ECON 1190 Problem Set 6: Regression Discontinuity

Claire Duquennois

Name: Rohan Krishnan

1 Empirical Analysis using Data from Manacorda, Miguel, & Vigorito (2011, American Economic Journal: Applied Economics)

This exercise uses data from Manacorda, Miguel, & Vigorito's paper, "Government Transfers and Political Support," published in the *American Economic Journal: Applied Economics* in 2011. This paper studies how receipt of a government anti-poverty cash transfer changes how beneficiary households support and view the government.

The data can be found on Edward Miguel's faculty website. Download and extract the contents from the `Government_Transfers_replication.zip` file.

2 Set up and constructing the data

The original data used in the paper is confidential. The authors instead provide the `reg_panes.dta` data file which is anonymized and created from the original data.

2.1 Question: Loading the Packages

Load any R packages you will be using:

Code:

```
library(haven)
library(dplyr)
library(stargazer)
library(lfe)
library(broom)
library(statar)
library(ggplot2)
```

2.2 Question: Open the `reg_panes.dta` file. To complete this problem set you will need the following variables from this data file:

Name	Description
aprobado	Ever received PANES 2005-2007
untracked07	Untracked in 2007
h_89	Supports current government 2007 [1 to 3]
hv34	Supports current government 2008 [1 to 3]
ind_reest	Predicted Income
newtreat	PANES eligibility

Drop all other variables. If needed, give the variables you are keeping more intuitive names.

Code:

```
data <- read_dta("~/Downloads/reg_panes.dta")

data1 <- data %>%
  select(aprobado, untracked07, h_89, hv34, ind_reest, newtreat)

is(data1)
```

```
## [1] "tbl_df"     "tbl"        "data.frame" "list"       "oldClass"
## [6] "vector"
```

```
data1 <- data.frame(data1)
is(data1)
```

```
## [1] "data.frame" "list"       "oldClass"   "vector"
```

```
colnames(data1) <- c("ever_received", "untracked07", "gov_support07", "gov_support08", "pred_inc", "eli
```

2.3 Question: The data as downloaded will require that you clean the variables of interest and construct a new dataset to generate the graphs. Start by generating the following cleaned variable:

-An indicator for receiving PANES that is NA if a respondent is untracked in 2007

Code:

```
data1$PANES_ind <- ifelse(data1$untracked07 == 1, NA,1)
```

2.4 Question: We are going to re-scale the variables that indicate support for the current government so that responses range from 0 to 1. To do this, tabulate the current variable to see how it is distributed and then generate a variable that will be NA if it is currently coded as 9, 0 if currently 2, 0.5 if currently 1 and 1 if currently 3. Do this for both the 2007 and 2008 variable.

Note: This is how the authors modify this variable in their code. It seems counter intuitive and does not correspond to the description of how this variable is coded in the survey questionnaire as reported in their appendix though it does correspond to their discussion in footnote 12. My guess is the transcription/translation of the survey question is incorrect.

Code:

```
table(data1$gov_support07)
```

```
##
##    1    2    3    9
##  397  122 1570  130
```

```
unique(data1$gov_support07)
```

```
## <labelled<double>[5]>:  comparación entre el gobierno actual y anterior
## [1] NA  3  9  1  2
##
## Labels:
##  value     label
##      1     igual
##      2      peor
##      3     mejor
##      9  ignorado
```

```
sum(is.na(data1$gov_support07))
```

```
## [1] 879
```

```
data1$gov_support07 <- ifelse(data1$gov_support07 == 9, NA,
                              ifelse(data1$gov_support07 == 2, 0,
                              ifelse(data1$gov_support07 == 1, 0.50,
                              ifelse(data1$gov_support07 == 3, 1, NA))))

table(data1$gov_support08)
```

```
##
##    1    2    3    9
##  476  159 1313  103
```

```
unique(data1$gov_support08)
```

```
## <labelled<double>[5]>: 84. en relación al gobierno anterior, ¿cree que el gobierno actual es...?
## [1] NA  1  3  2  9
##
## Labels:
##  value    label
##      1    igual
##      2     peor
##      3    mejor
##      9 ignorado
```

```r
sum(is.na(data1$gov_support08))
```

```
## [1] 1047
```

```r
data1$gov_support08 <- ifelse(data1$gov_support08 == 9, NA,
                              ifelse(data1$gov_support08 == 2, 0,
                              ifelse(data1$gov_support08 == 1, 0.50,
                              ifelse(data1$gov_support08 == 3, 1, NA))))
```

