Final

Janette Avelar

2022-06-06

Question 1

I have a problem. I want to maximize the number of students who think they will be getting an "A" (because those students give the best ratings), but I don't have enough time to do it. Please help me!

I think helping my students get an "A" (versus any other grade) is a function of 3 things: focusing my lectures, using clear and relevant examples, and being accessible to students outside of class (all continuous/numeric). Controlling for student GPA (also numeric), compute a logistic regression predicting the likelihood of getting an "A" (vs. "not an A") based on those 3 factors. Do those 3 teacher factors predict the likelihood of an "A" above and beyond GPA? What about each of them individually?

According to the logistic regression model, using focused lectures, clear and relevant examples, and being available predicts the likelihood of getting an "A" above and beyond GPA. However, these predictors individually are not as good at predicting the likelihood of getting an "A". Likelihood increases by 1.2 with focused lectures, 1.1 with clear and relevant examples, and 1.2 with availability while GPA increases likelihood by 2. Including GPA in our model significantly improves our predictions and is the best of the 4 predictors for determining the likelihood of getting an "A".

```
#data prep
ratings <- q1 %>%
  select(get_an_a, item16, item17, item20, gpa) %>%
  rename(focus = item16,
         examples = item17,
         availability = item20) %>%
  mutate(get_an_a = factor(get_an_a, levels = c(0, 1), labels = c("no", "yes")),
         gpa_c = gpa - mean(gpa)) %>%
  na.omit()
#run model
log_mod <- glm(get_an_a ~ focus + examples + availability, data = ratings, family = "binomial")
summary(log_mod)
```

```
##
## Call:
## glm(formula = get_an_a ~ focus + examples + availability, family = "binomial",
##     data = ratings)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9845  -0.8569  -0.7510   1.3835   2.0802
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.77233    0.39294  -7.055 1.72e-12 ***
## focus         0.15141    0.09404   1.610    0.107
## examples      0.11950    0.08998   1.328    0.184
## availability  0.18907    0.07344   2.575    0.010 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1674.9  on 1388  degrees of freedom
## Residual deviance: 1648.1  on 1385  degrees of freedom
## AIC: 1656.1
##
## Number of Fisher Scoring iterations: 4
```

```
#convert to odds
coef(log_mod) %>%
  exp()
```

```
##  (Intercept)        focus     examples availability
##   0.06251601   1.16347400   1.12693650   1.20812009
```

```
#run model controlling gpa
log_mod2 <- glm(get_an_a ~ focus + examples + availability + gpa_c, data = ratings, family = "binomial")
summary(log_mod2)
```

```
##
## Call:
## glm(formula = get_an_a ~ focus + examples + availability + gpa_c,
##     family = "binomial", data = ratings)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3560  -0.8487  -0.5865   1.1109   2.4024
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.91271    0.41694  -6.986 2.83e-12 ***
## focus         0.20878    0.09772   2.136   0.0326 *
## examples      0.09090    0.09383   0.969   0.3327
## availability  0.16109    0.07679   2.098   0.0359 *
## gpa_c         0.69452    0.06381  10.884  < 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: 1674.9  on 1388  degrees of freedom
## Residual deviance: 1509.8  on 1384  degrees of freedom
## AIC: 1519.8
```

```
##
## Number of Fisher Scoring iterations: 4
```

```
#convert to odds
coef(log_mod2) %>%
  exp()
```

```
##  (Intercept)        focus     examples availability        gpa_c
##   0.05432835   1.23217220   1.09515891   1.17478967   2.00274301
```

```
#compare the models
anova(log_mod, log_mod2, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: get_an_a ~ focus + examples + availability
## Model 2: get_an_a ~ focus + examples + availability + gpa_c
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      1385     1648.1
## 2      1384     1509.8  1    138.3 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on your findings from Part A, re-run the model with only the one best predictor of "get an A." Generate a plot with the probability of getting an A (on the y-axis) as a function of that one best predictor (on the x-axis).

