Problem set 8: The Health Insurance Subsidy Program

Answer key - PMAP 8521, Spring 2021

April 12, 2021

Contents

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)
`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.

```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
library(kableExtra)     # For fancy table formatting

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

# Turn off the messages that happen when you use group_by() %>% summarize()
options(dplyr.summarise.inform = FALSE)

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

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")),
         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", "Promotion"))) %>%
  # 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)
```

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.

Here's what you should do:

- Make a new dataset that only looks at eligible households (`eligible == "Eligible"`)
- Make a new dataset that only looks at eligible households *after* the experiment (`round == "After"`)
- Calculate the average health expenditures in treatment and control localities (`treatment_locality`) *before* the intervention (`round == "Before"`). Were expenditures fairly balanced across treatment and control groups before the intervention?
- Calculate the average health expenditures in treatment and control localities *after* the intervention (`round == "After"`)
- Determine the difference in average health expenditures across treatment and control *after* the intervention
- Using data *after* the intervention, use linear regression to determine the difference in means and statistical significance of the difference (hint: you'll want to use `health_expenditures ~ treatment_locality`). Use `lm_robust()` from the **estimatr** package and cluster by `locality_identifier` if you're feeling adventurous.
- Create another model that controls for the confounders: `age_hh + age_sp + educ_hh + educ_sp + female_hh + indigenous + hhsize + dirtfloor + bathroom + land + hospital_distance`. (Use `lm_robust()` again if you're brave.) Does the estimate of the causal effect change? Why or why not?
- Show the results from the two regressions in a side-by-side table if you want

First we'll make some new datasets that select specific rows:

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

hisp_eligible_after <- hisp_eligible %>%
  filter(round == "After")
```

Next we want to see the average health expenditures before and after the program, in both treatment and control localities. There are lots of ways we can do this—perhaps the easiest is to group by both `round` and `treatment_locality` and look at the average in each group. Here we can see that the treatment and control groups had similar expenditures before the intervention (around $14.50). After the intervention, those in the treatment group only spent an average of $7.84, while those in the control group spent nearly $18, making a difference of $10.

```
# Check balance of health expenditures before/after the intervention
hisp_eligible %>%
  group_by(round, treatment_locality) %>%
  summarize(avg_expenditures = mean(health_expenditures))
```

round	treatment_locality	avg_expenditures
Before	Control	14.57
Before	Treatment	14.49
After	Control	17.98
After	Treatment	7.84

Because this is an RCT, we can arguably talk about the program effect as *casual*. We can use regression to get an exact estimate of the causal effect, along with its significance. We can also control for other factors (though, according to the world of DAGs, we don't actually need to control for other variables when doing an RCT—random assignment allows us to cut out nodes from our DAG). Here are two models: one simple one with an indicator for treatment status, and one more complicated one with demographic controls. The standard errors in each are clustered by locality.

In both models, the HISP treatment causes a $10 decrease in health expenditures, and the effect is nearly identical in the more complicated model since all those control variables should be randomly distributed between the treatment and control groups, and thus not have any effect.

```
model_rct_simple <- lm_robust(health_expenditures ~ treatment_locality,
                              data = hisp_eligible_after,
                              clusters = locality_identifier)

model_rct_controls <- lm_robust(health_expenditures ~ treatment_locality +
                                  age_hh + age_sp + educ_hh + educ_sp +
                                  female_hh + indigenous + hhsize + dirtfloor +
                                  bathroom + land + hospital_distance,
                                data = hisp_eligible_after,
                                clusters = locality_identifier)

modelsummary(list("RCT simple" = model_rct_simple, "RCT controls" = model_rct_controls),
             stars = TRUE, gof_omit = 'IC|Log|Adj') %>%
  row_spec(3, background = "yellow")
```

	RCT simple	RCT controls
(Intercept)	17.981***	27.565***
	(0.309)	(0.873)
treatment_localityTreatment	-10.140***	-10.010***
	(0.399)	(0.346)
age_hh		0.041***
		(0.015)
age_sp		0.003
		(0.017)
educ_hh		-0.039
		(0.047)
educ_sp		-0.022
		(0.050)
female_hh		0.643
		(0.446)
indigenous		-1.905***
		(0.354)
hhsize		-1.603***
		(0.066)
dirtfloor		-1.849***
		(0.280)
bathroom		0.285
		(0.249)
land		0.038
		(0.038)
hospital_distance		-0.003
		(0.004)
R2	0.300	0.430
N	5629	5629
se_type	CR2	CR2

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Task 2: Inverse probability weighting and/or matching

Instead of using experimental data, we can estimate the causal effect using observational data alone by closing all the confounding backdoors. In this task, you should choose one of two approaches: inverse probability weighting or matching.

Do the following (for both approaches):

- Make a dataset based on `hisp` that only includes observations from after the intervention (`round == "After"`). Even though you technically have a column that indicates if the household was in the treatment group (`treatment_locality`), you're going to pretend that you don't have it This is now observational data—all you know is that a bunch of households participated in HISP and a bunch didn't.
- Run a naive model that estimates the effect of HISP enrollment on health expenditures (`health_expenditures ~ enrolled`) using this after-only observational data. What is the effect? Is this accurate? Why or why not?

First we'll make a dataset that only includes households after the intervention. This dataset is a mix of treatment and control households—we're pretending that we can't see the `treatment_locality` column and treating this as if it were observational data.

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

According to this naive model, which doesn't deal with selection bias or close any confounding backdoor paths, enrollment in the HISP is associated with almost $13 lower health expenditures, on average. However, we can't ascribe causality to this coefficient, since we haven't isolated or identified the causal pathway. We can do that with inverse probability weighting or with matching (or sometimes by simply including the backdoors as controls, but only if we assume all the backdoors are related to each other linearly).

