Lab4

Ben Tankus

4/22/2021

Lab 4 Assignment

Question

Here we will return to the UCBAdmissions dataset and fit a logistic regression model for the count of students admitted (out of the total applicants) for each combination of the factors gender and department. The usual drill applies – submit your answers as an RMarkdown document, following the instructions given in the previous labs.

(a) Construct an informative ggplot() of the empirical logits of admission proportion vs gender and department. It's up to you what aesthetics to map to which variables – there is more than one right answer here.

Some tips for part (a):

- You can use the `group` argument to geom_line() to connect points within a group – for instance, given a plot with Gender on the x axis and a variable called `eLogits` on the y, you could add `geom_line(aes(group = Dept, x = Gender, y = eLogits))`, where the slope of the connecting lines would correspond to the sign and magnitude of the difference in empirical logits (log of observed odds ratio) between genders within each department.
- You can also incorporate information about the total number of applicants to each department into your plot. For instance, supposing you had a data frame with separate columns containing counts of "Admitted" and "Rejected" by sex and department, you could map the number of applicants to the size of the plot points using `geom_point(size = Admitted + Rejected)`.

```
#to_cases
UCBAdmissions_case <- expand.dft(UCBAdmissions)
head(UCBAdmissions_case)
```

```
##      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
```

```
UCBAdmissions_case %>% count(Gender)
```

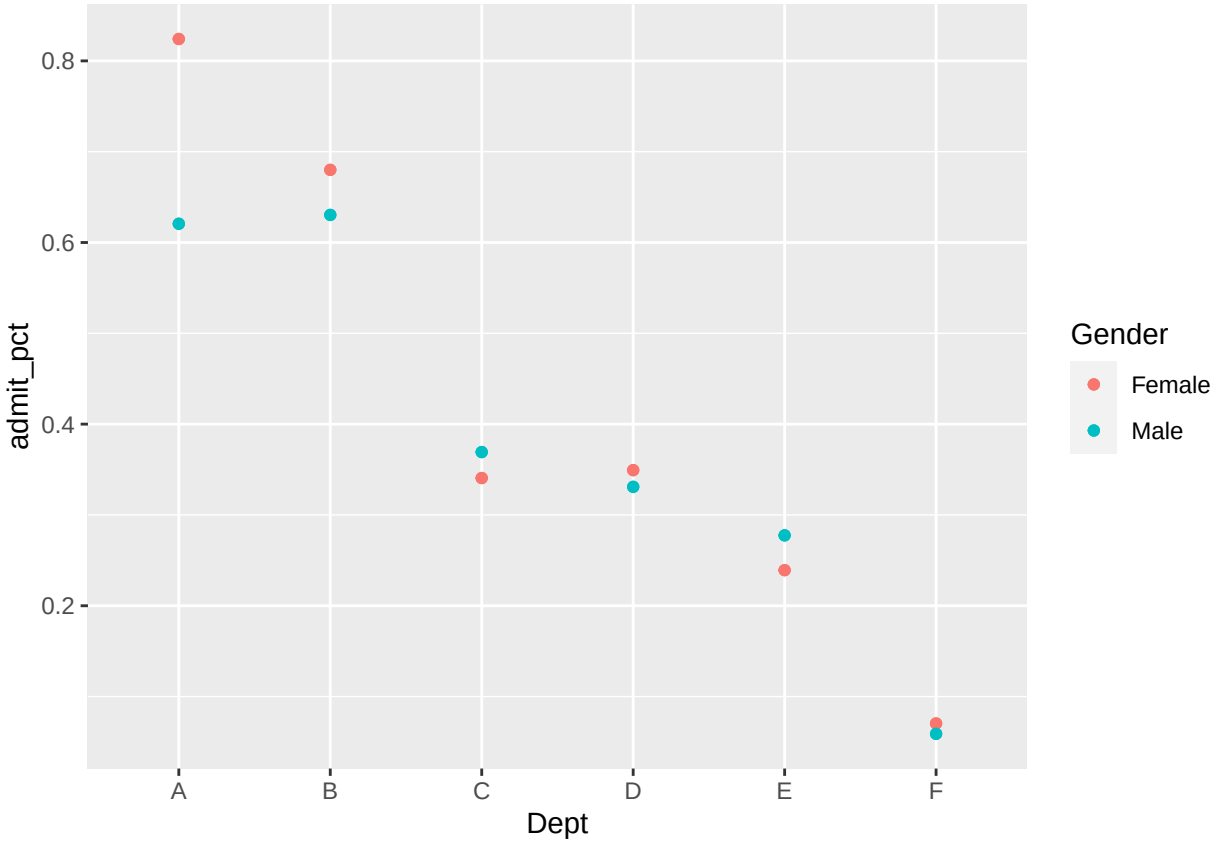
```
##   Gender    n
## 1 Female 1835
## 2   Male 2691
```

```
Admissions_by_gender <- UCBAdmissions_case %>%
  group_by(Admit, Dept, Gender) %>%
  summarize(counts =  n())
```

```
## `summarise()` regrouping output by 'Admit', 'Dept' (override with `.groups` argument)
```

```
Admissions_test <- pivot_wider(Admissions_by_gender, names_from = Admit, values_from = counts)

Admissions_test$admit_pct <- (Admissions_test$Admitted / (Admissions_test$Admitted + Admissions_test$Re
```

```
ggplot(data=Admissions_test, aes(x=Dept, y=admit_pct, color=Gender)) + geom_point() + geom_point(aes(gr
```



