Final Project Report
Predicting the 30-Days Mortality Rate of Patient with Covid-19

Yi Huang, Winnie Tam, Benjamin Unger, Wenyi Zhen

December 11, 2021

Contents

List of Tables

List of Figures

Abstract

This project applies the most popular classification methods to predict the mortality rate of patients with Covid-19 and survival analysis to estimate the survival time of these patients. In the United States, there is a significant increase in the 7-day average deaths after Thanksgiving 2021. To find the factors that have huge impacts on the mortality rate of patients who test positive in Covid-19 is our primary interest. The dataset is a public open resource from Harvard Dataverse Santorelli (2021). Despite the different methods, we found age, sex, IMD, renal disease, and cancer are the most important factors that have strongest effects on the death rate of patients with Covid-19.

1 Introduction

Covid-19 has had a tremendous impact on the world ever since its introduction in 2019. Not only has it caused the death of 5.29 million patients and counting around the world, but it has also affected the daily lives of everyone around the world. Due to the contagious nature of Covid-19, numerous workplaces have closed their doors and employees have transitioned to working from home. This, in turn, has affected the economy as well as everyone's livelihood. Despite the distribution of either the Pfizer, Moderna, or Johnson & Johnson vaccine, covid cases remain prevalent at a steady pace. According to the Centers for Disease Control and Prevention (CDC) *CDC Covid Data Tracker.* (2021), there has been a high of 1,500 deaths in a day in the United States in the past 30 days, with as many as 95,000 cases daily. Although 1,500 deaths from 95,000 cases are only a little more than 1.5%, numerically, a high of 1,500 deaths in a day is still an astounding number. This project examines the ethnic, demographic, socio-economic, and clinical risk factors associated with outcomes of COVID-19 positive hospital patients. The question that we are trying to answer is determining which factors among those specified in the dataset have strong impacts on the 30 days mortality of patients with Covid-19.

The dataset has a total of 582 observations. Each observation represents a Covid-19 positive patient that was either hospitalized in Bradford Hospital or Calderdale Hospital, 2 large hospitals located in the United Kingdoms. The categorical and binary variables are shown below:

Died30days: Y/N
Timeatrisk (actual time at risk): from 1-30 days

IMD (Index of Multiple Deprivation): level 1 - 5, a measure of poverty from least to most deprived.
Sex: Male/Female
Agecat: 18-49/50-59/60-69/70-79/80+
Ethnicity: White/SouthAsian/Other
Bmicat : Healthy weight/overweight/obese
Diabetes1:Yes/No
Diabetes2: Yes/No
Hypertension: Yes/No
CVD (Cardiovascular Disease): Yes/No
Asthma: Y/N
COPD (Chronic obstructive pulmonary disease, lung conditions that cause breathing difficulties.): Y/N
Cancer: Y/N
Renal Disease (kidney disease): Y/N

Link: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/RRCQEO

2 Methodology

This project uses logistic regression to investigate the relationship between the binary response variable (whether a patient died within 30 days) and all the categorical and binary predictors James (2021). Then we perform various models on the dataset, including LDA, Naive Bayes, KNN, Classification Trees, Random Forest, Boosting, as well as Support Vector Machine, to explain the data. The model with the minimum test error will be the best model. To avoid overfitting issues in the logistic regression model, Stepwise Selection and Lasso are used to regularize the model. Since KNN, boosting, and SVM can not handle categorical data directly, we apply ordinary encoding and dummy variable encoding to the dataset. Finally, we estimate the survival time for inpatients using survival analysis and compare the results with the best model obtained from the classification methods.

3 Results and Discussion

3.1 Logistic regression

Logistic regression is useful in estimating the probabilities of events and determining a relationship between the independent variables and the dependent variable. First, we manipulated our dataset to exclude the 1st column (id) and the 15th column (ICU). This is because the id column only gives us the number of observations so the id column is unnecessary. ICU is a dependent variable. However, the purpose of our research is to determine the factors that have the most impact on whether a patient died within 30 days of being admitted to the hospital (died30days). We only need the response variable died30days and timeatrisk in our dataset. After our dataset was cleaned, we performed logistic regression on the training data in order to predict the response variable (died30days) using all the other variables and obtained a test error of 0.2989691. Based on Table 1, age, type I diabetes, gender, and IMD are the top four important variables that affect the mortality rate a lot.

3.2 LDA, Naives Bayes, KNN

1. LDA (linear discriminant analysis) is a dimensionality reduction technique and is used to reduce the number of features. After performing LDA on the training data to predict died30days, the test error we obtained was 0.2920962.

2. Naive Bayes is a classification technique based on Bayes theorem and assumes independence between independent variables. We performed Naive Bayes on the training data and obtained a test error of 0.2817869.

3. In order for our code to run KNN, we had to first clean our data by ordinal encoding. For example, for the agecat variable, we assigned “18-49” as “1”, “50-59” as “2”, and so on. For our response variable died30days, we assigned “Yes” as “1” and “No” as “0”. After running through the method, the test error was 0.2852234.

