Problem Set 4

Francisco Brady

22 Mar 2024

Section I: Modeling Non-linear Relationships

In this part, you will be revisiting the dataset from Assignment 3, on union status and hourly wages. Please use the dataset unions.dta, provided under the Assignment 4 tab on Canvas. This time, we will be using the logged hourly wage, rather than the level, which is common in this literature.

The variables we will be using are:

1. `lnwage` – Logged hourly wage last year (in $). This was estimated by dividing wage and salary income by the approximate number of hours worked last year (weeks worked X usual hours worked per week). Observations with hourly wages less than $3 and more than $40 were excluded.
2. `union` - A dummy variable indicating whether the worker was a union member or covered by some other collective bargaining agreement.
3. `age` - Age in years.
4. `empsize` - The size of the firm the person works for. This was originally a categorical variable with ranges (e.g. 10-24, 25-99, etc) for which I have imputed the midpoint of the ranges, but just ignore that for now. Treat it as a continuous variable. For this assignment, I have also divided the firm size by 100, so a one-unit increase in `empsize` can be interpreted as a 100-person increase in the size of the firm.
5. And five, mutually exclusive variables indicating industry of employer:
    a. `Ind_retail` - binary variable indicating working in retail
    b. `Ind_personal` - binary variable indicating working in personal/service industry
    c. `Ind_health` - binary variable indicating working in health care industry
    d. `Ind_educ` - binary variable indicating working in education industry
    e. `Ind_govt` - binary variable indicating working in government

1. Regress the logged hourly wage on the following variables: age, union, empsize, ind_retail, ind_personal, ind_health, ind_educ. Put the results of this regression into column 1 of Table 1. Based on this model:

```
────────────────────────────────────────────────────────────────────────────────────
                          model 1         model 2         model 3         model 4   

────────────────────────────────────────────────────────────────────────────────────
Age                         0.006***        0.046***        0.046***        0.046***
                           (0.00)          (0.01)          (0.01)          (0.01)   
= 1 if in or cover~n        0.139***        0.120***        0.271***        0.116** 
                           (0.04)          (0.04)          (0.10)          (0.05)   
Firm size                   0.011***        0.010***        0.012***        0.010***
                           (0.00)          (0.00)          (0.00)          (0.00)   
ind_retail                 -0.432***       -0.409***       -0.409***       -0.410***
                           (0.05)          (0.05)          (0.05)          (0.05)   
ind_personal               -0.306***       -0.329***       -0.330***       -0.330***
                           (0.10)          (0.10)          (0.10)          (0.10)   
ind_health                 -0.080          -0.099*         -0.097*         -0.102*  
                           (0.06)          (0.06)          (0.06)          (0.06)   
ind_educ                   -0.149***       -0.137***       -0.144***       -0.138***
                           (0.05)          (0.05)          (0.05)          (0.05)   
agesquared                                 -0.000***       -0.000***       -0.000***
                                           (0.00)          (0.00)          (0.00)   
empsize_union                                              -0.016*                  
                                                           (0.01)                   
health_union                                                                0.025   
                                                                           (0.13)   
_cons                       2.516***        1.769***        1.764***        1.771***
                           (0.08)          (0.14)          (0.14)          (0.14)   
```

```
Adj. R-Squared              0.183        0.213        0.215        0.212
R-Squared                   0.189        0.219        0.222        0.219
Observations             1000.000     1000.000     1000.000     1000.000

* p<0.10, ** p<0.05, *** p<0.01
```

a. Interpret the coefficient on union and discuss its statistical significance.

The coefficient on union membership is 0.139, and it is statistically significant at the p<0.01 level. This implies that a transition from non-union government jobs to union government jobs is associated with an increase in hourly wage of around 14%, holding all else equal.

b. Interpret the coefficient on age and discuss its statistical significance.

