LRD paper Appendix C, Data Analysis

lrd authors

10 November, 2017

General dependencies.

```
#print(getwd())
library('knitr')
library(ggplot2)
library(xtable)
library(robustbase)
library(rdd)
if(!require('lrd')){
 source("lrd/R/functions.r")
 source("lrd/R/simulations.r")
 source("lrd/R/displaySim.r")
}
```

Initialization. If the variable `paperdir` is supplied, LaTeX code for the tables is saved there, for inclusion in the main paper; otherwise, the code is saved in the current working directory.

```
if(!exists(paperdir)) paperdir <- '.'

logit=function(x) log(x*.01/(1-x*.01))


ciChar <- function(ci,est=FALSE){
    ci <- round(ci,2)
    ci.out <- paste('(',round(ci[1],2),',',round(ci[2],2),')',sep='')
    if(est) ci.out <- c(ci.out,as.character(ci[3]))
    ci.out
}

round2 <- function(x) round(x,2)

nfunc <- function(bw) sum(abs(dat$R)<bw,na.rm=TRUE)

Wfunc <- function(W)
```

```
    paste0('[',round2(W[1]), ',',round2(W[2]),')')
```

Load data. This routine will download and unzip the Lind et al. replication material into the exdata subdirectory.

```
if(!is.element('dat',ls())){
    if (system.file(package="lrd")!="") {
        extdata_dir <- system.file("extdata", package="lrd")
    } else extdata_dir <- 'extdata'
    LSO_dta_location <- lrd::fetchLSOdata(extdata_dir)
    dat=foreign::read.dta(LSO_dta_location)
                                    #dat=subset(dat,left_school!=1)
    dat$dist_from_cut <- round(dat$dist_from_cut,2)
    dat$hsgrade_pct[dat$hsgrade_pct==100]=99.5
    dat$lhsgrade_pct=logit(dat$hsgrade_pct)
  #dat$age <- dat$age_at_entry>=19
    dat$R <- dat$dist_from_cut
    dat$Z <- dat$gpalscutoff
}
```

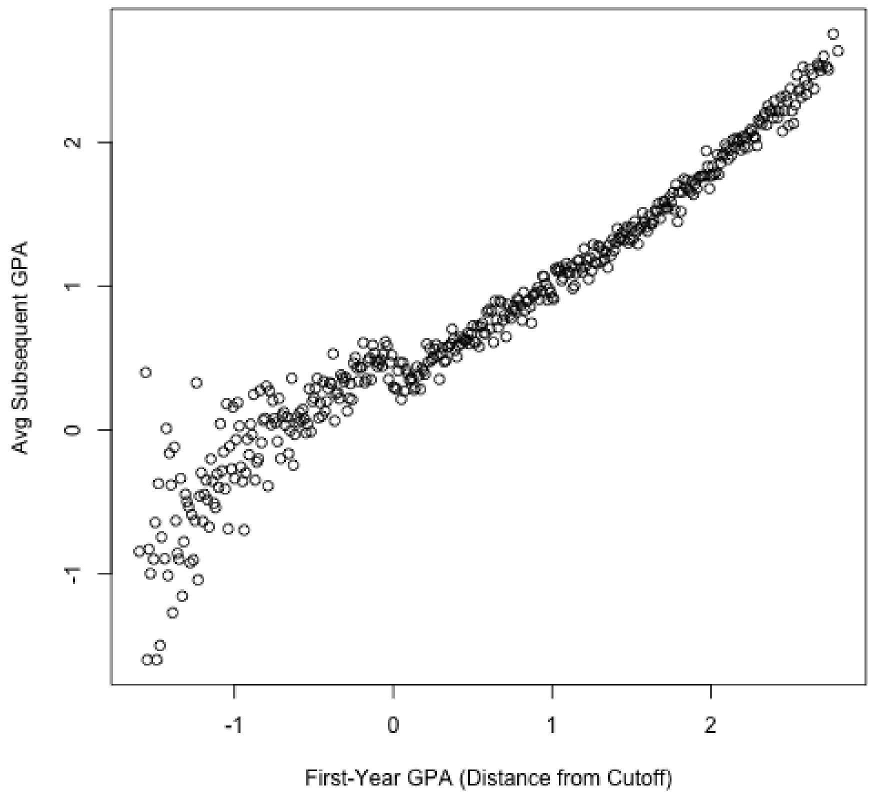
Total sample size, and number of "compliers" (students whose actual AP status matched what would have been predicted from first-year GPA)

```
ncomp <- with(dat,sum(gpalscutoff& !probation_year1))
ntot <- nrow(dat)
```