3.3 Classification Tree, Random Forest, Boosting, and SVM

1. Classification Tree

   Decision making trees are helpful to visualize the decision making process using an algorithm to generate rules in splitting our predictor space. In context with our research, our binary response variable assigns ”Yes” if the patient dies and ”No” if the patient does not die. Based on Figure 1 B, the age category predictor is the “root

node", which is the first split in the tree decision. This indicates that age category is a very important feature as it minimizes our cost function. Then the next three important variables are IMD, Gender, and Obese. Additionally, the test error obtained was 0.3195876.

2. Random Forest

   Random forest takes a collection of uncorrelated trees. "The majority votes" is an expression in which each tree predicts the classification of an observation and the majority wins. According to Figure 2 B, the number of trees that minimize our test error without being overly complex is between 95 and 99. Additionally, the test error obtained was 0.3092784.

3. Boosting

   The boosting method demonstrates that the features age, sex, imd, renal disease, and bmi category are the most influential when determining death with a relative influence of 48.48, 11.79, 11.03, and 4.91 respectively (Figure 3)B. The reason why the relative influence for diabetes1 is 0.00 is because most of the patients in the dataset do not have type I diabetes. Additionally, the test error obtained through boosting was 0.2920962. In summary, the test errors for tree, random forest, and boosting were 0.3195876, 0.3092784 and 0.2920962, respectively.

4. SVM

   To peform SVM on this dataset, we need data encoding methods such as ordinal encoding or dummy variable encoding since it can not handle categorical data directly Cerda (2018). The test error for SVM with ordinal encoding is 0.3127148, and for SVM with dummy variable encoding is 0.2818. While compar the test errors, the approach that provide a smaller test error would be the best.

3.4 Feature Selections (R^2, Adjusted R^2, BIC , C_p, Lasso)

In Logistic regression model, it often has overfitting issues when fitting the model with all categorical dataset. Thus, it is essential to apply feature selections.

1. R^2, Adjusted R^2, BIC (Bayesian Information Criterion) and C_p are 4 methods to evaluate the best model while penalizing the number of features. According to Figure 4 B below, the best subset selestion results contain all the predictors except the 80+ group in age category. However, the minimum BIC and C_P values from Figure 5B suggest that the best model contains only 5 variables, which might not be our opitimal

choice. These methods might not select all relevant predictors. Thus, we need to investigate more by applying Lasso to solve the overfitting issues.

2. Lasso Method is an effective regularization technique to reduce model complexity or the number of features. As a result our model will trade more bias for less variance. In Figure 6 B, we have two different lambdas: lambda.min (the lambda that gives the smallest cross-validated error) and lambda.1se (the lambda that gives the most regularized model). The lambda.min was found to be 0.009547284 whereas the lambda.1se was found to be 0.0423004. The test error obtained using the lambda.min was 0.2955326 and the test error obtained using lambda.1se was 0.3264605. As a result, given the test errors, the model with the lambda that gives the smallest cross-validated error is the best model (0.2955326) as opposed to the full model (0.2989691).

3.5 Survival Analysis

Survival Analysis or Time to Event Analysis is a statistical method that is often used to estimate the survival time for patients. In our project, we convert the two response variable status and time from died30days and time at risk. For example, if a patient dies in 15 days, then his/her survival time will be 15 days, and status as not censored. If a patient does not die in 30 days, then his/her status will be censored. The length of this study is 30days. We use 1 to denote censored and 0 to denote not censored.

1. Kaplan-Meier Curve

   We use Kaplan-Meier Curve to estimate the survival function for patient with Covid-19 in our study. Based on Figure 7 B, the estimated probability of survival for patients past 10 days is approximately 79%, and past 20 days drop to 74%. As time increases, the probability of survival for patients decreases.
   The probability of survival for patients decrease as age increase. For instance, at the end of 30 days, the probability of survival for the youngest age group (18-49) is 96%, and for oldest age group (80+) is only 43% as shown in Figure 8 B. As we add the predictor, sex, into the curves, there is a significant difference between female and male in Figure 9 B. By the end of the study, the probability of survival for female is 16% greater than male.

2. Log-rank Test

   In this section, we perform a Log-rank Test to compare the survival times between female and male. We want to check if there is any significant difference between two

groups.

H_0: No difference between Female and Male

H_a: There is a difference between different genders.

Test Statistics W = 11.67 with $P - Value < 0.001$.

Since $P - Value$ is smaller than significant level 0.01, we reject the null hypothesis. There is a significant difference between female and male.

3. Cox's Proportional Hazard Model

   From the summary of this model, we find the most important variable are age, diabetes1, Ethnicity, and Renal diseases. If all other variables are hold at the same level and set the reference as age group 18-49, the estimated hazard for patients at age of 80+ is 18 times greater than the reference group. Similarly, the estimate hazard for patients with Type I diabetes is 2.53 times greater than not having Type I diabetes. The estimate hazard of male patient is 1.8 times greater than female.

