Week 7 Example Solution

Katz

3/2/2021

This week we will be using a dataset on faculty burnout (to complement the prior examples of student stress and burnout, recognizing that we’re all in this together…). This is from the textbook, Smart Alex’s Task 2. I am copying those instructions here:

“Recent research has shown that lecturers are among the most stressed workers. A researcher wanted to know exactly what it was about being a lecturer that created this stress and subsequent burnout. She took 467 lecturers and administered several questionnaires to them that measured: Burnout (burnt out or not), Perceived Control (high score = low perceived control), Coping Style (high score = high ability to cope with stress), Stress from Teaching (high score = teaching creates a lot of stress for the person), Stress from Research (high score = research creates a lot of stress for the person) and Stress from Providing Pastoral Care (high score = providing pastoral care creates a lot of stress for the person). The outcome of interest was burnout, and Cooper, Sloan, and Williams’s (1988) model of stress indicates that perceived control and coping style are important predictors of this variable. The remaining predictors were measured to see the unique contribution of different aspects of a lecturer’s work to their burnout. Can you help her out by conducting a logistic regression to see which factors predict burnout? The data are in Burnout.dat.”

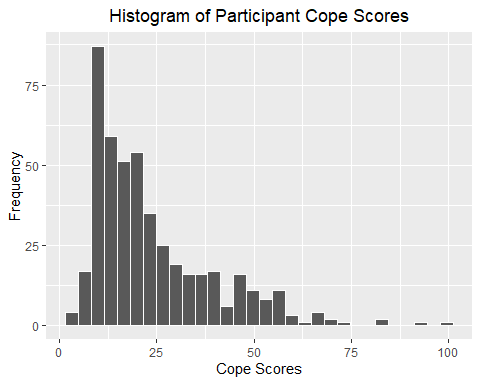
As usual, start by reading in the data.

burnout\_df <- read.delim("Burnout.dat")

Explore the data you have by using head(), with summary statistics and describe(), and exploratory plots.

burnout\_df %>%   
 ggplot(aes(x = cope)) +  
 geom\_histogram(color="white") +  
 labs(x = "Cope Scores",  
 y = "Frequency",  
 title = "Histogram of Participant Cope Scores") +  
 theme(plot.title = element\_text(hjust=0.5)) #this last line centers your title

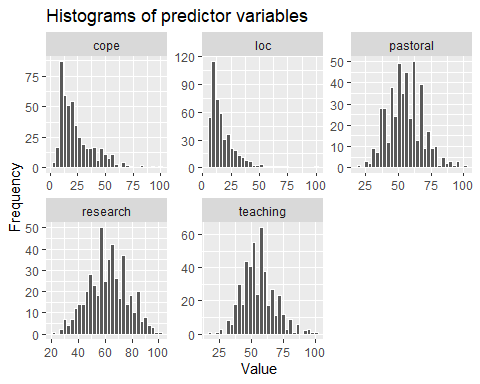
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



If you want a challenge, see if you can use pivot\_longer() to plot all variables at once (hint: this might be difficult with the binary outcome variable burnout)

burnout\_df %>% pivot\_longer(cols = loc:pastoral,   
 names\_to = "variable",   
 values\_to = "value") %>%   
 ggplot(aes(x = value)) +  
 geom\_histogram(color = "white") +  
 facet\_wrap(.~variable,   
 scales = "free") +  
 labs(x = "Value",  
 y = "Frequency",  
 title = "Histograms of predictor variables")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

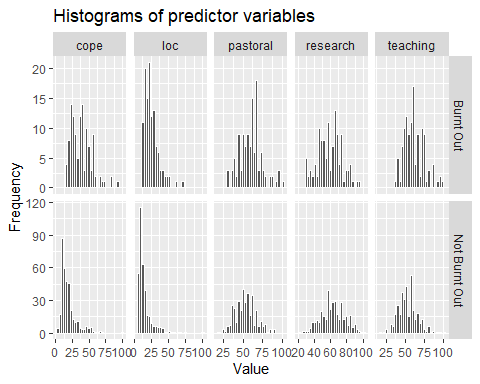


Make a few observations about the distribution of your predictor variables

It can also be a good idea to look at the distributions of your predictors across each level of the outcome variable

burnout\_df %>% pivot\_longer(cols = loc:pastoral,   
 names\_to = "variable",   
 values\_to = "value") %>%   
 ggplot(aes(x = value)) +  
 geom\_histogram(color = "white") +  
 facet\_grid(burnout~variable,   
 scales = "free") +  
 labs(x = "Value",  
 y = "Frequency",  
 title = "Histograms of predictor variables")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



Now you can start thinking about modeling the outcome…

Start with a simple model with two predictors you think might be most likely to be associated with the outcome.

But first, you will need to recode the outcome variable. You can do that in a couple of ways. If you use as.factor() then you may have less control over which is the “no” and which is “yes”, but you can try it and see…

burnout\_df <- burnout\_df %>% mutate(burnout\_factor = as.factor(burnout))

You can also recode it more explicitly as 0 and 1.

burnout\_df <- burnout\_df %>%   
 mutate(burnout\_bin = case\_when(burnout == "Burnt Out" ~ 1,  
 burnout == "Not Burnt Out" ~ 0))

Now that you have recoded the outcome variable, you can make your simple model. Note that I have written the burnout\_bin variable here rather than the original burnout variable.

simple\_model <- glm(burnout\_bin ~ teaching + research,   
 data = burnout\_df,  
 family = binomial(link = "logit"))

Now, look at the results of your model…

summary(simple\_model)

##   
## Call:  
## glm(formula = burnout\_bin ~ teaching + research, family = binomial(link = "logit"),   
## data = burnout\_df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6983 -0.7577 -0.6106 0.8744 2.1595   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.156536 0.764795 -4.127 3.67e-05 \*\*\*  
## teaching 0.048055 0.008997 5.341 9.23e-08 \*\*\*  
## research -0.010894 0.007678 -1.419 0.156   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 530.11 on 466 degrees of freedom  
## Residual deviance: 492.58 on 464 degrees of freedom  
## AIC: 498.58  
##   
## Number of Fisher Scoring iterations: 4

What do the coefficient values tell you? Write a few sentences practicing your interpretation. Also, what does the deviance tell you?

* Your observations here \*

Now try this again with a more complex model.

complex\_model <- glm(burnout\_bin ~ teaching + research + loc + cope,   
 data = burnout\_df,  
 family = binomial(link = "logit"))

Again, look at the model results.

summary(complex\_model)

##   
## Call:  
## glm(formula = burnout\_bin ~ teaching + research + loc + cope,   
## family = binomial(link = "logit"), data = burnout\_df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.76620 -0.52344 -0.31676 0.06889 2.46270   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.95473 0.95834 -3.083 0.00205 \*\*   
## teaching -0.08373 0.01697 -4.935 8.01e-07 \*\*\*  
## research 0.01748 0.01003 1.743 0.08130 .   
## loc 0.09563 0.01367 6.995 2.66e-12 \*\*\*  
## cope 0.13636 0.01548 8.811 < 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: 530.11 on 466 degrees of freedom  
## Residual deviance: 333.71 on 462 degrees of freedom  
## AIC: 343.71  
##   
## Number of Fisher Scoring iterations: 5

And now repeat the exercise of writing down your observations about the coefficients, what they mean, and what the deviance means (especially when compared with your simpler model)

* Your observations here \*