Project 2

Cara Johnston

2020-05-01

0. Introduction For this assignment, I chose to use the dataset 'Salaries' from the 'carData' package. There are 397 observations on 6 variables: rank, discipline, yrs.since.phd, yrs.service, sex and salary. Rank represents title, as in Professor, Assistant Professor or Associate Professor. Discipline has two levels representing department type, A (theoretical) or B (applied). The variable "yrs.since.phd" represents the number of years the individual has had their PhD. The variable "yrs.service" represent the number of years the individual has been teaching. The variable sex tells whether the individual is male or female. Lastly, the salary variable represents the person's nine-month salary, in dollars.

1. MANOVA

```
#Tests
man1<-manova(cbind(yrs.since.phd, yrs.service, salary)~rank, data=Prof_Salaries)
summary(man1)
```

```
##            Df  Pillai approx F num Df den Df    Pr(>F)
## rank        2 0.63281   60.633      6    786 < 2.2e-16 ***
## Residuals 394
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary.aov(man1)
```

```
##  Response yrs.since.phd :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## rank          2  32390 16194.8  191.18 < 2.2e-16 ***
## Residuals   394  33376    84.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response yrs.service :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## rank          2  24812   12406   115.9 < 2.2e-16 ***
## Residuals   394  42175     107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response salary :
##              Df     Sum Sq    Mean Sq F value    Pr(>F)
## rank          2 1.4323e+11 7.1616e+10  128.22 < 2.2e-16 ***
## Residuals   394 2.2007e+11 5.5855e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pairwise.t.test(Prof_Salaries$yrs.since.phd, Prof_Salaries$rank, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Prof_Salaries$yrs.since.phd and Prof_Salaries$rank
##
##           AsstProf AssocProf
## AssocProf 3.6e-10  -
## Prof      < 2e-16  < 2e-16
##
## P value adjustment method: none
```

```
pairwise.t.test(Prof_Salaries$yrs.service, Prof_Salaries$rank, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Prof_Salaries$yrs.service and Prof_Salaries$rank
##
##           AsstProf AssocProf
## AssocProf 2.0e-07  -
## Prof      < 2e-16  3.2e-13
##
## P value adjustment method: none
```

```
pairwise.t.test(Prof_Salaries$salary, Prof_Salaries$rank, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Prof_Salaries$salary and Prof_Salaries$rank
##
##           AsstProf AssocProf
## AssocProf 0.0016   -
## Prof      <2e-16   <2e-16
##
## P value adjustment method: none
```

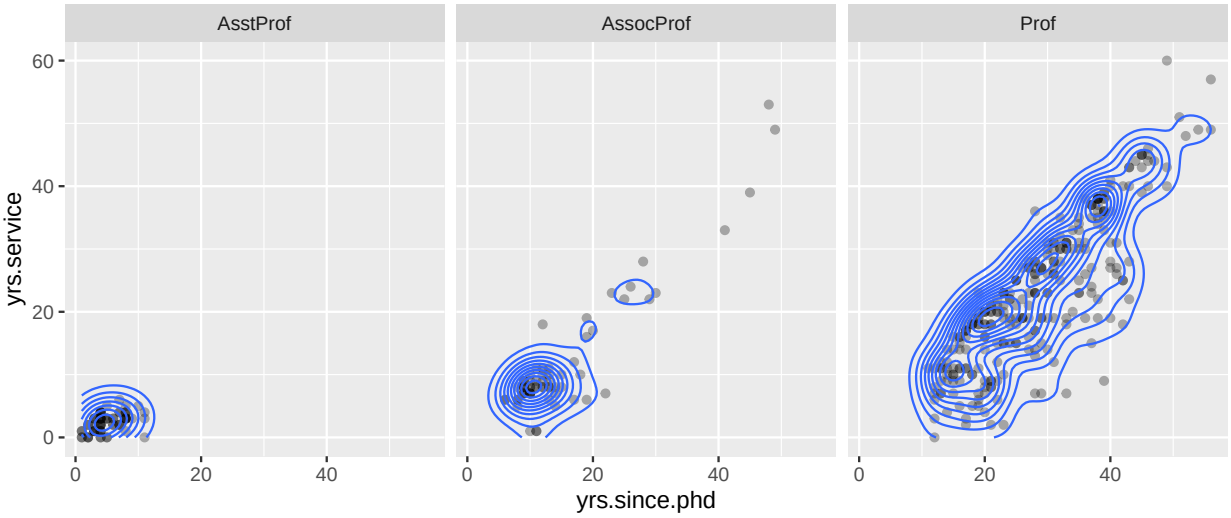
```
#Type I Error
1-(.95^13)
```

```
## [1] 0.4866579
```

```
#Bonferroni Correction
.05/13
```

```
## [1] 0.003846154
```

```
#Checking for Multivariate Normality
library(ggplot2)

ggplot(Prof_Salaries, aes(yrs.since.phd, yrs.service)) +  geom_point(alpha = .3) + geom_density_2d(h=10
```



