Problem Set 8

W. Hunter Giles

Set-up

```r
library(tidyverse)     # For ggplot, mutate(), filter(), and friends
library(broom)         # For converting models to data frames
library(estimatr)      # For lm_robust() and iv_robust()
library(modelsummary)  # For showing side-by-side regression tables
library(MatchIt)       # For matching
library(rdrobust)      # For nonparametric RD
library(rddensity)     # For nonparametric RD density tests
library(haven)         # For reading Stata files

set.seed(1234)  # Make any random stuff be the same every time you run this

# Round everything to 3 digits by default
options("digits" = 3)

# Turn off the message that happens when you use group_by() and summarize()
options(dplyr.summarise.inform = FALSE)

# Load raw data
hisp_raw <- read_stata("../data/evaluation.dta")

# Make nice clean dataset to use for the rest of the assignment
hisp <- hisp_raw %>%
  # Having a numeric 0/1 column is sometimes helpful for things that don't like
  # categories, like matchit()
  mutate(enrolled_num = enrolled) %>%
  # Convert these 0/1 values to actual categories
  mutate(eligible = factor(eligible, labels = c("Not eligible", "Eligible")),
```

```r
         enrolled = factor(enrolled, labels = c("Not enrolled", "Enrolled")),
         round = factor(round, labels = c("Before", "After")),
         treatment_locality = factor(treatment_locality, labels = c("Control", "Treatment"
         promotion_locality = factor(promotion_locality, labels = c("No promotion", "Promo
  # Get rid of this hospital column because (1) we're not using it, and (2) half
  # of the households are missing data, and matchit() complains if any data is
  # missing, even if you're not using it
  select(-hospital)
```

Background

The World Bank's *Impact Evaluation in Practice* has used a hypothetical example of a health insurance program throughout the book. This Health Insurance Subsidy Program (HISP) provides subsidies for buying private health insurance to poorer households, with the goal of lowering personal health expenditures, since people can rely on insurance coverage instead of paying out-of-pocket. Think of the HISP as a version of the Affordable Care Act (ACA, commonly known as Obamacare).

The dataset includes a number of important variables you'll use throughout this assignment:

Variable name	Description
`health_expenditures`	Out of pocket health expenditures (per person per year)
`eligible`	Household eligible to enroll in HISP
`enrolled`	Household enrolled in HISP
`round`	Indicator for before and after intervention
`treatment_locality`	Household is located in treatment community
`poverty_index`	1-100 scale of poverty
`promotion_locality`	Household is located in community that received random promotion
`enrolled_rp`	Household enrolled in HISP following random promotion

It also includes several demographic variables about the households. **Each of these are backdoor confounders between health expenditures participation in the HISP**:

Variable name	Description
`age_hh`	Age of the head of household (years)
`age_sp`	Age of the spouse (years)
`educ_hh`	Education of the head of household (years)
`educ_sp`	Education of the spouse (years)
`female_hh`	Head of household is a woman (1 = yes)

Variable name	Description
`indigenous`	Head of household speaks an indigenous language (1 = yes)
`hhsize`	Number of household members
`dirtfloor`	Home has a dirt floor (1 = yes)
`bathroom`	Home has a private bathroom (1 = yes)
`land`	Number of hectares of land owned by household
`hospital_distance`	Distance to closest hospital (km)

You will use each of the five main econometric approaches for estimating causal effects to measure the effect of HISP on household health expenditures. **Don't worry about conducting in-depth baseline checks and robustness checks.** For the sake of this assignment, you'll do the minimum amount of work for each method to determine the causal effect of the program.

