**Assignment 4**

**Due: Friday, March 22nd, 2024 11:59pm**

Please upload your assignment to Canvas under the Assignments tab.

For all questions requiring you to use Stata or R, please include your output as part of your homework. In Stata, you can highlight Stata output, right click, and select “copy as picture” to easily copy and paste Stata output into Word.

**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:
   1. *Ind\_retail-* binary variable indicating working in retail
   2. *Ind\_personal-* binary variable indicating working in personal/service industry
   3. *Ind\_health-* binary variable indicating working in health care industry
   4. *Ind\_educ-* binary variable indicating working in education industry
   5. *Ind\_govt-* binary variable indicating working in government
6. 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:
   1. Interpret the coefficient on union and discuss its statistical significance.

Being in a union is associated with ~14% increase. NOTE: when the log difference (coefficient) s larger than .1, which would indicate > 10% change, it is more accurate to do the log exponentiation than to interpret the coefficient as a straight 1-to-1 percentage change. Mention significance, etc.

* 1. Interpret the coefficient on age and discuss its statistical significance.

Significant at 99%, coefficient is 0.006 and implies ~.6% increase with additional year of age.

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

5\*.6 = 3.06%

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

5\*.6 = 3.06%

the point here is that they are the same!

1. 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:
   1. 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?

There is a slide about this in the lecture. If the sign on B\_1 is **positive**, and the sign on B\_2 is **negative**, this implies a decreasing slope and plateau to zero.

Coefficient is positive, coefficient on age^2 is negative, so b\_1 > 0, b\_2 < 0 this means that x increases in x at a decreasing rate.

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

Reference: lecture 14, slide 14

DY = (b 1 + 2b 2 X)DX

calculate this for 5

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

See above.

* 1. 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.

The second model makes more sense because when you are younger higher wage increases make sense, and it makes sense that wage increases decrease over time.

Given that coefficient on age^2 is significant, you cannot reject the null that would imply that the relationship is *only* linear, so including age^2 is a good idea.

1. 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.

Suggestion: Actually write out the model with the dummy variables plugged in.

Interaction interpretation: After adding empsize\_union variable, the empsize variable is now ONLY representing impacts of non-union employee size firms.

positive coeffs on union and emplyoer size, both positively related to hourly wage. For both we can reject the null hypothesis of no relationship with hourly wages.

For interaction term – was not significant, and is negative, which implies that growth in wage is 1.598% lower than an increase in 100 employee, non-unionized workplace, holding all else constant.

Employer\_size coefficient is telling us the coefficient for non\_union firms.

1. 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.

Two binary variables:

If non-union and healthcare, they have a .10 or 10% decrease, relative to others. The effect of being in a union is different for HC workers relative to non-HC workers, relative to government workers.

**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.

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**.
   1. Interpret the coefficient of the following variables:

* familyincome
  + on avg, an increase of 1k is associated with a decrease of .046 percentage points in teen births.
* Black
  + associated with 7.95 percentage point increase compared to white folks.
  + Mean teen birth: .1398, so .0795/.1398 = .5686
* frac\_marr\_parents
  + For each increase of 1 percentage point in the amount of time a person spends of their childhood with married parents, is associate with a 0.26 percentage point decrease in the likelihood of teen birth.

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.

* 1. 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?

This is the part of the hw where for stata users can use the predict command.

Check the end of the homework file.

There’s 275/6616 values are outside of 0-1, which is 4.16%

1. 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?

0, this is the advantage of logits

1. 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?

0, this is the advantage of probits

1. 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 models

Hispanic is not significant in lpm, but sig in logit and probit.

While the coefficients change across model, coefficients are positive.

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

1. 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?

Plug in all values:

.202 – 0.189 = 1.3 percentage points

1. Now answer the same question using the logit model. We would like you to do this two different ways for practice.
   1. Calculate the effect “long hand” using the fact that:

Plug in.

Confirm using the margins command



* 1. 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.)

1. 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.

Two models find effects in opposite directions. LPM implies as Parents Marriage increase, teen birth decrease, while Logit implies as parent’s marriage increases, teen birth also increases.

1. 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.

Black: The odds of black teen birth is 1.74 that of white individuals, holding all else constant.

Family Income: Since the odds ratio is essentially 1, the odds of teen birth remain the same regardless of income regardless of any unit increase in income, holding all else constant.

1. 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 the model you pick you might arrive at different policy conclusions.

- Notably family income, black, and some other variables are similar across the models.

