# Diabetes

Jillian Maw

4/6/2022

## HOMEWORK ANALYSIS #7 - DIABETES

Type 2 diabetes is a problem with your body that causes blood sugar levels to rise higher than normal (hyperglycemia) because your body does not use insulin properly. Specifically, your body can't make enough insulin to keep your blood sugar levels normal. Type 2 diabetes is associated with various health complications such as neuropathy (nerve damage), glaucoma, cataracts and various skin disorders. Early detection of diabetes is crucial to proper treatment so as to alleviate complications. The dataset `Diabetes.txt` contains information on 768 *women* who are at risk for diabetes. The dataset contains the following variables:

| Variable Name | Description |
|---|---|
| pregnant | Number of times pregnant |
| glucose | Plasma glucose concentration at 2 hours in an oral glucose tolerance test |
| diastolic | Diastolic blood pressure (mm Hg) |
| triceps | Triceps skin fold thickness (mm) |
| insulin | 2 hour serum insulin (mu U/ml) |
| bmi | Body mass index |
| pedigree | Numeric strength of diabetes in family line (higher numbers mean stronger history) |
| age | Age |
| diabetes | Does patient have diabetes (0 if "No", 1 if "Yes") |

Doctors hope to use the covariate information to diagnose if a patient has diabetes or not. **Note: many of the observations in this dataset contain values that can't occur (e.g. a BMI of 0). You will need to clean the dataset prior to your analysis.**

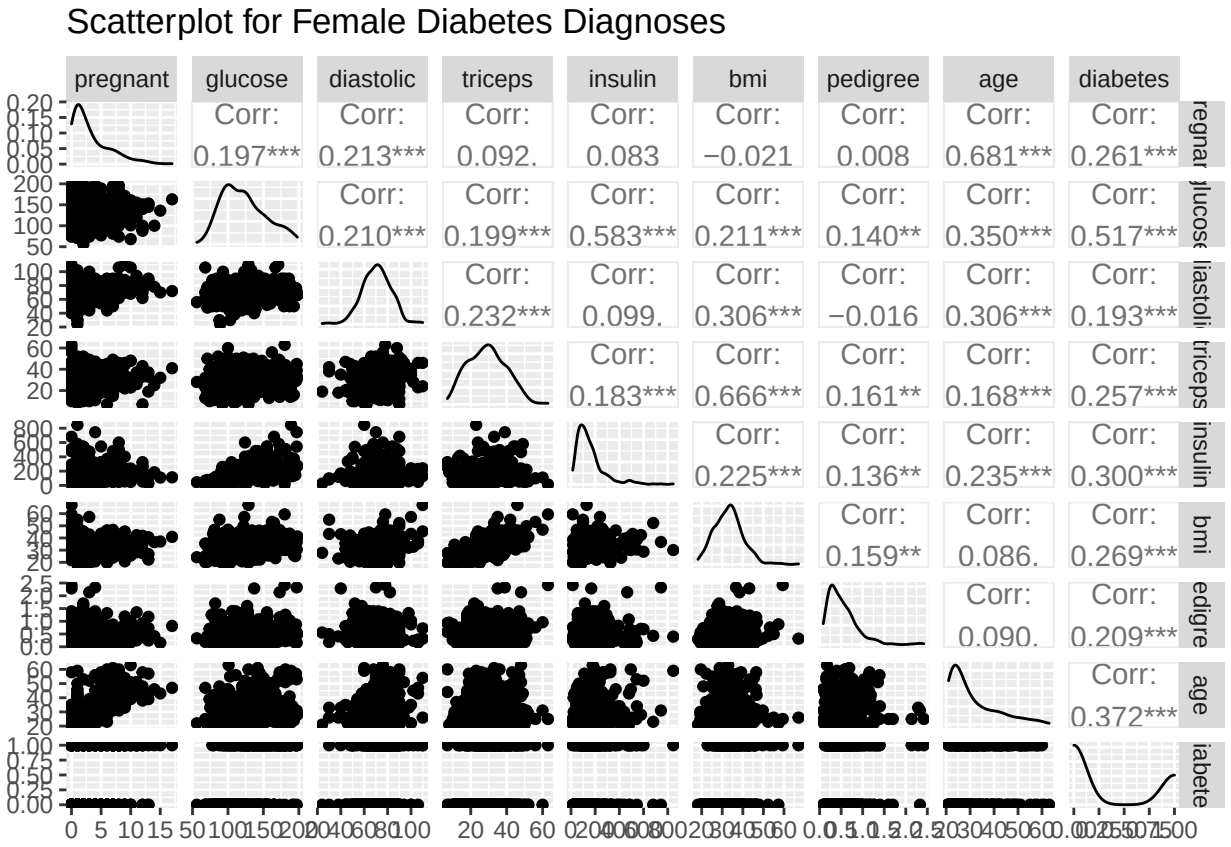
1. In your own words, summarize the overarching problem. Discuss how statistical modeling will be able to answer the posed questions.

We have been given a dataset about anonymous women, their medical measurements, and their diabetes diagnosis, and asked to create a statistical model that can determine a diagnosis for diabetes, for a woman. A diagnosis is a categorical variable, specifically a categorical variable with two possible values, "no" or "yes". We can encode these as "0" or "1", respectively. All of this means that we cannot use a multiple linear regression model, like we have heretofore done in the past; we must use a logistical regression model. Doctors hope to use our logistical regression model to predict if a female patient has diabetes or not, in order to help with the management of diabetes symptoms and prevent complications.