4 Conclusion

After comparing different methods, we conclude that being male, aged group, from deprived areas, with renal disease, and cancer are more likely to die from Covid-19 (Pradhan and Olsson., 2021). The difference in immune system function, and stress endurance level between males and females could be important determinants to explain the significant difference in probability of survival between females and male. The average accuracy for all classification models is around 70%, where the SVM with dummy variables has the smallest test error 0.2818, and the classification tree has the highest test error 0.3193 (Table 3).
Although some of the methods shrink the coefficient of Type I diabetes to zero, many methods list it as an important factor. We think it is important to know patient with Type I diabetes are at higher risk compare to those who do not have Type I diabetes.

References

CDC Covid Data Tracker. (2021), *Centers for Disease Control and Prevention* .

Cerda, Patricio, e. a. (2018), 'Similarity encoding for learning with dirty categorical variables', *Machine Learning* **107**(8-10), 1477–1494.

James, Gareth, e. a. (2021), *An Introduction to Statistical Learning: With Applications in R.*, Springer.

Pradhan, A. and Olsson., P.-E. (2021), 'Sex differences in severity and mortality from covid-19: are males more vulnerable?', *Biology of sex differences* **11**(1), 53.

Santorelli, Gillian, e. a. (2021), 'Ethnicity, pre-existing comorbidities, and outcomes of hospitalised patients with covid-19', *Harvard Dataverse,* .

A First section of Appendix.

Table 1: Summary of Logistic Regression Model

	Estimate	Std.Error	t value	Pr(> |t|)
Intercept	-19.10643	991.89749	-0.019	0.98463
sexMale	0.93953	0.32191	2.919	0.00352 **
agecat50-59	15.96034	991.89742	0.016	0.98716
agecat60-69	17.41906	991.89732	0.018	0.98599
agecat70-79	17.74660	991.89733	0.018	0.98573
agecat80+	18.42925	991.89732	0.019	0.98518
ethnicity3catSouth Asian	-0.30019	0.81294	-0.369	0.71193
ethnicity3catWhite	0.36708	0.77651	0.473	0.63640
imdIMD 2	-0.93974	0.41880	-2.244	0.02484 *
imdIMD 3	-0.37261	0.45425	-0.820	0.41206
imdIMD 4/5	-0.71592	0.44298	-1.616	0.10606
bmicatObese	0.31807	0.40989	0.776	0.43776
bmicatOverweight	0.36553	0.38434	0.951	0.34157
diabetes1Yes	1.43355	1.70150	0.843	0.39950
diabetes2Yes	0.32303	0.33677	0.959	0.33746
hypertensionYes	0.05384	0.30607	0.176	0.86035
cvdYes	0.24640	0.34189	0.721	0.47108
asthmaYes	-0.13032	0.47386	-0.275	0.78330
copdYes	-0.62027	0.38524	-1.610	0.10738
cancerYes	0.43277	0.54068	0.800	0.42347
renaldiseaseYes	0.40444	0.32118	1.259	0.20795
Null deviance:	366.15	on 290 degrees of freedom		
Residual deviance:	281.48	on 270 degrees of freedom		
AIC:	323.48			
Number of Fisher Scoring iterations: 17				

B Second section of Appendix.

Table 2: Summary of Cox's Proportional Hazard Model

Call: coxph(formula = Surv(time, status) sex + agecat + ethnicity3cat + imd + bmicat + diabetes1 + diabetes2 + hypertension + cvd + asthma + copd + cancer + renaldisease)

	coef	exp(coef)	se(coef)	z	p
sexMale	0.58566	1.79618	0.16134	3.630	0.000283
agecat50-59	1.54529	4.68935	0.65004	2.377	0.017443
agecat60-69	2.12781	8.39649	0.61447	3.463	0.000535
agecat70-79	2.53884	12.66496	0.62071	4.090	4.31e-05
agecat80+	2.89227	18.03412	0.61699	4.688	2.76e-06
ethnicity3catSouth Asian	0.44847	1.56591	0.44133	1.016	0.309545
ethnicity3catWhite	0.74270	2.10160	0.41717	1.780	0.075023
imdIMD 2	-0.24973	0.77901	0.20543	-1.216	0.224123
imdIMD 3	0.03202	1.03254	0.21625	0.148	0.882279
imdIMD 4/5	-0.12741	0.88037	0.22251	-0.573	0.566922
bmicatObese	0.14492	1.15594	0.20488	0.707	0.479376
bmicatOverweight	0.21639	1.24158	0.18449	1.173	0.240842
diabetes1Yes	0.92882	2.53151	0.62414	1.488	0.136712
diabetes2No	0.21059	1.23440	0.16613	1.268	0.204938
hypertensionYes	-0.10730	0.89825	0.15581	-0.689	0.491032
cvdYes	0.14174	1.15227	0.16669	0.850	0.395157
asthmaYes	-0.02758	0.97280	0.23166	-0.119	0.905231
copdYes	-0.16284	0.84973	0.19951	-0.816	0.414394
cancerYes	0.37920	1.46111	0.22121	1.714	0.086485
renaldiseaseYes	0.49168	1.63505	0.16021	3.069	0.002148
Likelihood ratio test=137.5	on 20 df,	p=¡ 2.2e-16			
n= 582,	number of events=	189			

