OLR_guestlecture

Sunny

March 5, 2019

Research Question

What's the reason(s) that influence the probability of a student applying to a graduate school?

H1: The higher the GPA, the more likely that a student apply to graduate school
H2: If one of the parents has a graduate degree, more likely the student want to apply to graduate school

Take a look at the data, we want to know how the possibility of apply to graduate school influenced by gpa and pared. We have 400 observations.

- apply (ordinal)unlikely, somewhatlikely, unlikely
- pared (binary) whether at least one parent has a graduate degree
- gpa (continuous) grade point average

```
score <- read.dta("https://stats.idre.ucla.edu/stat/data/ologit.dta") %>%
  as_data_frame() %>%
  dplyr::select(apply, pared, gpa) %>%
  mutate(apply=factor(apply, ordered=T),
         pared=factor(pared))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
score
```

```
## # A tibble: 400 x 3
##    apply           pared   gpa
##    <ord>           <fct> <dbl>
##  1 very likely     0      3.26
##  2 somewhat likely 1      3.21
##  3 unlikely        1      3.94
##  4 somewhat likely 0      2.81
##  5 somewhat likely 0      2.53
##  6 unlikely        0      2.59
##  7 somewhat likely 0      2.56
##  8 somewhat likely 0      2.73
##  9 unlikely        0      3
## 10 somewhat likely 1      3.5
## # ... with 390 more rows
```

Data visualization/exploration

Here we want to have a general idea about how the data look like, so that we can have a sense that whether our hypothesis make sense or not. But remember, this exploratory analysis can not provide any "statistical conclusion", it is simply describing the data. We will need to fit models and do the tests before writing any conclusion.

- Relationship between apply and pared? In the same application category, "very likely" has the highest percentage of having a parent who has graduate degree

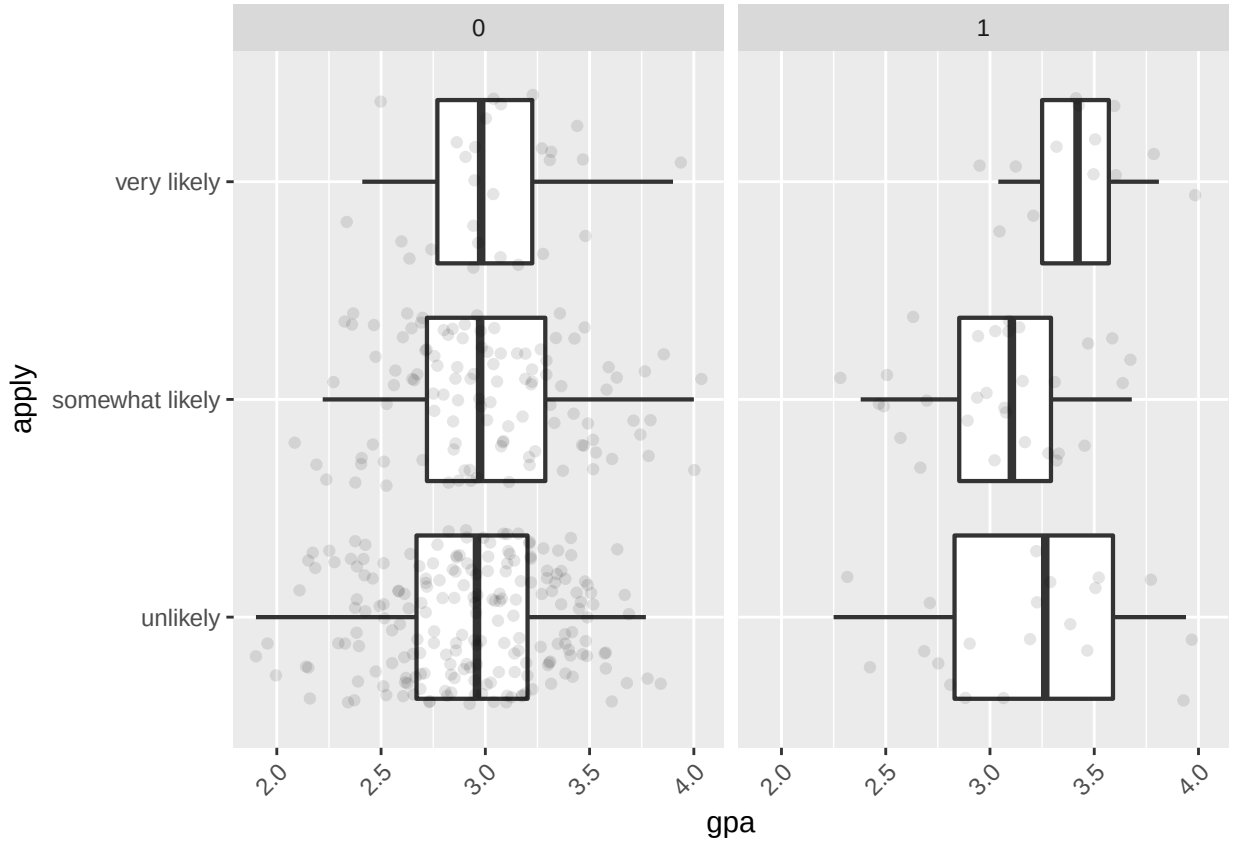
```
table <- ftable(xtabs(~apply+pared, data=score)) %>%
  as.data.frame.matrix()
rownames(table) <- c("unlikely", "somewhat likely", "very likely")
colnames(table) <- c("pared=0", "pared=1")
table$percent <- table[,2]/(table[,2]+table[,1])
table
```

```
##                 pared=0 pared=1    percent
## unlikely            200      20 0.09090909
## somewhat likely     110      30 0.21428571
## very likely          27      13 0.32500000
```