```
model_naive <- lm(health_expenditures ~ enrolled, data = hisp_after)
tidy(model_naive)
```

term	estimate	std.error	statistic	p.value
(Intercept)	20.7	0.124	167.1	0
enrolledEnrolled	-12.9	0.227	-56.8	0

If you're using inverse probability weighting, do the following:

- Use logistic regression to model the probability of enrolling in the HISP.
- Generate propensity scores for enrollment in the HISP.
- Add a new column to `enrolled_propensities` with `mutate()` that calculates the inverse probability weights using this formula (hint: "propensity" will be `p_enrolled`; "Treatment" will be `treatment_num`):

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

- Run a model that estimates the effect of HISP enrollment on health expenditures (`health_expenditures ~ enrolled`) using the `enrolled_propensities` data, weighting by your new inverse probability weights column. What is the causal effect of HISP on health expenditures? How does this compare to the naive model? Which do you believe more? Why?
- Show the results from the two regressions in a side-by-side table if you want

We can isolate the pathway between enrollment and health expenditures by using inverse probability weighting. We first use our backdoors to estimate a model that predicts the propensity (or probability) of enrolling in the program. We then plug each of our rows into that model to generate propensity scores. We then use those scores to generate weighted scores that are higher for observations that behave strangely (e.g. they had a low propensity score but enrolled in the program, or had a high propensity score and didn't enroll in the program). We finally run a regression model using those weights and find that enrolling in the HISP reduced health expenditures by $10.69, on average. This is a causal effect and not merely an association.

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

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

enrolled_ipw <- enrolled_propensities %>%
  mutate(ipw = (enrolled_num / p_enrolled) + ((1 - enrolled_num) / (1 - p_enrolled)))

model_ipw <- lm(health_expenditures ~ enrolled, data = enrolled_ipw, weights = ipw)
tidy(model_ipw)
```

term	estimate	std.error	statistic	p.value
(Intercept)	19.8	0.138	143.8	0
enrolledEnrolled	-10.7	0.196	-54.5	0

If you're using matching, do the following:

- Use `matchit()` to find the best matches for enrollment based on Mahalanobis nearest neighbor matching. The `matchit()` function can't work with categorical variables, so make sure you use `enrolled_num` instead of `enrolled`.

  It might take a minute to run the matching. If you include `cache=TRUE` in the chunk options, R will keep track of when the chunk changes; if you knit and there's been a change to the chunk, R will run the chunk, but if you knit and there's been no changes, R will use the previous output of the chunk and not actually run it.

- Run `summary(matched)` and see how many rows were matched and how many will be discarded.
- Use `match.data()` to store the results of the match as a new dataset.
- Run a model that estimates the effect of HISP enrollment on health expenditures (`health_expenditures ~ enrolled`) using the matched data, weighting by the `weights` column that `matchit()` generated. What is the causal effect of HISP on health expenditures? How does this compare to the naive model? Which do you believe more? Why?
- Show the results from the two regressions in a side-by-side table if you want

We can isolate the pathway between enrollment and health expenditures by using matching. We first use our backdoors to find sets of treated and untreated observations that are similar to each other based on the Mahalanobis distance of all their demographic characteristics (i.e. the confounding backdoors). We match with replacement so that an already-matched observation can be used again. All 2,965 treated observations are matched to 2,043 control observations, meaning that 4,906 rows will be discarded after matching.

