Causal Inference

Due June 8th.

Problem 1

In this problem set we use the data set draft.txt. It contains 10100 observations on four variables, log wages (mean 5.4319, column 1), year of birth (mean 51.5208, column 2), draft eligibility (mean 0.2752, column 3), and veteran status (mean 0.2215, column 4). To conduct the analysis, drop the observations for individuals born in 1950. For the following questions, write code to estimate the quantities of interest "manually" (DO NOT USE the canned functions, unless explicitly suggested.)

(a) Estimate the returns to veteran status by taking the difference in average log earnings for veterans and non-veterans.

```
> data <- read.table("draft.txt",
+                    header=F, sep = "")
>
> names(data)[names(data)=="V1"] <- "logwage"
> names(data)[names(data)=="V2"] <- "ybirth"
> names(data)[names(data)=="V3"] <- "draft"
> names(data)[names(data)=="V4"] <- "veteran"
>
> data <- subset(data, ybirth!=50)
> # the difference is simply
> mean(data$logwage[data$veteran==1])-mean(data$logwage[data$veteran==0])
[1] -0.05117761
```

(b) Estimate the difference through OLS. Calculate both the homoskedasticity based standard errors and the robust (White, heteroskedasticity-consistent) standard errors.

```
> Y <- data$logwage
> X <- cbind(1,data$veteran)
> N <- length(data[,1])
>
> b.ols <- solve(t(X)%*%X)%*%t(X)%*%Y
> e <- Y-X%*%b.ols
> s.sq <- as.numeric(t(e)%*%e)/(N-2)
> var.b.ols <-  solve(t(X)%*%X) * s.sq
> cbind(b.ols,rbind(sqrt(var.b.ols[1,1]),(sqrt(var.b.ols[2,2]))))
            [,1]        [,2]
[1,]  5.43028589 0.008821532
[2,] -0.05117761 0.019839416
>
> # White std.errors
> res.sq <- as.numeric(e)*as.numeric(e)
> Omega <- diag(res.sq)
> HCvar.b.ols <- solve(t(X)%*%X) %*% t(X) %*% Omega %*% X %*% solve(t(X)%*%X)
> cbind(b.ols,rbind(sqrt(HCvar.b.ols[1,1]),(sqrt(HCvar.b.ols[2,2]))))
            [,1]        [,2]
[1,]  5.43028589 0.008662335
[2,] -0.05117761 0.020886264
```

(c) Tabulate the frequencies of treatment status (veteran status) and treatment assignment (draft eligibility).

```
> table("veteran (D)"=data$veteran, "draft (Z)" =data$draft)
           draft (Z)
veteran (D)    0    1
          0 4662 1506
          1  918  602
```

(d) Tabulate the proportion of compliers, never-takers and always-takers under monotonicity. Could you identify compliers individually? Why or why not?

```
> # Never takers
> NT <- 1- mean(data[data$draft==1,]$veteran)
> NT
[1] 0.7144213
> # Always takers
> AT <- mean(data[data$draft==0,]$veteran)
> AT
[1] 0.1645161
> # Compliers
> C <- mean(data[data$draft==1,]$veteran) - mean(data[data$draft==0,]$veteran)
> C
[1] 0.1210626
> # Defiers (by monotonicity)
> D <-0
```

Compliers cannot be identified individually, because it is impossible to distinguish individual compliers in the treatment group from always-takers, and it is impossible to distinguish compliers in the non-treatment group from never-takers.

(e) Next we look at some instrumental variables estimates of the returns to veteran status. Estimate the local average treatment effect and its conventional (homoskedastic) two-stage-least-squares standard error.

```
> # IV
> Z <- cbind(1,data$draft)
> b.iv <- solve(t(Z)%*%X)%*%t(Z)%*%Y
> b.iv
           [,1]
[1,]  5.4643036
[2,] -0.2232354
>
> # 2SLS
> M <- Z%*%solve(t(Z)%*%Z)%*%t(Z)%*%X
> b.2sls <- solve(t(M)%*%M)%*%t(M)%*%Y
> b.2sls
           [,1]
[1,]  5.4643036
[2,] -0.2232354
>
> # s.e.
> r.2sls <- Y-X%*%b.2sls
> s.sq.2sls <- as.numeric(t(r.2sls)%*%r.2sls)/(N-2)
> vcov.2sls <- solve(t(Z)%*%X)%*%t(Z)%*%Z%*%solve(t(X)%*%Z)*drop(s.sq.2sls)
> vcov.2sls
              [,1]         [,2]
[1,]  0.0009079546 -0.004273466
[2,] -0.0042734655  0.021614739
> #equivalently:
> solve( t(M)%*%M ) * drop(s.sq.2sls)
              [,1]         [,2]
[1,]  0.0009079546 -0.004273466
[2,] -0.0042734655  0.021614739
>
> #standard errors:
```

```
> sqrt(diag(vcov.2sls))
[1] 0.03013228 0.14701952
```

(f) **Bonus**: Use the delta method for calculating the robust standard errors. You should begin by re-expressing the local average treatment effect as a function of various components for which you can calculate the component variances.

