Heart Disease: What is It and How to Predict It

Chad Huntebrinker

2025-05-09

# Abstract

Each year there are about 17.9 million lives lost to cardiovascular diseases, which is the leading cause of death globally (World Health Organization, 2023). One such type of cardiovascular disease is heart disease. In the United States, heart disease is the leading cause of death for men and women; in 2022, 702,880 people died from heart disease (Centers for Disease Control and Prevention, 2024). One of the problems with heart disease is that it can be difficult to have early detection due to it being related to multiple different medical and lifestyle factors. This results in many cases that might have been easily prevented or cured instead go undetected. Fortunately, with the help of data-modeling and using them to make predictions, we can create models that assist in forecasting and understanding heart disease. For this final project, I hope to answer the following research questions:

What are the most significant indicators that predict the presence of heart disease?

What are the most significant indicators for predicting the severity of the heart disease present?

Understanding which variables contribute most to the likelihood of heart disease could help in early screening, prevention, and treatment plans.

# Introduction

In order to find the predictors for heart disease, we must understand what heart disease is first. According to Mayo Clinic, heart disease is a range of conditions including blood vessel disease, irregular heartbeats, heart conditions that individuals are born with (called congenital heart defects), disease of the heart muscle, and heart valve disease. While there are multiple types of heart diseases, it boils down to a condition that impacts both the heart and the blood vessels by it.

I will be using the Heart Disease dataset from the UCI Machine Learning Repository. More specifically, I will be using the data collected at the Cleveland clinic as that’s the only one that has been used by machine learning researchers to date. The database was collected in the following way:

The reference group used to derive the model consisted of 303 consecutive patients referred for coronary angiography at the Cleveland Clinic between May 1981 and September 1984. No patient had a history or electrocardiographic evidence of prior myocardial infarction or known valvular or cardiomyopathic disease. All 303 patients provided a history and underwent physical examination, electrocardiogram at rest, serum cholesterol determination and fasting blood sugar determination as part of their routine evaluation. Historical data were recorded and coded without knowledge of noninvasive or angiographic test data. In addition, after giving informed consent, the patients underwent 3 noninvasive tests as part of a research protocol. The results of these tests (exercise electrocardiogram, thallium scintigraphy and cardiac fluoroscopy) were not interpreted until after the coronary angiograms had been read. These tests were analyzed and the results recorded without knowledge of the historical or angiographic results… The mean age of these patients was 54 years; 206 were men. Angiograms were interpreted by a cardiologist without knowledge of other test data (Deterano, Robert, 1998).

The dataset I will be using includes 303 observations with the following 14 variables: Age: Age of the patient in years

Sex: Male or female

Cp: Chest pain type (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)

Trestbps: Resting blood pressure in mmHg

Chol: Serum cholesterol in mg/dL

Fbs: Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)

Restecg: Resting electrocardiographic results (0 = normal, 1 = ST-T wave abnormality, 2 = probable left ventricular hypertrophy)

Thalach: Maximum heart rate during exercise

Exang: Exercise-induced angina (1 = yes, 0 = no)

Oldpeak: ST depression induced by exercise related to rest

Slope: Slope of the peak exercise ST segment (1 = upsloping, 2 = flat, 3 = downsloping)

Ca: Number of major vessels colored by fluoroscopy

Thal: Result from a Thallium stress test (3 = normal, 6 = fixed defect, 7 = reversible defect).

num: Presence of heart disease (0 = no heart disease, 1–4 = heart disease of increasing severity).

The CDC states that some key factors for heart disease are high cholesterol, high blood pressure, age, and sex (Centers for Disease Control and Prevention, 2024). I’ll be taking these experts’ opinions into consideration when building the model and will be starting the baseline model with these 4 variables. Other variables will be added later to see if they help in explaining heart disease factors. For example, not enough exercise can lead to heart disease and our data set includes a variable labeled “thalach” which measures the maximum heart rate achieved during exercise.

Due to the response variable being within a specific range (a number between 0 and 4), the response variable being ordinal, and having multiple variables to check their impact on the response variable, multiple cumulative logistic regression models will be fitted to the data. Another thing to note though was that while the severity of the heart disease present increased, the number of patients that had it decreased.

##   
## 0 1 2 3 4   
## 164 55 36 35 13

This made it difficult to train the model due to the lack of data. So to remedy this, a combination was done on value 1 and 2 (a mild case of heart disease, indicated by a 1) and 3 and 4 (a severe case of heart disease, indicated by a 2) to help balance out the data.

##   
## 0 1 2   
## 164 91 48

As a result, our model will be predicting whether there is no heart disease (0), a mild case of heart disease (1), or a severe case of heart disease (2) in a our new response variable “num3”.

# Methods

The first model created is with factors found on the CDC website: cholesterol, blood pressure, age, and sex.

##   
## Call:  
## vglm(formula = num3 ~ trestbps + chol + age + sex, family = cumulative(parallel = TRUE),   
## data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 6.687973 1.238216 5.401 6.62e-08 \*\*\*  
## (Intercept):2 8.372008 1.272945 6.577 4.80e-11 \*\*\*  
## trestbps -0.013619 0.006822 -1.996 0.045904 \*   
## chol -0.003787 0.002336 -1.621 0.104954   
## age -0.049714 0.014249 -3.489 0.000485 \*\*\*  
## sex -1.497985 0.288274 -5.196 2.03e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 547.7772 on 600 degrees of freedom  
##   
## Log-likelihood: -273.8886 on 600 degrees of freedom  
##   
## Number of Fisher scoring iterations: 5   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## trestbps chol age sex   
## 0.9864729 0.9962204 0.9515014 0.2235802

What’s found is that the p-value for the variable chol is greater than our cutoff of 0.05. As a result, an effort to remove it from the model will be tried by conducting a likelihood ratio test to see if the model improves without it. Our null hypothesis is that removing chol from the model improves it while the alternative hypothesis is that including chol improves the model.

