HW 8

SDS348 Spring 2021

Rose Hedderman

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")
head(admissions)
```

```
##      Admit Gender Dept
## 1 Admitted   Male    A
## 2 Admitted   Male    A
## 3 Admitted   Male    A
## 4 Admitted   Male    A
## 5 Admitted   Male    A
## 6 Admitted   Male    A
```

Question 1: (7 pts)

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?

```
# Create a binary variable coded as 0 and 1
admissions <- admissions %>%
  mutate(y = ifelse(Admit == "Admitted", 1, 0))
head(admissions)
```

```
##      Admit Gender Dept y
## 1 Admitted   Male    A 1
## 2 Admitted   Male    A 1
## 3 Admitted   Male    A 1
## 4 Admitted   Male    A 1
## 5 Admitted   Male    A 1
## 6 Admitted   Male    A 1
```

```
admissions %>%
  summarize(perc = y / sum(y))
```

```
##             perc
## 1   0.0005698006
## 2   0.0005698006
## 3   0.0005698006
## 4   0.0005698006
## 5   0.0005698006
## 6   0.0005698006
## 7   0.0005698006
## 8   0.0005698006
## 9   0.0005698006
## 10  0.0005698006
## 11  0.0005698006
## 12  0.0005698006
## 13  0.0005698006
## 14  0.0005698006
## 15  0.0005698006
## 16  0.0005698006
## 17  0.0005698006
## 18  0.0005698006
## 19  0.0005698006
## 20  0.0005698006
## 21  0.0005698006
## 22  0.0005698006
## 23  0.0005698006
## 24  0.0005698006
## 25  0.0005698006
## 26  0.0005698006
## 27  0.0005698006
## 28  0.0005698006
## 29  0.0005698006
## 30  0.0005698006
## 31  0.0005698006
## 32  0.0005698006
## 33  0.0005698006
## 34  0.0005698006
## 35  0.0005698006
## 36  0.0005698006
## 37  0.0005698006
## 38  0.0005698006
## 39  0.0005698006
## 40  0.0005698006
## 41  0.0005698006
## 42  0.0005698006
## 43  0.0005698006
## 44  0.0005698006
## 45  0.0005698006
## 46  0.0005698006
## 47  0.0005698006
## 48  0.0005698006
## 49  0.0005698006
## 50  0.0005698006
## 51  0.0005698006
## 52  0.0005698006
## 53  0.0005698006
## 54  0.0005698006
## 55  0.0005698006
## 56  0.0005698006
## 57  0.0005698006
## 58  0.0005698006
```

```
## 59  0.0005698006
## 60  0.0005698006
## 61  0.0005698006
## 62  0.0005698006
## 63  0.0005698006
## 64  0.0005698006
## 65  0.0005698006
## 66  0.0005698006
## 67  0.0005698006
## 68  0.0005698006
## 69  0.0005698006
## 70  0.0005698006
## 71  0.0005698006
## 72  0.0005698006
## 73  0.0005698006
## 74  0.0005698006
## 75  0.0005698006
## 76  0.0005698006
## 77  0.0005698006
## 78  0.0005698006
## 79  0.0005698006
## 80  0.0005698006
## 81  0.0005698006
## 82  0.0005698006
## 83  0.0005698006
## 84  0.0005698006
## 85  0.0005698006
## 86  0.0005698006
## 87  0.0005698006
## 88  0.0005698006
## 89  0.0005698006
## 90  0.0005698006
## 91  0.0005698006
## 92  0.0005698006
## 93  0.0005698006
## 94  0.0005698006
## 95  0.0005698006
## 96  0.0005698006
## 97  0.0005698006
## 98  0.0005698006
## 99  0.0005698006
## 100 0.0005698006
##  [ reached 'max' / getOption("max.print") -- omitted 4426 rows ]
```

Only 0.0569 % of applicants are admitted.

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?

```
# Fit a new regression model
fit1 <- glm(y ~ Gender, data = admissions, family = binomial(link = "logit"))
summary(fit1)
```

```
##
## Call:
## glm(formula = y ~ Gender, family = binomial(link = "logit"),
##     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
```

```
# Interpret coefficients as odds ratios
table(admissions$y, admissions$Gender)
```

```
##
##     Female Male
##   0   1278 1493
##   1    557 1198
```

```
# Odds of admission for male
odds_M = (557/1835) / (1278/1835)
# Odds of admission for F
odds_F = (1198/2691) / (1493/2691)
# Odds ratio of malignancy, M compared to S

print("odds_f")
```

```
## [1] "odds_f"
```

```
odds_F
```

```
## [1] 0.8024113
```

```
print("odds_m")
```

```
## [1] "odds_m"
```

```
odds_M
```

```
## [1] 0.4358372
```

```
print("odds m to f")
```

```
## [1] "odds m to f"
```

```
odds_F / odds_M
```

```
## [1] 1.84108
```

```
# Compare to the exponentiated coefficients of the model
exp(coef(fit1))
```

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

The p-value is pracically zero, so the results are significant. The odds of admittance for a male applicant are 1.84 times greater than the odds for a female applicant. The predicted probability for a female applicant is .8 while the odds of male admittance is 0.435. The odds ratio female to male is 1.84.

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?