(b) Based on your plot from (a), which variable (gender or department) appears to account for more of the variability in admissions? Explain.

According to the plot it looks like for a large part, gender tracks very closely to admission. Admission percentage however changes greatly based on department.

(c) Fit an appropriate (binomial) logistic regression model for admissions. What is the estimated dispersion parameter? Is there evidence of lack of fit?

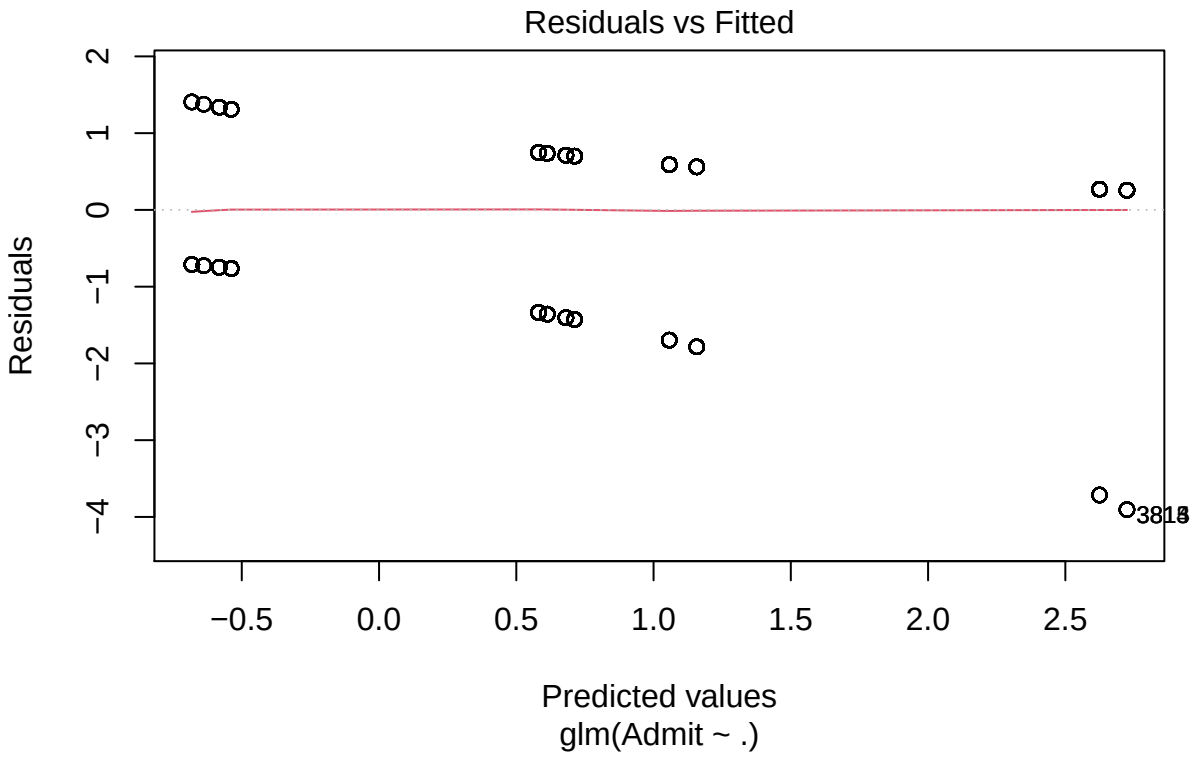
This model is fit fairly well using a binomial model as the dispersion parameter is very close to 1. It also converged in a small number of iterations, suggesting a good fit.