The coefficient on age is 0.006, and it is also statistically significant at the p<0.01 level. This implies that an a worker 1 year older can expect something like a .6% increase in their predicted hourly wage, holding everything else constant.

c. If age increases from 25 to 30, how is the hourly wage expected to change?

$$\begin{aligned}\Delta_{ln(wage)} &= \beta_1 \cdot (30 - 25)\\ \Delta_{ln(wage)} &= 0.006 \cdot (5)\\ \Delta_{ln(wage)} &= .03\end{aligned}$$

Between workers who are 25 and those who are 30, holding all else constant, the predicted hourly wage is around 3% higher on average.

d. If age increases from 40 to 45, how is the hourly wage expected to change?

$$\begin{aligned}\Delta_{ln(wage)} &= \beta_1 \cdot (45 - 40)\\ \Delta_{ln(wage)} &= 0.006 \cdot (5)\\ \Delta_{ln(wage)} &= .03\end{aligned}$$

Between workers who are 40 and those who are 45, holding all else constant, the predicted hourly wage is around 3% higher on average.

2. Now add a quadratic term for the age variable (note, you will have to make this variable yourself. Call it `agesquared`), keeping all other controls the same as in column 1. Put the results of this regression into column 2 of Table 1. Based on this model:

a. Look at the signs on the age and age-squared terms. Based only on this information, how would you describe the relationship between age and hourly wages?

The coefficient on age increased from 0.006 to 0.046 and it is significant at the p<0.01 level, as before. The coefficient on agesquared is -0.0004 and significant at the p<0.01 level. Because β_1 is positive and significant, and β_2 is negative and significant, this implies that log wages increase as age increases, but as a **decreasing** rate.

b. If age increases from 25 to 30, how is the hourly wage expected to change?

$$\begin{aligned}\Delta Y &= (\beta_1 + 2\beta_2 X)\Delta X\\ \Delta Y &= (0.046 + 2 * (-0.0005 * 25)) * (5)\\ \Delta Y &= .105\end{aligned}$$

So the wages of the government non-union worker is expected to increase by 10.5% on average, holding all else equal.

c. If age increases from 40 to 45, how is the hourly wage expected to change?

To do this, we need to calculate the predicted Y at both 40 and 45.

$$\Delta Y = (\beta_1 + 2\beta_2 X)\Delta X$$
$$\Delta Y = (0.046 + 2 * (-0.0005 * 40)) * (5)$$
$$\Delta Y = .03$$

So the wages of the government non-union worker is expected to increase by 3% on average, holding all else equal.

d. Which of the two specifications of age (the linear specification in column 1, or the quadratic in column 2), do you think best explains the relationship between age and hourly wages? Explain your reasoning in 1-2 sentences.

Including a squared age term is significant, which means that it is appropriate to include in the model. This makes intuitive sense because at the beginning of a career, extra years (age) can translate into wage gains, but the rate of wage increases, will taper off over time.

3. Now add an interaction term of employer size and union status (`empsize*union`) and put the results into column 3 of Table 1. Describe the relationship between hourly wages, employer size, and union status, incorporating the interaction terms and main effects. Your answer should include reference to a specific numerical illustration of the relationship and note the statistical significance of these relationships.

The coefficient on union has increased from 0.120 to 0.271, indicating that on average, switching from a non-union firm to union firm is associated with ~27% higher hourly wages, for government workers, holding all else equal. The interaction term on employer size and union is significant at a lower level (p<0.10). This implies that larger non-union government jobs are associated with slower wage increases. Employer size is still positive, indicating that increasing size of the firm and going from non-union to union are both associated with The interaction term between employer and union status is negative but only significant at p<0.10. This implies that a weak relationship exists, and that on average an increase of 100 employees and going from union to non-union firms is associated with 1.6% lower hourly wages.

4. Now go back to the model specified in column 2, and add an interaction term of whether the worker is in the health care industry and union status (`ind_health*union`). Put the results of this regression into column 4 of Table 1. Describe the relationship between hourly wages, working in the health care industry, and union status, incorporating the interaction terms and main effects. Your answer should include reference to a specific numerical illustration of the relationship and note the statistical significance of these relationships.

The coefficient for health care industry interacted with union is 0.025, however it is not positive. This implies that for a non-union health care worker, you cannot reject the hypothesis that hourly wages across union and non-union jobs are the same. The main effects for age and union are both still positive and significant, which implies that 1) as workers age, hourly wages increase for health care workers relative to non-union government workers, and 2) switching from non-union to union firms is associated with an 11.6% increase in hourly pay relative to non-union government workers, holding all else equal.

