Homework 3

Christy Hui

Due 12/07/2024

```r
# read in libraries
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.1
```

```r
library(sjPlot)
```

```
## Warning: package 'sjPlot' was built under R version 4.4.2

## Learn more about sjPlot with 'browseVignettes("sjPlot")'.
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.4.1

## Loading required package: lattice
```

```r
library(effects)
```

```
## Warning: package 'effects' was built under R version 4.4.2

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.4.1

## Use the command
##     lattice::trellis.par.set(effectsTheme())
##   to customize lattice options for effects plots.
## See ?effectsTheme for details.
```

```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.4.1
```

```r
library(nnet)
```

Problem 1

```r
# read in data and make every variable we work with the correct data type
liver = read.csv("liver23.csv")
liver$tx_fail = ifelse(liver$tx_fail == 0, "Success", "Failure")
liver$tx_fail = factor(liver$tx_fail, levels = c("Success", "Failure"))
liver$hgt_cm_don_calc.x = as.numeric(liver$hgt_cm_don_calc.x)
liver$bmi_don_calc.x = as.numeric(liver$bmi_don_calc.x)
liver$coronary_angio_don.x = as.factor(liver$coronary_angio_don.x)
liver$hist_hypertens_don.x = as.factor(liver$hist_hypertens_don.x)
liver$ethnicity_don = as.factor(liver$ethnicity_don)
```

Part 1

```
# create table of binary variable
table(liver$tx_fail)
```

```
##
## Success Failure
##   10000    3222
```

```
# show proportion as well
prop.table(table(liver$tx_fail))
```

```
##
##   Success   Failure
## 0.7563152 0.2436848
```

Around 24% of the data are failures, whereas around 76% are successes. While not ideal, this spread is not bad.

```
# create table of values for categorical variables
table(liver$coronary_angio_don.x)
```

```
##
##     N     U     Y
## 10613   633  1976
```

```
# show proportion as well
prop.table(table(liver$coronary_angio_don.x))
```

```
##
##          N          U          Y
## 0.80267736 0.04787475 0.14944789
```

Around 80% of the data are “No.” This may prove to be a challenge, as a large portion of the data is “No.”

```
# create table of values for categorical variables
table(liver$hist_hypertens_don.x)
```

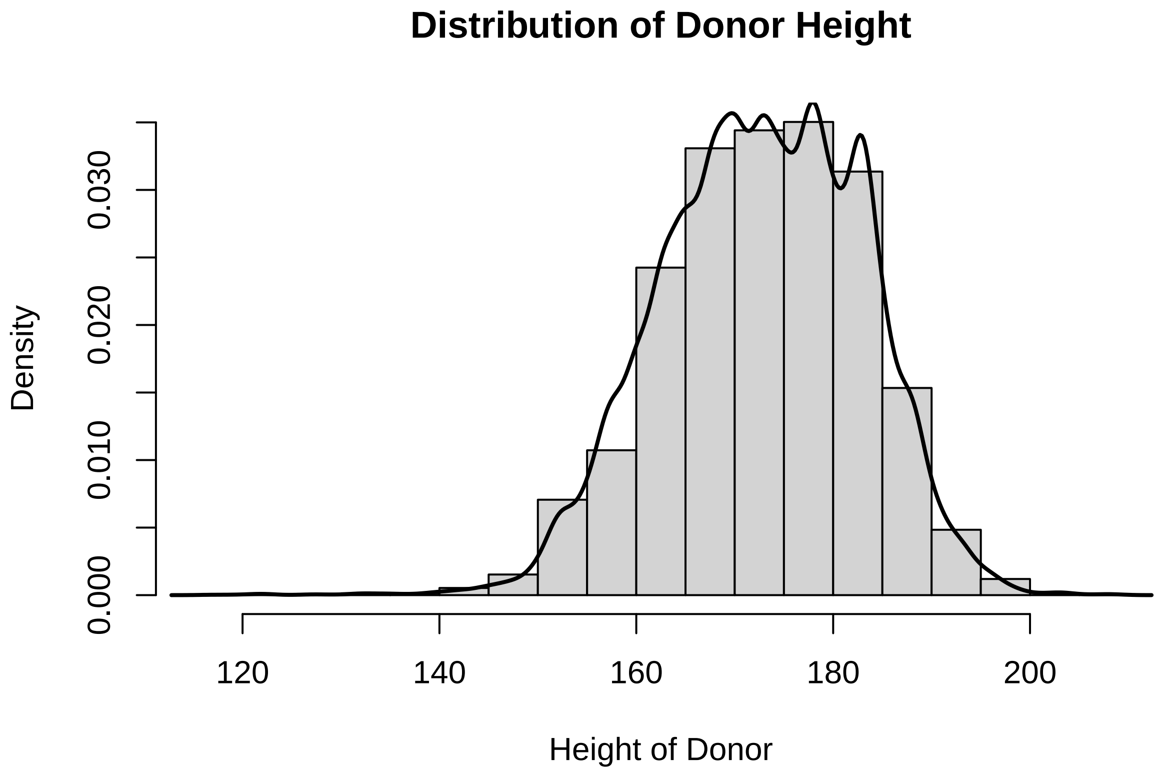
```
##
##     N     U     Y
## 10414   145  2663
```

```
# show proportion as well
prop.table(table(liver$hist_hypertens_don.x))
```

```
##
##          N          U          Y
## 0.78762668 0.01096657 0.20140675
```