- Relationship between apply and gpa? -> wow, higher gpa seems to assoiated with higher probability in applying

```
value <- score %>%
  group_by(apply) %>%
  summarize(mean(gpa))
value
```

```
## # A tibble: 3 x 2
##   apply           `mean(gpa)`
##   <ord>                 <dbl>
## 1 unlikely               2.95
## 2 somewhat likely        3.03
## 3 very likely            3.15
```



Model formulation

Option 1: OLS regression (violation of assumption that the response needs to be interval outcome)

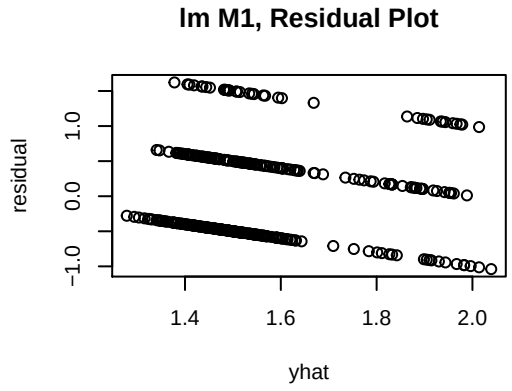
```
lm.M1 <- score %>%
  mutate(apply.c=as.numeric(score$apply)) %>%
  lm(apply.c~pared+gpa, data=.)
```

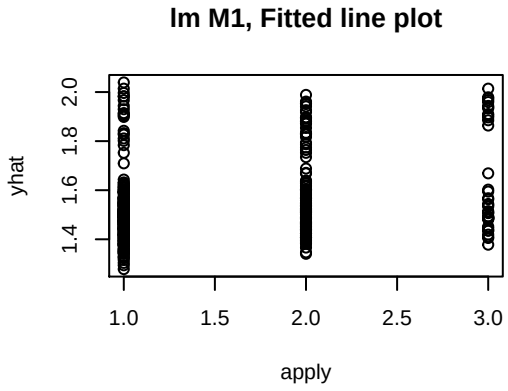
Let's take a look at the goodness of fit (ANOVA table).

