Sofia Guo Econ 142 PSET 8

Sofia Guo

4/8/2019

```
#load libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr)
library(reshape2)
library(ggplot2)
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

1. Construct mean values

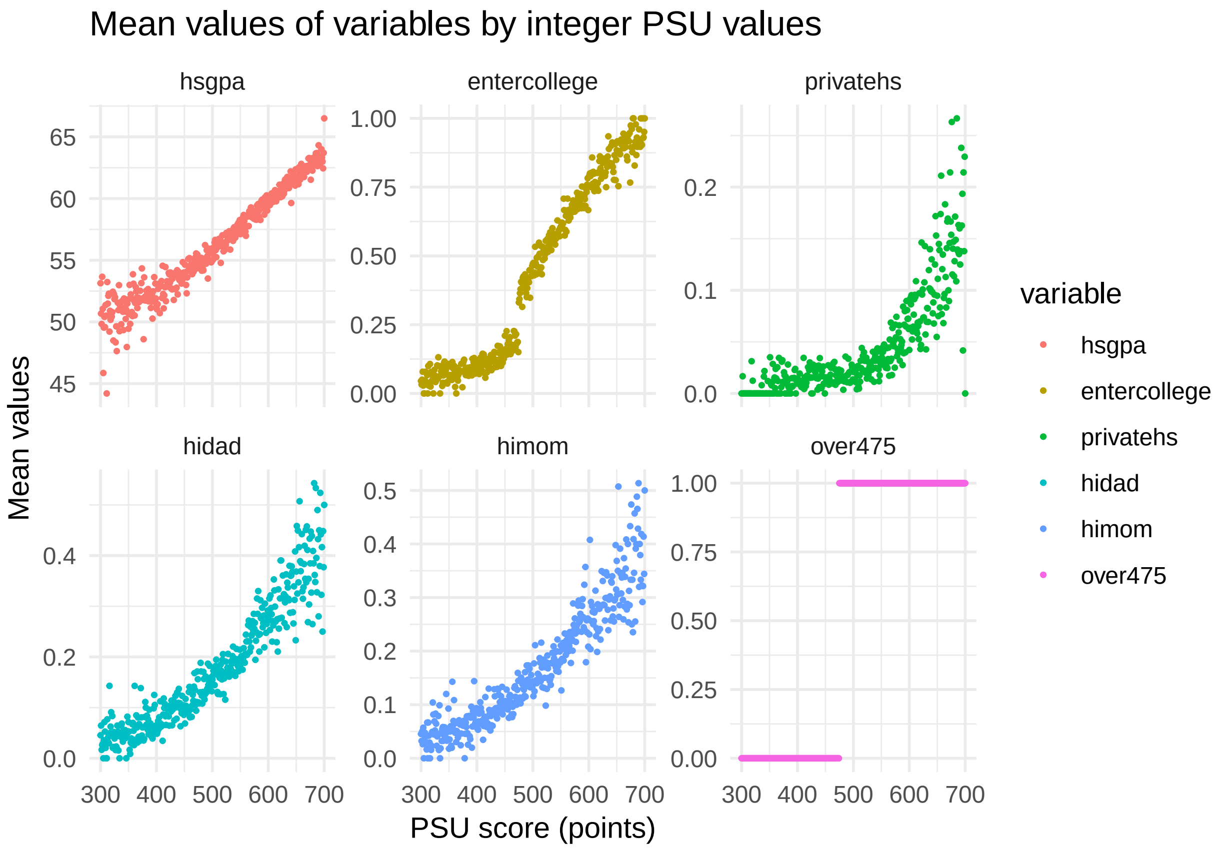
```
#load dataset
rd <- read.csv("/Users/sofia/Box/Cal (sofiaguo@berkeley.edu)/2018-19/Spring 2019/Econ 142/PSETS/PSET 8/

#construct mean values
rd_mean <- rd %>%
  group_by(as.integer(psu)) %>%
  summarize(hsgpa = mean(hsgpa),
entercollege = mean(entercollege),
privatehs = mean(privatehs),
hidad = mean(hidad),
himom = mean(himom),
over475 = mean(over475))

#melt df for graphing
rd_mean_melt <- melt(rd_mean, id.vars = "as.integer(psu)")

#plot the mean values as a function of PSU
ggplot(rd_mean_melt, aes(`as.integer(psu)`, value, group = variable, color = variable)) +
  geom_point(size = 0.5) + facet_wrap(~variable, scales = "free_y") +
  labs(x = 'PSU score (points)', y = 'Mean values',
```

