ECON 471

Fall 2020

Problem Set 5

Due Wednesday Dec 2, by Midnight CST

Name: Albert Wiryawan

Section: B3

1. Use the data in LOANAPP.TXT for this exercise. The binary variable to be explained is *approve*, which is equal to one if mortgage loan to an individual was approved. The key explanatory variable is *white*, a dummy variable equal to one if the applicant was white. The other applicants in the data set are black and Hispanic.
   To test for discrimination in the mortgage market, a linear probability model can be used:

$$approve = \beta_0 + \beta_1 white + other\ factors.$$

(i) If there is discrimination against minorities, and the appropriate factors have been controlled for, what is the sign of β_1?

For the above specified model, the parameter β_1 is expected to be positive as it is though that a white individual will have a greater probability of getting a loan approved.

(ii) Regress *approve* on *white* and report the results in the usual form. Interpret the coefficient on *white*. Is it statistically significant? Is it practically large?

$$approve = 0.7077922 + 0.2005957 white + other\ factors.$$
$$(0.01824) \qquad (0.01984)$$

The coefficient of white is 0.2005957 indicating that a white individual has a 20.06% higher chance of getting a loan approved. The p-value associated with this parameter is extremely close to 0 which is less than the typical 5% significance level. Which means that this coefficient is statistically significant at the 5% significance level

(iii) As controls, add the variables *hrat*, *obrat, , loanprc, unem, male, married, dep, sch, cosign, chist, pubrec, mortlat1, mortlat2*, and *vr*. What happens to the coefficient on *white*? Is there still evidence of discrimination against nonwhites?

The coefficient of white now decreases to 0.1283374 and has a p-value of 6.95e-11. Since the p-value is less than the 5% level of significance, the null hypothesis is rejected and there is evidence of statistical significance of the white parameter at the 5% significance level.

(iv) Now, allow the effect of race to interact with the variable measuring other obligations as a percentage of income (*obrat*). Is the interaction term significant?

The coefficient of (obrat * white) is 0.0080925 with a p-value of 0.000384 which is less than 0.05 at 5% level of significance indicating that the coefficient parameter of the interaction term obrat*white is statistically significant at 5% level of significance.

(v) Now, estimate a probit model of *approve* on *white*. Find the estimated probability of loan approval for both whites and nonwhites. How do these compare with the linear probability estimates in part (ii)?

P(approve = 1| white =0) = $\phi(0.54695 + 0.78395 * 0) = 0.707792$ (probability of loan approved for nonwhite)
P(approve = 1| white =1) =$\phi(0.54695 + 0.78395 * 1)$ =0.90875 (probability of loan approved for whites)

(vi) Now, add the variables *hrat*, *obrat, loanprc, unem, male, married, dep, sch, cosign, chist, pubrec, mortlat1, mortlat2*, and *vr* to the probit model. Is there statistically significant evidence of discrimination against nonwhites?

P(approve = 1| white =0) = $\phi(2.0623 + 0.520253 * 0) = 0.9804$ (probability of loan approved for nonwhite)

P(approve = 1| white =1) =$\phi(2.0623 + 0.520253 * 1)$ =0.995097 (probability of loan approved for whites)

(vii) Estimate the model from part (ii) by logit. Compare the coefficient on *white* to the probit estimate.

The model for logit is given by

$$approve = 0.8847 + 1.4094 white + other\ factors.$$

The parameter for white is given by 1.4094 and has a p-value close to zero which shows that added coefficient is still statistically significant showing being white adds to the probability of obtaining approval for a loan. The logit model also estimates this affect as stronger than that found in probit.

Code:

```{r}
# load data
setwd("/Users/albertwiryawan/Code/Class_GitRepos/Fall_2020/Econometrics/Problem Set #5/Data")
loanapp = read.table("loanapp.txt")
# attach database to r search path to make it easier to attach columns of tables to variables
attach(loanapp)
approve = V50
```

```
white = V59
#model = lm(approve ~ white)
#summary(model)
#coefficients(model)
# part iii
hrat = V26
obrat = V27
loanprc = V57
unem = V39
male = V48
married = V9
dep = V10
sch = V45
cosign = V35
chist= V55
pubrec= V25
mortlat1= V53
mortlat2= V54
vr= V44
#model1 = lm(approve ~ white + hrat +obrat + loanprc + unem + male + married + dep + sch + cosign
+ chist + pubrec + mortlat1 + mortlat2 + vr)
#coefficients(model1)
#summary(model1)
#part iv
model2 = lm(approve ~ white + hrat +obrat + loanprc + unem + male + married + dep + sch + cosign +
chist + pubrec + mortlat1 + mortlat2 + vr + I(obrat * white))
#summary(model2)
#coefficients(model2)
model3 = glm(approve ~ white, family=binomial(link="probit"))
model4 = glm(approve ~ white + hrat + obrat + loanprc + unem + male + married + dep + sch + cosign
+ chist + pubrec + mortlat1 + mortlat2 + vr, family=binomial(link="probit"))
model5 = glm(approve ~ white, family=binomial(link="logit"))
summary(model5)
```

2. An equation explaining chief executive officer salary is

$$\widehat{\log(salary)} = 4.59 + .257\log(sales) + .011roe + .158finance + .181consprod$$
$$(.30) \quad (.032) \qquad\qquad (.004) \qquad (.089) \qquad\qquad (.085)$$
$$-.283utility$$

$$(.099)$$

$$n = 209, \qquad R^2 = .357.$$

The data used are in CEOSAL1, where *finance*, *consprod*, and *utility* are binary variables indicating the financial, consumer products, and utilities industries. The base industry is transportation.

