Data Analysis Exam 3

Keltin Grimes

12/6/2021

Introduction

This report examines a dataset of medical patients in California. The variables contained in the dataset are summarized in the following table.

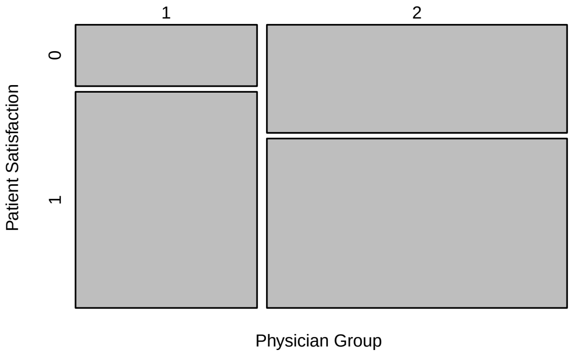
Variable Name	Description
pg	Physician Group (binary)
age	Age
sex	Sex (binary)
educ	Education (levels 1 to 6)
insu	Insurance Status (levels 1, 2, 3, 5)
drug	Drug Coverage Status (binary)
severity	Severity (levels 1 to 4)
com	Comorbidity
pcsd	Physical Comorbidity Scale
mcs.sd	Mental Comorbidity Scale
quality	Patient Satisfaction: 0 is not satsified, 1 is satsified

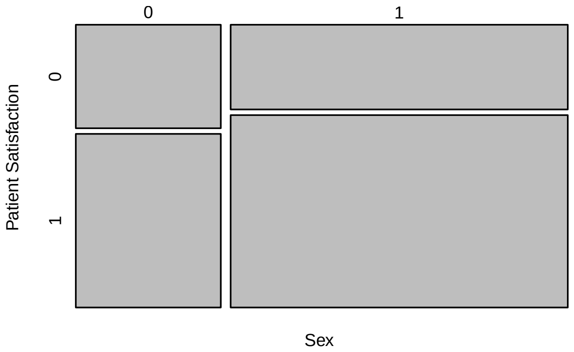
We use this data to achieve three goals:

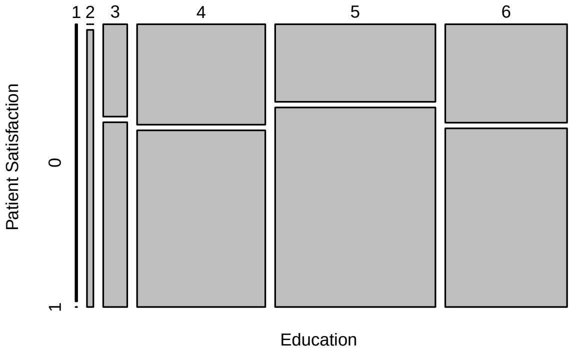
1. Predict patient satisfaction from the other variables.
2. Infer the causal effect of physician group on patient satisfaction.
3. Understand the relationship between the physician group, sex, and patient satisfaction.

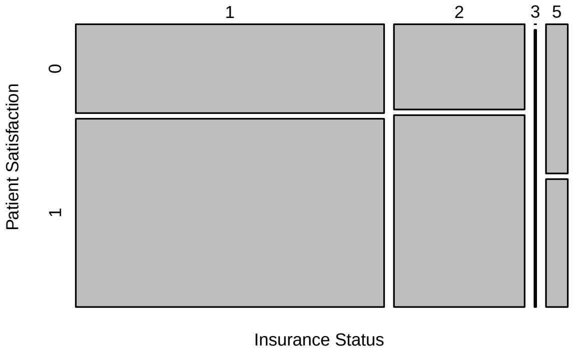
Exploratory Data Analysis

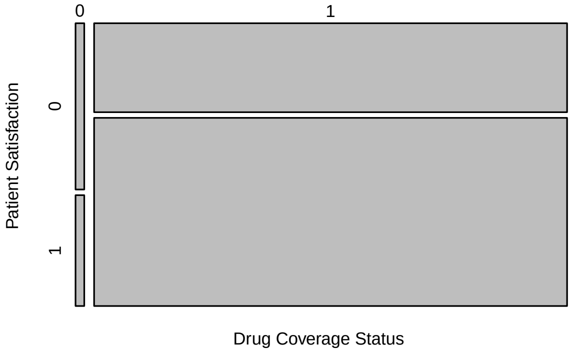
For each of the discrete covariates we have constructed a mosaic plot to examine their relationship with `quality`. The height of the bars show the conditional distribution of quality given the covariate, and the width of the columns show the marginal distribution of the covariate.