In order to use the delta method to get robust standard errors, start with the expression for the LATE estimator:

$$
\begin{aligned}
LATE =& \frac{E(Y|Z=1) - E(Y|Z=0)}{E(W|Z=1) - E(W|Z=0)} \\
=& \frac{\frac{E(YZ)}{E(Z)} - \frac{E(Y(1-Z))}{E(1-Z)}}{\frac{E(WZ)}{E(Z)} - \frac{E(W(1-Z))}{E(1-Z)}} \\
=& \frac{E(YZ) - E(Y)E(Z)}{E(WZ) - E(W)E(Z)} = f(E(YZ), E(WZ), E(Y), E(Z), E(W)) = \frac{Cov(Y,Z)}{Cov(W,Z)} = \beta_{IV}
\end{aligned}
$$

We know the asymptotic distribution of the moments:

$$
\sqrt{N}\left[\begin{bmatrix} \bar{YZ} \\ \bar{WZ} \\ \bar{Y} \\ \bar{Z} \\ \bar{W} \end{bmatrix} - \begin{bmatrix} E(YZ) \\ E(WZ) \\ E(Y) \\ E(Z) \\ E(W) \end{bmatrix}\right] \xrightarrow{d} \mathcal{N}\left[0, Cov\begin{bmatrix} Y_i Z_i \\ W_i Z_i \\ Y_i \\ Z_i \\ W_i \end{bmatrix}\right]
$$

To apply the delta method, derive the gradient of $f(\cdot)$:

$$
\nabla f(\cdot) = \begin{bmatrix} \frac{\partial f(\cdot)}{\partial E(YZ)} \\ \frac{\partial f(\cdot)}{\partial E(WZ)} \\ \frac{\partial f(\cdot)}{\partial E(Y)} \\ \frac{\partial f(\cdot)}{\partial E(Z)} \\ \frac{\partial f(\cdot)}{\partial E(W)} \end{bmatrix} = \begin{bmatrix} \frac{1}{E(WZ)-E(W)E(Z)} \\ \frac{-(E(YZ)-E(Y)E(Z))}{(E(WZ)-E(W)E(Z))^2} \\ \frac{-E(Z)}{E(WZ)-E(W)E(Z)} \\ \frac{E(W)E(YZ)-E(Y)E(WZ)}{(E(WZ)-E(W)E(Z))^2} \\ \frac{E(Z)(E(YZ)-E(Y)E(Z))}{(E(WZ)-E(W)E(Z))^2} \end{bmatrix}
$$

which can be estimated with the sample analogs:

$$
\nabla \hat{f}(\cdot) = \begin{bmatrix} 41.506095 \\ 9.265631 \\ -11.380704 \\ -226.801905 \\ -2.540576 \end{bmatrix}
$$

Thus, applying the Delta Method:

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0, \nabla f(\cdot)' Cov \nabla f(\cdot))$$

yields the robust standard error of the LATE estimate:

```
[LATE]                   -0.2232354
[Robust Std. Error]       0.1575931
```

The code for calculating this in R is as follows:

```
> n <- nrow(data)
>
> Z <- data$draft
> W <- data$veteran
> Y <- data$logwage
> YZ <- Y*Z
> WZ <- W*Z
>
> ddf <- data.frame(YZ,WZ,Y,Z,W)
> d.vcov <- (1/n)*var(ddf)
>
> YZ.bar <- mean(YZ)
> WZ.bar <- mean(WZ)
> Y.bar <- mean(Y)
> Z.bar <- mean(Z)
> W.bar <- mean(W)
>
> gradient <- matrix(c(
+   1/(WZ.bar - Z.bar*W.bar),
+   (Y.bar*Z.bar - YZ.bar)/(WZ.bar - Z.bar*W.bar)^2,
+   Z.bar/(Z.bar*W.bar-WZ.bar),
+   (YZ.bar*W.bar - WZ.bar*Y.bar)/(WZ.bar - Z.bar*W.bar)^2,
+   Z.bar*(YZ.bar-Y.bar*Z.bar)/(WZ.bar - Z.bar*W.bar)^2
+ ),ncol=1)
>
> var.delta <- t(gradient)%*%d.vcov%*%gradient
> se.delta <- sqrt(var.delta)
> se.delta
          [,1]
```

```
[1,] 0.1575931
```

Problem 2

Consider one of the original papers introducing the use of IV to estimate LATE: Angrist, Imbens, Rubin 1996. They list five assumptions required for IV estimators to be consistent for LATE. This problem considers two of these assumptions in the context of influential social science research using observational data.