```
# logistic regression
fit2 <- glm(y ~ Dept, data = admissions, family = binomial(link="logit"))
summary(fit2)
```

```
##
## Call:
## glm(formula = y ~ Dept, family = binomial(link = "logit"), 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
```

```
# odds ratio
exp(coef(fit2))
```

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

```
# Interpret coefficients as odds ratios
table(admissions$y, admissions$Dept)
```

```
##
##       A   B   C   D   E   F
##   0 332 215 596 523 437 668
##   1 601 370 322 269 147  46
```

```
# Odds of admission for Dept A
odds_A = (601/393) / (332/393)
# Odds of admission for Dept B
odds_B = (370/585) / (215/585)
# Odds of admission for Dept C
odds_C = (322/918) / (596/918)
# Odds of admission for Dept D
odds_D= (269/792) / (523/792)
# Odds of admission for Dept E
odds_E = (147/584) / (437/584)
# Odds of admission for Dept F
odds_F = (46/714) / (668/714)

odds_A
```

```
## [1] 1.810241
```

```
odds_B
```

```
## [1] 1.72093
```

```
odds_C
```

```
## [1] 0.5402685
```

```
odds_D
```

```
## [1] 0.5143403
```

```
odds_E
```

```
## [1] 0.3363844
```

```
odds_F
```

```
## [1] 0.06886228
```

Departments C, D, E and F all had significant effects on admission. Department F is the most selective. The sleast selective department was B. None of the departments had higher admission rates than Department A.

Question 2: (7 pts)

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?

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

```
##
## Call:
## glm(formula = y ~ Dept + Gender, family = binomial(link = "logit"),
##     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
```

```
# odds ratio
exp(coef(fit3))
```

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

For every one male applicant, probability of admission decreases 0.099. Department A and B are much less selective than any other departments. Contorlling for departments, there was not a significant effect on Gender on admissions for each department.

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`)?

```
# logistic regression
fit4 <- glm(y ~ Dept * Gender, data = admissions, family = binomial(link="logit"))
summary(fit4)
```

```
##
## Call:
## glm(formula = y ~ Dept * Gender, family = binomial(link = "logit"),
##     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
```

```
# odds ratio
exp(coef(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
```

Departments B, C, D, E, and F all favor male applicants.

Question 3: (5 pts)

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?

```
# calculate aic values for each model fit
summary(fit1)$aic
```

```
## [1] 5954.891
```

```
summary(fit2)$aic
```

```
## [1] 5201.02
```

```
summary(fit3)$aic
```

```
## [1] 5201.488
```

```
summary(fit4)$aic
```

```
## [1] 5191.284
```

The second model has the lowest AIC and has the better.

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

```
# use an anova analysis to determine deviance
anova(fit2, fit3, fit4, test = "LRT")
```

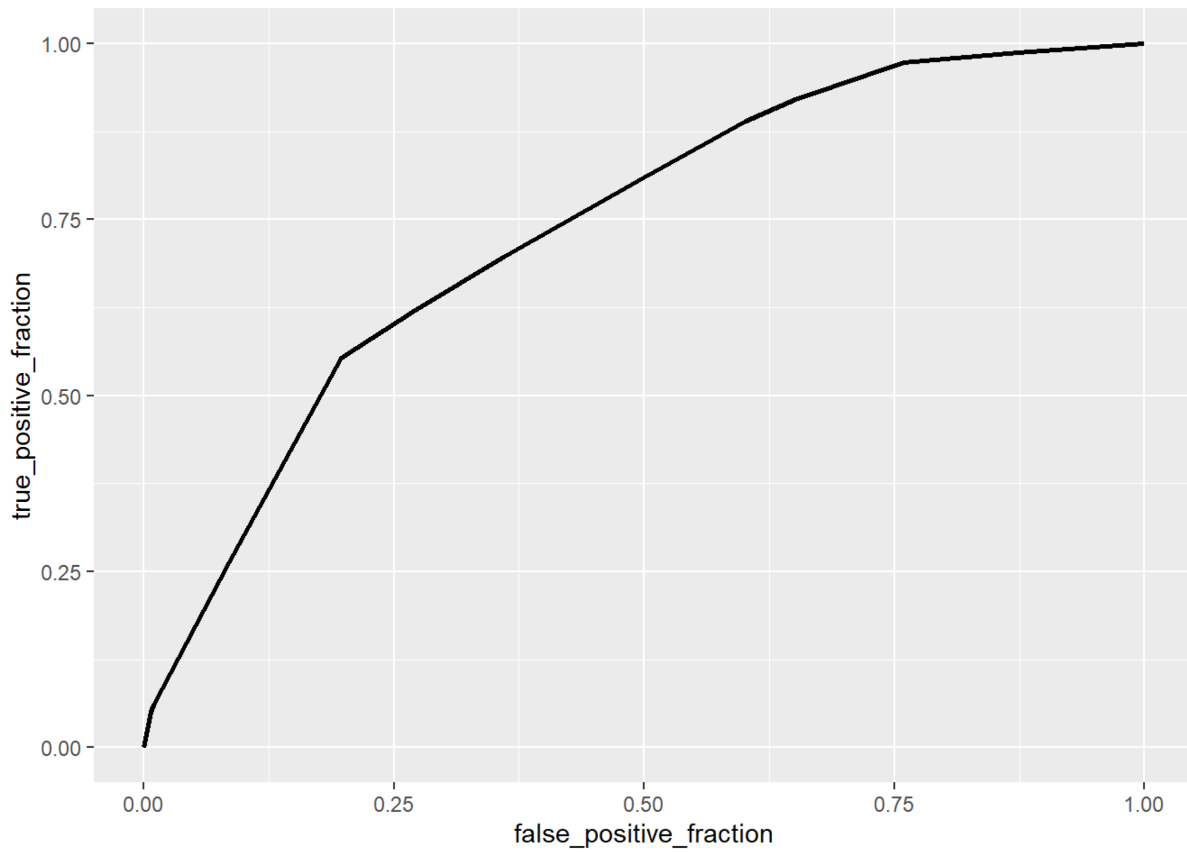
```
## Analysis of Deviance Table
##
## Model 1: y ~ Dept
## Model 2: y ~ Dept + Gender
## Model 3: y ~ Dept * Gender
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4520     5189.0
## 2      4519     5187.5  1   1.5312 0.215928
## 3      4514     5167.3  5  20.2043 0.001144 **
## ---
## Signif.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