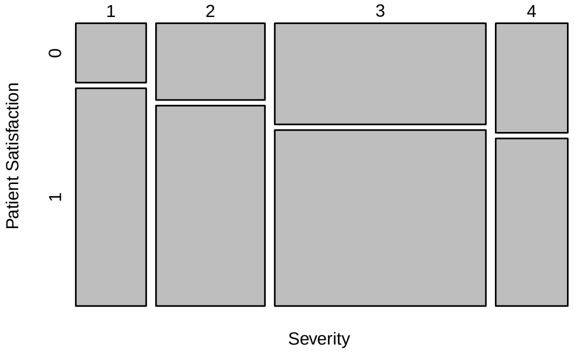






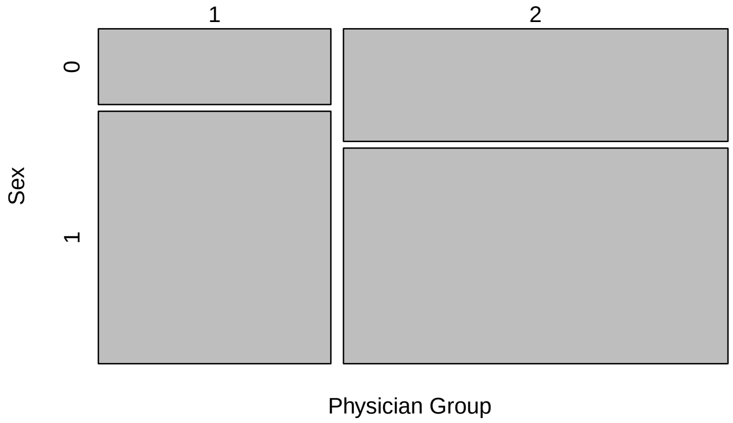




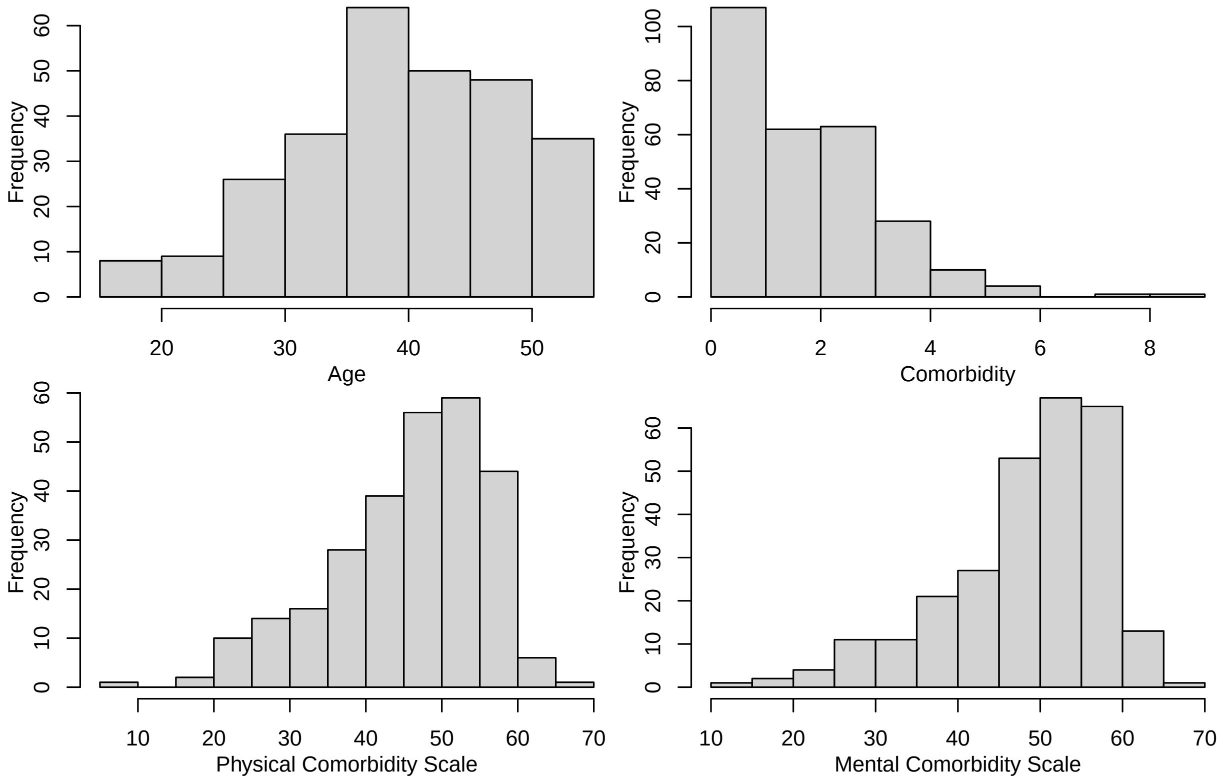


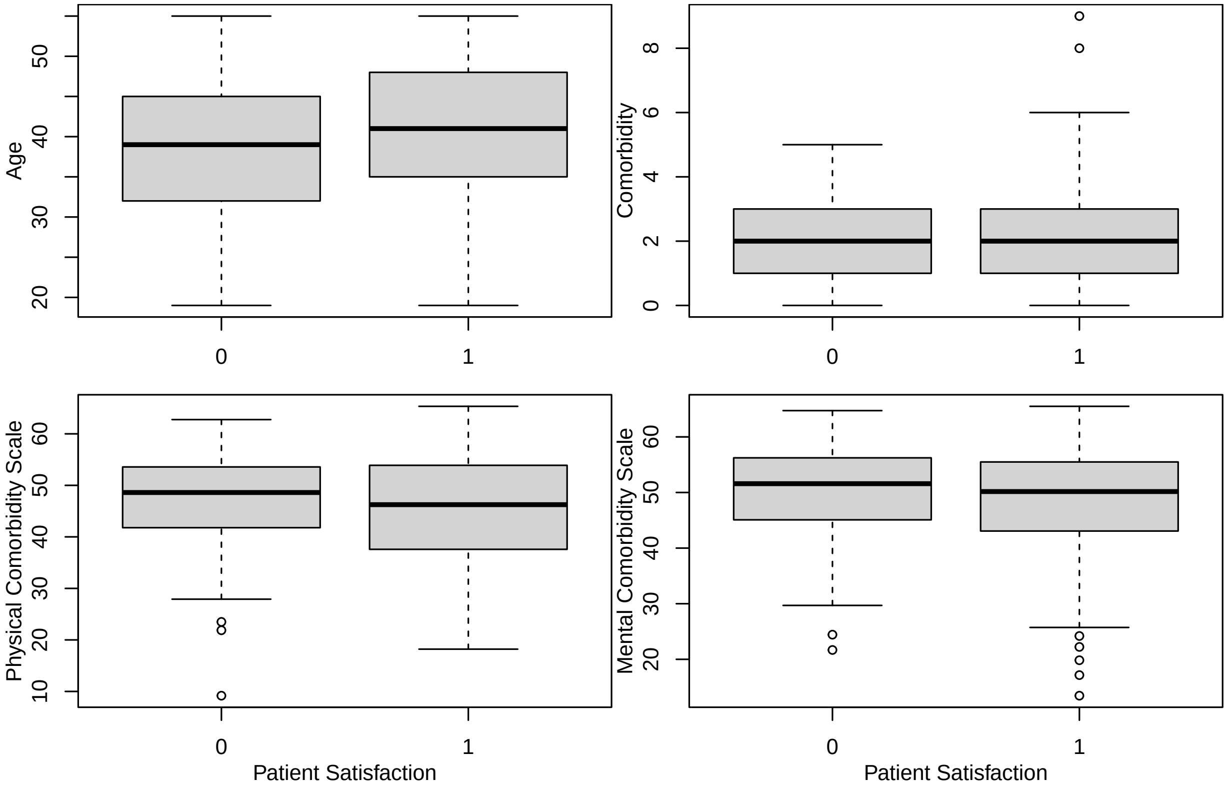
We can see a slight difference in the distribution of patient satisfaction for the different physician groups, and the patients are fairly evenly split across the physician groups. There is little differnce in patient satisfaction for the two sexes. Education levels 1 and 2 all have patient satisfaction 0 and 1, respectively, but there are very few such observations. Education levels 3, 4, 5, and 6 all have a similar distribution. Similarly with insurance status, the largest classes have similar satisfaction levels, while the rarest classes have very different levels. For severity, the proportion of patients satisfied with their treatment increases steadily from levels 1 through 4. Drug coverage status has two levels, and all but 5 observations are in category 1, which suggests this variable may not be of much use.