**CODING TIPS**

**Creating Regression Tables**

Stata: The command estout will output regression results into a regression table structured in the format requested. You will need to install the estout package first:

ssc install estout, replace

Then, run the following command after you run each regression in the problem set, changing the name and title of each regression accordingly:

estimates store reg1, title(Regression 1)

Once you have run all your regressions, use the code below to compile your table. See the “Regression Tables in Stata” slides on canvas for detailed instructions on how to modify your table to include all required elements.

estout reg1 reg2 reg3 reg4, cells(b(star fmt(3)) se(par fmt(2))) starlevels(\* 0.10 \*\* 0.05 \*\*\* 0.01) legend stats(r2 N, labels("R-Squared" "Observations")) label collabels(“”)

Alternatively, you can use the outreg2 command. You will need to install the package first:

ssc install outreg2

You can then store your regressions in the same way you would with the estout command, and create a regression table with the outreg2 command, which will create a table in excel:

est store m1

outreg2 [m1 m1 m2 m3 m4] using "pset4\_reg\_table.xls", replace dec(3)

R: the library/command “stargazer” allows you to create regression tables in the format requested. Remember to install it and load it first, using install.packages(“stargazer”) and library(“stargazer”). Also remember that you cannot use lm\_robust with stargazer; code to add robust errors is below. As ever, you will want tidyverse loaded. Haven and car may need to be loaded separately as well.

An example to use it would be the following:

m1 <- lm(y ~ x1, data=ps3\_df)

# Get adjusted standard errors

cov1 <- vcovHC(m1, type = "HC1")

r\_se1 <- sqrt(diag(cov1)) # Robust SEs

m2 <- lm(y ~ x1+ x2, data=ps3\_df)

# Get adjusted standard errors

cov2 <- vcovHC(m2, type = "HC1")

r\_se2 <- sqrt(diag(cov2)) # Robust SEs

stargazer(m1, m2, se=list(r\_se1, r\_se2), type = "text", header=FALSE, omit.stat = c("f","ser"), digits=2,column.sep.width="-5pt",title="Title",dep.var.caption="")

The type can be replaced for html or latex if you are using RMarkdown.

**CODING TIPS FOR PART I**

**Data Management and Linear Regression**

Creating the experience variable:

Stata:

gen agesquared=age^2

Setup for regression 1 (column 1 of your table):

Stata:

//regression

reg lnwage union age empsize ind\_educ ind\_retail ind\_personal ind\_health

R:

# create quadratic term

ps3\_df<- ps3\_df %>% mutate(exp2=exp^2)

#regression with factor variables

ps3\_df$married <- factor(ps3\_df$married)

ps3\_df$race <- factor(ps3\_df$race)

ps3\_df$male <- factor(ps3\_df$sex)

ps3\_df$cworker <- factor(ps3\_df$cworker)

m4 <- lm(l\_earnings ~ educ + exp2 + married + race + sex + cworker, data=ps3\_df)

# Adjust standard errors

cov4 <- vcovHC(m4, type = "HC1")

r\_se4 <- sqrt(diag(cov3)) # Robust SEs

Generating an interaction term:

Stata:

gen empsize\_union = empsize\*union

R:

ps3\_df$empsize\_union <- ps3\_df$empsize\*ps3\_df$union

**CODING TIPS FOR PART II**

**Logit and Probit Models**

Stata:

logit teenbirth familyincome black hisp other head\_educ frac\_marr\_parents, r

probit teenbirth familyincome black hisp other head\_educ frac\_marr\_parents, r

R:

model3 <- teenbirth ~ familyincome + black + hisp + other + head\_educ + frac\_marr\_parents

mlogit <- glm(model3, family = "binomial", data=ps3\_df)

# Adjust standard errors

cov4 <- vcovHC(m4, type = "HC1")

r\_se4 <- sqrt(diag(cov4)) # Robust SEs

ORcoef\_mymodel <- exp(coef(mymodel))  # calculating odds ratios from my logit model

mymodel\_robustse <- sqrt(diag(vcovHC(mymodel), type = "HC1"))  # calculating robust SEs for the coefficients like the example code, but in one step (without creating an intermediary object)

ORse\_mymodel <- ORcoef\_mymodel \* mymodel\_robustse  # creating a vector of robust SEs for the Odds Ratios, this is that extra step that Stata does as part of ", or"

mprobit <- glm(model3, family = binomial(link = “probit”), data=ps3\_df)

# Adjust standard errors

cov4 <- vcovHC(m4, type = "HC1")

r\_se4 <- sqrt(diag(cov4)) # Robust SEs

**Predictions and the Margins Command:**

Stata:

margins, at(familyincome=25 black = 0 hisp=1 other=0 head\_educ=12 frac\_marr\_parents=(.5 1))

R:

Library(margins)

#install first if needed

value\_list <- data.frame(

familyincome=25,

black=0,

hisp = 1,

other=0,

head\_educ = 12 ,

frac\_marr\_parents = 0.5,

)

# get predicted probability given specified values of the x's

prediction <- predict(mprobit,

newdata = value\_list,

type = "response",

se.fit = TRUE)

prediction

#what percent of predictions are out of range?

#(tibbles are a simplified dataframe)

yhat <- as\_tibble(predict(mprobit))

yhat %>% mutate(out=if\_else(value>1 | value < 0 ,1,0)) %>%

summarise(perc\_out=mean(out))