problemTwo

Daniel Emaasit

February 24, 2015

# Problem C17.5

Using the data in FERTIL1.RAW.

## Part (i)

Estimate a poisson regression model for kids, using the same variables in TABLE 13.1. Interpret the coefficient on y82.

# read the data into R  
library(haven)  
fertile\_data<-read\_dta("FERTIL1.dta")  
fertile\_data[]<-lapply(fertile\_data, as.numeric)  
fertile\_data[,7:20]<-lapply(fertile\_data[,7:20], as.factor)  
  
  
# estimate a poisson regression model  
library(MASS)  
poisson\_model1<-glm(kids~., family="poisson", data=fertile\_data)  
library(stargazer)

##   
## Please cite as:   
##   
## Hlavac, Marek (2014). stargazer: LaTeX code and ASCII text for well-formatted regression and summary statistics tables.  
## R package version 5.1. http://CRAN.R-project.org/package=stargazer

stargazer(poisson\_model1, type = "text")

##   
## =============================================  
## Dependent variable:   
## ---------------------------  
## kids   
## ---------------------------------------------  
## year 0.072\*\*   
## (0.030)   
##   
## educ -0.002   
## (0.021)   
##   
## meduc -0.002   
## (0.006)   
##   
## feduc -0.004   
## (0.007)   
##   
## age 0.191\*\*\*   
## (0.055)   
##   
## black1 0.347\*\*\*   
## (0.062)   
##   
## east1 0.084   
## (0.053)   
##   
## northcen1 0.137\*\*\*   
## (0.048)   
##   
## west1 0.074   
## (0.066)   
##   
## farm1 -0.030   
## (0.058)   
##   
## othrural1 -0.075   
## (0.070)   
##   
## town1 0.031   
## (0.049)   
##   
## smcity1 0.071   
## (0.062)   
##   
## y741 0.196   
## (0.305)   
##   
## y761 0.116   
## (0.286)   
##   
## y781 0.250   
## (0.310)   
##   
## y801 -0.156   
## (0.294)   
##   
## y821 -0.125   
## (0.321)   
##   
## y841   
##   
##   
## agesq -0.002\*\*\*   
## (0.001)   
##   
## y74educ -0.021   
## (0.027)   
##   
## y76educ -0.036   
## (0.027)   
##   
## y78educ -0.058\*\*   
## (0.028)   
##   
## y80educ -0.037   
## (0.027)   
##   
## y82educ -0.064\*\*   
## (0.027)   
##   
## y84educ -0.087\*\*\*   
## (0.028)   
##   
## Constant -8.436\*\*\*   
## (2.577)   
##   
## ---------------------------------------------  
## Observations 1,129   
## Log Likelihood -2,063.442   
## Akaike Inf. Crit. 4,178.885   
## =============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

The coefficient shows that the fertility was about 12.5% lower in this time period.

## Part (ii)

What is the estimated percentage difference in fertility between a black woman and a non black woman, holding other factors fixed?

(exp(0.347)-1)\*100

## [1] 41.48167

The estimated percentage difference is 41.5%

## Part (iii)

Obtain the estimated standard deviation .Is there evidence of over or underdispersion?

predicted\_values1<-predict(poisson\_model1, type="response", fertile\_data )

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =  
## ifelse(type == : prediction from a rank-deficient fit may be misleading

There is evidence of underdispersion.

## Part (iv)

Compute the fitted values from the Poisson regression and obtain the R-squared as the squared correlation between kids and estimated kids. Compare this with the R-squared for the linear regression model

fitted\_values<-predict(poisson\_model1, type="response", fertile\_data)

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =  
## ifelse(type == : prediction from a rank-deficient fit may be misleading

cor(fertile\_data$kids, fitted\_values)^2

## [1] 0.1314286

R-squared from the OLS model is bigger

# Problem C17.9

Use the data in APPLE.RAW for this exercise.

## Part (i)

Of the 660 families in the sample, how many report wanting none of the eco-labeled apples at the set price?

# read the data into R  
apple\_data<-read\_dta("APPLE.dta")  
  
# count the no of families wanting none of the ecolabeled apples  
library(dplyr)

##   
## Attaching package: 'dplyr'  
##   
## The following object is masked from 'package:MASS':  
##   
## select  
##   
## The following object is masked from 'package:stats':  
##   
## filter  
##   
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

apple\_data %>% filter(ecolbs==0) %>% count()

## Source: local data frame [1 x 1]  
##   
## n  
## 1 248

248 families report wanting none of the eco-labelled apples at the set price

## Part (ii)

Does the variable ecolbs seem to have a continous distribution over strictly positive values? What implications does your answer have for the suitability of a Tobit model for ecolbs?

# look at a sample of the data  
head(apple\_data$ecolbs, n=10)

## [1] 2.000000 2.000000 2.666667 0.000000 3.000000 3.000000 2.000000  
## [8] 1.666667 0.500000 2.333333

The variable ecolbs does not have a continous distribution. This type of distribution violates the latent variable formulation underlying the Tobit model.