Table 3: Table of Test Errors

	Test Error
Logistic Regression	0.2989691
Lasso Logistic	0.2955326
LDA	0.2920962
Naive Bayes	0.2818
KNN	0.2852234
SVM (Ordinal Encoding)	0.3127148
SVM (Dummy Variable Encoding)	0.2818000
Classification Tree	0.3195876
Random Forest	0.3092784
Boosting	0.2920962

Table 4: Order of Importence Variables

Logistic	Lasso Logistic	LDA	Naive Bayes	Tree	Boosting	Survival Analysis
Age	Age	Age	Age	Age	Age	Age
Diabetes1	Sex	Diabetes1	Sex	IMD	Sex	Diabetes1
Sex	IMD	Sex	IMD	Sex	IMD	Ethnicity
IMD	Renal Disease	Copd	Diabetes1	BMI	Copd	Renal diseases

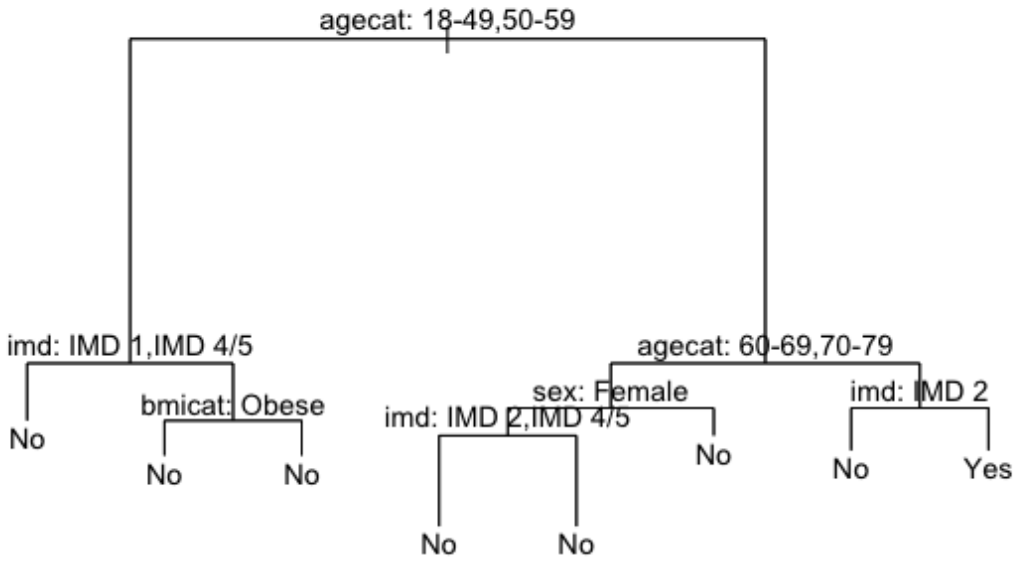

Figure 1: Classification Tree

