HW 7

Propensity Score Matching

Marwin Carmo

Contents

The data for this homework is stored in the `hbo.csv` file and contains information on tv shows, movies, etc. on HBO up to 2020. The variables we will be using today are:

- `type`: Whether the media is a TV show or a Movie/Other (factor)
- `rating`: The parental guidelines rating (e.g., G, PG, etc.; factor)
- `imdb_score`: The media's score on IMDB (numeric)
- `genre`: The media's genre (factor)

```
# Load required packages
library(dplyr)
library(MatchIt)
library(stargazer)

# Read in data, convert character vars to factors,
# and filter to years of interest
hbo <- read.csv("hbo.csv") |>
  select(-X) |>
  mutate(across(where(is.character), as.factor)) |>
  filter(year >= 1999)
```

Question 1 [1 pt.]

Recent decades have been talked about as a "golden age" of television shows. Inspired by this, we will examine whether or not TV shows (from 1999 to 2020) are rated more highly than other types of media on HBO.

With the `hbo` data, use a linear regression to assess the effect of `type` on `imdb_score`. Based on the results of the regression, state below whether TV shows are rated significantly more highly than other types of media, at the $p < .05$ level.

```r
mod1 <- lm(imdb_score ~ type, data = hbo)
summary(mod1)
```

```
##
## Call:
## lm(formula = imdb_score ~ type, data = hbo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0204 -0.5759  0.1241  0.7796  3.3241
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.57592    0.04032  163.11   <2e-16 ***
## typeTV       0.94448    0.08119   11.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.114 on 1012 degrees of freedom
## Multiple R-squared:  0.1179, Adjusted R-squared:  0.1171
## F-statistic: 135.3 on 1 and 1012 DF,  p-value: < 2.2e-16
```

- The model shows that TV shows have a higher IMDB score on average than other types of media ($b = 0.94$, 95% CI [0.79, 1.10], $t(1012) = 11.63$, $p < .001$),

Question 2

Part a [3 pts.]

We can't randomly assign media to be either a TV show or a movie, so let's redo the above analysis using a matched sample (use the functions from the `MatchIt` package). Use nearest neighbor matching, based on `rating` and `genre` (missing data for these variables were already filtered out for you, so you do not need to worry about that).

Display the matched regression output below.

```r
match1 <- matchit(type ~ rating + genre ,data=hbo, method='nearest')
match1_df <- match.data(match1)

mod2b <- lm(imdb_score ~ type, data = match1_df)
summary(mod2b)
```

```
##
## Call:
## lm(formula = imdb_score ~ type, data = match1_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0204 -0.5204  0.1796  0.7796  2.3116
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.58840    0.07020  93.847   <2e-16 ***
## typeTV       0.93200    0.09928   9.387   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.11 on 498 degrees of freedom
## Multiple R-squared:  0.1503, Adjusted R-squared:  0.1486
## F-statistic: 88.12 on 1 and 498 DF,  p-value: < 2.2e-16
```

Part b [1 pt.]

When matched on `rating` and `genre`, do TV shows appear to be more highly rated than other types of media, at the $p < .05$ significant level? How does this compare to the original analysis above?

- In the matched model TV shows are still significantly highly rated than other types of media ($b = 0.93$, 95% CI [0.74, 1.13], $t(498) = 9.39$, $p < .001$). In this model the coefficient for the show type is slighly lower than in the original model.

Question 3

Part a [3 pts.]

Re-do the propensity matched analysis, but this time use exact matching instead of nearest neighbor matching. Display the regression output below.

```
match2 <- matchit(type ~ rating + genre ,data=hbo, method='exact')
match2_df <- match.data(match2)

mod3a <- lm(imdb_score ~ type, data = match2_df)

summary(mod3a)
```

```
##
## Call:
## lm(formula = imdb_score ~ type, data = match2_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0723 -0.5812  0.1232  0.7188  3.3188
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.58117    0.04154  158.43   <2e-16 ***
## typeTV       0.99117    0.08219   12.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.087 on 918 degrees of freedom
## Multiple R-squared:  0.1368, Adjusted R-squared:  0.1358
## F-statistic: 145.4 on 1 and 918 DF,  p-value: < 2.2e-16
```

Part b [1 pt.]

Did any subgroups get dropped?

```
og_subgroups <-
  hbo |>
  distinct(rating, genre) |>
  mutate(subgroups = paste(rating, genre)) |>
```

```
  pull(subgroups)


exact_matched_subgroups <-
  match2_df |>
  distinct(rating, genre) |>
  mutate(subgroups = paste(rating, genre)) |>
  pull(subgroups)

length(setdiff(og_subgroups, exact_matched_subgroups))
```

```
## [1] 36
```

- Yes, 36 subgroups.

Question 4 [1 pt.]

Using the `stargazer()` function, create a table of the regressions in questions 1, 2 and 4. (1 pt.)

```
stargazer(mod1, mod2b, mod3a, type = "latex", header = FALSE,
          ci=TRUE, digits=2, title = "IMDB scores predicted by show type.")
```

Table 1: IMDB scores predicted by show type.

	Dependent variable:		
	imdb_score		
	(1)	(2)	(3)
typeTV	0.94***	0.93***	0.99***
	(0.79, 1.10)	(0.74, 1.13)	(0.83, 1.15)
Constant	6.58***	6.59***	6.58***
	(6.50, 6.65)	(6.45, 6.73)	(6.50, 6.66)
Observations	1,014	500	920
R^2	0.12	0.15	0.14
Adjusted R^2	0.12	0.15	0.14
Residual Std. Error	1.11 (df = 1012)	1.11 (df = 498)	1.09 (df = 918)
F Statistic	135.31*** (df = 1; 1012)	88.12*** (df = 1; 498)	145.42*** (df = 1; 918)
Note:			*p<0.1; **p<0.05; ***p<0.01