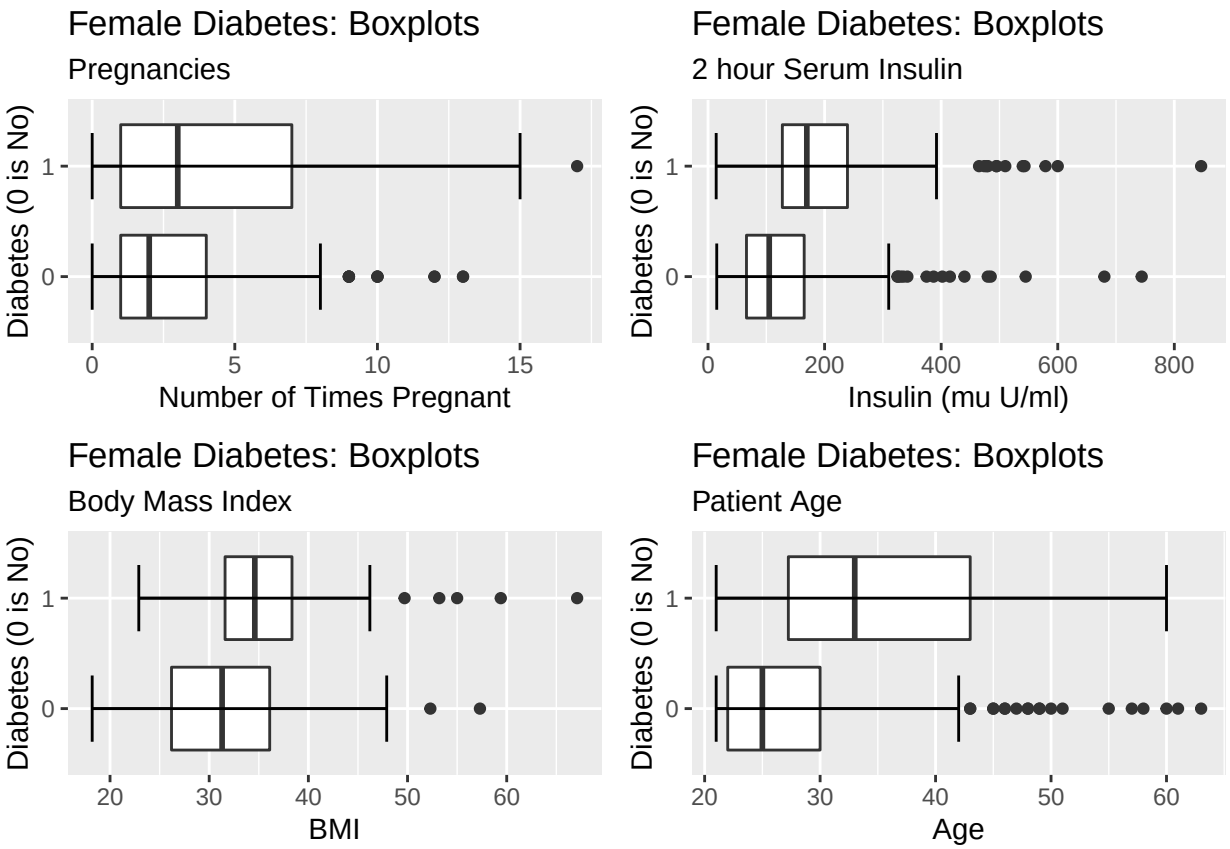
2. Explore the data using basic exploratory graphics and summary statistics. Include scatterplots with smooth curves to show the relationship between 2 covariates and the response (diabetes). Comment on any potential relationships you see through this exploratory analysis. Explain why traditional multiple linear regression methods are not suitable for this problem.

As stated earlier, a diagnosis of diabetes is a categorical variable, specifically a categorical variable with two possible values, "no" or "yes", that can be encoded as a "0" or "1", respectively. This means that we cannot use a multiple linear regression model, like we have heretofore done in the past; we would be unable to restrict the predictions to be between 0 or 1, the linear assumption needed for a linear regression model could be violated, the errors in prediction won't be normal, and the equal variance could also be violated. We must use a logistical regression model, to restrict our diagnosis to be between the two possible outcomes.

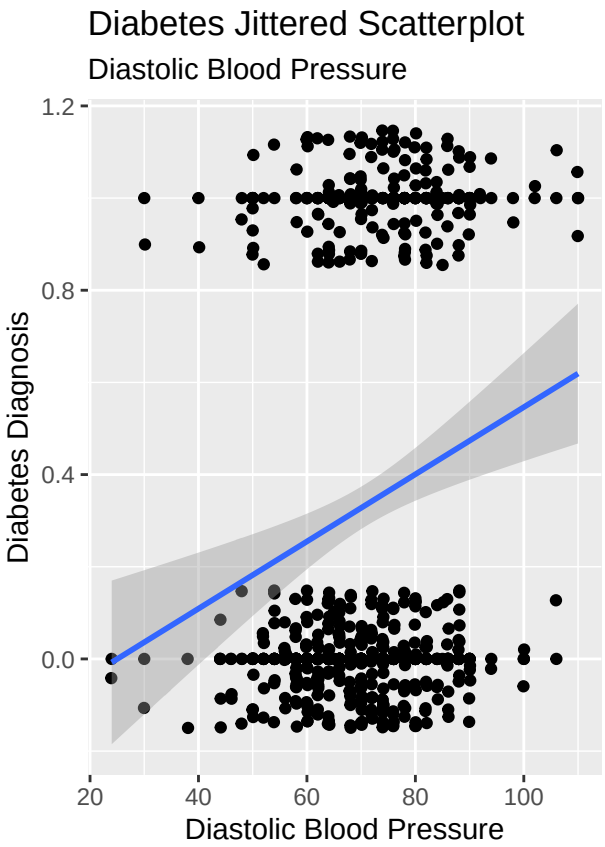
Below, we show the correlation color matrix plot and the scatterplot matrix showing a basic, visual overview of the data. As can be seen in the correlation matrix, there is strong correlation between diabetes and insulin, age and pregnancy, bmi and triceps, and also insulin and glucose. These explanatory variables potentially have interactions and collinearity that we will need to be aware of. The scatterplot matrix shows several of the variables have potential outliers. In fact, this data set had to be cleaned of missing measurements or measurements that just did not make sense (like the impossible bmi of 0). We also removed one outlier based on the age being 20 years older than the next oldest woman in the data set, to re-balance the difference in age distributions between those with a diabetes diagnosis and those without one.

