ST-518 HW 3

Chris Eckerson

R Question

1) GRE, GPA, & Prestige vs Admissions

(4 points) A researcher is interested in studying how (if?) GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of undergraduate institution are associated with admission into graduate school. The response variable, Admit, is a binary variable (1 == admit). This dataset can be found in admissions.csv available for download right underneath where you downloaded/opened this homework assignment. Treat the variables GRE and GPA as continuous, and treat RANK, which takes values 1 through 4, as a factor variable. A rank of 1 indicates that the student's undergraduate institution has the highest prestige, while a rank of 4 indicates that it has the lowest prestige.

```r
# Admit = log, gre & gpa = num, rank = factor w 1 = best.
admissions <- read_csv("../../Data/admissions.csv",
                       col_types = "innf")
ad <- admissions |> mutate(
  rank = factor(rank, levels = c(1, 2,3,4)) #, levels = c(4,3,2,1))
)
# str(admissions)
str(ad)
```

```
tibble [400 x 4] (S3: tbl_df/tbl/data.frame)
 $ admit: int [1:400] 0 1 1 1 0 1 1 0 1 0 ...
 $ gre  : num [1:400] 380 660 800 640 520 760 560 400 540 700 ...
 $ gpa  : num [1:400] 3.61 3.67 4 3.19 2.93 ...
 $ rank : Factor w/ 4 levels "1","2","3","4": 3 3 1 4 4 2 1 2 3 2 ...
```

a Fit models

(a) Fit a logistic regression model to these data, with the variable admit as the response and gpa, gre, and rank as explanatory variables. Fit another model without gre. Comment on how these models are different.

```
mod_full <- glm(admit ~ gpa + gre + rank,
            family = binomial(link = "logit"),
            data = ad)

mod_red <- glm(admit ~ gpa + rank,
            family = binomial(link = "logit"),
            data = ad)

s_red <- summary(mod_red)

s_full <- summary(mod_full)

s_red
```

```
Call:
glm(formula = admit ~ gpa + rank, family = binomial(link = "logit"),
    data = ad)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.4636     1.1003  -3.148 0.001645 **
gpa           1.0521     0.3102   3.392 0.000694 ***
rank2        -0.6810     0.3141  -2.168 0.030181 *
rank3        -1.3919     0.3419  -4.071 4.68e-05 ***
rank4        -1.5943     0.4152  -3.840 0.000123 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 462.88  on 395  degrees of freedom
AIC: 472.88

Number of Fisher Scoring iterations: 4
```

```
s_full
```

```
Call:
glm(formula = admit ~ gpa + gre + rank, family = binomial(link = "logit"),
    data = ad)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.989979   1.139951  -3.500 0.000465 ***
gpa          0.804038   0.331819   2.423 0.015388 *
gre          0.002264   0.001094   2.070 0.038465 *
rank2       -0.675443   0.316490  -2.134 0.032829 *
rank3       -1.340204   0.345306  -3.881 0.000104 ***
rank4       -1.551464   0.417832  -3.713 0.000205 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 458.52  on 394  degrees of freedom
AIC: 470.52

Number of Fisher Scoring iterations: 4
```

```
# Smaller mod first.
anova(mod_red, mod_full)
```

```
Analysis of Deviance Table

Model 1: admit ~ gpa + rank
Model 2: admit ~ gpa + gre + rank
  Resid. Df Resid. Dev Df Deviance
1       395     462.88
2       394     458.52  1   4.3578
```

```
s_red$coefficients |> exp()
```

```
              Estimate Std. Error     z value Pr(>|z|)
(Intercept) 0.03131623   3.005111  0.04294460 1.001646
gpa         2.86352334   1.363657 29.72120143 1.000694
rank2       0.50612454   1.369087  0.11443972 1.030641
rank3       0.24860056   1.407620  0.01705876 1.000047
rank4       0.20306061   1.514682  0.02150075 1.000123
```

```
# s_red$coefficients[1:5] |> exp()
s_full$coefficients |> exp()
```

```
             Estimate Std. Error     z value Pr(>|z|)
(Intercept) 0.0185001   3.126615  0.03019338 1.000465
gpa         2.2345451   1.393501 11.28099028 1.015507
gre         1.0022670   1.001095  7.92374081 1.039214
rank2       0.5089309   1.372302  0.11834270 1.033374
rank3       0.2617923   1.412423  0.02062602 1.000104
rank4       0.2119375   1.518665  0.02440100 1.000205
```

The reduced model doesn't have GRE, the reduced model's intercept is higher, gpa estimate is higher, and ranks are lower.

GPA, GRE, and all Ranks are significant in the full model, the same for the reduced except it doesn't have GRE.