MANOVA testing was performed, finding a significant result overall. Consequently, I ran univariate ANOVAs for each numeric variable. I found significant results for each variable (yrs.since.phd, yrs.service and salary) indicating that for each of these variables, at least one rank differs. Post-hoc t tests were also conducted, finding that all three ranks differed in years since PhD was obtained, years of service and salary. The findings remained the same even after adjusting for multiple comparisons (Bonferroni alpha = 0.0038). In total, 13 tests were performed, including 1 MANOVA, 3 ANOVAs, 9 t-tests. Due to the number of tests run, the probability of at least one type I error is 0.487. With the Bonferroni correction, the new alpha value is 0.0038. Using this new alpha, the differences between groups are still significant, as the p-value for each test was considerably lower than 0.0038.

MANOVA assumptions are as follows: 1. Random samples and independent observations 2. Multivariate normality of DVs 3. Homogeneity of within-group covariance matrices 4. Linear relationship among DVs 5. No extreme univariate or multivariate outliers 6. No multicollinearity

Most of the assumptions are difficult to test for, so I only tested for multivariate normality and homogeneity of covariances. Looking at the graphs for multivariate normality, we can assume that the data does not meet that assumption. Unfortunately, my code would not work to check for homogeneity of covariances, but it likely wasn't met. Most of the assumptions are difficult to meet, so I don't anticipate my data meeting most of them.

2. Randomization Test

```
#Randomization Test
rand<-vector()
for(i in 1:5000){
new<-data.frame(salary=sample(Prof_Salaries$salary),sex=Prof_Salaries$sex)
rand[i]<-mean(new[new$sex=="Female",]$salary)-
              mean(new[new$sex=="Male",]$salary)
}

mean(Prof_Salaries[Prof_Salaries$sex=="Female",]$salary)-
```

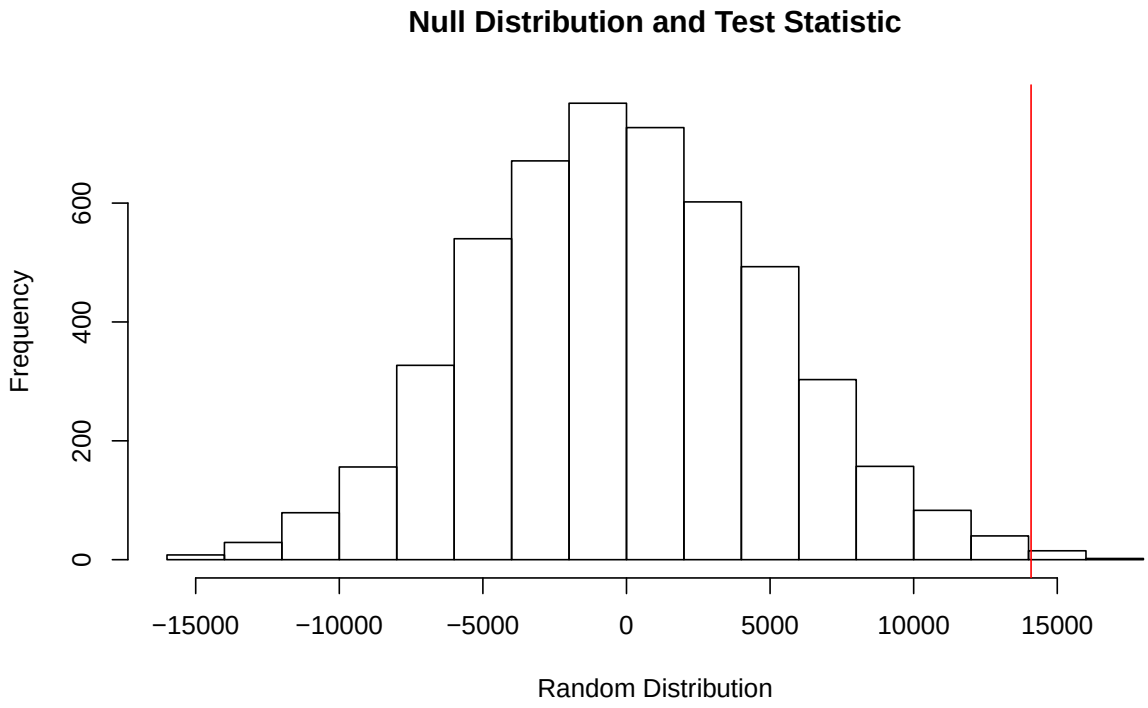
```
mean(Prof_Salaries[Prof_Salaries$sex=="Male",]$salary)
```

```
## [1] -14088.01
```

```
mean(rand> 14088.01 | rand< -14088.01)
```

```
## [1] 0.0048
```

```
#Visualization
{hist(rand,main="Null Distribution and Test Statistic",xlab="Random Distribution"); abline(v = 14088.01
```