Task 1: RCTs

To measure the effect of HISP accurately, World Bank researchers randomly assigned different localities (villages, towns, cities, whatever) to treatment and control groups. Some localities were allowed to join HISP; others weren't.

```r
hisp_eligible <- hisp %>%
  filter(eligible == "Eligible")

hisp_after <- hisp %>%
  filter(round == "After")
```

Below are the average health expenditures for the treatment and control group before the intervention. The control group had an average expenditure of 17.39, and the the treatment group had an average expenditure of 17.02.

```r
hisp %>%
  filter(round == "Before") %>%
  group_by(treatment_locality) %>%
  summarize(mean = mean(health_expenditures))
```

```
# A tibble: 2 x 2
  treatment_locality  mean
  <fct>              <dbl>
1 Control             17.4
2 Treatment           17.0
```

After the intervention, the control group's average expenditure increases to 20.1 and the the treatment group's average expenditure decreases to 13.7.

```r
df <- hisp %>%
  filter(round == "After") %>%
  group_by(treatment_locality) %>%
  summarize(mean = mean(health_expenditures))
df
```

```
# A tibble: 2 x 2
  treatment_locality  mean
  <fct>              <dbl>
1 Control             20.1
2 Treatment           13.7
```

The treatment group has an average health expenditure that is 6.41 less than the control group after the intervention.

```r
diff(df$mean) # Control - Treatment
```

```
[1] -6.41
```

The linear regression shows that the treatment group's expenditure after the intervention is 6.41 less than the control group's.

```r
lm_model <- lm_robust(health_expenditures ~ treatment_locality,
          data = hisp_after,
          clusters = locality_identifier)
tidy(lm_model)
```

```
                         term estimate std.error statistic  p.value conf.low
1                 (Intercept)    20.06     0.379      52.9 6.81e-48    19.30
2 treatment_localityTreatment    -6.41     0.504     -12.7 3.32e-23    -7.41
  conf.high    df             outcome
1     20.83  53.5 health_expenditures
2     -5.41 108.6 health_expenditures
```

The confounders slightly biased the treatment effect away from zero. When controlling for confounders the treatment effect is 6.12, meaning the treatment group spends 6.12 less than the control group after the intervention.

```r
full_linear_model <- lm_robust(health_expenditures ~ treatment_locality + age_hh + age_sp
          data = hisp_after,
          clusters = locality_identifier
          )
tidy(full_linear_model)
```

```
                          term estimate std.error statistic  p.value conf.low
1                  (Intercept) 28.95706   0.80870    35.807 5.46e-58  27.3522
2  treatment_localityTreatment -6.12955   0.40172   -15.258 8.37e-29  -6.9258
3                       age_hh  0.10801   0.01495     7.224 1.15e-10   0.0783
4                       age_sp  0.00799   0.01643     0.486 6.28e-01  -0.0246
5                      educ_hh  0.11265   0.04600     2.449 1.60e-02   0.0214
6                      educ_sp -0.00980   0.05009    -0.196 8.45e-01  -0.1091
7                    female_hh  1.08976   0.47396     2.299 2.37e-02   0.1489
8                   indigenous -2.80641   0.37524    -7.479 4.02e-11  -3.5515
9                       hhsize -2.38237   0.06408   -37.180 5.05e-62  -2.5094
10                   dirtfloor -3.04384   0.29840   -10.201 2.25e-17  -3.6355
11                    bathroom  0.97106   0.25513     3.806 2.41e-04   0.4650
12                        land  0.16545   0.04006     4.130 1.01e-04   0.0855
13           hospital_distance -0.00600   0.00454    -1.320 1.91e-01  -0.0151
   conf.high    df             outcome
1   30.56195  97.7 health_expenditures
2   -5.33334 108.9 health_expenditures
3    0.13769  96.9 health_expenditures
4    0.04059  99.6 health_expenditures
5    0.20387 104.6 health_expenditures
6    0.08953 104.5 health_expenditures
7    2.03059  95.8 health_expenditures
8   -2.06131  93.4 health_expenditures
9   -2.25531 104.4 health_expenditures
10  -2.45215 104.7 health_expenditures
11   1.47710 102.1 health_expenditures
12   0.24538  68.6 health_expenditures
13   0.00306  71.3 health_expenditures
```

```r
modelsummary(list(
  "Simple Regression" = lm_model,
  "Multiple Regression" = full_linear_model
),
title = "Health Expenditure on Helath Insurance Program")
```

