Assignment 1: Hansen AER 2013

Nathan Franz

2/27/2020

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.3.0 --
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## -- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(ggthemes)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
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
```

```r
library(sandwich)

setwd("C:/Users/Nathan/Documents/UT Courses/3. Spring 2019/Causal Inference/CausalInferenceCourseHomewo

# Load Hansen's data
dwi = read.csv("./data/hansen_dwi.csv")
```

Summary of Hansen's Project

Hansen investigates whether harsher punishments influence the rate of recidivism among drunk drivers, using Washington state administrative records on DUI stops from 1995 to 2011. His investigation is motivated by theoretical results, such as Becker's seminal "Crime and Punishment" paper from 1974, that posit that crime is rational behavior given the incentive structure for would-be criminals.

He uses a regression discontinuity design taking advantage of the strict 0.08% and 0.15% blood alcohol content (BAC) thresholds for DUI and aggravated DUI charges, respectively. He estimates the local average treatment effect (LATE) of the punishments associated with DUI on recidivism rate within the next four years.

He finds that having a BAC over the 0.08 DUI threshold reduces recidivism by as much as 2 percentage points among all tested drivers. He also finds that having a BAC over the 0.15 aggravated DUI threshold further reduces recidivism by as much as 1 percentage point.

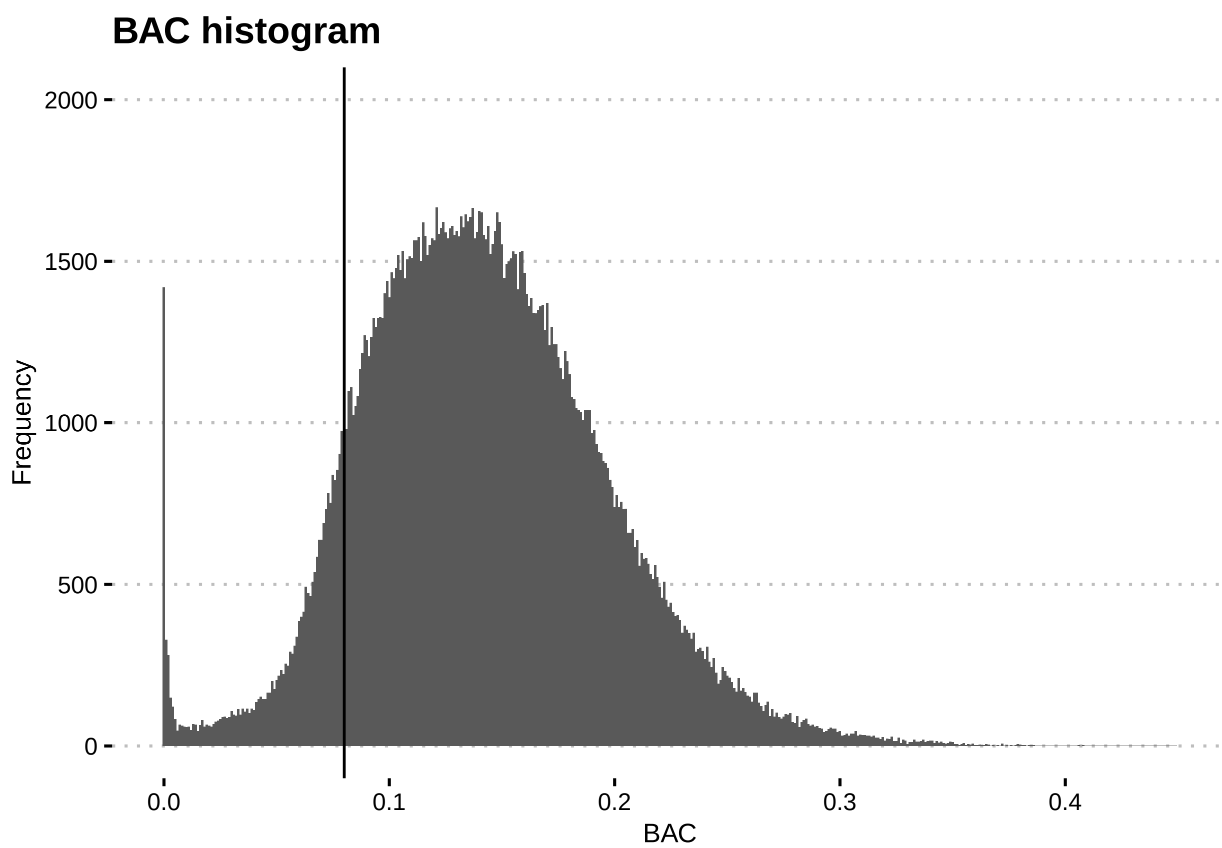
Replication

Density of the Running Variable

If people were capable of manipulating their blood alcohol content (BAC) with the precision required to fool the breathalyzer, the histogram of BAC would swell to the left of the legal threshold and drop to the right of it. That distortion of the distribution would occur because of the negative legal consequences for being charged with driving while intoxicated.

In the recreation of Hansen's BAC histogram below we do not see any evidence for sorting on the running variable. The characteristic density discontinuity is not present at BAC = 0.08, the black vertical line.