2.5 Question: Generate a variable that is the square of predicted income.

Code:

```
data1$pred_inc_sq <- (data1$pred_inc)^2
```

3 We start by reproducing the main figures (2,3,and 4) of the paper as good figures are key to any regression discontinuity paper.

3.1 Question: The data consists of over 3000 observations. How many points are plotted on these figures? How should we interpret the y axis? What does each point below the threshold represent? What does each point above the threshold represent?

Answer: There are 45 points are plotted in the figures. There are 30 for households below threshold and 15 for households above the threshold. Each point has about 43 observations in it.

The y axis represents the proportion of houses that received the PANES benefits for Figure 2. For Figures 3 and 4, the y axis is the proportion of houses within the cell's average support for the government for 2007 and 2008 respectively.

Each point below the threshold represents a group of ~43 households that fall below the PANES eligibility threshold (are eligible) based on predicted income. Each point above the threshold represents a group of ~43 households that fall above the PANES eligibility threshold (aren't eligible) based on predicted income.

3.2 Question: Why is the number of points above the threshold different from the number below?

Answer: The number of points above and below the threshold are different because there are around twice as many households below the threshold than above it in the data. The bins are designed to contain an approximately equal number of households per bin so there are more bins required below the threshold to ensure that each bin (above and below) contain the same number of observations.

3.3 Question: Replicating these figures will require restructuring our data and calculating the values that are plotted. Generate a variable that will indicate the percentile group the observation is in. Note the difference in the number of percentile groups above and below the threshold.

Note: you may find the `xtile` function in R useful.

Code:

```
restruc_data_below <- data1 %>%
  filter(pred_inc < 0) %>%
  mutate(group = xtile(pred_inc, n = 30))

restruc_data_above <- data1 %>%
  filter(pred_inc > 0) %>%
  mutate(group = (30 + xtile(pred_inc, n = 15)))

group_data <- rbind(restruc_data_below, restruc_data_above)

range(group_data$group)
```

```
## [1]  1 45
```

3.4 Question: For each of the percentile groups, calculate the mean of each of the variables we will use for plotting: predicted income, receipt of PANES, support for the government in 2007, and support for the government in 2008.

Code:

```
fig_data <- group_data %>%
  group_by(group) %>%
  summarize(
    mean_inc = mean(pred_inc),
    mean_rec = mean(ever_received),
    mean_supp07 = mean(gov_support07, na.rm = T),
    mean_supp08 = mean(gov_support08, na.rm = T)
  )
```

3.5 Question: Replicate figure 2. Make the figure as clear and informative as possible. You may want to create an indicator variable for percentiles above and below the threshold.

