Homework_2

Lindley Slipetz

7/4/2021

For this homework assignment, I have decided to use the Psychological Classification of Adult Male Inmates in Federal Prison in Indiana, 1986-1988 dataset. I will be predicting suicidal tendencies from race, sense of self, SES, parent's disability status, optimism abut the future, and orientation towards relationships. Let's load the data!

```
library(haven)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
prison_all<-read_dta("C:\\Users\\Owner\\Documents\\ICPSR\\MLE\\HW_2\\02370-0001-Data.dta")
```

Now we're going to do some data cleaning. There are two variables for parents' disability status (mom and dad), so we're going to do some manipulation to make it one variable with three levels:

0 = neither parent is disabled 1 = one parent is disabled 2 = both parents are disabled.

```
prison_all <- prison_all %>%
  mutate(mom_dis = ifelse(V140 == 1 | V140 == 2, 1, ifelse(V140 == 9 , NA, 0)))
prison_all <- prison_all %>%
  mutate(dad_dis = ifelse(V155 == 1 | V155 == 2, 1, ifelse(V155 == 9 , NA, 0)))
prison_all <- prison_all %>%
  mutate(dis = case_when(
    mom_dis == 0 & dad_dis == 0 ~ 0,
    mom_dis == 1 & dad_dis == 0 ~ 1,
    mom_dis == 0 & dad_dis == 1 ~ 1,
    mom_dis == 1 & dad_dis == 1 ~ 2
  ))
```

Just to make things easier on me, I'm going to subset to only the variables of interest. Then I'm going to rename them more intuitive names.

```
prison <- prison_all %>%
  select("V196", "RACE", "V107", "SES", "V266", "dis", "V21")
oldnames = c("V196", "RACE", "V107", "SES", "V266", "dis", "V21")
newnames = c("sui", "race", "self", "ses", "opt", "dis", "rel")
prison <- prison %>% rename_at(vars(all_of(oldnames)), ~ newnames)
```

Now let's handle missing data. We'll tranform the numeric values into NA and then do an na.omit.

```
prison$sui[prison$sui == 9] <- NA
prison$race[prison$race == 9] <- NA
prison$self[prison$self == 9] <- NA
prison$ses[prison$ses == 9] <- NA
prison$opt[prison$opt == 9] <- NA
prison$rel[prison$rel == 9] <- NA
prison_na <- na.omit(prison)
```

I need to recode the sui variable to zeros and ones.

```
prison_na <- prison_na %>%
  mutate(sui_bi = case_when(
    sui == 1  ~ 0,
    sui == 2 ~ 1,

  ))
```

The data is now ready to go. Let's look at some frequency tables.

```
table(prison_na$sui_bi)
```

```
##
##   0   1
## 254  20
```

```
table(prison_na$race)
```

```
##
##   1   2   3   4   5
## 195  64   6   6   3
```

```
table(prison_na$self)
```

```
##
##   1   2   3   5
##   4  47 222   1
```

```
table(prison_na$ses)
```

```
##
##   1   2   3   4
##  42  85 114  33
```

```
table(prison_na$opt)
```

```
##
##   1   2   3
##   3  85 186
```

```
table(prison_na$rel)
```

```
##
##   1   2   3   4   5
##   5 154  52  45  18
```

```
table(prison_na$dis)
```

```
##
##   0   1   2
## 191  70  13
```

We'll start with the linear probability model

```
ols <- lm(prison_na$sui_bi ~ prison_na$race + prison_na$self + prison_na$ses + prison_na$opt + prison_na
summary(ols)
```

```
##
## Call:
## lm(formula = prison_na$sui_bi ~ prison_na$race + prison_na$self +
##     prison_na$ses + prison_na$opt + prison_na$rel + prison_na$dis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20277 -0.09454 -0.05877 -0.03188  0.99958
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.13371    0.13900   0.962   0.3369
## prison_na$race -0.01611    0.02169  -0.743   0.4583
## prison_na$self  0.03146    0.03525   0.892   0.3730
## prison_na$ses  -0.02689    0.01805  -1.490   0.1374
## prison_na$opt  -0.01649    0.03284  -0.502   0.6161
## prison_na$rel  -0.01249    0.01636  -0.763   0.4458
## prison_na$dis   0.05231    0.02772   1.887   0.0603 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2592 on 267 degrees of freedom
## Multiple R-squared:  0.03211,    Adjusted R-squared:  0.01036
## F-statistic: 1.476 on 6 and 267 DF,  p-value: 0.1864
```

Well…none of the coefficients are significant and the adjusted R^2 is 0.0321 (or about 3.21% of the variance in suicidality is explained by the linear predictors). So…that's not good. Let's just hope this is a case of the linear probability model struggling with a binary variable.