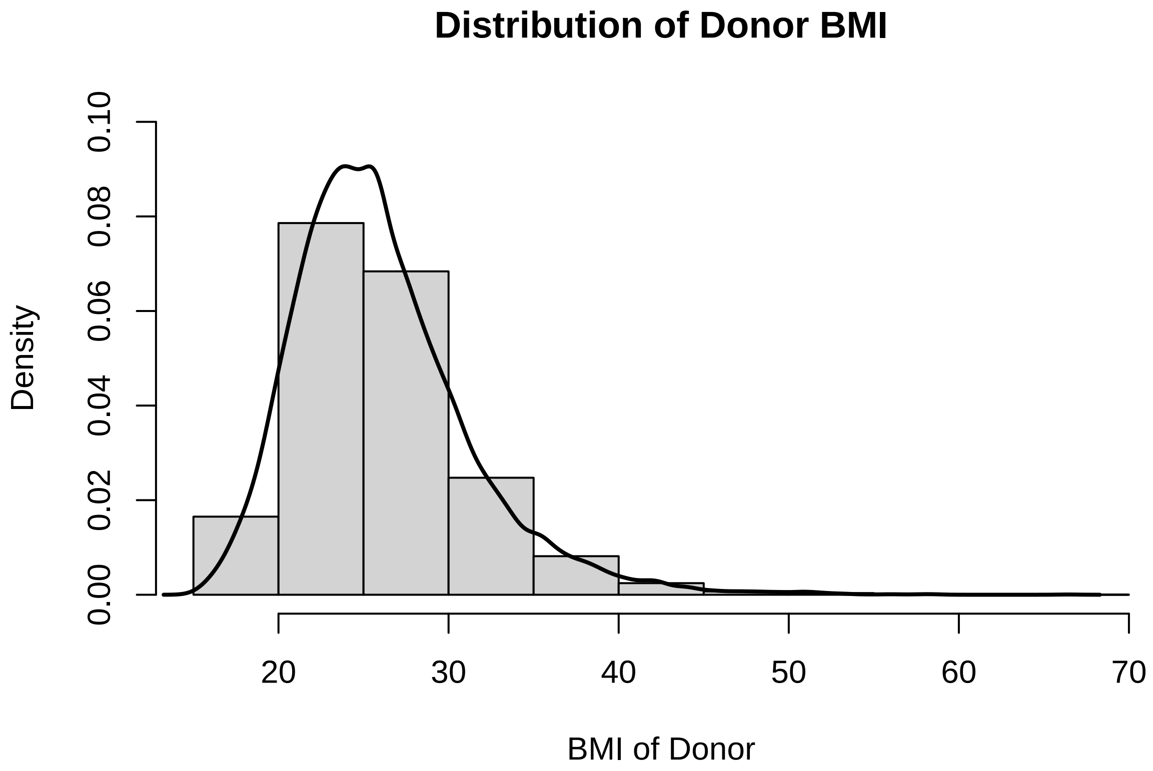
Around 79% of the data are “No.” This may prove to be a challenge, as a large portion of the data is “No.”

```
# create histogram of height values
hist(liver$hgt_cm_don_calc.x, xlab = "Height of Donor", main = "Distribution of Donor Height", freq = FA
lines(density(liver$hgt_cm_don_calc.x), lwd = 2)
```



The distribution of height seems to be, mostly, normal, which is promising.

```
# create histogram of BMI values
hist(liver$bmi_don_calc.x, xlab = "BMI of Donor", main = "Distribution of Donor BMI", freq = FALSE, yli
lines(density(liver$bmi_don_calc.x), lwd = 2)
```



The distribution of BMI seems to be right skew. This is a bit expected. This isn't too big of a deal, as it is not extremely right skew.

Part 2

```
# run logistic model to predict failure
liver_glm = glm(tx_fail ~ coronary_angio_don.x + hist_hypertens_don.x + hgt_cm_don_calc.x + bmi_don_cal
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(liver_glm)
```

```
##
## Call:
## glm(formula = tx_fail ~ coronary_angio_don.x + hist_hypertens_don.x +
##     hgt_cm_don_calc.x + bmi_don_calc.x, family = "binomial",
##     data = liver)
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            2.671619   0.422978   6.316 2.68e-10 ***
## coronary_angio_don.xU 18.578219 146.413996   0.127    0.899
## coronary_angio_don.xY -0.714100   0.072139  -9.899  < 2e-16 ***
## hist_hypertens_don.xU 19.046089 307.543642   0.062    0.951
## hist_hypertens_don.xY  1.320632   0.053513  24.679  < 2e-16 ***
## hgt_cm_don_calc.x     -0.026568   0.002343 -11.340  < 2e-16 ***
## bmi_don_calc.x         0.010079   0.004400   2.291    0.022 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14684  on 13221  degrees of freedom
## Residual deviance: 11452  on 13215  degrees of freedom
## AIC: 11466
##
## Number of Fisher Scoring iterations: 16
```

Part 3

```
# exponentiate coefficients
cbind(Estimate = exp(coef(liver_glm)))
```

```
##                             Estimate
## (Intercept)             1.446336e+01
## coronary_angio_don.xU 1.170625e+08
## coronary_angio_don.xY 4.896328e-01
## hist_hypertens_don.xU 1.869010e+08
## hist_hypertens_don.xY 3.745786e+00
## hgt_cm_don_calc.x     9.737820e-01
## bmi_don_calc.x        1.010130e+00
```

Keeping all else constant, the odds of a kidney transplant failure is 1.170625*10^8 times more likely for donors where it was unknown whether or not they died from coronary heart disease than donors who did not die of coronary heart disease.

Keeping all else constant, the odds of a kidney transplant failure is 51.03672% (1-4.896328e-01) less likely for donors who died from coronary heart disease than donors who did not die of coronary heart disease.

Keeping all else constant, the odds of a kidney transplant failure is 1.869010*10^8 times more likely for donors where it was unknown whether or not they had a history of hypertension than donors who did not have a history of hypertension.

Keeping all else constant, the odds of a kidney transplant failure is 3.745786 times more likely for donors who had a history of hypertension than donors who did not have a history of hypertension.

Keeping all else constant, for one unit increase in the height of the donor, the odds of a kidney transplant failure decreases by 0.026218 (1-0.9737820) units.

Keeping all else constant, for one unit increase in the BMI of the donor, the odds of a kidney transplant failure increases by 1.010130 units.

