R Code for Mastering 'Metrics

Jeffrey B. Arnold

Contents

Welcome 5
Install 5
License 5
Colonophon 5

I Chapter 1 9

1 National Health Interview Survey 11
1.1 References 13

2 RAND Health Insurance Experiment (HIE) 15
2.1 Table 1.3 15
2.2 Table 1.4 18
References 21

II Chapter 3 23

3 Minneapolis Domestic Violence Experiment 25
3.1 References 27

III Chapter 4 29

4 MLDA Regression Discontinuity 31
4.1 References 33

IV Chapter 5 35

5 Mississippi Bank Failures in the Great Depression 37
5.1 References 38

6 MLDA Difference-in-Difference 39
6.1 Table 5.2 39
6.2 Table 5.3 41
6.3 References 42

V Chapter 6 43

7 Twins and Returns to Schooling 45

References . . . 47

8 Child Labor Laws as an IV **49**
8.1 First stages and reduced forms . . . 49
8.2 IV returns . . . 50
References . . . 50

9 Quarter of Birth and Returns to Schooling **51**
9.1 Table 6.5. . . . 52
9.2 Figures . . . 55
References . . . 56

10 Sheepskin and Returns to Schooling **57**
10.1 Figure 1 . . . 57
10.2 Figure 2 . . . 58
References . . . 59

References **61**

Welcome

This work contains R code to reproduce many of the analyses in *Mastering 'Metrics* by Joshua D. Angrist and Jörn-Steffen Pischke (J. D. Angrist and Pischke 2014). This work provides R translations of the replication code available at masteringmetrics.com.

Install

To install all packages used in the examples in this work and the datasets from *Mastering 'Metrics* run

```
devtools::install_github("jrnold/masteringmetrics", subdir = "masteringmetrics")
```

License

Colonophon

The book is powered by https://bookdown.org which makes it easy to turn R markdown files into HTML, PDF, and EPUB.

This book was built with:

```
devtools::session_info(c("tidyverse"))
#> Session info -------------------------------------------------------------
#>  setting  value
#>  version  R version 3.4.4 (2018-03-15)
#>  system   x86_64, darwin15.6.0
#>  ui       X11
#>  language (EN)
#>  collate  en_US.UTF-8
#>  tz       America/Los_Angeles
#>  date     2018-04-08
#> Packages -----------------------------------------------------------------
#>  package      * version    date       source
#>  assertthat     0.2.0      2017-04-11 CRAN (R 3.4.0)
#>  backports      1.1.2      2017-12-13 CRAN (R 3.4.3)
#>  base64enc      0.1-3      2015-07-28 CRAN (R 3.4.0)
#>  BH             1.66.0-1   2018-02-13 CRAN (R 3.4.3)
#>  bindr          0.1.1.9000 2018-04-07 Github (krlmlr/bindr@b6e6fd6)
```

```
#>  bindrcpp       0.2.2.9000 2018-04-07 Github (krlmlr/bindrcpp@bd5ae73)
#>  broom          0.4.4      2018-03-29 cran (@0.4.4)
#>  callr          2.0.2      2018-02-11 CRAN (R 3.4.3)
#>  cellranger     1.1.0      2016-07-27 CRAN (R 3.4.0)
#>  cli            1.0.0      2017-11-05 cran (@1.0.0)
#>  colorspace     1.3-2      2016-12-14 CRAN (R 3.4.0)
#>  compiler       3.4.4      2018-03-15 local
#>  crayon         1.3.4      2017-09-16 CRAN (R 3.4.1)
#>  curl           3.2        2018-03-28 CRAN (R 3.4.4)
#>  DBI            0.8        2018-03-02 CRAN (R 3.4.3)
#>  dbplyr         1.2.1      2018-02-19 CRAN (R 3.4.3)
#>  debugme        1.1.0      2017-10-22 CRAN (R 3.4.2)
#>  dichromat      2.0-0      2013-01-24 CRAN (R 3.4.0)
#>  digest         0.6.15     2018-01-28 CRAN (R 3.4.3)
#>  dplyr          0.7.4.9003 2018-04-07 Github (tidyverse/dplyr@b7aaa95)
#>  evaluate       0.10.1     2017-06-24 CRAN (R 3.4.0)
#>  forcats        0.3.0      2018-02-19 cran (@0.3.0)
#>  foreign        0.8-69     2017-06-22 CRAN (R 3.4.4)
#>  ggplot2        2.2.1      2016-12-30 CRAN (R 3.4.0)
#>  glue           1.2.0      2017-10-29 CRAN (R 3.4.2)
#>  graphics     * 3.4.4      2018-03-15 local
#>  grDevices    * 3.4.4      2018-03-15 local
#>  grid           3.4.4      2018-03-15 local
#>  gtable         0.2.0      2016-02-26 CRAN (R 3.4.0)
#>  haven          1.1.1.9000 2018-04-08 Github (tidyverse/haven@746eb3e)
#>  highr          0.6        2016-05-09 CRAN (R 3.4.0)
#>  hms            0.4.2      2018-03-10 cran (@0.4.2)
#>  htmltools      0.3.6      2017-04-28 CRAN (R 3.4.0)
#>  httr           1.3.1      2017-08-20 CRAN (R 3.4.1)
#>  jsonlite       1.5        2017-06-01 CRAN (R 3.4.0)
#>  knitr          1.20       2018-02-20 CRAN (R 3.4.3)
#>  labeling       0.3        2014-08-23 CRAN (R 3.4.0)
#>  lattice        0.20-35    2017-03-25 CRAN (R 3.4.4)
#>  lazyeval       0.2.1      2017-10-29 CRAN (R 3.4.2)
#>  lubridate      1.7.3      2018-02-27 CRAN (R 3.4.3)
#>  magrittr       1.5        2014-11-22 CRAN (R 3.4.0)
#>  markdown       0.8        2017-04-20 CRAN (R 3.4.0)
#>  MASS           7.3-49     2018-02-23 CRAN (R 3.4.3)
#>  methods        3.4.4      2018-03-15 local
#>  mime           0.5        2016-07-07 CRAN (R 3.4.0)
#>  mnormt         1.5-5      2016-10-15 CRAN (R 3.4.0)
#>  modelr         0.1.1      2017-07-24 CRAN (R 3.4.1)
#>  munsell        0.4.3      2016-02-13 CRAN (R 3.4.0)
#>  nlme           3.1-137    2018-04-07 CRAN (R 3.4.4)
#>  openssl        1.0.1      2018-03-03 CRAN (R 3.4.3)
#>  parallel       3.4.4      2018-03-15 local
#>  pillar         1.2.1      2018-02-27 CRAN (R 3.4.3)
#>  pkgconfig      2.0.1      2017-03-21 CRAN (R 3.4.0)
#>  plogr          0.2.0      2018-03-25 CRAN (R 3.4.4)
#>  plyr           1.8.4      2016-06-08 CRAN (R 3.4.0)
#>  praise         1.0.0      2015-08-11 CRAN (R 3.4.0)
#>  psych          1.8.3.3    2018-03-30 cran (@1.8.3.3)
#>  purrr          0.2.4      2017-10-18 cran (@0.2.4)
```

```
#>  R6             2.2.2      2017-06-17 cran (@2.2.2)
#>  RColorBrewer   1.1-2      2014-12-07 CRAN (R 3.4.0)
#>  Rcpp           0.12.16    2018-03-13 cran (@0.12.16)
#>  readr          1.1.1      2017-05-16 CRAN (R 3.4.0)
#>  readxl         1.0.0      2017-04-18 CRAN (R 3.4.0)
#>  rematch        1.0.1      2016-04-21 CRAN (R 3.4.0)
#>  reprex         0.1.2      2018-01-26 CRAN (R 3.4.3)
#>  reshape2       1.4.3      2017-12-11 CRAN (R 3.4.3)
#>  rlang          0.2.0.9001 2018-04-08 Github (tidyverse/rlang@9c0637a)
#>  rmarkdown      1.9        2018-03-01 CRAN (R 3.4.3)
#>  rprojroot      1.3-2      2018-01-03 CRAN (R 3.4.3)
#>  rstudioapi     0.7        2017-09-07 CRAN (R 3.4.1)
#>  rvest          0.3.2      2016-06-17 CRAN (R 3.4.0)
#>  scales         0.5.0      2017-08-24 CRAN (R 3.4.1)
#>  selectr        0.4-1      2018-04-06 CRAN (R 3.4.4)
#>  stats        * 3.4.4      2018-03-15 local
#>  stringi        1.1.7      2018-03-12 cran (@1.1.7)
#>  stringr        1.3.0      2018-02-19 cran (@1.3.0)
#>  testthat       2.0.0      2017-12-13 cran (@2.0.0)
#>  tibble         1.4.2      2018-01-22 cran (@1.4.2)
#>  tidyr          0.8.0      2018-01-29 CRAN (R 3.4.3)
#>  tidyselect     0.2.4      2018-02-26 cran (@0.2.4)
#>  tidyverse      1.2.1      2017-11-14 cran (@1.2.1)
#>  tools          3.4.4      2018-03-15 local
#>  utf8           1.1.3      2018-01-03 cran (@1.1.3)
#>  utils        * 3.4.4      2018-03-15 local
#>  viridisLite    0.3.0      2018-02-01 CRAN (R 3.4.3)
#>  whisker        0.3-2      2013-04-28 CRAN (R 3.4.0)
#>  withr          2.1.2      2018-03-15 CRAN (R 3.4.4)
#>  xml2           1.2.0      2018-01-24 CRAN (R 3.4.3)
#>  yaml           2.1.18     2018-03-08 CRAN (R 3.4.4)
```

Part I

Chapter 1

Chapter 1

National Health Interview Survey

This reproduces the analyses in Table 1.1 of J. D. Angrist and Pischke (2014). which compares people with and without health insurance in the 2009 National Health Interview Survey (NHIS).

The code is derived from NHIS2009_hicompare.do.