```
#gpa alone model
log_mod3 <- glm(get_an_a ~ gpa_c, data = ratings, family = "binomial")
summary(log_mod3)
```

```
##
## Call:
## glm(formula = get_an_a ~ gpa_c, family = "binomial", data = ratings)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1826  -0.9035  -0.6692   1.1722   2.3821
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.01122    0.06578  -15.37   <2e-16 ***
## gpa_c        0.69730    0.06342   10.99   <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: 1674.9  on 1388  degrees of freedom
## Residual deviance: 1533.1  on 1387  degrees of freedom
## AIC: 1537.1
##
## Number of Fisher Scoring iterations: 4
```

```r
#convert to odds to see what's going on
coef(log_mod3) %>%
  exp()
```

```
## (Intercept)       gpa_c
##   0.3637764   2.0083142
```

```r
#calculate a threshold for our plot and predictions
thresh <- -log_mod3$coef[[1]] / log_mod3$coef[[2]]
thresh
```

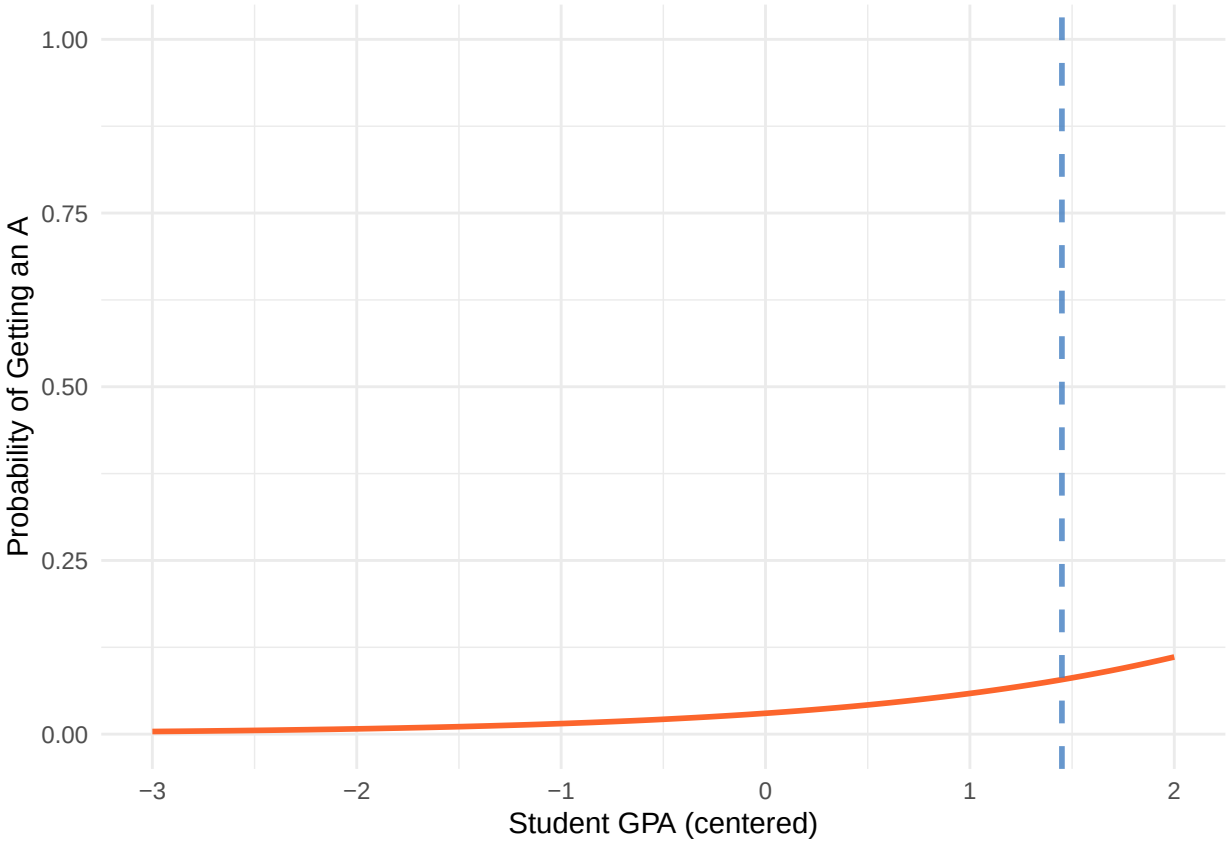
```
## [1] 1.450197
```

```r
#add preds
ratings$gpa_pred <- predict(log_mod3, type = "response")
#plot
ggplot(ratings, aes(x = gpa, y = gpa_pred)) +
  stat_smooth(method      = "glm",
              method.args = list(family = "binomial"),
              fullrange   = TRUE,
              se          = FALSE,
              color       = "#fc652c") +
  geom_vline(xintercept = thresh,
             linetype   = 2,
             color      = "#417dc1",
             size       = 1,
             alpha      = .8) +
  ylim(c(0, 1)) + #preds min .06 max .5
  xlim(c(-3, 2)) + #gpa min -2.53 max 1.47
  theme_minimal() +
  labs(x     = "Student GPA (centered)",
       y     = "Probability of Getting an A")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 1106 rows containing non-finite values (stat_smooth).
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```



Based on your findings from Part A, tell me how much the probability of getting an "A" will increase for a student with an average GPA if I improve by one unit in each of the 3 domains (give me separate probability changes in each domain, controlling for all other domains).