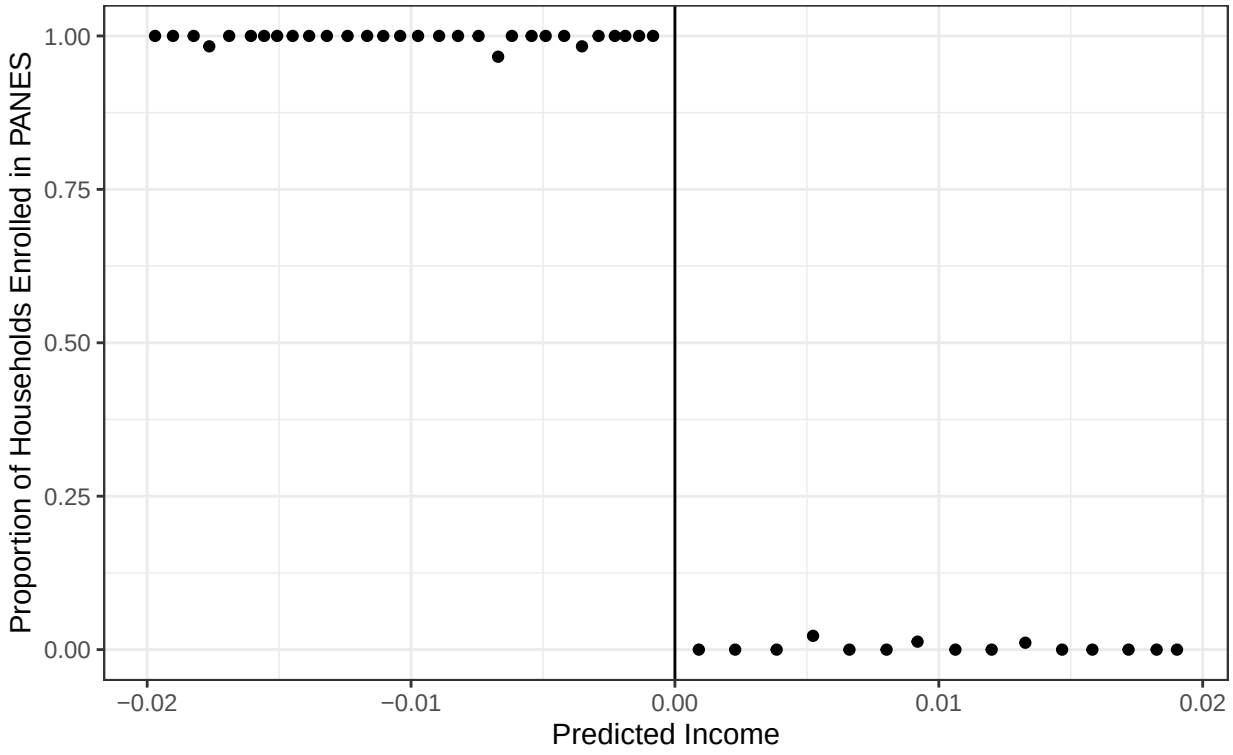
Code:

```
fig_data$thresh <- ifelse(fig_data$mean_inc < 0, 1, 0)
fig2 <- fig_data %>%
  ggplot(aes(x = mean_inc, y = mean_rec)) +
  geom_point() +
  geom_vline(xintercept = 0) +
  labs(x = "Predicted Income",
       y = "Proportion of Households Enrolled in PANES",
       title = "Figure 2: PANES Program Eligibility and Participation",
       subtitle = "Figure Replication") +
  theme_bw() +
  theme(plot.title = element_text(face = "bold"),
        plot.subtitle = element_text(face = "italic"))

fig2
```