```
# add new vars:
#   bac_min is the minimum of the two measured BACs (as used in the paper)
#   bac_overlimit is an indicator variable for whether BAC is over the 0.08% limit

dwi$over_limit <- as.numeric(eval(dwi$bac1 >= 0.08))



# recreate figure 1, histogram of bac_1

manipulation_histogram <- ggplot(data = dwi) +
  theme_clean() +
  theme(plot.background = element_blank()) +
  geom_histogram(aes(x=bac1), binwidth = 0.001) +
  geom_vline(xintercept = 0.08) +
  labs(title="BAC histogram") +
  xlab("BAC") +
  ylab("Frequency") +
  ylim(0,2000)
manipulation_histogram
```



Covariates by the Running Variable

```
# run regressions of covariates on running variable
lm_male_on_bac <- lm(male ~ I(bac1-0.08)*over_limit, data=dwi)
lm_white_on_bac <- lm(white ~ I(bac1-0.08)*over_limit, data=dwi)
lm_aged_on_bac <- lm(aged ~ I(bac1-0.08)*over_limit, data=dwi)
lm_acc_on_bac <- lm(acc ~ I(bac1-0.08)*over_limit, data=dwi)

models <- list(lm_male_on_bac, lm_white_on_bac, lm_aged_on_bac, lm_acc_on_bac)

# calculate robust standard errors for each regression

get_robust_ses <- function(fit) {
  sqrt(
    diag(
      sandwich::vcovHC(fit, type = "HC1")
    )
  )
}

# create Latex table

tbl1 <- stargazer(models, header = FALSE, style = "aer",
          title = "Regression Discontinuity Estimates for Covariates",
          column.labels = c("Male", "White","Age", "Accident"),
```

