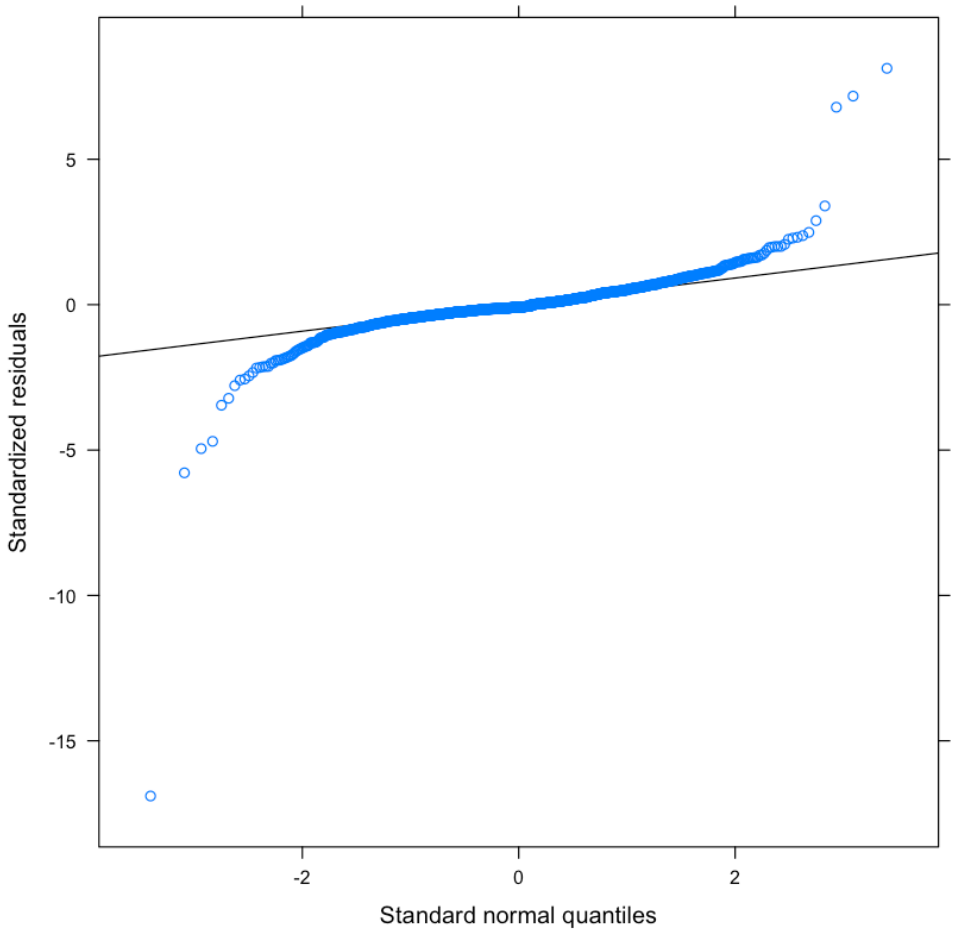

Figure 21: Model 4 QQplot

1.6 Model Evaluation

ROC curves were generated for all models, and the area under these curves was calculated as well. For the final model, the auroc was .9942, while it was .9937, .9631, and .964 for the other models respectively. This is a fantastic fit, and shows why I decided not test interaction terms.