For a student with average GPA, when holding all domains constant we end up with .35 odds of getting an "A". This improves by .09 with a unit increase in focus, by .04 with a unit increase in clear and relevant examples, and by .07 with a unit increase in availability.

How much would an average-GPA student improve if I put 2 units of effort into focusing my lectures, but no effort into the other 2 domains?

If we add 2 units of effort into focusing lectures, the odds improve by .19.

What is your recommendation?

Given these values, my recommendation is to put more effort into focusing lectures rather than focusing on all 3 domains.

```
#we need to center the other predictors to begin with "average" scores in each domain
#we'll also add
ratings <- ratings %>%
  mutate(focus_c = focus - mean(focus, na.rm = TRUE),
         examples_c = examples - mean(examples, na.rm = TRUE),
         availability_c = availability - mean(availability, na.rm = TRUE))
#run model with all centered predictors to create predictive model
log_mod_c <- glm(get_an_a ~ focus_c + examples_c + availability_c + gpa_c, data = ratings, family = "bi
summary(log_mod_c)
```

```
##
```

```
## Call:
## glm(formula = get_an_a ~ focus_c + examples_c + availability_c +
##     gpa_c, family = "binomial", data = ratings)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3560  -0.8487  -0.5865   1.1109   2.4024
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.03230    0.06689 -15.433   <2e-16 ***
## focus_c         0.20878    0.09772   2.136   0.0326 *
## examples_c      0.09090    0.09383   0.969   0.3327
## availability_c  0.16109    0.07679   2.098   0.0359 *
## gpa_c           0.69452    0.06381  10.884   <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: 1674.9  on 1388  degrees of freedom
## Residual deviance: 1509.8  on 1384  degrees of freedom
## AIC: 1519.8
##
## Number of Fisher Scoring iterations: 4
```

```
#get log odds
coef(log_mod_c) %>%
  exp()
```

```
##    (Intercept)        focus_c     examples_c availability_c          gpa_c
##       0.356187       1.232172       1.095159       1.174790       2.002743
```

```
#get coefficients to build formula
coefs <- coef(log_mod_c)
#save coefficients to build formulas
b0 <- coefs[1]
b1 <- coefs[2] #avg focus odds
b2 <- coefs[3] #avg examples odds
b3 <- coefs[4] #avg availability odds
b4 <- coefs[5] #avg gpa odds
#model with no increases; avg everything
exp(b0 + b1*0 + b2*0 + b3*0 + b4*0) #odds = .35
```

```
## (Intercept)
##    0.356187
```

```
#increasing by 1 unit of focus
exp(b0 + b1*1 + b2*0 + b3*0 + b4*0) #odds = .44
```

```
## (Intercept)
##   0.4388838
```

```
#increasing by 1 unit of examples
exp(b0 + b1*0 + b2*1 + b3*0 + b4*0) #odds = .39
```

```
## (Intercept)
##   0.3900814
```

```
#increasing by 1 unit of availability
exp(b0 + b1*0 + b2*0 + b3*1 + b4*0) #odds = .42
```

```
## (Intercept)
##   0.4184449
```

```
#increasing by 2 units of focus
exp(b0 + b1*2 + b2*0 + b3*0 + b4*0) #odds = .54
```

```
## (Intercept)
##   0.5407804
```

Question 3

I would like to understand the factors that contribute to student grades in 613, so I gathered data for the past 3 years. There have been 3 different instructors in that time, so my data are structured as students nested within instructors. I don't care about the specific instructors in my sample; I am interested to generalize to the population of all instructors, not in these 3 instructors in particular. I hypothesize that a student's grade in 613 is a function of their overall GPA and the number of hours they spend on homework per week. Further, across academic quarters, I think that average grades can differ randomly between instructors, that the effect of overall GPA will be moderated by the instructor's years of experience, and that the effect of hours per homework will be moderated by the instructor's average past ratings.