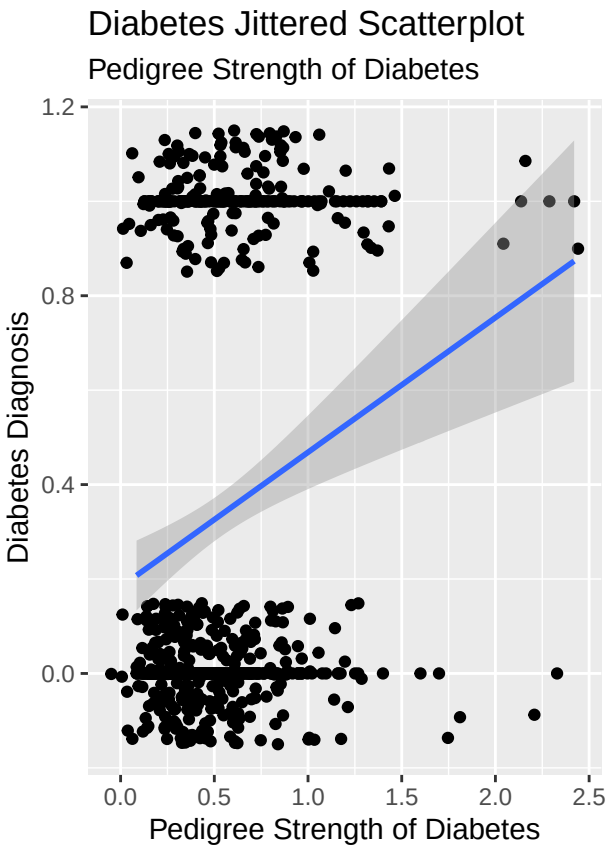



Below, we highlight four of the variables, pregnancy, insulin, body mass index, and age, that seemed to show outliers in the scatterplot matrix, in boxplot form, to show the differences in spread, according to diabetes diagnosis. We can see that removing the records of the 81-year-old woman, especially since she was not diagnosed with diabetes, improves the spread of the age variable to be more equal. We elected to leave the other outliers that were not as extreme in the other variables. We can see that all of these variables have a noticeable difference in means depending on the group of diabetes diagnosis.




Finally, we include two scatterplots with smooth curves to show the relationship between the covariate variable diastolic blood pressure and the response variable diabetes diagnosis, and the covariate variable pedigree and the response variable diabetes diagnosis. Both scatterplots make use of jitter to show that there are more spots on the 0 and 1 than would be seen without jitter. Both plot's logistic regression lines show a positive relationship between the variables and the diabetes diagnosis, meaning as diastolic blood pressure or pedigree went up, so did the probability of a positive diabetes diagnosis.




3. Use variable selection to choose which variables to use in a logistic regression model for `diabetes`. Provide a justification of your choice in criteria (AIC or BIC) and algorithm (forward vs. backward vs. exhaustive). What factors do you find are important in explaining the presence of diabetes?

Due to the number of explanatory variables in the data being less than 40, we can use the "best subset selection" variable selection procedure (also referred to as the exhaustive variable selection procedure) because it is the best method for minimizing the Aikake Information Criteria, the Bayesian Information Criteria, or the Predictive Error, and maximizes the Adjusted $R^2$ for a logistic regression model. These criteria are the best indicators for determining the fit of the model. In addition, the "best subset selection" procedure tests all possible variables with each other, allowing us to know with certainty that we have the best model. We decided to use the Aikake Information Criteria, or AIC, model comparison criterion, because we are looking to predict if a woman will have a positive diabetes diagnosis, and the AIC model comparison criterion is optimized for making predictions, compared to the Bayesian Information Criteria, or BIC, which is optimized for making inferences. After using the "best subset selection" variable selection method and the AIC model comparison criteria, we decided the most important variables to include were `glucose` (the measurement of plasma glucose concentration at 2 hours in an oral glucose tolerance test), `bmi` (the woman's body mass index), `pedigree` (the numeric strength of diabetes in a woman's family line, where higher numbers mean a stronger history), and `age` (the age of the woman).

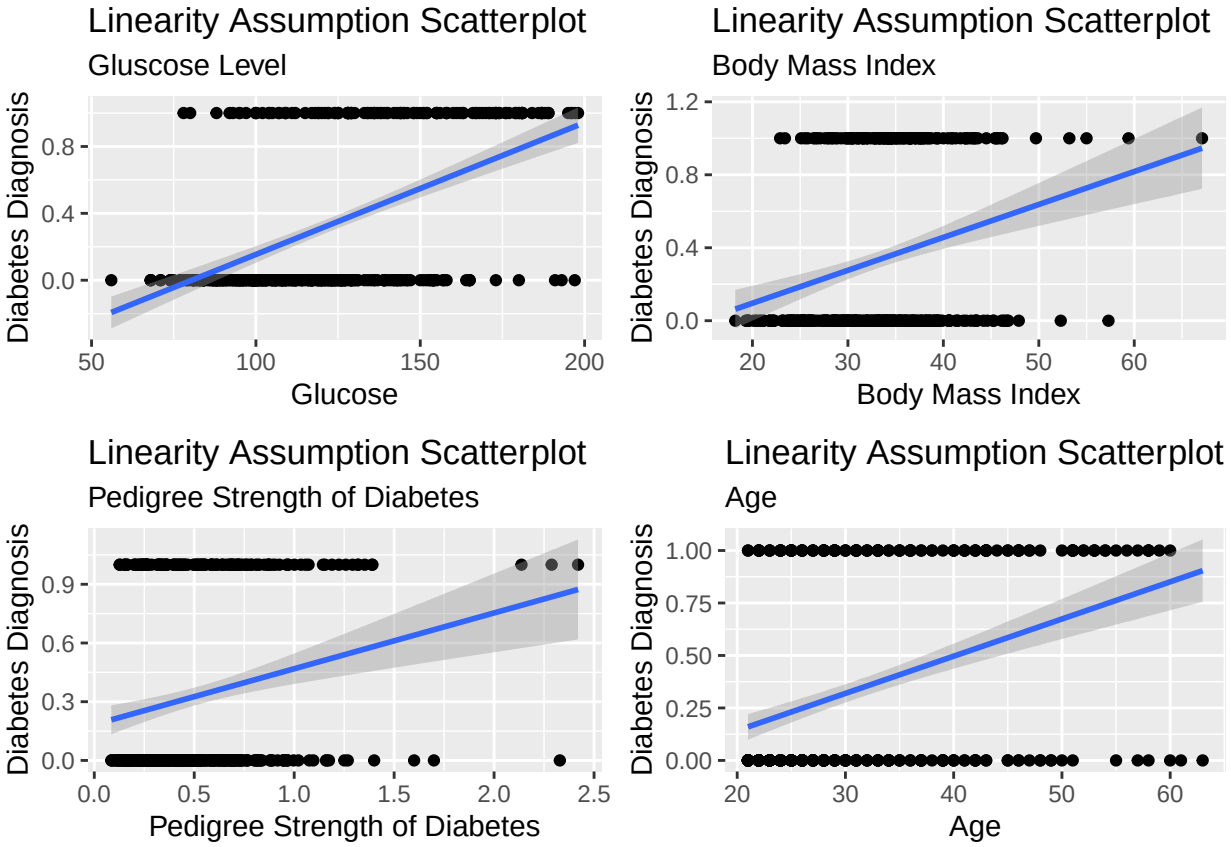
4. Write out a logistic regression model (using Greek letters) that includes your chosen covariates. Describe and justify any assumptions that you use in writing out your model.

$$y_i \overset{ind}{\sim} \text{Bern}(p_i)$$

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \sum_{j=1}^{P} x_{ij}\beta_j, \text{ where } \beta_j \text{ represents the following variables:}$$

- $\beta_1$, the woman's plasma glucose concentration, at 2 hours in an oral glucose tolerance test.
- $\beta_2$, the woman's body mass index.
- $\beta_3$, the woman's pedigree, or the numeric strength of diabetes in her family line. A higher number means a stronger history.
- $\beta_4$, the woman's age.