```
         covariate.labels = c("DUI"),
         omit = c("Constant", "bac1", "bac1:over_limit"),
         se = lapply(models, get_robust_ses),
         dep.var.caption = '',
         dep.var.labels.include = FALSE)
```

Table 1: Regression Discontinuity Estimates for Covariates

	Male	White	Age	Accident
	(1)	(2)	(3)	(4)
DUI	0.006	0.004	-1.115^{***}	-0.007^{**}
	(0.004)	(0.004)	(0.121)	(0.003)
Observations	214,558	214,558	214,558	214,558
R^2	0.000	0.001	0.013	0.021
Adjusted R^2	0.000	0.001	0.013	0.021
Residual Std. Error (df = 214554)	0.408	0.345	11.431	0.351
F Statistic (df = 3; 214554)	9.428^{***}	58.401^{***}	907.467^{***}	$1{,}566.761^{***}$

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

As shown in Table 1, only age shows a significant discontinuity at the BAC threshold.

The figure below shows linear model fits. Panel C in the linear model plots illuminates the significant imbalance found for the linear regression of age on BAC presented in the table above. It seems to be an artifact of including high values of BAC, which biases the model; all the average BAC levels (represented by empty dots) within .02 percentage points of the legal BAC threshold lie above the linear fit.

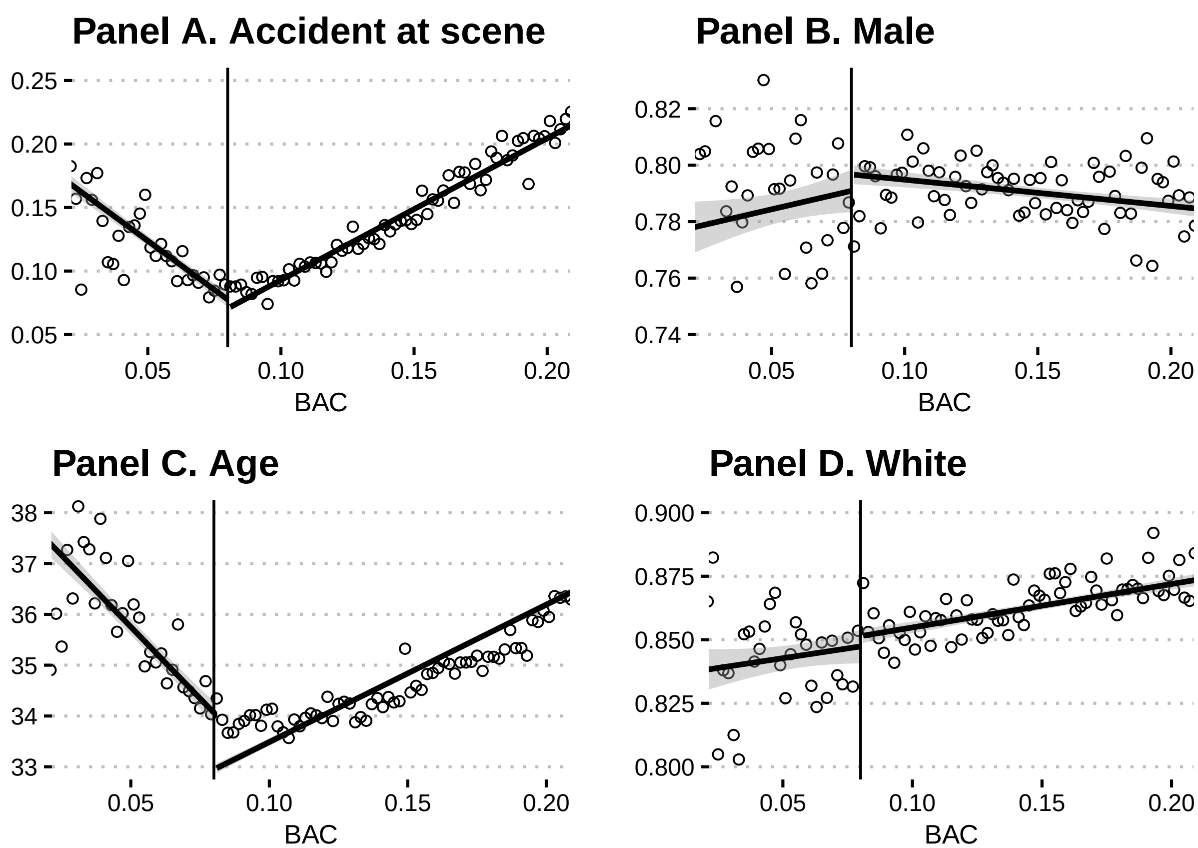
There are several differences between these plots and the corresponding plots presented in Hansen's paper. Obviously, the replication plots use only one threshold BAC level, rather than the two shown in the original. These influence our regression results by including data from farther away from the threshold than Hansen uses.

Additionally, Panel B in the replication plot shows a positive slope below the 0.08 BAC threshold, unlike the negative slope in the original. Also, Panel C in the original has an age axis that ranges from 0.34 to 0.38, three orders of magnitude smaller than the ranges in the replication.

