Nayak Project 2 - Final

Contents

Project 2

```
library(fivethirtyeight)
library(ggplot2)
library(car)
library(plotROC)
library(lmtest)
library(sandwich)
```

- **0. (5 pts)** Introduce your dataset and each of your variables (or just your main variables if you have lots) in a paragraph. What are they measuring? How many observations?

```
view(biopics)

biop <- biopics %>%
  na.omit(biopics) %>%
  select(-site) %>%
  view()
```

For this project I will be using the biopics dataset. The dataset lists biographical films (title), their box office numbers (box_office), country of origin (country), year the film was released (year_release), the type or profession of the subject being depicted (type_of_subject), the race of the subject (subject_race), whether or not the subject was a person of color (person_of_color), and if they were a male or female (subject_sex). This dataset wanted to analyze the number of biographical films released about people of color as well as their success in the box office. The NAs as well as the site variable were removed from the original dataset. The site variable indicated the IMDB page for the film for reference.

- **1. (15 pts)** Perform a MANOVA testing whether any of your numeric variables (or a subset of them, if including them all doesn't make sense) show a mean difference across levels of one of your categorical variables (3). If they do, perform univariate ANOVAs to find response(s) showing a mean difference across groups (3), and perform post-hoc t tests to find which groups differ (3). Discuss the number of tests you have performed, calculate the probability of at least one type I error (if unadjusted), and adjust the significance level accordingly (bonferroni correction) before discussing significant differences (3). Briefly discuss assumptions and whether or not they are likely to have been met (2).

```
#MANOVA
manova1 <- manova(cbind(box_office,year_release)~person_of_color, data = biop)
summary(manova1)
```

```
## Df Pillai approx F num Df den Df Pr(>F)
## person_of_color 1 0.020708 3.3093 2 313 0.03783 *
## Residuals 314
## ---
```

```
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

```
#Univariate ANOVAs
summary.aov(manova1)
```

```
## Response box_office :
## Df Sum Sq Mean Sq F value Pr(>F)
## person_of_color 1 6.2765e+15 6.2765e+15 5.5417 0.01918 *
## Residuals 314 3.5563e+17 1.1326e+15
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Response year_release :
## Df Sum Sq Mean Sq F value Pr(>F)
## person_of_color 1 251 251.43 1.3302 0.2496
## Residuals 314 59349 189.01
```

```
#5 post-hoc t-tests
pairwise.t.test(biop$box_office, biop$person_of_color, p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  biop$box_office and biop$person_of_color
##
##      FALSE
## TRUE 0.019
##
## P value adjustment method: none
```

```
pairwise.t.test(biop$year_release, biop$person_of_color, p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  biop$year_release and biop$person_of_color
##
##      FALSE
## TRUE 0.25
##
## P value adjustment method: none
```

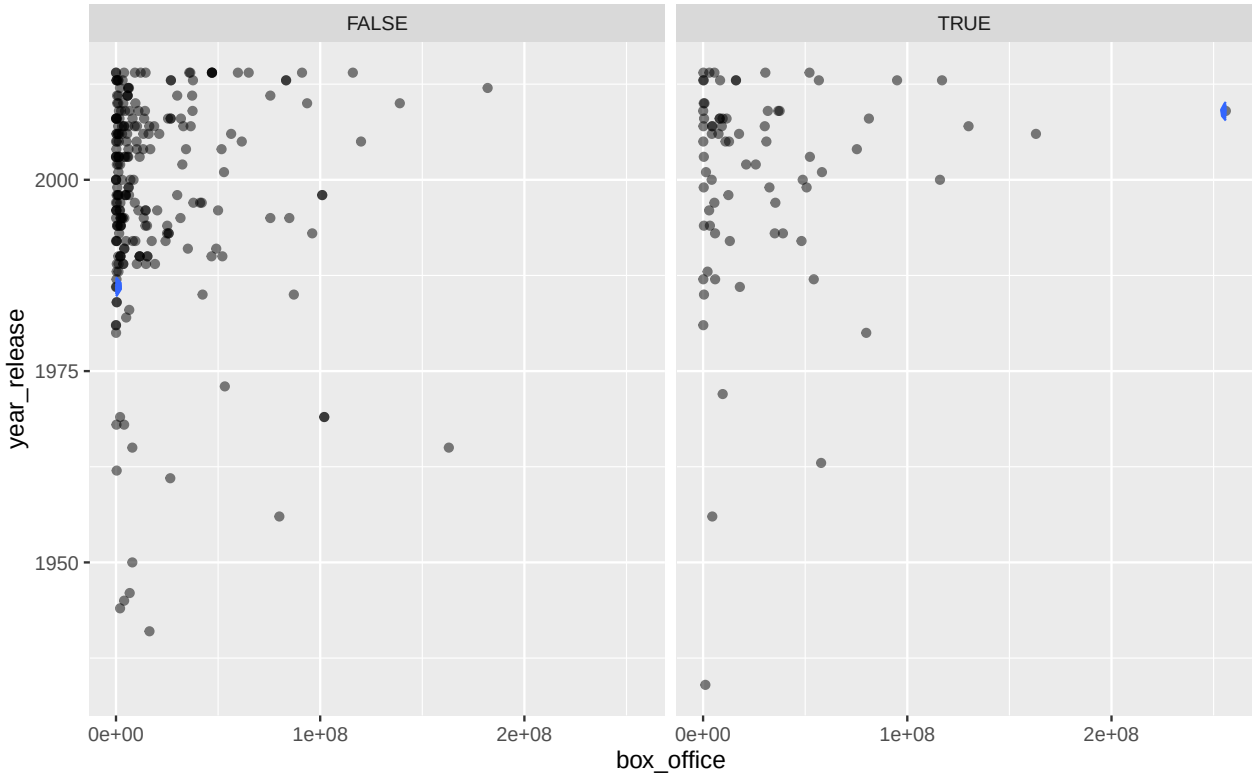
```
#Type I error
1-(0.95^2)
```

