# HW 8

SDS348 Spring 2021


## Name: Santhosh Saravanan

## EID: sks3648


**This homework is due on April 12, 2021 at 8am. Submit a pdf file on Gradescope.**

*For all questions, include the R commands/functions that you used to find your answer (show R chunk). Answers without supporting code will not receive credit. Write full sentences to describe your findings.*

---

**In this assignment, we will analyze some data from a famous case of alleged gender discrimination in admission to graduate programs at UC Berkeley in 1973. The three variables in the dataset are:**

- `Admit`: Admitted, Rejected
- `Gender`: Male, Female
- `Dept`: Departments A, B, C, D, E, F

```
admissions <- read.csv("https://raw.githubusercontent.com/laylaguyot/datasets/main//admissions.csv")
```

**Question 1: (7 pts)**

```
admissions <- admissions %>%
  mutate(y = ifelse(Admit == "Admitted", 1, 0))
table(admissions$y)
```

**1.1 (1 pt) First, create a dichotomous outcome variable y that is 1 if admitted, 0 otherwise. What percentage of the applicants were admitted?**

```
##
##    0    1
## 2771 1755
```

```
print(1755/(1755+2771))
```

```
## [1] 0.3877596
```

*About 38.77596% of people were admitted to graduate programs at UC Berkeley in 1973.*

```r
fit1 <- glm(y ~ Gender, data = admissions, family = "binomial")
summary(fit1)
```

**1.2 (3 pts) Predict y from Gender using a logistic regression. Is the effect significant? Interpret the effect: what is the odds ratio for admission to graduate school for women compared to men? What is the predicted probability of admission for a female applicant? for a male applicant?**

```
##
## Call:
## glm(formula = y ~ Gender, family = "binomial", data = admissions)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0855  -1.0855  -0.8506   1.2722   1.5442
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.83049    0.05077 -16.357   <2e-16 ***
## GenderMale   0.61035    0.06389   9.553   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6044.3  on 4525  degrees of freedom
## Residual deviance: 5950.9  on 4524  degrees of freedom
## AIC: 5954.9
##
## Number of Fisher Scoring iterations: 4
```

```r
exp(coefficients(fit1))
```

```
## (Intercept)  GenderMale
##   0.4358372   1.8410800
```

```r
oddsRatioWomenMen <- 1/1.8410800
```

```r
predict(fit1,newdata = data.frame(Gender="Female"),type = "response")
```

```
##         1
## 0.3035422
```

```r
predict(fit1,newdata = data.frame(Gender="Male"),type = "response")
```

```
##         1
## 0.4451877
```

*If you were a male applicant who applied to the graduate program at UC Berkeley in 1973 versus being a female applicant, changes the log odds of admission by 1.84108 and this is a significant effect (p-value $<$ 0.05). There is a significant effect of gender on the probability of getting admitted to the graduate school at UC Berkeley. The odds of acceptance to a graduate school program at UC Berkeley for female students is 0.54316 times what they are for male applicants.The predicted probability of admission for a female applicant is 30.35422% and the predicted probability of admissions for a male applicant is 44.51877%.*

```
fit2 <- glm(y ~ Dept, data = admissions, family = "binomial")
summary(fit2)
```