Since we are also interested in the relationship between physician group and sex, we also display a mosaic plot of these two variables below. We find that patients in physician group 1 are mostly of sex 1, while the proportion of patients in physician group 2 with sex 1 is slightly lower.



For the numeric covariates we first display their marginal distribution with a histogram, and then their conditional distribution given `quality` with a boxplot.





We can see that the marginal distributions are all unimodal, with comorbidity skewed right and the rest skewed left. The distributions of physical comorbidity and mental comorbidity appear to be quite similar. We can also see that the distributions of age, comorbidity, physical comorbidity, and mental comorbidity do not differ much across the different levels of patient satisfaction.

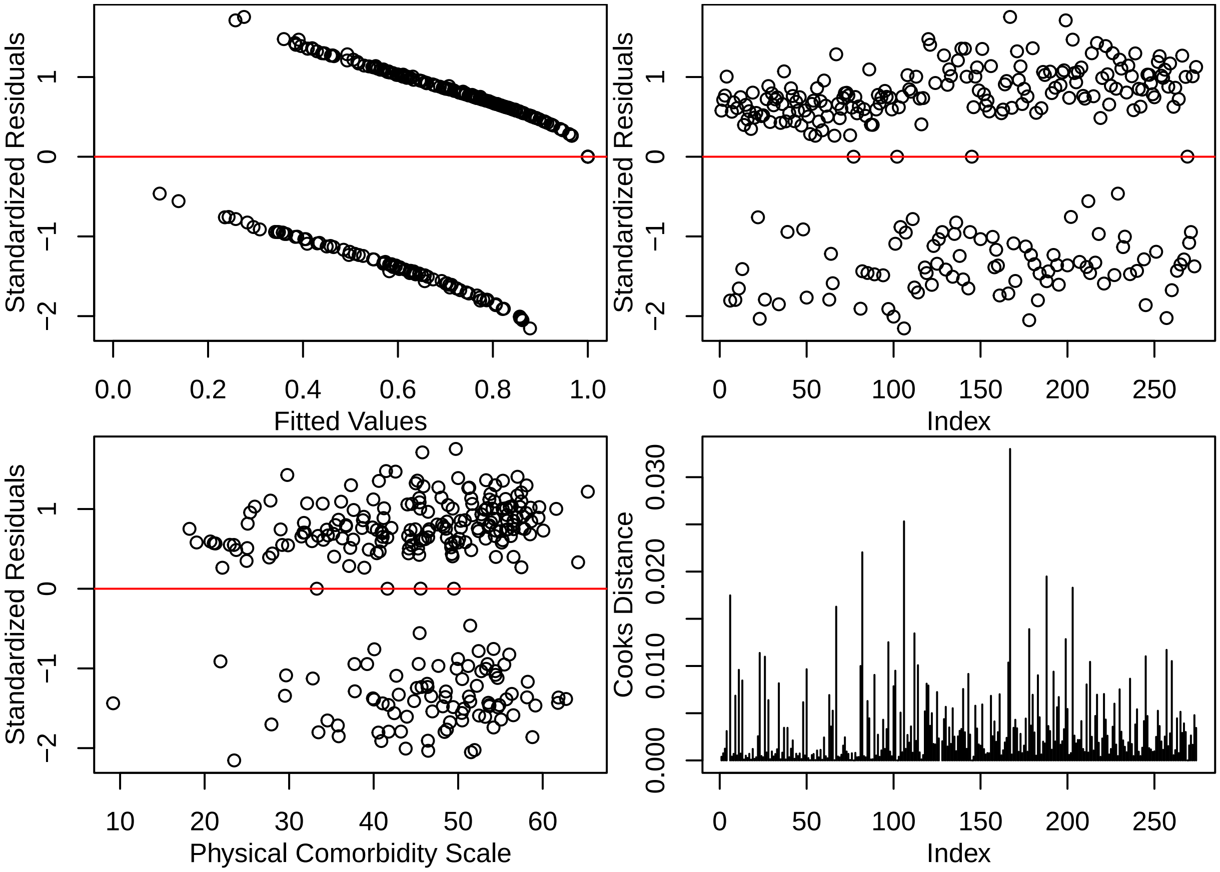
None of the covariates are very highly correlated. It is unlikely that we will have to model interactions between the covariates.