```
## [1] 0.0975
```

```
#bonferri adjustment
.05/2
```

```
## [1] 0.025
```

```
#Multivariate Normality
ggplot(biop, aes(x = box_office, y = year_release)) +
  geom_point(alpha = .5) +
  geom_density_2d(h=2) +
  facet_wrap(~person_of_color)
```




```
#Homogeniety of Variances and co-Variances
leveneTest(box_office~person_of_color,data=biop)
```

```
## Levene's Test for Homogeneity of Variance (center =
median)
## Df F value Pr(>F)
## group 1 4.4541 0.03561 *
## 314
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

```
leveneTest(year_release~person_of_color,data=biop)
```

```
## Levene's Test for Homogeneity of Variance (center =
median)
## Df F value Pr(>F)
## group 1 0.2997 0.5845
## 314
```

```
leveneTest(year_release+box_office~person_of_color,data=biop)
```

```
## Levene's Test for Homogeneity of Variance (center =
median)
## Df F value Pr(>F)
## group 1 4.4541 0.03561 *
## 314
## ---
```

```
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

A one-way MANOVA was conducted to determine the effect of whether the subject of a biographical film was a person of color (true or false) on the film's box office numbers and the date of release. Examination of bivariate density plots for each group revealed no stark deparures from multivariate normality. Examination of covariance matrices using the Levene test for each group revealed homogeniety for box office numbers but not for year of release. No univariate or multivariate outliers were evident and MANOVA was considered to be an appropriate analysis technique. Significant differences were found among the two options for at least one of the dependent variables. The MANOVA test resulted in a p-value of 0.0378, meaning that there is significant differences in the means of at least one variable on the subjects race. Univariate ANOVAs for each dependent variable were conducted as follow-up tests to the MANOVA, using the Bonferroni method for controlling Type I error rates for multiple comparisons. The univariate ANOVA for box office numbers was significant, while not significant for year of release: with a p-value of 0.019 and 0.249 respectively. Post hoc analysis was performed conducting pairwise comparisons to determine the difference in race for box office and release dates. People of color differed significantly from caucasians in terms of box office numbers but not in release year after adjusting for multiple comparisons (bonferroni 0.05/2 = 0.025). The chances of a type-I error was 9.75%.

- **2. (10 pts)** Perform some kind of randomization test on your data (that makes sense). This can be anything you want! State null and alternative hypotheses, perform the test, and interpret the results (7). Create a plot visualizing the null distribution and the test statistic (3).

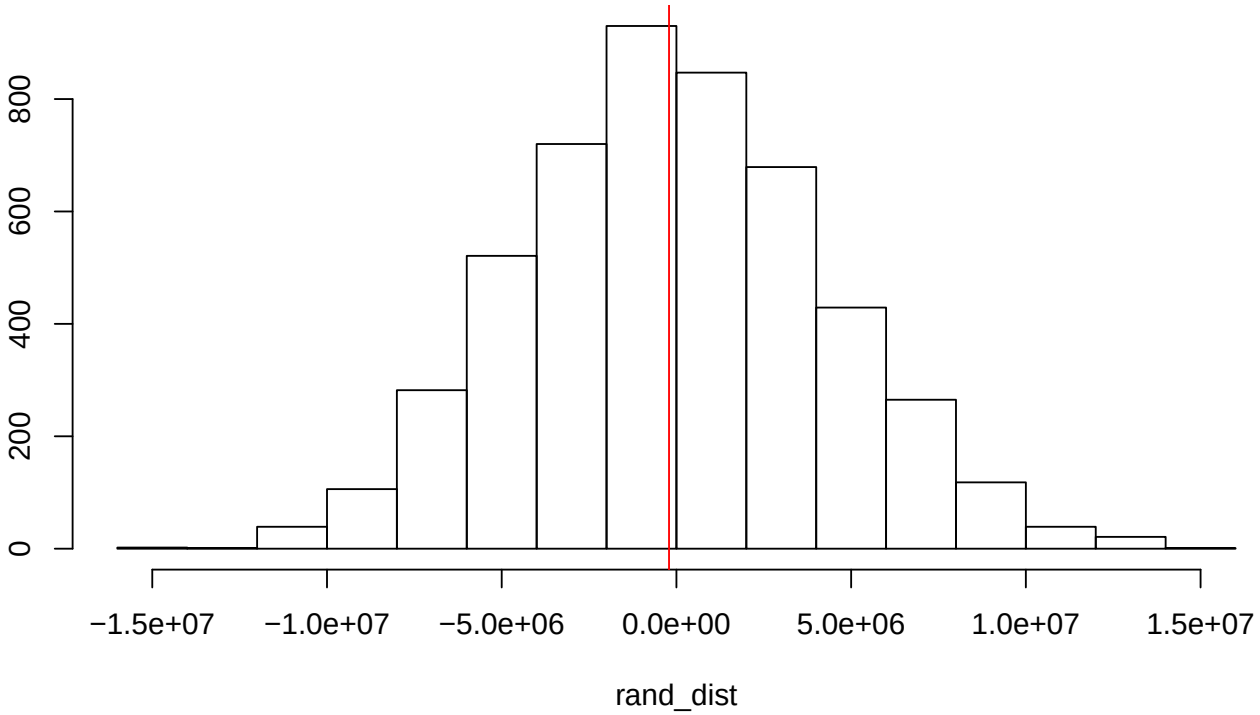
```
set.seed(348)