For the randomization test, I decided to test whether there were mean differences in salary for male and female professors. The null hypothesis (H0) is that the mean salary is the same for male vs. female professors. The alternative hypothesis (HA) is that the mean salary is different for male vs. female professors. After running the test, I got a p-value of 0.005, meaning that we can reject the null hypothesis and conclude that there is a difference in mean salary between male and female professors.

3. Linear Regression

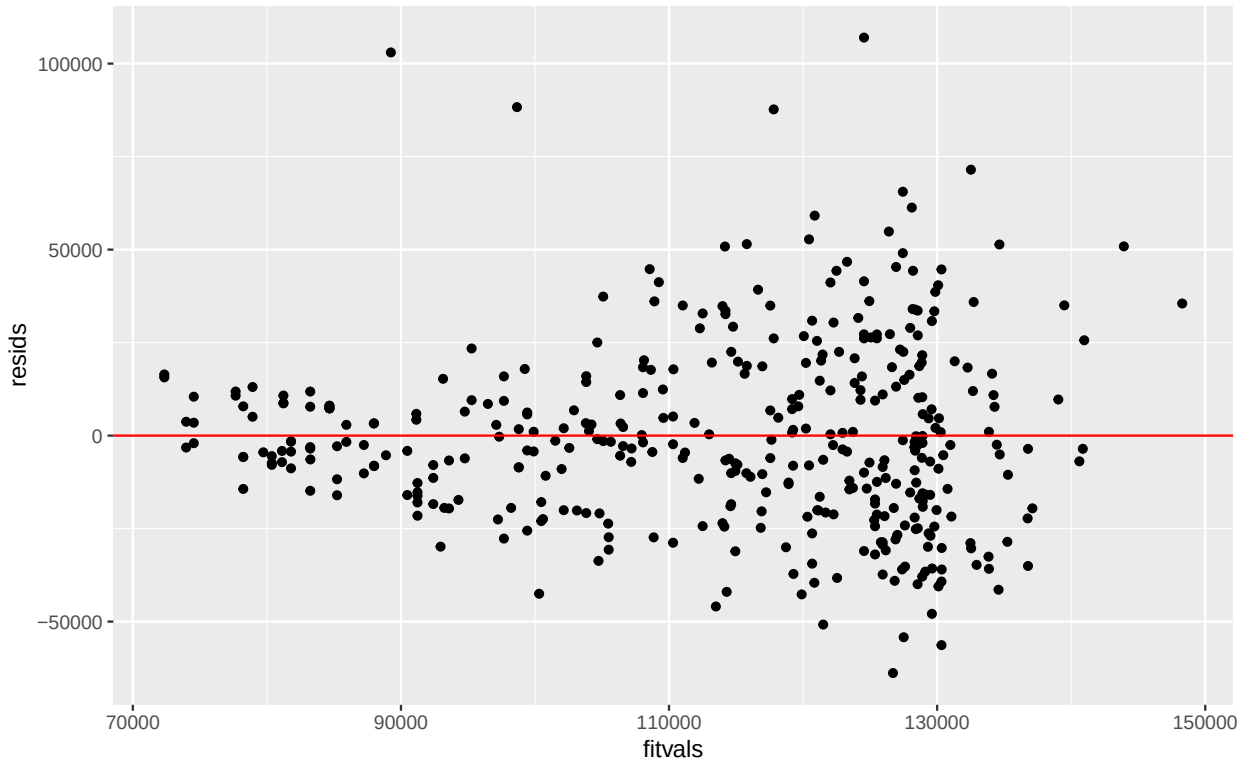
```
#Mean-Center Numeric Variables
Prof_Salaries$yrs.service_c <- Prof_Salaries$yrs.service - mean(Prof_Salaries$yrs.service, na.rm=T)

Prof_Salaries$yrs.since.phd_c <- Prof_Salaries$yrs.since.phd - mean(Prof_Salaries$yrs.since.phd, na.rm='

#Linear Regression Model
model <- lm(salary~yrs.service_c*yrs.since.phd_c, data=Prof_Salaries)
summary(model)
```

```
##
## Call:
## lm(formula = salary ~ yrs.service_c * yrs.since.phd_c, data = Prof_Salaries)
##
```