Figure 2: PANES Program Eligibility and Participation

Figure Replication



3.6 Question: What is the purpose of this figure and what should we take away from it?

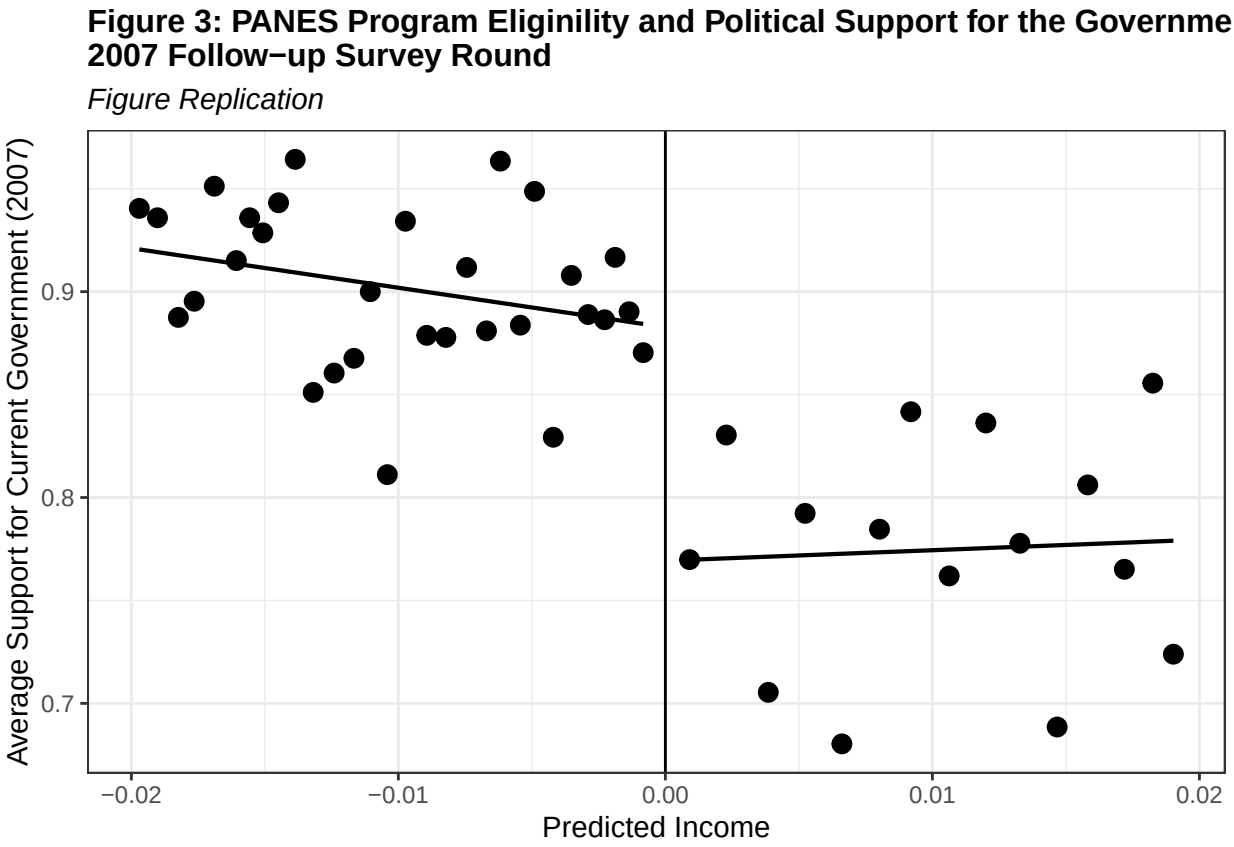
Answer: This figure shows how strictly the PANES program is enforced. Ideally, points to the left of the vertical line would be at 1 and points to the right would be at 0. We can see there is some small discrepancies in some of the points on both sides (likely due to how I binned using `xtile`); however, it appears that the PANES program was administered fairly strongly overall as the deviations are quite small.

3.7 Question: Replicate figures 3 and 4. Make these figures as clear and informative as possible (2pages).

Code:

