HW 4 Solutions

Frank Edwards

10/14/2019

Question 1

This analysis uses data from the Progresa conditional cash transfer program's implementation in Mexico around the 2000 presidential election. Precincts randomly gained eligibility for the program either 21 or 6 months before the election. We evaluate whether early eligibility increased votes for the incumbent PRI party, as we might expect that gaining access to a new cash-transfer program would buttress suppport for the incumbent political party prior to the election.

This is inline 2^n.

This is not inline

$$2^n = something x^3$$

```
control<-progresa %>%
  filter(treatment==0) %>%
  summarise(pri2000s = mean(pri2000s),
            t2000 = mean(t2000))

treatment_means<-progresa %>%
  filter(treatment==1) %>%
  summarise(pri2000s = mean(pri2000s),
            t2000 = mean(t2000)) %>%
  mutate(pri2000s = pri2000s - control$pri2000s,
         t2000 = t2000 - control$t2000)

pander(treatment_means)
```

pri2000s	t2000
3.622	4.27

When comparing means in the treatment group, which received the CCT program over a year before the control groups, we see a substantial difference in turnout and PRI support. The precincts receiving early Progresa implementation had turnout that was 3.6224955 points higher than late Progresa precincts. Early Progresa precincts had turnout that was 4.269676 points higher than late Progresa precicnts.

Next, I estimate a linear regression for each of the outcomes with the treatment as a predictor.

```
m_pri<-lm(pri2000s ~ treatment,
          data = progresa)

pander(coef(m_pri))
```

(Intercept)	treatment
34.49	3.622

The intercept term is equal to the mean value of PRI support across all precincts, and the treatment term is equal to the difference in PRI support between early and late Progresa precincts. Early Progresa precincts had higher levels of PRI support than late Progresa precicnts.

```
m_turn<-lm(t2000 ~ treatment,
           data = progresa)

pander(coef(m_turn))
```

(Intercept)	treatment
63.81	4.27

As before, when examining turnout, the regression results are identical to a comparison of means. In general, a single binary predictor will yield coefficients that are equal to the group means for each of the binary groups.

The results do support the author's original hypothesis. Early Progresa implementation is associated with higher turnout and higher support for the PRI in the 2000 election.

Question 2

```
m_pri_1<-lm(pri2000s ~ treatment +
            avgpoverty + pobtot1994 +
            votos1994 + pri1994 + pan1994 + prd1994,
          data = progresa)

m_turn_1<-lm(t2000 ~ treatment +
            avgpoverty + pobtot1994 +
            votos1994 + pri1994 + pan1994 + prd1994,
          data = progresa)
```

I estimate more complex models that include parameters for precinct poverty, population size, and prior election vote turnouts and vote shares (1994). This model estimates that, after controlling for population, poverty, and prior voting early Progresa precincts had 2.93 points higher PRI support than late Progresa precincts. This estimate is 0.69 points lower than the initial model.

For turnout, this model with additional predictors shows that precincts with early Progresa precincts had, on average, 4.55 points higher turnout than late Progesa precincts, holding population, poverty, and prior election results constant. This result is 0.28 points higher than the simpler model.

Question 3

I modify the models in question 2 to include more reasonable predictors; shares rather than counts for party votes, and a log transformation on population to adjust for the extreme skew in the distribution of precinct populations.

```
m_pri_2<-lm(pri2000s ~ treatment +
            avgpoverty + log(pobtot1994) +
            t1994 + pri1994s + pan1994s + prd1994s,
          data = progresa)

m_turn_2<-lm(t2000 ~ treatment +
            avgpoverty + log(pobtot1994) +
            t1994 + pri1994s + pan1994s + prd1994s,
          data = progresa)
```

```r
### compute differences in coefficients

out<-data.frame(model = c("Q2","Q3"),
                treatment_pri = c(coef(m_pri_1)[2], coef(m_pri_2)[2]),
                treatment_turn = c(coef(m_turn_1)[2], coef(m_turn_2)[2]))

pander(out)
```