Part 4

```
# exponentiate confidence intervals
suppressWarnings(exp(confint(liver_glm))) # had to run suppress warning because there were too many num
```

```
## Waiting for profiling to be done...
```

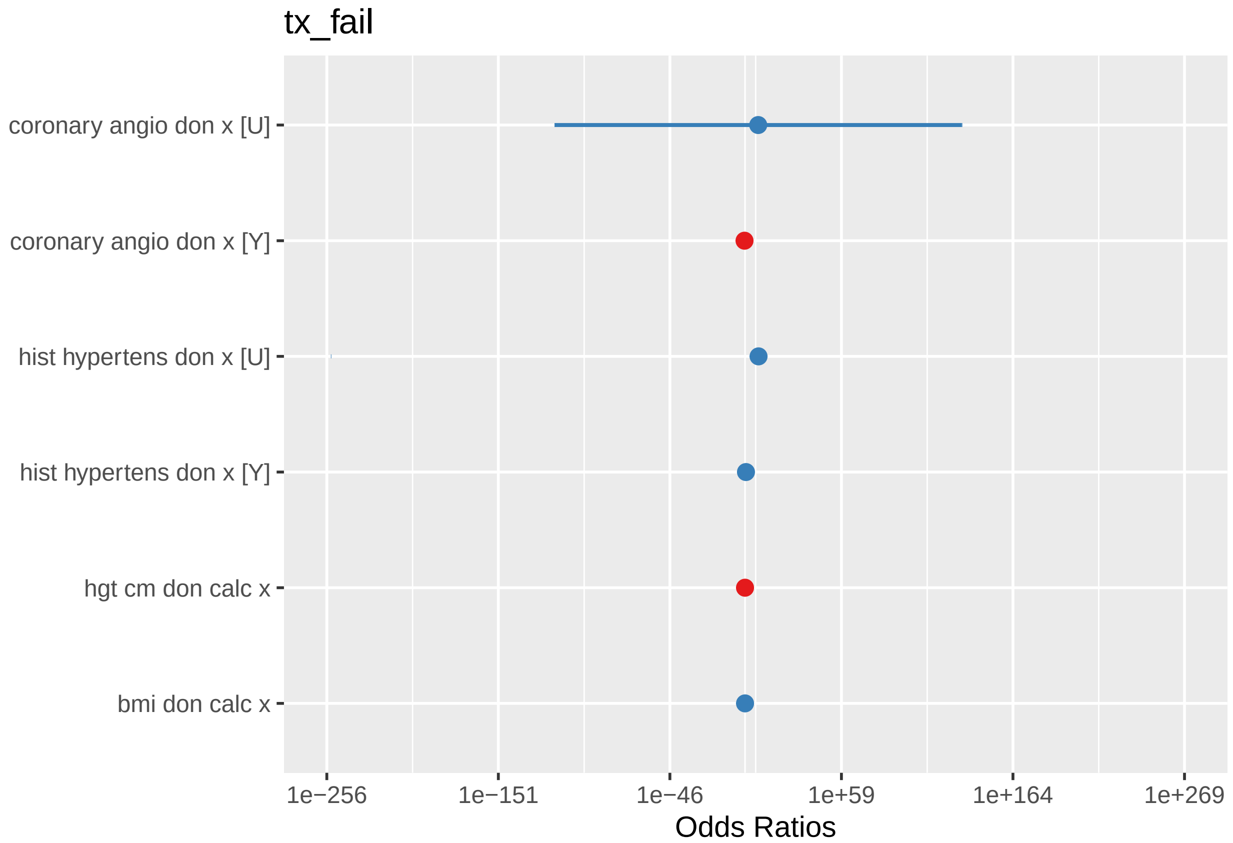
```
##                                2.5 %       97.5 %
## (Intercept)           6.314323e+00 3.315056e+01
## coronary_angio_don.xU 1.333328e+51 2.729940e+52
## coronary_angio_don.xY 4.243885e-01 5.631302e-01
## hist_hypertens_don.xU 5.202949e+21 4.012661e+61
## hist_hypertens_don.xY 3.372805e+00 4.160086e+00
## hgt_cm_don_calc.x      9.693146e-01 9.782585e-01
## bmi_don_calc.x         1.001431e+00 1.018855e+00
```

Part 5

Predictor	Odds	2.5%	97.5%	P-Value
Height	0.9737820	0.9693146	0.9782585	< 2e-16
BMI	1.010130	1.001431	1.018855	0.022
Coronary Angio Unknown	1.170625*10^8	1.333328-10^51	2.729940*10^52	0.899
Coronary Angio Yes	0.4896328	0.4243885	0.5631302	< 2e-16
Hypertension Unknown	1.869010*10^8	5.202949*10^21	4.012661*10^61	0.951
Hypertension Yes	3.745786	3.372805	4.160086	< 2e-16

Part 6

```r
# draw plot of odds
plot_model(liver_glm, p.val = "wald")
```



Part 7

Regarding the odds and confidence intervals, we only interpret the odds that are statistically significant. In this case, that would be the height, the state of having coronary heart disease, and having a history of hypertension. The odds for height and having coronary heart disease is less than 1. This means that height and having coronary disease is negatively associated with the failure of kidney transplants (albeit height is very close to 1). This is further supported by the confidence interval being under 1. Having hypertension is positively associated with the failure of kidney transplants, as evidenced from a higher odds ratio and 95% confidence interval being higher than 1.

Part 8

The null deviance shows how well our response is predicted by the model with just the intercept. The residual deviance, on the other hand, shows how well our response is predictors with the predictors in the model. The lower our scores, the better our models are.

In this case, our null deviance is 14684 on 13221 degrees of freedom and our residual deviance is 11452 on 13215 degrees of freedom. Since our residual deviance is lower than our null deviance, we can say our model does better with the predictors than without the predictors.