(a) Recall "Colonial Origins of Comparative Development" by Acemoglu, Johnson, and Robinson. They do not discuss the SUTVA assumption in their paper. Explain why this assumption might or might not be violated in practice, and provide real-world examples if you can.

SUTVA would be violated if one country's development outcomes might be affected by another country's institutional development. For example, we might imagine that institutional development in the (future) United States would directly affect the economies of all trading partners in Latin America (which are among AJR's cases).

(b) They also do not discuss the monotonicity assumption in their paper. Explain why this assumption might or might not be violated in practice, and provide real-world examples if you can.

This assumption would be violated if there were countries for which high settler mortality spurred the development of good institutions and low settler mortality spurred the development of extractive institutions. This could happen if, for example, the strength of the health system were a direct response to mortality, and if the health system then affected other institutional development.

(c) If you conclude that the assumptions are not violated, for what population parameter are their estimates consistent? If you conclude that the assumptions are violated, for what population parameter are their estimates consistent? Is this parameter of interest? Why or why not?

If the assumptions are not violated, their estimates are consistent for $LATE$. If the assumptions are violated, their estimates are consistent for

$$\frac{cov(GDP, Mortality)}{cov(Institutions, Mortality)}$$

which is not a parameter with clear theoretical meaning.

Problem 3

In this problem, we use the data set `Lee.dta` used in:

> Lee, David. "Randomized Experiments from Non-random Selection in U.S. House Elections," *Journal of Econometrics*, 142(2) 675-697.

The original paper can be found here: `http://www.princeton.edu/~davidlee/wp/RDrand.pdf`. The dataset contains 19857 observations with the following variables:

- `state`: state code
- `distnum`: congressional district number for each state
- `distid`: congressional district id (nationwide)
- `party`: party code (100 - Democrats and 200 - Republican)
- `partname`: party name
- `yearel`: year of election
- `origvote`: votes each candidate received
- `totvote`: total votes cast in each district in a given year
- `highestvote`: votes for candidate who received the largest votes in a district in a given year
- `sechighestvote`: votes for candidate who received second largest votes
- `officeexp`: terms served by a house representative.

Lee addressed the problem of measuring the electoral advantage of incumbency in the United States House of Representatives. In his article, the "incumbency advantage" is defined as the overall causal impact of being the current incumbent party in a district on the vote shares obtained in the district's election.