Table 3: Health Expenditure on Helath Insurance Program

	Simple Regression	Multiple Regression
(Intercept)	20.064	28.957
	(0.379)	(0.809)
treatment_localityTreatment	−6.406	−6.130
	(0.504)	(0.402)
age_hh		0.108
		(0.015)
age_sp		0.008
		(0.016)
educ_hh		0.113
		(0.046)
educ_sp		−0.010
		(0.050)
female_hh		1.090
		(0.474)
indigenous		−2.806
		(0.375)
hhsize		−2.382
		(0.064)
dirtfloor		−3.044
		(0.298)
bathroom		0.971
		(0.255)
land		0.165
		(0.040)
hospital_distance		−0.006
		(0.005)
Num.Obs.	9914	9914
R2	0.073	0.344
R2 Adj.	0.072	0.343
Std.Errors	by: locality_identifier	by: locality_identifier

Task 2: Inverse probability weighting and/or matching

Naive Model

According to the model below, the people who enrolled in the intervention had 12.9 less in health expenditures compared to the non-enrollees. However, this is an inaccurate representation because it includes both compilers and always-takers.

```
model.naive <- lm(health_expenditures ~ enrolled,
              data = hisp_after)
tidy(model.naive)
```

```
# A tibble: 2 x 5
  term             estimate std.error statistic p.value
  <chr>               <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)          20.7     0.124     167.        0
2 enrolledEnrolled    -12.9     0.227     -56.8       0
```

Inverse Probability Weighting

$$\frac{\text{Treatment}}{\text{Propensity}} + \frac{1 - \text{Treatment}}{1 - \text{Propensity}}$$

Logistic regression to model the probability of enrolling in the HISP based on demographic features.

```
model_logit <- glm(enrolled ~ age_hh + age_sp + educ_hh + educ_sp + female_hh + indigenous
                   data = hisp_after,
                   family = binomial(link = "logit"))
```

Below we fit the logistic regression model to get the probability of enrollment for each observation. When them mutate the probability to create the *inverse probability weighting ratio* (IPW). This ratio gives observations with with weird outcomes more weight.

```
enrolled_propensities <- augment_columns(model_logit, hisp_after,
                                         type.predict = "response") %>%
                                          rename(p_enrolled = .fitted)

enrolled_propensities <- enrolled_propensities %>%
  mutate(inverse_prob = (enrolled_num/p_enrolled)+((1-enrolled_num)/(1-p_enrolled))) %>%
  filter(inverse_prob <= 10)
```

A new linear model is used, but this time weights are included.

```r
ipw_model <- lm(health_expenditures ~ enrolled,
                             data = enrolled_propensities,
                             weights = inverse_prob)

tidy(ipw_model)
```

```
# A tibble: 2 x 5
  term             estimate std.error statistic p.value
  <chr>               <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)          19.8     0.134     148.        0
2 enrolledEnrolled    -11.0     0.194     -56.6       0
```

Below the naive model's results show that the effect of being enrolled in the program decreases the health expenditure by 12.87 on average ($p<.01$). The IPW model has a smaller coefficient magnitude of 11.00, meaning that participation in the program decrease health expenditure by 11.00 dollars per year. The IPW model can be assumed to be the causal effect, because it accounts for always-takers, and never-takers.

```r
modelsummary(list(
  "Naive" = model.naive,
  "IPW" = ipw_model
),
title = "Health Expenditures on Health Insurance Program")
```

Table 4: Health Expenditures on Health Insurance Program

	Naive	IPW
(Intercept)	20.707	19.830
	(0.124)	(0.134)
enrolledEnrolled	−12.867	−11.002
	(0.227)	(0.194)
Num.Obs.	9914	9869
R2	0.246	0.245
R2 Adj.	0.245	0.245
AIC	74 435.6	73 846.1
BIC	74 457.2	73 867.7
Log.Lik.	−37 214.778	−36 920.046
F	3225.402	3203.813
RMSE	10.33	13.41