Part 9

```
# calculate pseudo r-squared by following the formula: 1-(residual deviance/null deviance)
1-(liver_glm$deviance/liver_glm$null.deviance)
```

```
## [1] 0.2201033
```

Part 10

```
# make predictions binary using a 0.50% probability boundary
predictions = ifelse(predict(liver_glm) > 0.5, "Failure", "Success")
predictions = factor(predictions, levels = c("Success", "Failure"))

# craft confusion matrix
confusionMatrix(liver$tx_fail, predictions)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Success Failure
##    Success   10000       0
##    Failure    2448     774
##
##                Accuracy : 0.8149
##                  95% CI : (0.8081, 0.8214)
##     No Information Rate : 0.9415
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3235
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.8033
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 0.2402
##              Prevalence : 0.9415
##          Detection Rate : 0.7563
##    Detection Prevalence : 0.7563
##       Balanced Accuracy : 0.9017
##
##        'Positive' Class : Success
```

```
##
```

Part 11

```
# carry out 5-fold cross validation
set.seed(213)
train_control = trainControl(method = "cv", number = 5, classProbs = TRUE)
liver_cvglm = train(tx_fail ~ coronary_angio_don.x + hist_hypertens_don.x + hgt_cm_don_calc.x + bmi_don
                    data = liver,
                    method = "glm",
                    family = binomial,
                    trControl = train_control)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
# report results
liver_cvglm$results
```

```
##   parameter  Accuracy     Kappa  AccuracySD    KappaSD
## 1      none 0.8178038 0.3640265 0.001976407 0.01240588
```

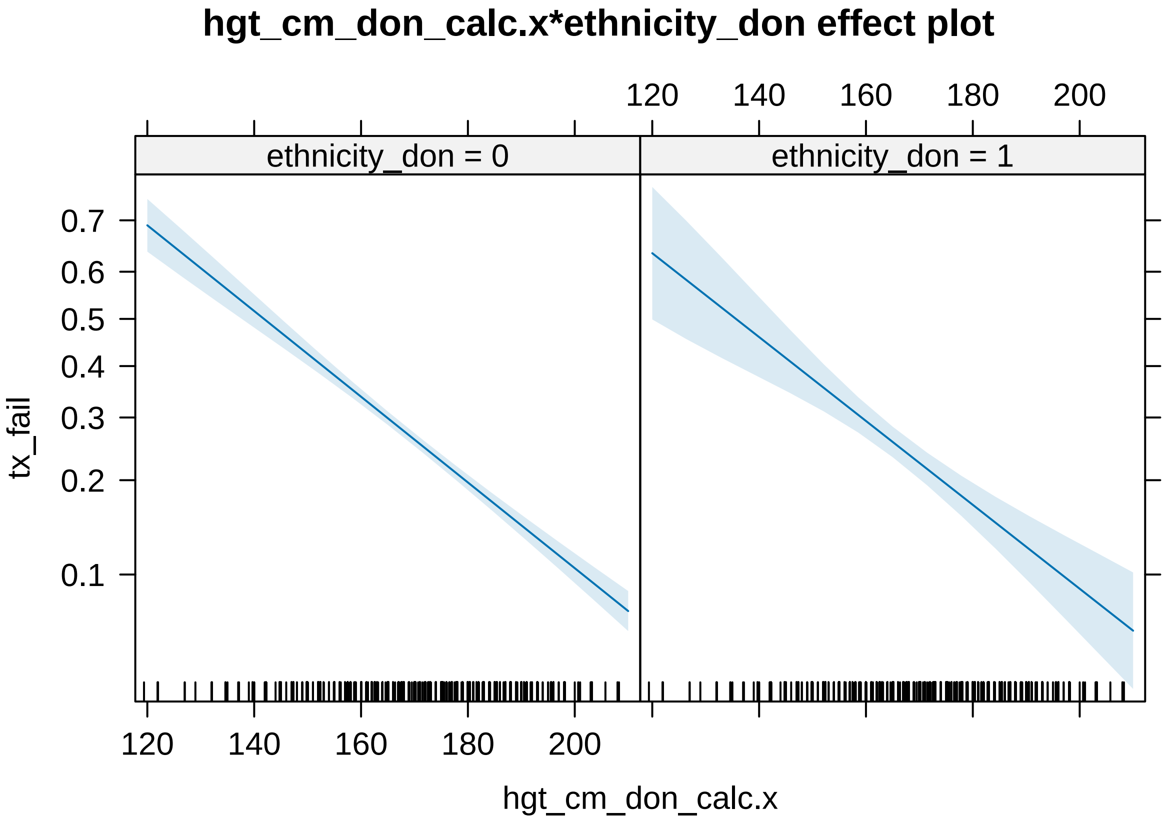
```
# confusion matrix of glm model
confusionMatrix(predict(liver_cvglm), liver$tx_fail)
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction Success Failure
##    Success    9850    2253
##    Failure     150     969
##
##                Accuracy : 0.8183
##                  95% CI : (0.8116, 0.8248)
##     No Information Rate : 0.7563
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3669
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9850
##             Specificity : 0.3007
##          Pos Pred Value : 0.8138
##          Neg Pred Value : 0.8660
##              Prevalence : 0.7563
##          Detection Rate : 0.7450
##    Detection Prevalence : 0.9154
##       Balanced Accuracy : 0.6429
##
##        'Positive' Class : Success
##
```

The accuracy is around 81%, which is quite good and much better than simply guessing whether or not a kidney transplant would fail or not (50%). The sensitivity is quite high, which indicates that our model is able to perform when guessing true positives. However, the model struggles with true negatives, due to a low specificity rate.

Part 12

```r
# replicate plot figure from homework
m1 = glm(tx_fail ~ hgt_cm_don_calc.x*ethnicity_don, data = liver, family="binomial")
plot(allEffects(m1))
```



Part A What is the relationship between height and transplant failure among ethnic groups? How does the height of the donor and interaction of ethnicity influence the likelihood of a kidney transplant failure?