**1.3 (3 pts) Predict y from Dept using a logistic regression. Which department(s) had a significant effect on admission? For which departments are odds of admission higher than department A? Which departments are the most selective? the least selective?**

```
##
## Call:
## glm(formula = y ~ Dept, family = "binomial", data = admissions)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4376  -0.9295  -0.3649   0.9572   2.3419
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.59346    0.06838   8.679   <2e-16 ***
## DeptB       -0.05059    0.10968  -0.461    0.645
## DeptC       -1.20915    0.09726 -12.432   <2e-16 ***
## DeptD       -1.25833    0.10152 -12.395   <2e-16 ***
## DeptE       -1.68296    0.11733 -14.343   <2e-16 ***
## DeptF       -3.26911    0.16707 -19.567   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6044.3  on 4525  degrees of freedom
## Residual deviance: 5189.0  on 4520  degrees of freedom
## AIC: 5201
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coefficients(fit2))
```

```
## (Intercept)       DeptB       DeptC       DeptD       DeptE       DeptF
##  1.81024096  0.95066362  0.29845113  0.28412811  0.18582302  0.03804039
```

```
oddsRatioAB <- 1/0.95066362
oddsRatioAC <- 1/0.29845113
oddsRatioAD <- 1/0.28412811
oddsRatioAE <- 1/0.18582302
oddsRatioAF <- 1/0.03804039
```

```
predict(fit2,newdata = data.frame(Dept="B"),type = "response")
```

```
##         1
## 0.6324786
```

```
predict(fit2,newdata = data.frame(Dept="C"),type = "response")
```

```
##         1
## 0.3507625
```

```
predict(fit2,newdata = data.frame(Dept="D"),type = "response")
```

```
##         1
## 0.3396465
```

```
predict(fit2,newdata = data.frame(Dept="E"),type = "response")
```

```
##         1
## 0.2517123
```

```
predict(fit2,newdata = data.frame(Dept="F"),type = "response")
```

```
##          1
## 0.06442577
```

*Departments C, D, E, and F all a significant effect on admission (p-values $< 0.05$). None of this departments have an odds of admission higher than department A as their odds ratio ( A to C,D,E, and F) are all greater than 1. The most selective departments are F, E, D, and C. The least selective department is B.*

**Question 2: (7 pts)**

```
fit3 <- glm(y ~ Dept + Gender, data = admissions, family = "binomial")
summary(fit3)
```

**2.1 (3 pts) Predict `y` from both `Gender` and `Dept` using a logistic regression. Interpret the coefficient for `Gender`. Controlling for the different departments, is there a significant effect of Gender on admissions? What is the corresponding odds ratio? What can you say about departments A and B compared to the other departments?**

```
##
## Call:
## glm(formula = y ~ Dept + Gender, family = "binomial", data = admissions)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4773  -0.9306  -0.3741   0.9588   2.3613
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.68192    0.09911   6.880 5.97e-12 ***
## DeptB       -0.04340    0.10984  -0.395    0.693
## DeptC       -1.26260    0.10663 -11.841  < 2e-16 ***
## DeptD       -1.29461    0.10582 -12.234  < 2e-16 ***
## DeptE       -1.73931    0.12611 -13.792  < 2e-16 ***
## DeptF       -3.30648    0.16998 -19.452  < 2e-16 ***
## GenderMale  -0.09987    0.08085  -1.235    0.217
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6044.3  on 4525  degrees of freedom
## Residual deviance: 5187.5  on 4519  degrees of freedom
## AIC: 5201.5
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coefficients(fit3))
```

```
## (Intercept)       DeptB       DeptC       DeptD       DeptE       DeptF
##  1.97767415  0.95753028  0.28291804  0.27400567  0.17564230  0.03664494
##  GenderMale
##  0.90495497
```

*If you were a male applicant who applied to the graduate program at UC Berkeley in 1973 versus being a female applicant, changes the log odds of admission by 0.9049551 and this is not a significant effect (p-value $>$ 0.05). However, controlling for departments as we saw in the previous problem, there is a significant effect of Gender on admissions. Departments A and B don't have a significant effect on admissions when either controlling for Gender and not controlling for Gender (p-values $<$ 0.05).One can hypothesize that Departments A and B are less selective than the other respective departments. This conclusion can further be reached with the odds of admission for Departments C, D, E, and F being much lower than Department A (reference group) and B.*

```
fit4 <- glm(y ~ Dept + Gender + (Dept*Gender), data = admissions, family = "binomial")
summary(fit4)
```