The assumptions made for this model include that the data set contains independent data and, when the dataset is graphed by log-odds, it is linear. We can assume the data is independent because no woman's measurements should affect another woman's. Even if there were women related to each other in the study, we can safely assume that the parent's genetics may influence a difference in sizes to allow for a reasonable assumption of independence between all women. When we check the linearity of the log-odds of the data set, we can create scatterplots of the explanatory variables, $\beta_1$ through $\beta_5$ as explained above, against the response variable, diabetes diagnosis, and check that the regression lines modeling those relationships are monotonic, or always increasing or always decreasing. Below are the scatterplots proving these relationships.



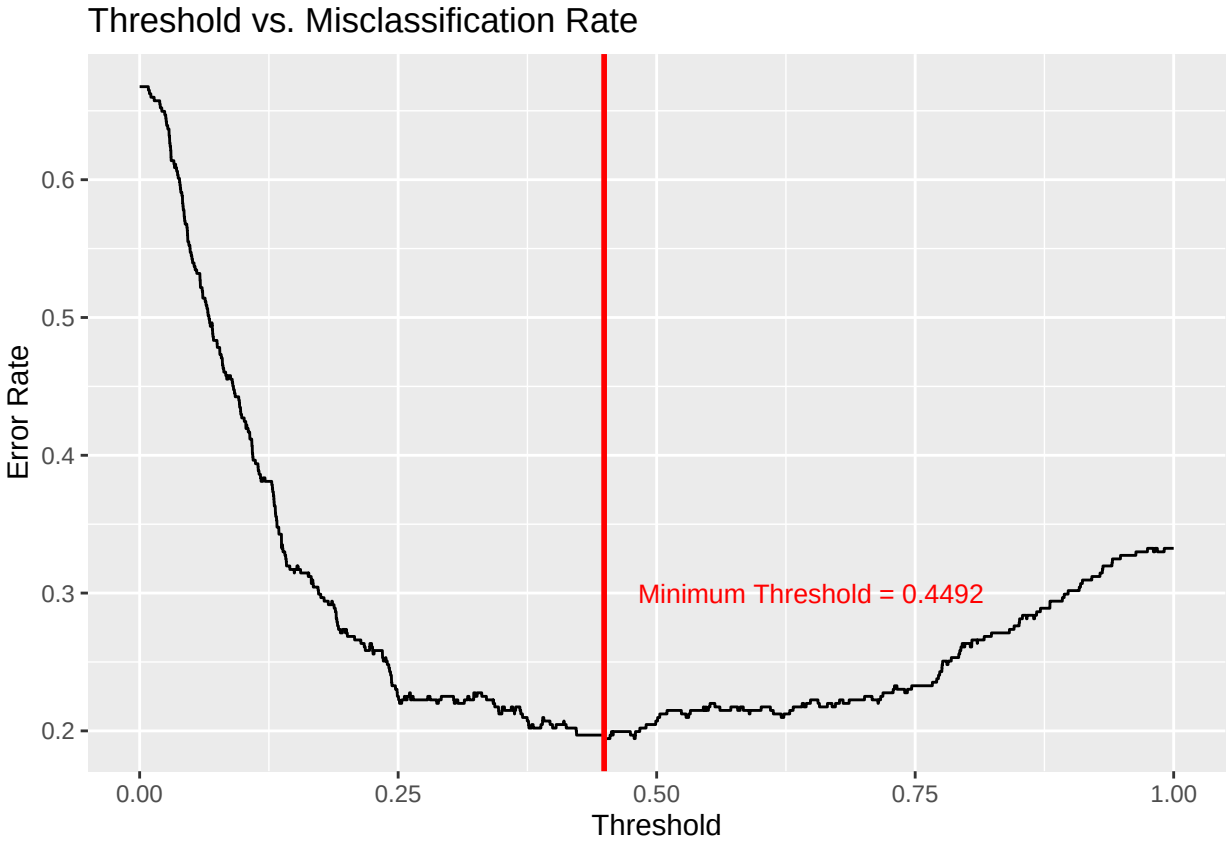

5. Fit the corresponding logistic regression model and give a 95% confidence interval for each effect therein. Interpret at least one (but not the intercept) of these intervals in the context of the problem.

We fit the model below with our best estimates of the coefficients $\beta_j$:

$\hat{y}_i \overset{ind}{\sim} \text{Bern}(p_i)$,$log(\frac{p_i}{1-p_i})$ = -10.2262 + 0.036 $x_i$ (glucose)+ 0.0728 $x_i$ (bmi)+ 1.0766 $x_i$ (pedigree)+ 0.0605 $x_i$ (age)·