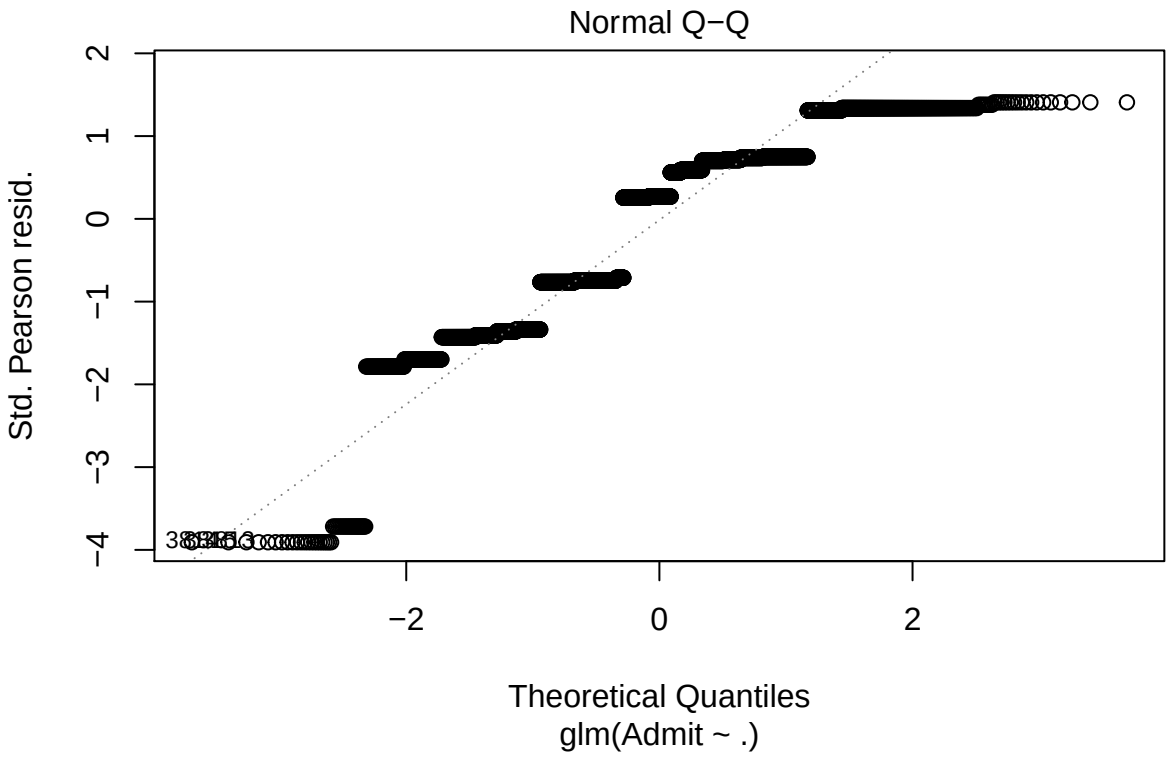
```
UCBA_case <- expand.dft(UCBAdmissions)

test1 <- glm(data = UCBA_case, Admit ~ .  , family = binomial)
test.fit <- summary(test1)
test.fit
```