Modeling

Since we fit models for which it is not feasible to retrieve predicted class probabilities, we do not examine the Area Under the ROC Curve, and instead use 10-fold cross-validation to analyze the performance of each model. We do not withhold a test set because there are only 276 observations in the dataset.

Baseline Model

Our baseline model will be a simple logistic regression model on all the covariates. In fitting the model we used Cook's distance to identify two outliers. Both of these happened to have `drug = 0`, which left only three observations in that category. We decided to omit these two observations, and drop the `drug` variable entirely because of the huge class disparities. We do this for all subsequent models and analysis. We analyze the residuals of the logistic regression model below.



There do not seem to be any issues with the residuals. After removing the two outliers mentioned previously the Cooks Distance plot does not indicate any more outliers. The indicator variables for `pg = 2`, `severity = 3`, and `severity = 4` are all significant at $\alpha = 0.05$, and all other variables are not significant. The model has a Null Deviance of 346.93.

Model Analysis

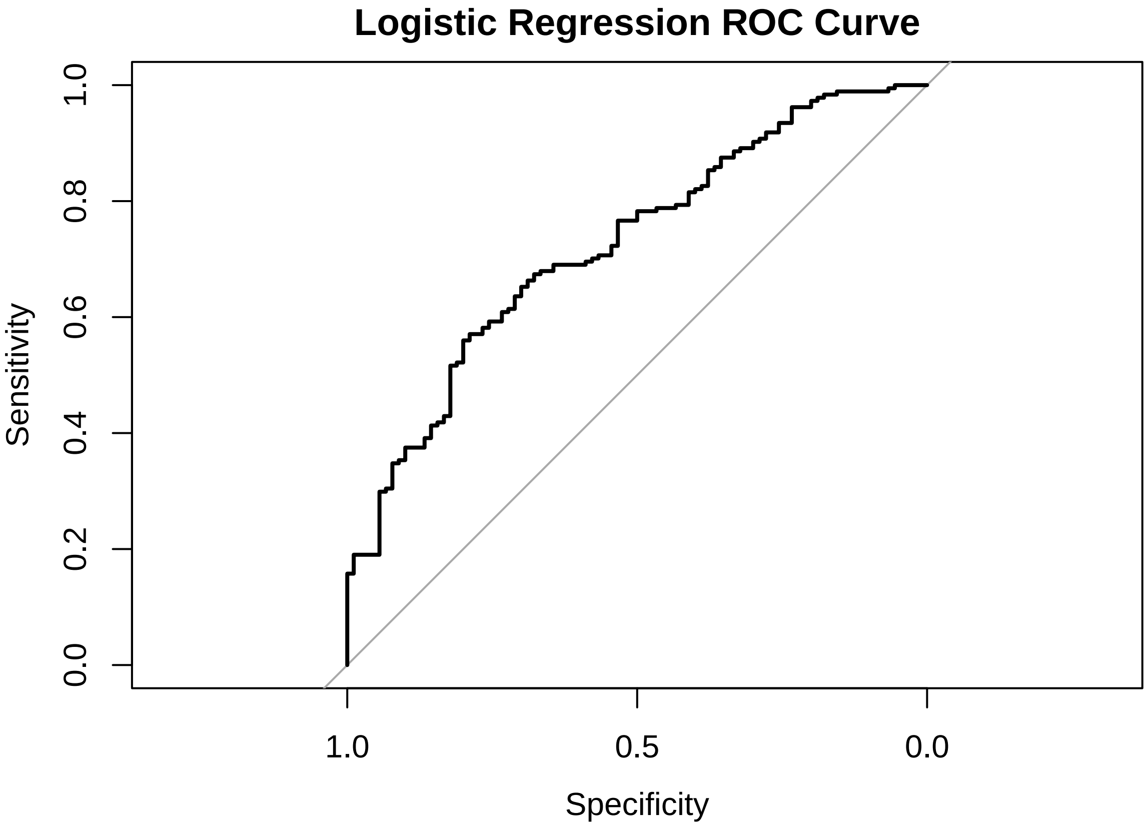
We build four models in addition to the logistic regression model: a Linear Discriminant Analysis (LDA) model, a Quadratic Discriminant Analysis (QDA) model, a Support Vector Machine (SVM) with a Gaussian kernel, and a random forest. We found no reason to modify the data further than the changes mentioned previously for any model. Note that we treat the discrete variables as numeric for the LDA and QDA models. For each model we perform 10-fold cross-validation and report the mean, variance, and 95% confidence interval of the classification errors below.