```
# recreate figure 2, panels A-D

dwi$bac_bucket <- cut(dwi$bac1, seq(from=0, to=0.450, by=0.002),
                      labels=(seq(from=0.001, to=0.449, by=0.002)),
                      include.lowest=TRUE)

dwi_sum <- dwi %>%
  group_by(bac_bucket) %>%
  summarize(
    male = mean(male),
    white = mean(white),
    acc = mean(acc),
    aged = mean(aged),
    over_limit = unique(over_limit)
  )
```

```r
dwi_sum$bac1 <- as.numeric(as.character(dwi_sum$bac_bucket))

# panel A - accident at scene

panel_a <- ggplot(data = dwi,
       aes(x = bac1,
           y = acc,
           group = over_limit)) +
  theme_clean() +
  theme(plot.background=element_blank()) +
  geom_point(data = dwi_sum, shape = 1) +
  geom_smooth(method = 'lm', color = 'black') +
  geom_vline(xintercept = 0.08) +
  #geom_vline(xintercept = 0.15) +
  labs(title="Panel A. Accident at scene") +
  xlab("BAC") +
  ylab("") +
  coord_cartesian(xlim = c(0.03, 0.2), ylim = c(0.05, 0.25))

# panel B - male

panel_b <- ggplot(data = dwi,
                  aes(x = bac1,
                      y = male,
                      group = over_limit)) +
  theme_clean() +
  theme(plot.background=element_blank()) +
  geom_point(data = dwi_sum, shape = 1) +
  geom_smooth(method = 'lm', color = 'black') +
  geom_vline(xintercept = 0.08) +
  #geom_vline(xintercept = 0.15) +
  labs(title="Panel B. Male") +
  xlab("BAC") +
  ylab("") +
  coord_cartesian(xlim = c(0.03, 0.2), ylim = c(0.74, 0.83))

# panel C - age

panel_c <- ggplot(data = dwi,
                  aes(x = bac1,
                      y = aged,
                      group = over_limit)) +
  theme_clean() +
  theme(plot.background=element_blank()) +
  geom_point(data = dwi_sum, shape = 1) +
  geom_smooth(method = 'lm', color = 'black') +
  geom_vline(xintercept = 0.08) +
  #geom_vline(xintercept = 0.15) +
  labs(title="Panel C. Age") +
  xlab("BAC") +
  ylab("") +
  coord_cartesian(xlim = c(0.03, 0.2), ylim = c(33, 38))
```

```
# panel D - white

panel_d <- ggplot(data = dwi,
                  aes(x = bac1,
                      y = white,
                      group = over_limit)) +
  theme_clean() +
  theme(plot.background=element_blank()) +
  geom_point(data = dwi_sum, shape = 1) +
  geom_smooth(method = 'lm', color = 'black') +
  geom_vline(xintercept = 0.08) +
  #geom_vline(xintercept = 0.15) +
  labs(title="Panel D. White") +
  xlab("BAC") +
  ylab("") +
  coord_cartesian(xlim=c(0.03, 0.2), ylim=c(0.8, 0.9))