Part B As the plot indicates, the patterns follow a very similar downward trend for ethnicity and transplant failure. Since both have similar negative slopes, we have a reason to believe that the relationship between donor height and transplant failure is consistent across both ethnic groups. As such, we have a reason to believe that there may be a nonexistent interaction effect.

Part 13

```r
# create new model by adding interaction effect
liver_glm2 = glm(tx_fail ~ coronary_angio_don.x + hist_hypertens_don.x + hgt_cm_don_calc.x + bmi_don_cal
                 data = liver,
                 family = "binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(liver_glm2)
```

```
##
## Call:
## glm(formula = tx_fail ~ coronary_angio_don.x + hist_hypertens_don.x +
##     hgt_cm_don_calc.x + bmi_don_calc.x + hgt_cm_don_calc.x *
##     ethnicity_don, family = "binomial", data = liver)
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        2.890760   0.453068   6.380 1.77e-10 ***
## coronary_angio_don.xU             18.578084 146.431474   0.127   0.8990
## coronary_angio_don.xY             -0.715012   0.072131  -9.913  < 2e-16 ***
## hist_hypertens_don.xU             19.042195 307.387135   0.062   0.9506
## hist_hypertens_don.xY              1.316722   0.053600  24.565  < 2e-16 ***
## hgt_cm_don_calc.x                 -0.027789   0.002524 -11.012  < 2e-16 ***
## bmi_don_calc.x                     0.010206   0.004398   2.321   0.0203 *
## ethnicity_don1                    -1.256240   1.191561  -1.054   0.2918
## hgt_cm_don_calc.x:ethnicity_don1   0.006946   0.007039   0.987   0.3238
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14684  on 13221  degrees of freedom
## Residual deviance: 11450  on 13213  degrees of freedom
## AIC: 11468
##
## Number of Fisher Scoring iterations: 16
```

Part 14

```r
# perform ANOVA to see if models are different
anova(liver_glm, liver_glm2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: tx_fail ~ coronary_angio_don.x + hist_hypertens_don.x + hgt_cm_don_calc.x +
##     bmi_don_calc.x
## Model 2: tx_fail ~ coronary_angio_don.x + hist_hypertens_don.x + hgt_cm_don_calc.x +
##     bmi_don_calc.x + hgt_cm_don_calc.x * ethnicity_don
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1     13215      11452
## 2     13213      11450  2   2.4547   0.2931
```

After running the ANOVA, we notice the p-value is greater than 0.05. As such, we fail to reject the null and state that we have reason to believe that the failure of a kidney transplant is not dependent on the combined effect of height and ethnicity. We recommend using the initial model.

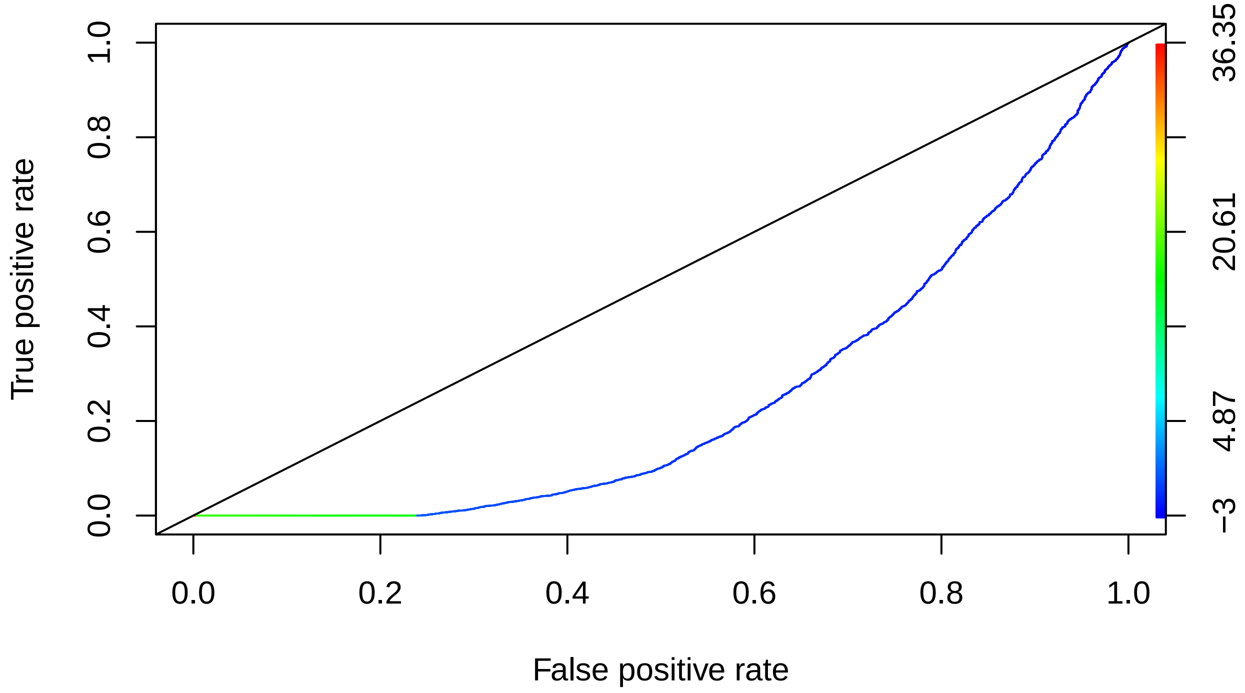
Part 15

```r
library(ROCR)
```

```
## Warning: package 'ROCR' was built under R version 4.4.2
```

```
# create vector of predictions for ROC curve
probpredictions = predict(liver_glm)
pred_liver_glm <- prediction(probpredictions, liver$tx_fail)

# initialize roc curve and plot
roc_curve <- performance(pred_liver_glm, "tpr","fpr")
plot(roc_curve, colorize=TRUE, col="blue")
abline(0, 1)
```



As shown in the above graphs, our model struggles to correctly classify true positives without a significant increase in false positives. This model seems to have low predictive power.