## Part (iii)

Estimate a Tobit model for ecolbs with ecoprc, regprc, faminc, and hhsize as explanatory variables. Which variables are significant at the 1% level?

# estimate a Tobit model  
library(AER)

## Loading required package: car  
## Loading required package: lmtest  
## Loading required package: zoo  
##   
## Attaching package: 'zoo'  
##   
## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric  
##   
## Loading required package: sandwich  
## Loading required package: survival  
## Loading required package: splines

tobit\_model1<-tobit(ecolbs~ecoprc+regprc+faminc+hhsize, data=apple\_data)  
stargazer(tobit\_model1, type="text")

##   
## ==========================================  
## Dependent variable:   
## ---------------------------  
## ecolbs   
## ------------------------------------------  
## ecoprc -5.705\*\*\*   
## (0.885)   
##   
## regprc 5.525\*\*\*   
## (1.066)   
##   
## faminc 0.006   
## (0.004)   
##   
## hhsize2 1.006\*\*   
## (0.456)   
##   
## hhsize3 0.769   
## (0.493)   
##   
## hhsize4 0.788   
## (0.506)   
##   
## hhsize5 1.233\*\*   
## (0.613)   
##   
## hhsize6 1.487   
## (0.934)   
##   
## hhsize7 1.225   
## (1.098)   
##   
## hhsize8 -0.609   
## (1.576)   
##   
## hhsize9 3.410   
## (3.450)   
##   
## Constant 0.624   
## (0.687)   
##   
## ------------------------------------------  
## Observations 660   
## Log Likelihood -1,263.248   
## Wald Test 54.416\*\*\* (df = 11)   
## ==========================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

The variables significant at the 1% level include:

## Part (iv)

Are faminc and hhsize jointly significant?

## Part (v)

Are the signs of the coeffficient on the price variables from part(iii) what you expect? Explain

The signs are as expected because as the price of the product increases the demand decreases

## Part (vi)

Test the hypothesis.

# calculate the t value  
calculated\_t\_value<-(-5.705+5.525)/(0.885+1.066)  
calculated\_t\_value

## [1] -0.09226038

The rule: Reject the null hypothesis if |calculated t value|>|critical t value|,

The decision: since the |calculated.t.value|<1.96, we FAIL to reject the null hypothesis

The inference: The ecopric is equal to the negative value of regprc

## Part (vii)

Obtain the estimates of E(ecolbs|x) for all observations in the sample. what are the smallest and largest values?

# predict the estimates  
estimated\_ecolbs<-predict(tobit\_model1, type="response",apple\_data)  
  
# find the range in the estimated values  
range(estimated\_ecolbs)

## [1] -2.146803 3.000000

## Part (viii)

Compute the squared correlation between ecolbs and estimated ecolbs.

# find the squared correlation between the actual and estimated values  
cor(apple\_data$ecolbs,estimated\_ecolbs)^2

## [1] 0.04467454

## Part (ix)

Now, estimate a linear model for ecolbs using the same explanatory variables from part (iii). Why are the OLS estimates so much smaller than the Tobit estimates? In terms of goodness-of-fit, is the Tobit model better than the linear model?

# estimate a linear model  
linear\_model<-lm(ecolbs~ecoprc+regprc+faminc+hhsize, data=apple\_data)  
stargazer(linear\_model, type = "text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## ecolbs   
## -----------------------------------------------  
## ecoprc -2.832\*\*\*   
## (0.592)   
##   
## regprc 2.953\*\*\*   
## (0.716)   
##   
## faminc 0.002   
## (0.003)   
##   
## hhsize2 0.524\*   
## (0.300)   
##   
## hhsize3 0.298   
## (0.325)   
##   
## hhsize4 0.370   
## (0.334)   
##   
## hhsize5 0.676   
## (0.411)   
##   
## hhsize6 0.481   
## (0.650)   
##   
## hhsize7 0.557   
## (0.761)   
##   
## hhsize8 -0.291   
## (0.973)   
##   
## hhsize9 1.980   
## (2.503)   
##   
## Constant 1.430\*\*\*   
## (0.461)   
##   
## -----------------------------------------------  
## Observations 660   
## R2 0.046   
## Adjusted R2 0.030   
## Residual Std. Error 2.488 (df = 648)   
## F Statistic 2.824\*\*\* (df = 11; 648)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

The OLS estimates are so much smaller because they are not scaled like the Tobit estimates. The OLS model fits this data better because its R-squared value is greater i.e 0.046 compared to 0.044

## Part (x)

Evaluate the following statement: "Because the R-squared from the Tobit model is so small, the estimated price effects are probably inconsisent."

I think this statement is incorrent.

# Problem C17.10

Use the data in SMOKE.RAW for this exercise.