model	treatment_pri	treatment_turn
Q2	2.928	4.549
Q3	0.2355	-0.153

As shown above, model 3 estimates substantially lower coefficients for the treatment variables than does model 2. The more reasonable model specification suggests that the control variables may be accounting for much of the variation in the differences in electoral outcomes across treatment and control precincts.

```r
## calculate coefficient of determination (R^2)
## see page 156
r2<-function(y, model){
  TSS<-sum((y - mean(y))^2)
  SSR<-sum(resid(model)^2)
  r2_out<-1-SSR/TSS
  return(r2_out)
}

### format table for output
out<-data.frame(model = c("Q2","Q3"),
                r2_pri = c(r2(progresa$pri2000s, m_pri_1),
                           r2(progresa$pri2000s, m_pri_2)),
                r2_turn = c(r2(progresa$t2000, m_turn_1),
                            r2(progresa$t2000, m_turn_2)))

pander(out)
```

model	r2_pri	r2_turn
Q2	0.2206	0.0785
Q3	0.5794	0.6921

The table above shows the coefficients of determination for the model specification in question 2, and the model specification in model 3. The model used in this question provides a dramatically stronger fit for the data than did the model in question 3.

The coefficient of determination indicates the proportion of total variance in the outcome explained by the regression model. It shows that in the model from question 2 for PRI vote share, the model explains about 22.0591131 percent of the variance in PRI vote share, while in the question 3 model, we explain 22.0591131 percent of the variance. The increase in goodness-of-fit for turnout is even more extreme.