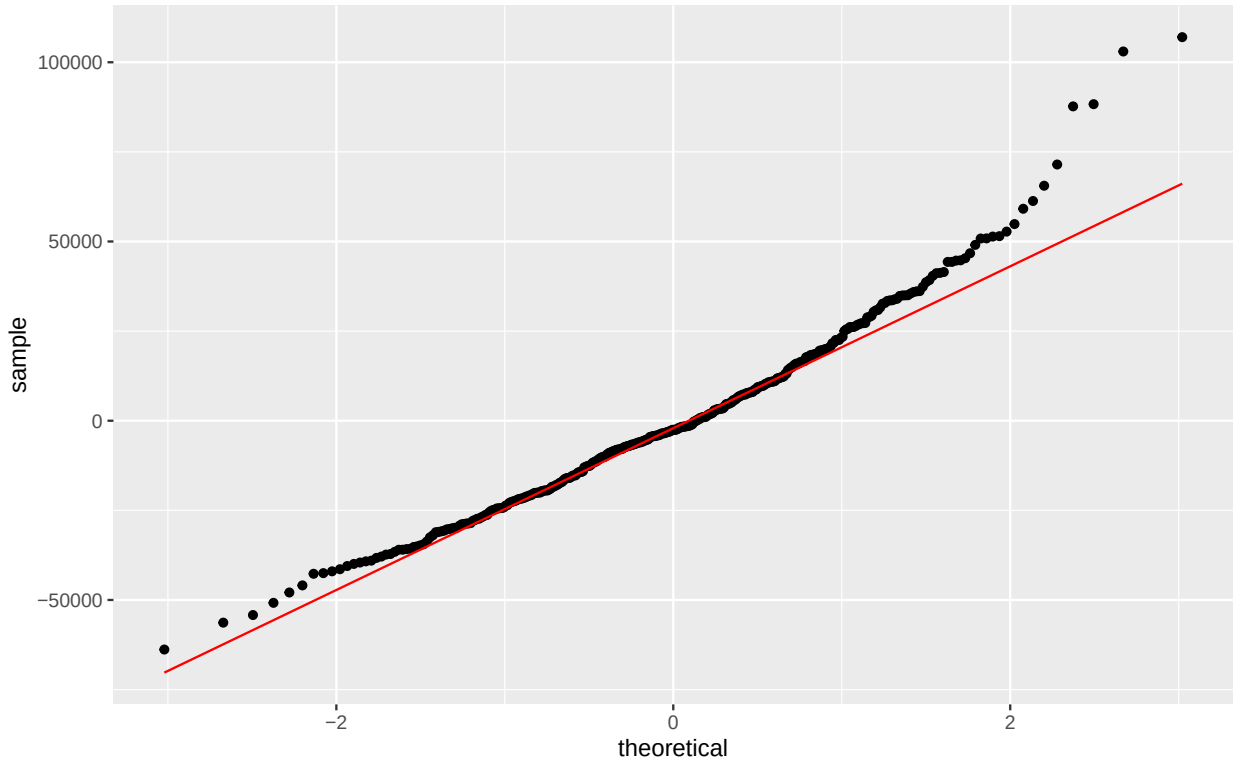
```
## Residuals:
##    Min     1Q Median     3Q    Max
## -63823 -17292  -2538  13158 107001
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  123533.470   1698.633  72.725  < 2e-16 ***
## yrs.service_c                   250.528    254.880   0.983    0.326
## yrs.since.phd_c                1056.086    242.975   4.346 1.76e-05 ***
## yrs.service_c:yrs.since.phd_c   -64.617      7.487  -8.630  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25120 on 393 degrees of freedom
## Multiple R-squared:  0.3177, Adjusted R-squared:  0.3125
## F-statistic: 60.99 on 3 and 393 DF,  p-value: < 2.2e-16
```

```
#Plot of Regression
new1<-Prof_Salaries
new1$yrs.service_c<-mean(Prof_Salaries$yrs.service_c)
new1$mean<-predict(model,new1)
new1$yrs.service_c<-mean(Prof_Salaries$yrs.service_c)+sd(Prof_Salaries$yrs.service_c)
new1$plus.sd<-predict(model,new1)
new1$yrs.service_c<-mean(Prof_Salaries$yrs.service_c)-sd(Prof_Salaries$yrs.service_c)
new1$minus.sd<-predict(model,new1)
newint<-new1%>%select(salary,yrs.since.phd_c,mean,plus.sd,minus.sd)%>%gather(yrs.service_c,value,-salary

mycols<-c("#619CFF","#F8766D","#00BA38")
names(mycols)<-c("-1 sd","mean","+1 sd")
mycols=as.factor(mycols)

ggplot(Prof_Salaries,aes(yrs.since.phd_c,salary),group=mycols)+geom_point()+geom_line(data=new1,aes(y=m
```


```
#Check Assumptions
resids<-model$residuals
fitvals<-model$fitted.values
ggplot()+geom_point(aes(fitvals,resids))+geom_hline(yintercept=0, color='red')
```



```
ggplot()+geom_qq(aes(sample=resids))+geom_qq_line(aes(sample=resids), color='red')
```




```
ks.test(resids, "pnorm", mean=0, sd(resids))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  resids
## D = 0.062195, p-value = 0.09271
## alternative hypothesis: two-sided
```

```
#with Robust Standard Errors
library(sandwich); library(lmtest)

summary(model)$coef[,1:2] #uncorrected SEs
```

```
##                                 Estimate  Std. Error
## (Intercept)                   123533.47023 1698.633174
## yrs.service_c                    250.52836  254.880140
## yrs.since.phd_c                 1056.08650  242.975151
## yrs.service_c:yrs.since.phd_c    -64.61694    7.487103
```

```
coeftest(model, vcov = vcovHC(model))[,1:2] #corrected SEs
```

```
##                                 Estimate Std. Error
## (Intercept)                   123533.47023 1974.96670
## yrs.service_c                    250.52836  310.70707
## yrs.since.phd_c                 1056.08650  294.53162
## yrs.service_c:yrs.since.phd_c    -64.61694   11.01044
```

```
#Calculation of R^2
(sum((Prof_Salaries$salary-mean(Prof_Salaries$salary))^2)-sum(model$residuals^2))/sum((Prof_Salaries$sa
```

```
## [1] 0.3176664
```