rand_dist<-vector()
for(i in 1:5000){
  new<-data.frame(box_office=sample(biop$box_office),poc=biop$person_of_color)
  rand_dist[i]<-mean(new[new$poc=="TRUE",]$box_office)-
    mean(new[new$poc=="FALSE",]$box_office)
}

mean(rand_dist)
```

```
## [1] -90457.6
```

```
{hist(rand_dist,main="",ylab=""); abline(v = -211475.5,col="red")}
```



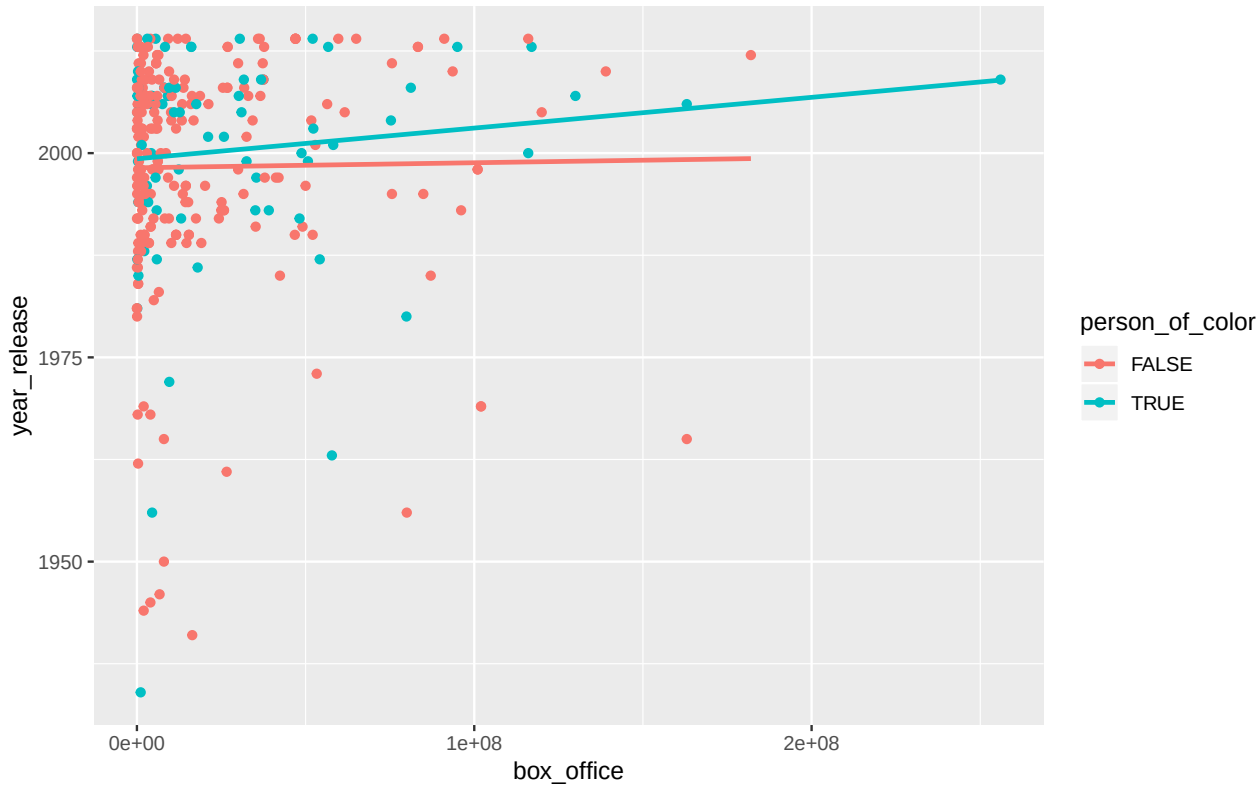
H0: The box office amount earned does not vary based on the race of the subject. HA: The box office amount earned does vary based on the race of the subject. According to the randomization test, we would reject the null hypothesis. The amount of money earned at the box office for biographical films does depend on the race of the film's subject.

- **3. (35 pts)** Build a linear regression model predicting one of your response variables from at least 2 other variables, including their interaction. Mean-center any numeric variables involved in the interaction.