Section II: Binary dependent variables

In this part of the assignment, you will be using data from the Panel Study of Income Dynamics (PSID) to examine teen birth rates. The simplified dataset contains the following information:

- `teenbirth` - an indicator variable =1 if the individual had a birth by the age of 19; zero otherwise.
- A set of mutually-exclusive indicators for race/ethnicity:
    - `white` - An indicator for whether the respondent identifies as White, non-Hispanic

    - `black` - An indicator for whether the respondent identifies as Black, non-Hispanic
    - `Hisp` - An indicator for whether the respondent identifies as Hispanic
    - `Other` - An indicator for whether the respondent identifies as an other race or ethnicity
- `head_educ` - The number of years of completed education for the parent of the individual
- `frac_marr_parents` - The fraction of childhood that the individual spent with married parents, continuous variable ranging from 0 to 1, with 0 representing individuals who never had married parents during childhood, and 1 representing individuals who always lived with married parents.
- `familyincome` - Average family income (in $1,000s) during childhood

Please create a second table (Table 2) to display all the regression output from Part II (as you did above for Part I). Remember to use robust standard errors throughout.

```
. use teen_births, clear

. quietly regress teenbirth black hisp other head_educ frac_marr_parents familyincome, rob

. eststo col1, title(LPM)

. quietly logit teenbirth black hisp other head_educ frac_marr_parents familyincome, robus

. eststo col2, title(logit)

. quietly probit teenbirth black hisp other head_educ frac_marr_parents familyincome, robu

. eststo col3, title(probit)

. quietly logit teenbirth black hisp other head_educ frac_marr_parents familyincome, robus

. eststo col4, title(odds)

. estout col1 col2 col3 col4, eform(0 0 0 1) cells(b(star fmt(3)) se(par fmt(2))) starleve
> .10 ** 0.05 *** 0.01) legend stats(r2_a r2 N, labels("Adj. R-Squared" "R-Squared" "Obser
> s")) label collabels("")

─────────────────────────────────────────────────────────────────────────────────────
                              LPM           logit          probit             odds

─────────────────────────────────────────────────────────────────────────────────────
main
black                       0.079***        0.555***        0.314***        1.742***
                           (0.01)          (0.10)          (0.05)          (0.17)
hisp                        0.071           0.648*          0.371**         1.911*
                           (0.05)          (0.33)          (0.19)          (0.64)
other                      -0.019          -0.373          -0.188           0.688
                           (0.05)          (0.75)          (0.37)          (0.51)
head_educ                  -0.015***       -0.078***       -0.045***        0.925***
                           (0.00)          (0.01)          (0.01)          (0.01)
frac_marr_parents          -0.026**         0.055           0.031           1.057
                           (0.01)          (0.09)          (0.05)          (0.09)
familyincome               -0.000***       -0.018***       -0.009***        0.982***
                           (0.00)          (0.00)          (0.00)          (0.00)
_cons                       0.324***       -0.517***       -0.339***        0.596***
                           (0.03)          (0.20)          (0.11)          (0.12)
─────────────────────────────────────────────────────────────────────────────────────
Adj. R-Squared              0.063
R-Squared                   0.064
Observations             6616.000        6616.000        6616.000        6616.000
─────────────────────────────────────────────────────────────────────────────────────
* p<0.10, ** p<0.05, *** p<0.01

.
```

1. Using a linear probability model, regress the variable teenbirth on the following variables: black, hisp, other, head_educ frac_marr_parents familyincome. This will be column 1.

a. Interpret the coefficient of the following variables:

- `familyincome`
- `black`
- `frac_marr_parents`

If the coefficient is statistically different than zero, you should make sure to discuss the magnitude of the estimated coefficient (this can be done in several ways. For example you can characterize the effect size in words, or standardize the coefficient). NOTE: For binary variables, make sure to (i) include mention of a reference category and (ii) note the effect in terms of both percentage points and percent. For a baseline, you can simply use the average probability of having a teen birth in the data.