**2.2 (4 pts) Predict `y` from both `Gender` and `Dept` using a logistic regression and include an *interaction* term. Compute the odds ratio for admission (Male vs. Female) in each department (A through F). Which departments favor male applicants (i.e., higher odds of admission for `Male`)?**

```
##
## Call:
## glm(formula = y ~ Dept + Gender + (Dept * Gender), family = "binomial",
##     data = admissions)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8642  -0.9127  -0.3821   0.9768   2.3793
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.5442     0.2527   6.110 9.94e-10 ***
## DeptB              -0.7904     0.4977  -1.588  0.11224
## DeptC              -2.2046     0.2672  -8.252  < 2e-16 ***
## DeptD              -2.1662     0.2750  -7.878 3.32e-15 ***
## DeptE              -2.7013     0.2790  -9.682  < 2e-16 ***
## DeptF              -4.1250     0.3297 -12.512  < 2e-16 ***
## GenderMale         -1.0521     0.2627  -4.005 6.21e-05 ***
## DeptB:GenderMale    0.8321     0.5104   1.630  0.10306
## DeptC:GenderMale    1.1770     0.2996   3.929 8.53e-05 ***
## DeptD:GenderMale    0.9701     0.3026   3.206  0.00135 **
## DeptE:GenderMale    1.2523     0.3303   3.791  0.00015 ***
## DeptF:GenderMale    0.8632     0.4027   2.144  0.03206 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6044.3  on 4525  degrees of freedom
## Residual deviance: 5167.3  on 4514  degrees of freedom
## AIC: 5191.3
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coefficients(fit4))
```

```
##      (Intercept)            DeptB            DeptC            DeptD
##       4.68421053       0.45365169       0.11029053       0.11461595
##            DeptE            DeptF       GenderMale DeptB:GenderMale
##       0.06711510       0.01616276       0.34921205       2.29803272
## DeptC:GenderMale DeptD:GenderMale DeptE:GenderMale DeptF:GenderMale
##       3.24461787       2.63817862       3.49825046       2.37068781
```

```
ProbDeptAFemale <- predict(fit4, newdata = data.frame(Gender = "Female", Dept = "A"), type = "response")
ProbDeptAMale <- predict(fit4, newdata = data.frame(Gender = "Male", Dept = "A"), type = "response")
ProbDeptBFemale <- predict(fit4, newdata = data.frame(Gender = "Female", Dept = "B"), type = "response")
ProbDeptBMale <- predict(fit4, newdata = data.frame(Gender = "Male", Dept = "B"), type = "response")
ProbDeptCFemale <- predict(fit4, newdata = data.frame(Gender = "Female", Dept = "C"), type = "response")
ProbDeptCMale <- predict(fit4, newdata = data.frame(Gender = "Male", Dept = "C"), type = "response")
ProbDeptDFemale <- predict(fit4, newdata = data.frame(Gender = "Female", Dept = "D"), type = "response")
ProbDeptDMale <- predict(fit4, newdata = data.frame(Gender = "Male", Dept = "D"), type = "response")
ProbDeptEFemale <- predict(fit4, newdata = data.frame(Gender = "Female", Dept = "E"), type = "response")
ProbDeptEMale <- predict(fit4, newdata = data.frame(Gender = "Male", Dept = "E"), type = "response")
ProbDeptFFemale <- predict(fit4, newdata = data.frame(Gender = "Female", Dept = "F"), type = "response")
ProbDeptFMale <- predict(fit4, newdata = data.frame(Gender = "Male", Dept = "F"), type = "response")
```

```
OddsDeptAFemale <- ProbDeptAFemale/(1-ProbDeptAFemale)
OddsDeptAMale <-  ProbDeptAMale/(1-ProbDeptAMale)
OddsDeptBFemale <- ProbDeptBFemale/(1-ProbDeptBFemale)
OddsDeptBMale <-  ProbDeptBFemale/(1-ProbDeptBFemale)
OddsDeptCFemale <-  ProbDeptCFemale/(1-ProbDeptCFemale)
OddsDeptCMale <-  ProbDeptCMale/(1-ProbDeptCMale)
OddsDeptDFemale <-  ProbDeptDFemale/(1-ProbDeptDFemale)
OddsDeptDMale <-  ProbDeptDMale/(1-ProbDeptDMale)
OddsDeptEFemale <-  ProbDeptEFemale/(1-ProbDeptEFemale)
OddsDeptEMale <-  ProbDeptEMale/(1-ProbDeptEMale)
OddsDeptFFemale <-  ProbDeptFFemale/(1-ProbDeptFFemale)
OddsDeptFMale <-  ProbDeptFMale/(1-ProbDeptFMale)
```

```
OddsDeptAMaleFemale <- OddsDeptAMale/OddsDeptAFemale
OddsDeptBMaleFemale <- OddsDeptBMale/OddsDeptBFemale
OddsDeptCMaleFemale <- OddsDeptCMale/OddsDeptCFemale
OddsDeptDMaleFemale <- OddsDeptDMale/OddsDeptDFemale
OddsDeptEMaleFemale <- OddsDeptEMale/OddsDeptEFemale
OddsDeptFMaleFemale <- OddsDeptFMale/OddsDeptFFemale
```