Create plots for Figure 1. First of the outcome (subsequent GPA):

```
figDat <- aggregate(dat[,c('nextGPA','lhsgrade_pct')],by=list(R=dat$R),
                    FUN=mean,na.rm=TRUE)
figDat$n <- as.vector(table(dat$R))
figDat <- within(figDat,n <- 2*n/max(n))

with(figDat,plot(R,nextGPA,xlab='First-Year GPA (Distance from Cutoff)',
                 ylab='Avg Subsequent GPA'))
```



then a covariate (High-School GPA):

```
with(figDat,plot(R,lhsgrade_pct,xlab='First-Year GPA (Distance from Cutoff)',
                 ylab='Avg logit(hsgrade_pct)'))
```

The McCrary density test failure and recovery described in Section 4.1

```
(mccrary1 <- rdd::DCdensity(dat$R,-0.005, bin=0.01,plot=FALSE) )
```

```
## [1] 0.000668
```

```
( mccraryDougnut <- rdd::DCdensity(dat$R[dat$R!=0],-0.005, bin=0.01,plot=FALSE) )
```

```
## [1] 0.154
```

main analysis

The sh method uses `lmrob`, which in turn requires a random seed. For confidence interval and estimation routines it's helpful to use the same seed throughout, as

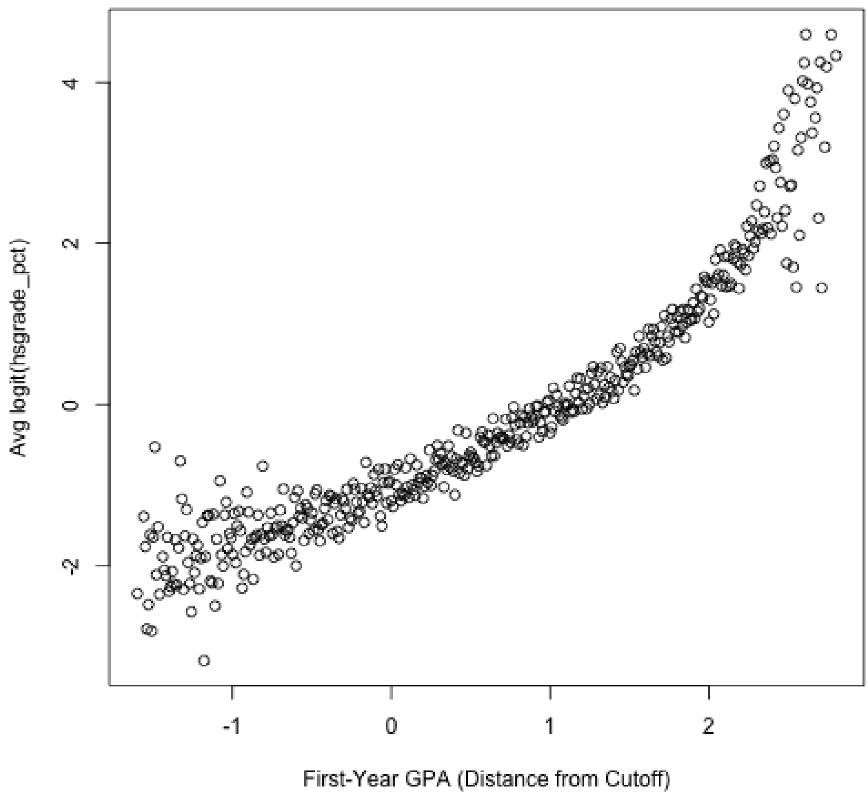


Figure 1: plot of chunk hs_gpaFig

this means the S-estimation initializers will always be sampling the same subsets of the sample.

```
set.seed(201705)
lmrob_seed <- .Random.seed


SHmain <- lrd::sh(subset(dat,R!=0),BW=0.5,outcome='nextGPA',Dvar='probation_year1')
unlist(SHmain)
```

```
##  p.value   CI.CI1   CI.CI2   CI.est       BW bal.pval       W1       W2
## 1.89e-11 1.69e-01 3.08e-01 2.38e-01 5.00e-01 1.00e+00 1.00e-02 5.00e-01
##        n
## 1.00e+04
```

```
# No-donut variant (not discussed in text)
SHnodo <- lrd::sh(dat, BW=0.5, outcome='nextGPA',Dvar='probation_year1')

SHdataDriven <- lrd::sh(dat=subset(dat,R!=0),outcome='nextGPA')
unlist(SHdataDriven)
```

```
##  p.value   CI.CI1   CI.CI2   CI.est       BW bal.pval       W1       W2
## 1.33e-19 1.68e-01 2.61e-01 2.15e-01 1.03e+00 1.54e-01 1.00e-02 1.03e+00
##        n
## 2.16e+04
```

```
SHcubic <- lrd::sh(dat=subset(dat,R!=0),BW=0.5,outcome='nextGPA',rhs='~Z+poly(R,3)')
unlist(SHcubic)
```

```
##  p.value   CI.CI1   CI.CI2   CI.est       BW bal.pval       W1       W2
## 6.33e-07 1.47e-01 3.37e-01 2.42e-01 5.00e-01 1.00e+00 1.00e-02 5.00e-01
##        n
## 1.00e+04
```

```
SHitt <- lrd::sh(dat=subset(dat,R!=0),BW=0.5,outcome='nextGPA', Dvar=NULL)
unlist(SHitt)
```

```
##  p.value   CI.CI1   CI.CI2   CI.est       BW bal.pval       W1       W2
## 1.89e-11 1.68e-01 3.07e-01 2.38e-01 5.00e-01 1.00e+00 1.00e-02 5.00e-01
##        n
## 1.00e+04
```

Create Table 1:

```
resultsTab <-
 do.call('rbind',
  lapply(list(main=SHmain,data_driven=SHdataDriven,cubic=SHcubic,ITT=SHitt),
   function(res) c(round2(res$CI[3]),
                   ciChar(res$CI[1:2]),
