Logistic Regression Modeling on Clinical Data for Estimation and Prediction related to Heart Disease

Arman Jahangiri
Mojtaba Kanani Sarcheshmeh
David Yang

December 6^{th}, 2023

Outline

Outline

Introduction

- **Dataset**:
  Clinical data related to heart disease diagnosis which includes patients' different attributes.

- **Number of variables**: 14

- **Response variable**: (target)
  Binary variable determining the presence of heart disease in the patient (0 refers to no disease)

- **Sampling method**: Convenience sampling.
  Patients from four separate hospitals during a specific time-frame were surveyed at the convenience of the researcher.

Introduction

- We have used the dataset to fit a generalized linear model in logistic family.
- We have used "Diagnosis of Heart Disease (target)" as the response variable and all other variables as predictor variables.
- The data consists of 5 continuous variables and 8 categorical variables, excluding the binary response variable.

Variable	**Explanation**	**Min**	**Max**
Age	Age of the patient	29	77
Sex	(1 = male, 0 = female)	0	1
cp	Chest Pain Type	0	3
trestbps	Resting Blood Pressure	94	200
chol	Serum Cholesterol	126	564
fbs	Fasting Blood Sugar (= 1 if ≥ 120)	0	1
restecg	Resting Electrocardiographic Results	0	2

Table 1: Description of Variables (Part 1)

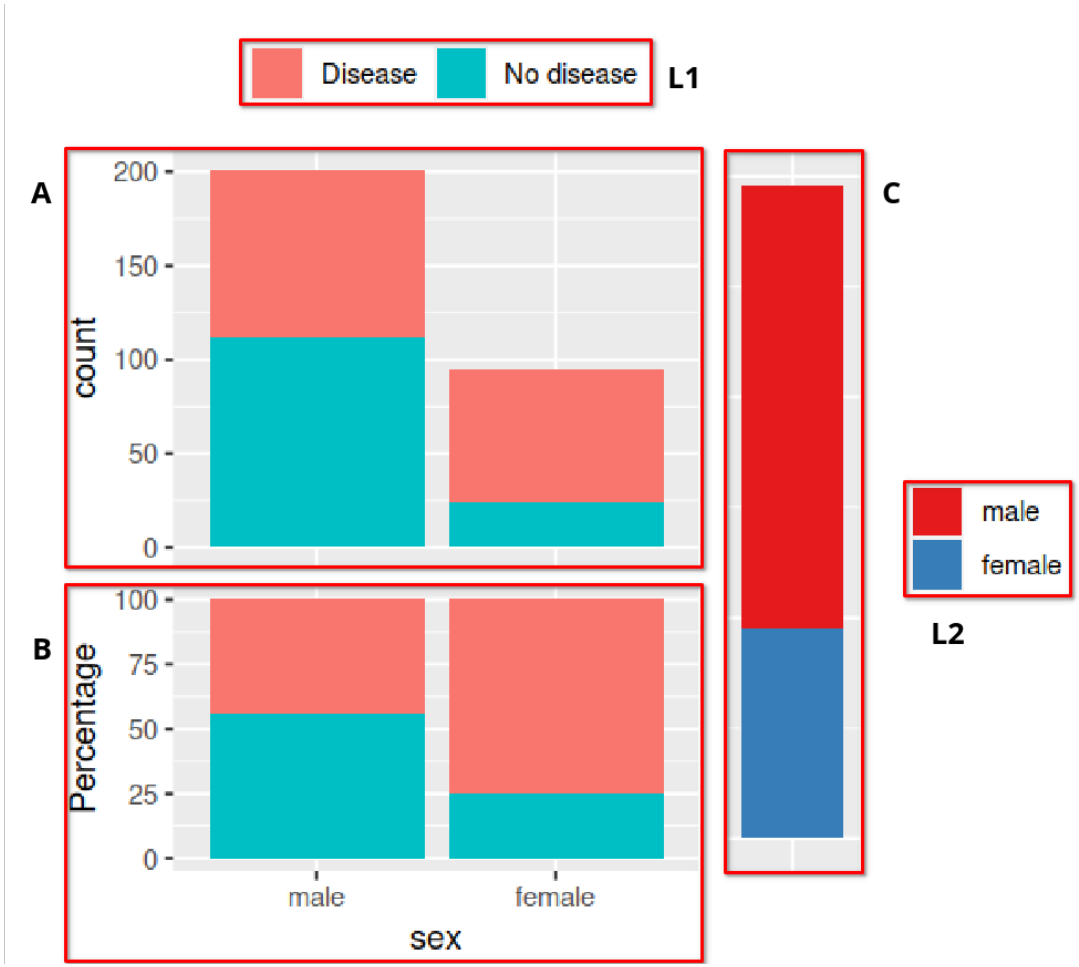
Variable	Explanation	Min	Max
thalach	Maximum Heart Rate Achieved	71	202
exang	Exercise-Induced Angina	0	1
oldpeak	ST Depression Induced by Exercise	0	6.2
slope	Slope of Peak Exercise ST Segment	0	2
ca	Number of Major Vessels	0	4
thal	Thalium Scintigraphy	0	3
target	Diagnosis of Heart Disease	0	1