Load the prerequisite packages.

```
library("tidyverse")
library("magrittr")
library("haven")
```

Load the data (originally from http://masteringmetrics.com/wp-content/uploads/2015/01/Data.zip), and adjust a few of the columns to account for differences in how Stata and R store data.

```
data("NHIS2009", package = "masteringmetrics")
```

Remove missing values.

```
NHIS2009 <- NHIS2009 %>%
  filter(marradult, perweight != 0) %>%
  group_by(serial) %>%
  mutate(hi_hsb = mean(hi_hsb1, na.rm = TRUE)) %>%
  filter(!is.na(hi_hsb), !is.na(hi)) %>%
  mutate(female = sum(fml)) %>%
  filter(female == 1) %>%
  select(-female)
```

For the sample only include married adults between 26 and 59 in age, and remove single person households.

```
NHIS2009 <- NHIS2009 %>%
  filter(between(age, 26, 59),
         marradult, adltempl >= 1)
```

Keep only single family households.

```
NHIS2009 <- NHIS2009 %>%
  group_by(serial) %>%
  filter(length(serial) > 1L) %>%
  ungroup()
```

Tables of wives and husbands by health insurance. status. The weighting following the "analytic" weights in the original `.do` file which weights observations by `perweight` and normalizes the weights so that the sub-samples of males and females have the same number as the original sample.

```
NHIS2009 %>%
  group_by(fml) %>%
  # normalize person weights to match number of observations in each
  # group
  mutate(perweight = perweight / sum(perweight) * n()) %>%
  group_by(fml, hi) %>%
  summarise(n_wt = sum(perweight)) %>%
  group_by(fml) %>%
  mutate(prop = n_wt / sum(n_wt))
#> # A tibble: 4 x 4
#> # Groups:   fml [2]
#>   fml      hi  n_wt  prop
#>   <lgl> <dbl> <dbl> <dbl>
#> 1 FALSE    0. 1281. 0.136
#> 2 FALSE    1. 8114. 0.864
#> 3 TRUE     0. 1131. 0.120
#> 4 TRUE     1. 8264. 0.880
```

Compare sample statistics of mean and women, with and without health insurance.

```
varlist <- c("hlth", "nwhite", "age", "yedu", "famsize", "empl", "inc")
NHIS2009_diff <- NHIS2009 %>%
  # rlang::set_attrs with NULL removes attributes from columns.
  # this avoids a warning from gather about differing attributes
  map_dfc(~ rlang::set_attrs(.x, NULL)) %>%
  select(fml, hi, one_of(varlist)) %>%
  gather(variable, value, -fml, -hi) %>%
  group_by(fml, hi, variable) %>%
  summarise(mean = mean(value, na.rm = TRUE), sd = sd(value, na.rm = TRUE)) %>%
  gather(stat, value, -fml, -hi, -variable) %>%
  unite(stat_hi, stat, hi) %>%
  spread(stat_hi, value) %>%
  mutate(diff = mean_1 - mean_0)
```

```
knitr::kable(NHIS2009_diff, digits = 3)
```

fml	variable	mean_0	mean_1	sd_0	sd_1	diff
FALSE	age	4.13e+01	4.42e+01	8.40e+00	8.61e+00	2.893
FALSE	empl	8.52e-01	9.22e-01	3.55e-01	2.68e-01	0.070
FALSE	famsize	4.06e+00	3.55e+00	1.54e+00	1.32e+00	-0.506
FALSE	hlth	3.70e+00	3.98e+00	1.01e+00	9.34e-01	0.278
FALSE	inc	4.36e+04	1.04e+05	3.57e+04	5.48e+04	60366.415
FALSE	nwhite	1.88e-01	2.00e-01	3.91e-01	4.00e-01	0.011
FALSE	yedu	1.12e+01	1.41e+01	3.47e+00	2.68e+00	2.919
TRUE	age	3.95e+01	4.22e+01	8.26e+00	8.65e+00	2.631
TRUE	empl	5.41e-01	7.58e-01	4.98e-01	4.29e-01	0.216
TRUE	famsize	4.07e+00	3.55e+00	1.54e+00	1.32e+00	-0.520
TRUE	hlth	3.61e+00	3.99e+00	1.02e+00	9.28e-01	0.382
TRUE	inc	4.36e+04	1.03e+05	3.52e+04	5.51e+04	59722.242
TRUE	nwhite	1.83e-01	2.02e-01	3.87e-01	4.01e-01	0.018
TRUE	yedu	1.14e+01	1.43e+01	3.50e+00	2.60e+00	2.913

1.1 References

- http://masteringmetrics.com/wp-content/uploads/2014/12/ReadMe_NHIS.txt
- http://masteringmetrics.com/wp-content/uploads/2015/01/NHIS2009_hicompare.do

Chapter 2

RAND Health Insurance Experiment (HIE)

This provides code replicates the Tables 1.3 and 1.4 of J. D. Angrist and Pischke (2014) which replicate the analyses from the RAND Health Insurance Experiment (Brook et al. 1983,Aron-Dine, Einav, and Finkelstein (2013)).

Load necessary libraries.

```
library("tidyverse")
library("broom")
library("haven")
library("rlang")
library("clubSandwich")
```

Function to calculate clustered standard errors and return a tidy data frame of the coefficients and standard errors.

```
cluster_se <- function(mod, cluster, type = "CR2") {
  vcov <- vcovCR(mod, cluster = cluster, type = type)
  coef_test(mod, vcov = vcov) %>%
    rownames_to_column(var = "term") %>%
    as_tibble() %>%
    select(term, estimate = beta, std.error = SE)
}
```

2.1 Table 1.3

J. D. Angrist and Pischke (2014) Table 1.3 presents demographic and baseline health characteristics for subjects of the RAND Health Insurance Experiment (HIE).

Load the `rand` data.

```
data("rand_sample", package = "masteringmetrics")
```

Calculate the number in each plan:

```
plantypes <- count(rand_sample, plantype)
```

```
knitr::kable(plantypes)
```

plantype	n
Catastrophic	759
Deductible	881
Coinsurance	1022
Free	1295

For each variable variables, estimate the the difference in means between heath insurance plan types.

```
varlist <- c("female", "blackhisp", "age", "educper",
             "income1cpi", "hosp", "ghindx", "cholest", "diastol",
             "systol", "mhi", "ghindxx",
             "cholestx", "diastolx", "systolx", "mhix")
```

Create column (1) with the mean and standard deviation of the "Catastrophic" plan,

```
catastrophic_stats <- rand_sample %>%
  filter(plantype == "Catastrophic") %>%
  select(one_of(varlist)) %>%
  gather(variable, value) %>%
  group_by(variable) %>%
  summarise(Mean = mean(value, na.rm = TRUE),
            `Std. Dev.` = sd(value, na.rm = TRUE))
```

```
knitr::kable(catastrophic_stats, digits = 3)
```

variable	Mean	Std. Dev.
age	3.24e+01	1.29e+01
blackhisp	1.72e-01	3.77e-01
cholest	2.07e+02	3.99e+01
cholestx	2.03e+02	4.21e+01
diastol	7.48e+01	1.10e+01
diastolx	7.88e+01	1.20e+01
educper	1.21e+01	2.88e+00
female	5.60e-01	4.97e-01
ghindx	7.09e+01	1.49e+01
ghindxx	6.85e+01	1.59e+01
hosp	1.15e-01	3.20e-01
income1cpi	3.16e+04	1.81e+04
mhi	7.38e+01	1.43e+01
mhix	7.55e+01	1.48e+01
systol	1.22e+02	1.65e+01
systolx	1.22e+02	1.87e+01

The difference in means between plans and the catastophic plan.

```
calc_diffs <- function(x) {
  # programmatically create the formula for lm
  f <- quo(!!sym(x) ~ plantype)
  mod <- lm(f, data = rand_sample)  # nolint
  out <- cluster_se(mod, cluster = rand_sample[["fam_identifier"]])
  out[["response"]] <- x
  out
}
```

```
plantype_diffs <- map_dfr(varlist, calc_diffs) %>%
  select(response, term, estimate, std.error) %>%
  mutate(term = str_replace(term, "^plantype", ""))
```

Create a table similar to J. D. Angrist and Pischke (2014) Table 1.3.

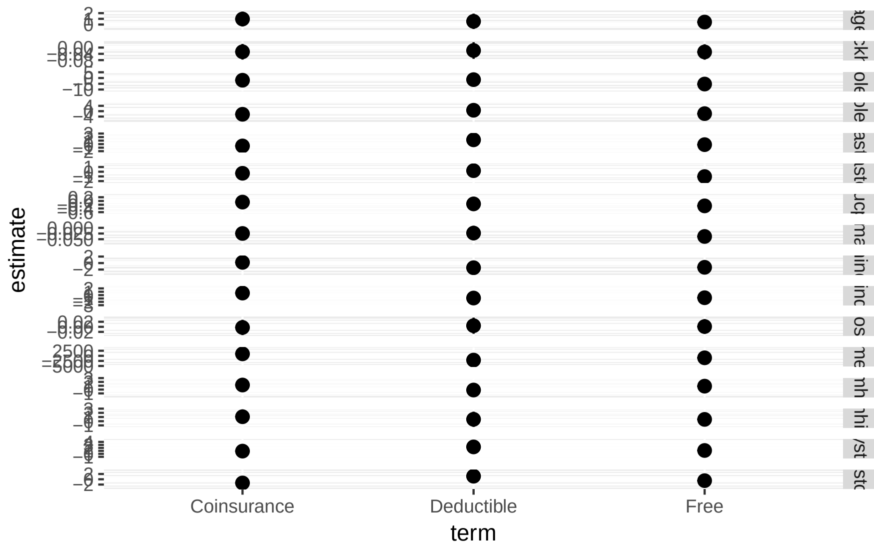
```
fmt_num <- function(x) {
  prettyNum(x, digits = 3, format = "f", big.mark = ",", drop0trailing = FALSE)
}

plantype_diffs %>%
  mutate(estimate = str_c(fmt_num(estimate), " (", fmt_num(std.error), ")")) %>%
  select(-std.error) %>%
  spread(term, estimate) %>%
  knitr::kable(digits = 3)
```

response	(Intercept)	Coinsurance	Deductible	Free
age	32.4 (0.485)	0.966 (0.655)	0.561 (0.676)	0.435 (0.614)
blackhisp	0.172 (0.0199)	-0.0269 (0.025)	-0.0188 (0.0266)	-0.0281 (0.0245)
cholest	207 (1.99)	-1.93 (2.76)	-1.42 (2.99)	-5.25 (2.7)
cholestx	203 (1.87)	-2.31 (2.47)	0.691 (2.58)	-1.83 (2.39)
diastol	74.8 (0.569)	-0.514 (0.786)	1.22 (0.831)	-0.143 (0.721)
diastolx	78.8 (0.466)	-0.335 (0.617)	0.219 (0.648)	-1.03 (0.588)
educper	12.1 (0.14)	-0.0613 (0.186)	-0.157 (0.191)	-0.263 (0.183)
female	0.56 (0.0118)	-0.0247 (0.0153)	-0.0231 (0.016)	-0.0379 (0.015)
ghindx	70.9 (0.694)	0.211 (0.922)	-1.44 (0.952)	-1.31 (0.872)
ghindxx	68.5 (0.702)	0.612 (0.903)	-0.869 (0.964)	-0.776 (0.867)
hosp	0.115 (0.0117)	-0.00249 (0.0152)	0.00449 (0.016)	0.00117 (0.0146)
income1cpi	31,603 (1,073)	970 (1,391)	-2,104 (1,386)	-976 (1,346)
mhi	73.8 (0.619)	1.19 (0.81)	-0.12 (0.822)	0.89 (0.766)
mhix	75.5 (0.696)	1.07 (0.872)	0.454 (0.911)	0.433 (0.826)
systol	122 (0.805)	0.907 (1.08)	2.32 (1.16)	1.12 (1.01)
systolx	122 (0.782)	-1.39 (0.986)	1.17 (1.06)	-0.522 (0.934)