##   
## Call:  
## vglm(formula = num3 ~ trestbps + age + sex, family = cumulative(parallel = TRUE),   
## data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 5.914384 1.117084 5.294 1.19e-07 \*\*\*  
## (Intercept):2 7.588924 1.152030 6.587 4.47e-11 \*\*\*  
## trestbps -0.014393 0.006793 -2.119 0.034110 \*   
## age -0.052237 0.014056 -3.717 0.000202 \*\*\*  
## sex -1.397762 0.277873 -5.030 4.90e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 550.2774 on 601 degrees of freedom  
##   
## Log-likelihood: -275.1387 on 601 degrees of freedom  
##   
## Number of Fisher scoring iterations: 5   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## trestbps age sex   
## 0.9857102 0.9491034 0.2471494

## Likelihood ratio test  
##   
## Model 1: num3 ~ trestbps + chol + age + sex  
## Model 2: num3 ~ trestbps + age + sex  
## #Df LogLik Df Chisq Pr(>Chisq)  
## 1 600 -273.89   
## 2 601 -275.14 1 2.5002 0.1138

What’s found is that the variables’ p-values for the simpler model are all less than 0.05 (and, thus, help to explain the presence and severity of heart disease). Also, our likelihood ratio test has a p-value of 0.1138. Since this is greater than 0.05, the null hypothesis is not rejected and the conclusion is that removing cholesterol from the model increases it’s effectiveness.

Now that there is a baseline model, the next step is to try including the rest of the variables to see what helps the model.

## $`cp model`  
## (Intercept):1 (Intercept):2 trestbps age sex   
## 6.378957e-14 1.319082e-17 1.677952e-03 1.402923e-03 1.237152e-07   
## cp   
## 2.172664e-12   
##   
## $`fbs model`  
## (Intercept):1 (Intercept):2 trestbps age sex   
## 1.397548e-07 5.795701e-11 3.433574e-02 1.967641e-04 4.839417e-07   
## fbs   
## 8.185616e-01   
##   
## $`restecg model`  
## (Intercept):1 (Intercept):2 trestbps age sex   
## 2.223864e-07 8.478554e-11 6.064962e-02 5.284022e-04 6.911703e-07   
## restecg   
## 1.290061e-02   
##   
## $`slope model`  
## (Intercept):1 (Intercept):2 trestbps age sex   
## 1.399845e-09 2.361932e-13 6.142289e-02 2.614299e-03 7.656522e-07   
## slope   
## 6.385133e-09   
##   
## $`ca model`  
## (Intercept):1 (Intercept):2 trestbps age sex   
## 1.520338e-03 1.377237e-05 7.480697e-03 7.774149e-01 2.605952e-05   
## ca0 ca1 ca2 ca3   
## 6.556351e-01 8.761180e-02 1.175330e-02 1.033644e-02   
##   
## $`exang model`  
## (Intercept):1 (Intercept):2 trestbps age sex   
## 1.567964e-07 3.059695e-11 5.160442e-02 8.917814e-04 6.567301e-06   
## exang   
## 4.881657e-11   
##   
## $`old peak model`  
## (Intercept):1 (Intercept):2 trestbps age sex   
## 8.504130e-06 2.618330e-09 1.761263e-01 1.040143e-02 1.073220e-05   
## oldpeak   
## 6.536853e-12   
##   
## $`thalach model`  
## (Intercept):1 (Intercept):2 trestbps age sex   
## 5.200913e-01 5.480627e-01 6.128789e-03 2.894217e-01 1.434202e-06   
## thalach   
## 3.539761e-10   
##   
## $`thal model`  
## (Intercept):1 (Intercept):2 trestbps age sex   
## 0.0185934648 0.0006387471 0.1362606608 0.0039461018 0.0491141589   
## thal3 thal6 thal7   
## 0.4260244984 0.8307200819 0.4744379289

According to the p-values, the model with fbs, restecg, ca, and thal are greater than 0.05. Thus, removal from consideration is the best next step. But the other variables seem like good predictors so the next step for them is to check their AIC scores

## AIC score of model 1 (includes age): 559.7772

## AIC score of model 2 (excludes age): 560.2774

## AIC score of model with cp: 497.9403

## AIC score of model with slope: 527.6302

## AIC score of model with exang: 515.7472

## AIC score of model with oldpeak: 509.8045

## AIC score of model with thalach: 518.4035

What’s found is that slope doesn’t have a huge drop in AIC like the others. So for the sake of keeping the model simple, slope will not be added to the model.

But the rest of the models improve when adding the new predictor variables; the next step would be to try adding all of them into our model.

##   
## Call:  
## vglm(formula = num3 ~ trestbps + age + sex + cp + exang + oldpeak +   
## thalach, family = cumulative(parallel = TRUE), data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 5.224128 1.894627 2.757 0.00583 \*\*   
## (Intercept):2 7.593980 1.921530 3.952 7.75e-05 \*\*\*  
## trestbps -0.017453 0.007584 -2.301 0.02137 \*   
## age -0.023285 0.016472 -1.414 0.15747   
## sex -1.381967 0.312185 -4.427 9.57e-06 \*\*\*  
## cp -0.750086 0.161357 -4.649 3.34e-06 \*\*\*  
## exang -0.677746 0.291341 -2.326 0.02000 \*   
## oldpeak -0.545514 0.117966 -4.624 3.76e-06 \*\*\*  
## thalach 0.017881 0.006622 2.700 0.00693 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 435.6682 on 597 degrees of freedom  
##   
## Log-likelihood: -217.8341 on 597 degrees of freedom  
##   
## Number of Fisher scoring iterations: 6   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## trestbps age sex cp exang oldpeak thalach   
## 0.9826979 0.9769840 0.2510842 0.4723259 0.5077604 0.5795439 1.0180421

What’s now found is that the p-value for age is greater than our cutoff of 0.05. A likelihood ratio test will be conducted to see if it should be included in the model or not. The null hypothesis is that the model removing age is better while the alternative hypothesis is that including age will help it. A comparison of their AIC scores will also be done.