Table 2: Description of Variables (Part 2)

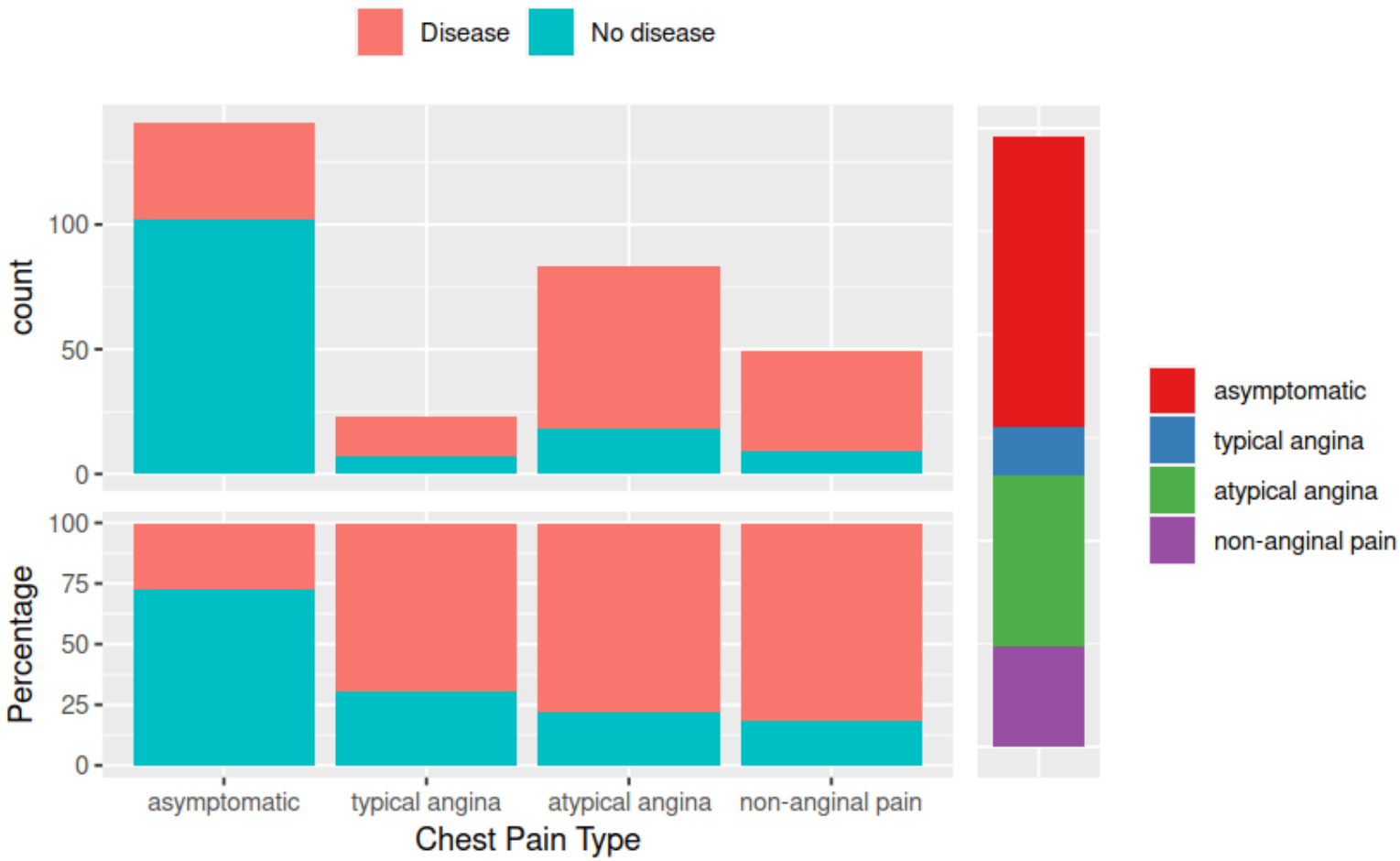
Outline

1. Categorical Variables

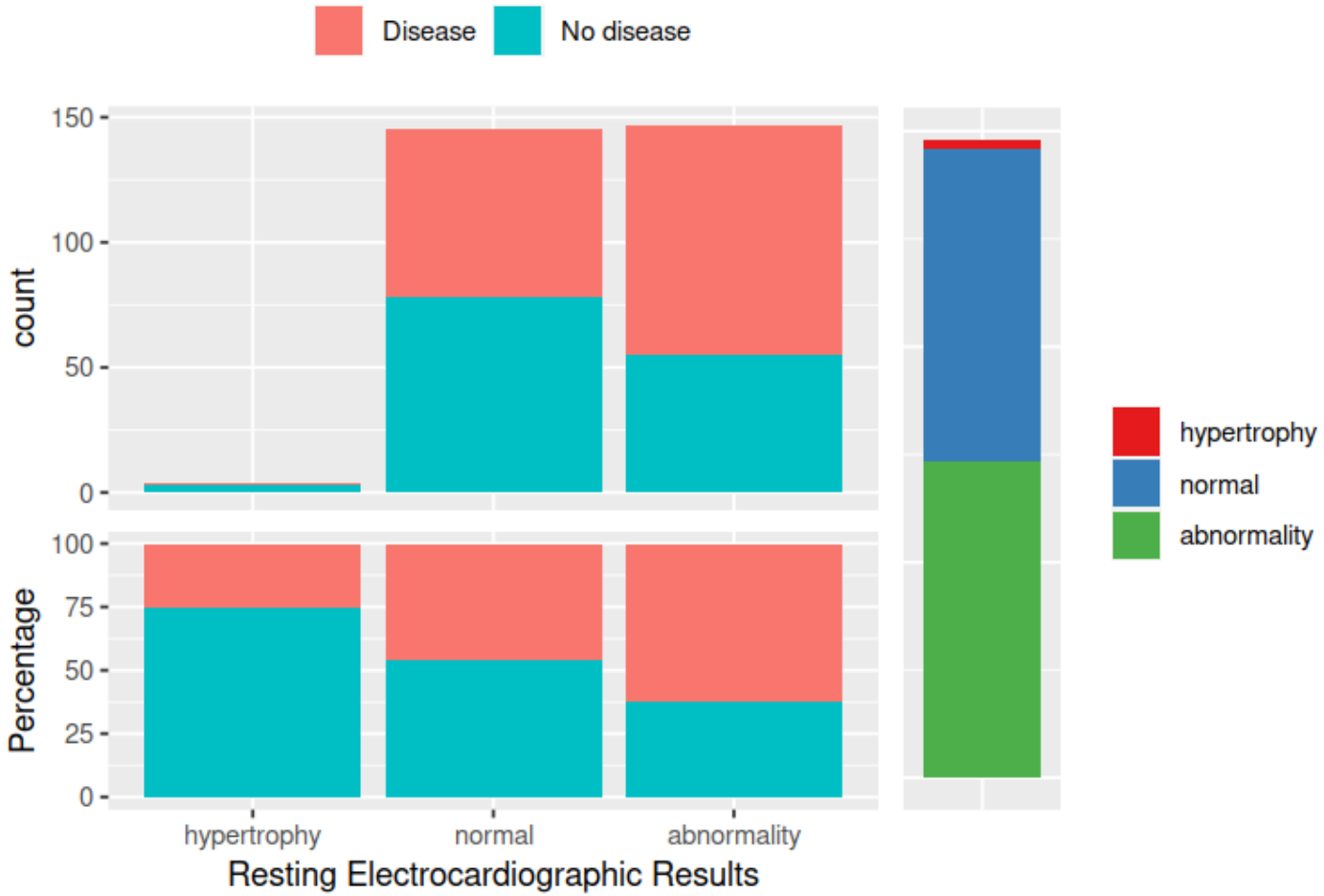
- The following charts provide insight about relationships between our categorical variables and the target variable.
- In figures **A** and **B**, the count and percentage of each of the categories of our predictor is compared to the target class; in part **C** the number of observations in each category is compared to each other. **L1** and **L2** are legends for the plot.