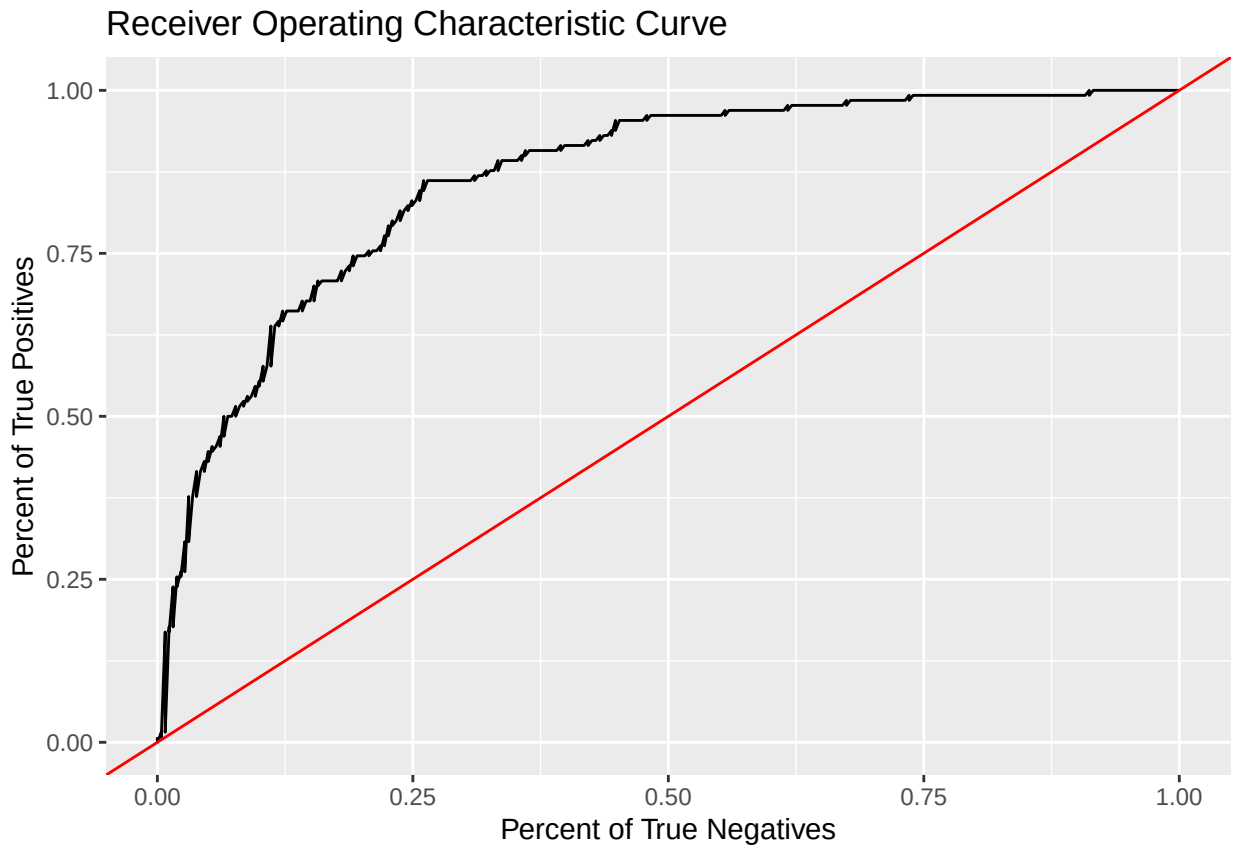
We are 95% confident that the true intercept for the above logistic regression model is between -99.9996% and -99.9723%, on average. We are 95% confident that as the plasma glucose concentration level increases by 1 the likelihood of a positive diabetes diagnosis increases between 2.687% and 4.7204%, on average. We are 95% confident that as the body mass index increases by 1 the likelihood of a positive diabetes diagnosis increases between 3.4488% and 12.0492%, on average. We are 95% confident that as the pedigree increases by 1 the likelihood of a positive diabetes diagnosis increases between 31.1299% and 581.5776%, on average. We are 95% confident that as the age increases by 1 the likelihood of a positive diabetes diagnosis increases between 3.3976% and 9.2743%, on average.

6. Determine an appropriate threshold for classification that minimizes the misclassification rate. Provide an appropriate plot showing that this is indeed the minimum.




As can be seen in the graph above, the threshold for classification that minimizes the misclassification rate is 0.4492, as this is were the relationship between the threshold and the error rate was lowest recorded.

7. Assess the model fit by building a confusion matrix from *all the data (i.e. not cross-validated)* using the classification threshold that you found in the previous problem and AUC for your logistic regression model. Comment on how well your model fits the data by its ability to correctly classify patients in the dataset. State your results in terms of sensitivity, specificity, positive predictive value, and negative predictive value.

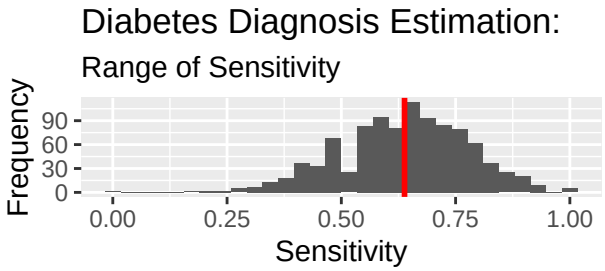


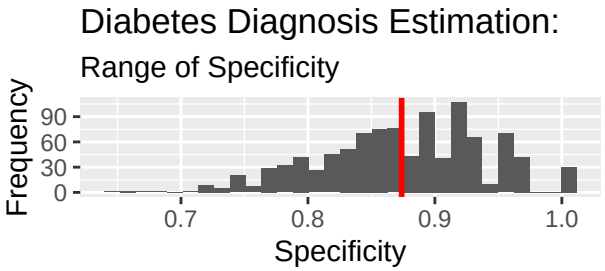

```
##           Actual
## Predicted   0   1 Sum
##       0   228  44 272
##       1    33  86 119
##       Sum 261 130 391
```

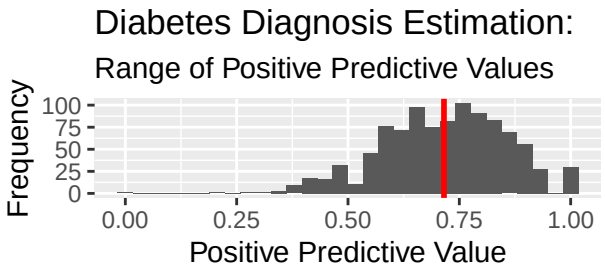
Our logistic regression model can create the above confusion matrix, with an area under the curve of the Receiver Operating Characteristic Curve of 0.8634, which says we classify fairly well across all thresholds. The confusion matrix also shows that we have a sensitivity (the percent of true positives) of 0.3384615, a specificity (the percent of true negatives) of 0.8735632, a positive predictive value (precision) of 0.6615385, and a negative predictive value of 0.8382353. This means 33.8461538% of people were correctly diagnosed with diabetes.

