**Module 6 HW**

Professor Nick Williams Fall 2024

Economics 4010, University of Cincinnati

## **What are you being asked to do:**

Estimate and interpret multiple regression models that include dummy variables, and test for significance

## Why is this important?

Creating, using, and interpreting models that include dummy variables is crucial in econometrics. We have practiced how to do this in class, and in our learning exercises. This homework gives you another opportunity to practice.

## Directions

* Turn in your answers in a Word document through Canvas.
* You **DO** need to turn in a copy of your R script for this homework.
* Please note that there are 2 questions on this homework.
* Make sure you look at my example R scripts from the lectures and learning exercise.
* In some circumstance, copying results from RStudio into Word is acceptable, but note that in many instances I ask you to interpret or explain. Below I make it clear when I want to write an answer.

Make sure you read and follow my directions in the companion Word document “Practicing Professionalism” that I distributed with the Module 2 HW. **Important: Lack of producing a neat and organized homework may result in a 10-point deduction from your total score!**

* The homework will be carefully graded out of 100 total points.

## Questions

1. A sample of house sales are given in the data ***housing\_canada.dta***. You have been tasked to investigate how the characteristics of the house affect the sales price. Variables in the data include:

price: sale price of a house, in Canadian dollars

lotsize: the lot size of a property in square feet

bedrooms: number of bedrooms

bathrms: number of full bathrooms

stories: number of stories excluding basement

driveway: dummy, 1 if the house has a driveway

recroom: dummy, 1 if the house has a recreational room

fullbase: dummy, 1 if the house has a full finished basement

gashw: dummy, 1 if the house uses gas for hot water heating

airco: dummy, 1 if there is central air conditioning

garagepl: number of garage places

prefarea: dummy, 1 if located in the preferred neighborhood of the city

As usual, before running any regressions you should use visualizations and descriptive statistics to make sure you understand the distributions of the variables in the data.

The dependent variable in each of the models should be *price*. The two models reported should be:

* Base: Include as independent variables: *lotsize, bedrooms, driveway, prefarea* and *garagepl.*
* GDum: Investigate whether the garage capacity has a nonlinear effect on sale price. To do this, create three dummies from the *garagepl* variable. The first dummy would be =1 if *garagepl* is equal to 1. The second dummy would be = 1 if *garagepl* is equal to 2. The third dummy would be = 1 if *garagepl* is equal to 3. **(Hint: What did I say in the class/synchro session about including both garagepl and the dummies?)**

Create a regressions table with 2 columns using “stargazer” **Turn in this table in your Word document.**

==========================================================

Dependent variable:

---------------------------------------------

price

(1) (2)

----------------------------------------------------------

Constant 1423.46 658.69

(4267.58) (4273.13)

p = 0.74 p = 0.88

lotsize 4.33 4.30

(0.44) (0.44)

p = 0.00 p = 0.00

bedrooms 9969.01 10005.38

(1173.19) (1169.55)

p = 0.00 p = 0.00

driveway 9664.08 9754.42

(2592.78) (2588.75)

p = 0.0003 p = 0.0002

prefarea 11534.77 11516.59

(2081.43) (2077.53)

p = 0.0000 p = 0.0000

garagepl 5528.26

(1064.52)

p = 0.0000

eq1 8720.49

(2144.38)

p = 0.0001

eq2 12076.72

(2357.16)

p = 0.0000

eq3 7122.85

(5931.13)

p = 0.24

----------------------------------------------------------

Observations 546 546

R2 0.46 0.46

F Statistic 90.31\*\*\* (df = 5; 540) 65.73\*\*\* (df = 7; 538)

* 1. In the Base model, interpret the coefficient on *garagepl* and discuss its statistical significance.

The coefficient on garagepl is 5528 which means that as the number of garage spaces increases by 1, all else equal, the price of the house will rise by $5,528 Canadian dollars. Since the pval is 0.000 we can reject the null hypothesis at all significance levels meaning that this is a statistically significant variable.

* 1. In the GDum model
     1. What is the “base” group in this model?

The base group is garagepl since it is the column we are parsing to get our 3 dummy vars, also know as the omitted group.

* + 1. Which of the garage capacity coefficients are individually significant? Briefly explain your answer.

Eq1 and eq2 with pvals of 0.0001 and 0.0000 respectively are both considered statistically significant for alphas >= 0.001 . eq3 has a pval of 0.24 meaning at the 5% significance level we fail to reject the null meaning it is statistically insignificant.

* + 1. Formally test that the set of the three coefficients on the garage capacity dummy variables are jointly equal to zero. You answer should include
       1. The null and alternative hypotheses you are testing

H0 = Heq1 = Heq2 = Heq3 = 0

HA = Heq1 = Heq2 = Heq3 != 0

* + - 1. The value of the F-statistic

10.914 is the F stat when doing lht(eq1,eq2,eq3)

* + - 1. Your decision about significance and the reason for your decision.