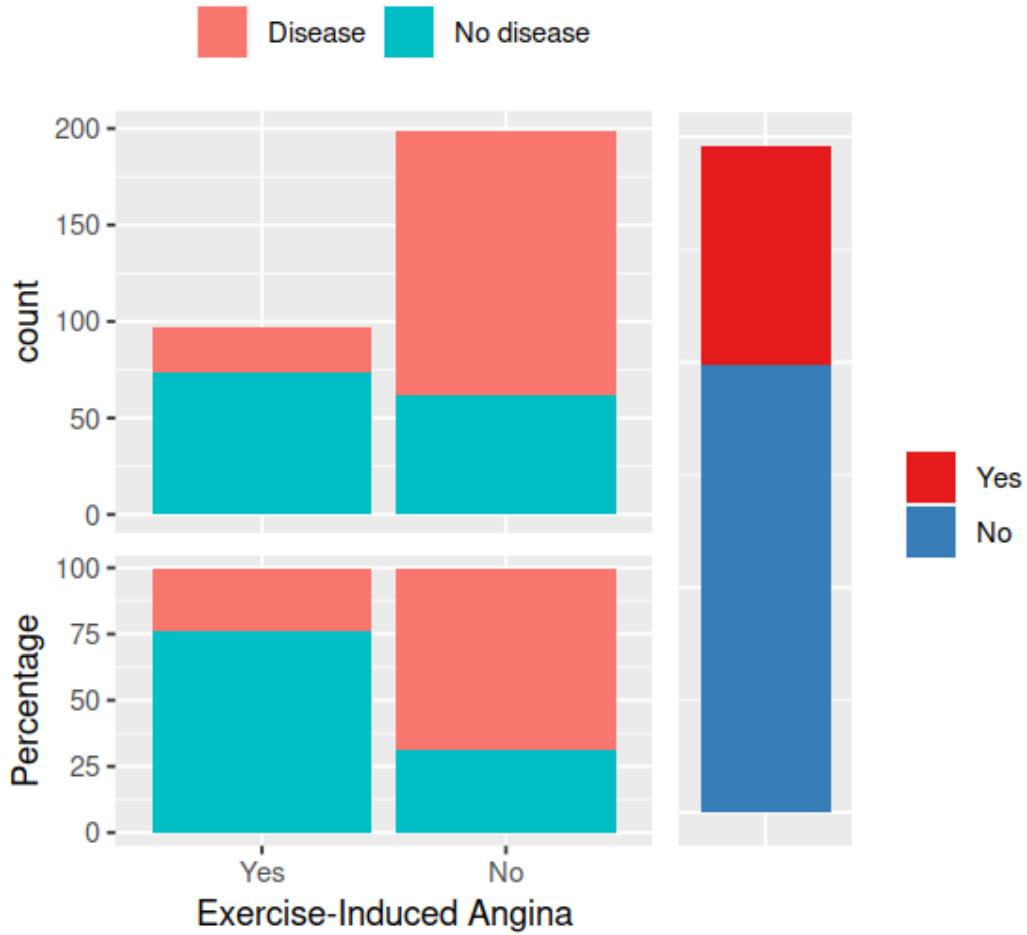
1. Categorical Variables



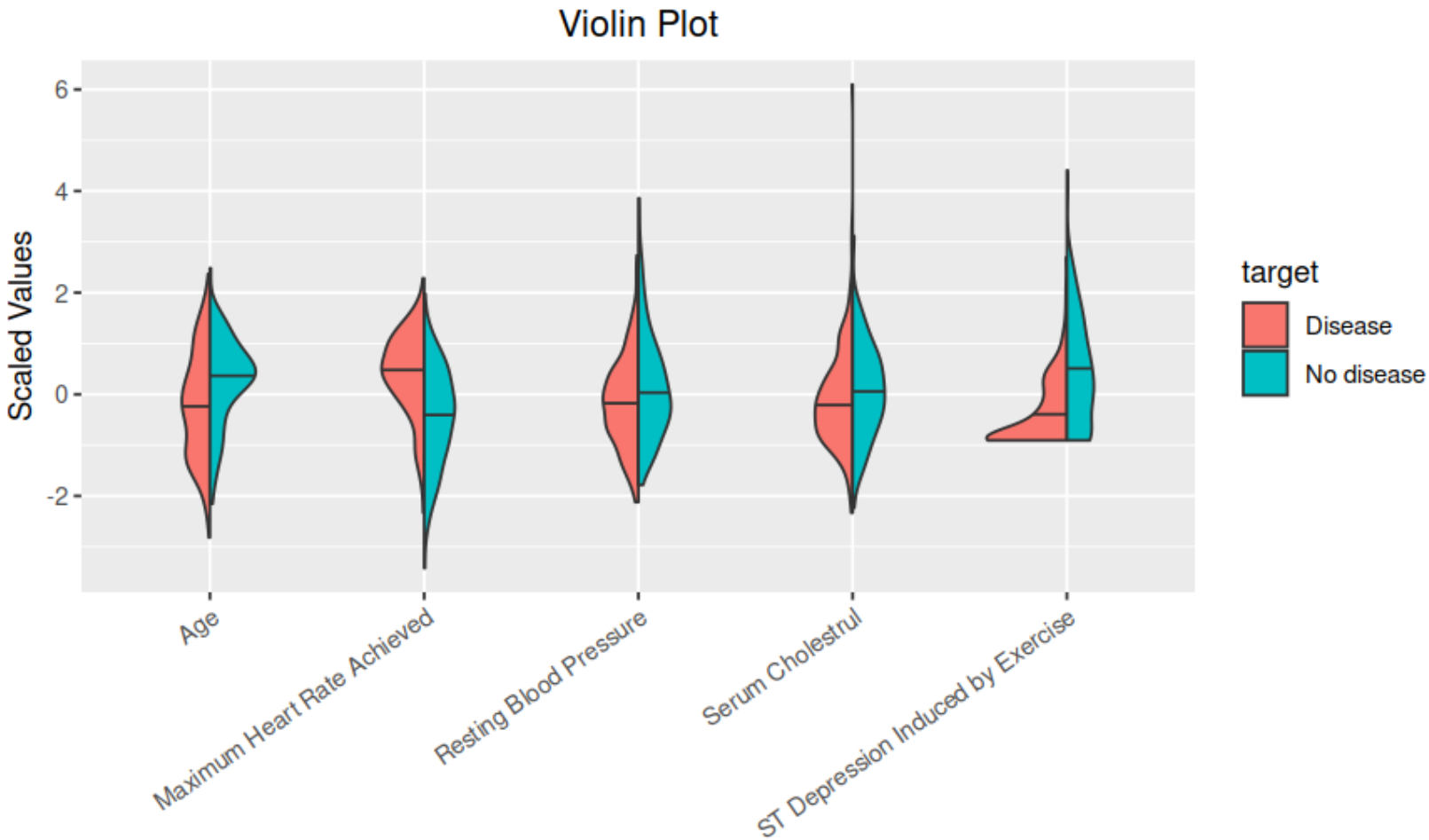
1. Categorical Variables



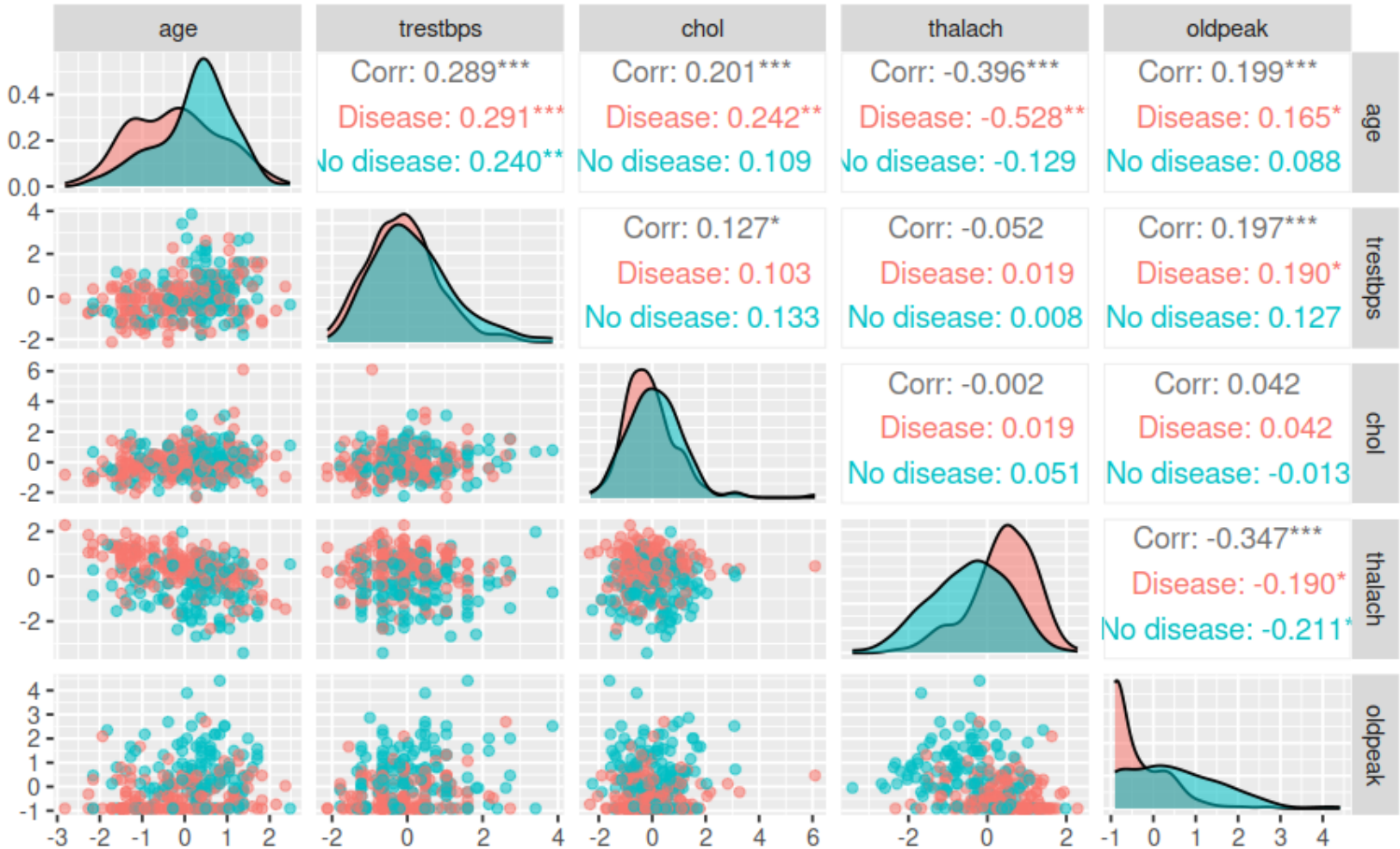
1. Categorical Variables



2. Numerical Variables



3. Pairs Plot - Correlation Matrix for Quantitative Predictors



Outline

Data cleaning

- Categorical variables that were number encoded were renamed with their corresponding descriptive strings so that model coefficients were informative and easily interpretable, specially in the visualization part.
- The NAs and missing values were checked for in the data. In our case, the heart disease prediction dataset contained no NAs or missing values.
- Duplicate entries (723) in our dataset were checked for and removed. We defined duplicates as rows that contained the same exact value for all 14 columns in our dataset.
- Rows containing erroneous entries (25 observations) were removed from our dataset. For example, this included values for categorical variables that fell outside of the grouping for its corresponding variable.