Interpretation of Coefficient Estimates: *Intercept*: Predicted salary for average number of years of service and average number of years since PhD is $123,533. *Years Service*: Controlling for number of years since PhD, for every one additional year of service, salary goes up by $251 on average. *Years since PhD*: Controlling for number of years of service, for every one additional year since PhD was received, salary increase by $1056 on average. *Interaction (yrs.service and yrs.since.phd)*: The differences in slope for salary is -64.6 for years of service and years since Phd was received.

Assumptions Linearity and normality appear to be met, though the data fails the homoskedasticity assumption, as shown in the graph where it appears to fan out.

Using Robust SEs In the corrected version using the robust SEs, the standard error values were higher than in the uncorrected version. This translates into smaller t-values and higher p-values for all variables in the model.

Variation explained by model I calculated R^2 to be 0.3177, which means that the model explains 31.77% of the variation in the outcome.

4. Regression with Bootstrapped SEs

```
model <- lm(salary~yrs.service_c*yrs.since.phd_c, data=Prof_Salaries)
summary(model)
```

```
##
## Call:
## lm(formula = salary ~ yrs.service_c * yrs.since.phd_c, data = Prof_Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -63823 -17292  -2538  13158 107001
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  123533.470   1698.633  72.725  < 2e-16 ***
## yrs.service_c                   250.528    254.880   0.983    0.326
## yrs.since.phd_c                1056.086    242.975   4.346 1.76e-05 ***
## yrs.service_c:yrs.since.phd_c   -64.617      7.487  -8.630  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25120 on 393 degrees of freedom
## Multiple R-squared:  0.3177, Adjusted R-squared:  0.3125
## F-statistic: 60.99 on 3 and 393 DF,  p-value: < 2.2e-16
```

```
#Normal-theory SEs
coeftest(model)[,1:2]
```

```
##                                 Estimate  Std. Error
## (Intercept)                 123533.47023 1698.633174
## yrs.service_c                  250.52836  254.880140
## yrs.since.phd_c               1056.08650  242.975151
## yrs.service_c:yrs.since.phd_c  -64.61694    7.487103
```

```r
#Robust SEs
coeftest(model, vcov=vcovHC(model))[,1:2]
```

```
##                                  Estimate Std. Error
## (Intercept)                  123533.47023 1974.96670
## yrs.service_c                   250.52836  310.70707
## yrs.since.phd_c                1056.08650  294.53162
## yrs.service_c:yrs.since.phd_c   -64.61694   11.01044
```

```r
#Bootstrapped SEs
samp_distn<-replicate(5000, {
  boot_dat <- sample_frac(Prof_Salaries, replace=T)
  model2 <- lm(salary~yrs.service_c*yrs.since.phd_c, data=boot_dat)
  coef(model2)
})

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