We then use this matched data to estimate the effect of the HISP on health expenditures and find that enrolling in the program reduces expenditures by $9.92, on average. This is a causal effect and not merely an association.

```
matched <- matchit(enrolled_num ~ age_hh + age_sp + educ_hh + educ_sp +
                     female_hh + indigenous + hhsize + dirtfloor + bathroom +
                     land + hospital_distance,
                   data = hisp_after,
                   method = "nearest", distance = "mahalanobis", replace = TRUE)

# I'm leaving this commented out in the knitted version because there's no way
# to filter any of the output (this is why we use all the tidyverse tools! Those
# packages all fit together really nicely! There's no way to use tidy() with
# stuff from matchit(), and it instead spits out a billion rows of
# not-so-helpful output.)
# summary(matched)

# Actually, there's a way to get just the table of matched and unmatched counts.
# The summary(matched) function returns a list of a bunch of stuff, and the
# counts are in an element of that list named "nn", so running this will show
# just that:
summary(matched)$nn
```

```
##               Control Treated
## All (ESS)        6949    2965
## All              6949    2965
## Matched (ESS)    1545    2965
## Matched          2043    2965
## Unmatched        4906       0
## Discarded           0       0
```

```
hisp_matched <- match.data(matched)
```

```
model_matched <- lm(health_expenditures ~ enrolled_num,
                    data = hisp_matched, weights = weights)
tidy(model_matched)
```

term	estimate	std.error	statistic	p.value
(Intercept)	17.76	0.177	100	0
enrolled_num	-9.91	0.230	-43	0

Here are the results from all three of these observational models:

```
modelsummary(list("Naive" = model_naive, "IPW" = model_ipw, "Matching" = model_matched),
             stars = TRUE, gof_omit = 'IC|Log|Adj') %>%
  row_spec(c(3, 5), background = "yellow")
```

	Naive	IPW	Matching
(Intercept)	20.707***	19.830***	17.756***
	(0.124)	(0.138)	(0.177)
enrolledEnrolled	-12.867***	-10.691***	
	(0.227)	(0.196)	
enrolled_num			-9.915***
			(0.230)
Num.Obs.	9914	9914	5008
R2	0.246	0.230	0.270
F	3225.402	2965.764	1851.203

* p < 0.1, ** p < 0.05, *** p < 0.01

Task 3: Diff-in-diff

Instead of using experimental data, we can estimate the causal effect using observational data alone with a difference-in-difference approach. We have data indicating if households were enrolled in the program (`enrolled`) and data indicating if they were surveyed before or after the intervention (`round`), which means we can find the differences between enrolled/not enrolled before and after the program.

Do the following:

- Make a new dataset based on `hisp` that only includes observations from the localities that were randomly chosen for treatment (`treatment_locality == "Treatment"`)
- Using that new dataset, run a regression model that estimates the difference-in-difference effect of being enrolled in the HISP program (huge hint: use `health_expenditures ~ enrolled + round + enrolled * round`). Use `lm_robust()` and cluster by `locality_identifier` if you're brave. What is the causal effect of HISP on health expenditures?
- Run a second model that estimates the difference-in-difference effect, but control for the following variables: `age_hh + age_sp + educ_hh + educ_sp + female_hh + indigenous + hhsize + dirtfloor + bathroom + land + hospital_distance`. (Again, cluster by `locality_identifier` if you're brave.) How does the causal effect change?
- Show the results from the two regressions in a side-by-side table if you want

First we'll make a dataset that only includes observations from the localities that were randomly chosen for treatment:

```
hisp_treatment <- hisp %>%
  filter(treatment_locality == "Treatment")
```

Next we can build a regression model to find the DiD estimator. In class, we made a 2 × 2 table to find the average outcome before and after the program for both treatment and control groups, and then found the difference of those differences. You don't need to do that by hand in real life—we can just use an interaction term of `enrolled * round` to find the DiD estimator. This has a couple advantages: (1) it's a heck of a lot easier, and (2) you can include control variables.

The causal effect in this approach is smaller than what we found in the RCT ($8.16 here vs. $10ish in the RCT), but that's because here we're not dealing with experimental data. Even though the models here use purely observational data, they do a pretty good job of estimating the causal effect. The control variables have very little influence on the causal effect (again, likely because they're randomly distributed across the treatment and control groups).

```
model_dd_simple <- lm_robust(health_expenditures ~ enrolled + round + enrolled * round,
                             data = hisp_treatment,
                             clusters = locality_identifier)

model_dd_controls <- lm_robust(health_expenditures ~ enrolled + round + enrolled * round +
                                 age_hh + age_sp + educ_hh + educ_sp +
                                 female_hh + indigenous + hhsize + dirtfloor +
                                 bathroom + land + hospital_distance,
                               data = hisp_treatment,
                               clusters = locality_identifier)

modelsummary(list("DiD simple" = model_dd_simple, "DiD controls" = model_dd_controls),
             stars = TRUE, gof_omit = 'IC|Log|Adj') %>%
  row_spec(7, background = "yellow")
```