Model	Mean Error	Variance of Errors	95% Confidence Interval
Logistic Regression	0.3286	0.0089	(0.2702, 0.3869)
LDA	0.3250	0.0085	(0.2679, 0.3821)
QDA	0.3431	0.0064	(0.2935, 0.3927)
SVM + Gaussian Kernel	0.3290	0.0068	(0.2779, 0.38)
Random Forest	0.3761	0.0082	(0.32, 0.4321)

Final Models

Model Selection

We can see from the cross-validation results that the confidence intervals for the errors are all overlapping. For this reason we suggest the logistic regression model as our final model, as it has the second lowest error while still being the most interpretable model. Since we are able to retrieve the predicted class probabilities from the logistic regression model, we display its Receiver Operating Characteristic curve on the full training data below.



The Area Under the ROC Curve is 0.7235. Recall from above that the indicator variables for `pg = 2`, `severity = 3`, and `severity = 4` are significant at $\alpha = 0.05$, and all other variables are not significant. The model has a training accuracy of 70.4%, and we display the confusion matrix below.

	Predicted 0	Predicted 1
True Value 0	29	61
True Value 1	20	164

Causal Effect Analysis

To infer the causal effect of physician group on patient satisfaction, we use the plug-in estimator with our final logistic regression model. We find that if we set patient group to zero, we expect a patient satisfaction of 0.784, and if we set patient group to one, we expect a patient satisfaction of 0.605.

Graphical Modeling

We fit a graphical model to examine the relationship between physician group, sex, and patient satisfaction. All second- and third-order interactions are not significant at $\alpha = 0.05$. This tells us that `pg` is independent of `sex` given `quality`, `pg` is independent of `quality` given `sex`, and `sex` independent of `quality` given `pg`.

Limitations

We chose to exclude the `drug` variable from our models due to there being so few observations in one of the two classes. It is possible that we threw away valuable information that may hurt our model's real-world performance, but based on the data available we believe doing so was the right choice.

Conclusion

We were able to construct a model to predict patient satisfaction that gave an Area Under the ROC Curve of 0.7235, which we believe is quite good given the somewhat limited data available. We estimated the causal effect of physician group on patient satisfaction, and found that setting physician group to 0 yields higher patient satisfaction on average, compared to setting physician group to 1. We also analyzed the relationship between physician group, sex, and patient satisfaction, and found that any two of these variables are independent given the third.