```
##   (Intercept) yrs.service_c yrs.since.phd_c yrs.service_c:yrs.since.phd_c
## 1    1921.944      301.2239        283.9978                      10.64156
```

The bootstrapped standard error values for all variables fell in between the original and robust standard error values, though they were closer to the robust SEs. This will also change the p-values, as they should now be slightly higher than the original p-values, and slightly smaller than the p-values determined using the robust SEs.

5. Logistic Regression

```r
fit<-glm(discipline~yrs.service+salary,data=Prof_Salaries,family=binomial(link="logit"))
coeftest(fit)
```

```
##
## z test of coefficients:
##
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) -1.1086e+00  4.2461e-01 -2.6108  0.009032 **
## yrs.service -4.1909e-02  9.2945e-03 -4.5090 6.514e-06 ***
## salary       1.7857e-05  4.0698e-06  4.3876 1.146e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
exp(coef(fit))
```

```
## (Intercept) yrs.service      salary
##   0.3300240   0.9589574   1.0000179
```

```r
#Confusion Matrix
probs<-predict(fit, type="response")
table(predict=as.numeric(probs>.5),truth=Prof_Salaries$discipline)%>%addmargins
```

```
##        truth
## predict   A   B Sum
##     0    82  38 120
##     1    99 178 277
##     Sum 181 216 397
```

```r
#Accuracy
(82+178)/397
```

```
## [1] 0.6549118
```

```
#Sensitivity (TPR)
178/216
```

```
## [1] 0.8240741
```

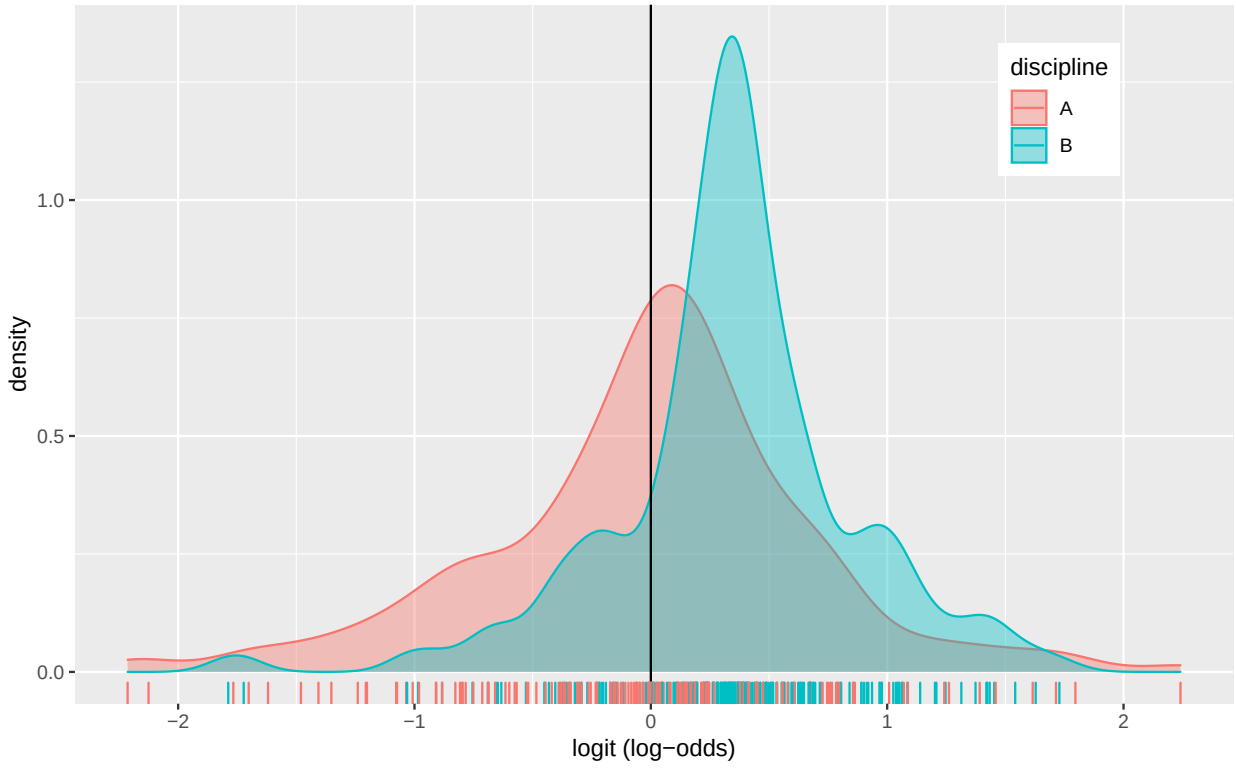
```
#Specificity (TNR)
82/181
```