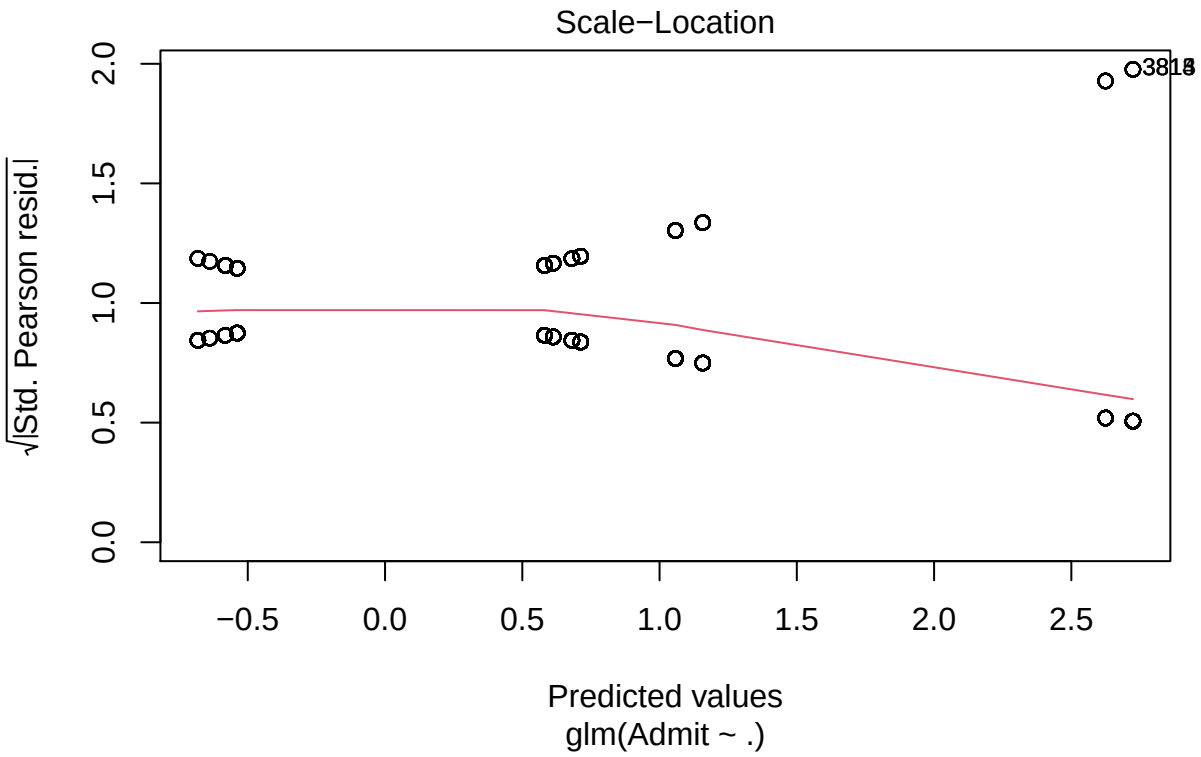
```
##
## Call:
## glm(formula = Admit ~ ., family = binomial, data = UCBA_case)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3613  -0.9588   0.3741   0.9306   1.4773
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.68192    0.09911  -6.880 5.97e-12 ***
## GenderMale   0.09987    0.08085   1.235    0.217
## 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 ***
## ---
## 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
```

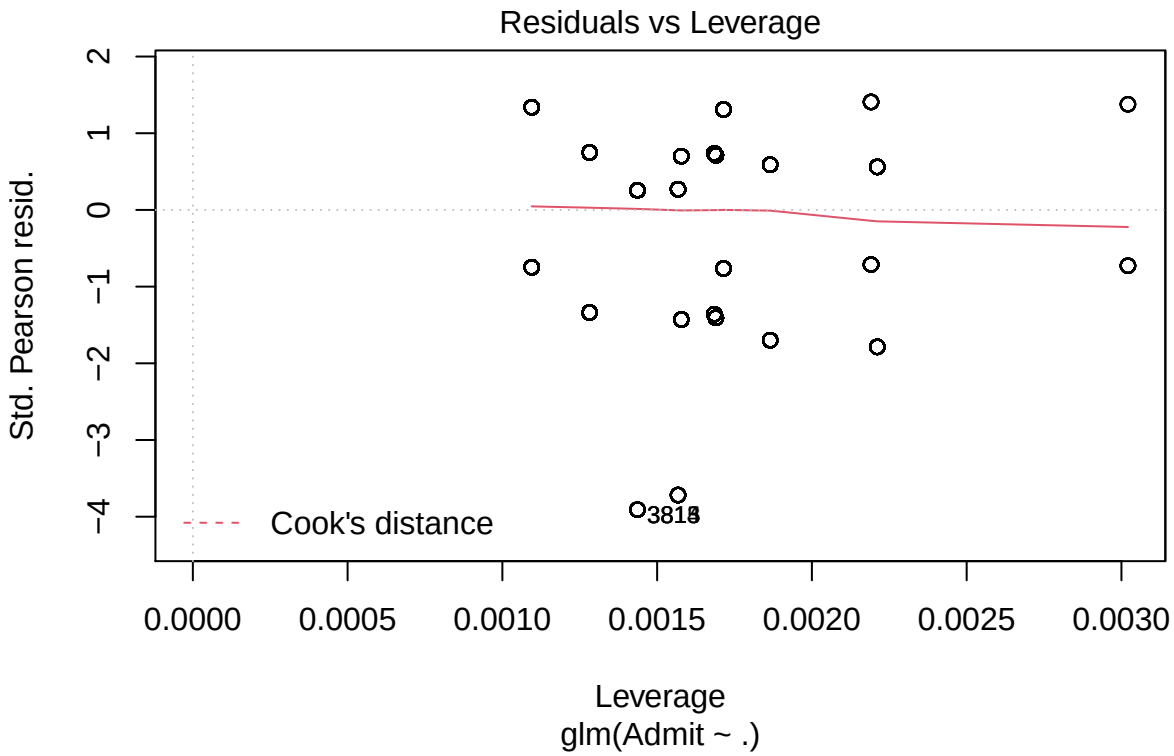
(d) Construct a plot of residuals vs fitted values (try just `plot`-ing your fitted model object). From this plot, can you identify a source for any fit problems encountered in part (c)?

```
plot(test1)
```