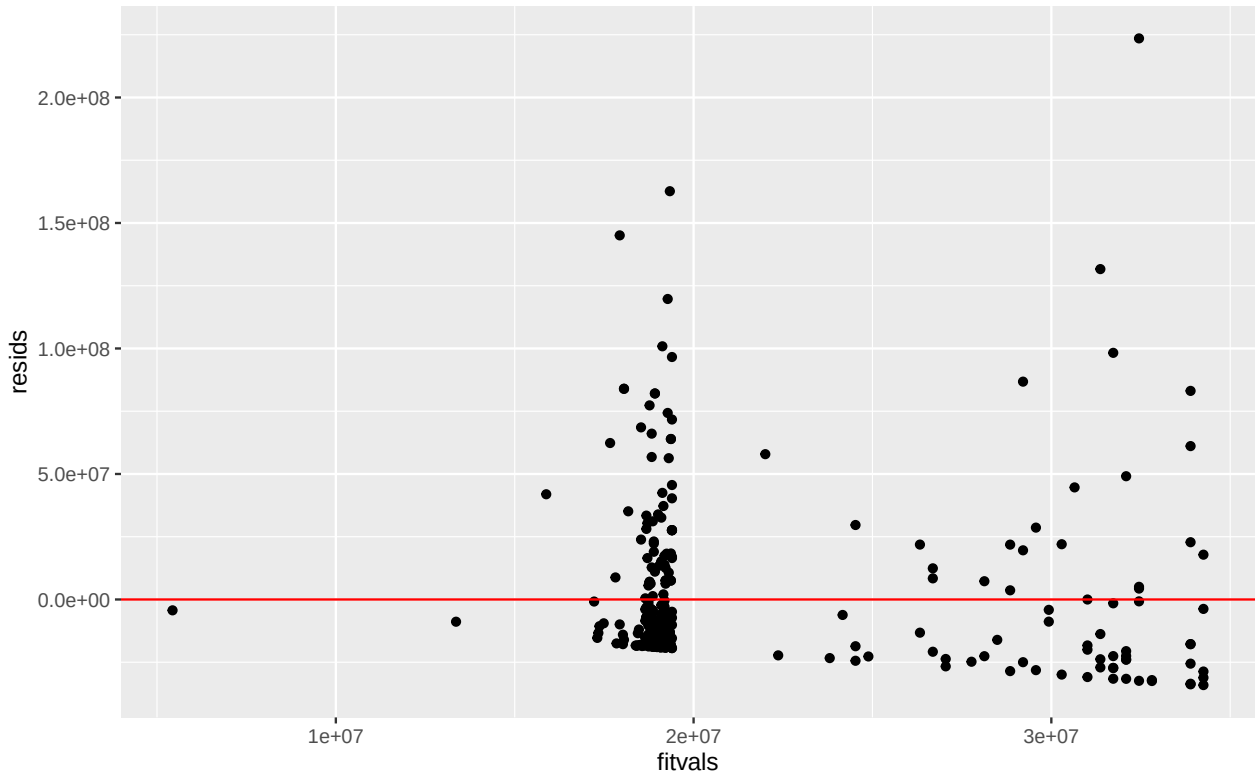
```
biop$yearc <- biop$year_release - mean(biop$year_release)
fitc <- lm(box_office ~ person_of_color+yearc+person_of_color*yearc, data = biop)
coef(fitc)
```

```
## (Intercept) person_of_colorTRUE yearc
## 18943412.29 9841751.87 29810.24
## person_of_colorTRUE:yearc
## 330404.88
```

```
ggplot(data.frame(biop), aes(box_office,year_release, color = person_of_color))+
  geom_point()+
  geom_smooth(method="lm",se=F)
```




```
#checking linearity and homoskedasticity
resids<-fitc$residuals
fitvals<-fitc$fitted.values
ggplot()+
  geom_point(aes(fitvals,resids))+
  geom_hline(yintercept=0, color='red')
```




```
bptest(fitc)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fitc
## BP = 5.7662, df = 3, p-value = 0.1236
```