##   
## Call:  
## vglm(formula = num3 ~ trestbps + sex + cp + exang + oldpeak +   
## thalach, family = cumulative(parallel = TRUE), data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 3.692286 1.550675 2.381 0.017262 \*   
## (Intercept):2 6.058450 1.576823 3.842 0.000122 \*\*\*  
## trestbps -0.019667 0.007425 -2.649 0.008079 \*\*   
## sex -1.331476 0.308502 -4.316 1.59e-05 \*\*\*  
## cp -0.736869 0.160701 -4.585 4.53e-06 \*\*\*  
## exang -0.642560 0.289512 -2.219 0.026455 \*   
## oldpeak -0.556168 0.117886 -4.718 2.38e-06 \*\*\*  
## thalach 0.020986 0.006228 3.370 0.000753 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 437.6474 on 598 degrees of freedom  
##   
## Log-likelihood: -218.8237 on 598 degrees of freedom  
##   
## Number of Fisher scoring iterations: 6   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## trestbps sex cp exang oldpeak thalach   
## 0.9805255 0.2640871 0.4786103 0.5259442 0.5734023 1.0212075

## Likelihood ratio test  
##   
## Model 1: num3 ~ trestbps + sex + cp + exang + oldpeak + thalach  
## Model 2: num3 ~ trestbps + age + sex + cp + exang + oldpeak + thalach  
## #Df LogLik Df Chisq Pr(>Chisq)  
## 1 598 -218.82   
## 2 597 -217.83 -1 1.9792 0.1595

## AIC score of model with age: 453.6682

## AIC score of model without age: 453.6474

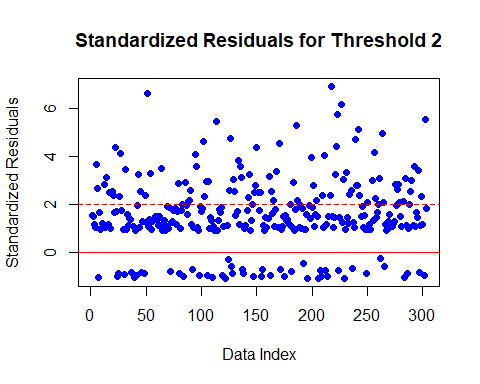
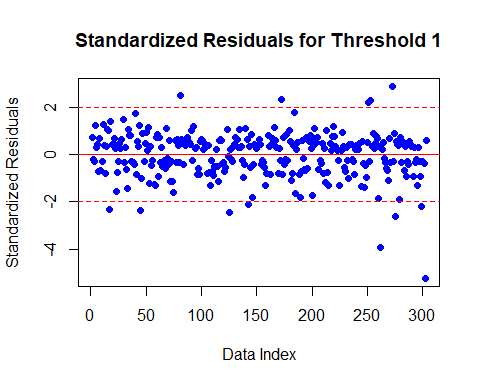
The AIC scores are almost identical and the p-value in the likelihood ratio test is 0.1595. Since the p-value is greater than 0.05, the null hypothesis is failed to be rejected and the model without age is better than the model with age. With these findings, the final model is almost completed. The last step is to check if there are any interaction variables that should be included. Specifically, an investigation was done to see if sex has any interaction with other predictors. A model with sex interaction terms was created with the following:

##   
## Call:  
## vglm(formula = num3 ~ sex \* trestbps + sex \* cp + sex \* exang +   
## sex \* oldpeak + sex \* thalach, family = cumulative(parallel = TRUE),   
## data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 10.003598 3.560366 NA NA   
## (Intercept):2 12.410797 3.587145 3.460 0.000541 \*\*\*  
## sex -10.469859 4.010032 -2.611 0.009030 \*\*   
## trestbps -0.033107 0.016058 -2.062 0.039229 \*   
## cp -1.301728 0.501002 -2.598 0.009370 \*\*   
## exang -0.710426 0.652256 -1.089 0.276073   
## oldpeak -0.901598 0.270490 -3.333 0.000859 \*\*\*  
## thalach 0.007028 0.015393 0.457 0.647990   
## sex:trestbps 0.022411 0.018289 1.225 0.220434   
## sex:cp 0.688456 0.528328 1.303 0.192546   
## sex:exang 0.210878 0.733705 0.287 0.773794   
## sex:oldpeak 0.461284 0.298940 1.543 0.122815   
## sex:thalach 0.020952 0.017036 1.230 0.218739   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 427.9347 on 593 degrees of freedom  
##   
## Log-likelihood: -213.9673 on 593 degrees of freedom  
##   
## Number of Fisher scoring iterations: 6   
##   
## Warning: Hauck-Donner effect detected in the following estimate(s):  
## '(Intercept):1'  
##   
##   
## Exponentiated coefficients:  
## sex trestbps cp exang oldpeak thalach   
## 2.837905e-05 9.674346e-01 2.720614e-01 4.914350e-01 4.059204e-01 1.007053e+00   
## sex:trestbps sex:cp sex:exang sex:oldpeak sex:thalach   
## 1.022664e+00 1.990639e+00 1.234762e+00 1.586109e+00 1.021173e+00

## [1] 453.9347

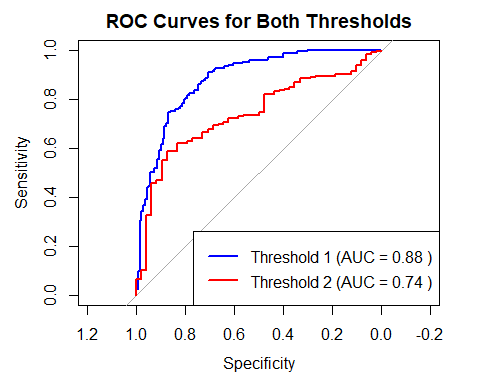
The p-values for all the interaction terms are greater than 0.05. Thus, the final model has no interaction terms and has the following terms: trestbps (resting blood pressure), sex, cp (chest pain type), exang (whether there was exercise induced angina), oldpeak (ST depression induced by exercise relative to rest), and thalach (maximum heart rate achieved during exercise).

The next step is to see how well this final model does at explaining and predicting the heart disease outcome within the data. The first step for that will be to look at the standardized residuals.



It’s important to note that this part of the investigation has two thresholds that will be checked with the model: the cases that had no heart disease compared to if they had a mild or severe case (threshold 1) and the cases that had no heart disease and a mild case of heart disease compared to a severe case of heart disease (threshold 2). What’s found is that the first threshold has some standardized residuals that are above the 2 and below the -2 mark but not many. This means that those residuals are well distributed around 0, meaning the model does a good job of differentiating between no disease and having a disease. For threshold 2, however, more outliers are found compared to the first threshold. This indicates the model has a tougher time distinguishing between the severity of the heart disease.

Next, the ROC curve and AUC (Area Under the Curve) will be looked at regarding these two thresholds.



What is found is the AUC is rather high for the first threshold and lower for the second. This indicates that the model does an effective job of detecting if a patient has heart disease; however, once again, the model struggles to detect how severe it is. This confirms what was found with the residuals as well.

# Results and Discussions

The model does a good job of predicting the presence of heart disease but not so much the severity of it. The biggest reason why that could be is because of the number of cases with or without heart disease. A quick look at the data will show that there is a large amount of cases with no heart disease, a decent size with a mild case, and a small amount with a severe case (at least when compared to the number of cases with no heart disease and a mild case of heart disease).

