**Module 7 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 variables that have been transformed using logs or that include quadratic terms

## Why is this important?

These transformations are frequently used in applications of econometrics to be able to model nonlinear relationships within the linear regression framework. We have discussed how to use and interpret log and quadratic transformations in class, and in our learning exercise. This homework gives you another opportunity to practice.

## Directions

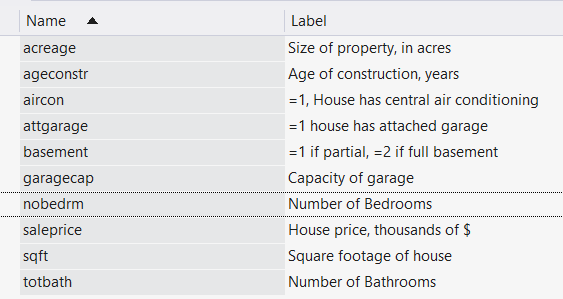
* Turn in your answers in a Word document through Canvas.
* You **DO NOT** need to turn in a copy of your R script for this homework.
* Please note that there are 2 questions on this homework.
* Make sure you look at my example R scripts from the lectures and learning exercise. I am NOT asking you to use any R code that we have not already used in at least one of those R scripts.
* 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\_ohio\_noneg.dta***. You have been tasked to investigate how the characteristics of the house affect the sales price. Variables in the data include:



Prior to your analysis, create the following new variables:

* The log of saleprice
* The log of acreage
* Two dummy variables for basement. The first dummy would be =1 if *basement* is equal to 1. The second dummy would be = 1 if *basement*  is equal to 2.

You should estimate three models:

* Model 1: Dependent variable = *saleprice*

Independent variables = *acreage, ageconstr, nobedrm, totbath, aircon, and your 2 basement dummies.*

* Model 2: Dependent variable = log(*saleprice)*

Independent variables = *acreage, ageconstr, nobedrm, totbath, aircon, and your 2 basement dummies.*

* Model 3: Dependent variable = log(*saleprice)*

Independent variables = log(*acreage), ageconstr, nobedrm, totbath, aircon, and your 2 basement dummies.*

Create a regressions table with 3 columns using “stargazer.” **Turn in this table in your Word document.**

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

Dependent variable:

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

saleprice ln\_saleprice

(1) (2) (3)

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

Constant -110.685 3.864 4.127

(10.832) (0.042) (0.041)

p = 0.000 p = 0.000 p = 0.000

acreage 12.044 0.053

(0.877) (0.003)

p = 0.000 p = 0.000

ln\_acre 0.155

(0.006)

p = 0.000

ageconstr 0.226 -0.003 -0.004

(0.076) (0.0003) (0.0003)

p = 0.004 p = 0.000 p = 0.000

nobedrm 6.556 0.068 0.064

(2.822) (0.011) (0.011)

p = 0.021 p = 0.000 p = 0.000

totbath 106.406 0.287 0.263

(2.103) (0.008) (0.008)

p = 0.000 p = 0.000 p = 0.000

aircon -0.805 0.190 0.243

(6.619) (0.026) (0.025)

p = 0.904 p = 0.000 p = 0.000

eq1 7.668 0.287 0.267

(5.833) (0.022) (0.022)

p = 0.189 p = 0.000 p = 0.000

eq2 32.878 0.325 0.294

(5.919) (0.023) (0.022)

p = 0.00000 p = 0.000 p = 0.000

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

Observations 5500 5500 5500

R2 0.511 0.545 0.572

F Statistic 820.657\*\*\* 938.359\*\*\* 1048.115\*\*\*

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

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

Standard errors in parentheses

* 1. In all three models interpret the coefficient on the *acreage* or log(*acreage*) variable.

Acreage is the size of the property in acres, regressed onto saleprice (or ln(saleprice)) in thousands of dollars.

For model 1 the coefficient on acreage is 12.044, all else equal, with a pval of 0.000 meaning that at any given alpha this would be considered statistically significant. So for every additional acre the saleprice goes up on average by $12,000.

For model 2 we are using the ln(saleprice) for dependent, with a coefficient of 0.053 and a pval of 0.000. This means that for every additional acre, saleprice increases by 5.3% in price.

For model 3 we are using ln(saleprice) and ln(acreage). The coefficient of ln\_acreage is 0.155 which means for every 1% increase of acreage of the property the saleprice goes up by 0.155%

* 1. In Model 1 and Model 2 interpret the coefficient on the dummy for when *basement* is equal to 1.

For model 1 the coefficient on basement == 1 is 7.668, which means that if the basement is partially finished then saleprice on average goes up by $7,668, however the pval is 0.189, so at most tests of significance this variable would be considered insignificant, but due to the fact the entire model passes significance at 0.01 level we will ignore. Model 2 has a coefficient of 0.287 for bathroom == 1. Which means that if the bathroom is partially finished, the saleprice is 28.7% higher, than if not.

