Lab 8 - LPM, logit, probits

Julieth Santamaria

March 29, 2019

Based on Bailey (2016)'s Real Econometrics, Chapter 12, Exercise 2.

Background

Public attitudes toward global warming influence the policy response to the issue. The dataset `EnvSurvey.cvs` provides data from a nationally representative survey of the U.S. public that asked multiple questions about the environment and energy. The table below describes the variables:

TABLE 12.8 Variables for Global Warming Data

Variable	Description
Male	Dummy variable = 1 for men
White	Dummy variable = 1 for whites
Education	Education, ranging from 1 for no formal education to 14 for professional/doctorate degree (treat as a continuous variable)
Income	Income, ranging from 1 for household income < $5,000 to 19 for household income > $175,000 (treat as a continuous variable)
Age	Age in years
Party7	Partisan identification, ranging from 1 for strong Republican, 2 for not-so-strong Republican, 3 leans Republican, 4 undecided/independent, 5 leans Democrat, 6 not-so-strong Democrat, 7 strong Democrat

Dataset

We want to estimate whether global warming is real and caused by humans (the dependent variable is `HumanCause`) using as independent variables sex, being white, education, income, age, and partisan identification.

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

Application

1. Use `lm()` to estimate the following linear probability model (LPM):

   $$HumanCause = \beta_0 + \beta_1 Male + \beta_2 White + \beta_3 Education + \beta_4 IncomeCategory + \beta_5 Age + \beta_6 Age^2 + \beta_7 Party7 + u$$

```
lpm1 <- lm(humancause~male + white + educ + incomecat, data=envsvy)
```

2. Add Age and Age squared to the regression above.

```
lpm2 <- lm(humancause~male + white + educ + incomecat + age + I(age^2), data=envsvy)
```

3. Now add partisan affiliation.

```
lpm3 <- lm(humancause~male + white + educ + incomecat + age + I(age^2)+party7, data=envsvy)
```

4. Before interpreting these results, let's make nice tables! To run the following lines you will need to save the "vartags_lab8.xlsx" file in your labs folder. You will also need to have tabLego installed in your computer. Folow the next steps every time you want to make nice tables:

```
    # 1. Import labels of the independent variables (You can also label some statistics).
    varTags <- read.xlsx("vartags_lab8.xlsx")

    # 2. Create the list of regressions you will include in the table
    regsdf <- addRegs(list(lpm1,lpm2,lpm3))

    # 3. Define how characteristics of your table
    tab1 <- textTable(reg.frame=regsdf,
                      regnums=c(1,2,3), # To set the order of the regressions
                      decs=2, # Sets the number of decimals to 2
                      var.tags=varTags # To use the labels
                      )

    # 4. Display your table
    textTablePrint(tab1,
                   before="Table 1: LMP models",
                   after=c("Significance: 0.01=***, 0.05=**, 0.1=*",
        "Note: The dependent variable is a dummy for whether the",
        "respondent thinks global warming is real and caused by humans.",
        "Heteroskedasticity-corrected standard errors in parenthesis."
        ))
```

```
## Table 1: LMP models
## ------------------------------------------------
##                              [a]      [b]      [c]
##                Constant  0.03     0.22**  -0.22**
##                          (0.07)   (0.11)   (0.10)
##                     Men -0.01    -0.01     0.02
##                          (0.02)   (0.02)   (0.02)
##                   White -0.08*** -0.08***  0.04
##                          (0.03)   (0.03)   (0.03)
##      Education category  0.03***  0.03***  0.03***
##                          (0.01)   (0.01)   (0.01)
##         Income category -0.00    -0.00     0.00
##                          (0.00)   (0.00)   (0.00)
##                     Age           -0.01**  -0.01***
##                                   (0.00)   (0.00)
```

```
## Partisan Identification                    0.09***
##                                           (0.00)
##             Age squared           0.00**   0.00**
##                                  (0.00)   (0.00)
##            Observations  1855     1855     1855
##               R-Squared 0.019    0.022    0.156
## ------------------------------------------------
## Significance: 0.01=***, 0.05=**, 0.1=*
## Note: The dependent variable is a dummy for whether the
## respondent thinks global warming is real and caused by humans.
## Heteroskedasticity-corrected standard errors in parenthesis.
```

5. How do you interpret the coefficients? Remember that the dependent is a dichotomous variable.

   Coefficient on `incomecat`: It is the change in the probability that a person says global warming has human cause for a one-unit change in the income category the person is in, holding everything else constant

6. How will the probability of an "average" woman saying global warming has human cause change if she goes from being undecided (party7=4) to strong republican (party7=1) (all else constant - at the means for continuous variables)?

$$\Delta P(\hat{Y}=1) = Final\ P(\hat{Y}=1) - Initial\ P(\hat{Y}=1)$$

$$= [\hat{\beta}_0 + \hat{\beta}_1(male=0) + ... + \beta_6 * 4] - [\hat{\beta}_0 + \hat{\beta}_1(male=0) + ... + \beta_6 * 1]$$

$$= \hat{\beta}_6 * 4 - \hat{\beta}_6 * 1$$

$$= \hat{\beta}_6 * 3$$

   The prob. of [an average woman] saying global warming has human cause will increase by $[\beta_6 * 3] * 100$ percentage points

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

```
prediction1 <- predict(lpm3, interval = "prediction")
```

```
## Warning in predict.lm(lpm3, interval = "prediction"): predictions on current data refer to _futu
```

```
summary(prediction1)
```

```
##       fit               lwr                upr
##  Min.   :-0.1814   Min.   :-1.04157   Min.   :0.6788
##  1st Qu.: 0.1835   1st Qu.:-0.67135   1st Qu.:1.0379
##  Median : 0.3355   Median :-0.51974   Median :1.1904
##  Mean   : 0.3358   Mean   :-0.51900   Mean   :1.1907
##  3rd Qu.: 0.4910   3rd Qu.:-0.36372   3rd Qu.:1.3459
##  Max.   : 0.7708   Max.   :-0.08509   Max.   :1.6266
```

   The minimum value of the fitted value is -0.1814, which does not make sense as a predicted probability. The maximum value of the fitted value is 0.7708, which makes sense as a probability (although the out-of-sample predictions would eventually exceed 1 for sufficiently high values of age, education, or income). The fact that predicted values are not bounded to be between 0 and 1 is one of the limits of the LPM approach, specially if we are interested in prediction.

8. Use a probit regression to estimate the probability of saying that global warming is real and caused by humans. Use the independent variables from part (a), including the age-squared variable. Compare statistical significance with LPM results.

```r
# Probit
probit <- glm(humancause~male + white + educ + incomecat + age +  I(age*age)+ party7,
              data=envsvy, family=binomial(link="probit"))
summary(probit)
```

```
##
## Call:
## glm(formula = humancause ~ male + white + educ + incomecat +
##     age + I(age * age) + party7, family = binomial(link = "probit"),
##     data = envsvy)
##
## 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 **
## I(age * age)  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
```

```r
# Logit
logit <- glm(humancause~male + white + educ + incomecat + age +  I(age*age)+ party7,
             data=envsvy, family=binomial(link="logit"))
summary(probit)
```

```
##
## Call:
## glm(formula = humancause ~ male + white + educ + incomecat +
##     age + I(age * age) + party7, family = binomial(link = "probit"),
##     data = envsvy)
##
## 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 **
## I(age * age)  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
```

9. Can we interpret the coefficients? No, we cannot. Next lab we are going to learn how to calculate marginal effects.