# Assemble panels into figure 2 replication
grid.arrange(panel_a, panel_b, panel_c, panel_d, nrow = 2)
```

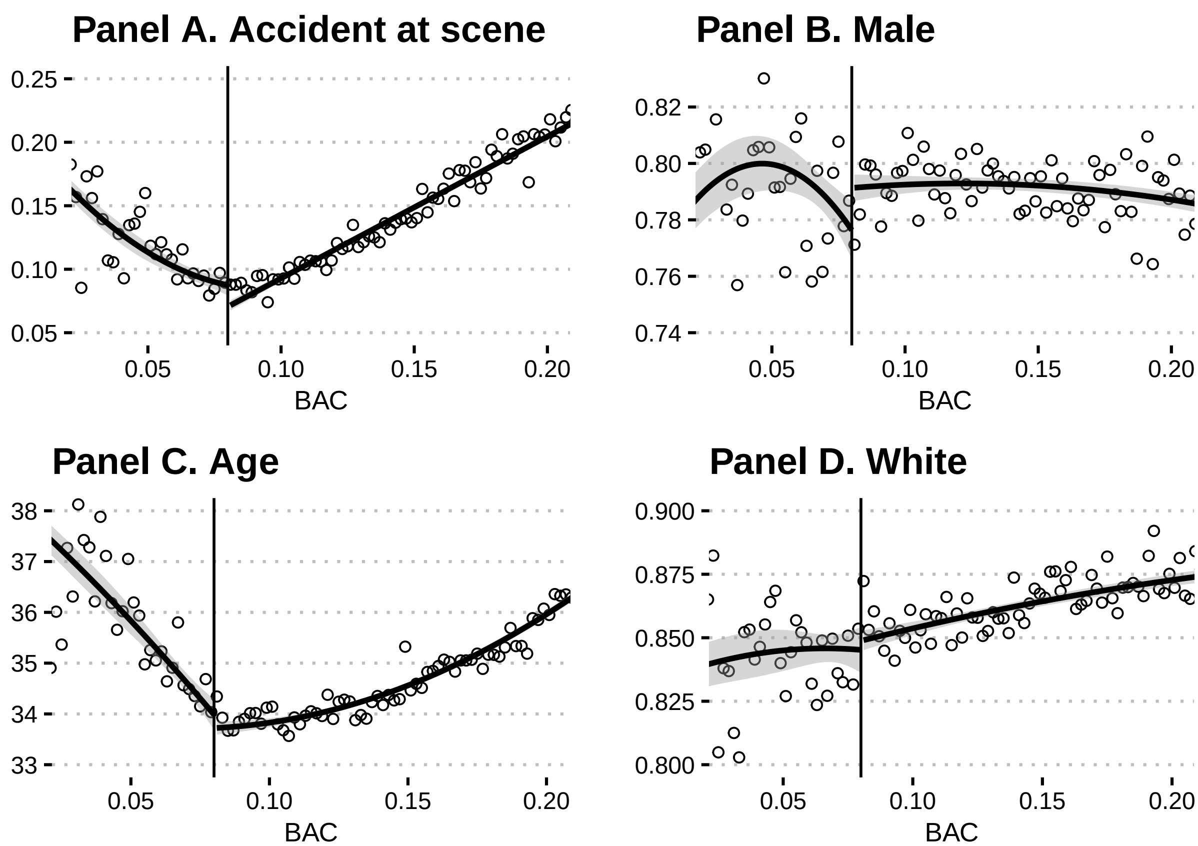


The bias introduced by the linear model for age above is fixed by using a quadratic model, presented in Panel C of the figure below. However, the quadratic model also introduces a potentially disqualifying amount of variance. For example, Panel B shows a questionable fit below the legal BAC threshold.

```
# recreate figure 2, panels A-D, but with quadratic model

# panel A - accident at scene
```

```r
panel_a_quadratic <- ggplot(data = dwi,
                   aes(x = bac1,
                       y = acc,
                       group = over_limit)) +
  theme_clean() +
  theme(plot.background=element_blank()) +
  geom_point(data = dwi_sum, shape = 1) +
  geom_smooth(method = 'lm', formula= y ~ poly(x,2), color = 'black') +
  geom_vline(xintercept = 0.08) +
  #geom_vline(xintercept = 0.15) +
  labs(title="Panel A. Accident at scene") +
  xlab("BAC") +
  ylab("") +
  coord_cartesian(xlim = c(0.03, 0.2), ylim = c(0.05, 0.25))

# panel B - male

panel_b_quadratic <- ggplot(data = dwi,
                   aes(x = bac1,
                       y = male,
                       group = over_limit)) +
  theme_clean() +
  theme(plot.background=element_blank()) +
  geom_point(data = dwi_sum, shape = 1) +
  geom_smooth(method = 'lm', formula= y ~ poly(x,2), color = 'black') +
  geom_vline(xintercept = 0.08) +
  #geom_vline(xintercept = 0.15) +
  labs(title="Panel B. Male") +
  xlab("BAC") +
  ylab("") +
  coord_cartesian(xlim = c(0.03, 0.2), ylim = c(0.74, 0.83))