- Family Income: The coefficient on family income is -0.000 and is statistically significant at the p<0.01 level. This implies that an increase of 1000 of family income has a small negative impact on teen birth.
- Black: The coefficient on Black is 0.079 and statistically significant at the p<0.01 level. This implies that relative to the reference group (white), on average Black racial identification is associated with an increase of 7.9 percentage points in the likelihood of a teen birth.
- Fraction of childhood with Married Parents: The coefficient on this variable is -0.026 and significant at the p<0.01 level. This implies that on average an individual with parents married their entire childhood (frac_marr_parents == 1) is associated with a decrease of 2.6 percentage points in the probability of teen births, relative to the white reference group with frac_marr_parents == 0.

b. Using the results from the regression above, calculate predicted values for each observation in the data set. What fraction of observations has predicted values outside of the range 0-1?

```
. quietly reg teenbirth black hisp other head_educ frac_marr_parents familyincome, robust

. predict predicted_values1
(option xb assumed; fitted values)
(6 missing values generated)

. sum predicted_values1 if predicted_values1 < 0 | predicted_values1 > 1

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
predicted_~1 |        275   -.0410155    .0737929  -.8012976  -6.64e-06

. local fraction = r(N) / _N

. di `fraction'
.04152824
```

4.15% of the predicted values are either greater than 1 or less than 0.

2. Estimate the model from question 1 above as a logit model. This will be column 2. Calculate the predicted values for each observation. What fraction of observations has predicted values outside the range 0-1?

```
. quietly logit teenbirth black hisp other head_educ frac_marr_parents familyincome, robus

. predict predicted_values2
(option pr assumed; Pr(teenbirth))
(6 missing values generated)

. sum predicted_values2 if predicted_values2 < 0 | predicted_values2 > 1

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
predicted_~2 |          0

. local fraction = r(N) / _N

. di `fraction'
0
```

None of the predicted values are greater than 1 or less than 0.

3. Estimate the model from question 1 above as a probit model. This will be column 3. Calculate the predicted values for each observation. What fraction of observations has predicted values outside the range 0-1?

```
. quietly probit teenbirth black hisp other head_educ frac_marr_parents familyincome, robu

. predict predicted_values3
(option pr assumed; Pr(teenbirth))
(6 missing values generated)

. sum predicted_values3 if predicted_values3 < 0 | predicted_values3 > 1

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
predicted_~3 |          0

. local fraction = r(N) / _N

. di `fraction'
0
```

None of the predicted values are greater than 1 or less than 0.

4. Compare the coefficients from the LPM model with the coefficients from the analogous probit and logit models. For which, if any, predictors, do you see a difference in terms of sign (positive or negative) or statistical significance across the three models?

- Family income has the same level of significance across all of the models. This implies that increasing family income is associated with lower probabilities of teen birth.
- Hispanic identification is not positive across all of the models, but more significant in the probit model than in the logit model. This implies that hispanic identification is most likely weakly associated with increased probability of teen birth.
- Fraction of childhood parents are married switches signs, and loses any significant from the linear probability model and the logit and probit.

For the questions below, imagine an individual with the following values of the variables:
- `familyincome` = 25
- `black` = 0
- `hisp` = 1
- `other` = 0
- `head_educ` = 12
- `frac_marr_parents` = .50

5. Using the results from the LPM model, how much would the probability of having a teen birth change if the individual spent all of their childhood with married parents?

Calculate `frac_marr_parents = .50`

$$Pr(teenbirth_{.50}) = 0.079(0) + 0.071(1) + (-0.019 * 0) + (-0.015 * 12) + (-0.026 * 0.5) + (-0.0005 * 25)$$
$$Pr(teenbirth_{.50}) = 0.1956026$$

Calculate `frac_marr_parents = 1`

$$Pr(teenbirth_{1}) = 0.079(0) + 0.071(1) + (-0.019 * 0) + (-0.015 * 12) + (-0.026 * 1) + (-0.0005 * 25)$$
$$Pr(teenbirth_{1}) = 0.1826618$$

Calculate `frac_marr_parents = 1` - `frac_marr_parents = .50`

$$Pr(teenbirth_{1}) - Pr(teenbirth_{.50}) = 0.1956026 - 0.1826618$$
$$Pr(teenbirth_{1}) - Pr(teenbirth_{.50}) = 0.01294075$$

