Chapter 12 Problem Set

Tianwei Liu

12/4/2019

2. Public attitudes toward global warming influence the policy response to the issue. The data set EnvSurvey.dta provides data from a nationally representative survey of the U.S. public that asked multiple questions about the environment and energy. Table 12.8 lists the variables.

```
load('Ch12_Exercise2_Global_warming.Rdata')
```

(a)Use an LPM to estimate the probability of saying that global warming is real and caused by humans (the dependent variable is HumanCause2). Control for sex, being white, education, income, age, and partisan identification.

```
reg2a <- lm(humancause ~ male + white + educ + incomecat + age + party7, data = dta)
summary(reg2a)
```

```
##
## Call:
## lm(formula = humancause ~ male + white + educ + incomecat + age +
##     party7, data = dta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6949 -0.3311 -0.1434  0.4424  1.0261
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.3840485  0.0768832  -4.995 6.43e-07 ***
## male         0.0206509  0.0203616   1.014    0.311
## white        0.0365759  0.0251690   1.453    0.146
## educ         0.0265781  0.0057513   4.621 4.08e-06 ***
## incomecat    0.0031728  0.0025808   1.229    0.219
## age         -0.0010053  0.0005982  -1.681    0.093 .
## party7       0.0867771  0.0050978  17.022  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4356 on 1848 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.1527, Adjusted R-squared:  0.1499
## F-statistic: 55.49 on 6 and 1848 DF,  p-value: < 2.2e-16
```

(i)Which variable has the most important influence on this opinion? Why?

partisan identification has the most important influence on this opinion. The first reason is that the coefficient on party7 is highly statistically significant (p-value is very very small), therefore, we are very confident in reject the null and conclude that partisan identification has an effect on the opinion.

(ii)What are the minimum and maximum fitted values from this model? Discuss implications briefly.

```r
min(predict(reg2a))
```

```
## [1] -0.1965759
```

```r
max(predict(reg2a))
```

```
## [1] 0.7463568
```

- The minimum fitted value from this model is -0.197. This value doesn't make sense because probability should be between 0 and 1.
- The maximum fitted value from this model is 0.746, meaning that in this model the highest probability of a person believing global warming is real and caused by human is 74.6%.

(iii)Add age-squared to the model. What is the effect of age? Use a simple sketch if necessary, with key point(s) identified.

```r
reg2a2 <- lm(humancause ~ male + white + educ + incomecat + age + I(age^2) + party7, data = dta)
summary(reg2a2)
```

```
##
## Call:
## lm(formula = humancause ~ male + white + educ + incomecat + age +
##     I(age^2) + party7, data = dta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7124 -0.3338 -0.1501  0.4316  1.0550
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.157e-01  1.017e-01  -2.121  0.03407 *
## male         1.978e-02  2.034e-02   0.973  0.33092
## white        3.506e-02  2.514e-02   1.395  0.16330
## educ         2.718e-02  5.748e-03   4.728 2.44e-06 ***
## incomecat    3.451e-03  2.579e-03   1.338  0.18112
## age         -9.293e-03  3.338e-03  -2.784  0.00542 **
## I(age^2)     8.433e-05  3.342e-05   2.524  0.01170 *
## party7       8.699e-02  5.091e-03  17.086  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4349 on 1847 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.1556, Adjusted R-squared:  0.1524
## F-statistic: 48.61 on 7 and 1847 DF,  p-value: < 2.2e-16
```

The effect of age is a function of age: it is the derivative of $-0.00929 * age + 0.0000843 * age^2$ which is equal to $-0.00929 + 0.000169 * age$.

(b)Use a probit model to estimate the probability of saying that global warming is real and caused by humans (the dependent variable is HumanCause2). Use the independent variables from part (a), including the age-squared variable.

```
reg2b <- glm(humancause ~ male + white + educ + incomecat + age + agesq + party7, family = binomial(lin
summary(reg2b)
```

```
##
## Call:
## glm(formula = humancause ~ male + white + educ + incomecat +
##     age + agesq + party7, family = binomial(link = "probit"),
##     data = dta)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6551  -0.8644  -0.5602   1.0361   2.4695
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.1189279  0.3244321  -6.531 6.52e-11 ***
## male         0.0627495  0.0642795   0.976  0.32897
## white        0.1038861  0.0777799   1.336  0.18167
## educ         0.0810659  0.0185483   4.371 1.24e-05 ***
## incomecat    0.0097614  0.0081337   1.200  0.23009
## age         -0.0286358  0.0104961  -2.728  0.00637 **
## agesq        0.0002569  0.0001055   2.436  0.01486 *
## party7       0.2646544  0.0169363  15.626  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2367.9  on 1854  degrees of freedom
## Residual deviance: 2061.8  on 1847  degrees of freedom
##   (17 observations deleted due to missingness)
## AIC: 2077.8
##
## Number of Fisher Scoring iterations: 4
```

(i)Compare statistical significance with LPM results.