	DiD simple	DiD controls
(Intercept)	20.791***	27.395***
	(0.174)	(0.561)
enrolledEnrolled	-6.302***	-1.513***
	(0.194)	(0.130)
roundAfter	1.513***	1.451***
	(0.360)	(0.359)
enrolledEnrolled × roundAfter	-8.163***	-8.161***
	(0.321)	(0.321)
age_hh		0.080***
		(0.011)
age_sp		-0.020
		(0.013)
educ_hh		0.060**
		(0.029)
educ_sp		-0.077**
		(0.034)
female_hh		1.104***
		(0.318)
indigenous		-2.312***
		(0.239)
hhsize		-1.995***
		(0.039)
dirtfloor		-2.300***
		(0.165)
bathroom		0.500***
		(0.160)
land		0.091***
		(0.029)
hospital_distance		-0.003
		(0.003)
R2	0.344	0.552
N	9919	9919
se_type	CR2	CR2

* p < 0.1, ** p < 0.05, *** p < 0.01

Task 4: RDD

Eligibility for the HISP is determined by income. Households that have an income of less than 58 on a standardized 1-100 scale (`poverty_index`) qualify for the program and are automatically enrolled. Because we have an arbitrary cutoff in a running variable, we can use regression discontinuity to measure the effect of the program on health expenditures.

Do the following:

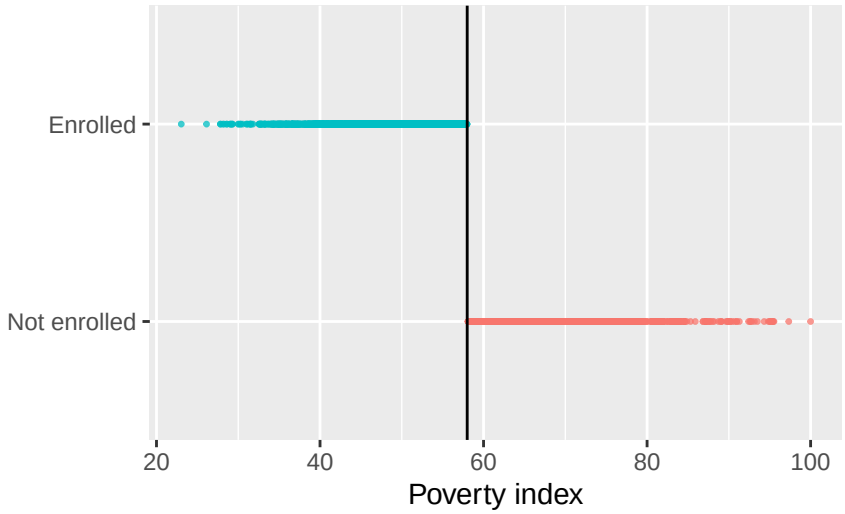
- Make a new dataset based on `hisp` that only includes observations from the localities that were randomly chosen for treatment (`treatment_locality == "Treatment"`)
- Use `mutate()` to add new variable that centers the poverty index variable at 58
- Determine if the discontinuity is sharp or fuzzy. (Hint: create a scatterplot with `poverty_index` on the x-axis, `enrolled` on the y-axis, and a vertical line at 58.)
- Determine if the distribution of the running variable (`poverty_index`) has a jump near the cutoff (it shouldn't). (Hint: create a histogram with `poverty_index` on the x-axis and a vertical line at 58. Use a McCrary test to see if there's a significant break in the distribution at 58.)
- Visualize the jump in outcome at the cutoff with a scatterplot (Hint: create a scatterplot with `poverty_index` on the x-axis, `health_expenditures` on the y-xis, color by `enrolled`, add a vertical line at 58, and add trendlines with `geom_smooth(method = "lm")`. You might want to adjust the size and transparency of the points with `geom_point(alpha = 0.2, size = 0.2)` or something similar.)
- Graphically, does it look like the HISP reduces health expenditures?
- Build a parametric regression model to estimate the size of the gap at the cutoff. You'll want to use the centered policy index variable to make it easier to interpret. You probably want to create a new dataset that only includes observations within some bandwidth that you choose (`filter(poverty_index_centered >= SOMETHING & poverty_index_centered <= SOMETHING)`). How big is the effect?
- Use `rdrobust()` from the **rdrobust** library to estimate the size of the gap nonparametrically. For the sake of simplicity, just use the default (automatic) bandwidth and kernel. How big is the effect?