* 1. Do you think acreage has a substantively important effect on the sale price of a house? Briefly explain.

Yes, I do, the average salesprice is $258k, with a median of $238k. with each additional acre property saleprice, on average, with all else equal, goes up by $12k, which is also 5.3% of property value. The only argument that could be made against is the majority of houses appear to be below 1 acre in size, so the amount that are affected by this metric is slim.

1. Based on Bailey Chapter 7, Number 3.
   1. Is the effect of age on fines non-linear? Assess this question by regressing *amount* onto *age,* a quadratic age term, and controls for *MPHover, Female, Black, and Hispani*c. Include your regression results in your answer.

|  |
| --- |
| ODEL INFO:  *Observations:* 31674 (36683 missing obs. deleted)  *Dependent Variable:* Amount  *Type:* OLS linear regression  MODEL FIT:  *F*(6,31667) = 5358.445, *p* = 0.000  *R² =* 0.504  *Adj. R² =* 0.504  *Standard errors:OLS*  ----------------------------------------------------  Est. S.E. t val. p  ----------------- -------- ------- --------- -------  (Intercept) 9.867 1.788 5.518 0.000  MPHover 6.873 0.039 177.313 0.000  Age -0.265 0.089 -2.982 0.003  age2 0.004 0.001 3.350 0.001  Black -1.920 1.018 -1.885 0.059  Hispanic 2.068 1.058 1.954 0.051  Female -3.510 0.475 -7.396 0.000  ---------------------------------------------------- |
|  |
| |  | | --- | |  | |

* + 1. Test that the coefficients on the age variables are individually significant.

Linear hypothesis test

Testing for individual significance we check the age and age^2 pvals to determine if they are above or below the significance level to reject the null. We will use a standard alpha of 0.05 Since age pval is 0.003, we can reject the null, assuring that age is statistically significant. Age^2 pval = 0.001, so once again we can reject the null.

* + 1. Test that the coefficients on the age variables are jointly significant.

Hypothesis:

Age = 0

age2 = 0

Model 1: restricted model

Model 2: Amount ~ MPHover + Age + age2 + Black + Hispanic + Female

Res.Df RSS Df Sum of Sq F Pr(>F)

1 31669 49740635

2 31667 49719369 2 21266 6.7725 0.001147 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

This is the linear hypothesis test done using age and age^2, as we can see they are jointly significant at the 0.01 level with an f stat of 6.7725

Beyond looking at the statistical significance of the variables, it can be more difficult to determine whether the nonlinear relationship is economically significant. The next two parts of this question suggest two different ways to potentially explore this issue.

* 1. Use predicted values to examine the relationship. Demonstrate this by sketching the relationship between age and ticket amount from the foregoing quadratic model: calculate the fitted value for a white male with 0 MPHover (probably not many people going zero miles over the speed limit got a ticket, but this simplifies calculations a lot) for ages equal to 20, 25, 30, 35, 40, and 70. Include the sketch and details about the calculations behind how you determined the values for your sketch. Note: You do not need to create this sketch in R. I would suggest that you simply use Excel to calculate and plot the values.

Below is a graph with ages 1 (which makes no sense) to age 80 and a chart with the above listed ages and ticket amounts for reference. The formula I used for this was;

TicketAmt = 9.867 + (-0.265)\*Age + 0.004\*Age^2

This is formula uses coefficients from the above regression listed below Q2, but assumes the driver is a white male going 0 MPHover, ie all categories except for age are nullified. Which leaves us with this. A positive quadratic curve, with the minimum ticket cost being $5.48 at age 33 for white males.

|  |  |
| --- | --- |
| age | amount |
| 20 | 6.167 |
| 25 | 5.742 |
| 30 | 5.517 |
| 35 | 5.492 |
| 40 | 5.667 |
| 70 | 10.917 |

* 1. Calculate the marginal effects of age. This can be done using either Method 1 or Method 2 discussed in the notes.
     1. Calculate the marginal effects for ages of 20, 35 and 70.

Age 20 Marginal effect is 11 cents, which is decreasing as age 19 m.a. was 12 cents as the curve flattens out. Age 35 marginal effect is 1 cent, and is increasing from the minimum of the graph where the slope is flat at age 34. Age 70 marginal effect is 30 cents, and we can see that after about 50 the slope really starts to get steep.

* + 1. Describe how these calculated marginal effects relate to the graph you produced in part b).

These confirm what we thought in part b. We have a positive quadratic where the minimum sits at 34 and on either side the slope increases as we move away from the min. Due to the fact of using the y intercept in part b we get slightly larger values on that graph with no negatives, but when measuring the change in slope it doesn’t really matter as much, here are the graphs when just using age and age^2 coefficient formula

The derivative is linear due to the fact that it’s a quadratic formula, and we can see that the derivative graph = 0 right above age 30