Problem 2

```
# read in diabetic csv
diabetic = read.csv("diabetic.csv")
```

Part 1

```
# recode total risk factor variable
diabetic$TotalRiskFactors = recode(diabetic$TotalRiskFactors, "
                                   0 = 'None';
                                   1 = 'One';
                                   2 = '2 or More';
                                   3 = '2 or More';
                                   4 = '2 or More';
                                   5 = '2 or More';
                                   6 = '2 or More';
                                   7 = '2 or More';
                                   8 = '2 or More';
                                   ")
diabetic$TotalRiskFactors = factor(diabetic$TotalRiskFactors, levels = c("None", "One", "2 or More"))
```

Part A

```
# recode diabetes variable
diabetic$Diabetes.new = recode(diabetic$Diabetes.new, "
                               0 = 'No';
                               1 = 'Yes';
                               ")
diabetic$Diabetes.new = as.factor(diabetic$Diabetes.new)
```

Part B

```
# recode smoking status variable
diabetic$SmokingStatus_NISTCode = recode(diabetic$SmokingStatus_NISTCode, "
                                         'FORMER' = 'FormerSmoker';
                                         'TRUE' = 'Smoker';
                                         'FALSE' = 'NonSmoker';
                                         ")
diabetic$SmokingStatus_NISTCode = factor(diabetic$SmokingStatus_NISTCode, levels = c("NonSmoker", "Forme
```

Part C

```
# recode age variable
diabetic$age.new = factor(diabetic$age.new, levels = c("low", "medium", "high"))
```

Part D

```
# clean diabetic data set from NAs
diabetic = diabetic[, c("age.new", "TotalRiskFactors", "Diabetes.new", "HypertensionDX", "SmokingStatus_
diabetic = na.omit(diabetic)
```

```
# make hypertensionDX factor
diabetic$HypertensionDX = as.factor(diabetic$HypertensionDX)
```

```
# create contingency table of age.new variable
table(diabetic$age.new)
```

Part E

```
##
##    low medium   high
##   1239   1427   1649
```

```
# create contingency table of TotalRiskFactors variable
table(diabetic$TotalRiskFactors)
```

```
##
##      None       One 2 or More
##      1790      1693       832
```

```
# create contingency table of Diabetes.new variable
table(diabetic$Diabetes.new)
```

```
##
##   No  Yes
## 3462  853
```

```
# create contingency table of HypertensionDX variable
table(diabetic$HypertensionDX)
```

```
##
##   no  yes
## 2267 2048
```

```
# create contingency table of SmokingStatus_NISTCode variable
table(diabetic$SmokingStatus_NISTCode)
```

```
##
##    NonSmoker FormerSmoker       Smoker
##         1867         1847          601
```

Part 2

```
# run multinomial model
diabetic_multi = multinom(age.new ~ TotalRiskFactors + Diabetes.new + HypertensionDX + SmokingStatus_NIS
```

```
## # weights:  24 (14 variable)
## initial  value 4740.512026
## iter  10 value 4260.991992
## iter  20 value 4108.350866
## final  value 4108.350692
## converged
```

```
summary(diabetic_multi)
```

```
## Call:
## multinom(formula = age.new ~ TotalRiskFactors + Diabetes.new +
##     HypertensionDX + SmokingStatus_NISTCode, data = diabetic)
##
## Coefficients:
##        (Intercept) TotalRiskFactorsOne TotalRiskFactors2 or More
## medium  -0.5158245           0.7201609                 0.6742832
## high    -0.8435857           0.9580151                 1.0635043
##        Diabetes.newYes HypertensionDXyes SmokingStatus_NISTCodeFormerSmoker
## medium       0.8848004          1.336294                         -0.3895716
## high         1.1620844          2.064695                         -0.6001646
##        SmokingStatus_NISTCodeSmoker
## medium                   -0.2048692
## high                     -1.2855500
##
```

```
## Std. Errors:
##        (Intercept) TotalRiskFactorsOne TotalRiskFactors2 or More
## medium  0.07981794          0.09496856                 0.1228790
## high    0.08530449          0.09988707                 0.1253497
##        Diabetes.newYes HypertensionDXyes SmokingStatus_NISTCodeFormerSmoker
## medium       0.1487110        0.09591240                         0.09115258
## high         0.1468778        0.09752606                         0.09318094
##        SmokingStatus_NISTCodeSmoker
## medium                    0.1209898
## high                      0.1420562
##
## Residual Deviance: 8216.701
## AIC: 8244.701
```

Part 3

```
# exponentiate odds
t(exp(coef(diabetic_multi)))
```

Exponentiated (Odds)

```
##                                       medium      high
## (Intercept)                        0.5970082 0.4301653
## TotalRiskFactorsOne                2.0547638 2.6065176
## TotalRiskFactors2 or More          1.9626257 2.8965034
## Diabetes.newYes                    2.4225008 3.1965893
## HypertensionDXyes                  3.8049162 7.8828954
## SmokingStatus_NISTCodeFormerSmoker 0.6773470 0.5487213
## SmokingStatus_NISTCodeSmoker       0.8147539 0.2764985
```

Keeping all else constant, the odds of a being medium age, compared to being low age, is 2.0547638 times more likely for people where they had one risk factor than people who did not have any risk factors.

Keeping all else constant, the odds of a being high age, compared to being low age, is 2.6065176 times more likely for people where they had one risk factor than people who did not have any risk factors.

Keeping all else constant, the odds of a being medium age, compared to being low age, is 1.9626257 times more likely for people where they had 2 or more risk factors than people who did not have any risk factors.

Keeping all else constant, the odds of a being high age, compared to being low age, is 2.8965034 times more likely for people where they had 2 or more risk factors than people who did not have any risk factors.

Keeping all else constant, the odds of a being medium age, compared to being low age, is 2.4225008 times more likely for people who have diabetes than people who do not have diabetes.

Keeping all else constant, the odds of a being high age, compared to being low age, is 3.1965893 times more likely for people who have diabetes than people who do not have diabetes.