                   W=Wfunc(res$W),
                   n=res$n)))
```

```r
colnames(resultsTab) <- c('Estimate','95\\% CI','$\\mathcal{W}$','n')


kable(resultsTab)
```

	Estimate	95% CI	$\mathcal{W}$	n
main	0.24	(0.17,0.31)	[0.01,0.5)	10014
data_driven	0.21	(0.17,0.26)	[0.01,1.03)	21593
cubic	0.24	(0.15,0.34)	[0.01,0.5)	10014
ITT	0.24	(0.17,0.31)	[0.01,0.5)	10014

```r
rownames(resultsTab) <- c('Main','Adaptive $\\mathcal{W}$','Cubic','ITT')

print(xtable(resultsTab),
      file=paste0(paperdir,"/tab-results.tex"), floating=F,
      sanitize.colnames.function=function(x) x,
      sanitize.rownames.function=function(x) x)
```

Results from two alternative methods, creating Table 2:

```r
CFT <- lrd::cft(subset(dat,R!=0),BW=NULL,outcome='nextGPA')
IK <- lrd::ik(subset(dat,R!=0),outcome='nextGPA')

altTab <-
 do.call('rbind',
  lapply(list(Limitless=SHitt,`Local Permutation`=CFT,`Local OLS`=IK),
   function(res) c(round2(res$CI[3]),
    ciChar(res$CI[1:2]),
    W=Wfunc(res$W),
    n=res$n)))



colnames(altTab) <- c('Estimate','95\\% CI','$\\mathcal{W}$','n')

kable(altTab)
```

	Estimate	95% CI	$\mathcal{W}$	n
Limitless	0.24	(0.17,0.31)	[0.01,0.5)	10014
Local Permutation	0.11	(0.05,0.17)	[0.01,0.18)	3436
Local OLS	0.23	(0.19,0.28)	[0.01,1.24)	25841

```r
print(xtable(altTab),
      file=paste0(paperdir,"/tab-alt.tex"), floating=F,
```

```
      sanitize.colnames.function=function(x) x)
```

Examine robustness weights

If there are regions of the data of high influence, the robust fitter should reject or downweight more frequently in those regions, and we'll see dips on the plot of robustness weights vs R.

Here is the plot corresponding to the main analysis presented in the paper.

```
lmrob_main <- lmrob(nextGPA~Z+R,
              offset=(SHmain$CI[3]*probation_year1),
              data=dat,subset=(R!=0 & abs(R)<.5),
              method='MM',
              control=lmrob.control(seed=lmrob_seed,
                           k.max=500, maxit.scale=500)
              )
```

Robustness weights are mostly near 1, never below .25.

```
robwts_main <- weights(lmrob_main, type="robustness")
summary(robwts_main)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.285   0.874   0.959   0.906   0.992   1.000
```

Not too much pattern to the robustness weights – although the lowest values do occur at slightly above the cutpoint, where we'd see savvy students whose rose above the cut due to savvyness.

```
ggp_main <- ggplot(data.frame(R=lmrob_main$model$R,
                              robweights=robwts_main),
                   aes(x=R,y=robweights))
ggp_main + geom_point(alpha=.1) + stat_smooth()
```

```
## `geom_smooth()` using method = 'gam'
```