```
## [1] 0.4530387
```

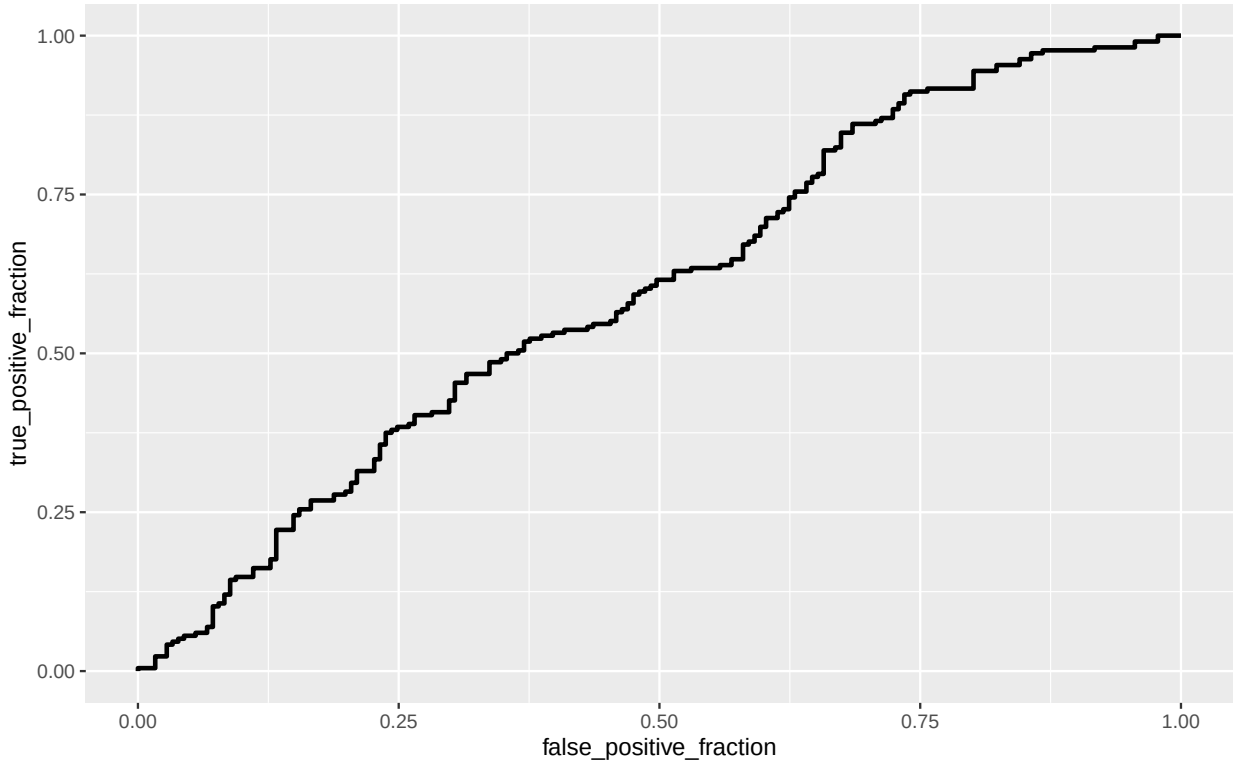
```
#Precision (PPV)
178/277
```

```
## [1] 0.6425993
```

```
#Density Plot
Prof_Salaries$logit<-predict(fit,type="link")

Prof_Salaries%>%ggplot()+geom_density(aes(logit,color=discipline,fill=discipline), alpha=.4)+
  theme(legend.position=c(.85,.85))+geom_vline(xintercept=0)+xlab("logit (log-odds)")+
  geom_rug(aes(logit,color=discipline))
```




```
#ROC Plot
library(plotROC)
ROCplot<-ggplot(Prof_Salaries)+geom_roc(aes(d=discipline,m=yrs.service+salary), n.cuts=0)
ROCplot
```