##   
## 0 1 2   
## 164 91 48

Since the number of cases with no heart disease compared to cases with heart disease (both mild and severe cases combined) are about the same, that would explain why the model is more effective at predicting the presence of heart disease compared to predicting the severity of it. Having data with more cases of heart disease (both mild and severe) would greatly improve the model’s effectiveness at predicting the severity of the heart disease. Overall, the model does a good job but it can be improved.

# Conclusions

The goal for this report was to find the variables that did the best job of explaining not only the presence of heart disease, but also the severity of it. In our final model, the following variables were used:

##   
## Call:  
## vglm(formula = num3 ~ trestbps + sex + cp + exang + oldpeak +   
## thalach, family = cumulative(parallel = TRUE), data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 3.692286 1.550675 2.381 0.017262 \*   
## (Intercept):2 6.058450 1.576823 3.842 0.000122 \*\*\*  
## trestbps -0.019667 0.007425 -2.649 0.008079 \*\*   
## sex -1.331476 0.308502 -4.316 1.59e-05 \*\*\*  
## cp -0.736869 0.160701 -4.585 4.53e-06 \*\*\*  
## exang -0.642560 0.289512 -2.219 0.026455 \*   
## oldpeak -0.556168 0.117886 -4.718 2.38e-06 \*\*\*  
## thalach 0.020986 0.006228 3.370 0.000753 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 437.6474 on 598 degrees of freedom  
##   
## Log-likelihood: -218.8237 on 598 degrees of freedom  
##   
## Number of Fisher scoring iterations: 6   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## trestbps sex cp exang oldpeak thalach   
## 0.9805255 0.2640871 0.4786103 0.5259442 0.5734023 1.0212075

trestbps: A 1 mmHg increase in resting blood pressure led to a 1.9% decrease in the odds of being in a lower severity category.

sex: Males have a 74% lower odds of being in a lower severity category.

cp: More severe chest pain is associated with a 52% decrease in the odds of being a less severe heart disease.

exang: Exercise-induced angina (a type of chest pain that happens during physical exertion) is associated with a 47% decrease in the odds of lower severity.

oldpeak: Each unit increase in ST depression was associated with a 43% decrease in the odds of being in a lower severity category.

thalach: Higher maximum heart rate was associated with a 2% increase in the odds of lower severity.

In short, higher blood pressure, males, more severe chest pain, excercise-induced angina, increase in ST depression, and a higher maximum heart rate when exercising seem to be linked to having higher odds of heart disease or a more severe disease. More specifically, it seemed like sex, chest pain type, and ST depression were the variables that had the most influence of indicating the presence and severity of heart disease. When seeing how well the model fits the data that was used, the model does a good job of differentiating between patients that have heart disease verses patients that do not. However, the model struggles when predicting how severe the heart disease is. This could be due to a lack of number of cases that have a mild and especially a severe case of heart disease. To improve the model, a good first step would be to increase the amount of cases with a severe and a mild case of heart disease in the data. This would allow the model to be better trained on predicting the severity of the disease. In the end, the variables seem to do a good job of explaining heart disease and our model does a good job of predicting it. However, it can be improved by having more data cases where heart disease is present.

# References

Centers for Disease Control and Prevention. “Heart Disease Facts.” Centers for Disease Control and Prevention, 11 Jan. 2024, <https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html>.

Detrano, Robert, et al. “International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease.” The American Journal of Cardiology, vol. 64, no. 5, 1989, pp. 304–310. <https://doi.org/10.1016/0002-9149(89)90524-9>.

Dua, Dheeru, and Casey Graff. “Heart Disease Data Set.” UCI Machine Learning Repository, University of California, Irvine, 1988, <https://archive.ics.uci.edu/dataset/45/heart+disease>.

Dua, Dheeru, and Casey Graff. “Heart Disease Data Set.” UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, 2019, <https://archive.ics.uci.edu/ml/datasets/heart+Disease>.

Mayo Clinic Staff. “Heart Disease.” Mayo Clinic, Mayo Foundation for Medical Education and Research, 12 Dec. 2024, <https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118>. Accessed 12 May 2025.

World Health Organization. “Cardiovascular Diseases.” World Health Organization, 2023, <https://www.who.int/health-topics/cardiovascular-diseases>.

# Appendix

#Chad Huntebrinker  
#Final Project  
library(readxl)  
library(VGAM)  
  
#First, load the dataset  
data\_set <- read\_excel("Heart\_Disease\_data.xlsx")  
colSums(is.na(data\_set))

## age sex cp trestbps chol fbs restecg thalach   
## 0 0 0 0 0 0 0 0   
## exang oldpeak slope ca thal num   
## 0 0 0 0 0 0

#Data was pretty spread out, so we needed to break them down into 3 categories instead of 5.  
data\_set$num3 <- cut(data\_set$num,  
 breaks = c(-1, 0, 2, 4),  
 labels = c(0, 1, 2),  
 ordered\_result = TRUE)  
  
  
#Variables that will be included due to the CDC: trestbps, chol, age, sex  
  
#Add the baseline predictors  
vglm\_model\_1 <- vglm(num3 ~ trestbps + chol + age + sex,   
 family = cumulative(parallel = TRUE), data = data\_set)  
summary(vglm\_model\_1)

##   
## Call:  
## vglm(formula = num3 ~ trestbps + chol + age + sex, family = cumulative(parallel = TRUE),   
## data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 6.687973 1.238216 5.401 6.62e-08 \*\*\*  
## (Intercept):2 8.372008 1.272945 6.577 4.80e-11 \*\*\*  
## trestbps -0.013619 0.006822 -1.996 0.045904 \*   
## chol -0.003787 0.002336 -1.621 0.104954   
## age -0.049714 0.014249 -3.489 0.000485 \*\*\*  
## sex -1.497985 0.288274 -5.196 2.03e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 547.7772 on 600 degrees of freedom  
##   
## Log-likelihood: -273.8886 on 600 degrees of freedom  
##   
## Number of Fisher scoring iterations: 5   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## trestbps chol age sex   
## 0.9864729 0.9962204 0.9515014 0.2235802

#Needed to remove chol because the p-value  
vglm\_model\_2 <- vglm(num3 ~ trestbps + age + sex,   
 family = cumulative(parallel = TRUE), data = data\_set)  
summary(vglm\_model\_2)