Drop in deviance is 4.3578. Indicating the full model may be better (p-val = 0.0368, drop in deviance test, df = 1).

```
# 1 - pchisq(4.3578, 1) #lower.tail = F
pchisq(4.3578, 1, lower.tail = F)
```

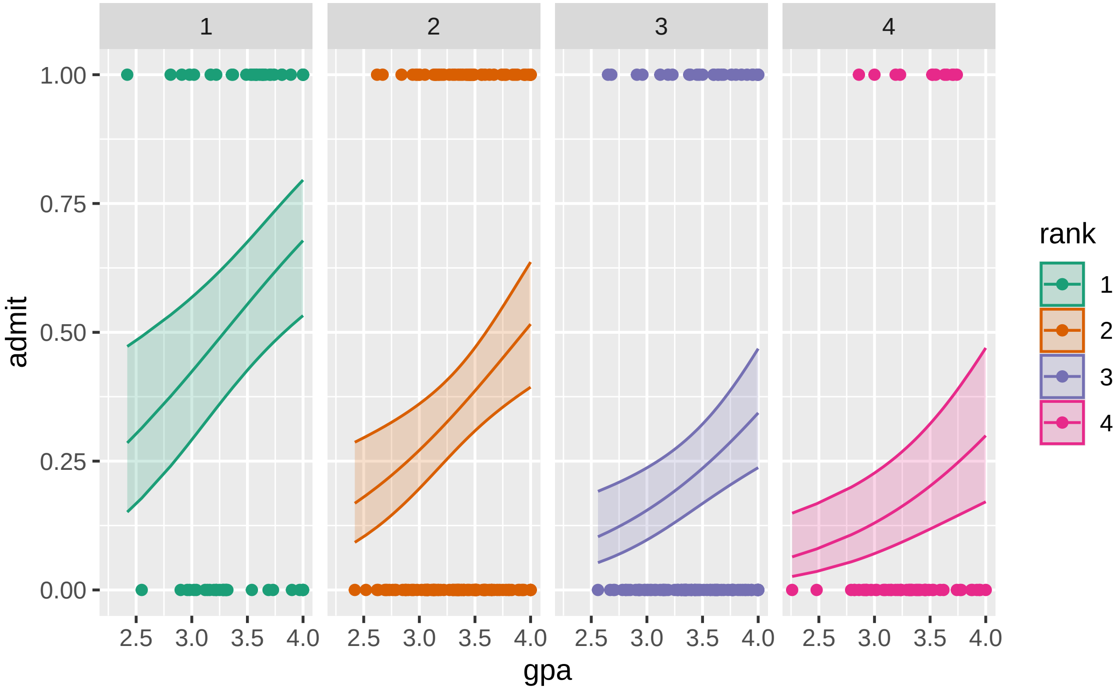
```
[1] 0.03683985
```

b. Scatter plot

(b) Produce a scatter plot of admit against GPA and overlay 4 separate fitted lines, one for each rank, from the regression with gpa and rank as explanatory variables.

```
logistic <- function(x){exp(x)/(1 + exp(x))}

# Obtain 95% pointnwise confidence bands from predict.glm()
```

```r
glm_pred <- predict.glm(mod_red, type="link", se.fit=TRUE)
low <- glm_pred$fit - 1.96 * glm_pred$se.fit
upp <- glm_pred$fit + 1.96 * glm_pred$se.fit

# back-transform everything to the data scale
glm_fit <- logistic(glm_pred$fit)
glm_lower <- logistic(low)
glm_upper <- logistic(upp)

# augment the Donner data frame
augment_df <- as.data.frame(cbind(ad, glm_fit, glm_lower, glm_upper))


augment_df |>  ggplot() +
  aes(y = admit, x = gpa, color = rank) +
  geom_point() +

  geom_line(aes(x = gpa,
        y = glm_fit,
         color = rank)) +

  geom_ribbon(
    aes(x = gpa,
        fill = rank,
        ymin = glm_lower,
        ymax = glm_upper),
    alpha = .2
  ) +

  scale_fill_brewer(palette ="Dark2") +
  scale_color_brewer(palette ="Dark2")+
  facet_grid(
      cols = vars(rank)
      )
```



c. Summary

(c) Write a short paragraph discussing your findings

It looks like the prestige of one's undergraduate institution has an effect on admissions to grad school. With a 4.0 GPA, someone from a prestigious school has between 53% & 80% probability of getting admitted. Whereas someone with the same gpa from a low ranked school has between 17% & 47% probability of getting admitted.

```
# augment_df |> filter(rank == 1 & gpa == 4.0)
# augment_df |> filter(rank == 4 & gpa == 4.0)
```

Conceptual Questions

2. Donner

(3 points) In 1846, the Donner and Reed families left Springfield, Illinois, for California by covered wagon. Along the way, more families and individuals joined the Donner Party, as it came to be known, until it reached its full size of 87 people. The group become stranded in

the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued in April 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

a. independence

(a) One assumption underlying the correct use of logistic regression is that observations are independent of each other. Is there some basis for thinking this assumption might be violated in the Donner Party data?