(i) Compute the approximate percentage in estimated salary between the utility and transportation industries, holding *sales* and *roe* fixed. Is the difference statistically significant at the 1% level?

H_0: $\mu_m = 0$

H_1: $\mu_m \neq 0$

t= (-.283-0)/ 0.099 = -2.86

with this test statistic, there is significance at the 1% level. The difference in estimated salary between the two sectors is significant.

(ii) Use equation (7.10) in the textbook (ch 7) to obtain the exact percentage difference in the estimated salary between the utility and transportation industries and compare this with the answer obtained in part (i).

u = [e^-.283 – 1] * 100 = -24.64%

This is the solved estimate for the coefficient estimate according to equation 7.10. It is smaller than 28.3% found in part 1.

(iii) What is the approximate percentage difference in estimated salary between the consumer products and finance industries? Write an equation that would allow you to test whether the difference is statistically significant.

The percentage difference in estimated salary between the consumer products and finance industries is (0.181-0.158) = 0.023. The proportionate difference in salary is 2.3%.

$$\log(\widehat{salary}) = a0 + a1\log(sales) + a3roe + a4consprod + a5utility + a6trans + e$$

Regress over this equation to test whether the difference is statistically significant. Since trans is a dummy variable to indicate transportation industry, the base group is finance and so the coefficient

will measure the difference between the consumer products and finance industries. This allows us to use the t-statistic on consprod.

3. Let *grad* be a dummy variable for whether a student-athlete at a large university graduates in five years. Let *hsGPA* and *SAT* be high school grade point average and SAT score, respectively. Let *study* be the number of hours spent per week in an organized study hall. Suppose that, using data on 420 student-athletes, the following logit model is obtained:

$$\hat{P}(grad = 1|hsGPA, SAT, study) = \Lambda(-1.17 + .24hsGPA + .00058\, SAT + .073study),$$

where $\Lambda(z) = \exp(z)/[1 + \exp(z)]$ is the logit function. Holding *hsGPA* fixed at 3.0 and *SAT* fixed at 1,200, compute the estimated difference in the graduation probability for someone who spent 10 hours per week in study hall and someone who spent 5 hours per week.

For 10 hours, 3.0gpa, and 1200 SAT

$$\Lambda(-1.17 + .24(3) + .00058\ (1200) + .073(10)) = \Lambda(0.976)$$

Given,

$$\Lambda(z) = \frac{\exp(0.976)}{1 + \exp(0.976)}$$

$$\Lambda(z) = 0.7263$$

For 5 hours, 3.0gpa, and 1200 SAT

$$\Lambda(z) = 0.64816$$

The difference in probability is thus 0.7263-0.64816 = 0.07814

4. Suppose you collect data from a survey on wages, education, experience and gender. In addition, you ask for information about marijuana usage. The original question is: "on how many separate occasions last month did you smoke marijuana?"
   (i) Write an equation that would allow you to estimate the effects of marijuana usage on wage, while controlling for other factors. You should be able to make statements such as, "smoking marijuana five more times per month is estimated to change wages by *x*%."

Log(wage) = $\beta_0 + \beta_1 MarijuanaUsage + \beta_2 educ + \beta_3 experience + \beta_4 experience^2 + \beta_5 male$

MarijuanaUsage = how many times marijuana is used by an employee per month
Educ = education level of the employee
Exper = experience level of an employee
Male = binary dummy variable (=1 for male)

It can be inferred that MarijuanUsage increase by 1 per month. As a result, the wage will change by $100\beta_1\%$. So when marijuana usage increases to 5 per month the estimated change is $500\beta_1\%$.

(ii) Write a model that would allow you to test whether drug usage has different effects on wages for men and women. How would you test that there are no differences in the effects of drug usage for men and women?

The model that is created to allow you to test whether drug usage has different effects on wages of men is given below.

$$\log(wage) = \beta_0 + \beta_1 MarijuanaUsage + \beta_2 educ + \beta_3 experience + \beta_4 experience^2 + \beta_5 male + \beta_6 male * MarjuanaUsage + u$$

The parameter of β_6 is tested by the hypothesis test below:
H_0: $\beta_6 = 0$

H_1: $\beta_6 \neq 0$

Then the t-statistic can be computed using the formula below

$$t = \widehat{\beta_6}/SE(\widehat{\beta_6})$$

Using a t-table if the obtained value is greater than the predetermined value at n-6 (n-5-1 where n is the number of observations) degrees of freedom. It can be concluded that there is statistically significant difference in the effects that the drug usage has on men and women.

(iii) Suppose you think it is better to measure marijuana usage by putting people into one of four categories: nonuser, light user (1 to 5 times per month), moderate user (6 to 10 times per month), and heavy user (more than 10 times per month). Now, write a model that would allow you to estimate the effects of marijuana usage on wage.

$$\log(wage) = \beta_0 + \delta_1 lightuser + \delta_2 moderateuser + \delta_3 heavyuser + \beta_2 educ + \beta_3 experience + \beta_4 experience^2 + \beta_5 male + u$$

Binary dummy variables for the type of user is given by the specified variable lightuser, moderateuser, heavyuser. Nonuser don't need to be specified by the model as it will only depend on other parameters for log(wage)

(iv) Using the model in part (iii), explain in detail how to test the null hypothesis that marijuana usage has no effect on wage. Be very specific and include a careful listing of degrees of freedom.

To test for the null hypothesis that marijuana has no effect on wage the following statements are made

H_0: $\delta_1 = \delta_2 = \delta_3 = 0$

H_1: $\delta_1 \neq \delta_2 \neq \delta_3 \neq 0$

An F-test can now be conducted. There are 3 restrictions and 8 unrestricted coefficients. Thus, the degrees of freedom in the numerator is 3 while the denominator is 8.