QMB 6316: R for Business Analytics
Department of Economics
College of Business
University of Central Florida
Fall 2024

Assignment 3

Due Friday, November 22, 2024 at 11:59 PM
in digital form on Webcourses.

Question 1:

Use the script `tractor_analysis_A3.R` to produce a more accurate model for used tractor prices with the data from `tractor_full_A3.csv`. Note that this dataset contains a different record of tractor sales than in `tractor_full.csv`, the dataset created in class. The sales data in this new dataset were taken from a town in the midwest, far north of Florida, so the pricing dynamics may well be different. The dataset includes the following variables.

$saleprice_i$	=	the price paid for tractor i in dollars
$horsepower_i$	=	the horsepower of tractor i
age_i	=	the number of years since tractor i was manufactured
$enghours_i$	=	the number of hours of use recorded for tractor i
$diesel_i$	=	an indicator of whether tractor i runs on diesel fuel
fwd_i	=	an indicator of whether tractor i has four-wheel drive
$manual_i$	=	an indicator of whether tractor i has a manual transmission
$johndeere_i$	=	an indicator of whether tractor i is manufactured by John Deere
cab_i	=	an indicator of whether tractor i has an enclosed cab
$spring_i$	=	an indicator of whether tractor i was sold in April or May
$summer_i$	=	an indicator of whether tractor i was sold between June and September
$winter_i$	=	an indicator of whether tractor i was sold between December and March

In this exercise, we will follow an approach different than that taken in class. We will first estimate a model with our choices of functional form, then consider exclusions of insignificant variables from the full model. Note that this approach allows for inclusion of possibly irrelevant variables and avoids excluding any relevant variables.

a) Estimate a regression model that uses all available variables. Make sure to choose a reasonable transformation of the dependent variable, such as the logarithmic transformation, if necessary. Does this seem to be a reasonable model, as opposed to using average tractor prices? That is, is the R^2 or the $\bar{R}^2$ reasonably high?

b) Which variables seem to help explain used tractor prices? Which have positive and negative relationships with tractor prices? Did you find any variables that do not have statistically significant coefficients under this specification?

c) Before making any reductions to your model, revise the model specification first by considering nonlinear specifications. A used tractor dealer tells you that overpowered used tractors are hard to sell, since they consume more fuel. Tractor prices often increase with horsepower, up to a point, but beyond that they decrease. To incorporate this advice, you create and include a variable for squared horsepower.

   i) Hypothesize the signs for the coefficients on horsepower under this scenario. What should be the sign of the coefficient for horsepower? What sign do you expect for squared horsepower?

   ii) Perform 1-sided t-tests of the hypotheses for each of the horsepower coefficients. That is, are the t-statistics higher than 1.96, with the same sign as you hypothesized?

   iii) Compare the model with or without the quadratic functional form for horsepower. Which model do you recommend? Be sure to cite evidence to support your choice. Specifically, check the four specification criteria: statistical significant t-statistics, an increase in R^2, a good theoretical justification, and no large change in the other coefficients.

At each stage, use the best model that you have found so far, incorporating the findings from the answer to the previous question.

d) Again, use the best model that results from the answer to the previous question to address further questions. Use the model you have so far to test that the time of year has no effect on the sale of tractors. That is, test whether the t-statistics on the seasonal indicators are statistical significant. Is there evidence that tractor prices follow a seasonal pattern?

e) Finally, consider another modification to your model. Some say that John Deere tractors hold their value longer than other tractors. This raises the question of whether an additional hour of use affects the value of a John Deere tractor differently than for tractors of other brands. Test the hypothesis, at the 5% level of significance, that the slope for engine hours differs by brand. You can test this by including an interaction with `age:johndeere` in the regression. Does an additional hour of use affect the price of a John Deere tractor differently than tractors of other brands?