# panel C - age

panel_c_quadratic <- ggplot(data = dwi,
                   aes(x = bac1,
                       y = aged,
                       group = over_limit)) +
  theme_clean() +
  theme(plot.background=element_blank()) +
  geom_point(data = dwi_sum, shape = 1) +
  geom_smooth(method = 'lm', formula= y ~ poly(x,2), color = 'black') +
  geom_vline(xintercept = 0.08) +
  #geom_vline(xintercept = 0.15) +
  labs(title="Panel C. Age") +
  xlab("BAC") +
  ylab("") +
  coord_cartesian(xlim = c(0.03, 0.2), ylim = c(33, 38))

# panel D - white

panel_d_quadratic <- ggplot(data = dwi,
                   aes(x = bac1,
```

```r
                         y = white,
                         group = over_limit)) +
  theme_clean() +
  theme(plot.background=element_blank()) +
  geom_point(data = dwi_sum, shape = 1) +
  geom_smooth(method = 'lm', formula= y ~ poly(x,2), color = 'black') +
  geom_vline(xintercept = 0.08) +
  #geom_vline(xintercept = 0.15) +
  labs(title="Panel D. White") +
  xlab("BAC") +
  ylab("") +
  coord_cartesian(xlim=c(0.03, 0.2), ylim=c(0.8, 0.9))

# Assemble panels into figure 2 replication, but quadratic
grid.arrange(panel_a_quadratic, panel_b_quadratic,
             panel_c_quadratic, panel_d_quadratic, nrow = 2)
```



Regression Discontinuity

```r
# Estimate equation 1 with recidivism as the outcome,
#   using BAC data in [0.03,0.13]

# Make linear models, control for bac1 linearly,
#                      interact bac1 with cutoff linearly,
```

```r
#                      interact bac1 with cutoff linearly and quadratically

linear_control <- lm(recidivism ~ 1 + male + white + aged + acc +
                       bac1 + over_limit,
   data = dwi[dwi$bac1 >= 0.03 & dwi$bac1 <= 0.13,])

linear_interact <- lm(recidivism ~ 1 + male + white + aged + acc +
                        bac1*over_limit,
   data = dwi[dwi$bac1 >= 0.03 & dwi$bac1 <= 0.13,])

quadratic_interact <- lm(recidivism ~ 1 + male + white + aged + acc +
                            bac1*over_limit + I(bac1^2)*over_limit,
   data = dwi[dwi$bac1 >= 0.03 & dwi$bac1 <= 0.13,])

# Make a list of the models from above for table

rd_panel_a_models = list(linear_control, linear_interact, quadratic_interact)

# Create Latex table for R Markdown

stargazer(rd_panel_a_models, header = FALSE, style = "aer",
          title = "Regression Discontinuity Estimates for the Effect of DUI on Recidivism, BAC in [0.03
          column.labels = c("Linear Control", "With Interaction","With Quadratic Interaction"),
          covariate.labels = c("DUI"),
          omit = c("Constant", "male", "white", "aged", "acc", "bac1",
                   "I(bac1^2)", "bac1:over_limit", "over_limit:I(bac1^2)"),
          se = lapply(rd_panel_a_models, get_robust_ses),
          dep.var.caption = '',
          dep.var.labels.include = FALSE)
```

Table 2: Regression Discontinuity Estimates for the Effect of DUI on Recidivism, BAC in [0.03, 0.13]

	Linear Control	With Interaction	With Quadratic Interaction
	(1)	(2)	(3)
DUI	-0.027^{***}	-0.059^{***}	0.113
	(0.004)	(0.015)	(0.084)
Observations	89,967	89,967	89,967
R^2	0.004	0.004	0.004
Adjusted R^2	0.004	0.004	0.004
Residual Std. Error	0.308 (df = 89960)	0.308 (df = 89959)	0.308 (df = 89957)
F Statistic	54.029^{***} (df = 6; 89960)	46.995^{***} (df = 7; 89959)	37.092^{***} (df = 9; 89957)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