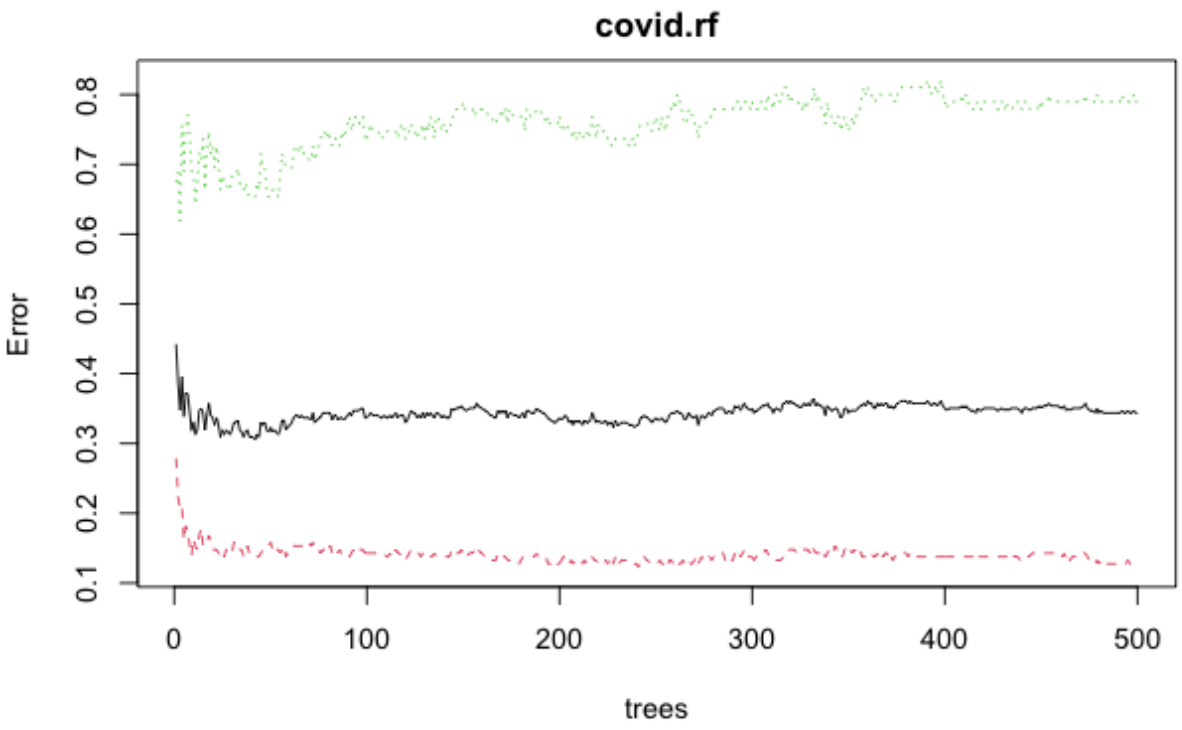

Figure 2: Random Forest

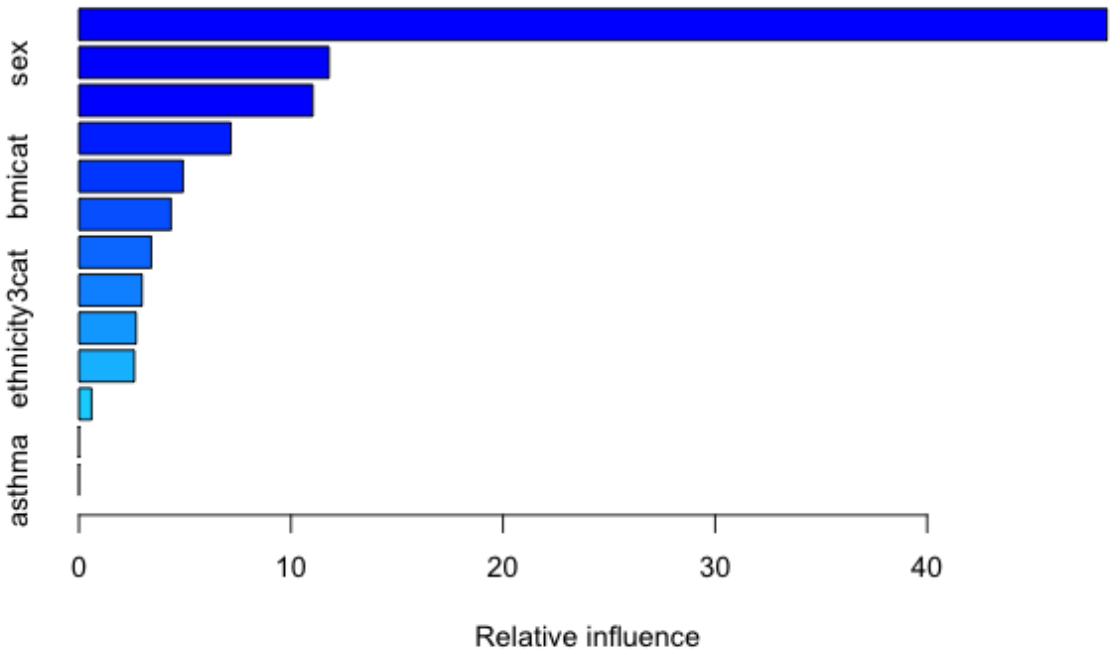

Figure 3: Boosting

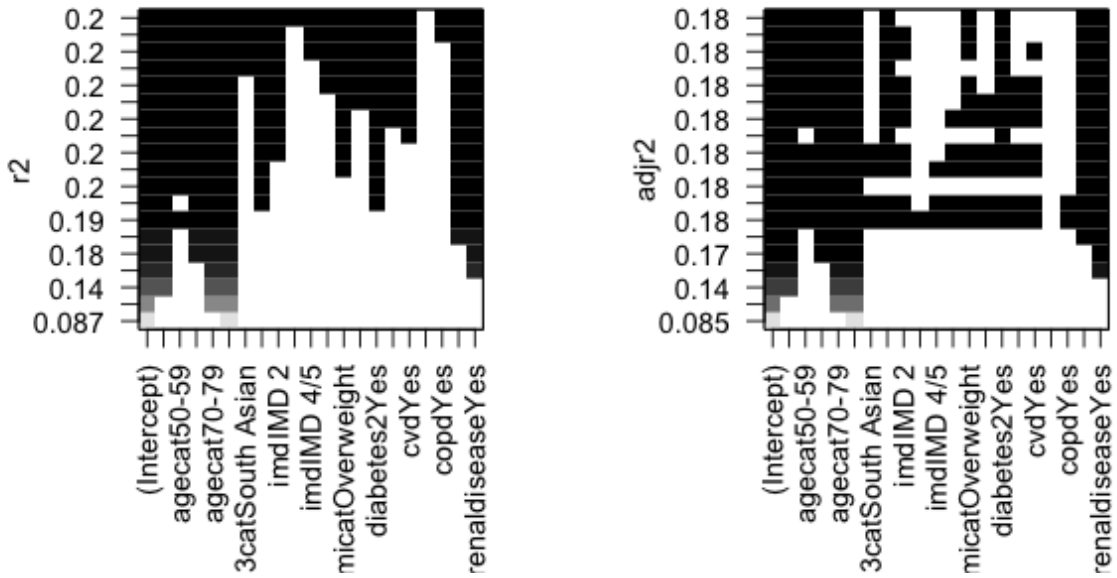

Figure 4: R^2 and AdjustedR^2

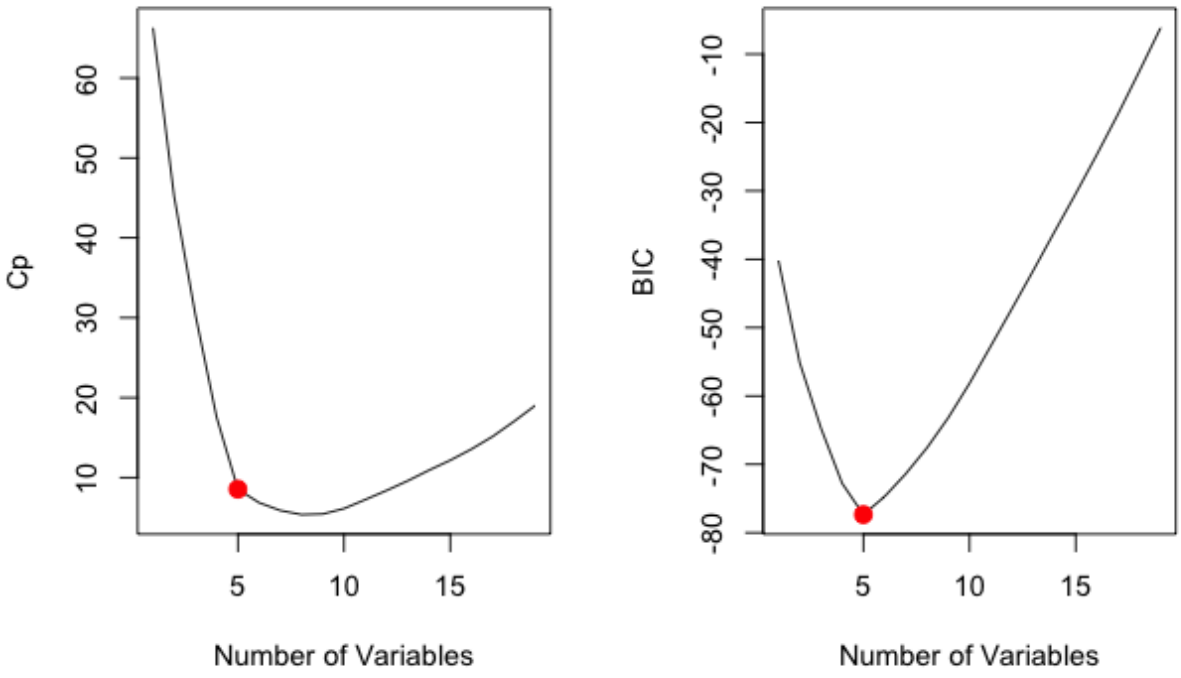

Figure 5: BIC and C_p

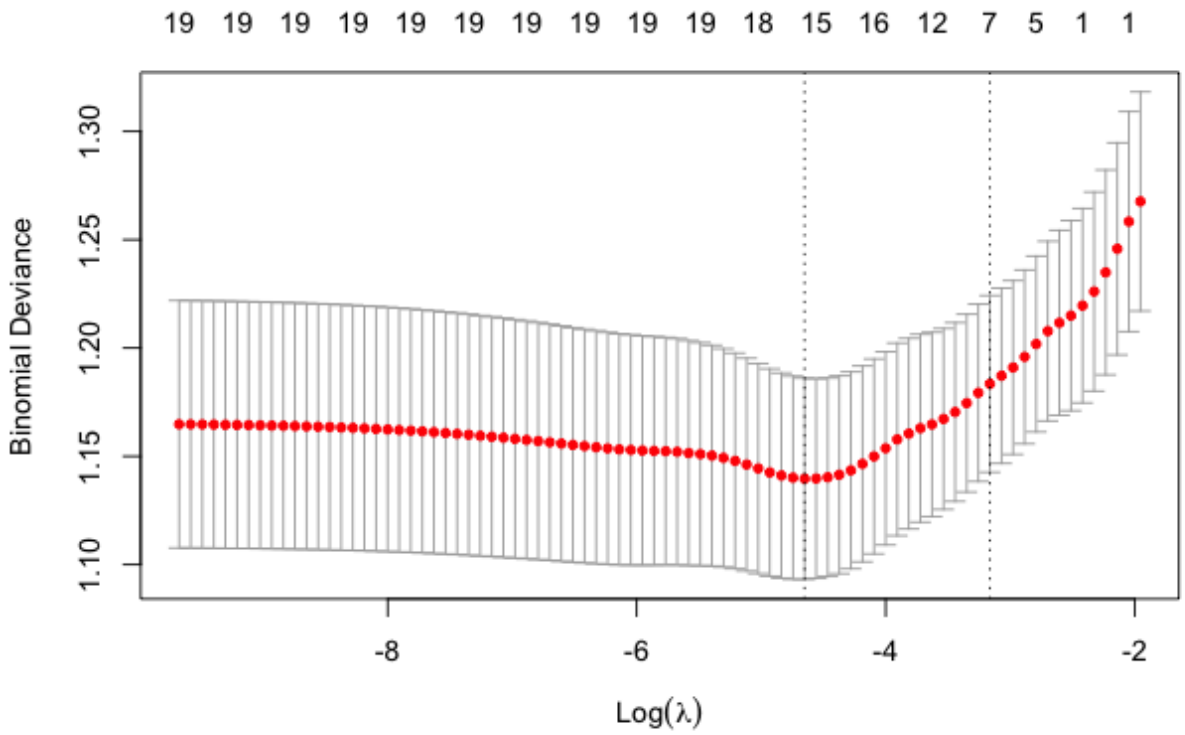

Figure 6: Binomial Deviance against Log(λ)