- After data cleaning, our dataset contained **296** observations. ($n = 296$)

Data Standardization

- We scaled all the 6 continuous variables (”age”, ”trestbps”, ”chol”, ”thalach”, ”oldpeak”, ”ca”) using the Z-score standardization technique.
- Standardized variables can contribute to better numerical stability in the computations involved in regression analysis.
- Standardization helps in reducing the impact of multicollinearity.
- When using regularization techniques like Ridge or Lasso regression, scaling becomes importantsince the penalty term is influenced by the scale of the variables.

Outline

Logistic regression (David)

- Logistic regression was used since the response variable ("target") was binary.
- The logit link function was used (the probit model performed slightly worse):

$$\eta = logit(p) = log(\frac{p}{1-p})$$

- We performed three main processes to build logistic regression models for our dataset:
  1. Variable selection with the full model
  2. Variable selection with the reduced model
  3. Variable selection via Lasso regression

(Process 1): Building from the full model (David)

- We started with the full linear model which consists of all 13 predictors $\eta = \sum_{j=1}^{13} \beta_j X_j$

- We found that "ca" had a quadratic relationship with our response via the Wald test which we added to our full model. **(Model 1)**

- From scientific literature, we found that certain variables in our predictors were correlated and hence we added their corresponding interaction terms to the basic full model. **(Model 2)**
  - Females tend to have have lower blood pressure
  - Males tend to have higher levels of resting blood sugar

(Process 1): Other full models with two-way interactions (David)

We explored other methods to decide which two-way interaction terms to include in our model:

- Backward stepwise selection on Full Model with all two-way interaction terms **(Model 3)**
- Forward stepwise selection with the maximum endpoint model as the Full model with all two-way interaction terms
  - Starting from the Model 2. This gave **(Model 4)**.
  - Starting from the intercept model. This gave **(Model 5)**.