First we'll make a dataset that only includes observations from the localities that were randomly chosen for treatment, and we'll center the poverty index at 58

```
hisp_rdd <- hisp %>%
  filter(treatment_locality == "Treatment") %>%
  mutate(poverty_index_centered = poverty_index - 58)
```

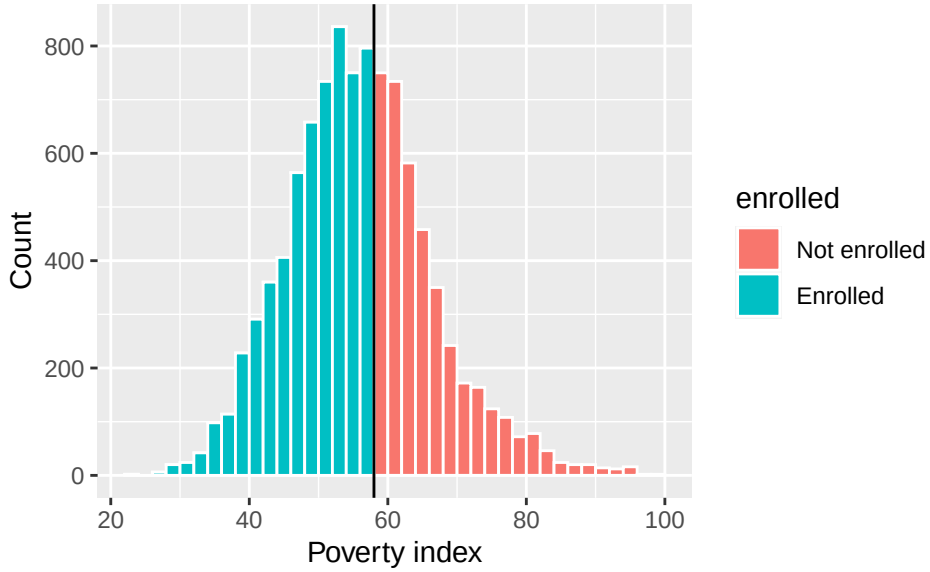
Next we'll check how sharp the discontinuity is. Here's it sharp (yay fake data). In real life, it would like be fuzzy. If that's the case, you'll need to use an instrument to account for the endogeneity in the choice to seek out treatment—often this can just be assignment to treatment or use of treatment.

```
ggplot(hisp_rdd, aes(x = poverty_index, y = enrolled, color = enrolled)) +
  geom_point(size = 0.5, alpha = 0.5) +
  geom_vline(xintercept = 58) +
  guides(color = FALSE) +
  labs(x = "Poverty index", y = NULL)
```



Now that we know this is a sharp design, we need to see if the running variable is evenly distributed around the cutoff—we don't want bunching up around the threshold. We can see this in a histogram—the distribution looks okay around 58:

```
ggplot(hisp_rdd, aes(x = poverty_index, fill = enrolled)) +
  geom_histogram(binwidth = 2, color = "white", boundary = 58) +
  geom_vline(xintercept = 58) +
  labs(x = "Poverty index", y = "Count")
```



We can test this more formally with a McCrary test. Here, the test t-statistic for a discontinuity in the distribution is 0.937 (p = 0.349), which is not statistically significant. We can confirm this with a McCrary plot.

```
density_check <- rddensity(X = hisp_rdd$poverty_index, c = 58)
summary(density_check)
```

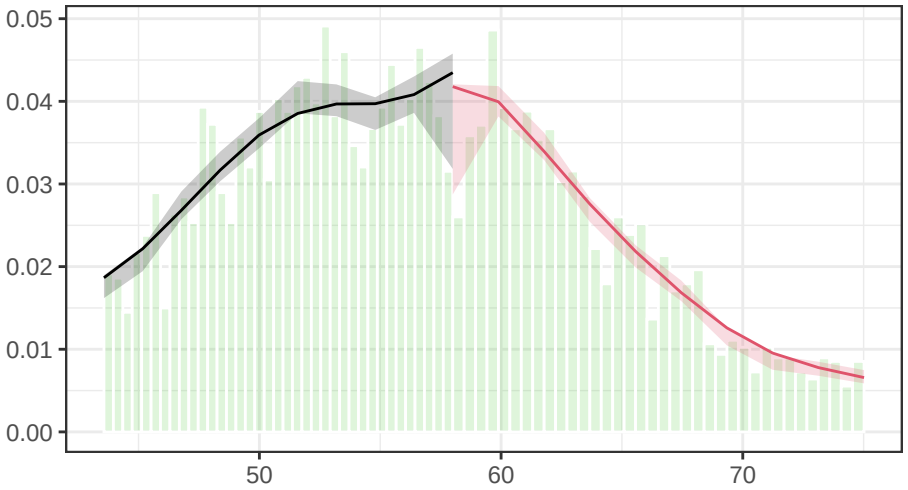
```
##
## Manipulation testing using local polynomial density estimation.
##
## Number of obs =       9919
```

```
## Model =                  unrestricted
## Kernel =                 triangular
## BW method =              estimated
## VCE method =             jackknife
##
## c = 58                   Left of c           Right of c
## Number of obs            5929                3990
## Eff. Number of obs       1858                1992
## Order est. (p)           2                   2
## Order bias (q)           3                   3
## BW est. (h)              4.81                5.673
##
## Method                   T                   P > |T|
## Robust                   -0.6822             0.4951
```

```
## Warning in summary.CJMrddensity(density_check): There are repeated observations. Point estimates and
## massPoints=FALSE to suppress this feature.
```

```
##
## P-values of binomial tests (H0: p=0.5).
##
## Window Length / 2          <c     >=c    P>|T|
## 0.114                       16      14    0.8555
## 0.229                       58      32    0.0080
## 0.343                       98      66    0.0152
## 0.457                      148     122    0.1280
## 0.572                      186     160    0.1789
## 0.686                      242     202    0.0641
## 0.801                      274     254    0.4083
## 0.915                      320     288    0.2086
## 1.029                      362     318    0.0991
## 1.144                      428     376    0.0720
```