Coefficients preserve statistical significance across models. In both LPM and Probit model, education and partisan identification are strongly statistically significant, with p-values smaller than 0.001 level of significance. Age and age-squared are both significant across the two models.

(ii)What are the minimum and maximum fitted values from this model? Discuss implications briefly.

```
min(predict(reg2b))
```

```
## [1] -2.033224
```

```
max(predict(reg2b))
```

```
## [1] 0.83148
```

```r
pnorm(-2.0332)
```

```
## [1] 0.02101616
```

```r
pnorm(0.83148)
```

```
## [1] 0.7971487
```

The minimum and maximum fitted values are not probabilities. The minimum/maximum probability in our model is calculated using the CDF ($\Pr(\text{min/max}) = \phi$ (min/max fitted values)). Given our calculation above, the model predicts the minimum probability is 0.021 and the maximum probability is 0.797.

(iii)Use the observed-value, discrete-differences approach to indicate the effect of partisan identification on the probability of saying global warming is real and caused by humans. For simplicity, simulate the effect of an increase of one unit on this seven-point scale (as opposed to the effect of one standard deviation, as we have done for continuous variables in other cases). Compare to LPM and "marginal-effects" interpretations.

```r
## Observed-value, discrete differences approach
p11 = pnorm(reg2b$coefficients[1] + reg2b$coefficients[2]*dta$male + reg2b$coefficients[3]*dta$white + r

p21 = pnorm(reg2b$coefficients[1] + reg2b$coefficients[2]*dta$male + reg2b$coefficients[3]*dta$white + r

describe(p21-p11)
```

```
## p21 - p11 :  Male  Format:%8.0g
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1872        0     1845        1  0.08673   0.0202  0.04835  0.05725
##      .25      .50      .75      .90      .95
##  0.07502  0.09289  0.10269  0.10499  0.10519
##
## lowest : 0.01746790 0.02287502 0.02712239 0.02924676 0.02950951
## highest: 0.10527429 0.10527434 0.10527438 0.10527443 0.10527448
```

```r
## Marginal effects approach
probitmfx(formula = humancause ~ male + white + educ + incomecat + age + agesq + party7, data = dta, atr
```

```
## Call:
## probitmfx(formula = humancause ~ male + white + educ + incomecat +
##     age + agesq + party7, data = dta, atmean = FALSE)
##
## Marginal Effects:
##                 dF/dx   Std. Err.       z     P>|z|
## male       1.9757e-02  2.0207e-02  0.9777  0.328210
## white      3.2300e-02  2.3826e-02  1.3557  0.175196
## educ       2.5552e-02  5.7626e-03  4.4342 9.243e-06 ***
## incomecat  3.0768e-03  2.5607e-03  1.2015  0.229539
## age       -9.0262e-03  3.2900e-03 -2.7435  0.006079 **
## agesq      8.0983e-05  3.3099e-05  2.4467  0.014416 *
## party7     8.3421e-02  4.2508e-03 19.6246 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## dF/dx is for discrete change for the following variables:
##
## [1] "male"  "white"
```

The coefficient of the party7 variable in the LPM model is 0.087; the effect of partisan identification is 0.0867 in the observed values, discrete-differences model and 0.834 in the marginal effects model. The effects estimated using the three models are essentially the same.

(iv)Use the observed-value, discrete-differences approach to indicate the effect of being male on the probability of saying global warming is real and caused by humans. Compare to LPM and "marginal-effects" interpretations.

```
## Observed-value, discrete differences approach
p12 = pnorm(reg2b$coefficients[1] + reg2b$coefficients[2]*0 + reg2b$coefficients[3]*dta$white + reg2b$c

p22 = pnorm(reg2b$coefficients[1] + reg2b$coefficients[2]*1 + reg2b$coefficients[3]*dta$white + reg2b$c

describe(p22-p12)
```

```
## p22 - p12 :  White  Format:%8.0g
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1872        0     1827        1  0.01976 0.005738 0.009578 0.011696
##      .25      .50      .75      .90      .95
## 0.015938 0.021658 0.024514 0.024926 0.025008
##
## lowest : 0.003377057 0.003994542 0.004819413 0.005238834 0.005465260
## highest: 0.025029302 0.025029313 0.025029323 0.025029330 0.025029331
```

The marginal effect model is shown in the previous question.