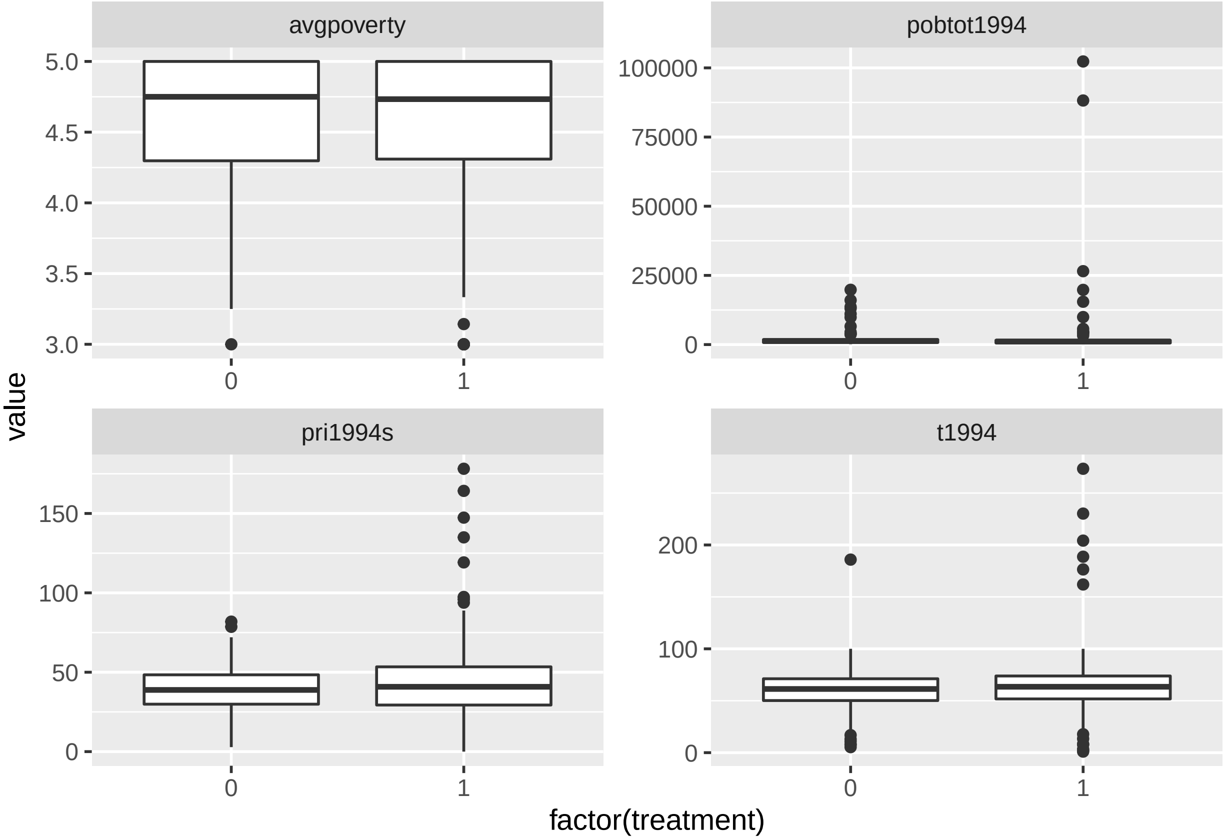
Question 4

Was assignment to treatment truly random? Here, I explore the difference in each of the focal predictors across the early and late Progresa precincts. The boxplot below visualizes these distributions.

```
### gather all variables into treatment, variable, value structure
### for facet plotting

plot_dat<-progresa %>%
  select(treatment, t1994, pri1994s, avgpoverty, pobtot1994) %>%
  gather(key = "variable", value = "value", -treatment)

ggplot(plot_dat,
       aes(x = factor(treatment), y = value)) +
  geom_boxplot() +
  facet_wrap(~variable, scales = "free")
```



The boxplot doesn't clearly indicate large differences between the treatment and control group, although very high PRI support districts do appear to more often be in the treatment group.

Question 5

```
m_pri_3<-lm(pri2000v ~ treatment +
            avgpoverty + log(pobtot1994) +
            t1994r + pri1994v + pan1994v + prd1994v,
          data = progresa)

m_turn_3<-lm(t2000r ~ treatment +
            avgpoverty + log(pobtot1994) +
            t1994r + pri1994v + pan1994v + prd1994v,
          data = progresa)
```

```
### new coefficients
## PRI vote share, official
coef(m_pri_3)[2]
```

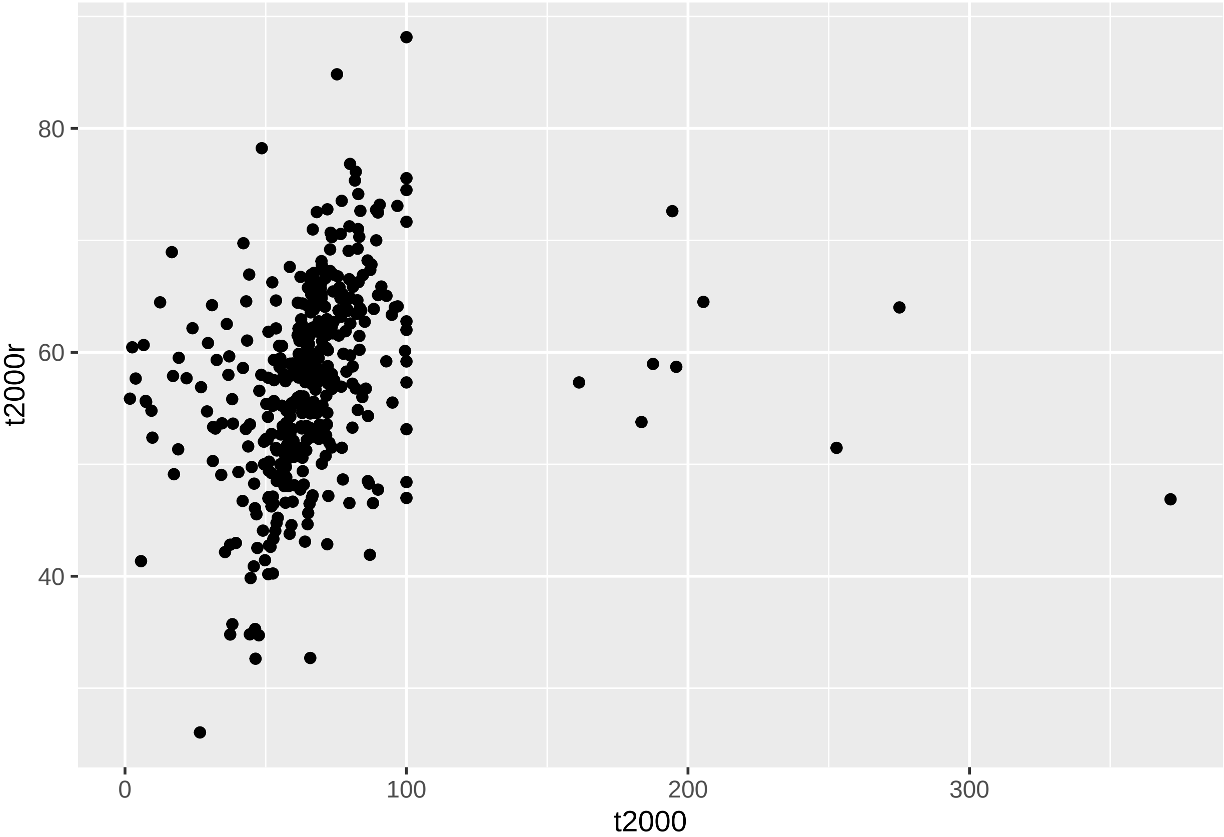
```
## treatment
## 0.8017207
```

```
## Turnout, official
coef(m_turn_3)[2]
```

```
## treatment
## -1.080943
```