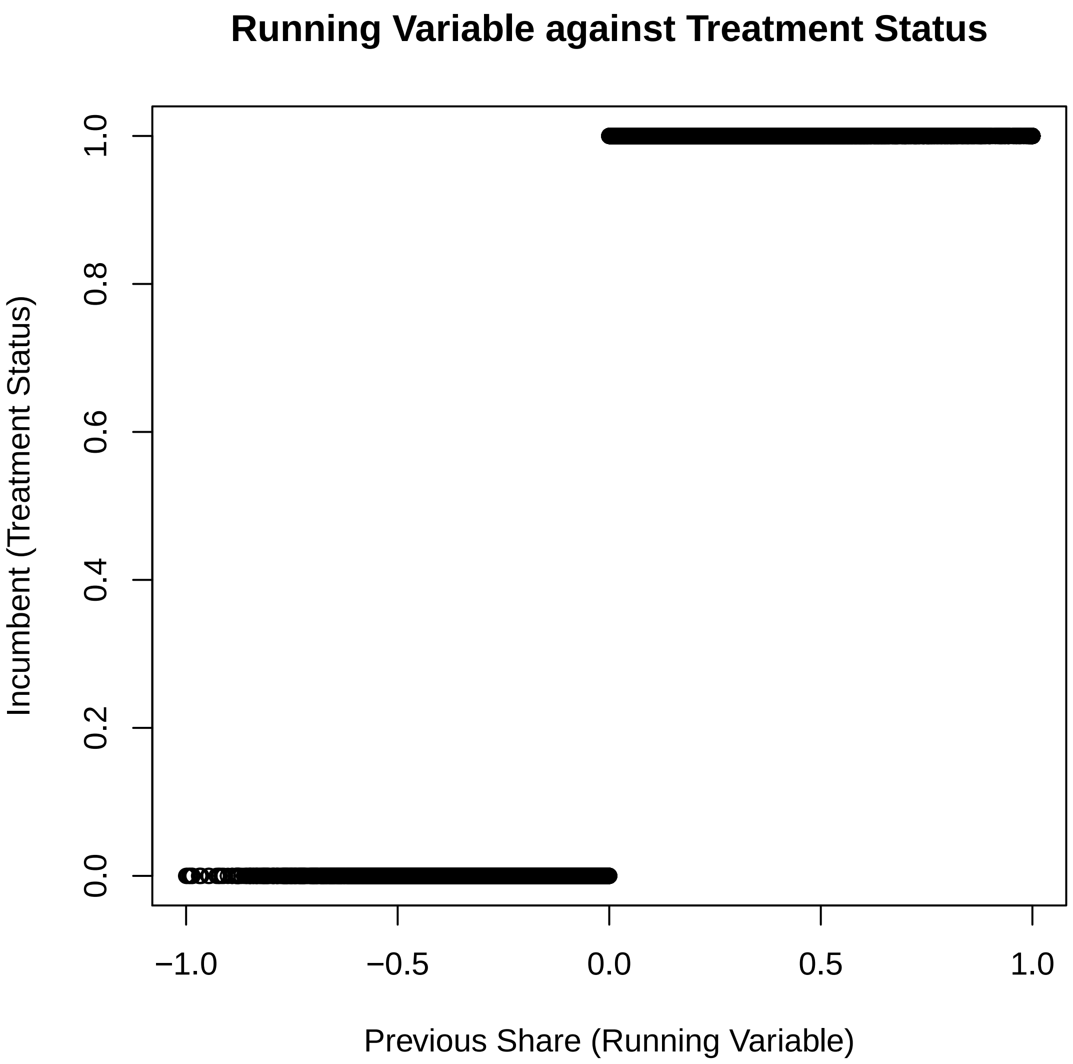
(a) Write down a reasonable RDD model for this problem. Be sure to explain what each variable in the model is. Also please give the interpretation for *each* coefficient in the model. Is this a sharp RDD, or a fuzzy one? What are the appropriate identification assumptions? How could you test the plausibility of these identification assumptions as far as possible?

The forcing variable (vote share in the previous election, X_i) will be used to quantify deviations from the threshhold of victory ($c = 0.5$). The dependent variable (Y_i) is the vote share in the current election. The treatment (party incumbency) D is 1 if $X_i - c > 0$ and 0 if $X_i - c < 0$. One simple RDD model would allow for different slopes on either side of the threshold: $E[Y|X, D] =$

$\beta_0 + \beta_1(X - c) + \beta_2(D) + \beta_3[(X - c)(D)] + u_i$. The coefficient on D will be our estimate of the average effect of the treatment at $X = c$.

(b) Generate the appropriate dependent, forcing, and treatment (i.e. party incumbency status) variables. Plot your forcing variable by treatment status to check whether your forcing variable properly defines the treatment status.

```
> ##Create winner dummy variable
> data$winner<-ifelse(data$origvote==data$highestvote,1,0)
> ##Create distance from threshhold variable
> data$voteshare<-ifelse(data$winner==1,(data$origvote-data$sechighestvote)/data$totvote,
> ##Sort data by district id, year and party, run loop to generate previous share
> sort.data<-data[order(data$distid,data$yearel,data$party),]
> previousshare<-c()
> for (i in 1:19857){
+   sub1<-sort.data[i,]
+   sub<-subset(sort.data,sort.data$party==sub1$party & sort.data$yearel==sub1$yearel-2 &
+   any<-nrow(sub); previousshare[i]<-ifelse(any==0,NA,sub$voteshare)
+ }
> incumbent<-ifelse(previousshare>0,1,0)
> new<-cbind(sort.data,previousshare,incumbent)
> plot(new$previousshare,new$incumbent, xlab="Previous Share (Running Variable)", ylab="I
```



(c) Using a bandwidth of 25% (±0.25) around the winning threshold, estimate the RDD with no additional covariates and assuming different linear slopes for the treated and untreated. Report the coefficient estimate corresponding to your treatment effect and its standard error. What is the proper interpretation of this value? That is, what does it actually identify?