```r
# Estimate equation 1 with recidivism as the outcome,
#   using BAC data in [0.055,0.105]

# Make linear models, control for bac1 linearly,
#                      interact bac1 with cutoff linearly,
#                      interact bac1 with cutoff linearly and quadratically
```

```r
linear_control_narrow <- lm(recidivism ~ 1 + male + white + aged + acc +
                       bac1 + over_limit,
                     data = dwi[dwi$bac1 >= 0.055 & dwi$bac1 <= 0.105,])

linear_interact_narrow <- lm(recidivism ~ 1 + male + white + aged + acc +
                        bac1*over_limit,
                      data = dwi[dwi$bac1 >= 0.055 & dwi$bac1 <= 0.105,])

quadratic_interact_narrow <- lm(recidivism ~ 1 + male + white + aged + acc +
                           bac1*over_limit + I(bac1^2)*over_limit,
                         data = dwi[dwi$bac1 >= 0.055 & dwi$bac1 <= 0.105,])

# Make a list of the models from above for table

rd_panel_b_models = list(linear_control_narrow, linear_interact_narrow,
                         quadratic_interact_narrow)

# Create Latex table for R Markdown

stargazer(rd_panel_b_models, header = FALSE, style = "aer",
          title = "Regression Discontinuity Estimates for the Effect of DUI on Recidivism, BAC in [0.05
          column.labels = c("Linear Control", "With Interaction","With Quadratic Interaction"),
          covariate.labels = c("DUI"),
          omit = c("Constant", "male", "white", "aged", "acc", "bac1",
                   "I(bac1^2)", "bac1:over_limit", "over_limit:I(bac1^2)"),
          se = lapply(rd_panel_b_models, get_robust_ses),
          dep.var.caption = '',
          dep.var.labels.include = FALSE)
```

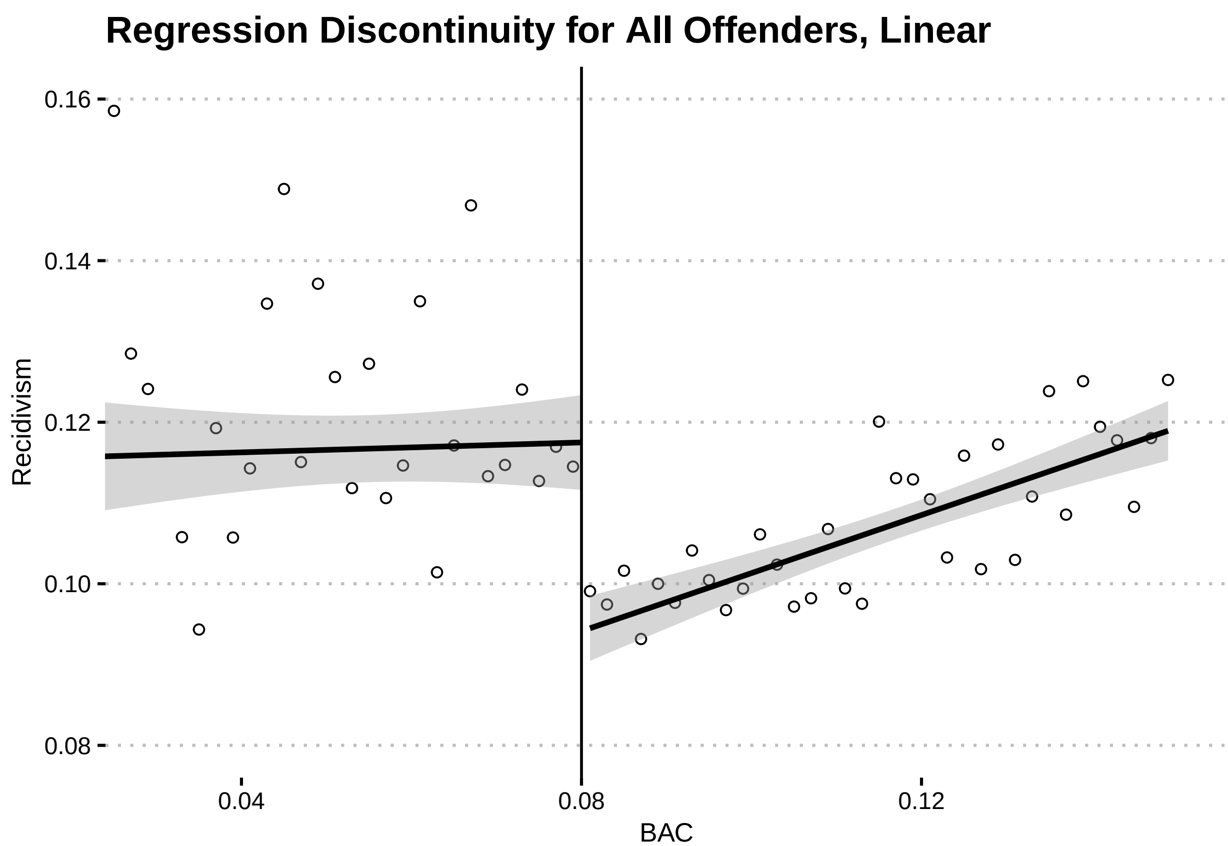
Table 3: Regression Discontinuity Estimates for the Effect of DUI on Recidivism, BAC in [0.055, 0.105]

	Linear Control	With Interaction	With Quadratic Interaction
	(1)	(2)	(3)
DUI	-0.022^{***}	-0.069^{**}	0.270
	(0.006)	(0.034)	(0.409)
Observations	47,205	47,205	47,205
R^2	0.004	0.004	0.004
Adjusted R^2	0.004	0.004	0.004
Residual Std. Error	0.306 (df = 47198)	0.306 (df = 47197)	0.306 (df = 47195)
F Statistic	31.384^{***} (df = 6; 47198)	27.198^{***} (df = 7; 47197)	21.232^{***} (df = 9; 47195)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