There may be some violations of this assumption. The party was composed of families and groups that traveled together. Those groups are not independent. Families will share similar genetics. Their health may be affected as a group by their diets before the incident. Groups that travel together would share a similar effect.

```r
donner = case2001
# donner |> head()
```

b. The over 50's

(b) Why should one be reluctant to draw conclusions about the ratio of male and female odds of survival for Donner Party members over 50? (Hint: Look again the graph of the Donner Party data from lecture, where status is plotted against age.)

There are fewer observations in the over 50 crowd, especially among the female group. This makes inference outside the bounds of the data precarious.

c. Survival is 50%

(c) In this week's lecture, it was found that the estimated logistic regression equation is:

- $\widehat{logit}(p) = 1.63 - 0.078 Age + 1.60 Female$, where Female is an indicator variable equal to one for females and zero for males.

What is the age at which the estimated probability of survival is 50% for women? What about for men?

```r
f <- 1
# logOdds = 1.63 - 0.078*a + 1.60*f
agef = (1.63 + 1.60*f - log(1)) / (0.078)

f <- 0
```

```
agem = (1.63 + 1.60*f - log(1)) / (0.078)

agef |> round()
```

```
[1] 41
```

```
agem |> round()
```

```
[1] 21
```

For the Donner party, at 41 years old the estimated probability of survival for women is 50%. It's 21 for men.

3. The common brushtail possum

(3 points) The common brushtail possum of the Australia region is a bit cuter than its distant cousin, the American opossum. We consider 104 brushtail possums from two regions in Australia, where the possums may be considered a random sample from a larger population. The first region is Victoria, and the second region consists of New South Wales and Queensland.

We use logistic regression to differentiate between possums in these two regions. The outcome variable population takes value 1 if the possum is from Victoria and 0 if it is from New South Wales and Queensland. Five predictors are considered: sex_male, an indicator for a possum being male, head_length, skull_width, total_length, and tail_length. A full and reduced logistic model are summarized in the following table:

```
----------  Full Model -----------------------|--- Reduced Model--------------------------
            Estimate  SE      Z      Pr(> |Z|) | Estimate      SE          Z         Pr(> |Z|)
(Intercept)  39.2349   11.5368 3.40    0.0007    |   33.5095     9.9053      3.38      0.0007
sex_male     -1.2376    0.6662  -1.86   0.0632    | -1.4207      0.6457     -2.20      0.0278
head_length  -0.1601    0.1386  -1.16   0.2480    |
skull_width  -0.2012    0.1327  -1.52   0.1294    | -0.2787      0.1226     -2.27      0.0231
total_length 0.6488     0.1531   4.24   0.0000    |  0.5687      0.1322      4.30      0.0000
tail_length  -1.8708    0.3741  -5.00   0.0000    | -1.8057      0.3599     -5.02      0.0000
```

```
full_m <- c(39.2349, -1.2376, -0.1601, -0.2012, 0.6488, -1.8708)
red_m <- c(33.5095, -1.4207, -0.2787, 0.5687, -1.8057)
```

a. The remaining estimates

(a) The variable head_length was taken out for the reduced model based on its p-value in the full model. Why did the remaining estimates change between the two models?

There was less information available to use in the equation. While head length may have been essentially useless, it was still being used in the full model. Without it in there, whatever effect it was estimated to have had to be taken up by the other variables.

b. Probability it's from Victoria?

(b) Suppose we see a male possum with a 65 mm wide skull, a 32 cm long tail, and a total length of 80 cm. If we know this possum was captured in the wild in Australia, what is the probability that this possum is from Victoria (using the reduced model)?

a male possum
65 mm wide skull 32 cm long tail of 80 cm total length

```
red_m
```

```
[1] 33.5095 -1.4207 -0.2787  0.5687 -1.8057
```

```
M <- 1
sw <- 65
t <- 32
len <- 80

# ln(odds) <- 33.5095 -1.4207*M -0.2787*sw + 0.5687*len - 1.8057*t

odds <- exp(33.5095) + exp(-1.4207*M) + exp(-0.2787*sw) + exp(0.5687*len) + exp(-1.8057*t)
odds
```

```
[1] 5.736731e+19
```

```
prob <- odds/(1 + odds)
prob*100
```

```
[1] 100
```

The probability of this possum being from Victoria is estimated to be 100%. I am assuming that these were all given in the correct dimensions. If they were all supposed to be in mm or cm, then it could be different.

```
M <- 1
sw <- 65/1000
t <- 32/100
len <- 80/100

odds <- exp(33.5095) + exp(-1.4207*M) + exp(-0.2787*sw) + exp(0.5687*len) + exp(-1.8057*t)
odds
```

```
[1] 3.572654e+14
```

```
prob <- odds/(1 + odds)
prob*100
```

```
[1] 100
```

Making all of the measurements in meters, didn't change it.