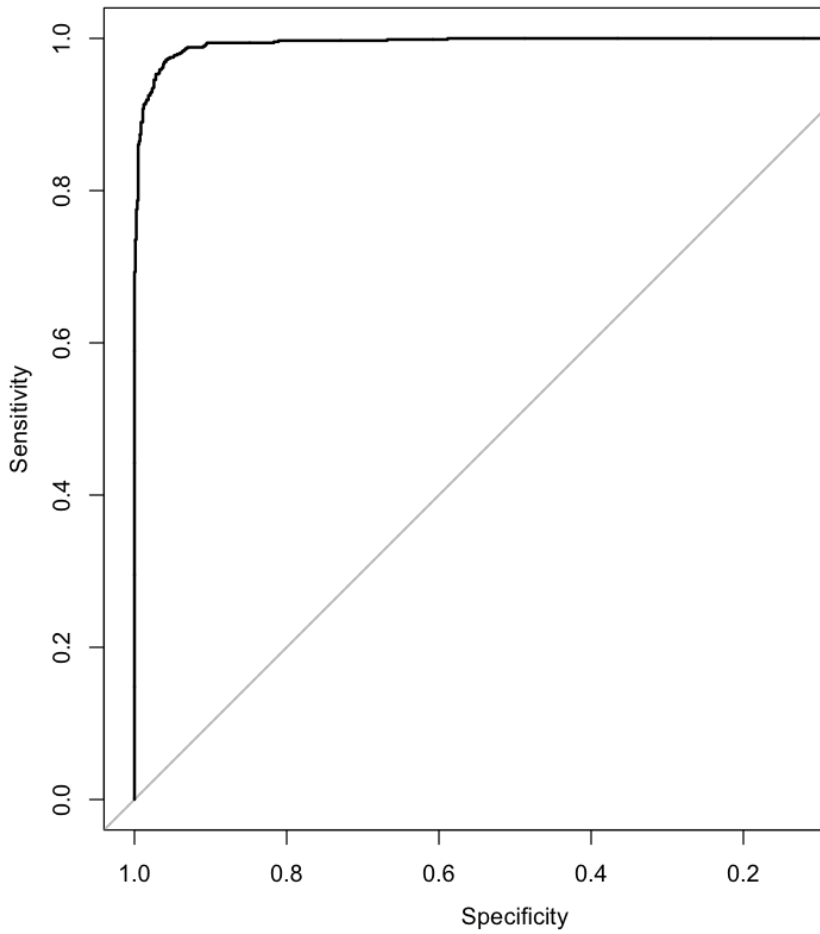

Figure 22: ROC Curve Final Model(model 2)