*Departments C and E favor male applicants because the final OddsDept ratios (males:females) are greater than 1.* ### Question 3: (5 pts)

```
AIC(fit1,fit2,fit3,fit4)
```

**3.1 (1 pt) According to the Akaike information criterion (AIC), which of the four models we created to predict y seem to be a better fit?**

```
##      df      AIC
## fit1  2 5954.891
## fit2  6 5201.020
## fit3  7 5201.488
## fit4 12 5191.284
```

*The smaller the AIC, the better the fit. In this case, the model from 2.2 is the best fit.*

```
anova(fit4, fit3, fit2, test = "LRT")
```

**3.2 (1 pt) According to the analysis of deviance below, which of the three models included seem to significantly lower the deviance?**

```
## Analysis of Deviance Table
##
## Model 1: y ~ Dept + Gender + (Dept * Gender)
## Model 2: y ~ Dept + Gender
## Model 3: y ~ Dept
```

```
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4514     5167.3
## 2      4519     5187.5 -5 -20.2043 0.001144 **
## 3      4520     5189.0 -1  -1.5312 0.215928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

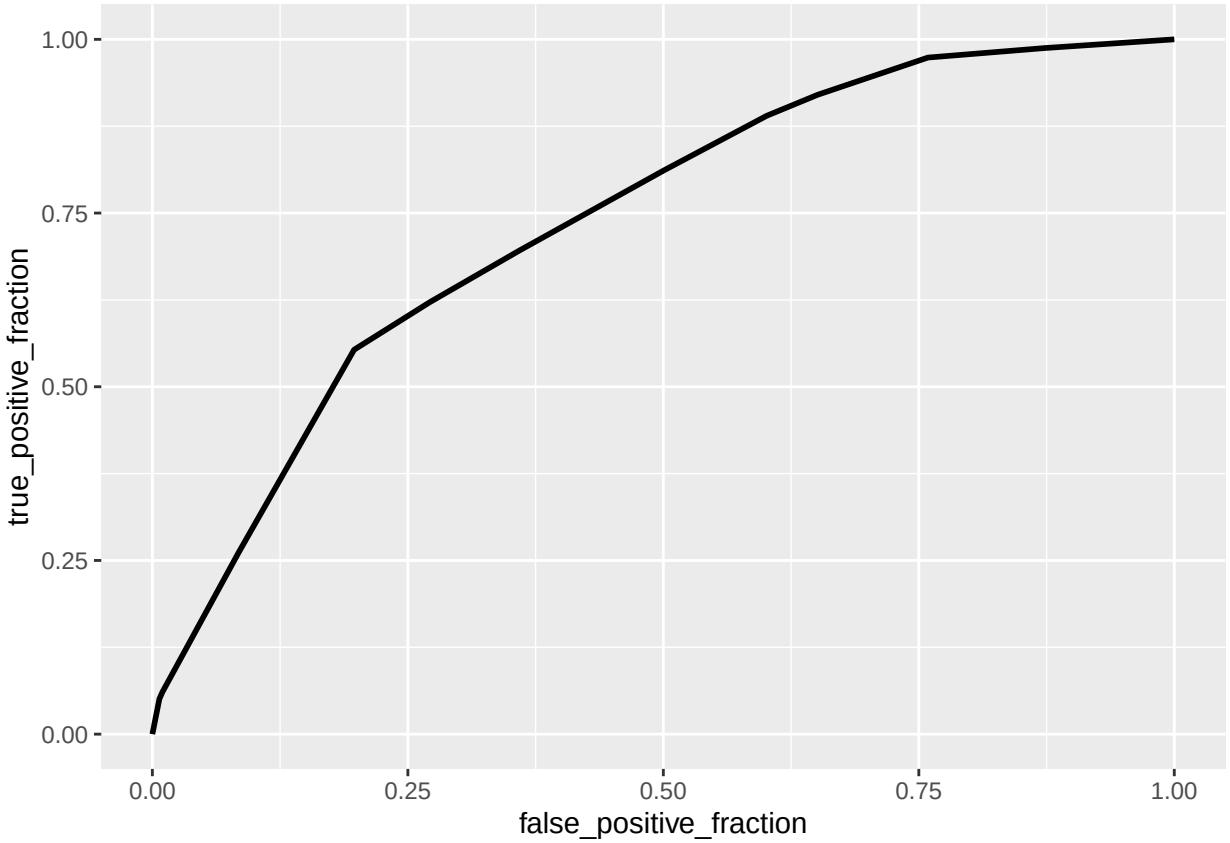
```
anova(fit4,fit3,fit1,test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ Dept + Gender + (Dept * Gender)
## Model 2: y ~ Dept + Gender
## Model 3: y ~ Gender
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      4514     5167.3
## 2      4519     5187.5 -5    -20.2  0.001144 **
## 3      4524     5950.9 -5   -763.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Fit 1 seemed to significantly lower the deviance.*

```
# your code goes here (make sure to add comments)
admissions$prob <- predict(fit4, type = "response")
ROCplot <- ggplot(admissions) +
  geom_roc(aes(d = y, m = prob), n.cuts = 0)
ROCplot
```