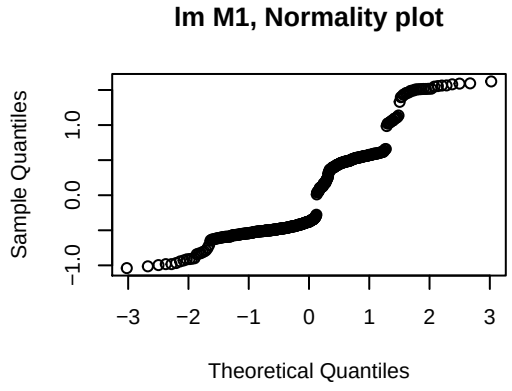
```
anova(lm.M1) # get the Anova table for the model, including the F test
```

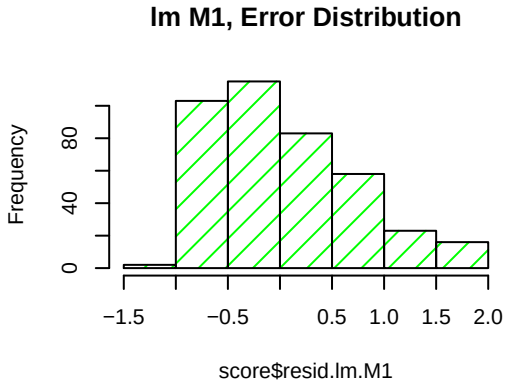
```
## Analysis of Variance Table
##
## Response: apply.c
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## pared       1   8.588  8.5879 20.2831 8.793e-06 ***
## gpa         1   2.322  2.3225  5.4853   0.01967 *
## Residuals 397 168.090  0.4234
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the ANOVA table, our predictors (pared and gpa) are all significant in the t-test. Also, the model F-test is significant. Those mean our linear regression is very nice? No, because "the tests are only reliable when the assumptions of linear regression were met". So, before being too happy about the results, let's take a look at whether the assumptions were met or not (see below).





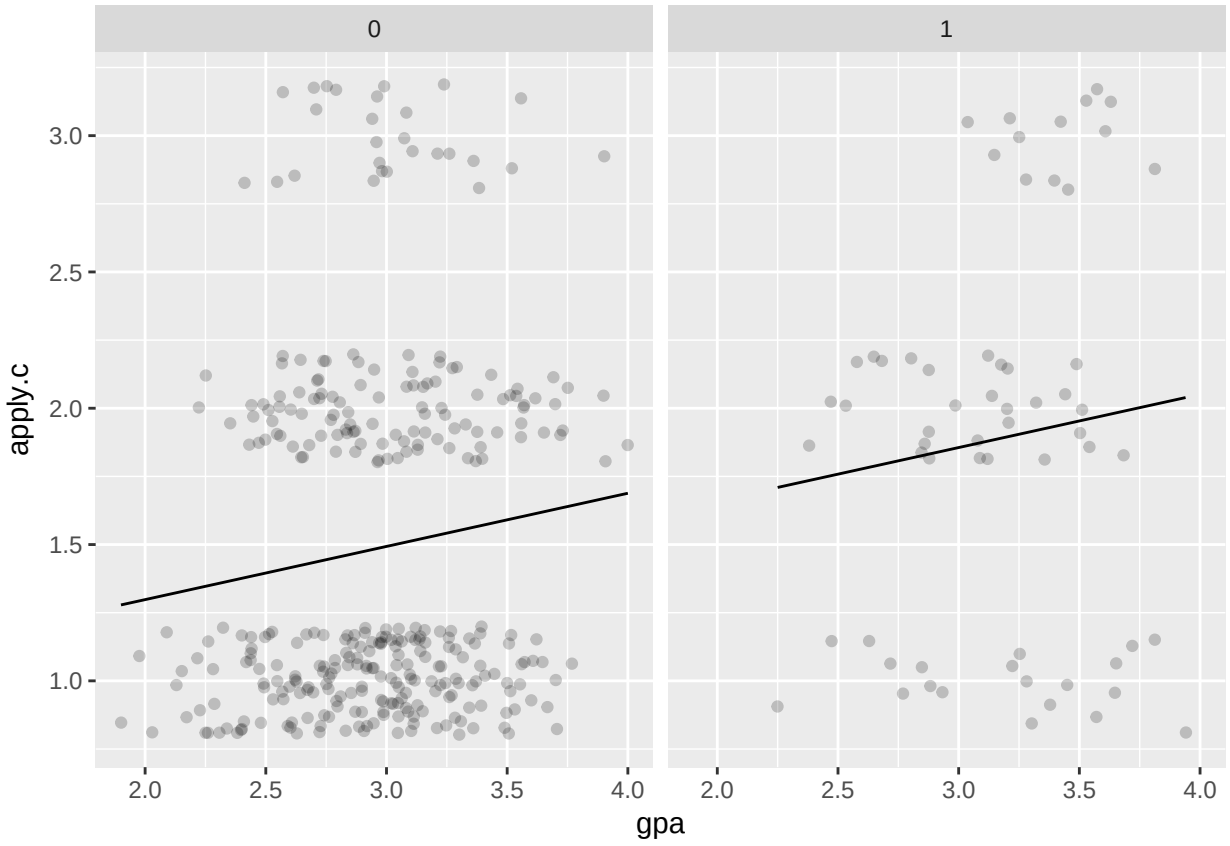




- Normality plot: Normality assumption is not met as the points are not in a straight line. The estimated coefficients are biased.
- Residual plot: Equal variance assumption is not met as there is a shape in the residual plot. The estimated variance of the coefficients are biased.

Combined together, all the coefficient t-tests, and model F-test are biased (not correct). (Oops!) But we can still take a look how the fitted line looks like in our data.