The coefficient of the male variable in the LPM model is 0.0198 (difference of means because male is a dummy variable); the effect of male is 0.01976 in the observed values, discrete-differences model and 0.01976 in the marginal effects model. The effects estimated using the three models are the same, this is because for dummy variables, the marginal effects model also uses observed values, discrete-differences approach to calculate the marginal effect of that specific dummy variable.

(c)The survey described in this item also included a survey experiment in which respondents were randomly assigned to different question wordings for an additional question about global warming. The idea was to see which frames were most likely to lead people to agree that the earth is getting warmer. The variable we analyze here is called WarmAgree. It records whether respondents agreed that the earth's average temperature is rising. The experimental treatment consisted of four different ways to phrase the question.

- The variable Treatment equals 1 for people who were asked "Based on your personal experiences and observations, do you agree or disagree with the following statement: The average temperature on earth is getting warmer."

- The variable Treatment equals 2 for people who were given the following information before being asked if they agreed that the average temperature of the earth is getting warmer: "The following figure [Figure 12.10] shows the average global temperature compared to the average temperature from 1951–1980. The temperature analysis comes from weather data from more than 1,000 meteorological stations around the world, satellite observations of sea surface temperature, and Antarctic research station measurements."

- The variable Treatment equals 3 for people who were given the following information before being asked if they agreed that average temperature of the earth is getting warmer: "Scientists working at the National Aeronautics and Space Administration (NASA) have concluded that the average global temperature has

increased by about a half degree Celsius compared to the average temperature from 1951–1980. The temperature analysis comes from weather data from more than 1,000 meteorological stations around the world, satellite observations of sea surface temperature, and Antarctic research station measurements."

• The variable Treatment equals 4 for people who were simply asked "Do you agree or disagree with the following statement: The average temperature on earth is getting warmer." This is the control group.

Which frame was most effective in affecting opinion about global warming?

```
dta$treatment1 = dta$treatment == 1
dta$treatment2 = dta$treatment == 2
dta$treatment3 = dta$treatment == 3
reg3 <- glm(warmagree ~ treatment1 + treatment2 +treatment3, family = binomial(link='probit'), data = d
summary(reg3)
```

```
##
## Call:
## glm(formula = warmagree ~ treatment1 + treatment2 + treatment3,
##     family = binomial(link = "probit"), data = dta)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6459  -1.4641   0.8573   0.8811   0.9156
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.40591    0.06150   6.600 4.12e-11 ***
## treatment1TRUE  0.09695    0.08595   1.128  0.25933
## treatment2TRUE  0.24341    0.08790   2.769  0.00562 **
## treatment3TRUE  0.05698    0.08523   0.669  0.50378
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2309.3  on 1871  degrees of freedom
## Residual deviance: 2300.9  on 1868  degrees of freedom
## AIC: 2308.9
##
## Number of Fisher Scoring iterations: 4
```

```
probitmfx(formula = warmagree ~ treatment1 + treatment2 +treatment3, data = dta, atmean = FALSE)
```

```
## Call:
## probitmfx(formula = warmagree ~ treatment1 + treatment2 + treatment3,
##     data = dta, atmean = FALSE)
##
## Marginal Effects:
##                    dF/dx Std. Err.      z    P>|z|
## treatment1TRUE 0.033493  0.029270 1.1443 0.252505
## treatment2TRUE 0.082493  0.028604 2.8839 0.003928 **
## treatment3TRUE 0.019791  0.029362 0.6740 0.500301
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## dF/dx is for discrete change for the following variables:
##
## [1] "treatment1TRUE" "treatment2TRUE" "treatment3TRUE"
```

- Choosing treatment question 4 as the reference category, i ran two models. The first is the probit model, and the second is the marginal effect model. In both models, the coefficients on all variables (treatment1, treatment2, and treatment3) are positive, which means that these three treatment questions outperform statement 4 (the probability of saying yes to the other three questions are higher than to question 4).
- Take a closer look, I found that the coefficient on treatment2 is statistically significant at 0.01 level for both models (as the p-value associated with treatment2 is smaller than 0.01). This means that the probability of saying yes to treatment question 4 is significantly higher than to question 4. Therefore, the second frame (treatment2) is most effective in affecting opinion on global warming. By our marginal effects model, respondents are 8.2% more likely to say yes to the second frame than the last frame.