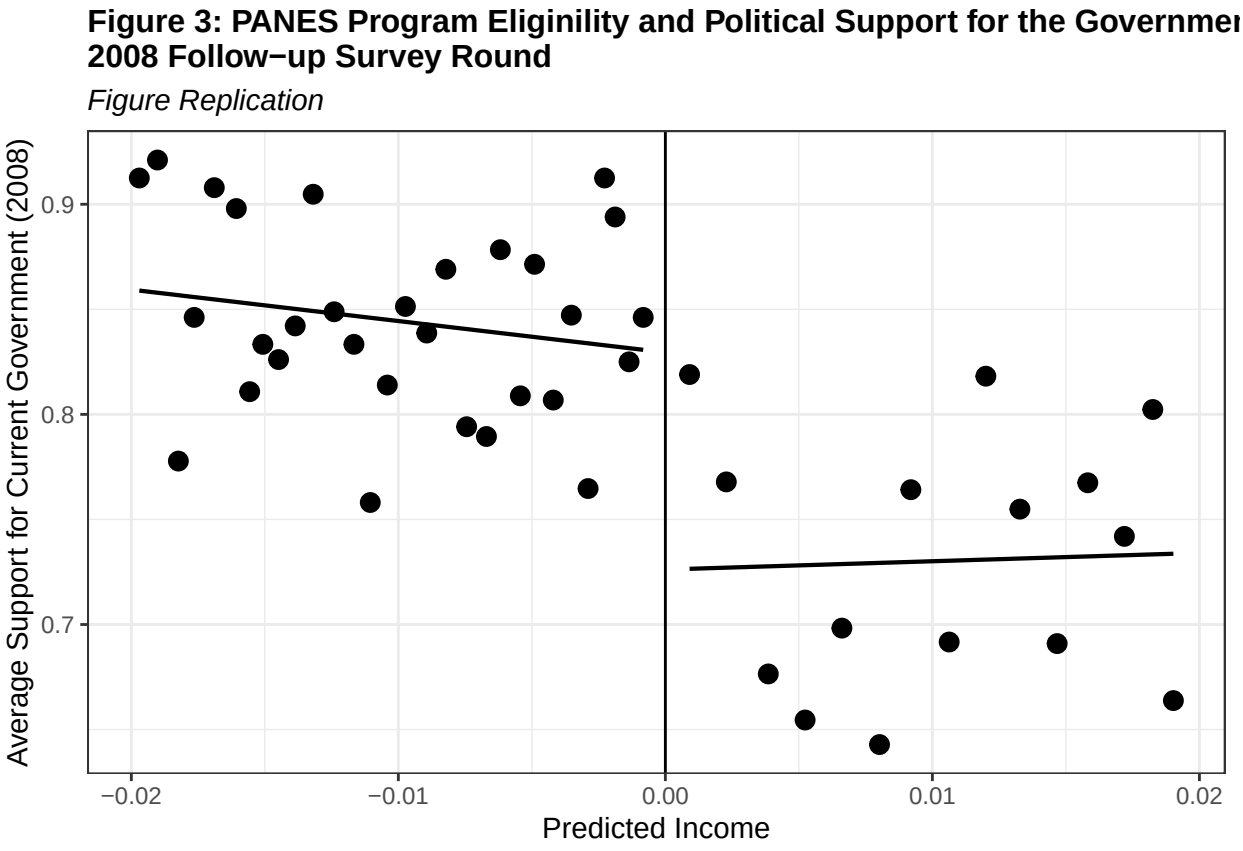
```
fig3 <- fig_data %>%
  ggplot(aes(x = mean_inc, y = mean_supp07)) +
  geom_point(size = 3) +
  geom_vline(xintercept = 0) +
  geom_smooth(method = "lm", data = fig_data %>% filter(mean_inc < 0),
              se = FALSE, color = "black", linewidth = 0.75) +
  geom_smooth(method = "lm", data = fig_data %>% filter(mean_inc > 0),
              se = FALSE, color = "black", linewidth = 0.75) +
  labs(x = "Predicted Income",
       y = "Average Support for Current Government (2007)",
       title = "Figure 3: PANES Program Eliginility and Political Support for the Government,\n2007 Foll
       subtitle = "Figure Replication") +
  theme_bw() +
  theme(plot.title = element_text(face = "bold", size = 12),
        plot.subtitle = element_text(face = "italic"))
fig3
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



```
fig4 <- fig_data %>%
  ggplot(aes(x = mean_inc, y = mean_supp08)) +
  geom_point(size = 3) +
  geom_vline(xintercept = 0) +
  geom_smooth(method = "lm", data = fig_data %>% filter(mean_inc < 0),
              se = FALSE, color = "black", linewidth = 0.75) +
  geom_smooth(method = "lm", data = fig_data %>% filter(mean_inc > 0),
              se = FALSE, color = "black", linewidth = 0.75) +
  labs(x = "Predicted Income",
       y = "Average Support for Current Government (2008)",
       title = "Figure 3: PANES Program Eliginility and Political Support for the Government,\n2008 Foll
       subtitle = "Figure Replication") +
  theme_bw() +
  theme(plot.title = element_text(face = "bold", size = 12),
        plot.subtitle = element_text(face = "italic"))
fig4
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