When we fit without omitting R=0 students, here is the best fitting version of the model.

```
lmrob_nodo <- lmrob(nextGPA~Z+R,
              offset=(SHnodo$CI[3]*probation_year1),
              data=dat,subset=(abs(R)<.5),
              method='MM',
              control=lmrob.control(seed=lmrob_seed,
                                k.max=500, maxit.scale=500)
      )

robwts_nodo <- weights(lmrob_nodo, type="robustness")
```

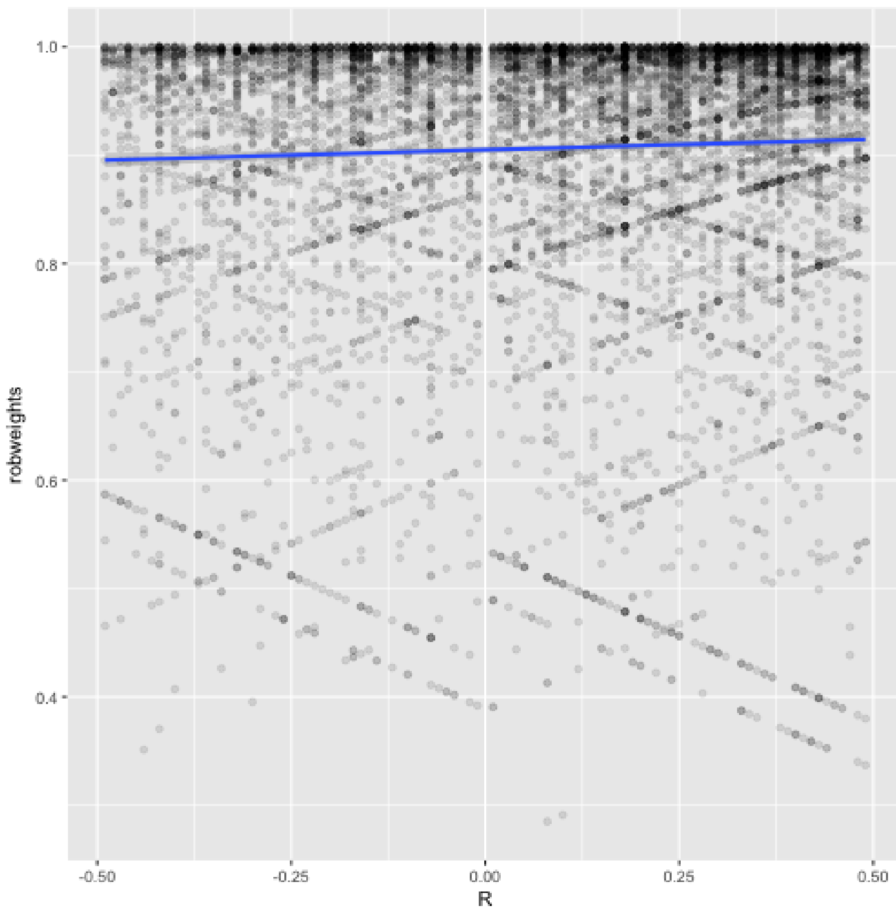


Figure 2: plot of chunk unnamed-chunk-10

Do the observations at R=0 stand out? With no donut, robustness weights have a slight tendency to be lower among observations at R=0.

```
by(robwts_nodo,
   lmrob_nodo$model$R==0, summary)
```

```
## lmrob_nodo$model$R == 0: FALSE
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.287   0.874   0.959   0.907   0.992   1.000
## --------------------------------------------------------
## lmrob_nodo$model$R == 0: TRUE
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.389   0.860   0.954   0.889   0.989   1.000
```

```
t.test(wt~atcut, data.frame(wt=robwts_nodo,
                            atcut=(lmrob_nodo$model$R==0)),
       var.equal=F, alternative="g")
```

```
##
##  Welch Two Sample t-test
##
## data:  wt by atcut
## t = 2, df = 200, p-value = 0.04
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.00151     Inf
## sample estimates:
## mean in group FALSE  mean in group TRUE
##               0.907               0.889
```

The plot is similar to that of the main analysis, with some low robustness weight observations as R=0 but also plenty of ordinary weight observations there.

```
ggp_nodo <- ggplot(data.frame(R=lmrob_nodo$model$R,
                              robweights=robwts_nodo),
                   aes(x=R,y=robweights))
ggp_nodo + geom_point(alpha=.1) + stat_smooth()
```

```
## `geom_smooth()` using method = 'gam'
```