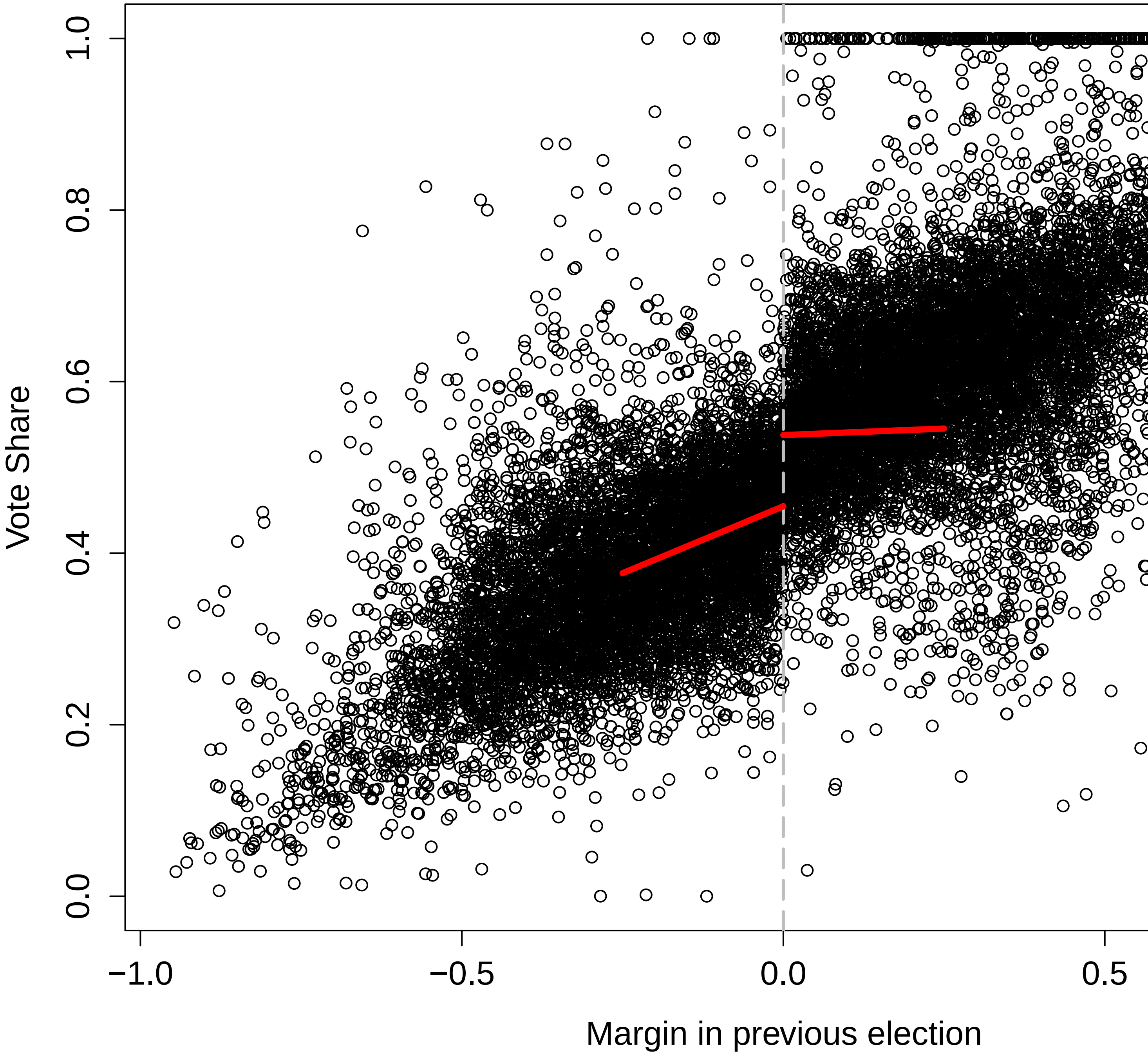
```
> trim<-subset(new,previousshare<0.25 & previousshare>-0.25)
> voteshareabs<-trim$origvote/trim$totvote; trim<-cbind(trim,voteshareabs)
> reg3 <-lm(voteshareabs~previousshare+incumbent+previousshare:incumbent, data=trim)
> summary(reg3)

Call:
lm(formula = voteshareabs ~ previousshare + incumbent + previousshare:incumbent,
    data = trim)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.52008 -0.05844 -0.00337  0.05232  0.61123

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             0.454567   0.003164 143.662   <2e-16 ***
previousshare           0.311610   0.022290  13.980   <2e-16 ***
incumbent               0.083064   0.004410  18.834   <2e-16 ***
previousshare:incumbent 0.030514   0.030931   0.987    0.324
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.09847 on 7920 degrees of freedom
Multiple R-squared:  0.4277,Adjusted R-squared:  0.4275
F-statistic:  1973 on 3 and 7920 DF,  p-value: < 2.2e-16
```

Under the assumptions of (a), we are identifying the average treatment effect of incumbency for those at the threshhold.