```
    title = 'Mean values of variables by integer PSU values') +
  theme_minimal()
```



We see that there is a "sharp" discontinuity at PSU = 475 for the variable *over475*, but all other variables except *entercollege* are relatively smooth over the running variable.

2. Fit local linear regressions using different bandwidths

Regress one of the outcome variables on the following X's:

1. *constant*
2. *psu*
3. $over475$
4. a 4th variable $= X_4 = (psu - 475) * over475$

Fit the model:

$$y = \beta_1 + \beta_2 psu + \beta_3 over475 + \beta_4 X_4 + \varepsilon$$

so that β_3 measures the jump in y at 475 points, $\beta_2 =$ slope of line to left of 475, $\beta_2 + \beta_4 =$ slope to the right of 475.

2(a). Run regressions with 10 point bandwith

```r
#construct X4
rd_mean_x4 <- rd_mean%>%
  mutate(X4 = (`as.integer(psu)` - 475)*over475)

#specify bandwidth
band = 10
rd_mean_10 <- rd_mean_x4 %>%
  filter(`as.integer(psu)` >= 475 - band & `as.integer(psu)` <= 475 + band -1)

#run regressions
reg_rd_ec_10 <- lm(entercollege ~ `as.integer(psu)` + over475 + X4, data = rd_mean_10)
reg_rd_hg_10 <- lm(hsgpa ~ `as.integer(psu)` + over475 + X4, data = rd_mean_10)
reg_rd_hd_10 <- lm(hidad ~ `as.integer(psu)` + over475 + X4, data = rd_mean_10)
reg_rd_hm_10 <- lm(himom ~ `as.integer(psu)` + over475 + X4, data = rd_mean_10)
```

```r
#display table
stargazer(reg_rd_ec_10,
          reg_rd_hg_10,
          reg_rd_hd_10,
          reg_rd_hm_10,
          type = "latex", title = "10 point bandwidth RD estimates",
          header = F,
          font.size = "small",
          multicolumn = F,
          column.sep.width = '0.1pt',
          single.row = T)
```

Table 1: 10 point bandwidth RD estimates

	Dependent variable:			
	entercollege	hsgpa	hidad	himom
	(1)	(2)	(3)	(4)
‘as.integer(psu)‘	-0.004 (0.003)	0.066 (0.048)	0.002 (0.003)	0.005^{**} (0.002)
over475	0.172^{***} (0.026)	-0.002 (0.394)	-0.024 (0.026)	-0.036^{**} (0.016)
X4	0.012^{**} (0.005)	-0.150^{**} (0.068)	0.002 (0.004)	-0.002 (0.003)
Constant	2.028 (1.506)	23.519 (22.594)	-0.654 (1.490)	-2.273^{**} (0.922)
Observations	20	20	20	20
R^2	0.931	0.236	0.106	0.355
Adjusted R^2	0.918	0.093	-0.062	0.234
Residual Std. Error (df = 16)	0.029	0.437	0.029	0.018
F Statistic (df = 3; 16)	72.343^{***}	1.650	0.629	2.932^{*}
Note:				*p<0.1; **p<0.05; ***p<0.01

2(b). Fit bandwidth of 20

```r
#specify bandwidth
band = 20
rd_mean_20 <- rd_mean_x4 %>%
  filter(`as.integer(psu)` >= 475 - band & `as.integer(psu)` <= 475 + band -1)

#run regressions
```

```r
reg_rd_ec_20 <- lm(entercollege ~ `as.integer(psu)` + over475 + X4, data = rd_mean_20)
reg_rd_hg_20 <- lm(hsgpa ~ `as.integer(psu)` + over475 + X4, data = rd_mean_20)
reg_rd_hd_20 <- lm(hidad ~ `as.integer(psu)` + over475 + X4, data = rd_mean_20)
reg_rd_hm_20 <- lm(himom ~ `as.integer(psu)` + over475 + X4, data = rd_mean_20)
```