```
#checking for normality
shapiro.test(resids)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resids
## W = 0.69948, p-value < 2.2e-16
```

```
#Robust standard errors
coeftest(fitc, vcov=vcovHC(fitc))
```

```
##
## t test of coefficients:
##
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18943412 1954185 9.6938 < 2e-16 ***
## person_of_colorTRUE 9841752 5167856 1.9044 0.05778 .
## yearc 29810 187188 0.1593 0.87357
## person_of_colorTRUE:yearc 330405 323955 1.0199 0.30856
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

```
#proportion of variance
SST <- sum((biop$box_office-mean(biop$box_office))^2) #SS Total
SSR <- sum((fitc$fitted.values-mean(biop$box_office))^2) #SS Regression
SSR/SST
```

```
## [1] 0.02267162
```

A linear regression was performed predicting box office numbers from the interaction between the race of the subject and the year of release. Year of release was the only numeric predictor variable so it was centered. The interaction gave an intercept coefficient of 18943412.29, meaning that for every non-poc film, every increase in year in general increased the box office number by that amount. The model did not follow homoskedasticity, but did have significant normality and linearity. Fitting for robust standard errors made the fit significant, with an intercept coefficient of 28785164 for people of color.

- **4. (5 pts)** Rerun same regression model (with interaction), but this time compute bootstrapped standard errors. Discuss any changes you observe in SEs and p-values using these SEs compared to the original SEs and the robust SEs)

```
fitc <- lm(box_office ~ person_of_color+yearc+person_of_color*yearc, data = biop)
coef(fitc)
```

```
## (Intercept) person_of_colorTRUE yearc
## 18943412.29 9841751.87 29810.24
## person_of_colorTRUE:yearc
## 330404.88
```

```
samp_distn<-replicate(5000, {
boot_dat <- sample_frac(biop, replace=T)
fitboot <- lm(box_office ~ person_of_color+yearc+person_of_color*yearc, data = boot_dat)
coef(fitboot)
})

#Robust SEs
samp_distn %>% t %>% as.data.frame %>% summarize_all(sd)
```

```
## (Intercept) person_of_colorTRUE yearc
person_of_colorTRUE:yearc
## 1 1936635 5100256 188305.7 346283.6
```

The regression was redone with an interaction between the predictors subject_sex and year released. Computing for robust standard errors, the intercept was reduced to 1909026 for caucasion people. This indicates that while the release year on its own does not affect the success of a movie with a person of color subject, when interacting with the box office numbers, the year does become slightly more indicative of the success of the movie.

- **5. (40 pts)** Perform a logistic regression predicting a binary categorical variable (if you don't have one, make/get one) from at least two explanatory variables (interaction not necessary).

```
newbiop <- biop %>%
  mutate(poc = ifelse(person_of_color == "TRUE", 1, 0)) %>%
  glimpse()
```

```
## Observations: 316
## Variables: 15
## $ title <chr> "12 Years a Slave", "21", "24 Hour Party
People", "42", "A Dangerous...
## $ country <chr> "US/UK", "US", "UK", "US", "Canada/UK",
"Canada/UK", "Canada/UK", "U...
```

```
## $ year_release <int> 2013, 2008, 2002, 2013, 2011, 2011,
2011, 2007, 2007, 2004, 2001, 19...
## $ box_office <dbl> 5.67e+07, 8.12e+07, 1.13e+06,
9.50e+07, 5.70e+06, 5.70e+06, 5.70e+06...
## $ director <chr> "Steve McQueen", "Robert Luketic",
"Michael Winterbottom", "Brian He...
## $ number_of_subjects <int> 1, 1, 1, 1, 3, 3, 3, 2, 2, 1,
1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ subject <chr> "Solomon Northup", "Jeff Ma", "Tony
Wilson", "Jackie Robinson", "Car...
## $ type_of_subject <chr> "Other", "Other", "Musician",
"Athlete", "Academic", "Academic", "Ac...
## $ race_known <chr> "Known", "Known", "Known", "Known",
"Known", "Known", "Known", "Know...
## $ subject_race <chr> "African American", "Asian
American", "White", "African American", "...
## $ person_of_color <lgl> TRUE, TRUE, FALSE, TRUE, FALSE,
FALSE, FALSE, FALSE, TRUE, FALSE, TR...
## $ subject_sex <chr> "Male", "Male", "Male", "Male",
"Male", "Male", "Female", "Male", "F...
## $ lead_actor_actress <chr> "Chiwetel Ejiofor", "Jim
Sturgess", "Steve Coogan", "Chadwick Bosema...
## $ yearc <dbl> 14.177215, 9.177215, 3.177215, 14.177215,
12.177215, 12.177215, 12.1...
## $ poc <dbl> 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0,
1, 1, 0, 0, 0, 1, 0, 0,...
```

```
pocfit <- glm(poc~subject_sex+box_office,data=newbiop, family = "binomial")
coeftest(pocfit)
```

```
##
## z test of coefficients:
##
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.5947e+00 2.9701e-01 -5.3691 7.912e-08 ***
## subject_sexMale 3.4304e-01 3.2559e-01 1.0536 0.2921
## box_office 7.7071e-09 3.5729e-09 2.1571 0.0310 *
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

```
exp(coef(pocfit))
```

```
##     (Intercept) subject_sexMale      box_office
##        0.202970        1.409229        1.000000
```

A logistical regression was done to predict the liklihood of a subject being a person of color based on the subject's sex and box office data. The regression found that caucasian males had a higher chance of gaining a higher box office number.