The F Stat is larger than DF and there are 3 stars next to it indicating that it is significant at the 0.01 significance level. It also has a very small pval 5.731e-07 which we can assume will be way less than whatever significance level we choose

* 1. Given the results across the two models, how would you answer the questions as to whether garages have a statistically significant effect on sale prices? How about a substantively important effect? Provide an answer in a short paragraph.

Garages certainly have a statistically significant as well as substantively significant impact on sales price, for shared and independent reasons. As far as their statistical significance on sales price this would be due to them passing the F-Test as well as the pval for the dummy regression being small enough to reject the null with certainty. Though it could be noted that when the number of garage spots equals 3 it does fail the pval significance test at .1 . So it could maybe be said it is less significant a metric once there are 2 garage spots but in large it is a statistically significant variable. The coefficient on garagepl in the first regression is $5528.26, all else, equal, which is a pretty big number for a house price to jump with an additional garage place. And for GDum regression which regresses individual number of garage spots, with the addition of 1 garage spot, that is going from 0 spots to 1 spot, all else equal, the salesprice increases by $8720 (statistically significant) which is a huge jump, and then when going from 1 to 2, all else equal, the sales price goes up on average by 12,000 dollars. Finally the when it goes from 2 to 3 garage spots, the salesprice rises by 7k on average, individually this metric is statistically insignificant but nonetheless these are pretty big jumps in sales price considering the average salesprice in the dataset is 68k.

1. Based on Bailey Chapter 6, Number 4. This problem is asking you to investigate the amount of traffic ticket fines received by different demographic groups. The dependent variable is *Amount*. The questions below ask you to run two different regressions. Create a table with 2 columns using “stargazer” for these regressions, **turn in this table in your Word document.**

======================================================================

Dependent variable:

-----------------------------------------------------

Amount

(1) (2)

----------------------------------------------------------------------

Constant 4.85 2.34

(0.98) (1.10)

p = 0.0000 p = 0.04

MPHover 6.88 7.02

(0.04) (0.05)

p = 0.00 p = 0.00

Age 0.03 0.03

(0.02) (0.02)

p = 0.13 p = 0.13

MPHover:black 0.30

(0.16)

p = 0.06

MPHover:hispanic 0.81

(0.16)

p = 0.0000

MPHover:female -0.75

(0.09)

p = 0.00

black -2.03 -7.32

(1.02) (3.00)

p = 0.05 p = 0.02

hispanic 1.93 -12.46

(1.06) (3.08)

p = 0.07 p = 0.0001

female -3.54 8.99

(0.47) (1.52)

p = 0.00 p = 0.00

----------------------------------------------------------------------

Observations 31674 31674

R2 0.50 0.51

F Statistic 6425.81\*\*\* (df = 5; 31668) 4043.47\*\*\* (df = 8; 31665)

======================================================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Standard errors in parentheses

* 1. Use an OLS regression to assess whether there are differences in fines for African-Americans and Hispanics. Your regression should include the dummy for female, as well as controls for age and miles per hour. Also include dummies for African American and Hispanic. Briefly interpret these two dummies, discuss their statistical and economic significance.

The coefficient on African American is -2.029 and the coefficient on Hispanic is 1.932. This means that for a given ticket the average amount is 2 dollars less for an African American and about 2 dollars more for a Hispanic American, with p values of 0.046 and 0.068 respectively. This kind of is dicey due to the fact that it lies so close to the typical significance level of 5%. But nonetheless for a significance level of 0.05 the African American coefficient is deemed statistically significant and we are not able to reject the null for the Hispanic American coefficient. Though because there is minor uncertainty in each I think we can assume that these due play a role in deciding amount but just not as big of a role as other factors. I think these could be deemed statistically significant, though 2 dollars is not a lot I think it is enough of an amount to be mentioned. If it was less than a dollar I would probably change my mind.

* 1. As expected in the first regression, individuals pay a higher fine the more they are exceeding the speed limit. But is the increase in the fine with *MPHover* the same for women, African-Americans and Hispanics? Building from the model in part a), use a 2nd OLS regression to assess this question. (Hint: This would mean interacting these dummies with the miles over the speed limit variable). Using your estimates, demonstrate and briefly explain how much the fines will go up with MPH over the speed limit for these three demographic groups.

For the 3 demographic groups, the fine changes slightly from the base group. The only p value with ambiguity is African American which is 0.059, take what you want from that, depending on your significance level it could reject or fail to reject the null. As far as the coefficients, the base is 7.022 dollars for every 1 MPH increase in speed, all else equal. For the same speed, interacted with also being an African American you pay 0.295 dollars more (~30cents), for a Hispanic American you pay 0.811 dollars more (~81 cents) and for being a female you pay -0.746 (75 cents less!).