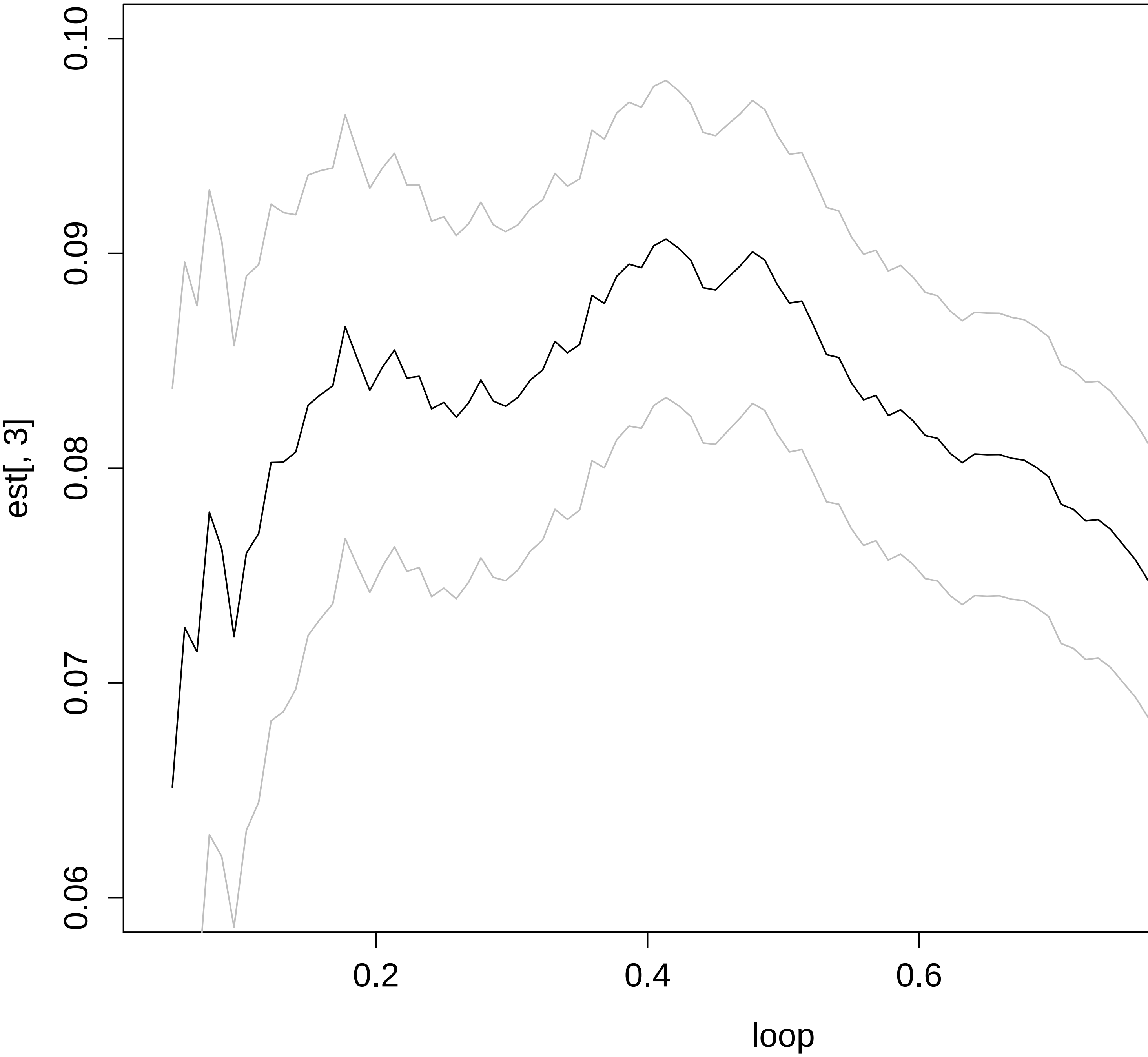
(d) Draw an RDD plot of your analysis. That is, show a scatter plot of the data, overlayed by the fitted values from your models on each side of the threshold.

```
> plot(trim$previousshare,trim$voteshareabs, xlab="Margin in previous election", ylab="Vo
> abline(v=0, col="gray", lwd=2, lty=5)
> lines(x=c(-0.25,0), y=c(summary(reg3)$coeff[1,1]-0.25*(summary(reg3)$coeff[2,1]),summar
>
> lines(x=c(0,0.25)
+       , y=c(summary(reg3)$coeff[1,1]+summary(reg3)$coeff[3,1]
+             ,summary(reg3)$coeff[1,1]+summary(reg3)$coeff[3,1]+0.25*(summary(reg3)$coef
```