When using the official turnout and vote share measures, we see that early Progresa is estimated to slightly increase PRI vote share in the average precinct, and slightly decrease turnout in the average precinct. These results are similar to the results obtained in the model for question 3.

```
ggplot(progresa,
       aes(x = t2000, y = t2000r)) +
  geom_point()
```



I plot the official and poplulation size turnout measures against each other above in the scatterplot. These two variables are only weakly correlated with each other. Given that they should be measuring the same underlying construct - voter turnout, this is troubling.

Question 6

In this question, I consider whether the effect of early Progresa varies across precincts as a function of poverty. To do so, I include interactions between early Progresa and precinct poverty, modeled as both a linear and quadratic term for poverty.

```
m_pri_4<-lm(pri2000v ~
              treatment * (avgpoverty + I(avgpoverty^2))+
              log(pobtot1994),
            data = progresa)

m_turn_4<-lm(t2000r ~
              treatment * (avgpoverty + I(avgpoverty^2))+
              log(pobtot1994),
             data = progresa)

progresa<- progresa %>%
  mutate(yhat_pri = fitted(m_pri_4),
         yhat_turn = fitted(m_turn_4))

# ggplot(progresa,
#        aes(x = avgpoverty, y = yhat_pri, color = factor(treatment))) +
#   geom_point(alpha =0.5)
#
#
#
#
# , color = 2, alpha =  0.5,
#              position = position_jitter(width = 0.1)) +
#   geom_point(aes(y = yhat_turn), color =3, alpha = 0.5,
#              position = position_jitter(width = 0.1)) +
#   xlab("Observed poverty (jittered)") +
#   ylab("Predicted values") +
#   labs(title = "Predicted turnout (green) and PRI vote share (red)",
#        subtitle = "Facetted by timing of Progresa exposure, 1 = Early") +
```

This figure shows the results of models include terms that allow the timing of Progresa exposure to have variable impacts on turnout and PRI voteshare in low and high poverty precincts. This model also allows the relationship between poverty, the treatment, and the outcome to have a non-linear relationship, either allowing for a tapering of the relationship for high poverty, or a greater impact of the treatment in high poverty zones that a linear relationship would expect.

The figure shows the predicted impact of early Progresa based on the described model specification. Based on this visual, we do see that poverty is closely related to PRI vote share in both treatment and control, but that the exposure to treatment (early Progresa) does not appear to have much of a predicted impact on turnout or vote share.