Keeping all else constant, the odds of a being medium age, compared to being low age, is 3.8049162 times more likely for people who have a history of hypertension than people who do not have a history of hypertension.

Keeping all else constant, the odds of a being high age, compared to being low age, is 7.8828954 times more likely for people who have a history of hypertension than people who do not have a history of hypertension.

Keeping all else constant, the odds of a being medium age, compared to being low age, is 0.6773470 times more likely for people who have a history of being a former smoker than people who have never smoked before.

Keeping all else constant, the odds of a being high age, compared to being low age, is 0.5487213 times more likely for people who have a history of being a former smoker than people who have never smoked before.

Keeping all else constant, the odds of a being medium age, compared to being low age, is 0.8147539 times more likely for people who are smokers than people who have never smoked before.

Keeping all else constant, the odds of a being high age, compared to being low age, is 0.2764985 times more likely for people who are smokers than people who have never smoked before.

```
# exponentiate confidence intervals for medium age level
exp(confint(diabetic_multi))[, , "medium"]
```

```
##                                        2.5 %    97.5 %
## (Intercept)                        0.5105510 0.6981061
## TotalRiskFactorsOne                1.7057859 2.4751373
## TotalRiskFactors2 or More          1.5425623 2.4970789
## Diabetes.newYes                    1.8100098 3.2422532
## HypertensionDXyes                  3.1528573 4.5918309
## SmokingStatus_NISTCodeFormerSmoker 0.5665288 0.8098423
## SmokingStatus_NISTCodeSmoker       0.6427466 1.0327926
```

```
# exponentiate confidence intervals for high age level
exp(confint(diabetic_multi))[, , "high"]
```

```
##                                        2.5 %    97.5 %
## (Intercept)                        0.3639352 0.5084482
## TotalRiskFactorsOne                2.1430714 3.1701856
## TotalRiskFactors2 or More          2.2655634 3.7031547
## Diabetes.newYes                    2.3969790 4.2629421
## HypertensionDXyes                  6.5113559 9.5433334
## SmokingStatus_NISTCodeFormerSmoker 0.4571262 0.6586696
## SmokingStatus_NISTCodeSmoker       0.2093024 0.3652676
```

Part 4

```
# find non exponentiated odds to fill in table
t(coef(diabetic_multi))
```

Non Exponentiated Odds (Log of Odds)

```
##                                         medium       high
## (Intercept)                        -0.5158245 -0.8435857
## TotalRiskFactorsOne                 0.7201609  0.9580151
## TotalRiskFactors2 or More           0.6742832  1.0635043
## Diabetes.newYes                     0.8848004  1.1620844
## HypertensionDXyes                   1.3362940  2.0646953
## SmokingStatus_NISTCodeFormerSmoker -0.3895716 -0.6001646
## SmokingStatus_NISTCodeSmoker       -0.2048692 -1.2855500
```

```
z_scores = summary(diabetic_multi)$coefficients/summary(diabetic_multi)$standard.errors
p_values = 2 * (1 - pnorm(abs(z_scores)))
t(p_values)
```

Calculate P-Values

```
##                                           medium         high
## (Intercept)                          1.029785e-10 0.000000e+00
## TotalRiskFactorsOne                  3.375078e-14 0.000000e+00
## TotalRiskFactors2 or More            4.079465e-08 0.000000e+00
## Diabetes.newYes                      2.684713e-09 2.442491e-15
## HypertensionDXyes                    0.000000e+00 0.000000e+00
## SmokingStatus_NISTCodeFormerSmoker 1.921347e-05 1.188052e-10
## SmokingStatus_NISTCodeSmoker         9.040271e-02 0.000000e+00
```

Medium

Predictor	Odds	Log of Odds	97.5%	P-Value
One Risk vs None	2.0547638	0.7201609	2.4751373	3.375078e-14
Two or More Risks vs None	1.9626257	0.6742832	2.4970789	4.079465e-08
Has Diabetic (Yes vs No)	2.4225008	0.8848004	3.2422532	2.684713e-09
HypertensionDX (Yes vs No)	3.8049162	1.3362940	4.5918309	0
Former Smoker vs Non Smoker	0.6773470	-0.3895716	0.8098423	1.921347e-05
Smoker vs Non Smoker	0.8147539	-0.2048692	1.0327926	9.040271e-02

High

Predictor	Odds	Log of Odds	97.5%	P-Value
One Risk vs None	2.6065176	0.9580151	3.1701856	0
Two or More Risks vs None	2.8965034	1.0635043	3.7031547	0
Has Diabetic (Yes vs No)	3.1965893	1.1620844	4.2629421	2.442491e-15
HypertensionDX (Yes vs No)	7.8828954	2.0646953	9.5433334	0
Former Smoker vs Non Smoker	0.5487213	-0.6001646	0.6586696	1.188052e-10
Smoker vs Non Smoker	0.2764985	-1.2855500	0.3652676	0

Part 5

Part A Assuming a high risk factor is "Two or More Risks vs None:"

- People who have a higher risk factor have 1.9626257 times the odds of being a medium age vs a low age compared to people who have no risk.
- People who have a higher risk factor have 2.8965034 times the odds of being a high age vs a low age compared to people who have no risk.

Part B

- People who are diabetic have 2.4225008 times the odds of being a medium age vs a low age compared to people who are not diabetic.
- People who are diabetic have 3.1965893 times the odds of being a high age vs a low age compared to people who are not diabetic.

Part C

- People who are smokers have 0.8147539 times the odds of being a medium age vs a low age compared to people who are non smokers.
- People who are smokers have 0.2764985 times the odds of being a high age vs a low age compared to people who are non smokers.

Part D The 95% confidence interval is over 1 for both medium and high ages regarding the odds of diabetic type II. Since the 95% confidence interval is greater than 1, having diabetes, compared to not having diabetes, increase the odds of being in the medium and high age groups.