3.8 Question: Interpret these figures. What should we take away from them?

Answer: Figure 3 shows the difference in political support for the current government above and below the PANES threshold in 2007. Specifically, average support is higher among those below the threshold. Below the threshold, increased predicted income predicts a slight decrease in government support holding all else constant. Above the threshold, increased predicted income predicts a extremely slight increase in support for the current government. This makes sense since, above the threshold, higher predicted income households are likely well off enough to not particularly care about assistance programs while the households who are right along the threshold may feel like they are missing out.

The same interpretation can be made for Figure 4 except we are looking at support for the current government in 2008 rather than 2007.

3.9 Question: Replicate the results of the three regressions estimated in the first column of table 1. Present your results in a table. Interpret the coefficients.

Code:

```
reg11 <- felm(ever_received ~ eligibility, data = data1)
reg12 <- felm(gov_support07 ~ eligibility, data = data1)
reg13 <- felm(gov_support08 ~ eligibility, data = data1)

#Paper does not mention use of robust se
stargazer(reg11, reg12, reg13, se = list(reg11$se, reg12$se, reg13$se),
          type = "latex", no.space = TRUE, title = "Table 1 Col 1 Replication",
          header = FALSE)
```

Table 2: Table 1 Col 1 Replication

	Dependent variable:		
	ever_received	gov_support07	gov_support08
	(1)	(2)	(3)
eligibility	0.995***	0.129***	0.118***
	(0.002)	(0.012)	(0.014)
Constant	0.003**	0.772***	0.728***
	(0.001)	(0.009)	(0.011)
Observations	3,098	2,089	1,948
R^2	0.989	0.050	0.034
Adjusted R^2	0.989	0.049	0.033
Residual Std. Error	0.051 (df = 3096)	0.280 (df = 2087)	0.313 (df = 1946)
Note:			*p<0.1; **p<0.05; ***p<0.01

Answer: The first regression tells us that a house that is eligible for PANES benefits is predicted to have a statistically significant increase of 0.995 in the probability of ever receiving benefits. The constant tells us that a household that was ineligible for PANES had a base probability of receiving PANES of 0.003.

The second regression tells us that a household that is eligible for PANES benefits is predicted to have a statistically significant increase of 0.129 points (12.9 percentage points) in their support for the current government in 2007. The constant tells us that a household that was ineligible for PANES had a base support of the current government in 2007 of 0.772.

The second regression tells us that a household that is eligible for PANES benefits is predicted to have a statistically significant increase of 0.118 (11.8 percentage points) points in their support for the current government in 2008 The constant tells us that a household that was ineligible for PANES had a base support of the current government in 2008 of 0.728.

3.10 Question: Write down the specifications used in row 2 of columns 1,2 and 3 of table 1.

Answer:

Row 2 Column 1: $GovernmentSupport2007 = \beta_0 + \beta_1 * Eligible$

Row 2 Column 2: $GovernmentSupport2007 = \beta_0 + \beta_1 * (Eligible \times PredictedIncome)$

Row 2 Column 3: $GovernmentSupport2007 = \beta_0 + \beta_1 * (Eligible \times PredictedIncome^2)$

3.11 Question: Replicate the results reported in row 2 of Table 1 columns 1, 2 and 3. Explain the difference between these specifications and interpret their coefficients. (2 pages)

Hint: the variables listed in the table above after newtreat are the controls you will want to include.