4.Are members of Congress more likely to meet with donors than with mere constituents? To answer this question, Kalla and Broockman (2015) conducted a field experiment in which they had political activists attempt to schedule meetings with 191 congressional offices regarding efforts to ban a potentially harmful chemical. The messages the activists sent out were randomized. Some messages described the people requesting the meeting as "local constituents," and others described the people requesting the meeting as "local campaign donors." Table 12.10 describes two key variables from the experiment.

```
load('Ch12_Exercise4_Congress_donors.RData')
```

(a)Before we analyze the experimental data, let's suppose we were to conduct an observational study of access based on a sample of Americans and ran a regression in which the dependent variable indicates having met with a member of Congress and the independent variable was whether the individual donated money to a member of Congress. Would there be concerns about endogeneity? If so, why?

There shouldn't be concerns about endogeneity, because this study is a RCT. In RCTs, by definition, the X variables are independent of the error term.

(b)Use a probit model to estimate the effect of the donor treatment condition on probability of meeting with a member of Congress. Interpret the results.

```
dta$meetMC <- dta$staffrank == 5
reg4 <- glm(meetMC ~ treat_donor, family = binomial(link='probit'), data = dta)
summary(reg4)
```

```
##
## Call:
## glm(formula = meetMC ~ treat_donor, family = binomial(link = "probit"),
##     data = dta)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4033  -0.4033  -0.2187  -0.2187   2.7370
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.9841     0.2418  -8.205 2.31e-16 ***
## treat_donor   0.5663     0.3336   1.698   0.0895 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 66.425  on 190  degrees of freedom
## Residual deviance: 63.495  on 189  degrees of freedom
## AIC: 67.495
##
## Number of Fisher Scoring iterations: 6
```

```
pnorm(reg4$coefficients[1])
```

```
## (Intercept)
##  0.02362205
```

```
pnorm(reg4$coefficients[1] + reg4$coefficients[2])
```

```
## (Intercept)
##    0.078125
```

The probit model tells us the probability of meeting with a member of congress for local constituents is 2.4%, and the probability of meeting with a congress member for campaign donors is 7.8%. However, as the coefficient on donor treatment variable is not statistically significant at the conventional 0.05 level (p-value is $0.09 > 0.05$), the difference between the two groups (donors and non-donors) is not significant.

(c)What factors are missing from the model? What does this omission mean for our results?

- donation amount. Congress members are more likely to meet with donors who have donated a larger amount because they are the people that are important for him/her.
- Omitting donation amount would give rise to omitted variable bias. OVB would bias our estimates.

(d)Use an LPM to make your estimate. Interpret the results. Assess the correlation of the fitted values from the probit model and LPM.

```
reg4b <- lm(meetMC ~ treat_donor, data = dta)
summary(reg4b)
```

```
##
## Call:
## lm(formula = meetMC ~ treat_donor, data = dta)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07812 -0.07812 -0.02362 -0.02362  0.97638
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02362    0.01772   1.333   0.1842
## treat_donor  0.05450    0.03062   1.780   0.0766 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1997 on 189 degrees of freedom
## Multiple R-squared:  0.01649,    Adjusted R-squared:  0.01129
## F-statistic: 3.169 on 1 and 189 DF,  p-value: 0.07664
```

Intepretation: being a donar is associated with 0.0545 higher likelihood of meeting with a member of congress than non-donor constituents. The difference is consistent with the prediction of the difference in probability of the Probit model.

(e)Use an LPM to assess the probability of meeting with a senior staffer (defined as staffrank > 2).

```
dta$meetSenior <- dta$staffrank > 2
reg5 <- lm (meetSenior ~ treat_donor, data = dta)
summary(reg5)
```

```
##
## Call:
## lm(formula = meetSenior ~ treat_donor, data = dta)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18750 -0.18750 -0.05512 -0.05512  0.94488
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.05512    0.02611   2.111  0.03609 *
## treat_donor  0.13238    0.04511   2.935  0.00375 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2942 on 189 degrees of freedom
## Multiple R-squared:  0.04359,    Adjusted R-squared:  0.03853
## F-statistic: 8.613 on 1 and 189 DF,  p-value: 0.003751
```

The probability of meeting a senior staffer for local constituents (non-donors) is 5.5%, and the probability of meeting a senior staffer for donors is $5.5\% + 13.2\% = 18.7\%$. As p-value is less than 0.01, the coefficient on donor is statistically significant, the probability of meeting a senior staffer for donors is significantly higher than the probability for non-donors.