```
calc_auc(ROCplot)
```

```
##   PANEL group       AUC
## 1     1    -1 0.5980919
```

```
#Class Diag Function
class_diag <- function(probs,truth){
  #CONFUSION MATRIX: CALCULATE ACCURACY, TPR, TNR, PPV
  tab<-table(factor(probs>.5,levels=c("FALSE","TRUE")),truth)
  acc=sum(diag(tab))/sum(tab)
  sens=tab[2,2]/colSums(tab)[2]
  spec=tab[1,1]/colSums(tab)[1]
  ppv=tab[2,2]/rowSums(tab)[2]
  if(is.numeric(truth)==FALSE & is.logical(truth)==FALSE) truth<-as.numeric(truth)-1
    #CALCULATE EXACT AUC
  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<-Prof_Salaries[sample(nrow(Prof_Salaries)),]
folds<-cut(seq(1:nrow(Prof_Salaries)),breaks=k,labels=F)
diags<-NULL
for(i in 1:k){
  train<-data[folds!=i,]
  test<-data[folds==i,]
  truth<-test$discipline

  fit2<-glm(discipline~yrs.service+salary,data=train,family="binomial")

  probs<-predict(fit2,newdata = test,type="response")

  diags<-rbind(diags,class_diag(probs,truth))
}

summarize_all(diags,mean)
```

```
##         acc      sens      spec       ppv       auc
## 1 0.6197436 0.7954489 0.4386117 0.6282468 0.6729663
```

Controlling for salary, I found that years of service has a significant negative impact on discipline. Controlling for years of service, I found that salary has a significant positive impact on discipline. By exponentiating the coefficients, I was able to identify more specific predictions of discipline. I found that the odds of being in discipline A is 0.330 for a professor with 0 years of service and no salary. Controlling for salary, for every one year increase in service, the odds of being in discipline A increase by 0.959. Controlling for years of service, for every dollar increase in salary, odds of being in discipline A increase by a factor of 1.0.

Using the confusion matrix, I calculated the Accuracy, Sensitivity (TPR), Specificity (TNR), and Recall (PPV) of the model. Accuracy is 0.65, representing the proportion of correctly identified disciplines. Sensitivity is 0.82, representing the true positive rate, or the probability in which discipline A is correctly identified. Specificity is 0.45, representing the true negative rate, or the probability in which discipline B is correctly identified. Recall/Precision is 0.64. Overall, these numbers help us see how well this model can predict a professor's discipline based off of their salary and years of service.

After generating the ROC plot, I was able to calculate the AUC of the model. The AUC was calculated to be 0.598, meaning the model is a bad predictor of a professor's discipline.

After performing the 10-fold cross validation, I found that accuracy = 0.620, sensitivity = 0.795 and recall = 0.628.

6. Lasso Regression

```
library(glmnet)
y <- as.matrix(Prof_Salaries$discipline)
x <- model.matrix(discipline~ -1+., data=Prof_Salaries)

set.seed(1234)
cv2<-cv.glmnet(x,y,family='binomial')
lasso2<-glmnet(x,y,family='binomial',lambda=cv2$lambda.1se)
coef(lasso2)
```

```
## 11 x 1 sparse Matrix of class "dgCMatrix"
##                              s0
## (Intercept)     -1.486458e-01
## rankAsstProf     .
## rankAssocProf    .
## rankProf         .
```

```
## yrs.since.phd   -1.579462e-02
## yrs.service      .
## sexMale          .
## salary           5.412581e-06
## yrs.service_c    .
## yrs.since.phd_c -5.355404e-03
## logit            3.484447e-01
```

```
#10-fold CV
set.seed(1234)
k=10
data<-Prof_Salaries[sample(nrow(Prof_Salaries)),]
folds<-cut(seq(1:nrow(Prof_Salaries)),breaks=k,labels=F)
diags<-NULL
for(i in 1:k){
  train<-data[folds!=i,]
  test<-data[folds==i,]
  truth<-test$discipline

  fit3<-glm(discipline~yrs.since.phd+salary,data=train,family="binomial")

  probs<-predict(fit3,newdata = test,type="response")

  diags<-rbind(diags,class_diag(probs,truth))
}

summarize_all(diags,mean)
```

```
##         acc      sens      spec       ppv       auc
## 1 0.6330128 0.7755812 0.4811869 0.6388663 0.7142284
```

Based on the results of the Lasso regression, I retained the variables yrs.since.phd and salary. After performing the 10-fold CV, I found the out-of-sample accuracy to be 0.633, which is poor, but slightly better than the out-of-sample accuracy found in the logistic regression in part 5 (0.62).