Plot the difference-in-means of each plantype vs. catastrophic insurance.

```
ggplot(filter(plantype_diffs, term != "(Intercept)"),
              aes(x = term, y = estimate,
                  ymin = estimate - 2 * std.error,
                  ymax = estimate + 2 * std.error)) +
  geom_hline(yintercept = 0, colour = "white", size = 1) +
  geom_pointrange() +
  facet_grid(response ~ ., scales = "free_y")
```



2.2 Table 1.4

Replicate J. D. Angrist and Pischke (2014) Table 1.4 which presents health outcome and health expenditure results from the RAND HIE.

```
data("rand_person_spend", package = "masteringmetrics")
```

Correlate year variable from annual expenditures data to correct calendar year in order to adjust for inflation.

```
rand_person_spend <- mutate(rand_person_spend,
                            expyear = indv_start_year + year - 1)
```

Adjust spending for inflation. The CPI adjustment values below are based on the June CPI from 1991 (see table found at http://www.seattle.gov/financedepartment/cpi/historical.htm).

```
cpi <- tribble(
  ~ year, ~ cpi,
  1973, 3.07,
  1974, 2.76,
  1975, 2.53,
  1976, 2.39,
  1977, 2.24,
  1978, 2.09,
  1979, 1.88,
  1980, 1.65,
  1981, 1.5,
  1982, 1.41,
  1983, 1.37,
  1984, 1.31,
  1985, 1.27
)
```

```
rand_person_spend <- left_join(rand_person_spend,
                               cpi, by = c("expyear" = "year")) %>%
  mutate(out_inf = outsum * cpi,
         inpdol_inf = inpdol * cpi)
```

Add a total spending variable.

```
rand_person_spend <- mutate(rand_person_spend,
                            tot_inf = inpdol_inf + out_inf)
```

Add a variable for any health insurance (free, Individual deductible, or cost-sharing):

```
rand_person_spend <- mutate(rand_person_spend,
                                 any_ins = plantype != "Catastrophic")
```

Count the number of observations in each plan-type,

```
count(rand_person_spend, plantype)
#> # A tibble: 4 x 2
#>   plantype         n
#>   <fct>        <int>
#> 1 Catastrophic  3724
#> 2 Deductible    4175
#> 3 Cost Sharing  5464
#> 4 Free          6840
```

and any-insurance,

```
count(rand_person_spend, any_ins)
#> # A tibble: 2 x 2
#>   any_ins     n
#>   <lgl>   <int>
#> 1 FALSE    3724
#> 2 TRUE    16479
```

Create a list of response variables.

```
varlist <- c("ftf", "out_inf", "totadm", "inpdol_inf", "tot_inf")
```

Calculate the mean and standard deviation for those receiving catastrophic insurance.

```
rand_person_spend %>%
  filter(plantype == "Catastrophic") %>%
  select(one_of(varlist)) %>%
  gather(response, value) %>%
  group_by(response) %>%
  summarise(Mean = mean(value, na.rm = TRUE),
            `Std. Dev.` = sd(value, na.rm = TRUE))
#> # A tibble: 5 x 3
#>   response       Mean `Std. Dev.`
#>   <chr>         <dbl>       <dbl>
#> 1 ftf          2.78         5.50
#> 2 inpdol_inf 388.        2308.
#> 3 out_inf    248.         488.
#> 4 tot_inf    636.        2535.
#> 5 totadm       0.0991       0.379
```

Calculate the difference in means between plans and the catastophic plan.

```
calc_diffs <- function(x) {
  # programmatically create the formula
  f <- quo(!!sym(x) ~ plantype)

  mod <- lm(f, data = rand_person_spend)  # nolint
```

```
  out <- cluster_se(mod, cluster = rand_person_spend[["fam_identifier"]])
  out[["response"]] <- x
  out
}
```

```
person_diffs <- map_dfr(varlist, calc_diffs) %>%
  select(response, term, estimate, std.error) %>%
  mutate(term = str_replace(term, "^plantype", ""))
```

Standard errors are clustered by family identifier using the **clubSandwich** package.

Print the table. If this were an actual publication, I'd make it nicer.

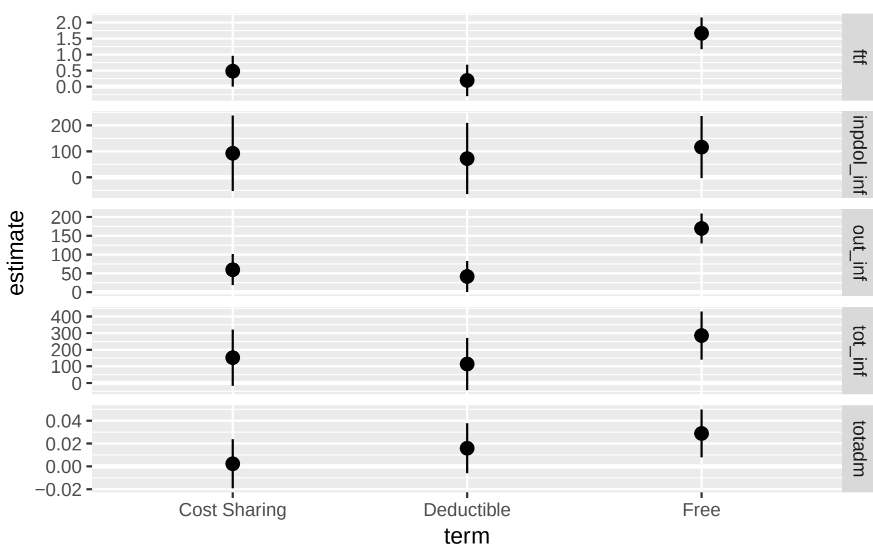
```
fmt_num <- function(x) {
  prettyNum(x, digits = 3, format = "f", big.mark = ",", drop0trailing = FALSE)
}

person_diffs %>%
  mutate(estimate = str_c(fmt_num(estimate), " (", fmt_num(std.error), ")")) %>%
  select(-std.error) %>%
  spread(term, estimate) %>%
  knitr::kable(digits = 3)
```

response	(Intercept)	Cost Sharing	Deductible	Free
ftf	2.78 (0.178)	0.481 (0.24)	0.193 (0.247)	1.66 (0.248)
inpdol_inf	388 (44.9)	92.5 (72.8)	72.2 (68.6)	116 (59.8)
out_inf	248 (14.8)	59.8 (20.7)	41.8 (20.8)	169 (19.9)
tot_inf	636 (54.5)	152 (84.6)	114 (79.1)	285 (72.4)
totadm	0.0991 (0.00785)	0.0023 (0.0108)	0.0159 (0.0109)	0.0288 (0.0105)

Additionally we could plot the difference-in-means of each plan type vs. catastrophic insurance.

```
ggplot(filter(person_diffs, term != "(Intercept)"),
              aes(x = term, y = estimate,
                  ymin = estimate - 2 * std.error,
                  ymax = estimate + 2 * std.error)) +
  geom_hline(yintercept = 0, colour = "white", size = 1) +
  geom_pointrange() +
  facet_grid(response ~ ., scales = "free_y")
```



References

- https://www.icpsr.umich.edu/icpsrweb/NACDA/studies/6439/version/1
- http://masteringmetrics.com/wp-content/uploads/2015/01/ReadMe_RAND.txt
- http://masteringmetrics.com/wp-content/uploads/2015/01/Code.zip

Part II

Chapter 3

Chapter 3

Minneapolis Domestic Violence Experiment

This replicates Table 3.3 of *Mastering 'Metrics*, which replicates the Minneapolis Domestic Violence Experiment (Sherman and Berk 1984,J. D. Angrist (2006)).