Save results:

```
save(list=ls(),file=paste0('RDanalysis-',format(Sys.time(),"%m%d%H%M"),'.RData'))
```

Session information

```
sessionInfo()
```

```
## R version 3.3.1 (2016-06-21)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X 10.12.6 (Sierra)
##
```

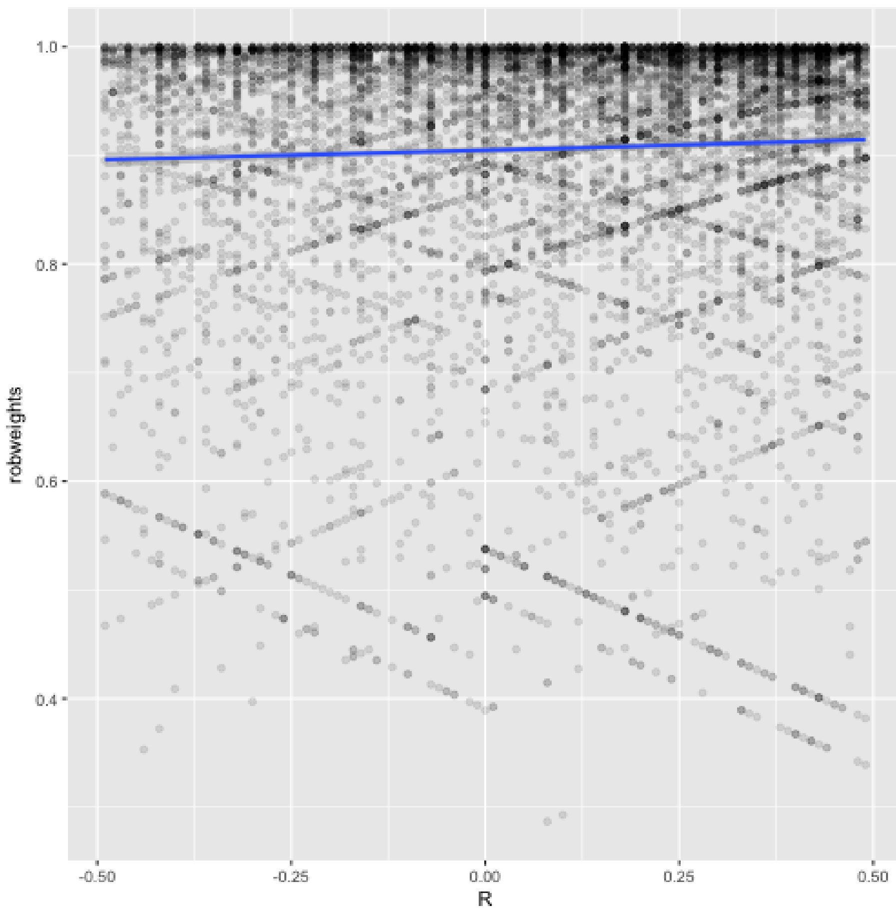


Figure 3: plot of chunk unnamed-chunk-13

```
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lrd_0.0.0.9000    rdd_0.57          Formula_1.2-1
##  [4] AER_1.2-4         survival_2.40-1   car_2.1-4
##  [7] lmtest_0.9-34     zoo_1.7-13        sandwich_2.3-4
## [10] robustbase_0.92-7 xtable_1.8-2      ggplot2_2.2.1
## [13] knitr_1.15.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.8        highr_0.6          nloptr_1.0.4
##  [4] DEoptimR_1.0-8     plyr_1.8.4         tools_3.3.1
##  [7] digest_0.6.10      lme4_1.1-12        evaluate_0.10
## [10] tibble_1.3.4       gtable_0.2.0       nlme_3.1-128
## [13] lattice_0.20-33    mgcv_1.8-15        rlang_0.1.2
## [16] Matrix_1.2-6       parallel_3.3.1     yaml_2.1.13
## [19] SparseM_1.77       stringr_1.1.0      MatrixModels_0.4-1
## [22] rprojroot_1.2      grid_3.3.1         nnet_7.3-12
## [25] foreign_0.8-66     rmarkdown_1.5      minqa_1.2.4
## [28] magrittr_1.5       backports_1.1.1    scales_0.4.1
## [31] htmltools_0.3.5    MASS_7.3-45        splines_3.3.1
## [34] rsconnect_0.5      pbkrtest_0.4-6     colorspace_1.2-6
## [37] labeling_0.3       quantreg_5.29      stringi_1.1.1
## [40] lazyeval_0.2.0     munsell_0.4.3
```