##   
## Call:  
## vglm(formula = num3 ~ trestbps + age + sex, family = cumulative(parallel = TRUE),   
## data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 5.914384 1.117084 5.294 1.19e-07 \*\*\*  
## (Intercept):2 7.588924 1.152030 6.587 4.47e-11 \*\*\*  
## trestbps -0.014393 0.006793 -2.119 0.034110 \*   
## age -0.052237 0.014056 -3.717 0.000202 \*\*\*  
## sex -1.397762 0.277873 -5.030 4.90e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 550.2774 on 601 degrees of freedom  
##   
## Log-likelihood: -275.1387 on 601 degrees of freedom  
##   
## Number of Fisher scoring iterations: 5   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## trestbps age sex   
## 0.9857102 0.9491034 0.2471494

lrtest(vglm\_model\_1, vglm\_model\_2)

## Likelihood ratio test  
##   
## Model 1: num3 ~ trestbps + chol + age + sex  
## Model 2: num3 ~ trestbps + age + sex  
## #Df LogLik Df Chisq Pr(>Chisq)  
## 1 600 -273.89   
## 2 601 -275.14 1 2.5002 0.1138

#Try adding the other variables  
  
model\_cp <- vglm(num3 ~ trestbps + age + sex + cp,   
 family = cumulative(parallel = TRUE),   
 data = data\_set)  
  
model\_fbs <- vglm(num3 ~ trestbps + age + sex + fbs,   
 family = cumulative(parallel = TRUE),   
 data = data\_set)  
  
model\_restecg <- vglm(num3 ~ trestbps + age + sex + restecg,   
 family = cumulative(parallel = TRUE),   
 data = data\_set)  
  
model\_thalach <- vglm(num3 ~ trestbps + age + sex + thalach,   
 family = cumulative(parallel = TRUE),   
 data = data\_set)  
  
model\_exang <- vglm(num3 ~ trestbps + age + sex + exang,   
 family = cumulative(parallel = TRUE),   
 data = data\_set)  
  
model\_oldpeak <- vglm(num3 ~ trestbps + age + sex + oldpeak,   
 family = cumulative(parallel = TRUE),   
 data = data\_set)  
  
model\_slope <- vglm(num3 ~ trestbps + age + sex + slope,   
 family = cumulative(parallel = TRUE),   
 data = data\_set)  
  
model\_ca <- vglm(num3 ~ trestbps + age + sex + ca,   
 family = cumulative(parallel = TRUE),   
 data = data\_set)  
  
model\_thal <- vglm(num3 ~ trestbps + age + sex + thal,   
 family = cumulative(parallel = TRUE),   
 data = data\_set)  
  
summary(model\_cp)

##   
## Call:  
## vglm(formula = num3 ~ trestbps + age + sex + cp, family = cumulative(parallel = TRUE),   
## data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 10.564362 1.408571 7.500 6.38e-14 \*\*\*  
## (Intercept):2 12.543959 1.468502 8.542 < 2e-16 \*\*\*  
## trestbps -0.022634 0.007204 -3.142 0.00168 \*\*   
## age -0.048486 0.015180 -3.194 0.00140 \*\*   
## sex -1.588159 0.300337 -5.288 1.24e-07 \*\*\*  
## cp -1.123425 0.159965 -7.023 2.17e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 485.9403 on 600 degrees of freedom  
##   
## Log-likelihood: -242.9701 on 600 degrees of freedom  
##   
## Number of Fisher scoring iterations: 5   
##   
## Warning: Hauck-Donner effect detected in the following estimate(s):  
## '(Intercept):1'  
##   
##   
## Exponentiated coefficients:  
## trestbps age sex cp   
## 0.9776198 0.9526707 0.2043013 0.3251641

summary(model\_fbs) # No good

##   
## Call:  
## vglm(formula = num3 ~ trestbps + age + sex + fbs, family = cumulative(parallel = TRUE),   
## data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 5.937851 1.127675 5.266 1.40e-07 \*\*\*  
## (Intercept):2 7.612889 1.162466 6.549 5.80e-11 \*\*\*  
## trestbps -0.014554 0.006878 -2.116 0.034336 \*   
## age -0.052431 0.014083 -3.723 0.000197 \*\*\*  
## sex -1.401347 0.278455 -5.033 4.84e-07 \*\*\*  
## fbs 0.074000 0.322585 0.229 0.818562   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 550.2255 on 600 degrees of freedom  
##   
## Log-likelihood: -275.1128 on 600 degrees of freedom  
##   
## Number of Fisher scoring iterations: 5   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## trestbps age sex fbs   
## 0.9855515 0.9489194 0.2462650 1.0768063

summary(model\_restecg) # No good

##   
## Call:  
## vglm(formula = num3 ~ trestbps + age + sex + restecg, family = cumulative(parallel = TRUE),   
## data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 5.817240 1.123110 5.180 2.22e-07 \*\*\*  
## (Intercept):2 7.522474 1.158756 6.492 8.48e-11 \*\*\*  
## trestbps -0.012881 0.006866 -1.876 0.060650 .   
## age -0.049027 0.014145 -3.466 0.000528 \*\*\*  
## sex -1.380459 0.278104 -4.964 6.91e-07 \*\*\*  
## restecg -0.295713 0.118927 -2.487 0.012901 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 544.0201 on 600 degrees of freedom  
##   
## Log-likelihood: -272.01 on 600 degrees of freedom  
##   
## Number of Fisher scoring iterations: 5   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## trestbps age sex restecg   
## 0.9872016 0.9521556 0.2514630 0.7440007

summary(model\_slope)