Code Appendix

```
## ---- warning=FALSE, message=FALSE, echo=FALSE--------------------------------
## Import Packages
library(knitr)
library(dplyr)
library(pROC)
library(MASS)
library(e1071)
library(randomForest)

## Set Seed
set.seed(401)

## Load Data
df_numeric = read.table("asthma.txt", header=TRUE)
df_numeric$quality = as.factor(df_numeric$quality)

df = read.table("asthma.txt", header=TRUE)
df$pg = as.factor(df$pg)
df$sex = as.factor(df$sex)
df$educ = as.factor(df$educ)
df$insu = as.factor(df$insu)
df$drug = as.factor(df$drug)
df$severity = as.factor(df$severity)
df$quality = as.factor(df$quality)


## ---- warning=FALSE, message=FALSE, echo=FALSE--------------------------------
# Introduction

## Create Table of Variable Descriptions
variables = names(df)
descriptions = c(
  "Physician Group (binary)",
  "Age",
  "Sex (binary)",
  "Education (levels 1 to 6)",
  "Insurance Status (levels 1, 2, 3, 5)",
  "Drug Coverage Status (binary)",
  "Severity (levels 1 to 4)",
  "Comorbidity",
  "Physical Comorbidity Scale",
  "Mental Comorbidity Scale",
  "Patient Satisfaction: 0 is not satsified, 1 is satsified"
)
variable_info = data.frame("Variable Name"=variables, "Description"=descriptions,
                           check.names=FALSE)
kable(variable_info)


## ---- warning=FALSE, message=FALSE, echo=FALSE, fig.dim=c(8,7)----------------
## Plot Discrete Variables
par(mfrow=c(3,2),oma = c(0, 1, 0, 0), mar = c(3, 2, 0, 1), mgp = c(2, 1, 0), xpd = NA)
```

```
mosaicplot(table(df$pg, df$quality), main=NA, xlab="Physician Group",
           ylab="Patient Satisfaction", cex.axis = 1)
mosaicplot(table(df$sex, df$quality), main=NA, xlab="Sex",
           ylab="Patient Satisfaction", cex.axis = 1)
mosaicplot(table(df$educ, df$quality), main=NA, xlab="Education",
           ylab="Patient Satisfaction", cex.axis = 1)
mosaicplot(table(df$insu, df$quality), main=NA, xlab="Insurance Status",
           ylab="Patient Satisfaction", cex.axis = 1)
mosaicplot(table(df$drug, df$quality), main=NA, xlab="Drug Coverage Status",
           ylab="Patient Satisfaction", cex.axis = 1)
mosaicplot(table(df$severity, df$quality), main=NA, xlab="Severity",
           ylab="Patient Satisfaction", cex.axis = 1)


## ---- warning=FALSE, message=FALSE, echo=FALSE, out.width="70%", fig.align="center"----
## Plot pg vs. sex
mosaicplot(table(df$pg, df$sex), main=NA, xlab="Physician Group", ylab="Sex", cex.axis = 1)


## ---- warning=FALSE, message=FALSE, echo=FALSE, fig.dim=c(8,5)----------------
## Plot Numeric Variables
par(mfrow=c(2,2),oma = c(0, 1, 0, 0), mar = c(3, 2, 0, 1), mgp = c(2, 1, 0), xpd = NA)
hist(df$age, breaks=10, main=NA, xlab="Age")
hist(df$com, breaks=10, main=NA, xlab="Comorbidity")
hist(df$pcsd, breaks=10, main=NA, xlab="Physical Comorbidity Scale")
hist(df$mcs.sd, breaks=10, main=NA, xlab="Mental Comorbidity Scale")


## ---- warning=FALSE, message=FALSE, echo=FALSE, fig.dim=c(8,5)----------------
## Plot Numeric Variables Given Quality
par(mfrow=c(2,2),oma = c(0, 1, 0, 0), mar = c(3, 2, 0, 1), mgp = c(2, 1, 0), xpd = NA)
boxplot(age ~ quality, data=df, xlab=NA, ylab="Age")
boxplot(com ~ quality, data=df, xlab=NA, ylab="Comorbidity")
boxplot(pcsd ~ quality, data=df, xlab="Patient Satisfaction", ylab="Physical Comorbidity Scale")
boxplot(mcs.sd ~ quality, data=df, xlab="Patient Satisfaction", ylab="Mental Comorbidity Scale")


## ---- warning=FALSE, message=FALSE, echo=FALSE-------------------------------
## Variable Correlation
cor_table = cor(droplevels(df_numeric[,-11]))


## ---- warning=FALSE, message=FALSE, echo=FALSE-------------------------------
# Modeling

## Remove Outliers, Drop `drug`, and Fit Logistic Regression
df_numeric = df_numeric[-c(51,218), -6]
df = df[-c(51,218), -6]
out = glm(quality ~ ., data=df, family="binomial")

## Plot Residuals and Cooks Distance
par(mfrow=c(2,2),oma = c(0, 1, 0, 0), mar = c(3, 2, 0, 1), mgp = c(2, 1, 0), xpd = NA)
plot(fitted(out), rstudent(out), xlab="Fitted Values", ylab="Standardized Residuals")
```

```
abline(h=0, col="red", xpd=FALSE)
plot(rstudent(out), ylab="Standardized Residuals")
abline(h=0, col="red", xpd=FALSE)
plot(df$pcsd, rstudent(out), xlab="Physical Comorbidity Scale", ylab="Standardized Residuals")
abline(h=0, col="red", xpd=FALSE)