```
#Confusion Matrix
newbiop2<-newbiop%>%mutate(prob=predict(pocfit, type="response"),
                                  prediction=ifelse(prob>.5,1,0))
classify2<-newbiop2 %>%
  transmute(prob,prediction,poc=poc)
table(prediction=newbiop2$prediction,poc=newbiop2$poc) %>%
```

```
  addmargins()
```

```
##           poc
## prediction   0   1 Sum
##        0   239  74 313
##        1     1   2   3
##        Sum 240  76 316
```

```
#Accuracy
(239+2)/316
```

```
## [1] 0.7626582
```

```
#Sensitivity (TPR)
239/313
```

```
## [1] 0.7635783
```

```
#Specificity (TNR)
2/3
```

```
## [1] 0.6666667
```

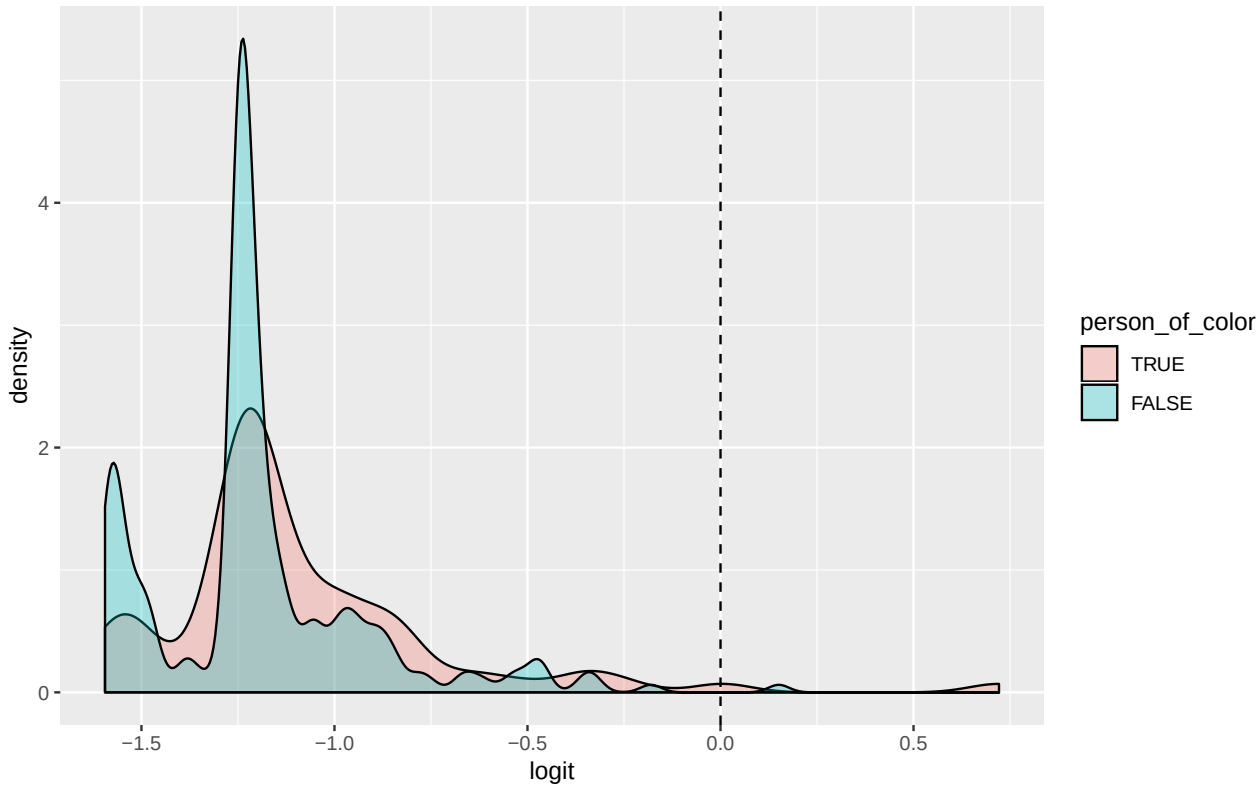
```
#Recall (PPV)
238/240
```