6. Now answer the same question using the logit model. We would like you to do this two different ways for practice.

a. Calculate the effect "long hand" using the fact that:

$$Pr(Y=1|X)F(\beta_0+\sum_{k=1}^{K}\beta_k x_k)=\frac{1}{1+e^{-(\beta_0+\sum_{k=1}^{K}\beta_k x_k)}}$$

General Formula

$$Pr(Y=1|X)=\frac{1}{1+e^{-(\beta_0+\beta_1 Black+\beta_2 Hisp+\beta_3 Other+\beta_4 headeduc+\beta_5 marrparents+\beta_6 famincome)}}$$

`frac_marr_parents = .50`

\$\$

$$Pr(Y=1|X=.50)=\frac{1}{1+e^{-(-0.517+(0.555*0)+(0.648*1)+(-0.373*0)+(-0.078*12)+(0.055*.50)+(-0.018*25)}}$$
$$Pr(Y=1|X=.50)=0.2276325$$

\$\$

`frac_marr_parents = 1`

$$Pr(Y=1|X=1)=\frac{1}{1+e^{-(-0.517+(0.555*0)+(0.648*1)+(-0.373*0)+(-0.078*12)+(0.055*.50)+(-0.018*25)}}$$
$$Pr(Y=1|X=1)=0.2325338$$

Change:

$$\Delta Pr(\hat{Y}=1)=Pr(Y=1|X=1)-Pr(Y=1|X=.50)$$
$$\Delta Pr(\hat{Y}=1)=0.2325338-0.2276325$$
$$\Delta Pr(\hat{Y}=1)=0.004901257$$

b. Now do the same calculation using the margins command in Stata/R. (note: for the margins command to work, you need to re-run the logit model, then find the predicted value.)

```
. quietly logit teenbirth black hisp other head_educ frac_marr_parents familyincome, robus

. margins, at(black=0 hisp=1 other=0 head_educ=12 frac_marr_parents= (.50 1) familyincome=

Adjusted predictions                                    Number of obs = 6,616
Model VCE: Robust

Expression: Pr(teenbirth), predict()
1._at: black             =  0
       hisp              =  1
       other             =  0
       head_educ         = 12
       frac_marr_pare~s  = .5
       familyincome      = 25
2._at: black             =  0
       hisp              =  1
       other             =  0
       head_educ         = 12
       frac_marr_pare~s  =  1
       familyincome      = 25

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   std. err.      z    P>|z|     [95% conf. interval]
```

```
-------------+----------------------------------------------------------------
         _at |
          1  |   .2276324   .0576313     3.95   0.000      .114677    .3405877
          2  |   .2325336   .0585373     3.97   0.000     .1178026    .3472646
------------------------------------------------------------------------------

. di .2325336 - .2276324
.0049012
```

7. How do the effects calculated in 5 (based on LPM) and 6 (based on the Logit) differ from each other? Briefly discuss in a sentence or two.

The LPM model finds a much larger effect and it is significant, which factors into the expected probability calculation.

8. Redo the logit model shown in column 2, but instead of presenting the coefficients, present the odds ratios associated with each predictor. This should be column 4. Interpret the odds ratios associated with the predictors familyincome and black. One sentence should be sufficient for each predictor.

Family income: The odds ratio of family income is .982, which implies that an increase of 1000 in family income would decrease the odds of teen birth by a little bit less than 2%, holding all else equal.
Black: Relative to white teens, being Black increases the probability of a teen birth by 72%, holding all else equal.

9. Briefly describe what you found most interesting about the results of the analyses above from a substantive policy perspective. One short paragraph should be sufficient.

Depending on what model you pick to estimate the coefficients, you may arrive at different conclusions. This could affect the policy implications and approach. If you choose the LPM, then you will arrive at the conclusion that encouraging parents of children to get and stay married will reduce the likelihood of teen births. If you choose a logit or probit model, that is not the most important conclusion to reach. In terms of policy implications, you may conclude that policies that support increasing family income or increasing educational attainment are more important and worth advocating.