(Process 2): Building from the reduced model (David)

- From the basic full model, we used the Wald test to remove insignificant coefficients and added a quadratic term for the predictor "ca" (Wald test) **(Model 6)**.
- We added all possible two-way interactions to Model 6 to generate **Model 7**.
- To explore the effect of the intercept, we removed the intercept from Model 7 to generate **Model 8**.

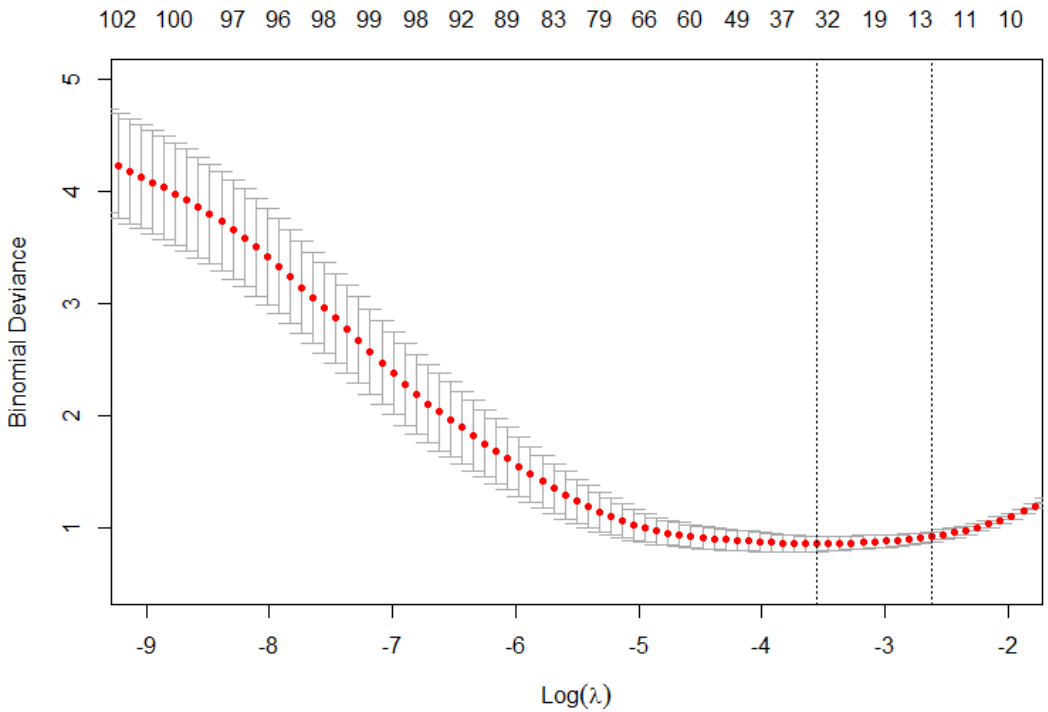
(Process 2): Other reduced models with two-way interactions (David)

- We explored other methods to decide which two-way interaction terms to include in our model.
- Backward stepwise selection on the Reduced model with all possible two-way interaction terms **Model 9**
- Forward stepwise selection on with the maximum endpoint model as the reduced model with all possible two-way interaction terms
  - Starting from Model 6.
  - Starting from the intercept model
  - Both steps generated the **same** model which we named **Model 10**
- From scientific literature we found that "sex" and "trestbps" was correlated and hence we added their interaction term to Model 6 to generate **Model 11**.

(Process 3): Lasso regression

- In this section, we made use of lasso regression technique to perform variable selection on the data with all the 2-way interaction terms and ca^2. **(Model 12)**
- The Cross-Validation method offered $\lambda = 0.01977$ to be optimal, which resulted in 41 variables, and %58 of explained deviance
- The ridge penalty shrinks the coefficients of correlated predictors towards each other while the lasso tends to pick one of them and discard the others. This is why we have preferred LASSO over Ridge to conduct automatic variable selection.

(Process 3): Lasso regression



Over-Dispersion Check

- We checked for existence of over-dispersion, and we found that the dispersion parameter is less than 1.
- Hence the model is not suspect, and we did not use quasi-binomial models.