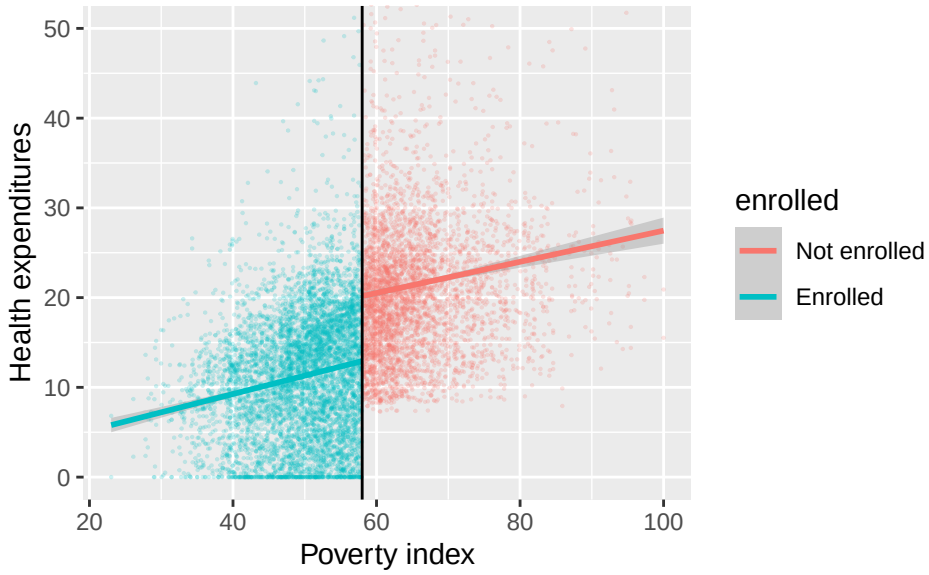
```
mccrary_test <- rdplotdensity(density_check, X = hisp_rdd$poverty_index)
```



Now that we've checked our main assumptions, we can check for jumps in health expenditures around the poverty index cutoff. It looks like there's a gap!

```
ggplot(hisp_rdd, aes(x = poverty_index, y = health_expenditures, color = enrolled)) +
  geom_point(alpha = 0.2, size = 0.1) +
```

```
  geom_smooth(method = "lm") +
  geom_vline(xintercept = 58) +
  labs(x = "Poverty index", y = "Health expenditures") +
  coord_cartesian(ylim = c(0, 50))  # Zoom in on the 0-50 range
```



We can measure the size of that gap a couple different ways. First, we can use a parametric (i.e. regular linear regression) model based on a bandwidth of 5 points around the cutoff. (I chose 5 arbitrarily here.) We use the centered poverty index variable—we don't need to, but it makes interpretation of the intercept term a little easier. According to this model, being enrolled in the program causes a decrease of $7.01 in health expenditures, and the effect is statistically significant.