Load necessary packages.

```
library("tidyverse")
```

Load the MDVE data.

```
data("mdve", package = "masteringmetrics")
```

Randomized assignments (i.e. what are police assigned to do) are in the `assigned` column. Actual outcomes (i.e. what action do the police actually take) is in the `outcome` column. gen outcome = "Arrest" if T_FINAL == 1 replace outcome = "Advise" if T_FINAL == 2 replace outcome = "Separate" if T_FINAL == 3 replace outcome = "Other" if T_FINAL == 4 gen total = 1

```
mdve <- mutate(mdve,
               assigned = case_when(
      T_RANDOM == 1 ~ "Arrest",
      T_RANDOM == 2 ~ "Advise",
      T_RANDOM == 3 ~ "Separate"
    ),
      outcome = case_when(
        T_FINAL == 1 ~ "Arrest",
        T_FINAL == 2 ~ "Advise",
        T_FINAL == 3 ~ "Separate",
        T_FINAL == 4 ~ "Other"
      ),
      coddled_a = assigned != "Arrest",
      coddled_o = outcome != "Arrest"
    ) %>%
  filter(outcome != "Other")
```

Assigned and delivered treatments in the MDVE:

```
mdve_summary <-
  mdve %>%
  count(assigned, outcome) %>%
  group_by(assigned) %>%
```

```
  mutate(p = n / sum(n))
print(mdve_summary, n = nrow(mdve_summary))
#> # A tibble: 8 x 4
#> # Groups:   assigned [3]
#>   assigned outcome      n      p
#>   <chr>    <chr>    <int>  <dbl>
#> 1 Advise   Advise      84 0.778
#> 2 Advise   Arrest      19 0.176
#> 3 Advise   Separate     5 0.0463
#> 4 Arrest   Arrest      91 0.989
#> 5 Arrest   Separate     1 0.0109
#> 6 Separate Advise       5 0.0439
#> 7 Separate Arrest      26 0.228
#> 8 Separate Separate    83 0.728
```

Assigned proportions in the MDVE:

```
mdve_assigned <- mdve %>%
  count(assigned) %>%
  mutate(p = n / sum(n))
mdve_assigned
#> # A tibble: 3 x 3
#>   assigned     n     p
#>   <chr>    <int> <dbl>
#> 1 Advise     108 0.344
#> 2 Arrest      92 0.293
#> 3 Separate   114 0.363
```

Delivered treatments in the MDVE:

```
mdve_outcome <- mdve %>%
  count(outcome) %>%
  mutate(p = n / sum(n))
mdve_outcome
#> # A tibble: 3 x 3
#>   outcome      n     p
#>   <chr>    <int> <dbl>
#> 1 Advise      89 0.283
#> 2 Arrest     136 0.433
#> 3 Separate    89 0.283
```

Probability of being coddled, given being assigned the coddled treatment:

```
mdve_coddled <- mdve %>%
  count(coddled_a, coddled_o) %>%
  group_by(coddled_a) %>%
  mutate(p = n / sum(n))
mdve_coddled
#> # A tibble: 4 x 4
#> # Groups:   coddled_a [2]
#>   coddled_a coddled_o     n      p
#>   <lgl>     <lgl>     <int>  <dbl>
#> 1 FALSE     FALSE        91 0.989
#> 2 FALSE     TRUE          1 0.0109
#> 3 TRUE      FALSE        45 0.203
#> 4 TRUE      TRUE        177 0.797
```

IV first stage,

$$E[D_i|Z_i = 1] - E[D_i|Z_i = 0].$$

```
filter(mdve_coddled, coddled_o, coddled_a)$p -
  filter(mdve_coddled, coddled_o, !coddled_a)$p
#> [1] 0.786
```

The response variable is not provided, so the full 2SLS is not estimated here.

3.1 References

- http://masteringmetrics.com/wp-content/uploads/2015/02/MDVE_Table33.do
- http://masteringmetrics.com/wp-content/uploads/2015/02/ReadMe_MDVE.txt

Part III

Chapter 4

Chapter 4

MLDA Regression Discontinuity

MLDA Regression Discontinuity (based on data from Carpenter and Dobkin (2011)) from Chapter 4 of *Mastering 'Metrics*, Table 4.1 and Figures 4.2, 4.4, and 4.5 in Mastering Metrics. These present sharp RD estimates of the effect of the minimum legal drinking age (MLDA) on mortality.

Load libraries.

```
library("tidyverse")
library("haven")
library("rlang")
library("broom")
library("lmtest")
library("sandwich")
```

Load MLDA data

```
data("mlda", package = "masteringmetrics")
```

Add an indicator variable for individuals over 21 years of age.

```
mlda <- mutate(mlda,
               age = agecell - 21,
               over21 = as.integer(agecell >= 21))
```

Add a variable for other causes of death.

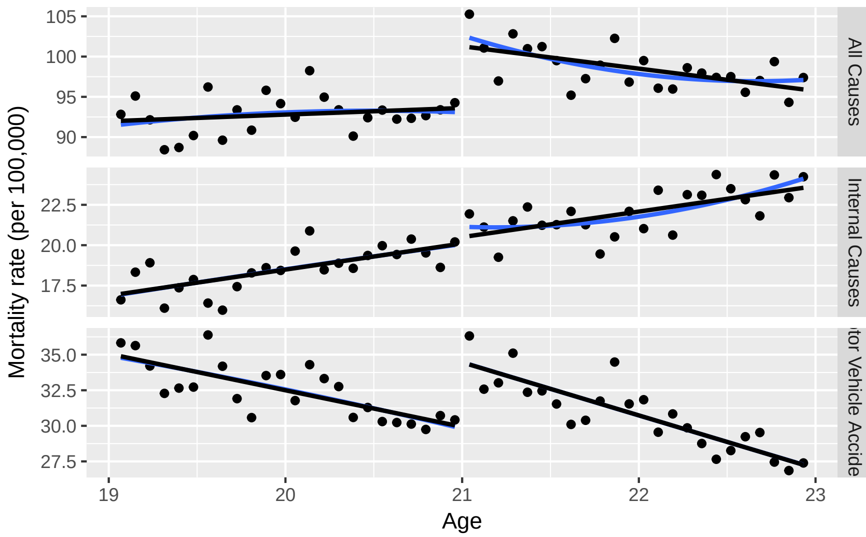
```
mlda <- mutate(mlda, ext_oth = external - homicide - suicide - mva)
```

For "all causes", "motor vehicle accidents", and "internal causes" deaths plot the linear and quadratic trends on each side of age 21.

```
varlist <- c("all" = "All Causes",
             "mva" = "Motor Vehicle Accidents",
             "internal" = "Internal Causes")
```

```
mlda %>%
  select(agecell, over21, one_of(names(varlist))) %>%
  gather(response, value, -agecell, -over21, na.rm = TRUE) %>%
  mutate(response = recode(response, !!!as.list(varlist))) %>%
  ggplot(aes(x = agecell, y = value)) +
  geom_point() +
  geom_smooth(mapping = aes(group = over21), se = FALSE, method = "lm",
              formula = y ~ poly(x, 2)) +
```

```
  geom_smooth(mapping = aes(group = over21), se = FALSE, method = "lm",
              formula = y ~ x, color = "black") +
  facet_grid(response ~ ., scales = "free_y") +
  labs(y = "Mortality rate (per 100,000)", x = "Age")
```




```
responses <- c("all" = "All deaths",
               "mva" = "Motor vehicle accidents",
               "suicide" = "Suicide",
               "homicide" = "Homocide",
               "ext_oth" = "Other external causes",
               "internal" = "All internal causes",
               "alcohol" = "Alcohol")
```

Define a function to run four regressions for a given response variable, y.

```
run_reg <- function(y) {
  mods <- list(
    "Ages 19-22, Linear" =
      lm(quo(!!sym(y) ~ age * over21), data = mlda),
    "Ages 19-22, Quadratic" =
      lm(quo(!!sym(y) ~ poly(age, 2, raw = TRUE) * over21), data = mlda),
    "Ages 20-21, Linear" =
      lm(quo(!!sym(y) ~ age * over21),
             data = filter(mlda, agecell >= 20, agecell <= 22)),
    "Ages 20-21, Quadratic" =
      lm(quo(!!sym(y) ~ poly(age, 2, raw = TRUE) * over21),
             data = filter(mlda, agecell >= 20, agecell <= 22))
  )
  out <- tibble(
    model_name = names(mods),
    model = mods,
    ages = rep(c("19-22", "20-21"), each = 2),
    trend = rep(c("Linear", "Quadratic"), 2),
    model_num = seq_along(mods)
  ) %>%
    mutate(coefs = map(model, ~ tidy(coeftest(.x, vcovHC(.x))))) %>% # nolint
    unnest(coefs, .drop = FALSE) %>%
```

```
    filter(term == "over21") %>%
    select(model_name, model, term, estimate, std.error) %>%
    mutate(response = y)
  # sample size = df.residuals + residuals
  out[["obs"]] <- map_dfr(mods, glance) %>%
    mutate(obs = df.residual + df) %>%
    pluck("obs")
  out
}

mlda_regs <- map_dfr(names(responses), run_reg) %>%
  mutate(response = recode(response, !!!as.list(responses)))
```

```
mlda_regs %>%
  select(model_name, response, estimate, std.error) %>%
  gather(stat, value, estimate, std.error) %>%
  spread(model_name, value) %>%
  knitr::kable()
```

response	stat	Ages 19-22, Linear	Ages 19-22, Quadratic	Ages 20-21, Linear	Ages 20-21, Quac
Alcohol	estimate	0.442	0.799	0.740	
Alcohol	std.error	0.213	0.431	0.360	
All deaths	estimate	7.663	9.548	9.753	
All deaths	std.error	1.374	2.231	2.279	
All internal causes	estimate	0.392	1.073	1.692	
All internal causes	std.error	0.592	0.931	0.877	
Homocide	estimate	0.104	0.200	0.164	-
Homocide	std.error	0.394	0.604	0.590	
Motor vehicle accidents	estimate	4.534	4.663	4.759	
Motor vehicle accidents	std.error	0.731	1.366	1.385	
Other external causes	estimate	0.838	1.797	1.414	
Other external causes	std.error	0.413	0.673	0.606	
Suicide	estimate	1.794	1.814	1.724	
Suicide	std.error	0.530	0.950	0.881	

The robust standard errors using the HC3 standard errors from `sandwich::vcovHC` and differ from those reported in *Mastering 'Metrics*.