plot(cooks.distance(out), type="h", ylab="Cooks Distance")


## ---- warning=FALSE, message=FALSE, echo=FALSE--------------------------------
Conf = function(input, alpha) {
  se = sd(input)/sqrt(length(input))
  z = -qnorm(alpha/2)
  left  = round(mean(input) - z*se, digits=4)
  right = round(mean(input) + z*se, digits=4)
  c.i.s = paste0("(",left,", ",right,")")
  return(c.i.s)
}

## Randomly Partition Data
k = 10
indices = sample(1:nrow(df))
splits = split(indices, cut(seq_along(indices), k, labels=FALSE))

## Linear Regression Model
scores_log = rep(0, times=k)
for (i in 1:k) {
  train = df_numeric[-splits[[i]],]
  out_log = glm(quality ~ ., data=train, family="binomial")
  preds = predict(out_log, df_numeric[splits[[i]], -10], type="response")
  preds = ifelse(preds > 0.5, 1, 0)
  scores_log[i] = sum(df_numeric[splits[[i]], 10] != preds)/length(splits[[i]])
}
conf_log = Conf(scores_log, 0.05)

## LDA Model
scores_lda = rep(0, times=k)
for (i in 1:k) {
  train = df_numeric[-splits[[i]],]
  out_lda = lda(quality ~ ., data=train)
  preds = predict(out_lda, df_numeric[splits[[i]], -10])$class
  scores_lda[i] = sum(df_numeric[splits[[i]], 10] != preds)/length(splits[[i]])
}
conf_lda = Conf(scores_lda, 0.05)

## QDA Model
scores_qda = rep(0, times=k)
for (i in 1:k) {
  train = df_numeric[-splits[[i]],]
  out_lda = qda(quality ~ ., data=train)
  preds = predict(out_lda, df_numeric[splits[[i]], -10])$class
  scores_qda[i] = sum(df_numeric[splits[[i]], 10] != preds)/length(splits[[i]])
}
conf_qda = Conf(scores_qda, 0.05)
```

```
## SVM with Gaussian Kernel Model
scores_svm = rep(0, times=k)
for (i in 1:k) {
  train = df[-splits[[i]],]
  out_svm = svm(quality ~., data=train, type="C-classification")
  preds = predict(out_svm, df[splits[[i]], -10])
  scores_svm[i] = sum(df[splits[[i]], 10] != preds)/length(splits[[i]])
}
conf_svm = Conf(scores_svm, 0.05)

## Random Forest
scores_rf = rep(0, times=k)
for (i in 1:k) {
  train = df[-splits[[i]],]
  out_rf = randomForest(quality ~ ., data=train, ntree = 500)
  preds = predict(out_rf, df[splits[[i]], -10])
  scores_rf[i] = sum(df[splits[[i]], 10] != preds)/length(splits[[i]])
}
conf_rf = Conf(scores_rf, 0.05)

## Cross-Validation Summary Table
models = c(
  "Logistic Regression",
  "LDA",
  "QDA",
  "SVM + Gaussian Kernel",
  "Random Forest"
)
cv_means = c(mean(scores_log), mean(scores_lda), mean(scores_qda), mean(scores_svm), mean(scores_rf))
cv_means = round(cv_means, digits=4)
cv_vars = c(var(scores_log), var(scores_lda), var(scores_qda), var(scores_svm), var(scores_rf))
cv_vars = round(cv_vars, digits=4)
cv_confs = c(conf_log, conf_lda, conf_qda, conf_svm, conf_rf)
cv_results = data.frame("Model"=models, "Mean Error"=cv_means,
                        "Variance of Errors"=cv_vars,
                        "95% Confidence Interval"=cv_confs, check.names=FALSE)
kable(cv_results)


## ---- warning=FALSE, message=FALSE, echo=FALSE--------------------------------
# Final Models

## Logistic Regression ROC Curve
R = roc(df$quality, fitted(out))
plot(R, main="Logistic Regression ROC Curve")
auc_log = auc(R)


## ---- warning=FALSE, message=FALSE, echo=FALSE--------------------------------
## Confusion Matrix
preds = predict(out, df[,-10], type="response")
preds = ifelse(preds > 0.5, 1, 0)
confusion_matrix = table(df$quality, preds)
```

```
confusion_df = data.frame(" "=c("True Value 0", "True Value 1"),
                          "Predicted 0"=confusion_matrix[,1],
                          "Predicted 1"=confusion_matrix[,2],
                          check.names=FALSE)
rownames(confusion_df) = c()
kable(confusion_df)


## ---- warning=FALSE, message=FALSE, echo=FALSE--------------------------------
## Causal Inference
final_model = out
grid = c(1, 2)
estimates = rep(0, 2)

XAtemp = df[, -10]
for (i in 1:2) {
  XAtemp$pg = grid[i]
  XAtemp$pg = factor(x=XAtemp$pg, levels=c(1,2))
  hat_vals = predict(final_model, XAtemp, type="response")
  estimates[i] = mean(hat_vals)
}


## ---- warning=FALSE, message=FALSE, echo=FALSE--------------------------------
## Graphical Model
counts = df %>% count(pg, sex, quality)
out_glm = glm(n ~ pg*sex*quality, data=counts, family="poisson")


## ----code = readLines(knitr::purl(knitr::current_input(), documentation = 1)), echo = T, eval = F----
## NA
```