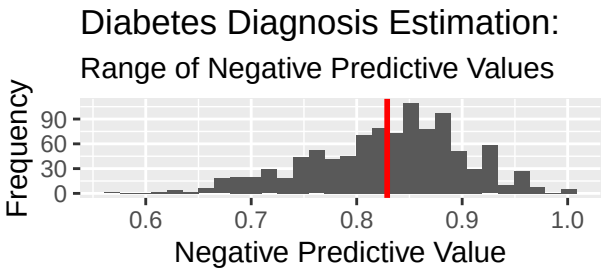
8. Assess the predictive ability of your model by running a cross-validation study where you classify the "test" patients using the threshold you found above. Report your results in terms of the average sensitivity, specificity, positive predictive value, and negative predictive value for the test sets.

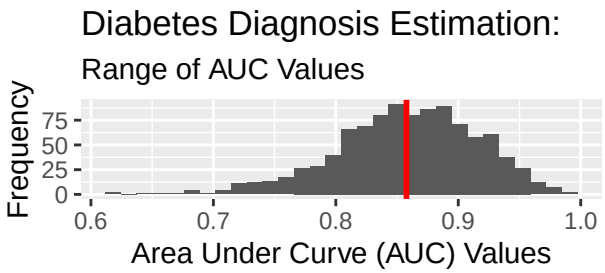
To test the predictive ability of the logistic regression model we created, we can run a cross-validation study classifying patients as either diabetic or not, using the previously found threshold of 0.4492. We found the cross validation generally produced average results of sensitivity, specificity, positive predicted value, and negative predicted value of 0.6383, 0.8738, 0.7155, and 0.8289, respectively. To help with proving the cross validation, we included graphics below that illustrate the 1000 tests done to prove the predictive ability of the new logistic regression model.




9. Predict the probability of diabetes for the following patient: pregnant= 1, glucose= 90, diastolic=62, triceps= 18, insulin= 59, bmi= 25.1, pedigree= 1.268, and age= 25. Do you think patient has diabetes? Why or why not?

For a woman who is 25 years old, has a pedigree (or numeric strength of diabetes in her family line) of 1.628, a body mass index of 25.1, and a plasma glucose concentration (at 2 hours in an oral glucose tolerance test) of 90, we predict the probability that she has diabetes is 9.2338%. Therefore, the probability of her having diabetes is low, due to our model's reported low prediction response being well below our threshold of 44.92%.