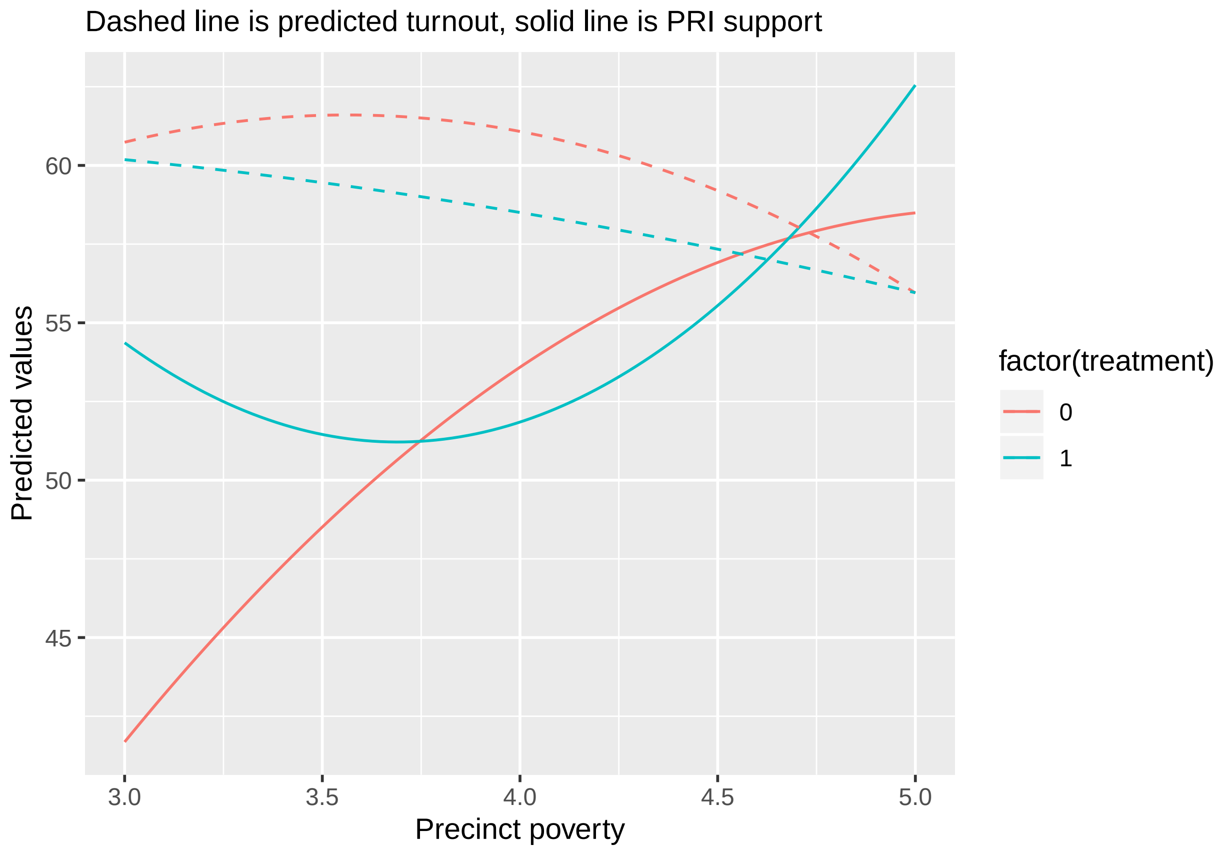
The plot below illustrates the shape of the estimated regression curve for both outcomes, PRI support and voter turnout. I plot the Progresa timing variable in color, and indicate the outcome with variable line types, solid for PRI support and dashed for turnout. As above, the models don't provide clear support for the hypothesis - exposure to early Progresa did not have a dramatic effect on voter behavior. Poverty is a strong predictor of PRI support, however.

```
poverty<-seq(min(progresa$avgpoverty),
             max(progresa$avgpoverty),
             length.out = 100)

predict_data<-data.frame(avgpoverty = rep(poverty, 2),
                    pobtot1994 = rep(rep(mean(progresa$pobtot1994,100)),2),
                    treatment = rep(c(0,1), each =100))
```

```
predict_data<-predict_data %>%
  mutate(yhat_pri = predict(m_pri_4, newdata = predict_data),
         yhat_turnout = predict(m_turn_4, newdata = predict_data))

ggplot(predict_data,
  aes(x=avgpoverty, color = factor(treatment))) +
  geom_line(aes(y=yhat_pri), lty=1) +
  geom_line(aes(y=yhat_turnout), lty=2) +
  labs(subtitle = "Dashed line is predicted turnout, solid line is PRI support") +
  ylab("Predicted values") +
  xlab("Precinct poverty")
```


Dashed line is predicted turnout, solid line is PRI support
Predicted values
Precinct poverty
factor(treatment)