(f)Use an LPM to assess the probability of meeting with a low-level staffer (defined staffrank $= 1$).

```
dta$meetlow <- dta$staffrank == 1
reg6 <- lm (meetlow ~ treat_donor, data = dta)
summary(reg6)
```

```
##
## Call:
## lm(formula = meetlow ~ treat_donor, data = dta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.1260 -0.1260 -0.1260 -0.1094  0.8906
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.12598    0.02902   4.341 2.31e-05 ***
## treat_donor -0.01661    0.05014  -0.331    0.741
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3271 on 189 degrees of freedom
## Multiple R-squared:  0.0005803,  Adjusted R-squared:  -0.004708
## F-statistic: 0.1097 on 1 and 189 DF,  p-value: 0.7408
```

This model tells us the probability for local constituents (non-donors) of meeting with a low-level staffer is 12.6%. The coefficient on donor dummy variable is -0.0166 but it is not significant becasue the p-value associated with the coefficient is 0.74, larger than the conventional 0.05 level of significance. This means being a donor does not really affect the probability of meeting with a low-level staff.

POLICY MEMO

"Ban the Box" (BTB) policies restrict employers from asking about applicants' criminal histories on job applications and are often presented as a means of reducing unemployment among black men, who disproportionately have criminal records.To assess the impact of BTB, scholars submitted online job applications on behalf of fictitious job applicants to low-skill, entry-level job openings both before and after BTB went into effect in New Jersey and New York City. Each application signaled that the applicant was either black or white by using names that are strongly linked to one race or the other. Race and other characteristics (such as education or whether someone had committed a crime) were all randomly assigned when crafting the application.

Question: Did the difference between call-back rates for Black and White applications change after the "ban the box" policy went into effect? The data is available in BanTheBox.dta.
Here is a subset of variables that we will use. You may not need all of them; you won't need additional variables. - response: Application rec'd positive response - crimbox: Application has Box - black: Applicant is Black - white: Applicant is White - ged: Applicant has GED (equivalent to graduating from high school) - crime: Applicant has criminal record - box_white: Box x White - nocrime_box: Applicant has no criminal reord x Box

```
## Load the dataset
btb <- read_dta('BanTheBox.dta')
```

Model 1(given BTB is passed): response ~ black + ged

```
## LPM
reg7 <- lm(response ~ black + ged , data = btb[btb$crimbox == 0, ])
summary(reg7)
```

```
##
## Call:
## lm(formula = response ~ black + ged, data = btb[btb$crimbox ==
##     0, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.1373 -0.1336 -0.1070 -0.1032  0.8968
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.137293   0.005220  26.302  < 2e-16 ***
## black       -0.030341   0.005981  -5.073 3.97e-07 ***
## ged         -0.003719   0.005981  -0.622    0.534
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3249 on 11803 degrees of freedom
## Multiple R-squared:  0.002201,   Adjusted R-squared:  0.002032
## F-statistic: 13.02 on 2 and 11803 DF,  p-value: 2.252e-06
```

```
## Probit model
reg7b <- glm (response ~ black + ged, family = binomial(link='probit'), data = btb[btb$crimbox == 0, ])
summary(reg7b)
```

```
##
## Call:
## glm(formula = response ~ black + ged, family = binomial(link = "probit"),
##     data = btb[btb$crimbox == 0, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5438  -0.5352  -0.4752  -0.4672   2.1303
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.09182    0.02550 -42.821  < 2e-16 ***
## black       -0.15199    0.03000  -5.066 4.07e-07 ***
## ged         -0.01857    0.02994  -0.620    0.535
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8672.9  on 11805  degrees of freedom
## Residual deviance: 8646.9  on 11803  degrees of freedom
## AIC: 8652.9
##
## Number of Fisher Scoring iterations: 4
```

```
pnorm(reg7b$coefficients[1])
```

```
## (Intercept)
##    0.137456
```

```
pnorm(reg7b$coefficients[1] + reg7b$coefficients[2])
```

```
## (Intercept)
##   0.1067841
```