Code:

```
reg21 <- reg12
reg22 <- felm(gov_support07 ~ eligibility + pred_inc + eligibility*pred_inc, data = data1)
reg23 <- felm(gov_support07 ~ eligibility + pred_inc + pred_inc_sq +
                eligibility*pred_inc + eligibility*pred_inc_sq, data = data1)

stargazer(reg21, reg22, reg23, se = list(reg21$se, reg22$se, reg23$se),
          type = "latex", no.space = TRUE, title = "Table 1 Row 2 Col 1-3 Replication",
          header = FALSE)
```

Table 3: Table 1 Row 2 Col 1-3 Replication

	Dependent variable:		
	gov_support07		
	(1)	(2)	(3)
eligibility	0.129***	0.110***	0.130***
	(0.012)	(0.026)	(0.040)
pred_inc		−0.011	0.812
		(1.646)	(6.736)
pred_inc_sq			−40.457
			(321.113)
eligibility:pred_inc		−1.916	2.377
		(2.170)	(9.072)
eligibility:pred_inc_sq			292.215
			(433.228)
Constant	0.772***	0.772***	0.769***
	(0.009)	(0.019)	(0.030)
Observations	2,089	2,089	2,089
R^2	0.050	0.051	0.051
Adjusted R^2	0.049	0.049	0.049
Residual Std. Error	0.280 (df = 2087)	0.280 (df = 2085)	0.280 (df = 2083)
Note:			*p<0.1; **p<0.05; ***p<0.01

Answer: Each subsequent specification adds a polynomial of `pred_inc` and an interaction term between that polynomial and `eligibility` (polynomials of 0, 1, and 2). It is important to note that the previous polynomial and interaction terms are included in the next regression.

The first regression tells us that a household that is eligible for PANES benefits is predicted to have a statistically significant increase of 0.129 points (12.9 percentage points) in their support for the current government in 2007.

The second regression tells us that a household that is eligible for PANES benefits is predicted to have a statistically significant increase of 0.110 points (11.0 percentage points) in their support for the current government in 2007 holding predicted income and interacting predicted income with eligibility.

The third regression tells us that a household that is eligible for PANES benefits is predicted to have a statistically significant increase of 0.130 points (13.0 percentage points) in their support for the current government in 2007 holding predicted income and squared predicted income and interacting predicted income with eligibility and squared predicted income with eligibility.

All regressions have a coefficient for `eligibility`that is significant at the 1% level, meaning we can reject the null hypothesis that there is no relationship between eligbility for PANES and support for the current government in 2007.

3.12 Question: What is the point of including all of these specifications?

Answer: These specifications allow us to control for any non-linear relationships between predicted income and government support (polynomial terms) and any relationships between `eligibility` and the various orders of predicted income (interaction terms). By doing so, we can confirm that the coefficient for `eligibility` is a valid estimate and we are not absorbing other types of effects on government support.

3.13 Question: Using the coefficients estimated above, write out the function you would use to predict the probability a household supports the current government based on their predicted income score:

a) If they are eligible for the transfer using the results from column 1.

$$gsupp07 = 0.772 + 0.129 \times 1 = 0.901$$

b) If they are not eligible for the transfer using the results from column 1.

$$gsupp07 = 0.772 + 0.129 \times 0 = 0.772$$

c) If they are eligible for the transfer using the results from column 2.

$$gsupp07 = 0.772 + 0.011 \times 1 - 0.011 = 1.916 \times 1 = -1.144$$

d) If they are not eligible for the transfer using the results from column 2.

$$gsupp07 = 0.772 + 0.011 \times 0 - 0.011 - 1.916 \times 0 = 0.761$$

e) If they are eligible for the transfer using the results from column 3.

$$gsupp07 = 0.769 + 0.130 \times 1 + 0.812 - 40.457 + 2.377 \times 1 + 292.25 \times 1 = 255.881$$

f) If they are not eligible for the transfer using the results from column 3.

$$gsupp07 = 0.769 + 0.230 \times 0 + 0.812 - 40.457 + 2.377 \times 0 + 292.25 \times 0 = 38.876$$

Answer:

3.14 Question: How narrow is the "bandwidth" used by the authors. Why does this matter? Check that the results are robust to a narrower bandwidth.