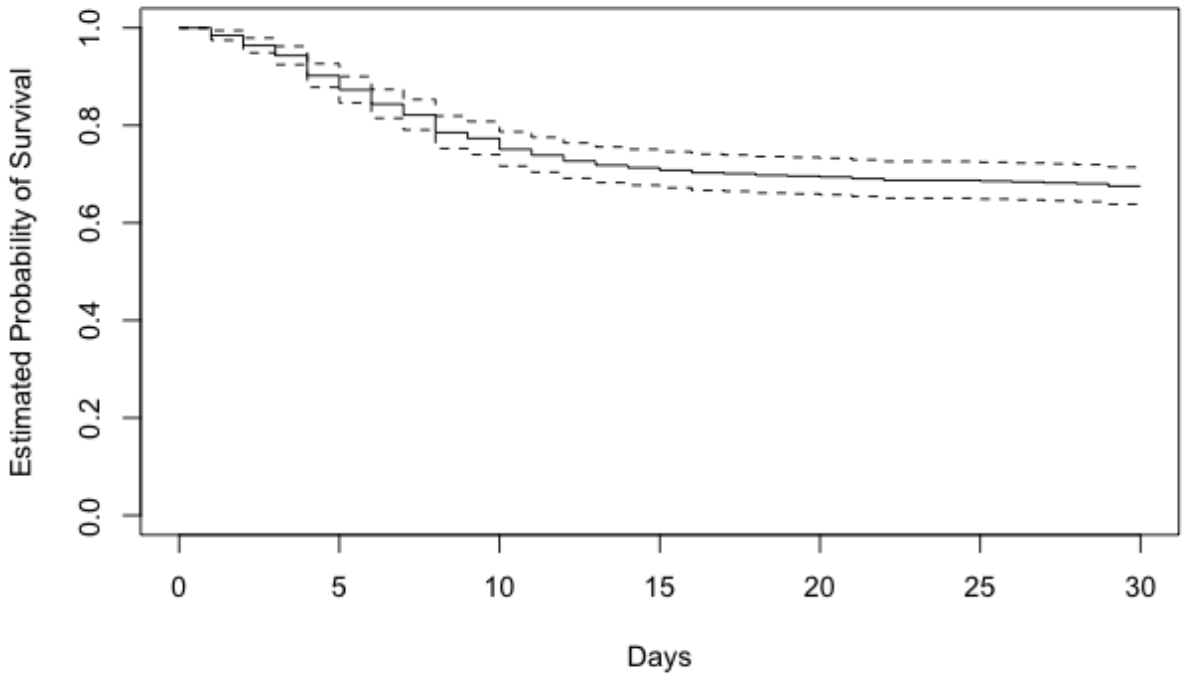

Figure 7: Kaplan-Meier Curve

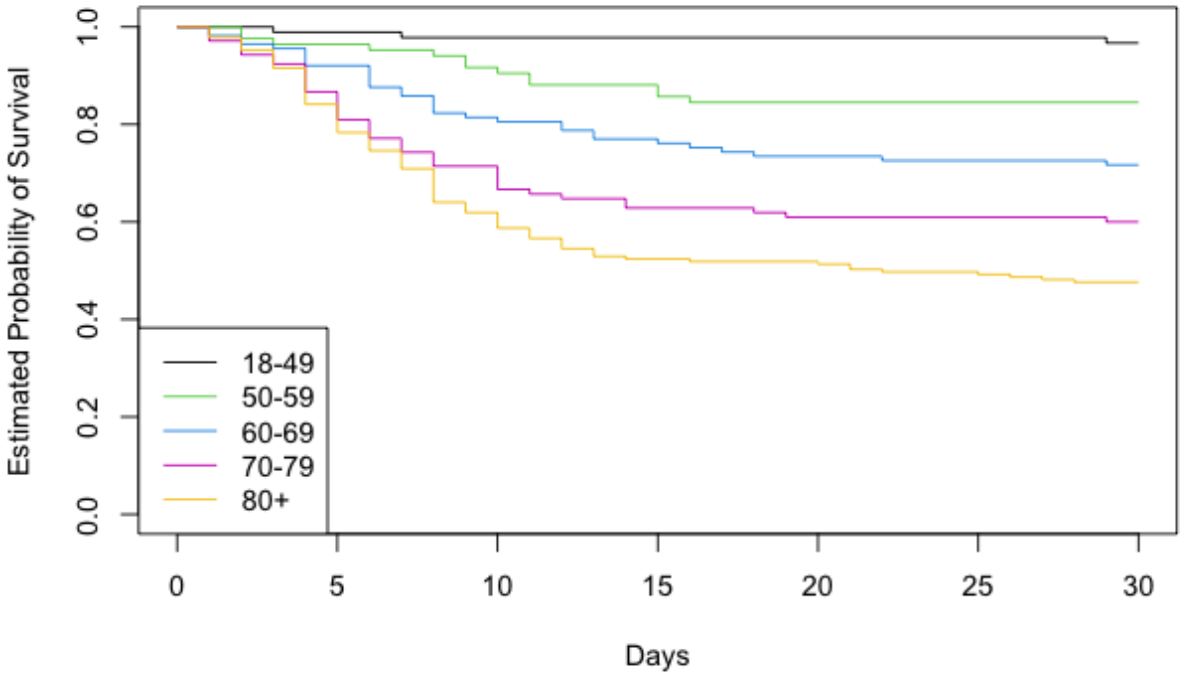

