ST 518 - Homework 3

Nora Quick

R Question:

1. 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
adm <- read.csv("admissions.csv")
head(adm)
```

```
##   admit gre  gpa rank
## 1     0 380 3.61    3
## 2     1 660 3.67    3
## 3     1 800 4.00    1
## 4     1 640 3.19    4
## 5     0 520 2.93    4
## 6     1 760 3.00    2
```

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.

```r
glm_gre <- glm(admit ~ gre + gpa + rank, family = binomial(link = "logit"), data = adm)
summary(glm_gre)
```

```
##
## Call:
## glm(formula = admit ~ gre + gpa + rank, family = binomial(link = "logit"),
##     data = adm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5802  -0.8848  -0.6382   1.1575   2.1732
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.449549   1.132846  -3.045  0.00233 **
## gre          0.002294   0.001092   2.101  0.03564 *
## gpa          0.777014   0.327484   2.373  0.01766 *
## rank        -0.560031   0.127137  -4.405 1.06e-05 ***
## ---
## 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: 459.44  on 396  degrees of freedom
## AIC: 467.44
##
## Number of Fisher Scoring iterations: 4
```

```r
glm_no <- glm(admit ~ gpa + rank, family = binomial(link = "logit"), data = adm)
summary(glm_no)
```

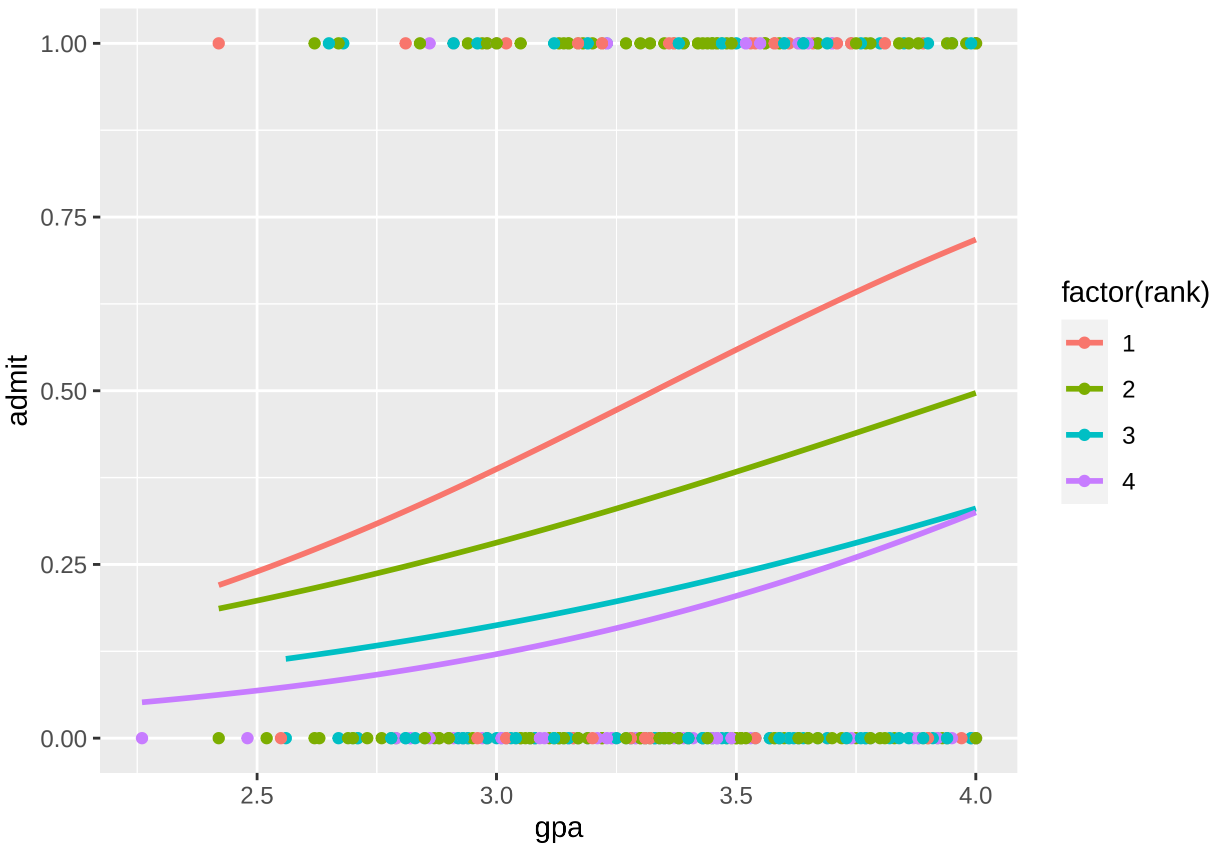
```
##
## Call:
## glm(formula = admit ~ gpa + rank, family = binomial(link = "logit"),
##     data = adm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4599  -0.8974  -0.6657   1.1516   2.1781
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.8826     1.0916  -2.641 0.008273 **
## gpa           1.0270     0.3064   3.352 0.000801 ***
## rank         -0.5822     0.1263  -4.609 4.05e-06 ***
## ---
## 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: 463.93  on 397  degrees of freedom
## AIC: 469.93
##
## Number of Fisher Scoring iterations: 3
```

Without gre the p-values of gpa and rank become much more significant. This leads me to think that gre is not an important variable to admitance.

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

```r
ggplot(adm, aes(x=gpa, y=admit, color=factor(rank))) +
  geom_point() +
  geom_smooth(method=glm, method.args = list(family = "binomial"), se=FALSE)
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



c. Write a short paragraph discussing your findings.

As we could have guessed the higher the ranking of the school (the more prestegious) the higher admitance. GPA is also an important factor to admitance. The higher it is the more likely someone will get admitted. Almost shockingly the GRE does not play a large factor into someone getting admitted.

Conceptual Questions:

2. 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. 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 inthe Donner Party data?

I would think that families would make the data not so independent. The parents of families might have sacrificed themselves more so that their children could go on making their death dependent on family.

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.)

Older men would be "more capable" to go out on explorations to find food or help. They would also be head of households and therefore be responsible for large parties of people.

Additionally, based on the graph there are a lot more older men in general.

c. In this week's lecture, it was found that the estimated logisitic regression equation is: 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?

As the TA suggested I imputted 0.5 into both equations from the lecture slides for survival. I inputted 0.5 to both and found that for women the age was 35 and for men it was 14.5.

3. 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:

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?

The estimates changed between the two models because head_length was not imporant.

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)?

There is a 49% chance that it is from Australia.