Code:

```
narrow_data <- data1[data1$pred_inc < 0.01 & data1$pred_inc > -0.01,]

narr_reg_rec <- felm(ever_received ~ eligibility, data = narrow_data)

stargazer(narr_reg_rec, se = list(narr_reg_rec$se),
          type = "latex", no.space = TRUE, title = "Robustness Check",
          header = FALSE)
```

Table 4: Robustness Check

	Dependent variable:
	ever_received
eligibility	0.992***
	(0.003)
Constant	0.005*
	(0.003)
Observations	1,511
R^2	0.984
Adjusted R^2	0.984
Residual Std. Error	0.063 (df = 1509)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
narr_reg_g070 <- felm(gov_support07 ~ eligibility, data = narrow_data)
narr_reg_g071 <- felm(gov_support07 ~ eligibility + pred_inc + eligibility*pred_inc, data = narrow_data)
narr_reg_g072 <- felm(gov_support07 ~ eligibility + pred_inc + pred_inc_sq+
                        eligibility*pred_inc + eligibility*pred_inc_sq, data = narrow_data)
```

```
## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or not positive definite
```

```
narr_reg_g08 <- felm(gov_support08 ~ eligibility, data = narrow_data)

stargazer(narr_reg_g070, narr_reg_g071, narr_reg_g072, narr_reg_g08,
          se = list(narr_reg_g070$se, narr_reg_g071$se, narr_reg_g072$se, narr_reg_g08$se),
          type = "latex", no.space = TRUE, title = "Robustness Check", header = FALSE)
```

Answer:

The authors use a bandwidth of +/- 0.02. If it is too big, the results become harder to justify because the observations on either side become more different from each other. The smaller threshold coefficients are still highly significant at the 1% level, meaning the results are robust to a narrower bandwidth.

Table 5: Robustness Check

	Dependent variable:			
	gov_support07			gov_support08
	(1)	(2)	(3)	(4)
eligibility	0.129***	0.114***	0.114***	0.121***
	(0.017)	(0.037)	(0.037)	(0.021)
pred_inc		−0.439	−0.439	
		(4.585)	(4.585)	
pred_inc_sq				
eligibility:pred_inc		−2.167	−2.167	
		(6.287)	(6.287)	
eligibility:pred_inc_sq				
Constant	0.769***	0.771***	0.771***	0.719***
	(0.013)	(0.028)	(0.028)	(0.016)
Observations	1,007	1,007	1,007	937
R^2	0.051	0.052	0.052	0.034
Adjusted R^2	0.051	0.049	0.049	0.033
Residual Std. Error	0.274 (df = 1005)	0.275 (df = 1003)	0.275 (df = 1003)	0.321 (df = 935)

Note: *p<0.1; **p<0.05; ***p<0.01

3.15 Question: The authors attribute these effects to the causal effect of receiving the government transfers. What is the implied assumption behind this interpretation?

Answer: There is an implied assumption of continuity around the threshold. Namely, the relationship, absent an effect of transfers, should be smooth and continuous around the threshold. They are also assuming observations around the threshold are as good as randomly assigned.

3.16 Question: What evidence do they provide to support this assumption?

Answer: They compare households above and below the threshold in the paper to show that they are not significantly different. They look at variables like age, education, income, and gener. They did not find a significant difference between any above or below the threshold. From these findings, they were able to conclude that the distribution of households around the threshold was effectively random.

3.17 Question: Was this threshold eligibility score specifically designed for this particular program? Why does this matter?

Answer: The score was specifically designed for this program. This allows researchers to know exactly where to set the threshold and ensures much stricter adherence to the assignmnet of benefits. It also protects the research design from exogenous effects like political considerations and allows for more statistically accurate results.

4 Submission instructions:

1) Knit your assignment in PDF (It should be 26 pages long).
2) Make sure you have ONE question and answer per page (this allows gradescope to easily find your answers).
3) Upload your assignment PDF to gradescope.