Homework-3

Lindley Slipetz

7/11/2021

For this homework, I will be using the Childhood adversity and traumatic stress among inpatients at a psychiatric hospital in the Baltimore area from 1993-1995. The data include diagnoses, psychological symptoms, physical and sexual abuse, post-traumatic stress disorder, self-destructive behavior, and demographic data. I will be predicting suicidality (an ordered variable) from gender, race, self-harm, SES, mood disorder diagnosis, history of neglect, positive affect, and psychoticism.

I'm loading the data and packages.

```
#install.packages("brant")
#install.packages("patchwork")
require(brant) # for brant test
require(ggplot2)
require(MASS) # for polr() & mvrnorm()
require(patchwork) # for combining graphs
require(tidyverse)
full_data <- read.table(file = 'G:\\My Drive\\ICPSR\\ML\\HW_2\\36168-0001-Data.tsv', sep = '\t', header
```

Now, I'm going to turn race into a binary variable (it's currently white, black, and other. There are very few observations in the other category, so I'm turning it into a binary variable of white and other).

```
full_data <- full_data %>%
  mutate(race = case_when(
    RACE == 0 ~ 0,
    RACE == 1 ~ 1,
    RACE == 3 ~ 1
  ))
```

Here, I subset the data to only the variables I'm interested in.

```
subset_data <- full_data %>%
  select(SISDB_SUIC, SEX, race, SISDB_SHARM, SES, MOODDX, NEGLECT, PASUM, SCL_PSY )
```

Now I'm going to look at the amount of missing data and figure out what I'm going to do.

```
df <- as.data.frame(
  cbind(
    lapply(
      lapply(subset_data, is.na), sum)
    )
  )

rownames(subset(df, df$V1 != 0))
```

```
## [1] "SISDB_SUIC"  "race"        "SISDB_SHARM" "PASUM"       "SCL_PSY"
```

Okay. "SISDB_SUIC", "race", "SISDB_SHARM", "PASUM", and "SCL_PSY" all have missing data. Let's

see how much of problem it is.

```
sum(is.na(subset_data$SISDB_SUIC))
```

```
## [1] 10
```

```
sum(is.na(subset_data$race))
```

```
## [1] 3
```

```
sum(is.na(subset_data$SISDB_SHARM))
```

```
## [1] 10
```

```
sum(is.na(subset_data$PASUM))
```

```
## [1] 2
```

```
sum(is.na(subset_data$SCL_PSY))
```

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

That's not that much missing data (at least to me). I think we'd be safe to just omit the data with NA.

```
complete_data <- na.omit(subset_data)
```

Now let's try OLS with our data.

```
ols <- lm(complete_data$SISDB_SUIC ~ complete_data$SEX + complete_data$race + complete_data$SISDB_SHARM
summary(ols)
```

```
##
## Call:
## lm(formula = complete_data$SISDB_SUIC ~ complete_data$SEX + complete_data$race +
##     complete_data$SISDB_SHARM + complete_data$SES + complete_data$MOODDX +
##     complete_data$NEGLECT + complete_data$PASUM + complete_data$SCL_PSY)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.76261 -0.43551  0.00315  0.35912  2.26530
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.282564   0.325948   3.935 0.000116 ***
## complete_data$SEX         -0.257019   0.134724  -1.908 0.057909 .
## complete_data$race        -0.445105   0.165733  -2.686 0.007869 **
## complete_data$SISDB_SHARM  0.553183   0.055233  10.015  < 2e-16 ***
## complete_data$SES          0.001089   0.003729   0.292 0.770657
## complete_data$MOODDX       0.131624   0.123926   1.062 0.289510
## complete_data$NEGLECT     -0.024522   0.044329  -0.553 0.580774
## complete_data$PASUM        0.010498   0.036869   0.285 0.776144
## complete_data$SCL_PSY      0.115364   0.070728   1.631 0.104501
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8428 on 193 degrees of freedom
## Multiple R-squared:  0.4971, Adjusted R-squared:  0.4762
## F-statistic: 23.84 on 8 and 193 DF,  p-value: < 2.2e-16
```