## Appendix of Code

```
knitr::opts_chunk$set(echo = FALSE, include = FALSE)
library(vroom) #faster data reading
library(DataExplorer) #for plot explorations
library(ggplot2) #for professional looking graphics
library(GGally) #for ggpairs
library(MASS) #for standardized residuals
library(normtest) #for JB-test
library(lmtest) #for BP-test
library(car) #for variance inflation factors & added-variable plots
library(caret) #for confusion matrix
library(bestglm) #for variable selection procedures
```

```r
library(knitr) #for pretty tables with kable
# library(kableExtra) #for if kable needs to be landscape
library(gridExtra) #for making grids on same row
library(pROC) #for making a ROC curve on step 7
options(scipen = 999) #for preventing scientific notation
diagnosis = vroom("~/R programming/STAT_330/Diabetes.txt", show_col_types = FALSE)
#diastolic, insulin, and bmi cannot be 0, triceps and glucose probably cannot be 0
#remove 0s from this data set, we start with 768 obs
diagnosis <- subset(diagnosis, glucose != 0) #now 763 obs, removed 5 obs
diagnosis <- subset(diagnosis, diastolic != 0) #now 728 obs, removed 35 obs
diagnosis <- subset(diagnosis, triceps != 0) #now 534 obs, removed 194 obs
diagnosis <- subset(diagnosis, insulin != 0) #now 393 obs, removed 141 obs
diagnosis <- subset(diagnosis, bmi != 0) #now 392 obs, removed 1 obs
#what if I drop the max(diagnosis$age)?
diagnosis <- subset(diagnosis, age != max(diagnosis$age))
#total of 376 obs removed; may need to remove other values
#plot_intro(diagnosis)
#plot_missing(diagnosis)
plot_correlation(diagnosis)
ggpairs(diagnosis, progress = FALSE, title = 'Scatterplot for Female Diabetes Diagnoses')
preg.box <- ggplot(data=diagnosis, mapping=aes(x = pregnant, y = factor(diabetes))) +
  geom_boxplot() +
  xlab('Number of Times Pregnant') +
  ylab('Diabetes (0 is No)') +
  ggtitle('Female Diabetes: Boxplots', subtitle = 'Pregnancies') +
  stat_boxplot(geom ='errorbar', width = 0.6)

# gluc.box <- ggplot(data=diagnosis, mapping=aes(x = glucose, y = factor(diabetes))) +
#   geom_boxplot() +
#   xlab('Gluscose Level') +
#   ylab('Diabetes (0 is No)') +
#   ggtitle('Female Diabetes: Boxplots',
#           subtitle = 'Plasma Glucose Concentration (2 hours, oral tolerance test)') +
#   stat_boxplot(geom ='errorbar', width = 0.6)

# dia.box <- ggplot(data=diagnosis, mapping=aes(x = diastolic, y = factor(diabetes))) +
#   geom_boxplot() +
#   xlab('Diastolic Blood Pressure (mm Hg)') +
#   ylab('Diabetes (0 is No)') +
#   ggtitle('Female Diabetes: Boxplots', subtitle = 'Diastolic Blood Pressure') +
#   stat_boxplot(geom ='errorbar', width = 0.6)

# tri.box <- ggplot(data=diagnosis, mapping=aes(x = triceps, y = factor(diabetes))) +
#   geom_boxplot() +
#   xlab('Tricep Skin Fold Thickness (mm)') +
#   ylab('Diabetes (0 is No)') +
#   ggtitle('Female Diabetes: Boxplots', subtitle = 'Tricep Skin Fold Thickness') +
#   stat_boxplot(geom ='errorbar', width = 0.6)

ins.box <- ggplot(data=diagnosis, mapping=aes(x = insulin, y = factor(diabetes))) +
  geom_boxplot() +
  xlab('Insulin (mu U/ml)') +
  ylab('Diabetes (0 is No)') +
```

```
  ggtitle('Female Diabetes: Boxplots', subtitle = '2 hour Serum Insulin') +
  stat_boxplot(geom ='errorbar', width = 0.6)

bmi.box <- ggplot(data=diagnosis, mapping=aes(x = bmi, y = factor(diabetes))) +
  geom_boxplot() +
  xlab('BMI') +
  ylab('Diabetes (0 is No)') +
  ggtitle('Female Diabetes: Boxplots', subtitle = 'Body Mass Index') +
  stat_boxplot(geom ='errorbar', width = 0.6)

# ped.box <- ggplot(data=diagnosis, mapping=aes(x = pedigree, y = factor(diabetes))) +
#   geom_boxplot() +
#   xlab('Diabetes Pedigree Strength') +
#   ylab('Diabetes (0 is No)') +
#   ggtitle('Female Diabetes: Boxplots',
#           subtitle = 'Numeric Strength of Diabetes in Pedigree') +
#   stat_boxplot(geom ='errorbar', width = 0.6)

age.box <- ggplot(data=diagnosis, mapping=aes(x = age, y = factor(diabetes))) +
  geom_boxplot() +
  xlab('Age') +
  ylab('Diabetes (0 is No)') +
  ggtitle('Female Diabetes: Boxplots', subtitle = 'Patient Age') +
  stat_boxplot(geom ='errorbar', width = 0.6)

# print(preg.box)
# print(gluc.box)
# print(dia.box)
# print(tri.box)
# print(ins.box)
# print(bmi.box)
# print(ped.box)
# print(age.box)

grid.arrange(preg.box, ins.box, bmi.box, age.box, nrow = 2)
diastolic.scatter <- ggplot(data = diagnosis, mapping=aes(y=diabetes, x=diastolic)) +
  geom_point() +
  geom_jitter(width=0.15,height=0.15) +
  xlab('Diastolic Blood Pressure') +
  ylab('Diabetes Diagnosis') +
  ggtitle('Diabetes Jittered Scatterplot', subtitle = 'Diastolic Blood Pressure') +
  geom_smooth(method="glm")

pedigree.scatter <- ggplot(data = diagnosis, mapping=aes(y=diabetes, x=pedigree)) +
  geom_point() +
  geom_jitter(width=0.15,height=0.15) +
  xlab('Pedigree Strength of Diabetes') +
  ylab('Diabetes Diagnosis') +
  ggtitle('Diabetes Jittered Scatterplot', subtitle = 'Pedigree Strength of Diabetes') +
  geom_smooth(method="glm")

suppressMessages(grid.arrange(diastolic.scatter, pedigree.scatter, nrow = 1))
vs.res <- bestglm(as.data.frame(diagnosis), IC="AIC", method="exhaustive",
```

```
                  TopModels=5, family=binomial)
vs.res$BestModel
pred.logreg <- glm(diabetes~glucose+bmi+pedigree+age, diagnosis, family = binomial)
gluc.scatter <- ggplot(data = diagnosis, mapping=aes(y=diabetes, x=glucose)) +
  geom_point() +
  #geom_jitter(width=0.15,height=0.15) +
  xlab('Glucose') +
  ylab('Diabetes Diagnosis') +
  ggtitle('Linearity Assumption Scatterplot', subtitle = 'Gluscose Level') +
  geom_smooth(method="glm")

bmi.scatter <- ggplot(data = diagnosis, mapping=aes(y=diabetes, x=bmi)) +
  geom_point() +
  #geom_jitter(width=0.15,height=0.15) +
  xlab('Body Mass Index') +
  ylab('Diabetes Diagnosis') +
  ggtitle('Linearity Assumption Scatterplot', subtitle = 'Body Mass Index') +
  geom_smooth(method="glm")

pedigree2.scatter <- ggplot(data = diagnosis, mapping=aes(y=diabetes, x=pedigree)) +
  geom_point() +
  #geom_jitter(width=0.15,height=0.15) +
  xlab('Pedigree Strength of Diabetes') +
  ylab('Diabetes Diagnosis') +
  ggtitle('Linearity Assumption Scatterplot',
          subtitle = 'Pedigree Strength of Diabetes') +
  geom_smooth(method="glm")

age.scatter <- ggplot(data = diagnosis, mapping=aes(y=diabetes, x=age)) +
  geom_point() +
  #geom_jitter(width=0.15,height=0.15) +
  xlab('Age') +
  ylab('Diabetes Diagnosis') +
  ggtitle('Linearity Assumption Scatterplot', subtitle = 'Age') +
  geom_smooth(method="glm")

suppressMessages(grid.arrange(gluc.scatter, bmi.scatter, pedigree2.scatter,