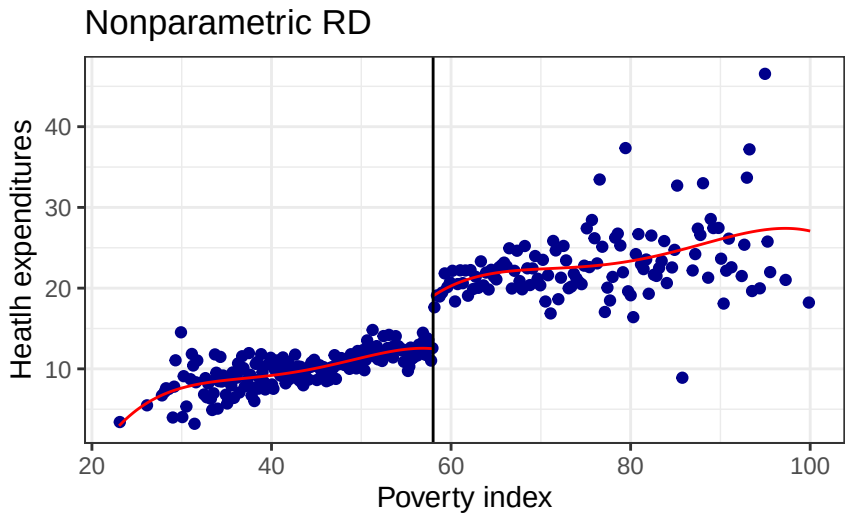
```
hisp_rdd_bw5 <- hisp_rdd %>%
  filter(poverty_index_centered >= -5 & poverty_index_centered <= 5)

model_rdd_parametric <- lm(health_expenditures ~ poverty_index_centered + enrolled,
                           data = hisp_rdd_bw5)
tidy(model_rdd_parametric)
```

term	estimate	std.error	statistic	p.value
(Intercept)	19.90	0.314	63.38	0.000
poverty_index_centered	0.20	0.098	2.05	0.041
enrolledEnrolled	-7.01	0.560	-12.52	0.000

We can also use a nonparametric model (i.e. not linear regression, but a more flexible curvy line). We'll use the default automatic bandwidth and kernel—in real life, you'd experiment with different bandwidths and kernel types to see how sensitive your effect is to those choices. Here, `rdrobust()` decided to use a triangular kernel with a bandwidth of 6.071, and the size of the gap at the cutoff is $6.47. This gap is statistically significant ($p < 0.001$).

```
# Plot nonparametric gap
rdplot(y = hisp_rdd$health_expenditures,
       x = hisp_rdd$poverty_index,
       c = 58,
```

```
       x.label = "Poverty index", y.label = "Heatlh expenditures",
       title = "Nonparametric RD")
```




```
# Build nonparametric model
model_rdd_nonparametric <- rdrobust(y = hisp_rdd$health_expenditures,
                                    x = hisp_rdd$poverty_index,
                                    c = 58)
```

```
## [1] "Mass points detected in the running variable."
```

```
summary(model_rdd_nonparametric)
```

```
## Call: rdrobust
##
## Number of Obs.                 9919
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                 5929        3990
## Eff. Number of Obs.            2498        2130
## Order est. (p)                    1           1
## Order bias  (q)                   2           2
## BW est. (h)                   6.359       6.359
## BW bias (b)                  10.803      10.803
## rho (h/b)                     0.589       0.589
## Unique Obs.                     717         669
##
## ===============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## ===============================================================================
##   Conventional     6.523     0.512    12.729     0.000     [5.519 , 7.528]
##         Robust         -         -    10.590     0.000     [5.236 , 7.614]
## ===============================================================================
```

Task 5: IVs/2SLS

Finally, we can use an instrument to remove the endogeneity from the choice to enroll in the HISP and estimate the causal effect from observational data. As you read in chapter 5, World Bank evaluators randomly selected households to receive encouragement to enroll in HISP. You can use this encouragement as an instrument for enrollment.

Do the following:

- Create a dataset based on `hisp` that only includes observations from after the intervention (`round == "After"`)
- Build a naive regression model that estimates the effect of HISP enrollment on health expenditures. You'll need to use the `enrolled_rp` variable instead of `enrolled`, since we're measuring enrollment after the encouragement intervention. (Hint: you'll want to use `health_expenditures ~ enrolled_rp`.) What does this naive model tell us about the effect of enrolling in HISP?
- Check the relevance, exclusion, and exogeneity of promotion (`promotion_locality`) as an instrument. For relevance, you'll want to run a model that predicts enrollment based on promotion (hint: `enrolled_rp ~ promotion_locality`) and check (1) the significance of the coefficient and (2) the F-statistic. For exclusion and exogeneity, you'll have to tell a convincing story that proves promotion influences health expenditures *only through* HISP enrollment.
- Run a 2SLS regression model with promotion as the instrument. You can do this by hand if you want (i.e. run a first stage model, extract predicted enrollment, and use predicted enrollment as the second stage), *or* you can just use the `iv_robust()` function from the **estimatr** library. (Hint: you'll want to use `health_expenditures ~ enrolled_rp | promotion_locality` as the formula). After removing the endogeneity from enrollment, what is the casual effect of enrollment in the HISP on health expenditures?
- Show the results from the two regressions in a side-by-side table if you want