The fourth model, the interaction term momdel, seems to significantly lower deviance.

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?

```
# Model 4 is the best fit (fit4)
admissions$prob <- predict(fit4, type = "response")

ROCplot1 <- ggplot(admissions) +
  geom_roc(aes(d = y, m = prob), n.cuts = 0)
ROCplot1
```




```
#AUC
calc_auc(ROCplot1)
```

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

On average, 73.7% of the time male applicants will have higher acceptance rates than female applicants.

Question 4: (6 pts)

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?

```
# sorted by female count
ad_info <- admissions %>%
  group_by(Dept) %>%
  summarize(m_ct = sum(Gender == 'Male'),
            fm_ct = sum(Gender == 'Female'),
            m_accept = (sum(Gender == 'Male' & y == 1)/sum(Gender == 'Male')),
            f_accept = (sum(Gender == 'Female' & y == 1)/sum(Gender == 'Female'))) %>%
  arrange(desc(fm_ct))

# sortedby male count
head(ad_info)
```

```
## # A tibble: 6 x 5
##   Dept   m_ct fm_ct m_accept f_accept
##   <fct> <int> <int>    <dbl>    <dbl>
## 1 C       325   593   0.369    0.341
## 2 E       191   393   0.277    0.239
## 3 D       417   375   0.331    0.349
## 4 F       373   341   0.0590   0.0704
## 5 A       825   108   0.621    0.824
## 6 B       560    25   0.630    0.68
```

```
ad_info <- admissions %>%
  group_by(Dept) %>%
  summarize(m_ct = sum(Gender == 'Male'),
            fm_ct = sum(Gender == 'Female'),
            m_accept = (sum(Gender == 'Male' & y == 1)/sum(Gender == 'Male')),
            f_accept = (sum(Gender == 'Female' & y == 1)/sum(Gender == 'Female'))) %>%
  arrange(desc(m_ct))
head(ad_info)
```

```
## # A tibble: 6 x 5
##   Dept   m_ct fm_ct m_accept f_accept
##   <fct> <int> <int>    <dbl>    <dbl>
## 1 A       825   108   0.621    0.824
## 2 B       560    25   0.630    0.68
## 3 D       417   375   0.331    0.349
## 4 F       373   341   0.0590   0.0704
## 5 C       325   593   0.369    0.341
## 6 E       191   393   0.277    0.239
```

The majority of women applied to Dept C and E, while the majority of men applied to Dept A and B. Department A accepted 62% of men and 82% of women that applied. Department B accepted 63% of men and 68% of women that applied. Department C accepted 36% of men and 34% of women that applied. Department E accepted 27.7% of men and 23.9% of women that applied.

4.2 (2 pts) Review the first example from the Wikipedia article (https://en.wikipedia.org/wiki/Simpson%27s_paradox) about the Simpson's paradox. Write a conclusion for this assignment.

In conclusion, the four departments were biased towards women six were biased towards men during admissions. However, Bickel found that women were likely to apply to more competitive departments with low rates of admission while men were more likely to apply to less competitive departments among qualified applicants.

```
##       sysname        release        version       nodename        machine 
##     "Windows"       "10 x64"  "build 19042"     "ROSE-XPS"       "x86-64" 
##         login           user effective_user 
##       "roseh"        "roseh"        "roseh" 
```