                              age.scatter, nrow = 2))
diag.beta.0 <- round(as.numeric(coef(pred.logreg)["(Intercept)"]), digits = 4)
diag.beta.1 <- round(as.numeric(coef(pred.logreg)["glucose"]), digits = 4)
diag.beta.2 <- round(as.numeric(coef(pred.logreg)["bmi"]), digits = 4)
diag.beta.3 <- round(as.numeric(coef(pred.logreg)["pedigree"]), digits = 4)
diag.beta.4 <- round(as.numeric(coef(pred.logreg)["age"]), digits = 4)
int.min.ci <- suppressMessages(round(
  as.numeric(100 * (exp(confint(pred.logreg)[1,1]) - 1)), digits = 4))
int.max.ci <- suppressMessages(round(
  as.numeric(100 * (exp(confint(pred.logreg)[1,2]) - 1)), digits = 4))
glu.min.ci <- suppressMessages(round(
  as.numeric(100 * (exp(confint(pred.logreg)[2,1]) - 1)), digits = 4))
glu.max.ci <- suppressMessages(round(
  as.numeric(100 * (exp(confint(pred.logreg)[2,2]) - 1)), digits = 4))
bmi.min.ci <- suppressMessages(round(
  as.numeric(100 * (exp(confint(pred.logreg)[3,1]) - 1)), digits = 4))
```

```r
bmi.max.ci <- suppressMessages(round(
  as.numeric(100 * (exp(confint(pred.logreg)[3,2]) - 1)), digits = 4))
ped.min.ci <- suppressMessages(round(
  as.numeric(100 * (exp(confint(pred.logreg)[4,1]) - 1)), digits = 4))
ped.max.ci <- suppressMessages(round(
  as.numeric(100 * (exp(confint(pred.logreg)[4,2]) - 1)), digits = 4))
age.min.ci <- suppressMessages(round(
  as.numeric(100 * (exp(confint(pred.logreg)[5,1]) - 1)), digits = 4))
age.max.ci <- suppressMessages(round(
  as.numeric(100 * (exp(confint(pred.logreg)[5,2]) - 1)), digits = 4))
diab.probs <- predict.glm(pred.logreg, type="response")
thresh <- seq(from=0, to=1, length=10000)
#Empty vector to hold misclassification rates
misclass <- rep(NA,length=length(thresh))
for(i in 1:length(thresh)) {
  #If probability greater than threshold then 1 else 0
  my.classification <- ifelse(diab.probs>thresh[i], 1, 0)

  # calculate the pct where my classification not eq truth
  misclass[i] <- mean(my.classification!=diagnosis$diabetes)
}
#Find threshold which minimizes misclassification
threshold <- thresh[which.min(misclass)]
threshold <- round(threshold, digits = 4)

#how do I do a threshold vs error plot?
thresh.plot <- ggplot() +
  geom_line(aes(x=thresh, y=misclass)) +
  xlab("Threshold") +
  ylab("Error Rate") +
  ggtitle("Threshold vs. Misclassification Rate") +
  geom_vline(xintercept = threshold, col = "red", lwd = 1) +
  annotate("text", x = threshold + 0.2, y = 0.3,
           label = paste("Minimum Threshold =", threshold), col = "red", size = 3.5)
print(thresh.plot)

#ROC (Receiver Operating Characteristic) Curve
my.roc <- roc(diagnosis$diabetes, diab.probs, levels = c(0,1), direction = "<")
thresh.check <- round(auc(my.roc)[1], digits = 4)
#or should I use > instead?
roc.plot <- ggplot() +
  geom_line(aes(x=1-my.roc[["specificities"]], y=my.roc[["sensitivities"]])) +
  geom_abline(intercept=0, slope=1, color ='red') +
  xlab("Percent of True Negatives") +
  ylab("Percent of True Positives") +
  ggtitle("Receiver Operating Characteristic Curve")
print(roc.plot)

#confusion matrix
best.class <- ifelse(diab.probs>threshold, 1, 0)
#table(diagnosis$diabetes, best.class)
confusion.mat <- addmargins(table("Predicted" = best.class, "Actual" = diagnosis$diabetes))
confusion.mat
```

```r
# Choose number of CV studies to run in a loop & test set size
set.seed(86)
n.cv <- 1000
n.test <- round(.1*nrow(diagnosis))

## Set my threshold for classifying
cutoff <- threshold

## Initialize matrices to hold CV results
sens <- rep(NA, n.cv)
spec <- rep(NA, n.cv)
ppv <- rep(NA, n.cv)
npv <- rep(NA, n.cv)
auc <- rep(NA, n.cv)

## Begin for loop
for(cv in 1:n.cv){
  ## Separate into test and training sets
  test.obs <- sample(1:nrow(diagnosis), n.test)
  test.set <- diagnosis[test.obs,]
  train.set <- diagnosis[-test.obs,]

  ## Fit best model to training set
  train.model <- glm(diabetes~glucose+bmi+pedigree+age,data=train.set,family=binomial)

  ## Use fitted model to predict test set
  pred.probs <- predict.glm(train.model,newdata=test.set, type="response")
  #response gives probabilities

  ## Classify according to threshold
  test.class <- ifelse(pred.probs>cutoff, 1, 0)

  ## Create a confusion matrix
  conf.mat <- as.data.frame.matrix(
    addmargins(table("Actual"=factor(
      test.set$diabetes, levels=c(0,1)),
      "pred" = factor(test.class, levels=c(0,1)))))

  ## Pull of sensitivity, specificity, PPV and NPV using bracket notation
  sens[cv] <- conf.mat[2,2]/conf.mat[2,3]
  spec[cv] <- conf.mat[1,1]/conf.mat[1,3]
  ppv[cv] <- conf.mat[2,2]/conf.mat[3,2]
  npv[cv] <- conf.mat[1,1]/conf.mat[3,1]

  ## Calculate AUC
  auc[cv] <- auc(roc(test.set$diabetes, pred.probs, levels = c(0,1), direction = "<"))
} #End for-loop

mean.sens <- round(mean(sens), digits = 4)
mean.spec <- round(mean(spec), digits = 4)
mean.ppv <- round(mean(ppv), digits = 4)
mean.npv <- round(mean(npv), digits = 4)
mean.auc <- round(mean(auc), digits = 4)
```

```
CV.sens <- ggplot() +
  geom_histogram(mapping=aes(x=sens)) +
  xlab('Sensitivity') +
  ylab('Frequency') +
  ggtitle('Diabetes Diagnosis Estimation:',
          subtitle = 'Range of Sensitivity') +
  geom_vline(xintercept = mean.sens, col = "red", lwd = 1)

CV.spec <- ggplot() +
  geom_histogram(mapping=aes(x=spec)) +
  xlab('Specificity') +
  ylab('Frequency') +
  ggtitle('Diabetes Diagnosis Estimation:',
          subtitle = 'Range of Specificity') +
  geom_vline(xintercept = mean.spec, col = "red", lwd = 1)

CV.ppv <- ggplot() +
  geom_histogram(mapping=aes(x=ppv)) +
  xlab('Positive Predictive Value') +
  ylab('Frequency') +
  ggtitle('Diabetes Diagnosis Estimation:',
          subtitle = 'Range of Positive Predictive Values') +
  geom_vline(xintercept = mean.ppv, col = "red", lwd = 1)

CV.npv <- ggplot() +
  geom_histogram(mapping=aes(x=npv)) +
  xlab('Negative Predictive Value') +
  ylab('Frequency') +
  ggtitle('Diabetes Diagnosis Estimation:',
          subtitle = 'Range of Negative Predictive Values') +
  geom_vline(xintercept = mean.npv, col = "red", lwd = 1)

CV.auc <- ggplot() +
  geom_histogram(mapping=aes(x=auc)) +
  xlab('Area Under Curve (AUC) Values') +
  ylab('Frequency') +
  ggtitle('Diabetes Diagnosis Estimation:',
          subtitle = 'Range of AUC Values') +
  geom_vline(xintercept = mean.auc, col = "red", lwd = 1)

suppressMessages(grid.arrange(CV.sens, CV.spec, CV.ppv, CV.npv, CV.auc, nrow=3))
newdf <- data.frame(glucose= 90, bmi= 25.1, pedigree= 1.268, age= 25)
pred.prob <- predict.glm(pred.logreg, newdata=newdf, type="response")
pred.perc <- round(100 * as.numeric(pred.prob), digits = 4)
#End of homework's code
```