Outline

Model comparison metrics (David)

The following metrics were investigated for model comparison:

- Residual deviance
- Akaike information criterion (AIC)
- McFadden's Pseudo-R^2
- Likelihood ratio test (comparison to the minimal model)
- Mean-squared prediction error (MSE)
  - Computed via cross-validation with 10 folds for 10 iterations

Selected Models (David)

Model Metrics for Model Comparison

Model	Deviance	AIC	q	R2	lrt_pval	MSE
Model 1	178.21	218.21	19	0.56	3.09e-38	0.12
Model 2	174.19	224.19	24	0.57	2.18e-36	0.12
Model 3	0.00	118.00	58	1.00	3.79e-54	0.22
Model 4	0.00	106.00	52	1.00	8.61e-57	0.18
Model 5	145.44	199.44	26	0.64	4.86e-41	0.12
Model 6	193.34	217.34	11	0.53	5.52e-40	0.11
Model 7	141.91	265.91	62	0.65	3.63e-27	0.18
Model 8	141.91	265.91	61	0.65	8.03e-28	0.18
Model 9	193.34	217.34	11	0.53	5.52e-40	0.11
Model 10	186.49	214.49	13	0.54	4.22e-40	0.11
Model 11	191.77	217.77	12	0.53	1.19e-39	0.11
Model 12	236.62	0.00	41	0.42	6.15e-18	0.66

Outline

Final model: reason for selection

- The final model we selected is **Model 5** (forward stepwise algorithm from intercept model)
- The model contains the following **coefficients** (10): thal, ca, cp, slope, sex, ca^2, oldpeak, trestbps, thalach, chol
- The model contains the following **two-way interactions** (7): thal:ca, slope:trestbps, thal;oldpeak, slope:oldpeak, slope:sex, sex:oldpeak, trestbps:chol

Final model: confidence intervals

Confidence intervals - Part 1

	2.5 %	97.5 %
(Intercept)	-8.2715	-2.3654
factor(thal)normal	-20.1729	215.7082
factor(thal)reversable defect	0.9988	3.1807
ca	1.7209	3.6448
factor(cp)atypical angina	-3.8968	-1.4104
factor(cp)non-anginal pain	-2.6526	-0.0844
factor(cp)typical angina	-5.0925	-1.5796
factor(slope)flat	2.7296	8.6153
factor(slope)upsloping	3.3983	18.2094
factor(sex)male	3.0715	9.0023
I(ca^2)	-1.4479	-0.4077
oldpeak	-6.1414	-1.1812
trestbps	-0.1417	1.1154
thalach	-1.3023	-0.1542
chol	-0.0137	1.0360

Final model: confidence intervals

Confidence intervals - Part 2

	2.5 %	97.5 %
factor(thal)normal:ca	-22.0445	309.4746
factor(thal)reversable defect:ca	-2.1022	-0.1259
factor(slope)flat:trestbps	-0.3480	1.5094
factor(slope)upsloping:trestbps	-6.8918	-1.2133
factor(thal)normal:oldpeak	-2.1616	16.4485
factor(thal)reversable defect:oldpeak	0.5519	3.4193
factor(slope)flat:oldpeak	1.0176	4.5627
factor(slope)upsloping:oldpeak	0.2300	7.1003
factor(slope)flat:factor(sex)male	-7.0118	-1.1682
factor(slope)upsloping:factor(sex)male	-21.2604	-4.5967
factor(sex)male:oldpeak	0.4662	4.3731
trestbps:chol	-0.0837	0.9121

Outline

Conclusion

- In order to double-check the relations found between the variables from the data visualization, we made use of the interaction models to further verify those interactions in model selection techniques (backward, forward, lasso, etc.)
- Based on our analysis of the heart disease dataset, conducted the process of data exploration, preprocessing, and model building. We considered various logistic regression models, including those with interaction terms and exponents to consider the relations of the predictors as well. The models were evaluated using different metrics such as deviance, AIC, pseudo-R^2, likelihood ratio test, and mean-squared prediction error.
- After comparison of these models, we selected the final model based on its performance metrics and the number of variables it had, but one can use other models based on different interests.
- In conclusion, the logistic regression model provides insights into the relationships between various patient attributes and the likelihood of heart disease. The selected model can serve as a tool for understanding and predicting heart disease based on a given new patient. Further research and validation may enhance the robustness of our findings.

Outline

References

[1] Dataset: Heart Disease Dataset (Public Health Dataset)
https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data

[2] Simmons B. (2021). Investigating on Heart Disease Datasets and Building Predictive Models, A Thesis submitted to the Graduate Faculty of Elizabeth City State University

[3] Gareth James, Daniela Witten, An introduction to Statistical Learning with Application R: Spring New York. Wickham and Grollemu

[4] Annette J. Dobson, Adrian G. Barnett (2018), An Introduction to Generalized Linear Models, Fourth Edition: Spring New York. Wickham and Grollemu; CRC Press

Thank You for your attention!

Questions...