SES is significant with positive coefficient meaning that lower SES is associated with higher suicidality (the

scale used for SES has higher scores meaning lower SES). The neglect scale score has a significant negative coefficient meaning that as childhood neglect increases, suicidality decreases. That's interesting. There is also a significant positive coefficient for positive affect. This means that those interviewed who reported more frequent happiness are more likely to have attempted suicide. Again, these are not relationships you'd expect to find. Let's see how the ordered model does.

```
out1 <- polr(as.ordered(complete_data$SISDB_SUIC) ~ complete_data$SEX + complete_data$race + complete_da
             data = complete_data, method = "logistic", Hess = TRUE)
summary(out1)
```

```
## Call:
## polr(formula = as.ordered(complete_data$SISDB_SUIC) ~ complete_data$SEX +
##     complete_data$race + complete_data$SISDB_SHARM + complete_data$SES +
##     complete_data$MOODDX + complete_data$NEGLECT + complete_data$PASUM +
##     complete_data$SCL_PSY, data = complete_data, Hess = TRUE,
##     method = "logistic")
##
## Coefficients:
##                               Value Std. Error t value
## complete_data$SEX         -0.596394   0.327629 -1.8203
## complete_data$race        -1.038138   0.387794 -2.6770
## complete_data$SISDB_SHARM  1.314108   0.164179  8.0041
## complete_data$SES         -0.002964   0.009278 -0.3195
## complete_data$MOODDX       0.300781   0.311839  0.9645
## complete_data$NEGLECT     -0.068077   0.111475 -0.6107
## complete_data$PASUM        0.042235   0.089563  0.4716
## complete_data$SCL_PSY      0.349701   0.179772  1.9452
##
## Intercepts:
##     Value   Std. Error t value
## 0|1 -1.0948  0.7951    -1.3769
## 1|2  0.5951  0.7930     0.7504
## 2|3  1.7514  0.8077     2.1684
##
## Residual Deviance: 376.0843
## AIC: 398.0843
```

Odds ratio

```
exp(coef(out1))
```

```
##         complete_data$SEX        complete_data$race complete_data$SISDB_SHARM
##                 0.5507942                 0.3541134                 3.7214292
##         complete_data$SES      complete_data$MOODDX     complete_data$NEGLECT
##                 0.9970401                 1.3509136                 0.9341888
##       complete_data$PASUM     complete_data$SCL_PSY
##                 1.0431399                 1.4186429
```

Let's interpret the coefficients that were significant for OLS. For SES, the odds of high suicidality vs low to mid suicidality are 0.997 times lower for those with high SES. For neglect, the odds of high suicidality vs low to mid suicidality are 0.934 times lower for those with high neglect scores. For positive affect, the odds of high suicidality vs low to mid suicidality are 1.043 times higher for those with high positive affect scores. Let's look at the graphs

```
set.seed(1234)
simbt <- mvrnorm(n = 1000, mu = c(out1$coefficients, out1$zeta), Sigma = vcov(out1))
simb <- simbt[, 1:8] # 1000 * 5 matrix of simulated coefficients
```

```r
simt <- simbt[, 7:10] # 1000 * 3 matrix of simulated cutpoints
```

```r
xbc <- (simb[, 1] * mean(complete_data$SEX)
        + simb[, 2] * mean(complete_data$race)
        + simb[, 3] * mean(complete_data$SISDB_SHARM)
        + simb[, 4] * mean(complete_data$SES)
        + simb[, 5] * 1
        + simb[, 6] * mean(complete_data$NEGLECT)
        + simb[, 7] * mean(complete_data$PASUM)
        + simb[, 8] * mean(complete_data$SCL_PSY))
xbn <- (simb[, 1] * mean(complete_data$SEX)
        + simb[, 2] * mean(complete_data$race)
        + simb[, 3] * mean(complete_data$SISDB_SHARM)
        + simb[, 4] * mean(complete_data$SES)
        + simb[, 5] * 0
        + simb[, 6] * mean(complete_data$NEGLECT)
        + simb[, 7] * mean(complete_data$PASUM)
        + simb[, 8] * mean(complete_data$SCL_PSY))
```

