Spring 2021 **Firstname M. Lastname**

University of Central Florida
College of Business

QMB 6911
Capstone Project in Business Analytics

Solutions: Problem Set #10

1 Data Description

This analysis follows the script `Tractor_Reg_Model.R` to produce a more accurate model for used tractor prices with the data from `TRACTOR7.csv` in the `Data` folder. 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

I will revisit the recommended linear model from Problem Set #7, which was supported in Problem Sets #8 and #9 by considering other nonlinear specifications within a Generalized Additive Model.

Then I will further investigate this nonlinear relationship by considering the issue of sample selection: John Deere may produce tractors of specific qualities based on their perceived value to typical John Deere customers, in ways that are not represented by the variables in the dataset.

	Model 1
(Intercept)	8.72792^{***}
	(0.10602)
horsepower	0.01112^{***}
	(0.00107)
squared_horsepower	-0.00001^{***}
	(0.00000)
age	-0.03233^{***}
	(0.00358)
enghours	-0.00004^{***}
	(0.00001)
diesel	0.20350^{*}
	(0.09805)
fwd	0.26539^{***}
	(0.05820)
manual	-0.15015^{*}
	(0.06189)
johndeere	0.31872^{***}
	(0.07186)
cab	0.48345^{***}
	(0.07003)
R^2	0.80591
Adj. R^2	0.79935
Num. obs.	276

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Tab. 1: Quadratic Model for Tractor Prices

2 Linear Regression Model

A natural staring point is the recommended linear model from Problem Set #7.

2.1 Quadratic Specification for Horsepower

In the demo for Problem Set #7, we considered the advice of a used tractor dealer who reported that overpowered used tractors are hard to sell, since they consume more fuel. This implies that tractor prices often increase with horsepower, up to a point, but beyond that they decrease. To incorporate this advice, I created and included a variable for squared horsepower. A decreasing relationship for high values of horsepower is characterized by a positive coefficient on the horsepower variable and a negative coefficient on the squared horsepower variable.

The results of this regression specification are shown in Table 1. The squared horsepower variable has a coefficient of $-2.081e - 05$, which is nearly ten times as large as the standard error of $2.199e - 06$, which is very strong evidence against the null hypothesis of a positive or zero coefficient. I conclude that the log of the sale price does decline for large values of horsepower.

With the squared horsepower variable, the $\bar{R}^2$ is 0.764, indicating that it is a much stronger model than the others we considered. The F-statistic is large, indicating that it is a better candidate than the simple average log sale price. The new squared horsepower variable is statistically significant and the theory behind it is sound, since above a certain point, added horsepower may not improve performance but will cost more to operate. This new model is much improved over the previous models with a linear specification for horsepower. Next, I will attempt to improve on this specification, as we did for Problem Set #8.

3 Sample Selection

3.1 Predicting the Selection into Samples

The specification in Table 1 assumes a quadratic functional form for the relationship between price and horsepower, without selecting into samples by brand. To investigate this relationship further, consider the set of variables that are related to whether or not John Deere makes a tractor with the characteristics observed in the dataset.

3.2 Estimating a Sample Selection Model