Please write down the L1 and L2 equations to test my hypothesis and also the single-equation form of the model [don't actually run the model!]. Be sure to indicate what each predictor is, what the meaning of its parameter is (i.e., the "b" or "g" attached to each predictor), and what the variance of each error term is and what that variance represents.

Level 1 Model:

$$Y_{ij} = b_{0j} + b_{1j}GPA_{ij} + b_{2j}hours_{ij} + e_{ij}$$

where:
* b_{0j} = predicted student grade holding all else constant (intercept)
* b_{1j} = predicted change in student grade with each increase in GPA
* b_{2j} = predicted change in student grade with each increase in hours spent doing homework
* e_{ij} = residual variance is error at the individual level

Level 2 Model:

$$b_{0j} = g_{00} + g_{01}instructor_j + g_{02}years_j + g_{03}avg_j + u_{0j}$$

where:
* g_{00} = overall intercept and expected grade when hours on homework and GPA are 0 (or average, assuming we center)
* g_{01} = predicted change in intercept based on instructor assuming grades differ (main effect)
* g_{02} = predicted change in intercept based on number of years of experience

* g_{03} = predicted change in intercept based on the average past ratings * u_{0j} = variance is random error for intercepts

$$b_{1j} = g_{10} + g_{11}instructor_j + g_{12}years_j + g_{13}avg_j + u_{1j}$$

where: * g_{10} = overall slope and expected increase as a function of GPA and hours spent on homework
* g_{11} = predicted change in slope based on effect of instructor
* g_{12} = predicted change in slope based on effect of years teaching on GPA (interaction)
* g_{13} = predicted change i slope based on effect of average past ratings on hours spent on homework (interaction)
* u_{ij} = variance is random error for the slope

Single-equation form:

$Y_{ij} = g_{00}+g_{01}instructor_j+g_{02}years_j+g_{03}avg_j+g_{10}+g_{11}instructor_j+g_{12}years_j*GPA_{ij}+g_{13}avg_j*hours_{ij}+(e_{ij}+u_{0j}+u_{ij})$

Question 4

In a study on the efficacy of the Drug Abuse Resistance Education (DARE) program, researchers used a MLM to examine students nested within schools that did or did not receive the DARE intervention. Their final model for marijuana use was:

$$Marijuana_use_{ij} = b_{0j} + b_{1j} * PreDARE + e_{ij}$$

where:

$$b_{0j} = 0.033 - 0.044 * DARE_j + u_{0j}$$
$$b_{1j} = 0.098 + u_{1j}$$

In the equations, the DV is marijuana use, and the predictors are pre-DARE marijuana use and the DARE intervention (DARE = 0 if the school did not receive the intervention, and DARE = 1 if it did). The only parameters that were significantly different than zero were 0.098, var(u_0j), and var(e_ij). Label each of the parameters with the appropriate "g", explain the meaning of each, and interpret the results as you would in a real paper.

Parameters:
* g_{00} = 0.033, which is the overall intercept and expected marijuana use when pre-DARE use was 0 in control group
* g_{01} = -0.044, which is the expected change in intercept for the group that received the intervention
* g_{10} = 0.098, which is the overall slope and expected increase in use as a function of pre-DARE use

To study the efficacy of the DARE program, a multilevel model (MLM) was conducted to examine the effects of the program on groups of students. These students came from different schools, and schools that did not implement the DARE program were used as a control to test its effects. The results from the model indicated that decreased expected marijuana use scores for students who received the intervention were not significantly different from 0. However, the model found that the expected increase in use as a function of pre-intervention marijuana use was significantly different from 0, indicating that for each additional expected unit prior to the intervention, marijuana use would increase by 0.098. The variance in random error for the intercepts and at the individual level were also significantly different from 0.

These findings indicate that individual use prior to the intervention was a stronger predictor for post-intervention use than the intervention itself. That is, students who were more likely to be using marijuana prior were more likely to continue or increase their use (albeit to a small degree). Additionally, the significant variance in error for the intercepts and at the individual level seem to indicate that our model did not capture the variance well. There is an unknown predictor we did not account for creating a large amount of variance among individuals that exists outside of the control and intervention groups which are affecting the outcomes. Thus, an MLM model may not have been the most appropriate model to use with this data, or alternately, additional predictors should be identified in future analyses.