First we'll make a dataset that only includes observations from the localities that were randomly chosen for treatment:

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

Next we'll build a naive regression model that estimates the effect of being enrolled in HISP on health expenditure. This gives an effect (-$12.70), but the coefficient is wrong—the `enrolled_rp` variable has inherent endogeneity that we have to get rid of (i.e. some people self selected into the program and other omitted variables might explain why they did).

```
model_iv_naive <- lm(health_expenditures ~ enrolled_rp,
                     data = hisp_after)
tidy(model_iv_naive)
```

term	estimate	std.error	statistic	p.value
(Intercept)	20.6	0.124	165.9	0
enrolled_rp	-12.7	0.229	-55.5	0

To excise the endogeneity from `enrolled_rp`, we can use an instrument: whether or not someone received encouragement to enroll in HISP. We need to make sure the instrument meets our three criteria:

1. ***Relevance*: Promotion should have some sort of significant relationship with enrollment, and the model should have an F-statistic of at least 10. That is definitely the case here. Receiving encouragement/promotion is associated with a 0.4 unit increase in enrollment (which is technically measured on a 0-2 scale, so it's hard to interpret here), and that rela-**

tionship is statistically significant ($p < 0.001$). The F-statistic is 2485, which is definitely bigger than 104 (and definitely bigger than 10). I'd say this is a relevant instrument.

```
model_iv_check <- lm(enrolled_rp ~ promotion_locality,
                     data = hisp_after)
tidy(model_iv_check)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.084	0.006	14.4	0
promotion_localityPromotion	0.408	0.008	49.8	0

```
glance(model_iv_check)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.2	0.2	0.407	2485	0	1	-5158	10322	10343	1643	9912	9914

2. ***Exclusion*: Promotion should have an effect on health expenditures, but *only through* enrollment and for no other reasons. I think the story here makes sense. Getting a letter in the mail encouraging you to enroll in HISP probably doesn't make you spend less on health insurance on its own. Maybe it could cause you to look for a cheaper health insurance plan, but probably not. The only way you're going to spend less on health expenses after receiving advertising for HISP is if you then sign up for HISP.**
3. ***Exogeneity*: Promotion shouldn't be related to other variables (both omitted and present) that cause health expenditures. That's probably the case here. Unlike other examples we've seen like parental health status being unrelated to health or SES or other behaviors (which clearly isn't the case), it's unlikely that receiving marketing materials is related with other factors that influence health spending.**

We'll call this a good instrument.

Now we can run a 2SLS model using promotion as an instrument to remove the endogeneity from enrollment. In previous problem sets you did this process by hand, running the first stage, extracting predicted `enrolled_rp` from the model, and then using predicted `enrolled_rp` in the second stage to find the effect of enrollment on health expenditures. We can use `iv_robust()` to do all of that for us behind the scenes. The effect of the program after using two-stage least squares with promotion as an instrument is -9.5, meaning that the HISP program causes a decrease of $9.50 in health expenditures. The effect is statistically significant.

```
model_2sls <- iv_robust(health_expenditures ~ enrolled_rp | promotion_locality,
                        data = hisp_after)
tidy(model_2sls)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
(Intercept)	19.6	0.181	108.8	0	19.3	20.00	9912	health_expenditures
enrolled_rp	-9.5	0.516	-18.4	0	-10.5	-8.49	9912	health_expenditures

Task 6: Summary

You just calculated a bunch of causal effects. Which one do you trust the most? Why?

Here's a summary of all the causal effects we just calculated:

Approach	Effect
RCT (simple)	-10.14
RCT (controls)	-10.01
Inverse probability weighting	-10.69
Mahalanobis- distance matching	-9.92
Diff-in-diff (simple)	-8.16
Diff-in-diff (controls)	-8.16
RDD (parametric)	-7.01
RDD (nonparametric)	-6.47
IV/2SLS	-9.5

In theory, the *true* effect is the RCT estimate of $10.14, since that's based on a gold-standard experimental research design. The other approaches are all based on non-experimental observational data, but they get fairly close to the RCT effect. In this situation, IPW and matching do fairly well, which is reassuring. IV/2SLS gets fairly close—but that doesn't mean IV is always the best. If I had to rely only on observational data (i.e. I had no idea the true effect was $10), I'd probably go with RDD or diff-in-diff. For me personally, all the contortions you have to go through to justify an IV sometimes strain credulity. RDD feels much cleaner—there's an arbitrary line, and people around the line are roughly the same. Diff-in-diff also feels cleaner, since there are clear before/after treatment/control groups (though you have to assume parallel trends in the two groups, and there's no way to do that with the data we have).

In the end, it's clear that the HISP has an effect, but it's difficult (perhaps impossible) to know which approach is the best.