4.1 References

- http://masteringmetrics.com/wp-content/uploads/2015/01/master_cd_rd.do
- http://masteringmetrics.com/wp-content/uploads/2015/01/ReadMe_MLDA.txt

Part IV

Chapter 5

Chapter 5

Mississippi Bank Failures in the Great Depression

A difference-in-difference analysis of Mississippi bank failures during the Great Depression (Richardson and Troost 2009). This replicates Figures 5.1–5.3 in *Mastering 'Metrics*.

```
library("tidyverse")
library("lubridate")
```

Load the `banks` data.

```
data("banks", package = "masteringmetrics")
```

Only use yearly data in the difference-in-difference estimates. Use the number of banks on July 1st of each year.

```
banks <- banks %>%
  filter(month(date) == 7L, mday(date) == 1L) %>%
  mutate(year = year(date)) %>%
  select(year, matches("bi[ob][68]"))
```

Generate the counterfactual using the difference between the number of banks in district 8 and district 6.

```
banks <- banks %>%
  arrange(year) %>%
  mutate(diff86 = bib8[year == 1930] - bib6[year == 1930],
         counterfactual = if_else(year >= 1930, bib8 - diff86, NA_integer_)) %>%
  select(-diff86)
```

Plot the lines of the Distinct 8 banks in business, District 6 banks in business, and the District 6 counterfactual. This is equivalent to Figure 5.3 of J. D. Angrist and Pischke (2014).

```
select(banks, year, bib8, bib6, counterfactual) %>%
  gather(variable, value, -year, na.rm = TRUE) %>%
  mutate(variable = recode(variable, bib8 = "8th district",
                           bib6 = "6th district",
                           counterfactual = "Counterfactual")) %>%
  ggplot(aes(x = year, y = value, colour = variable)) +
  geom_point() +
  geom_line() +
  ylab("Number of Banks in Business") +
```

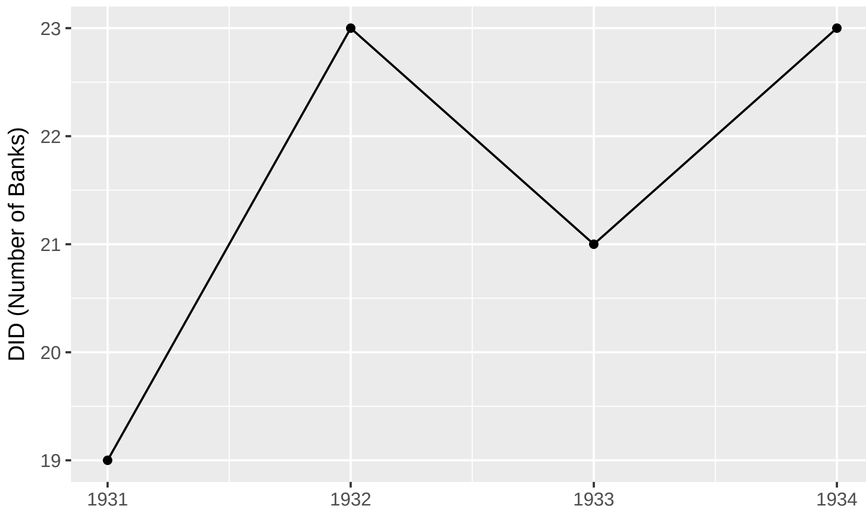
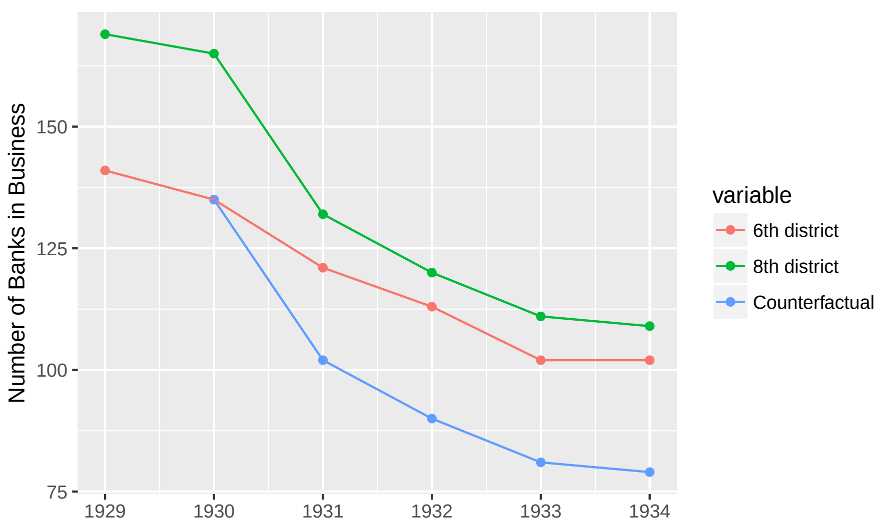


Figure 5.1: Difference between Eighth District and Sixth District Counterfactuals

```
  xlab("")
```



Plot the difference-in-difference estimate for all years after 1930.

```
ggplot(filter(banks, year > 1930), aes(x = year, y = bib6 - counterfactual)) +
  geom_point() +
  geom_line() +
  ylab("DID (Number of Banks)") +
  xlab("")
```

5.1 References

- http://masteringmetrics.com/wp-content/uploads/2015/02/master__banks.do
- http://masteringmetrics.com/wp-content/uploads/2015/02/ReadMe__BankFailures.txt

Chapter 6

MLDA Difference-in-Difference

Difference-in-difference estimates of the effect of the minimum legal drinking age (MLDA) on mortality (Mouchel, Williams, and Zador 1987; Norberg, Bierut, and Grucza 2009). This replicates the analyses in Tables 5.2 and 5.3 in *Mastering 'Metrics*.

Load necessary libraries.

```
library("tidyverse")
library("haven")
library("rlang")
library("broom")
library("clubSandwich")
```

```
data("deaths", package = "masteringmetrics")
```

In these regressions, we will use both indicator variables for year as well as a trend, so make a factor version of the `year` variable.

```
deaths <- mutate(deaths, year_fct = factor(year))
```

6.1 Table 5.2

Regression DD Estimates of MLDA-Induced Deaths among 18-20 year-olds, from 1970-1983

```
dtypes <- c("all" = "All deaths",
            "MVA" = "Motor vehicle accidents",
            "suicide" = "Suicide",
            "internal" = "All internal causes")
```

Estimate the DD for MLDA for all causes of death in 18-20 year olds. Run the regression with `lm` and calculate the cluster robust standard errors using `sandwich::vcovCL`. Subset the data.

```
data <- filter(deaths, year <= 1983, agegr == "18-20 yrs", dtype == "all")
```

Run the OLS model.

```
mod <- lm(mrate ~ 0 + legal + state + year_fct, data = data)
```

Calculate cluster robust coefficients. These are calculated using a different method than Stata uses, and thus will be slightly different than those reported in the book.

```
vcov <- vcovCR(mod, cluster = data[["state"]],
               type = "CR2")
coef_test(mod, vcov = vcov) %>%
  rownames_to_column(var = "term") %>%
  as_tibble() %>%
  select(term, estimate = beta, std.error = SE) %>%
  filter(term == "legal") %>%
  knitr::kable(digits = 2)
```

term	estimate	std.error
legal	10.8	4.48

Function to calculate clustered standard errors and return a tidy data frame of the coefficients and standard errors.

```
cluster_se <- function(mod, cluster, type = "CR2") {
  vcov <- vcovCR(mod, cluster = cluster, type = "CR2")
  coef_test(mod, vcov = vcov) %>%
    rownames_to_column(var = "term") %>%
    as_tibble() %>%
    select(term, estimate = beta, std.error = SE)
}
```

```
run_mlda_dd <- function(i) {
  data <- filter(deaths, year <= 1983, agegr == "18-20 yrs", dtype == i) # nolint
  mods <- tribble(
    ~ name, ~ model,
    "No trends, no weights",
    lm(mrate ~ 0 + legal + state + year_fct, data = data),
    "Time trends, no weights",
    lm(mrate ~ 0 + legal + year_fct + state + state:year, data = data),
    "No trends, weights",
    lm(mrate ~ 0 + legal + year_fct + state, data = data, weights = pop),
    # nolint start
    # "Time trends, weights",
    #   lm(mrate ~ 0 + legal + year_fct + state + state:year,
    #      data = data, weights = pop)
    # nolint end
  ) %>%
    mutate(coefs = map(model, ~ cluster_se(.x, cluster = data[["state"]],
                                           type = "CR2"))) %>%
    unnest(coefs) %>%
    filter(term == "legal") %>%
    mutate(response = i) %>%
    select(name, response, estimate, std.error)
}
```

```
mlda_dd <- map_df(names(dtypes), run_mlda_dd)
```

```
mlda_dd %>%
  knitr::kable(digits = 2)
```

name	response	estimate	std.error
No trends, no weights	all	10.80	4.48
Time trends, no weights	all	8.47	4.74
No trends, weights	all	12.41	4.78
No trends, no weights	MVA	7.59	2.43
Time trends, no weights	MVA	6.64	2.47
No trends, weights	MVA	7.50	2.30
No trends, no weights	suicide	0.59	0.57
Time trends, no weights	suicide	0.47	0.74
No trends, weights	suicide	1.49	0.92
No trends, no weights	internal	1.33	1.53
Time trends, no weights	internal	0.08	1.80
No trends, weights	internal	1.89	1.83

6.2 Table 5.3

Regression DD Estimates of MLDA-Induced Deaths among 18-20 year-olds, from 1970-1983, controlling for Beer Taxes. This is the analysis presented in J. D. Angrist and Pischke (2014) Table 5.3.

```
run_beertax <- function(i) {
  data <- filter(deaths, year <= 1983, agegr == "18-20 yrs",
                 dtype == i, !is.na(beertaxa))
  out <- tribble(
    ~ name, ~ model,
    "No time trends",
    lm(mrate ~ 0 + legal + beertaxa + year_fct + state, data = data),
    "Time trends",
    lm(mrate ~ 0 + legal + beertaxa + year_fct + state + state:year,
       data = data)
  ) %>%
    # calc culstered standard errors
    mutate(coefs = map(model, ~ cluster_se(.x, data[["state"]]))) %>%
    unnest(coefs) %>%
    filter(term %in% c("legal", "beertaxa")) %>%
    mutate(response = i) %>%
    select(response, name, term, estimate, std.error)
}
```

```
beertax <- map_df(names(dtypes), run_beertax)
```

```
beertax %>%
  knitr::kable(digits = 2)
```

response	name	term	estimate	std.error
all	No time trends	legal	10.98	4.60
all	No time trends	beertaxa	1.51	9.02
all	Time trends	legal	10.03	4.57
all	Time trends	beertaxa	-5.52	30.40
MVA	No time trends	legal	7.59	2.51
MVA	No time trends	beertaxa	3.82	5.27
MVA	Time trends	legal	6.89	2.47
MVA	Time trends	beertaxa	26.88	18.76
suicide	No time trends	legal	0.45	0.58
suicide	No time trends	beertaxa	-3.05	1.61
suicide	Time trends	legal	0.38	0.72
suicide	Time trends	beertaxa	-12.13	8.28
internal	No time trends	legal	1.46	1.56
internal	No time trends	beertaxa	-1.36	3.02
internal	Time trends	legal	0.88	1.68
internal	Time trends	beertaxa	-10.31	10.90

Note: I had trouble getting `sandwich::vcovCL` to estimate clustered standard errors for this regression.