## Part (i)

How many people in the sample do not smoke at all? what fraction of people claim to smoke 20 cigarettes a day? Why do you think there is a pile up of people at 20 cigarettes?

# read the data into R  
smoke\_data<-read\_dta("SMOKE.dta")  
  
# count the number of people who do not smoke  
smoke\_data %>% filter(cigs==0) %>% count()

## Source: local data frame [1 x 1]  
##   
## n  
## 1 497

# count the number of people who smoke 20 cigarettes  
smoke\_data %>% filter(cigs==20) %>% count()

## Source: local data frame [1 x 1]  
##   
## n  
## 1 101

Apparently there are 20 cigarettes in a pack, so that could be a reason why there's a pile up.

## Part (ii)

Does cigs seem a good candidate for having a condidtional Poisson distribution?

No, cigs does not seem a good candidate because Poisson distribution does not allow for such focal points that in the cigs variable.

## Part (iii)

Estimate a Poisson regression model for cigs, including log(cigpric), log(income),white, educ, age, and age2 as explanatory variables. What are the estimated price and income elasticities??

# convert string variables to factor variables  
smoke\_data$white<-as.factor(smoke\_data$white)  
smoke\_data$age<-as.numeric(smoke\_data$age)  
smoke\_data$cigs<-as.numeric(smoke\_data$cigs)  
smoke\_data$restaurn<-as.factor(smoke\_data$restaurn)  
  
# estimate a poisson regression model  
poisson\_model2<-glm(cigs~lcigpric+lincome+white+educ+age+agesq, family="poisson", data=smoke\_data)  
stargazer(poisson\_model2, type="text")

##   
## =============================================  
## Dependent variable:   
## ---------------------------  
## cigs   
## ---------------------------------------------  
## lcigpric -0.355\*\*   
## (0.144)   
##   
## lincome 0.085\*\*\*   
## (0.020)   
##   
## white1 -0.002   
## (0.037)   
##   
## educ -0.060\*\*\*   
## (0.004)   
##   
## age 0.115\*\*\*   
## (0.005)   
##   
## agesq -0.001\*\*\*   
## (0.0001)   
##   
## Constant 1.463\*\*   
## (0.615)   
##   
## ---------------------------------------------  
## Observations 807   
## Log Likelihood -8,184.030   
## Akaike Inf. Crit. 16,382.060   
## =============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

The price elasticity is -0.578 and the income elasticity is 0.004

## Part (iv)

Using the maximum likelihood standard errors, are the price and income variables sististically significant at the 5% level?

The price variable is significant while the income variable is not significant.

## Part (v)

Obtain the estimate of the variance. What is the estimated standard deviation? How should you adjust the standard errors from part (iv)?

The standard errors should be adjusted by multiplying by the estimated standard deviation.

## Part (vi)

Using the adjusted standard errors, are the price and income elasticities now statistically different from zero? Explain.

The price elasticity is statistically significant while the income elasticity is not significant

## Part (vii)

Are the education and age variables significant using the more robust standard errors? How do you interpret the coefficient on educ?

The education and age variables are still significant. The coefficient on education implies that an increase in education by one year reduces the number of cigaretters consumed by 5.5%.

## Part (viii)

Obtain the fitted values from the poisson regression model. Find the minimum and maximum values and discuss how well the exponential model predicts heavy cigarette smoking.

# obtain the fitted values  
predicted\_values<-predict(poisson\_model2, type="response", smoke\_data)  
  
# find the minimum and maximum values  
range(predicted\_values)

## [1] 0.5150124 18.8356664

It is very difficult predicting heavy smoking using the available explanatory variables.

## Part (ix)

Using the fitted values from part (viii), obtain the squared correlation coefficient between esimated y and y.

# obtain the square correlation coefficeitn  
cor(predicted\_values, smoke\_data$cigs)^2

## [1] 0.04320565

## Part (x)

Estimate a linear model for cigs by OLS, using the explanatory variables. Does the linear model or exponential model provide a better fit? Is either R-squared very large?

# estimate a linear regression model  
linear\_model2<-lm(cigs~lcigpric+lincome+white+educ+age+agesq, data=smoke\_data)  
stargazer(linear\_model2, type="text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## cigs   
## -----------------------------------------------  
## lcigpric -2.901   
## (5.747)   
##   
## lincome 0.754   
## (0.730)   
##   
## white1 -0.205   
## (1.458)   
##   
## educ -0.514\*\*\*   
## (0.168)   
##   
## age 0.782\*\*\*   
## (0.161)   
##   
## agesq -0.009\*\*\*   
## (0.002)   
##   
## Constant 5.767   
## (24.079)   
##   
## -----------------------------------------------  
## Observations 807   
## R2 0.045   
## Adjusted R2 0.038   
## Residual Std. Error 13.459 (df = 800)   
## F Statistic 6.300\*\*\* (df = 6; 800)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

R-squared values for both are small and almost the same.