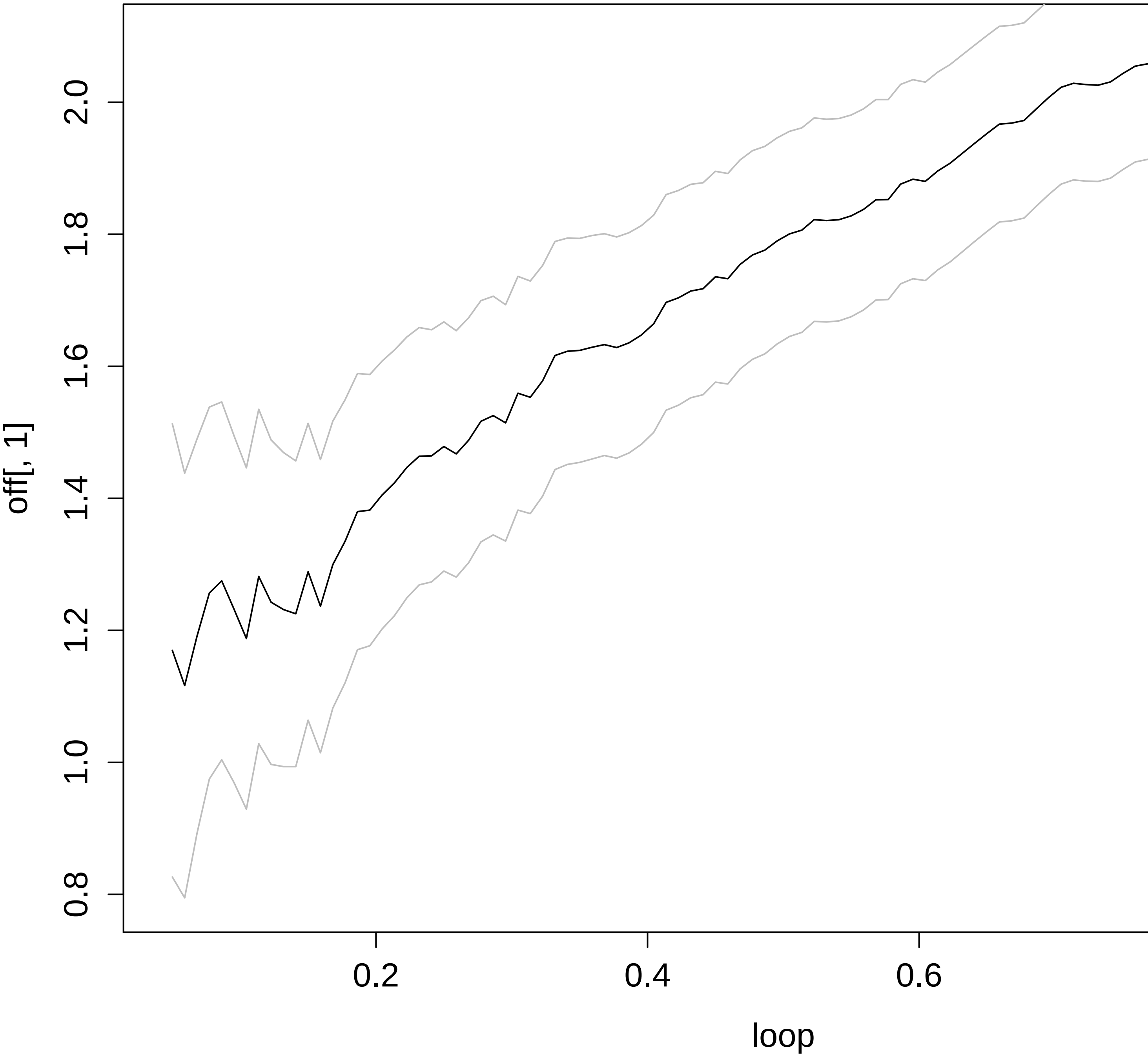
(e) Repeat the estimation of (c) for the bandwith range 5% - 95%. Plot the resulting estimates along with their 95% confidence intervals, with bandwidth on the x-axis.

```
> loop <- seq(from=0.05,to=.95,length.out=100)
> est <- matrix(NA,ncol=3,nrow=100)
> for (i in 1:100){
+   trim<-subset(new,previousshare<loop[i] & previousshare>-loop[i])
+   voteshareabs<-trim$origvote/trim$totvote
+   trim<-cbind(trim,voteshareabs)
+   m<-lm(voteshareabs~previousshare+incumbent+previousshare:incumbent, data=trim)
+   est[i,2]<-m$coefficients[3]
+   est[i,3]<-confint(m, level=0.95)[3,2]
+   est[i,1]<-confint(m, level=0.95)[3,1]
+ }
>
>
> plot(loop,est[,3], type=’l’,ylim=c(.06,.1), col=’gray’)
> lines(loop,est[,1], type=’l’, col=’gray’)
> lines(loop,est[,2], type=’l’)
```