6.3 References

- http://masteringmetrics.com/wp-content/uploads/2015/01/analysis.do
- http://masteringmetrics.com/wp-content/uploads/2015/01/ReadMe_MLDA_DD.txt

Part V

Chapter 6

Chapter 7

Twins and Returns to Schooling

Estimates of the returns to schooling for Twinsburg twins (Ashenfelter and Krueger 1994; Ashenfelter and Rouse 1998). This replicates the analysis in Table 6.2 of *Mastering 'Metrics*.

```
library("tidyverse")
library("sandwich")
library("lmtest")
library("AER")
```

Load `twins` data.

```
data("pubtwins", package = "masteringmetrics")
```

Run a regression of log wage on controls (Column 1 of Table 6.2).

```
mod1 <- lm(lwage ~ educ + poly(age, 2) + female + white, data = pubtwins)
coeftest(mod1, vcov = sandwich)
#>
#> t test of coefficients:
#>
#>                Estimate Std. Error t value Pr(>|t|)
#> (Intercept)      1.1791     0.1631    7.23  1.3e-12 ***
#> educ             0.1100     0.0104   10.54  < 2e-16 ***
#> poly(age, 2)1    4.9643     0.5697    8.71  < 2e-16 ***
#> poly(age, 2)2   -4.2957     0.5919   -7.26  1.1e-12 ***
#> female          -0.3180     0.0397   -8.00  5.4e-15 ***
#> white           -0.1001     0.0679   -1.47     0.14
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: The `age` coefficients are different (but equivalent) to those reported in the Table due to the the use of `poly(age, .)`, which calculates orthogonal polynomials.

Run regression of the difference in log wage between twins on the difference in education (Column 2 of Table 6.2).

```
mod2 <- lm(dlwage ~ deduc, data = filter(pubtwins, first == 1))
coeftest(mod2, vcov = sandwich)
#>
#> t test of coefficients:
#>
#>              Estimate Std. Error t value Pr(>|t|)
```

```
#> (Intercept)   0.0296     0.0275    1.07   0.2835
#> deduc         0.0610     0.0198    3.09   0.0022 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Run a regression of log wage on controls, instrumenting education with twin's education (Column 3 of Table 6.2).

```
mod3 <- ivreg(lwage ~ educ + poly(age, 2) + female + white |
                . - educ + educt, data = pubtwins)
summary(mod3, vcov = sandwich, diagnostics = TRUE)
#>
#> Call:
#> ivreg(formula = lwage ~ educ + poly(age, 2) + female + white |
#>     . - educ + educt, data = pubtwins)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -1.69585 -0.29218  0.00494  0.26262  2.47060
#>
#> Coefficients:
#>               Estimate Std. Error t value Pr(>|t|)
#> (Intercept)     1.0636     0.2113    5.03  6.2e-07 ***
#> educ            0.1179     0.0137    8.62  < 2e-16 ***
#> poly(age, 2)1   5.0367     0.5805    8.68  < 2e-16 ***
#> poly(age, 2)2  -4.2897     0.5928   -7.24  1.3e-12 ***
#> female         -0.3149     0.0403   -7.81  2.2e-14 ***
#> white          -0.0974     0.0682   -1.43     0.15
#>
#> Diagnostic tests:
#>                   df1 df2 statistic p-value
#> Weak instruments   1 674    796.30  <2e-16 ***
#> Wu-Hausman         1 673      0.92    0.34
#> Sargan             0  NA        NA      NA
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.507 on 674 degrees of freedom
#> Multiple R-Squared: 0.338,   Adjusted R-squared: 0.333
#> Wald test: 56.8 on 5 and 674 DF,  p-value: <2e-16
```

Note: The coefficient for years of education is slightly different than that reported in the book.

Run a regression of the difference in wage, instrumenting the difference in years of education with twin's education (Column 4 of Table 6.2).

```
mod4 <- ivreg(dlwage ~ deduc | deduct,
              data = filter(pubtwins, first == 1))
summary(mod4, vcov = sandwich, diagnostics = TRUE)
#>
#> Call:
#> ivreg(formula = dlwage ~ deduc | deduct, data = filter(pubtwins,
#>     first == 1))
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
```

```
#> -2.0423 -0.3111 -0.0274  0.2471  2.0824
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   0.0274     0.0277    0.99   0.3237
#> deduc         0.1070     0.0339    3.15   0.0018 **
#>
#> Diagnostic tests:
#>                  df1 df2 statistic p-value
#> Weak instruments   1 338     85.15  <2e-16 ***
#> Wu-Hausman         1 337      4.12   0.043 *
#> Sargan             0  NA        NA      NA
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.512 on 338 degrees of freedom
#> Multiple R-Squared: 0.0132,  Adjusted R-squared: 0.0103
#> Wald test: 9.94 on 1 and 338 DF,  p-value: 0.00176
```

Note: The coefficient for years of education is slightly different than that reported in the book.

References

- http://masteringmetrics.com/wp-content/uploads/2015/02/ReadMe_Twinsburg.txt
- http://masteringmetrics.com/wp-content/uploads/2015/02/twins.do

Chapter 8

Child Labor Laws as an IV

2SLS estimates of the returns to schooling using child labor laws as instruments for years of schooling (Acemoglu and Angrist 2000). This replicates Table 6.3 of *Mastering 'Metrics*.

```
library("AER")
library("sandwich")
library("clubSandwich")
library("tidyverse")
library("broom")
```

Load the `child_labor` data.

```
data("child_labor", package = "masteringmetrics")
child_labor <- mutate(child_labor,
                      year = factor(year),
                      yob_fct = factor(yob),
                      sob = factor(sob))
```

8.1 First stages and reduced forms

Column 1. Years of Schooling.

```
mod1 <- lm(indEduc ~ year + yob_fct + sob + cl7 + cl8 + cl9,
           data = child_labor, weights = weight)
# coef_test(mod1, vcov = vcovCR(mod1, cluster = child_labor[["sob"]]))
```

Column 2. Years of Schooling. State of birth dummies x linear year of birth trends.

```
mod2 <- lm(indEduc ~ year + yob_fct + sob + sob:yob + cl7 + cl8 + cl9,
           data = child_labor, weights = weight)
# coef_test(mod2, vcov = vcovCR(mod2, cluster = child_labor[["sob"]]))
```

Column 3. Log weekly wages.

```
mod3 <- lm(lnwkwage ~ year + yob_fct + sob + cl7 + cl8 + cl9,
           data = child_labor, weights = weight)
# coef_test(mod3, vcov = vcovCR(mod1), cluster = child_labor[["state"]])
```

Column 4. Log weekly wages. State of birth dummies x linear year of birth trends.

```
mod4 <- lm(lnwkwage ~ year + yob_fct + sob + sob:yob + cl7 + cl8 + cl9,
           data = child_labor, weights = weight)
# coef_test(mod4, vcov = vcovCR(mod2), cluster = child_labor[["state"]])
```

8.2 IV returns

Column 3. Log weekly wages.

```
mod5 <- ivreg(lnwkwage ~ year + yob_fct + sob + indEduc |
               . - indEduc + cl7 + cl8 + cl9,
              data = child_labor, weights = weight)
# coef_test(mod5, vcov = vcovCR(mod1), cluster = child_labor[["state"]])
```

Column 4. Log weekly wages. State of birth dummies x linear year of birth trends.

```
mod6 <- ivreg(lnwkwage ~ year + yob_fct + sob + sob:yob + indEduc |
               . - indEduc + cl7 + cl8 + cl9,
              data = child_labor, weights = weight)
# coef_test(mod6, vcov = vcovCR(mod2), cluster = child_labor[["state"]])
```

References

- http://masteringmetrics.com/wp-content/uploads/2015/02/ReadMe_ChildLaborLaws.txt
- http://masteringmetrics.com/wp-content/uploads/2015/02/AA_regs.do

Chapter 9

Quarter of Birth and Returns to Schooling

This replicates Tables 6.4 and 6.5, and Figures 6.1 and 6.2 of *Mastering 'Metrics*. These present an IV analysis of the returns to schooling using quarters of birth (QOB) as instruments for years of schooling (J. D. Angrist and Krueger 1991).

```
library("AER")
library("sandwich")
library("lmtest")
library("tidyverse")
library("broom")
```

Load `twins` data.

```
data("ak91", package = "masteringmetrics")
```

Some cleaning of the data.

```
ak91 <- mutate(ak91,
               qob_fct = factor(qob),
               q4 = as.integer(qob == "4"),
               yob_fct = factor(yob))
```

Table 6.4. IV recipe for returns to schooling using a single QOB instrument. Regress log wages on 4th quarter.

```
mod1 <- lm(lnw ~ q4, data = ak91)
coeftest(mod1, vcov = sandwich)
#>
#> t test of coefficients:
#>
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  5.89827    0.00136 4329.13   <2e-16 ***
#> q4           0.00681    0.00274    2.48    0.013 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regress years of schooling on 4th quarter.

```
mod2 <- lm(s ~ q4, data = ak91)
coeftest(mod2, vcov = sandwich)
```

```
#>
#> t test of coefficients:
#>
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 12.74731    0.00661    1929  < 2e-16 ***
#> q4           0.09212    0.01316       7  2.6e-12 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

IV regression of log wages on years of schooling, with 4th quarter as an instrument for years of schooling.

```
mod3 <- ivreg(lnw ~ s | q4, data = ak91)
coeftest(mod3, vcov = sandwich)
#>
#> t test of coefficients:
#>
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)    4.955      0.358   13.85   <2e-16 ***
#> s              0.074      0.028    2.64   0.0083 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

9.1 Table 6.5.

Regression Estimates of Returns to Schooling using Quarter of Birth Instruments

Column 1. OLS

```
mod4 <- lm(lnw ~ s, data = ak91)
coeftest(mod4, vcov = sandwich)
#>
#> t test of coefficients:
#>
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 4.995182   0.005074     984   <2e-16 ***
#> s           0.070851   0.000381     186   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Column 2. IV with only the 4th quarter as an instrument.

```
mod5 <- ivreg(lnw ~ s | q4, data = ak91)
summary(mod5, vcov = sandwich, diagnostics = TRUE)
#>
#> Call:
#> ivreg(formula = lnw ~ s | q4, data = ak91)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -8.7765 -0.2393  0.0713  0.3326  4.6536
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)    4.955      0.358   13.85   <2e-16 ***
```

```
#> s              0.074      0.028    2.64   0.0083 **
#> 
#> Diagnostic tests:
#>                     df1    df2 statistic p-value    
#> Weak instruments      1 329507     48.99 2.6e-12 ***
#> Wu-Hausman            1 329506      0.01    0.91    
#> Sargan                0     NA        NA      NA    
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> 
#> Residual standard error: 0.638 on 329507 degrees of freedom
#> Multiple R-Squared: 0.117,   Adjusted R-squared: 0.117 
#> Wald test: 6.97 on 1 and 329507 DF,  p-value: 0.00829
```

The argument `diagnostics = TRUE` will run an F-test on the first stage which is reported as the "Weak instruments" diagnostic.