```
## [1] 0.9916667
```

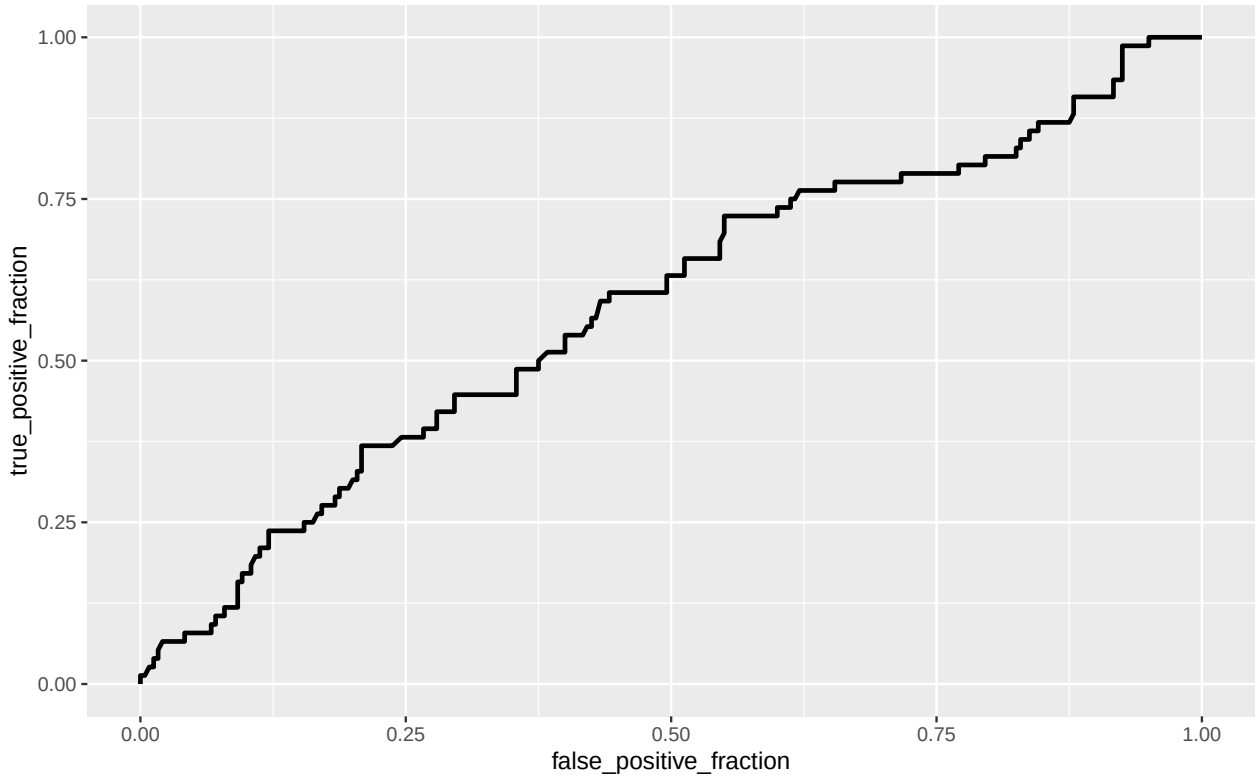
A confusion matrix was created and found an accuracy of 0.7627, a sensitivity of 0.7636, a specificity of 0.667, and a recall value of 0.9917. This indicates that the accuracy of the model is decent but not great.

```
#Density of log-odds
biop$logit<-predict(pocfit) #get predicted log-odds
biop$person_of_color<-factor(biop$person_of_color,levels=c("TRUE","FALSE"))

ggplot(biop,aes(logit, fill=person_of_color))+
  geom_density(alpha=.3)+
  geom_vline(xintercept=0,lty=2)
```




```
#ROC Curve
ROCplot1 <- ggplot(newbiop)+geom_roc(aes(d=poc,m=box_office), n.cuts=0)
ROCplot1
```



```
calc_auc(ROCplot1)
```

```
##   PANEL group       AUC
## 1     1    -1 0.5850055
```

An ROC curve was generated as well as an AUC value of 0.585. This AUC score falls under the "bad" category.

```
class_diag<-function(probs,truth){
  if(is.numeric(truth)==FALSE & is.logical(truth)==FALSE)
    truth<-as.numeric(truth)-1
  tab<-table(factor(probs>.5,levels=c("FALSE","TRUE")),truth)
  prediction<-ifelse(probs>.5,1,0)
  acc=mean(truth==prediction)
  sens=mean(prediction[truth==1]==1)
  spec=mean(prediction[truth==0]==0)
  ppv=mean(truth[prediction==1]==1)

ord <- order(probs, decreasing = TRUE)
probs <- probs[ord]; truth <- truth[ord]
TPR = cumsum(truth)/max(1,sum(truth))
FPR = cumsum(!truth)/max(1,sum(!truth))
dup <- c(probs[-1]>=probs[-length(probs)], FALSE)
TPR <- c(0,TPR[!dup],1); FPR <- c(0,FPR[!dup],1)
n <- length(TPR)
auc <- sum( ((TPR[-1]+TPR[-n])/2) * (FPR[-1]-FPR[-n]) )
data.frame(acc,sens,spec,ppv,auc)
}