```
lm.M1.plot <- score %>%
  ggplot(aes(y=apply.c, x=gpa))+
    geom_jitter(alpha=.2, height=.2) +
    geom_line(aes(x=gpa, y=yhat.lm.M1))+
    facet_grid(.~score$pared)
lm.M1.plot
```



Option 2: Classifier (loss the information contained in the ordering)

```
#Give it a try! :D
```

Option 3: Ordinal logistic regression (Sounds great!)

```
olr.M2 <- polr(apply~pared+gpa, data=score) #Reduced model, without interaction.
olr.M5 <- olr.M2 #My final selection according to the above observations.

score <- score %>%
  cbind(., olr.M5.prob.=predict(olr.M5, ., type="probs")) %>% #estimated probability
```

```
  cbind(., olr.M5.decision=predict(olr.M5))                    #estimated result according to the probab
```

Goodness of fit: usually look for AIC, likelihood, and Residual Deviance.

```
gf <- table(score$apply, score$olr.M5.decision)
accuracy <- (gf[1]+gf[5]+gf[9])/400

print(paste0("AIC = ", AIC(olr.M5), ".")) # The smaller the better
print(paste0("Log likelihood = ", logLik(olr.M5), ".")) # The larger the better
print(paste0("Overall accuracy = ", accuracy, ".")) # The larger the better
```

```
## [1] "AIC = 725.063788308407."
```

```
## [1] "Log likelihood = -358.531894154204."
```

```
## [1] "Overall accuracy = 0.5775."
```

The values look nice. There are actually more more more statistis that we can use to access the goodness of fit for the OLR model. Here we demonstrated the most basic three. :) Now if we think the model is good enough, we can interpret the coefficient.