Column 3. OLS. Controls for year of birth.

```
mod6 <- lm(lnw ~ s + yob_fct, data = ak91)
coeftest(mod6, vcov = sandwich)
#> 
#> t test of coefficients:
#> 
#>              Estimate Std. Error t value Pr(>|t|)    
#> (Intercept)  5.017348   0.006019  833.65  < 2e-16 ***
#> s            0.071081   0.000381  186.34  < 2e-16 ***
#> yob_fct1931 -0.006387   0.005123   -1.25  0.21251    
#> yob_fct1932 -0.014838   0.005052   -2.94  0.00331 ** 
#> yob_fct1933 -0.017583   0.005068   -3.47  0.00052 ***
#> yob_fct1934 -0.020999   0.005062   -4.15  3.3e-05 ***
#> yob_fct1935 -0.032895   0.005039   -6.53  6.7e-11 ***
#> yob_fct1936 -0.031781   0.004970   -6.39  1.6e-10 ***
#> yob_fct1937 -0.036712   0.004894   -7.50  6.4e-14 ***
#> yob_fct1938 -0.036890   0.004856   -7.60  3.1e-14 ***
#> yob_fct1939 -0.048164   0.004833   -9.96  < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Column 4. IV reg using only the 4th quarter as an instrument. Controls for year of birth.

```
mod7 <- ivreg(lnw ~ s + yob_fct | q4 + yob_fct, data = ak91)
summary(mod7, vcov = sandwich, diagnostics = TRUE)
#> 
#> Call:
#> ivreg(formula = lnw ~ s + yob_fct | q4 + yob_fct, data = ak91)
#> 
#> Residuals:
#>     Min      1Q  Median      3Q     Max 
#> -8.7785 -0.2346  0.0719  0.3405  4.6687 
#> 
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)    
#> (Intercept)  4.96599    0.35393   14.03   <2e-16 ***
#> s            0.07520    0.02841    2.65   0.0081 ** 
#> yob_fct1931 -0.00696    0.00647   -1.08   0.2819
```

```
#> yob_fct1932 -0.01557    0.00708   -2.20   0.0279 *
#> yob_fct1933 -0.01855    0.00833   -2.23   0.0259 *
#> yob_fct1934 -0.02209    0.00909   -2.43   0.0151 *
#> yob_fct1935 -0.03425    0.01061   -3.23   0.0012 **
#> yob_fct1936 -0.03338    0.01208   -2.76   0.0057 **
#> yob_fct1937 -0.03857    0.01368   -2.82   0.0048 **
#> yob_fct1938 -0.03910    0.01596   -2.45   0.0143 *
#> yob_fct1939 -0.05053    0.01705   -2.96   0.0030 **
#>
#> Diagnostic tests:
#>                     df1    df2 statistic p-value
#> Weak instruments      1 329498     47.73 4.9e-12 ***
#> Wu-Hausman            1 329497      0.02    0.88
#> Sargan                0     NA        NA      NA
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.638 on 329498 degrees of freedom
#> Multiple R-Squared: 0.117,   Adjusted R-squared: 0.117
#> Wald test: 1.81 on 10 and 329498 DF,  p-value: 0.054
```

Column 4. IV reg using all quarters as instruments. Controls for year of birth.

```
mod8 <- ivreg(lnw ~ s + yob_fct | qob_fct + yob_fct, data = ak91)
summary(mod8, vcov = sandwich, diagnostics = TRUE)
#>
#> Call:
#> ivreg(formula = lnw ~ s + yob_fct | qob_fct + yob_fct, data = ak91)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -8.9945 -0.2544  0.0676  0.3509  4.8425
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  4.59174    0.25057   18.32  < 2e-16 ***
#> s            0.10525    0.02012    5.23  1.7e-07 ***
#> yob_fct1931 -0.01111    0.00591   -1.88  0.05988 .
#> yob_fct1932 -0.02089    0.00623   -3.35  0.00080 ***
#> yob_fct1933 -0.02556    0.00698   -3.66  0.00025 ***
#> yob_fct1934 -0.03007    0.00742   -4.05  5.1e-05 ***
#> yob_fct1935 -0.04414    0.00836   -5.28  1.3e-07 ***
#> yob_fct1936 -0.04501    0.00930   -4.84  1.3e-06 ***
#> yob_fct1937 -0.05207    0.01034   -5.04  4.7e-07 ***
#> yob_fct1938 -0.05518    0.01184   -4.66  3.1e-06 ***
#> yob_fct1939 -0.06780    0.01259   -5.39  7.2e-08 ***
#>
#> Diagnostic tests:
#>                     df1    df2 statistic p-value
#> Weak instruments      3 329496     32.32  <2e-16 ***
#> Wu-Hausman            1 329497      2.98   0.084 .
#> Sargan                2     NA      3.26   0.196
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#>
#> Residual standard error: 0.647 on 329498 degrees of freedom
#> Multiple R-Squared: 0.0905,  Adjusted R-squared: 0.0905
#> Wald test: 3.79 on 10 and 329498 DF,  p-value: 3.9e-05
```

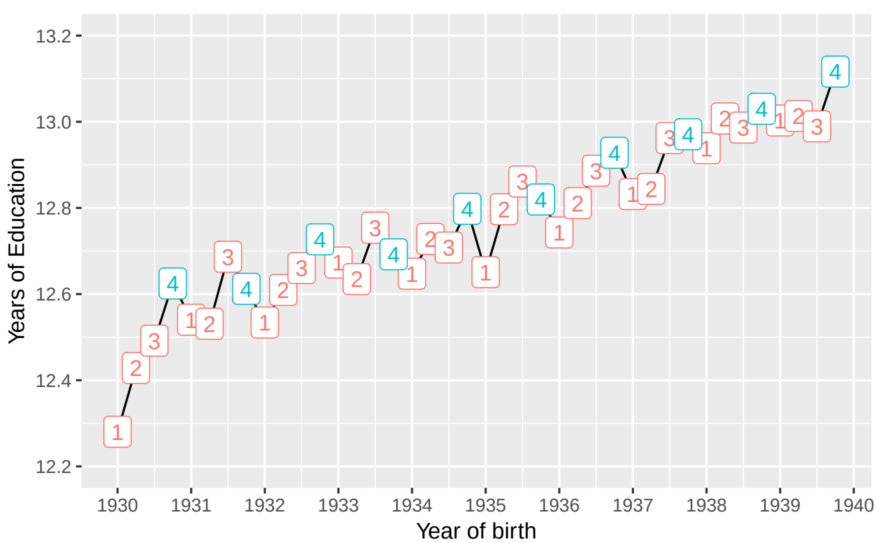
9.2 Figures

Summarize the average wages by age:

```
ak91_age <- ak91 %>%
  group_by(qob, yob) %>%
  summarise(lnw = mean(lnw), s = mean(s)) %>%
  mutate(q4 = (qob == 4))
```

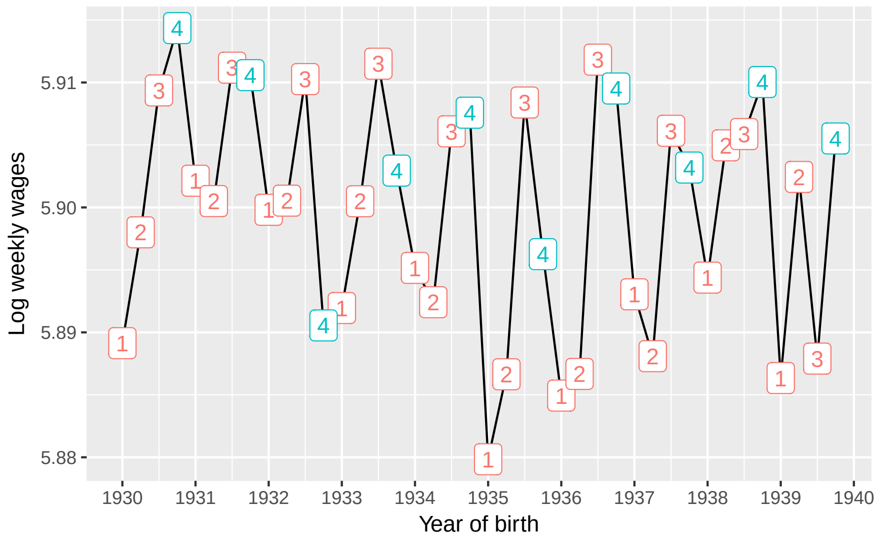
Average years of schooling by quarter of birth for men born in 1930-39 in the 1980 US Census.

```
ggplot(ak91_age, aes(x = yob + (qob - 1) / 4, y = s)) +
  geom_line() +
  geom_label(mapping = aes(label = qob, color = q4)) +
  theme(legend.position = "none") +
  scale_x_continuous("Year of birth", breaks = 1930:1940) +
  scale_y_continuous("Years of Education", breaks = seq(12.2, 13.2, by = 0.2), limits = c(12.2, 13.2))
```



Average log wages by quarter of birth for men born in 1930-39 in the 1980 US Census.