#10-fold CV
set.seed(1234)
k=10

data <- newbiop %>% sample_frac
folds <- ntile(1:nrow(data),n=10)

diags<-NULL
for(i in 1:k){
train <- data[folds!=i,]
test <- data[folds==i,]
truth <- test$poc
fit3 <- glm(poc ~ box_office + subject_sex, data = newbiop, family=binomial)
probs <- predict(fit3, newdata=test, type="response")
diags<-rbind(diags,class_diag(probs,truth))
}
diags%>%summarize_all(mean)
```

```
##         acc       sens      spec ppv       auc
## 1 0.7622984 0.02678571 0.9964286 NaN 0.5819354
```

A 10-fold CV was performed, giving an AUC of 0.608, which is poor, but better than the original AUC calculated. This analysis has an accuracy of 0.7627, a sensitivity of 0.03, and does not give a recall value.

- **6. (10 pts)** Choose one variable you want to predict (can be one you used from before; either binary or continuous) and run a LASSO regression inputting all the rest of your variables as predictors. Choose

lambda to give the simplest model whose accuracy is near that of the best (i.e., `lambda.1se`). Discuss which variables are retained. Perform 10-fold CV using this model: if response in binary, compare model's out-of-sample accuracy to that of your logistic regression in part 5; if response is numeric, compare the residual standard error (at the bottom of the summary output, aka RMSE): lower is better fit!

```
library(glmnet)

y <- as.matrix(newbiop$poc)
x <- model.matrix(poc~country+year_release+box_office+number_of_subjects+type_of_subject+subject_race+p
x <- scale(x)

cv <- cv.glmnet(x,y, family="binomial")
lasso <- glmnet(x,y, family="binomial", lambda=cv$lambda.1se)
coef(lasso)
```

```
## 50 x 1 sparse Matrix of class "dgCMatrix"
##                                            s0
## (Intercept)                        -4.395898
## countryCanada/UK                    .
## countryUK                           .
## countryUS                           .
## countryUS/Canada                    .
## countryUS/UK                        .
## countryUS/UK/Canada                 .
## year_release                        .
## box_office                          .
## number_of_subjects                  .
## type_of_subjectActivist             .
## type_of_subjectActor                .
## type_of_subjectActress              .
## type_of_subjectActress / activist   .
## type_of_subjectArtist               .
## type_of_subjectAthlete              .
## type_of_subjectAthlete / military   .
## type_of_subjectAuthor               .
## type_of_subjectAuthor (poet)        .
## type_of_subjectComedian             .
## type_of_subjectCriminal             .
## type_of_subjectGovernment           .
## type_of_subjectHistorical           .
## type_of_subjectJournalist           .
## type_of_subjectMedia                .
## type_of_subjectMedicine             .
## type_of_subjectMilitary             .
## type_of_subjectMusician             .
## type_of_subjectOther                .
## type_of_subjectPolitician           .
## type_of_subjectSinger               .
## type_of_subjectTeacher              .
## type_of_subjectWorld leader         .
## subject_raceAfrican American        .
## subject_raceAsian                   .
## subject_raceAsian American          .
## subject_raceCaribbean               .
```

```
## subject_raceHispanic (Latin American)  .
## subject_raceHispanic (Latina)          .
## subject_raceHispanic (Latino)          .
## subject_raceHispanic (White)           .
## subject_raceIndian                     .
## subject_raceMediterranean              .
## subject_raceMiddle Eastern             .
## subject_raceMiddle Eastern (White)     .
## subject_raceMulti racial               .
## subject_raceNative American            .
## subject_raceWhite                      .
## person_of_colorTRUE                    6.302513
## subject_sexMale                        .
```

```
set.seed(1234)
k=10

data <- newbiop %>%
  sample_frac
folds <- ntile(1:nrow(data),n=10)
diags<-NULL

for(i in 1:k){
  train <- data[folds!=i,]
  test <- data[folds==i,]
  truth <- test$poc
  fit6 <- glm(poc~person_of_color,
              data=train, family = "binomial")
  probs <- predict(fit6, newdata = test, type = "response")
  diags <- rbind(diags,class_diag(probs,truth))
}
diags%>%summarize_all(mean)
```

```
##   acc sens spec ppv auc
## 1   1    1    1   1   1
```

The variable to be predicted was the liklihood of being a person of color based on the other variables. Several variables had to be removed due to having a different categorical response per observation, or having the same answer for each observation. A LASSO regression was performed and found that no other variable was significant to predict the poc binary variable. The only variable that returned as significant was the person_of_color variable, which is the non-binary version of the poc variable. As a result, the LASSO regression gave an accuracy, sensitivity, and recall value of 1.