Residuals vs Fitted
Residuals
Predicted values
glm(Admit ~ .)
2
1
0
−1
−2
−3
−4
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
381


Normal Q−Q
Std. Pearson resid.
2
1
0
−1
−2
−3
−4
−2
0
2
Theoretical Quantiles
glm(Admit ~ .)


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
2.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
Predicted values
glm(Admit ~ .)



The residuals vs fitted values looks like a straightforward plot with minimal sources of fit issues.

(d) Refit the binomial model above, but excluding the data from department A. Now what is the estimated dispersion parameter? Based onthe p-value, what would you conclude about the effect of Gender on admissions (to departments other than A) using this model?

```
test1 <- glm(data = UCBA_case, Admit ~ .  , family = binomial, subset = Dept != 'A')

summary(test1)
```

```
##
## Call:
## glm(formula = Admit ~ ., family = binomial, data = UCBA_case,
##     subset = Dept != "A")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3483  -0.9567   0.3675   0.9164   1.4155
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.51349    0.11936  -4.302 1.69e-05 ***
## GenderMale  -0.03069    0.08676  -0.354    0.724
## DeptC        1.14008    0.12188   9.354  < 2e-16 ***
## DeptD        1.19456    0.11984   9.968  < 2e-16 ***
## DeptE        1.61308    0.13928  11.581  < 2e-16 ***
```

```
## DeptF        3.20527    0.17880  17.927  < 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: 4511.1  on 3592  degrees of freedom
## Residual deviance: 3974.2  on 3587  degrees of freedom
## AIC: 3986.2
##
## Number of Fisher Scoring iterations: 5
```

When deptartment A is removed the p-value for department B changes dirastically. This likely means admittance rates for A and B are dependant.

(e) The approach in part (d) allowed us to keep the binomial likelihood model, but only by performing an unprincipled exclusion of some apparently-legitimate data that happened to be "outlying".

To avoid this, we'll refit the model for all departments with the quasibinomial family.

Using the quasibinomial model for all departments, what do you conclude about the effect of Gender on admissions? Support your conclusion by constructing and interpreting a 95% confidence interval for

$P_{diff} = [P(Admit|(Department, Male)) - P(Admit|(Department, Female))]$

That is, construct an interval on the model scale, then backtransform to the data scale. Be careful with the direction (male higher or female higher) of the observed difference in conditional probability of admission.

Gender does not have a statistically significant effect on admission (p-value 0.22). as seen by the summary below.

```
test1 <- glm(data = UCBA_case, Admit ~ .   , family = quasibinomial)

summary(test1)
```

```
##
## Call:
## glm(formula = Admit ~ ., family = quasibinomial, data = UCBA_case)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3613  -0.9588   0.3741   0.9306   1.4773
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.68192    0.09918  -6.875 7.03e-12 ***
## GenderMale   0.09987    0.08090   1.234    0.217
## DeptB        0.04340    0.10992   0.395    0.693
## DeptC        1.26260    0.10671  11.832  < 2e-16 ***
## DeptD        1.29461    0.10590  12.225  < 2e-16 ***
## DeptE        1.73931    0.12620  13.782  < 2e-16 ***
## DeptF        3.30648    0.17010  19.438  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.001447)
##
##     Null deviance: 6044.3  on 4525  degrees of freedom
## Residual deviance: 5187.5  on 4519  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```