(f) Report a similar plot to (e) using `officeexp`, the terms served by a house representative, as an outcome. What does the identification assumption in (a) suggest about the distribution of this variable at the threshold? Does it correspond to your findings?

```
> loop <- seq(from=0.05,to=.95,length.out=100)
> off <- matrix(NA,ncol=3,nrow=100)
> for (i in 1:100){
+   trim<-subset(new,previousshare<loop[i] & previousshare>-loop[i])
+   voteshareabs<-trim$origvote/trim$totvote
+   trim<-cbind(trim,voteshareabs)
+   m<-lm(officeexp~previousshare+incumbent+previousshare:incumbent, data=trim)
+   off[i,2]<-m$coefficients[3]
+   off[i,3]<-confint(m, level=0.95)[3,2]
+   off[i,1]<-confint(m, level=0.95)[3,1]
+ }
>
> plot(loop,off[,1], type=’l’, col=’gray’)
> lines(loop,off[,3], type=’l’, col=’gray’)
> lines(loop,off[,2], type=’l’)
```



Under the assumptions of (a), the number of terms served by a representative should not vary discontinuously at the threshold. However, as the plot indicates, incumbency right at the threshold is associated with more than one term on average, and with even more terms as the bandwith is broadened.

(g) Finally, let's run placebo tests where you compute a false incumbency status at other hypothetical threshold values, c^*. Conduct four such tests for $c^* = -0.15, -0.075, 0.075, 0.15$ (given that the actual threshold is $c = 0$, i.e. winners got positive and losers got negative vote share margins). For each placebo test, restrict the sample to the left or the right of the actual cutoff value (i.e. compute the effects at the negative placebo thresholds within the untreated (loser) sample and the effects at the positive placebo thresholds within the treated (winner) samples only). Report your placebo effect estimates along with standard errors. Do the results strengthen or weaken your confidence in the findings for the incumbency effect at the actual cutoff value?

```
> losers<-subset(trim,incumbent==0)
> winners<-subset(trim,incumbent==1)
> placebo.neg.015<-ifelse(losers$previousshare>-.15,1,0)
> reg.1<-lm(voteshareabs~previousshare+placebo.neg.015+previousshare:placebo.neg.015, data
> placebo.neg.075<-ifelse(losers$previousshare>-.075,1,0)
> reg.2<-lm(voteshareabs~previousshare+placebo.neg.075+previousshare:placebo.neg.075, data
> placebo.075<-ifelse(winners$previousshare>.075,1,0)
> reg.3<-lm(voteshareabs~previousshare+placebo.075+previousshare:placebo.075, data=winners
> placebo.015<-ifelse(winners$previousshare>.15,1,0)
> reg.4<-lm(voteshareabs~previousshare+placebo.015+previousshare:placebo.015, data=winners
> plac.1<-as.vector(summary(reg.1)$coeff[3,1:2]);plac.2<-as.vector(summary(reg.2)$coeff[3
> rbind(plac.1,plac.2,plac.3,plac.4)
               [,1]        [,2]
plac.1 -0.015659775 0.005161880
plac.2 -0.003997424 0.006107565
plac.3 -0.003795730 0.007925556
plac.4 -0.011948733 0.006406606
```

None of the placebo thresholds generates a result, which strengthens our confidence in the finding of an incumbency effect at the actual cutoff value.