**3.3 (3 pts) Consider the model that you believe has the best fit (you can use the two previous questions to help you decide which of the four models it should be!). Save the predicted probabilities of admission for each applicant in the `admission` dataset. Plot the ROC curve and compute the AUC. Using the rules of thumb discussed in lecture, what does the area under the curve indicates?**




```
# Calculate AUC
calc_auc(ROCplot)
```

```
##   PANEL group       AUC
## 1     1    -1 0.7372103
```

*The AUC is the overall measure of model performance as the higher the area under the curve, the better prediction power the model has. According to the rule of thumb, the model is fair in terms of its prediction power as the AUC score is between 0.7 and 0.8.*

**Question 4: (6 pts)**

```
chunks <- admissions %>% select(Dept,Gender) %>% count(Gender,Dept) %>% group_by(Dept)
sampleFrame <- admissions %>% count(y,Gender,Dept) %>% filter(y == 1)
sampleFrame$totalNum <- chunks$n
sampleFrame <- sampleFrame %>% mutate(Admitted=n/totalNum)
chunks$Admitted <- sampleFrame$Admitted
chunks <- rename(chunks,count = n)
chunks <- chunks[order(chunks$count, decreasing = TRUE),]
rm(sampleFrame)
```

**4.1 (4 pts) Using `dplyr` functions on the dataset `admissions`, create a dataframe with counts of applicants of each gender in each department (e.g., number of males who applied to department**

**A) and also the percent of applicants admitted of each gender in each department. Sort the count variable in descending order. What top 2 departments did the majority of women apply to? What about the majority of men? What about the respective selectivity (percent of admitted applicants) in these departments?** *Departments C and E were the two departments that a majority of women applied to. The majority of men applied to Departments A and B. The two departments with high selectivity rates were Departments A and B. The two departments with the lowest selectivity rates were Departments E and F.*

**4.2 (2 pts) Review the first example from the Wikipedia article about the Simpson's paradox. Write a conclusion for this assignment.** *With respect to this paradox, men applying to these departments were more likely to be admitted compared to their women counterparts. There is a massive disparity between the admit rates compared to men and women when taking into account the Admited column rates ( 62% compared to 34%).*

```
##                                       sysname
##                                       "Linux"
##                                       release
##                                 "5.4.0-70-generic"
##                                       version
## "#78-Ubuntu SMP Fri Mar 19 13:29:52 UTC 2021"
##                                      nodename
##                           "MechaChungus-linux64"
##                                       machine
##                                      "x86_64"
##                                         login
##                                 "OmniLordSanta"
##                                          user
##                                 "OmniLordSanta"
##                                  effective_user
##                                 "OmniLordSanta"
```