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, using Tobit models for sample selection.

2.1.1 Separate Models by Brand

To test for many possible differences in models by brand of tractor, Table 2 shows the estimates for two separate models by brand of tractor. Model 1 shows the estimates for the full sample, Model 2 shows the estimates from the full model for John Deere tractors and Model 4 represents all other brands. Models 3 and 5 show the estimates from a reduced version of each model, in which all coefficients are statistically significant. The coefficients appear similar across the two subsamples. Notable differences include the statistical significance for the indicators for four-wheel drive, manual transmission and an enclosed cab. These features seem to change the value of other tractors, but perhaps these coefficients are not measured accurately for the small sample of 39 John Deere tractors.

We can also test for all of the differences at the same time by using an F-test. In this case, the full, unrestricted model has $K = 2 \times 9 = 18$ parameters, one for each variable in two models. The test that all of the coefficients are the same has $M = 9-1 = 8$ restrictions. The one restriction fewer accounts for the John Deere indicator in the full model, which allows for two separate intercepts. The F-statistic has a value of

$$\frac{(RSS_M - RSS)/M}{RSS/(N-K-1)} = \frac{(42.15882 - 41.1432)/3}{41.1432/263} = 0.7929991.$$

This is also a very low value for the F-statistic. There is no evidence to reject the null that all coefficients are equal across both samples and conclude that the John Deere indicator should be the only brand difference left in the model.

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	8.72792***	8.86706***	9.03796***	8.77320***	8.90792***
	(0.10602)	(0.22409)	(0.16430)	(0.12450)	(0.08769)
horsepower	0.01112***	0.01502***	0.01580***	0.01032***	0.01057***
	(0.00107)	(0.00250)	(0.00223)	(0.00119)	(0.00119)
squared_horsepower	−0.00001***	−0.00002***	−0.00002***	−0.00001***	−0.00001***
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
age	−0.03233***	−0.03038**	−0.03295***	−0.03164***	−0.03283***
	(0.00358)	(0.00914)	(0.00738)	(0.00399)	(0.00392)
enghours	−0.00004***	−0.00006*	−0.00006**	−0.00004***	−0.00004***
	(0.00001)	(0.00002)	(0.00002)	(0.00001)	(0.00001)
diesel	0.20350*	0.08485		0.18218	
	(0.09805)	(0.18242)		(0.11984)	
fwd	0.26539***	0.12882		0.29072***	0.30003***
	(0.05820)	(0.15529)		(0.06308)	(0.06296)
manual	−0.15015*	0.06749		−0.17919**	−0.14668*
	(0.06189)	(0.17288)		(0.06743)	(0.06413)
johndeere	0.31872***				
	(0.07186)				
cab	0.48345***	0.32344	0.38517*	0.51732***	0.52756***
	(0.07003)	(0.17555)	(0.16365)	(0.07696)	(0.07688)
R^2	0.80591	0.91993	0.91606	0.77992	0.77769
Adj. R^2	0.79935	0.89858	0.90334	0.77220	0.77090
Num. obs.	276	39	39	237	237

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

Tab. 2: Separate Models by Brand

	Model 1	Model 2
(Intercept)	-0.77781^{*}	-0.51461
	(0.38590)	(0.26756)
horsepower	0.00019	
	(0.00434)	
squared_horsepower	0.00000	
	(0.00001)	
age	0.00852	
	(0.01433)	
enghours	0.00003	
	(0.00004)	
diesel	-0.97286^{**}	-1.01766^{**}
	(0.35873)	(0.34206)
fwd	0.16677	
	(0.24458)	
manual	0.63033^{*}	0.73176^{**}
	(0.29427)	(0.27747)
cab	-0.67541^{*}	-0.44114^{*}
	(0.30337)	(0.21490)
AIC	220.18598	214.08158
BIC	252.76958	228.56319
Log Likelihood	-101.09299	-103.04079
Deviance	202.18598	206.08158
Num. obs.	276	276

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

Tab. 3: Probit Models for Brand Selection of Tractors

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.

Table 3 shows the estimates for a probit model to predict the selection into samples by brand name. Model 1 in Table 3 shows a preliminary probit model to predict the selection indicator, with all the other explanatory variables in the model. John Deere tractors are more likely to be gasoline-powered, have manual transmissions, and less likely to have an enclosed cab. Model 2 shows the result of a variable-reduction exercise to eliminate variables that are not statistically significant. These estimates provide a concise but useful model to indicate the tractor designs that would be favored by John Deere engineers and customers. This model is used to specify the selection equation of the sample selection estimates discussed next.

3.2 Estimating a Sample Selection Model