##   
## Call:  
## vglm(formula = num3 ~ trestbps + age + sex + slope, family = cumulative(parallel = TRUE),   
## data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 7.148222 1.180452 6.055 1.40e-09 \*\*\*  
## (Intercept):2 8.985927 1.226492 7.327 2.36e-13 \*\*\*  
## trestbps -0.013092 0.006999 -1.870 0.06142 .   
## age -0.043604 0.014487 -3.010 0.00261 \*\*   
## sex -1.410888 0.285378 -4.944 7.66e-07 \*\*\*  
## slope -1.171424 0.201749 -5.806 6.39e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 515.6302 on 600 degrees of freedom  
##   
## Log-likelihood: -257.8151 on 600 degrees of freedom  
##   
## Number of Fisher scoring iterations: 5   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## trestbps age sex slope   
## 0.9869936 0.9573330 0.2439266 0.3099253

summary(model\_ca) # No good

##   
## Call:  
## vglm(formula = num3 ~ trestbps + age + sex + ca, family = cumulative(parallel = TRUE),   
## data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 5.188758 1.636433 3.171 0.00152 \*\*   
## (Intercept):2 7.236555 1.664548 4.347 1.38e-05 \*\*\*  
## trestbps -0.019352 0.007235 -2.675 0.00748 \*\*   
## age -0.004495 0.015899 -0.283 0.77741   
## sex -1.244287 0.295877 -4.205 2.61e-05 \*\*\*  
## ca0 -0.526627 1.180916 -0.446 0.65564   
## ca1 -2.054376 1.202703 -1.708 0.08761 .   
## ca2 -3.097776 1.229537 -2.519 0.01175 \*   
## ca3 -3.245275 1.265527 -2.564 0.01034 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 485.5847 on 597 degrees of freedom  
##   
## Log-likelihood: -242.7924 on 597 degrees of freedom  
##   
## Number of Fisher scoring iterations: 5   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## trestbps age sex ca0 ca1 ca2 ca3   
## 0.98083410 0.99551550 0.28814640 0.59059392 0.12817274 0.04514952 0.03895784

summary(model\_exang)

##   
## Call:  
## vglm(formula = num3 ~ trestbps + age + sex + exang, family = cumulative(parallel = TRUE),   
## data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 6.143780 1.171495 5.244 1.57e-07 \*\*\*  
## (Intercept):2 8.057284 1.212776 6.644 3.06e-11 \*\*\*  
## trestbps -0.013611 0.006993 -1.946 0.051604 .   
## age -0.048747 0.014671 -3.323 0.000892 \*\*\*  
## sex -1.317200 0.292240 -4.507 6.57e-06 \*\*\*  
## exang -1.692524 0.257438 -6.575 4.88e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 503.7472 on 600 degrees of freedom  
##   
## Log-likelihood: -251.8736 on 600 degrees of freedom  
##   
## Number of Fisher scoring iterations: 5   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## trestbps age sex exang   
## 0.9864809 0.9524222 0.2678844 0.1840544

summary(model\_oldpeak)

##   
## Call:  
## vglm(formula = num3 ~ trestbps + age + sex + oldpeak, family = cumulative(parallel = TRUE),   
## data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 5.171688 1.161633 4.452 8.50e-06 \*\*\*  
## (Intercept):2 7.137776 1.198841 5.954 2.62e-09 \*\*\*  
## trestbps -0.009633 0.007121 -1.353 0.1761   
## age -0.037329 0.014569 -2.562 0.0104 \*   
## sex -1.277035 0.290112 -4.402 1.07e-05 \*\*\*  
## oldpeak -0.779293 0.113477 -6.867 6.54e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 497.8045 on 600 degrees of freedom  
##   
## Log-likelihood: -248.9023 on 600 degrees of freedom  
##   
## Number of Fisher scoring iterations: 5   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## trestbps age sex oldpeak   
## 0.9904134 0.9633589 0.2788630 0.4587304

summary(model\_thalach)

##   
## Call:  
## vglm(formula = num3 ~ trestbps + age + sex + thalach, family = cumulative(parallel = TRUE),   
## data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 -0.986777 1.534157 -0.643 0.52009   
## (Intercept):2 0.919490 1.530786 0.601 0.54806   
## trestbps -0.019366 0.007066 -2.741 0.00613 \*\*   
## age -0.016294 0.015381 -1.059 0.28942   
## sex -1.390561 0.288487 -4.820 1.43e-06 \*\*\*  
## thalach 0.037427 0.005966 6.273 3.54e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 506.4035 on 600 degrees of freedom  
##   
## Log-likelihood: -253.2017 on 600 degrees of freedom  
##   
## Number of Fisher scoring iterations: 5   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## trestbps age sex thalach   
## 0.9808201 0.9838379 0.2489355 1.0381363

summary(model\_thal) # No good

##   
## Call:  
## vglm(formula = num3 ~ trestbps + age + sex + thal, family = cumulative(parallel = TRUE),   
## data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 4.198503 1.783877 2.354 0.018593 \*   
## (Intercept):2 6.171102 1.807267 3.415 0.000639 \*\*\*  
## trestbps -0.010663 0.007157 -1.490 0.136261   
## age -0.043260 0.015008 -2.882 0.003946 \*\*   
## sex -0.611653 0.310863 -1.968 0.049114 \*   
## thal3 1.120493 1.407632 0.796 0.426024   
## thal6 -0.314640 1.471808 -0.214 0.830720   
## thal7 -1.009809 1.411773 -0.715 0.474438   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 489.5091 on 598 degrees of freedom  
##   
## Log-likelihood: -244.7546 on 598 degrees of freedom  
##   
## Number of Fisher scoring iterations: 5   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## trestbps age sex thal3 thal6 thal7   
## 0.9893934 0.9576622 0.5424534 3.0663654 0.7300514 0.3642886

#Model with fbs, restecg, ca, and thal are no good but the other ones seem like good predictors.  
#Check the AICs of all of them  
AIC(vglm\_model\_1)

## [1] 559.7772

AIC(vglm\_model\_2)

## [1] 560.2774

AIC(model\_cp)

## [1] 497.9403

AIC(model\_slope) #Doesn't have a huge improvement

## [1] 527.6302

AIC(model\_exang)

## [1] 515.7472

AIC(model\_oldpeak)

## [1] 509.8045

AIC(model\_thalach)

## [1] 518.4035

#Slope doesn't have a huge drop in AIC, so we're on the fence with it.  
#For the sake of keeping the model simple, we'll keep it off the model.  
  