```
ggplot(ak91_age, aes(x = yob + (qob - 1) / 4, y = lnw)) +
  geom_line() +
  geom_label(mapping = aes(label = qob, color = q4)) +
  scale_x_continuous("Year of birth", breaks = 1930:1940) +
  scale_y_continuous("Log weekly wages") +
  theme(legend.position = "none")
```



References

- http://masteringmetrics.com/wp-content/uploads/2015/02/ReadMe_QOB.txt
- http://masteringmetrics.com/wp-content/uploads/2015/02/ak91.do

Chapter 10

Sheepskin and Returns to Schooling

This replicates Figures 6.3 and 6.4 of *Mastering 'Metrics*. These analyses use a fuzzy RD design to analyze the "sheepskin effects" of a high school diploma (Clark and Martorell 2014).

```
library("tidyverse")
```

Load `sheepskin` data.

```
data("sheepskin", package = "masteringmetrics")
```

Create indicator variable for passing the test.

```
sheepskin <- mutate(sheepskin, test_lcs_pass = (minscore >= 0))
```

10.1 Figure 1

Figure 1. Regression discontinuity

```
mod1_lhs <- lm(receivehsd ~ poly(minscore, 4),
               data = filter(sheepskin, minscore < 0), weights = n)
mod1_rhs <- lm(receivehsd ~ poly(minscore, 4),
               data = filter(sheepskin, minscore >= 0), weights = n)
```

Append fitted values to the original dataset

```
fig1_data <- sheepskin %>%
  select(minscore, receivehsd, n) %>%
  modelr::add_predictions(mod1_lhs, var = "fit_hsd2_l") %>%
  mutate(fit_hsd2_l = if_else(minscore > 0, NA_real_, fit_hsd2_l)) %>%
  modelr::add_predictions(mod1_rhs, var = "fit_hsd2_r") %>%
  mutate(fit_hsd2_r = if_else(minscore < 0, NA_real_, fit_hsd2_r))
```

Figure 6.3.

```
ggplot(fig1_data, aes(x = minscore)) +
  geom_vline(xintercept = 0, color = "white", size = 2) +
  geom_point(mapping = aes(y = receivehsd), color = "gray") +
  geom_line(mapping = aes(y = fit_hsd2_l)) +
  geom_line(mapping = aes(y = fit_hsd2_r)) +
  scale_x_continuous("Test Scores Relative to Cutoff",
                     breaks = seq(-30, 15, by = 5), limits = c(-30, 15)) +
```

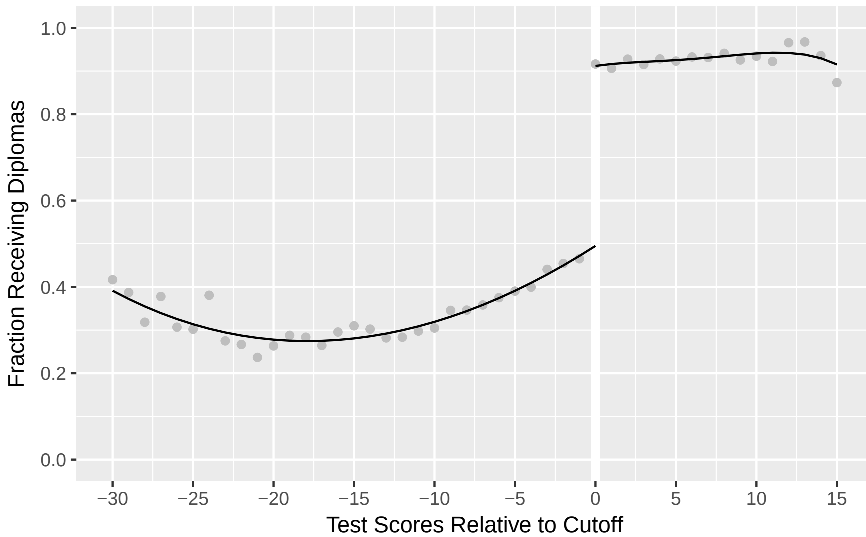


Figure 10.1: Last-chance exams and Texas sheepskin

(#fig:fig.6.3)

```
  scale_y_continuous("Fraction Receiving Diplomas",
                     breaks = seq(0, 1, by = 0.2), limits = c(0, 1))
```

10.2 Figure 2

```
mod2_lhs <- lm(avgearnings ~ poly(minscore, 4),
               data = filter(sheepskin, minscore < 0),
               weights = n)
mod2_rhs <- lm(avgearnings ~ poly(minscore, 4),
               data = filter(sheepskin, minscore >= 0), weights = n)
```

Append fitted values to the original dataset

```
fig2_data <- sheepskin %>%
  select(minscore, avgearnings, n) %>%
  modelr::add_predictions(mod2_lhs, var = "fit_l") %>%
  mutate(fit_l = if_else(minscore > 0, NA_real_, fit_l)) %>%
  modelr::add_predictions(mod2_rhs, var = "fit_r") %>%
  mutate(fit_r = if_else(minscore < 0, NA_real_, fit_r))
```

Figure 6.4.

```
ggplot(fig2_data, aes(x = minscore)) +
  geom_vline(xintercept = 0, color = "white", size = 2) +
  geom_point(mapping = aes(y = avgearnings), color = "gray") +
  geom_line(mapping = aes(y = fit_l)) +
  geom_line(mapping = aes(y = fit_r)) +
  scale_x_continuous("Test Scores Relative to Cutoff",
                     breaks = seq(-30, 15, by = 5), limits = c(-30, 15)) +
  scale_y_continuous("Annual Earnings", breaks = seq(8000, 18000, by = 2000),
                     limits = c(8000, 18000), labels = scales::comma_format())
```

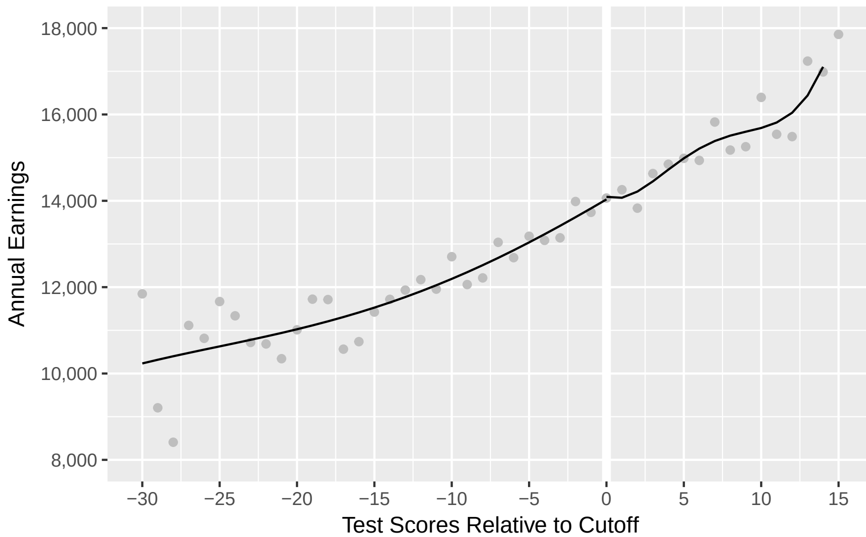


Figure 10.2: The effect of last-chance exam scores on earnings
(#fig:fig.6.4)

References

- http://masteringmetrics.com/wp-content/uploads/2015/02/ReadMe_Sheepskin.txt
- http://masteringmetrics.com/wp-content/uploads/2015/02/cm_graphs.do

References

Acemoglu, Daron, and Joshua Angrist. 2000. "How Large Are Human-Capital Externalities? Evidence from Compulsory Schooling Laws." *NBER Macroeconomics Annual*. doi:10.1086/654403.

Angrist, Joshua D. 2006. "Instrumental Variables Methods in Experimental Criminological Research: What, Why and How." *Journal of Experimental Criminology*. doi:10.1007/s11292-005-5126-x.

Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*. http://www.jstor.org/stable/2937954.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2014. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton UP. https://press.princeton.edu/titles/10363.html.

Aron-Dine, Aviva, Liran Einav, and Amy Finkelstein. 2013. "The RAND Health Insurance Experiment, Three Decades Later." *Journal of Economic Perspectives*. doi:10.1257/jep.27.1.197.

Ashenfelter, Orley, and Alan Krueger. 1994. "Estimates of the Economic Return to Schooling from a New Sample of Twins." *American Economic Review*. http://www.jstor.org/stable/2117766.

Ashenfelter, Orley, and Cecilia Rouse. 1998. "Income, Schooling, and Ability: Evidence from a New Sample of Identical Twins." *Quarterly Journal of Economics*. doi:10.1162/003355398555577.

Brook, Robert H., Jr. John E. Ware, William H. Rogers, Emmett B. Keeler, Allyson R. Davies, Cathy A. Donald, George A. Goldberg, Kathleen N. Lohr, Patricia C. Masthay, and Joseph P. Newhouse. 1983. "Does Free Care Improve Adults' Health? — Results from a Randomized Controlled Trial." *New England Journal of Medicine*. doi:10.1056/NEJM198312083092305.

Carpenter, Christopher, and Carlos Dobkin. 2011. "The Minimum Legal Drinking Age and Public Health." *Journal of Economic Perspectives*. doi:10.1257/jep.25.2.133.

Clark, Damon, and Paco Martorell. 2014. "The Signaling Value of a High School Diploma." *Journal of Political Economy*. doi:10.1086/675238.

Mouchel, William Du, Allan F. Williams, and Paul Zador. 1987. "Raising the Alcohol Purchase Age: Its Effects on Fatal Motor Vehicle Crashes in Twenty-Six States." *Journal of Legal Studies*. http://www.jstor.org/stable/724480.

Norberg, Karen E., Laura J. Bierut, and Richard A. Grucza. 2009. "Long-Term Effects of Minimum Drinking Age Laws on Past-Year Alcohol and Drug Use Disorders." *Alcoholism: Clinical and Experimental Research*. doi:10.1111/j.1530-0277.2009.01056.x.

Richardson, Gary, and William Troost. 2009. "Monetary Intervention Mitigated Banking Panics During the Great Depression: Quasi-Experimental Evidence from a Federal Reserve District Border, 1929–1933." *Journal of Political Economy*. doi:10.1086/649603.

Sherman, Lawrence W., and Richard A. Berk. 1984. "The Specific Deterrent Effects of Arrest for Domestic Assault." *American Sociological Review*. http://www.jstor.org/stable/2095575.