```r
# Replicate Panel A of figure 3 from paper
dwi_rd_sum <- dwi[dwi$bac1<0.15,] %>%
  group_by(bac_bucket) %>%
  summarize(
    recidivism = mean(recidivism),
    over_limit = unique(as.numeric(eval(bac1 >= 0.08)))
```

```
  )

dwi_rd_sum$bac1 <- as.numeric(as.character(dwi_rd_sum$bac_bucket))

# RD - panel A, linear
rd_linear <- ggplot(data = dwi[dwi$bac1<0.15,],
                    aes(x = bac1,
                        y = recidivism,
                        group = over_limit)) +
  theme_clean() +
  theme(plot.background=element_blank()) +
  geom_point(data = dwi_rd_sum, shape = 1) +
  geom_smooth(method = 'lm', color = 'black') +
  geom_vline(xintercept = 0.08) +
  labs(title="Regression Discontinuity for All Offenders, Linear") +
  xlab("BAC") +
  ylab("Recidivism") +
  coord_cartesian(xlim = c(0.03, 0.15), ylim = c(0.08, 0.16))
rd_linear
```




```
# RD - panel A, quadratic
rd_quadratic <- ggplot(data = dwi[dwi$bac1<0.15,],
                       aes(x = bac1,
                           y = recidivism,
                           group = over_limit)) +
  theme_clean() +
```

```
  theme(plot.background=element_blank()) +
  geom_point(data = dwi_rd_sum, shape = 1) +
  geom_smooth(method = 'lm', formula= y ~ poly(x,2), color = 'black') +
  geom_vline(xintercept = 0.08) +
  labs(title="Regression Discontinuity for All Offenders, Quadratic") +
  xlab("BAC") +
  ylab("Recidivism") +
  coord_cartesian(xlim = c(0.03, 0.15), ylim = c(0.08, 0.16))
rd_quadratic
```