#But the rest of the models improve when we add the new predictor variables.  
#So we will try adding them in.  
  
vglm\_model\_3 <- vglm(num3 ~ trestbps + age + sex + cp + exang + oldpeak + thalach,  
 family = cumulative(parallel = TRUE),  
 data = data\_set)  
summary(vglm\_model\_3)

##   
## Call:  
## vglm(formula = num3 ~ trestbps + age + sex + cp + exang + oldpeak +   
## thalach, family = cumulative(parallel = TRUE), data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 5.224128 1.894627 2.757 0.00583 \*\*   
## (Intercept):2 7.593980 1.921530 3.952 7.75e-05 \*\*\*  
## trestbps -0.017453 0.007584 -2.301 0.02137 \*   
## age -0.023285 0.016472 -1.414 0.15747   
## sex -1.381967 0.312185 -4.427 9.57e-06 \*\*\*  
## cp -0.750086 0.161357 -4.649 3.34e-06 \*\*\*  
## exang -0.677746 0.291341 -2.326 0.02000 \*   
## oldpeak -0.545514 0.117966 -4.624 3.76e-06 \*\*\*  
## thalach 0.017881 0.006622 2.700 0.00693 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 435.6682 on 597 degrees of freedom  
##   
## Log-likelihood: -217.8341 on 597 degrees of freedom  
##   
## Number of Fisher scoring iterations: 6   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## trestbps age sex cp exang oldpeak thalach   
## 0.9826979 0.9769840 0.2510842 0.4723259 0.5077604 0.5795439 1.0180421

#The p-value for age was greater than 0.05, so let's try removing it from the model.  
  
  
vglm\_model\_4 <- vglm(num3 ~ trestbps + sex + cp + exang + oldpeak + thalach,  
 family = cumulative(parallel = TRUE),  
 data = data\_set)  
summary(vglm\_model\_4)

##   
## Call:  
## vglm(formula = num3 ~ trestbps + sex + cp + exang + oldpeak +   
## thalach, family = cumulative(parallel = TRUE), data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 3.692286 1.550675 2.381 0.017262 \*   
## (Intercept):2 6.058450 1.576823 3.842 0.000122 \*\*\*  
## trestbps -0.019667 0.007425 -2.649 0.008079 \*\*   
## sex -1.331476 0.308502 -4.316 1.59e-05 \*\*\*  
## cp -0.736869 0.160701 -4.585 4.53e-06 \*\*\*  
## exang -0.642560 0.289512 -2.219 0.026455 \*   
## oldpeak -0.556168 0.117886 -4.718 2.38e-06 \*\*\*  
## thalach 0.020986 0.006228 3.370 0.000753 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 437.6474 on 598 degrees of freedom  
##   
## Log-likelihood: -218.8237 on 598 degrees of freedom  
##   
## Number of Fisher scoring iterations: 6   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## trestbps sex cp exang oldpeak thalach   
## 0.9805255 0.2640871 0.4786103 0.5259442 0.5734023 1.0212075

AIC(vglm\_model\_3)

## [1] 453.6682

AIC(vglm\_model\_4)

## [1] 453.6474

#The AICs are basically the same, but model 4 (without age) is simpler so we'll keep that  
  
#Next, we'll test interaction between predictors in model\_4.  
#Specifically, the sex of the individual with it:  
vglm\_model\_interaction <- vglm(num3 ~ sex \* trestbps + sex \* cp + sex \* exang + sex \* oldpeak + sex \* thalach,  
 family = cumulative(parallel = TRUE),  
 data = data\_set)  
  
summary(vglm\_model\_interaction)

##   
## Call:  
## vglm(formula = num3 ~ sex \* trestbps + sex \* cp + sex \* exang +   
## sex \* oldpeak + sex \* thalach, family = cumulative(parallel = TRUE),   
## data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 10.003598 3.560366 NA NA   
## (Intercept):2 12.410797 3.587145 3.460 0.000541 \*\*\*  
## sex -10.469859 4.010032 -2.611 0.009030 \*\*   
## trestbps -0.033107 0.016058 -2.062 0.039229 \*   
## cp -1.301728 0.501002 -2.598 0.009370 \*\*   
## exang -0.710426 0.652256 -1.089 0.276073   
## oldpeak -0.901598 0.270490 -3.333 0.000859 \*\*\*  
## thalach 0.007028 0.015393 0.457 0.647990   
## sex:trestbps 0.022411 0.018289 1.225 0.220434   
## sex:cp 0.688456 0.528328 1.303 0.192546   
## sex:exang 0.210878 0.733705 0.287 0.773794   
## sex:oldpeak 0.461284 0.298940 1.543 0.122815   
## sex:thalach 0.020952 0.017036 1.230 0.218739   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 427.9347 on 593 degrees of freedom  
##   
## Log-likelihood: -213.9673 on 593 degrees of freedom  
##   
## Number of Fisher scoring iterations: 6   
##   
## Warning: Hauck-Donner effect detected in the following estimate(s):  
## '(Intercept):1'  
##   
##   
## Exponentiated coefficients:  
## sex trestbps cp exang oldpeak thalach   
## 2.837905e-05 9.674346e-01 2.720614e-01 4.914350e-01 4.059204e-01 1.007053e+00   
## sex:trestbps sex:cp sex:exang sex:oldpeak sex:thalach   
## 1.022664e+00 1.990639e+00 1.234762e+00 1.586109e+00 1.021173e+00

AIC(vglm\_model\_interaction)

## [1] 453.9347

#The AIC is basically the same and the p-values disqualify them. So no interaction between sex and the other  
#variables.  
  
#Next we'll move onto how well the model fits and explains the data.  
  
#Chad Huntebrinker  
#Final Project  
library(readxl)  
library(VGAM)  
library(pROC)  
  
#First, load the dataset  
data\_set <- read\_excel("Heart\_Disease\_data.xlsx")  
  
#Data was pretty spread out, so we needed to break them down into 3 categories instead of 5.  
data\_set$num3 <- cut(data\_set$num,  
 breaks = c(-1, 0, 2, 4),  
 labels = c(0, 1, 2),  
 ordered\_result = TRUE)  
  
#Use the model we found last script  
vglm\_final\_model <- vglm(num3 ~ trestbps + sex + cp + exang + oldpeak + thalach,  
 family = cumulative(parallel = TRUE),  
 data = data\_set)  
  
#And we will also try parallel = FALSE to cover our bases  
vglm\_final\_model\_nonparallel <- vglm(num3 ~ trestbps + sex + cp + exang + oldpeak + thalach,  
 family = cumulative(parallel = TRUE),  
 data = data\_set)  
summary(vglm\_final\_model)