Figure 8: Kaplan-Meier Curve for Age

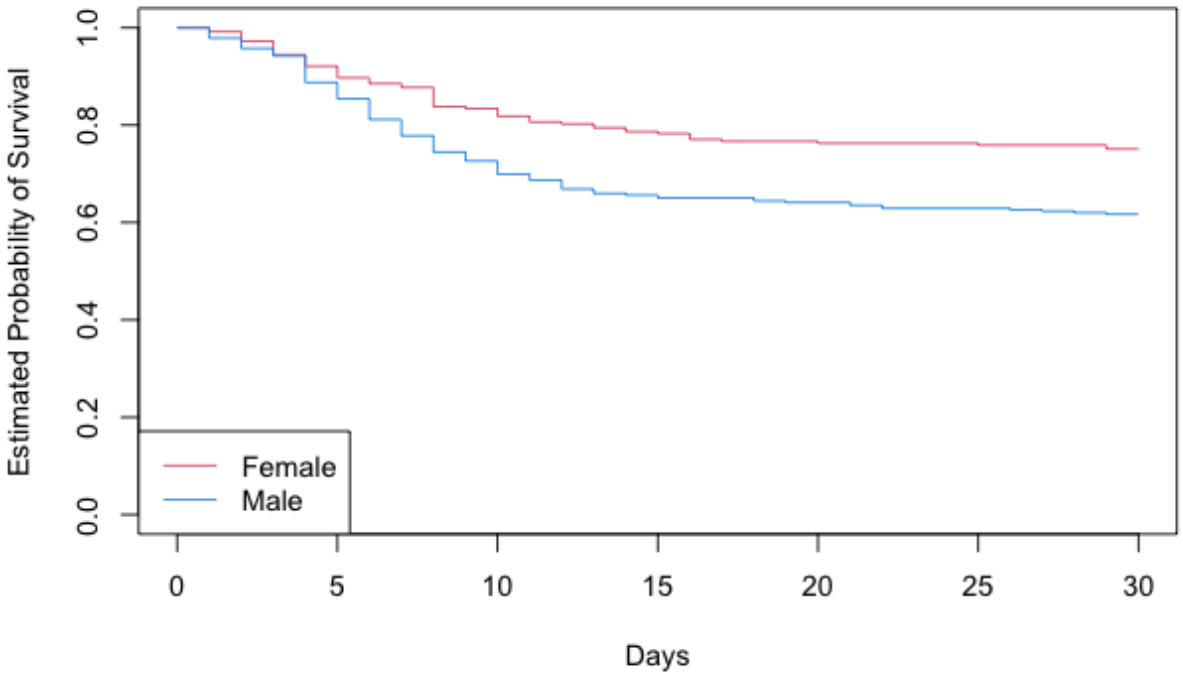

Figure 9: Kaplan-Meier Curve for Sex