```r
#display table
stargazer(reg_rd_ec_20,
          reg_rd_hg_20,
          reg_rd_hd_20,
          reg_rd_hm_20,
          type = "latex", title = "20 point bandwidth RD estimates",
          header = F,
          font.size = "small",
          multicolumn = F,
          column.sep.width = '0.1pt',
          single.row = T)
```

Table 2: 20 point bandwidth RD estimates

	Dependent variable:			
	entercollege	hsgpa	hidad	himom
	(1)	(2)	(3)	(4)
‘as.integer(psu)‘	0.001 (0.001)	0.033 (0.023)	0.002^{**} (0.001)	0.002^{***} (0.001)
over475	0.172^{***} (0.018)	-0.212 (0.369)	-0.017 (0.016)	-0.023^{*} (0.011)
X4	0.002 (0.002)	-0.005 (0.032)	-0.001 (0.001)	0.001 (0.001)
Constant	-0.139 (0.517)	39.429^{***} (10.494)	-0.927^{**} (0.445)	-0.948^{***} (0.324)
Observations	40	40	40	40
R^2	0.936	0.183	0.319	0.587
Adjusted R^2	0.931	0.115	0.263	0.553
Residual Std. Error (df = 36)	0.029	0.583	0.025	0.018
F Statistic (df = 3; 36)	175.392^{***}	2.690^{*}	5.633^{***}	17.069^{***}
Note:				*p<0.1; **p<0.05; ***p<0.01

Compared to part (a) estimates with a bandwidth of 10, I found that β_3 when y is *entercollege* is exactly the same at 0.172 (0.026). However, the other estimates change pretty drastically, especially for β_1 and β_4.

2(c). Fit bandwidths from 5 to 50

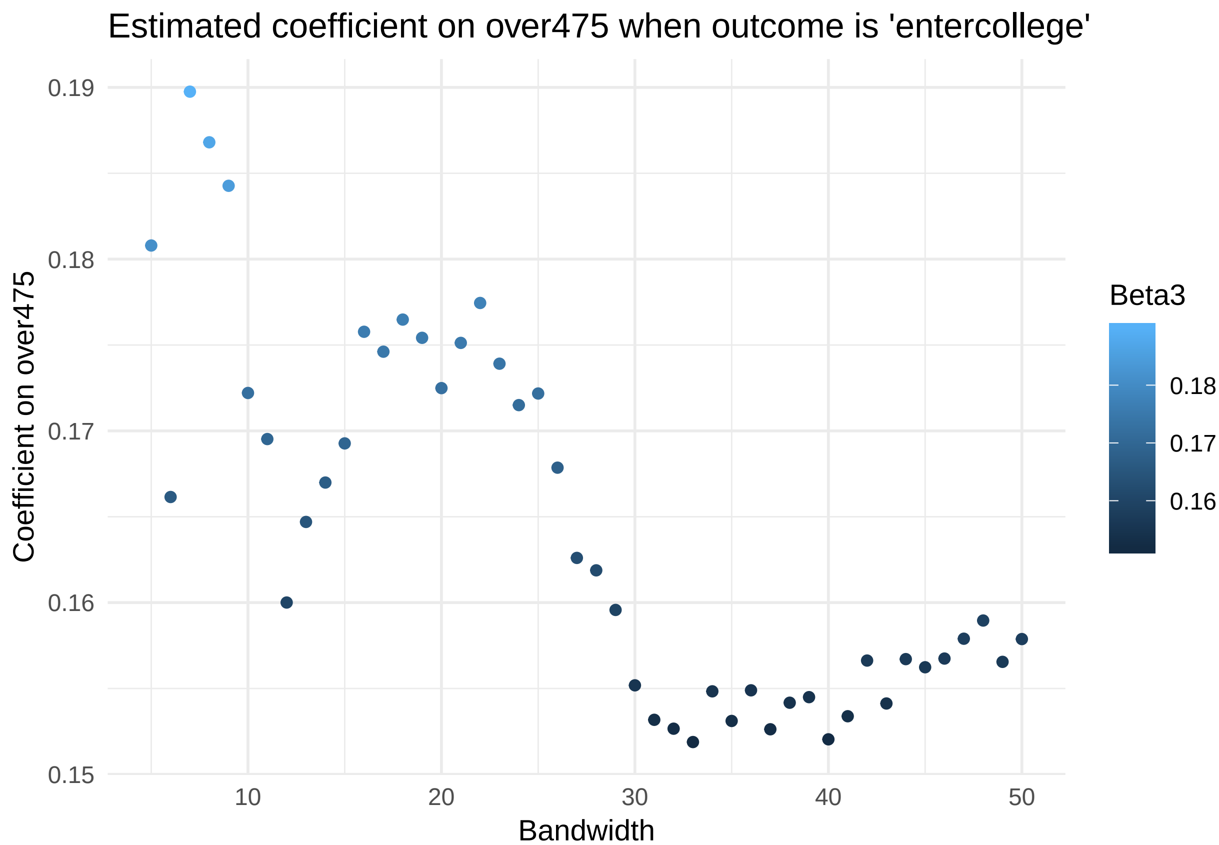
```r
#specify bandwidth

reg_list_5_50 <- list()

for(i in 5:50){
reg_df <- rd_mean_x4 %>%
  filter(`as.integer(psu)` >= 475 - i & `as.integer(psu)` <= 475 + i -1)
#run regressions
reg_list_5_50[[i]] <- coefficients(lm(entercollege ~ `as.integer(psu)` + over475 + X4, data = reg_df))[
}

beta_3_5_50 <- data.frame("bandwidth" = 5:50, "Beta3" = unlist(reg_list_5_50))
```

```r
#graph the results
ggplot(beta_3_5_50, aes(bandwidth, Beta3, color = Beta3)) +
  geom_point() +
  labs(x = "Bandwidth", y = "Coefficient on over475", title = "Estimated coefficient on over475 when ou
  theme_minimal()
```



We see here that the identical coefficient of 0.172 is present when the bandwidth is 10 and 20. As the bandwidth increases, the estimated β_3 generally decreases but at a non-linear rate (and actually increases a little as the bandwidth approaches 50). This wave-like pattern is super interesting and might suggest that the underlying data is non-linear which would cause estimates the fluctuate as the bandwith changes.