##   
## Call:  
## vglm(formula = num3 ~ trestbps + sex + cp + exang + oldpeak +   
## thalach, family = cumulative(parallel = TRUE), data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 3.692286 1.550675 2.381 0.017262 \*   
## (Intercept):2 6.058450 1.576823 3.842 0.000122 \*\*\*  
## trestbps -0.019667 0.007425 -2.649 0.008079 \*\*   
## sex -1.331476 0.308502 -4.316 1.59e-05 \*\*\*  
## cp -0.736869 0.160701 -4.585 4.53e-06 \*\*\*  
## exang -0.642560 0.289512 -2.219 0.026455 \*   
## oldpeak -0.556168 0.117886 -4.718 2.38e-06 \*\*\*  
## thalach 0.020986 0.006228 3.370 0.000753 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 437.6474 on 598 degrees of freedom  
##   
## Log-likelihood: -218.8237 on 598 degrees of freedom  
##   
## Number of Fisher scoring iterations: 6   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## trestbps sex cp exang oldpeak thalach   
## 0.9805255 0.2640871 0.4786103 0.5259442 0.5734023 1.0212075

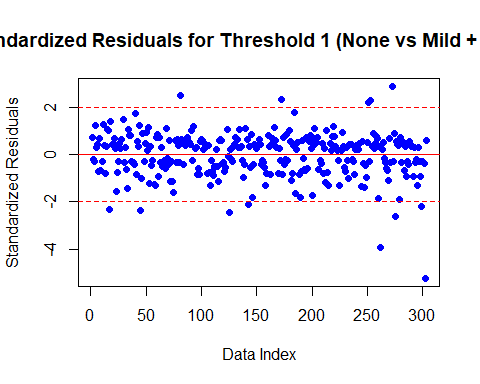
summary(vglm\_final\_model\_nonparallel)

##   
## Call:  
## vglm(formula = num3 ~ trestbps + sex + cp + exang + oldpeak +   
## thalach, family = cumulative(parallel = TRUE), data = data\_set)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 3.692286 1.550675 2.381 0.017262 \*   
## (Intercept):2 6.058450 1.576823 3.842 0.000122 \*\*\*  
## trestbps -0.019667 0.007425 -2.649 0.008079 \*\*   
## sex -1.331476 0.308502 -4.316 1.59e-05 \*\*\*  
## cp -0.736869 0.160701 -4.585 4.53e-06 \*\*\*  
## exang -0.642560 0.289512 -2.219 0.026455 \*   
## oldpeak -0.556168 0.117886 -4.718 2.38e-06 \*\*\*  
## thalach 0.020986 0.006228 3.370 0.000753 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 437.6474 on 598 degrees of freedom  
##   
## Log-likelihood: -218.8237 on 598 degrees of freedom  
##   
## Number of Fisher scoring iterations: 6   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## trestbps sex cp exang oldpeak thalach   
## 0.9805255 0.2640871 0.4786103 0.5259442 0.5734023 1.0212075

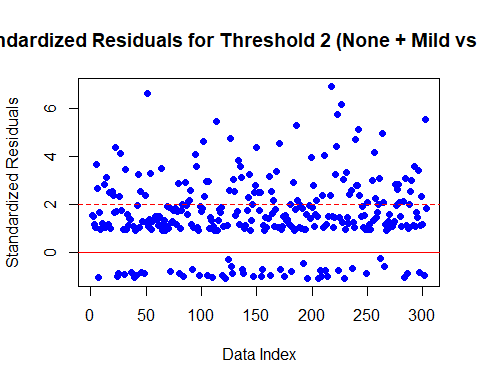
#Since the non\_parallel p\_values are most all greater than 0.05, we will stick with parallel one.  
#Parallel = TRUE means that we’re assuming that the odds ratios are consistent across severity thresholds.  
  
AIC(vglm\_final\_model)

## [1] 453.6474

#Now we will check how well the model fits to the data  
  
#First, let's check on some of the residuals  
#Calculate standardized residuals  
  
pred\_prob <- predict(vglm\_final\_model, type = "response")  
  
#Define observed outcomes for each threshold  
threshold\_1 <- data\_set$num3 < 1 # "None" vs "Mild + Severe"  
threshold\_2 <- data\_set$num3 < 2 # "None + Mild" vs "Severe"  
  
#Calculate response residuals for each threshold  
residuals\_threshold\_1 <- threshold\_1 - pred\_prob[, 1]  
residuals\_threshold\_2 <- threshold\_2 - pred\_prob[, 2]  
  
#Standardized the residuals  
std\_residuals\_threshold\_1 <- residuals\_threshold\_1 / sqrt(pred\_prob[, 1] \* (1 - pred\_prob[, 1]))  
std\_residuals\_threshold\_2 <- residuals\_threshold\_2 / sqrt(pred\_prob[, 2] \* (1 - pred\_prob[, 2]))  
  
#Plot for threshold 1  
plot(std\_residuals\_threshold\_1,   
 main = "Standardized Residuals for Threshold 1 (None vs Mild + Severe)",   
 ylab = "Standardized Residuals",   
 xlab = "Data Index",   
 pch = 16, col = "blue")  
abline(h = 0, col = "red")  
abline(h = c(-2, 2), col = "red", lty = 2)



#Plot for threshold 2  
plot(std\_residuals\_threshold\_2,   
 main = "Standardized Residuals for Threshold 2 (None + Mild vs Severe)",   
 ylab = "Standardized Residuals",   
 xlab = "Data Index",   
 pch = 16, col = "blue")  
abline(h = 0, col = "red")  
abline(h = c(-2, 2), col = "red", lty = 2)



#Try to find a similarity between the outlie data points but unable to find it.  
outliers\_SRT2 <- which(abs(std\_residuals\_threshold\_2) > 2)  
outlier\_data\_set <- data\_set[outliers\_SRT2, ]  
  
  
#Let's check the ROC curve and the AUC  
  
#ROC for first threshold  
roc1 <- roc(threshold\_1, pred\_prob[, 1], plot = TRUE, col = "blue",   
 main = "ROC Curves for Both Thresholds", lwd = 2)

## Setting levels: control = FALSE, case = TRUE

## Setting direction: controls < cases

auc1 <- auc(roc1)  
  
#ROC for second threshold  
roc2 <- roc(threshold\_2, pred\_prob[, 2], plot = TRUE, col = "red", lwd = 2, add = TRUE)

## Setting levels: control = FALSE, case = TRUE

## Setting direction: controls > cases

auc2 <- auc(roc2)  
  
#Add legend  
legend("bottomright", legend = c(paste("Threshold 1 (AUC =", round(auc1, 2), ")"),   
 paste("Threshold 2 (AUC =", round(auc2, 2), ")")),   
 col = c("blue", "red"), lwd = 2)