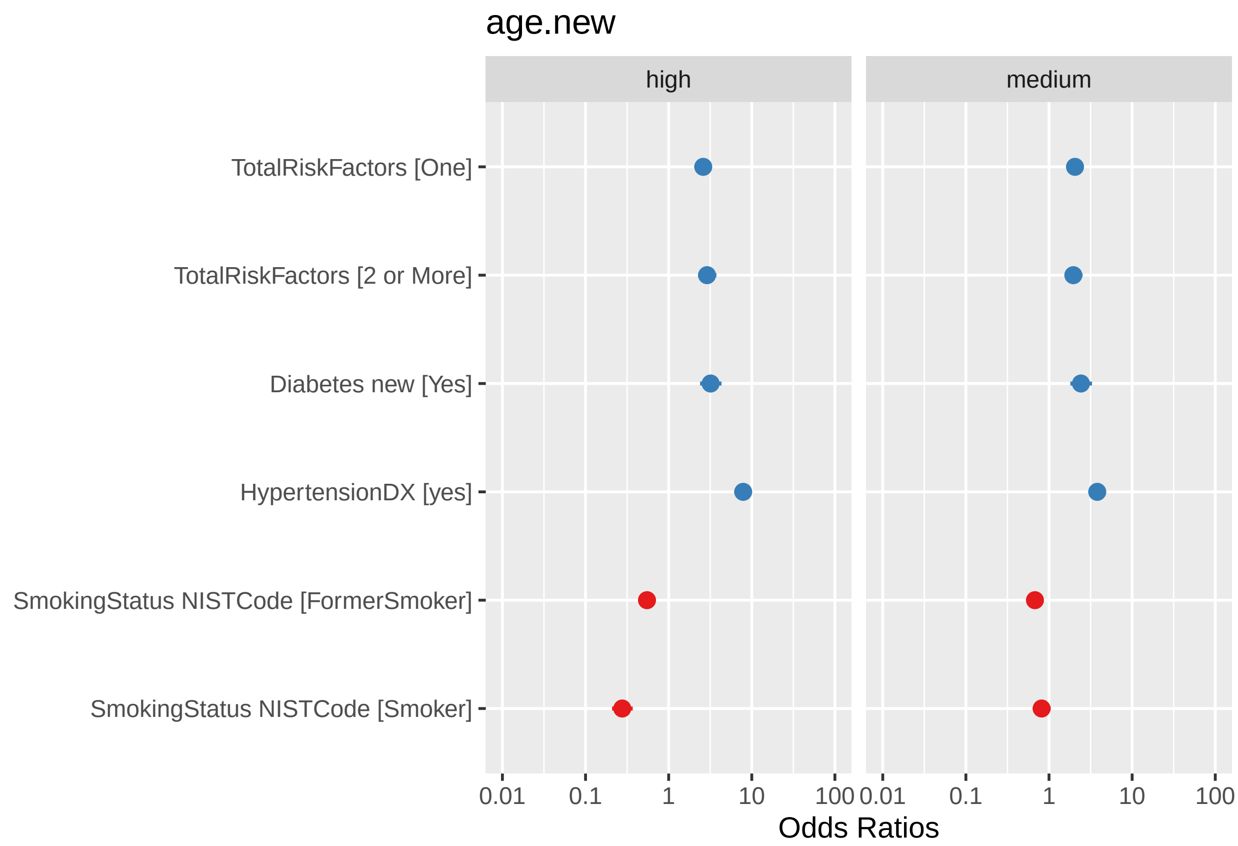
Part E The 95% confidence interval is under 1 for both medium and high ages regarding the odds of being a smoker. Since the 95% confidence interval is less than 1, being a smoker, compared to not being a smoker, recudes the odds of being in the medium and high age groups.

Part 6

$$P(HighAgeGroup) = \frac{1}{e^{-(-0.8435857+0.9580151*OneRisk+1.0635043*TwoOrMoreRisks+1.1620844*HasDiabeticYes+2.0646953*Hypertension}}$$

Part 7

```r
plot_model(diabetic_multi)
```



Part 8

For the high age group, the total risk factors (one and 2 or more compared to having no risk factors), having diabetes (compared to not having diabetes), and having high hypertension (compared to not having a history of high hypertension) are positively associated with age (as shown by the odds ratios being higher than 1). On the other hand, being a former smoker and being a smoker (compared to not having a history of smoking) are negative associated.

Part 9

```
predictions = predict(diabetic_multi, diabetic)
confusionMatrix(predictions, diabetic$age.new)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  low medium high
##     low     916    513  282
##     medium  126    260  226
##     high    197    654 1141
##
## Overall Statistics
##
##                Accuracy : 0.537
##                  95% CI : (0.5219, 0.5519)
##     No Information Rate : 0.3822
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3014
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: low Class: medium Class: high
## Sensitivity              0.7393       0.18220      0.6919
## Specificity              0.7415       0.87812      0.6808
## Pos Pred Value           0.5354       0.42484      0.5728
## Neg Pred Value           0.8760       0.68485      0.7813
## Prevalence               0.2871       0.33071      0.3822
## Detection Rate           0.2123       0.06025      0.2644
## Detection Prevalence     0.3965       0.14183      0.4616
## Balanced Accuracy        0.7404       0.53016      0.6864
```

Part 10

The model performs better than randomly guessing or guessing with no information (38.22%) due to the accuracy rate of 53.7%. This is further exemplified by our p-value of less than 0.05, which confirms that our model is statistically significant. The model guesses low and high true positives decently well, as shown by the relatively high (73.93% and 69.19%) sensitivity rates. The model also guesses low and high true negatives decently well, as shown by the relatively high (74.15% and 68.08%) specificity rates. It guesses true negatives for mediums very well, as shown by the 87.81% specificity rate. However, the model performs terrible for true positives, as shown by the 18.22% sensitivity rating. All in all, the model performs decently, as shown by the accuracy rating (and further exemplified by the sensitivity and specificity rates). However, the model struggles to distinguish medium ages. This may be due to an overlap in features between the low and high age categories.