- After the BTB policy is passed, employers do not have information on applicant's criminal history, therefore I did not include crime as an independent variable, because this is not relevant for employers' decision. I run a LPM to see whether crime is an irrelevant variable or not in this case.

```
reg7_ir <- lm(response ~ black + crime + ged , data = btb[btb$crimbox == 0, ])
summary(reg7_ir)
```

```
##
## Call:
## lm(formula = response ~ black + crime + ged, data = btb[btb$crimbox ==
##     0, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.1412 -0.1337 -0.1107 -0.1032  0.9006
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.141228   0.006093  23.178  < 2e-16 ***
## black       -0.030543   0.005983  -5.105 3.35e-07 ***
## crime       -0.007490   0.005983  -1.252    0.211
## ged         -0.003811   0.005981  -0.637    0.524
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3249 on 11802 degrees of freedom
## Multiple R-squared:  0.002333,   Adjusted R-squared:  0.00208
## F-statistic: 9.201 on 3 and 11802 DF,  p-value: 4.464e-06
```

- As we see in the above model, crime is an irrelevant variable because we fail to reject the null hypothesis that crime has an effect (p-value associated with crime is 0.21, which is greater than the conventional 0.05 level of significance). Therefore, I excluded crime in my analysis.
- In the probit model, white applicants have roughly probability of 13.7% to receive a call-back while black applicants's probability of receiving a call-back is 10.7%. The 3% difference is consistent with the coefficient on black dummy variable in the LPM model. As this coefficient is very statistically significant (p-value smaller than 0.001 level of significance), we conclude that there is a difference between black and white candidates, and that black candidates are 3% less likely to receive a call-back from employers.

Model 2(given BTB is still in practice and the box is still required in application): response ~ black + crime + ged

```
reg72 <- lm(response ~ black + crime + ged, data =btb[btb$crimbox == 1, ])
summary(reg72)
```

```
##
## Call:
## lm(formula = response ~ black + crime + ged, data = btb[btb$crimbox ==
##     1, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14388 -0.12607 -0.09404 -0.07624  0.92472
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1260731  0.0112675  11.189  < 2e-16 ***
## black       -0.0009548  0.0113394  -0.084    0.933
## crime       -0.0498378  0.0113400  -4.395 1.15e-05 ***
## ged          0.0178038  0.0113456   1.569    0.117
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3108 on 3003 degrees of freedom
## Multiple R-squared:  0.00732,    Adjusted R-squared:  0.006328
## F-statistic: 7.381 on 3 and 3003 DF,  p-value: 6.32e-05
```

```
## Probit model
reg72b <- glm (response ~ black + crime + ged, family = binomial(link='probit'), data = btb[btb$crimbox
summary(reg72b)
```

```
##
## Call:
## glm(formula = response ~ black + crime + ged, family = binomial(link = "probit"),
##     data = btb[btb$crimbox == 1, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5600  -0.5158  -0.4400  -0.4018   2.2643
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.152502   0.059554 -19.352  < 2e-16 ***
## black       -0.003445   0.061211  -0.056    0.955
## crime       -0.269354   0.061641  -4.370 1.24e-05 ***
## ged          0.094843   0.061230   1.549    0.121
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2072.3  on 3006  degrees of freedom
## Residual deviance: 2050.2  on 3003  degrees of freedom
## AIC: 2058.2
##
## Number of Fisher Scoring iterations: 4
```

```
pnorm(reg72b$coefficients[1])
```

```
## (Intercept)
##   0.1245575
```

```
pnorm(reg72b$coefficients[1] + reg72b$coefficients[3])
```

```
## (Intercept)
##  0.07753408
```

- Before the BTB policy is enacted, applicants are required to provide their criminal information, therefore in this model crime is a relevant variable.
- In this model given crime box is on application, race starts to play a little effect. Across both the LPM and the Probit model, the coefficient black is not statistically significant, with an almost 0 z-value. Therefore, we fail to reject the null hypothesis that there is a difference between black and white applicants.
- However, crime has a really significant effect. Applicants with no criminal history has a 12.5% chance of receiving a call-back while the probability for applicants who have committed crimes is roughly 7.8%. The difference is consistent with the coefficient on crime in the LPM model. As the coefficient on crime variable is very statistically significant (p-value smaller than 0.001 level of significance), crime absolutely plays a part in employer's decision.

** Conclusion **

- Before BTB is in effect, I have found no evidence that employers discriminate against black people. Employers are less likely to hire people with criminal history, but they are equally likely to hire blacks compared with whites.
- After BTB is in effect, because employers do not have information regarding whether a person has committed crimes before, they possibly assume that black people are more likely to commit a crime than white people, therefore they become less likely to hire black people.
- Interestingly, whether or not a person has a GED (graduated from high school) does not appear to have an effect on the probability of receiving a call-back. This may be because the study randomly applies for only low-skilled, entry-level jobs that do not require higher education.