Logit model

```
out1 <- glm(prison_na$sui_bi ~ prison_na$race + prison_na$self + prison_na$ses + prison_na$opt + prison_
            data = prison_na, family = binomial, x = TRUE)
summary(out1)
```

```
##
## Call:
## glm(formula = prison_na$sui_bi ~ prison_na$race + prison_na$self +
##     prison_na$ses + prison_na$opt + prison_na$rel + prison_na$dis,
##     family = binomial, data = prison_na, x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7888  -0.4104  -0.3387  -0.2744   2.7319
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.4215     2.1177  -0.671    0.502
## prison_na$race  -0.3558     0.4339  -0.820    0.412
## prison_na$self   0.4064     0.5447   0.746    0.456
## prison_na$ses   -0.3980     0.2787  -1.428    0.153
## prison_na$opt   -0.2021     0.4682  -0.432    0.666
## prison_na$rel   -0.2356     0.2899  -0.813    0.416
## prison_na$dis    0.6215     0.3538   1.757    0.079 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 143.20  on 273  degrees of freedom
## Residual deviance: 134.86  on 267  degrees of freedom
## AIC: 148.86
##
## Number of Fisher Scoring iterations: 6
```

```
nullmod <- glm(prison_na$sui_bi ~ 1, family="binomial")
1-logLik(out1)/logLik(nullmod)
```

```
## 'log Lik.' 0.05826228 (df=7)
```

Again we see no significant coefficients and a pseudo-R^2 (McFadden's) of 0.0583. Even though I haven't done the probit yet, we can safely say that this set of predictors does not do a good job of explaining the variance in suicidality.

Probit

```
out2 <- glm(prison_na$sui_bi ~ prison_na$race + prison_na$self + prison_na$ses + prison_na$opt + prison_
            data = prison_na, family = binomial(link = 'probit'), x = TRUE)
summary(out2)
```

```
##
## Call:
## glm(formula = prison_na$sui_bi ~ prison_na$race + prison_na$self +
##     prison_na$ses + prison_na$opt + prison_na$rel + prison_na$dis,
##     family = binomial(link = "probit"), data = prison_na, x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7646  -0.4211  -0.3339  -0.2736   2.7266
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.8121     1.0514  -0.772   0.4399
## prison_na$race  -0.1688     0.2028  -0.832   0.4053
## prison_na$self   0.1538     0.2691   0.571   0.5677
## prison_na$ses   -0.1849     0.1368  -1.352   0.1764
## prison_na$opt   -0.1059     0.2353  -0.450   0.6526
## prison_na$rel   -0.1065     0.1348  -0.790   0.4292
## prison_na$dis    0.3391     0.1827   1.856   0.0635 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 143.20  on 273  degrees of freedom
## Residual deviance: 135.05  on 267  degrees of freedom
## AIC: 149.05
##
## Number of Fisher Scoring iterations: 6
```

```
nullmod_2 <- glm(prison_na$sui_bi ~ 1, family="binomial")
1-logLik(out1)/logLik(nullmod_2)
```

```
## 'log Lik.' 0.05826228 (df=7)
```

As expected, the probit also does not have significant coefficients and a pseudo-R^2 of 0.0583 (unsurprisingly the same as logit).

```
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
#install.packages("devtools")
#devtools::install_github("ChandlerLutz/starpolishr")
library(starpolishr)
star.out <- stargazer(out1, out2, ols)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Jul 05, 2021 - 10:58:09 AM

When I tried to add the row “star_insert_row(star.out,”R^2 and Pseudo R^2 & 0.0583 & 0.0583 & 0.0321 \\“, insert.after = 40)” I kept getting an error, so unfortunately the table doesn’t include that. I think we’ve learned two things from these models. First, this set of predictors does a bad job in explaining the variance in suicidality. I have two ideas for how to fix it. Maybe we should use a rare events binary model because those that assent to suicidal ideation are rare in the sample. We could also choose better predictors. We also see that logit and probit are comparable, which is exactly what was expected.

Table 1:

	Dependent variable:		
	sui_bi		
	logistic	*probit*	*OLS*
	(1)	(2)	(3)
race	−0.356	−0.169	−0.016
	(0.434)	(0.203)	(0.022)
self	0.406	0.154	0.031
	(0.545)	(0.269)	(0.035)
ses	−0.398	−0.185	−0.027
	(0.279)	(0.137)	(0.018)
opt	−0.202	−0.106	−0.016
	(0.468)	(0.235)	(0.033)
rel	−0.236	−0.107	−0.012
	(0.290)	(0.135)	(0.016)
dis	0.621*	0.339*	0.052*
	(0.354)	(0.183)	(0.028)
Constant	−1.422	−0.812	0.134
	(2.118)	(1.051)	(0.139)
Observations	274	274	274
R^2			0.032
Adjusted R^2			0.010
Log Likelihood	−67.428	−67.527	
Akaike Inf. Crit.	148.856	149.054	
Residual Std. Error			0.259 (df = 267)
F Statistic			1.476 (df = 6; 267)
Note:			*p<0.1; **p<0.05; ***p<0.01