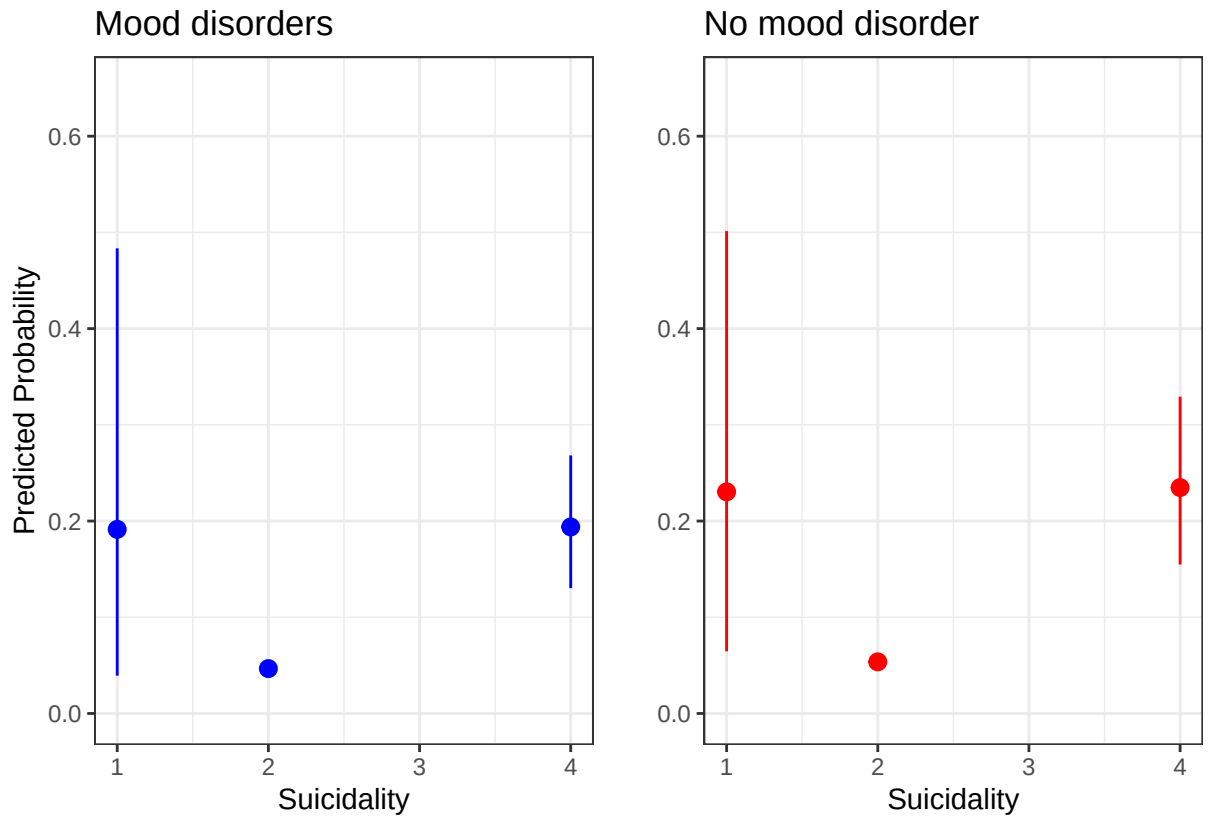
```r
res_pr_mood <- matrix(NA, nrow = 8, ncol = 3) # matrix to store results
rownames(res_pr_mood) <- paste0(rep(c("mood", "non-mood"), each = 4), "-",
                                rep(c(1:4), times = 2))
colnames(res_pr_mood) <- c("Mean", "Lower", "Upper")
cut <- cbind(-Inf, simt, Inf) # 1000 * 5 matrix of simulated cutpoints
for (j in 1:4){ # for each value of the dependent variable...
  # mood = 1
  pr_c <- plogis(cut[, j + 1] - xbc) - plogis(cut[, j] - xbc) # vector simulated predicted probs
  res_pr_mood[j, 1] <- mean(pr_c) # simulated mean
  res_pr_mood[j, 2:3] <- quantile(pr_c, probs = c(0.025, 0.975)) # simulated 95% CI
  # mood = 0
  pr_n <- plogis(cut[, j + 1] - xbn) - plogis(cut[, j] - xbn)
  res_pr_mood[j + 4, 1] <- mean(pr_n)
  res_pr_mood[j + 4, 2:3] <- quantile(pr_n, probs = c(0.025, 0.975))
}
res_pr_mood # display results
```

```
##                   Mean        Lower       Upper
## mood-1       0.19136797  0.039295777 0.483464594
## mood-2       0.04664625 -0.007648657 0.126288497
## mood-3      -0.17727936 -0.493840053 0.002521392
## mood-4       0.19388274  0.130282166 0.268205822
## non-mood-1   0.23037518  0.064578109 0.501395664
## non-mood-2   0.05365551 -0.008999478 0.133516895
## non-mood-3 -0.20423578 -0.511016156 0.004126662
## non-mood-4   0.23479213  0.154806881 0.329234067
```

```r
res_pr_mood <- as.data.frame(res_pr_mood)
res_pr_mood$DV <- rep(c(1:4), times = 2)
g1 <- ggplot(res_pr_mood[1:4,]) +
  geom_pointrange(aes(x = DV, y = Mean, ymin = Lower, ymax = Upper), color = "blue") +
  xlab("Suicidality") + ylab("Predicted Probability") + ggtitle("Mood disorders") +
  ylim(0.0, 0.65) + theme_bw()
g2 <- ggplot(res_pr_mood[5:8,]) +
  geom_pointrange(aes(x = DV, y = Mean, ymin = Lower, ymax = Upper), color = "red") +
```

```
  xlab("Suicidality") + ylab("") + ggtitle("No mood disorder") +
  ylim(0.0, 0.65) + theme_bw()
g1 | g2 # combine plots
```

```
## Warning: Removed 1 rows containing missing values (geom_pointrange).
## Warning: Removed 1 rows containing missing values (geom_segment).
## Warning: Removed 1 rows containing missing values (geom_pointrange).
## Warning: Removed 1 rows containing missing values (geom_segment).
```