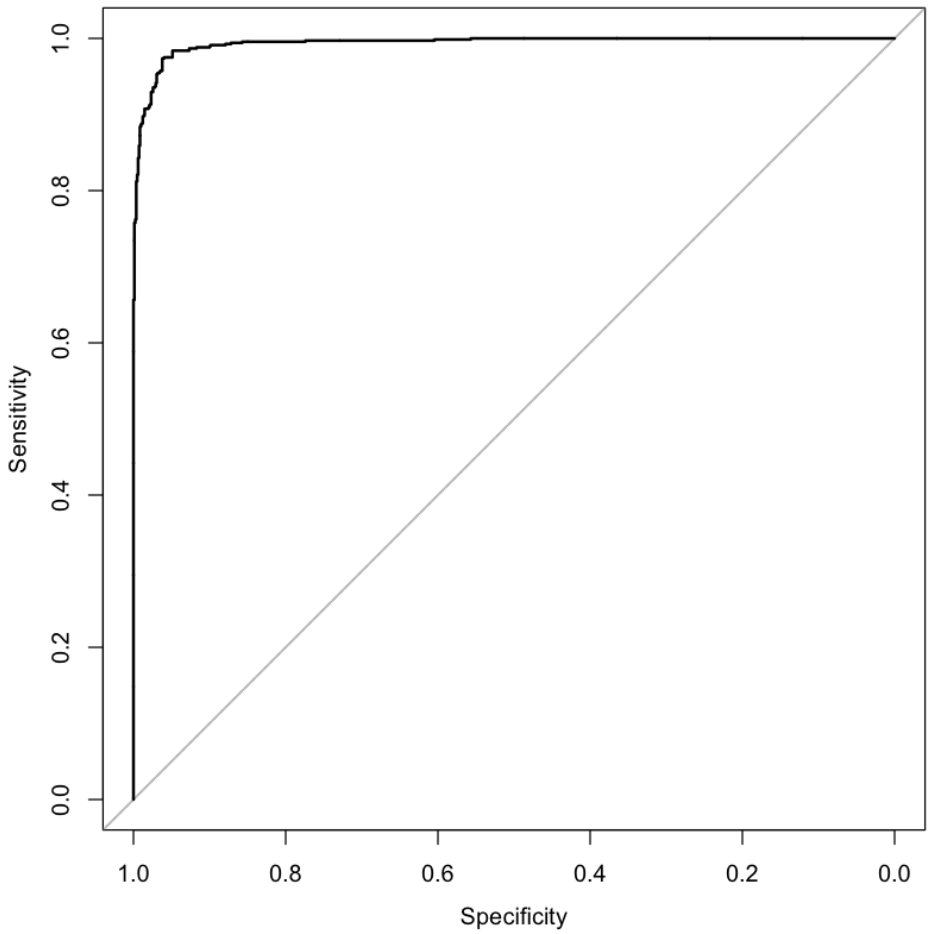

Figure 23: ROC Curve Model 1

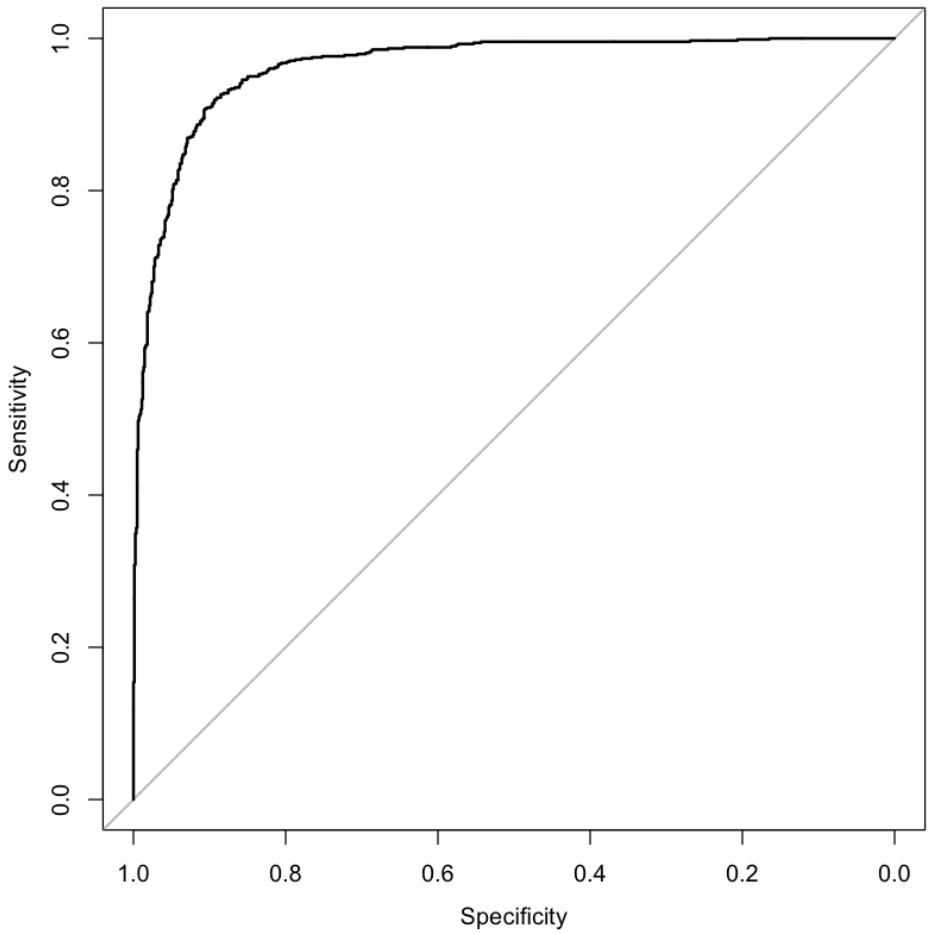

Figure 24: ROC Curve Model 3