```
coef(summary(olr.M5)) # coefficients
```

```
##
## Re-fitting to get Hessian
```

```
##                                 Value Std. Error  t value
## pared1                      1.0457078  0.2656427 3.936520
## gpa                         0.6042468  0.2539454 2.379436
## unlikely|somewhat likely    2.1762687  0.7670896 2.837046
## somewhat likely|very likely 4.2715846  0.7921502 5.392392
```

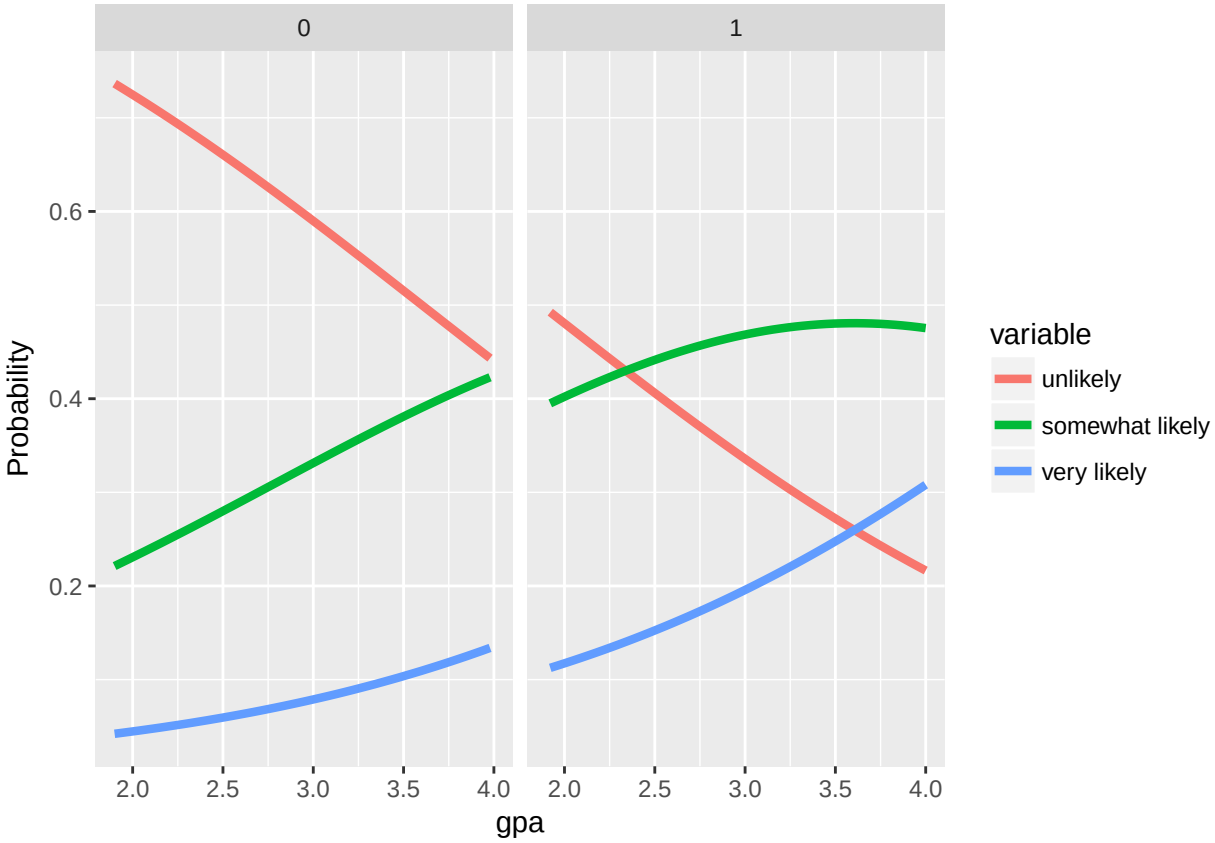
- Example of interpretation (slope): For gpa, one unit increase in gpa, there is a 0.6 increase in the expected value of apply (in log odds scale), given that all of the other variables in the model are constant. A positive slope indicates a tendency for tahe response level to increase as the predictor increases.
- Example of interpretation (intercept): 2.17 is the expected log odds of being "unlikely" versus "somewhat likely" and "very likeli" combined when all the predictors are 0.

```
exp(coef(olr.M5))     # proportional odds ratios
```

```
##   pared1      gpa
## 2.845412 1.829873
```

- Example of interpretation (proportional odds ratio): For one unit increase in gpa, the odds of moving from a lower categories to a high category is multiplied by 1.82.

Let's take a look how the fitted line looks like in our data.


0
1
Probability
0.6
0.4
0.2
2.0
2.5
3.0
3.5
4.0
gpa
variable
unlikely
somewhat likely
very likely