There doesn't seem to be a difference between those with mood disorders and those without (which is exactly what OLS told us). Let's look at the results as a table.

```
res <-matrix(NA, nrow = 9, ncol = 3)
res[1,] <- c("Variables", "OLS", "OLM")
res[2:9,1] <- colnames(complete_data[,-1])
for(a in 2:9){
  res[a,2] <- round(ols$coefficients[a], 3)
}
for(b in 2:9){
  res[b,3] <- round(out1$coefficients[b-1], 3)
}
res
```

```
##       [,1]          [,2]     [,3]
##  [1,] "Variables"   "OLS"    "OLM"
##  [2,] "SEX"         "-0.257" "-0.596"
```

```
##  [3,] "race"        "-0.445" "-1.038"
##  [4,] "SISDB_SHARM" "0.553"  "1.314"
##  [5,] "SES"         "0.001"  "-0.003"
##  [6,] "MOODDX"      "0.132"  "0.301"
##  [7,] "NEGLECT"     "-0.025" "-0.068"
##  [8,] "PASUM"       "0.01"   "0.042"
##  [9,] "SCL_PSY"     "0.115"  "0.35"
```

From the table of coefficients, we cannot really tell much. OLM coefficients cannot be directly interpretted, even their signs. This is because the sign may differ across levels of the variable. Let's see if the cut points are significant.

Now I am going to calculate the change in predicted probability for each value of suicidality due to changing race for an otherwise average respondent.

```
Xb0 <- (coef(out1)[1] * mean(complete_data$SEX)
        + coef(out1)[2] * 0
        + coef(out1)[3] * mean(complete_data$SISDB_SHARM)
        + coef(out1)[4] * mean(complete_data$SES)
        + coef(out1)[5] * mean(complete_data$MOODDX)
        + coef(out1)[6] * mean(complete_data$NEGLECT)
        + coef(out1)[7] * mean(complete_data$PASUM)
        + coef(out1)[8] * mean(complete_data$SCL_PSY))
Xb1 <- (coef(out1)[1] * mean(complete_data$SEX)
        + coef(out1)[2] * 1
        + coef(out1)[3] * mean(complete_data$SISDB_SHARM)
        + coef(out1)[4] * mean(complete_data$SES)
        + coef(out1)[5] * mean(complete_data$MOODDX)
        + coef(out1)[6] * mean(complete_data$NEGLECT)
        + coef(out1)[7] * mean(complete_data$PASUM)
        + coef(out1)[8] * mean(complete_data$SCL_PSY))
```

```
## quality = 1
mean((plogis(out1$zeta[1] - Xb1) - plogis(-Inf - Xb1)) -
       (plogis(out1$zeta[1] - Xb0) - plogis(-Inf - Xb0)))
```

```
## [1] 0.08703575
```

```
## quality = 2
mean((plogis(out1$zeta[2] - Xb1) - plogis(out1$zeta[1] - Xb1)) -
       (plogis(out1$zeta[2] - Xb0) - plogis(out1$zeta[1] - Xb0)))
```

```
## [1] 0.1451166
```

```
## quality = 3
mean((plogis(out1$zeta[3] - Xb1) - plogis(out1$zeta[2] - Xb1)) -
       (plogis(out1$zeta[3] - Xb0) - plogis(out1$zeta[2] - Xb0)))
```

```
## [1] 0.00547618
```

```
## quality = 4
mean((plogis(Inf - Xb1) - plogis(out1$zeta[3] - Xb1)) -
       (plogis(Inf - Xb0) - plogis(out1$zeta[3] - Xb0)))
```

```
## [1] -0.2376286
```

The changes in predicted probability tell us that there is a positive change suicidality with a change in race for the average person for the the lowest two categories, but there is a negative change in suicidality with a change in race for the average person for high values of suicidality.

I did a graph of the change in predictive probabilities above.

Let's test if the cut points are significant.

```
#install.packages("aod")
library(aod)
wald.test(Sigma = vcov(out1), b = c(coef(out1), out1$zeta), Terms = 2:9)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 103.9, df = 8, P(> X2) = 0.0
```

The wald test is significant, meaning the cut points are statistically different from the null.