Spring 2023 **Firstname M. Lastname**

University of Central Florida
College of Business

QMB 6911
Capstone Project in Business Analytics

Solutions: Problem Set #7

1 Data Description

This analysis follows the script `Tractor_Nonlinearity.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 first estimate a model with nonlinear functional forms, and then consider exclusions of insignificant variables from the full model, with nonlinearity taken into account. This approach allows for inclusion of possibly relevant variables and avoids excluding any irrelevant variables.

	Model 1	Model 2
(Intercept)	8.86^{***}	8.89^{***}
	(0.13)	(0.12)
horsepower	0.01^{***}	0.01^{***}
	(0.00)	(0.00)
age	-0.03^{***}	-0.04^{***}
	(0.00)	(0.00)
diesel	0.37^{***}	0.12
	(0.11)	(0.10)
fwd	0.34^{***}	0.35^{***}
	(0.07)	(0.07)
hp_cat50-100		0.41^{***}
		(0.09)
hp_cat100-150		0.69^{***}
		(0.13)
hp_cat150-200		0.62^{**}
		(0.20)
hp_cat200-250		0.69^{**}
		(0.24)
hp_cat250+		0.04
		(0.40)
R^2	0.68	0.75
Adj. R^2	0.68	0.75
Num. obs.	276	276

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

Tab. 1: Log. of Tractor Prices

2 Example with a Categorical Numerical Variable

Suppose for a moment that the horsepower variable were measured as a categorical variable. I collected the observations into bins of with 50 horsepower up to 250, with the rest in the last category. The results in Table 1 show the regression estimates with this categorical variable included in the regression instead of the continuous variable.

Notice that the 50-100-horsepower group is 40% more valuable than those with 50 or less. The relationship levels off between 100 and 250 horsepower. Tractors with more than 250 horsepower are not statistically worth

more than those with 0-50 horsepower, all else equal. This suggests a nonlinear relationship between horsepower and the tractor price.

Since we do, in fact, have the horsepower variable measured as a continuous variable, we can do better.

3 Nonlinear Model Specifications

3.1 Quadratic Specification for Horsepower

Now suppose that a used tractor dealer reports 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.

If we expect a decreasing relationship for high values of horsepower, this would be 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 2.

The squared horsepower variable has a coefficient of $-1.409e{-}05$, which is around six times as large as the standard error of $2.261e-06$, 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 $\bar{R}^2$ has increased substantially to 0.808, indicating that it is a much stronger model. The F-statistic is even larger than before, indicating that it is still 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.

This improved model affords an opportunity to reconsider other variables in the previous models. these models all include an indicator that the tractor has an enclosed cab, which is also statistically significant. The seasonal indicators in Models 1 and 2 are not statistically significant under this specification neither.

	Model 1	Model 2	Model 3
(Intercept)	8.88115^{***}	8.72555^{***}	8.72792^{***}
	(0.11626)	(0.11156)	(0.10602)
horsepower	0.00489^{***}	0.01115^{***}	0.01112^{***}
	(0.00039)	(0.00107)	(0.00107)
age	-0.02962^{***}	-0.03206^{***}	-0.03233^{***}
	(0.00382)	(0.00359)	(0.00358)
enghours	-0.00004^{***}	-0.00004^{***}	-0.00004^{***}
	(0.00001)	(0.00001)	(0.00001)
diesel	0.30937^{**}	0.21453^{*}	0.20350^{*}
	(0.10409)	(0.09854)	(0.09805)
fwd	0.26715^{***}	0.27526^{***}	0.26539^{***}
	(0.06281)	(0.05876)	(0.05820)
manual	-0.16323^{*}	-0.15308^{*}	-0.15015^{*}
	(0.06637)	(0.06209)	(0.06189)
johndeere	0.29829^{***}	0.30972^{***}	0.31872^{***}
	(0.07735)	(0.07236)	(0.07186)
spring	-0.03992	-0.04892	
	(0.06955)	(0.06506)	
summer	-0.06520	-0.05729	
	(0.06819)	(0.06379)	
winter	0.03697	0.04596	
	(0.07634)	(0.07141)	
cab	0.66974^{***}	0.47786^{***}	0.48345^{***}
	(0.06759)	(0.07031)	(0.07003)
squared_horsepower		-0.00001^{***}	-0.00001^{***}
		(0.00000)	(0.00000)
R^2	0.77921	0.80761	0.80591
Adj. R^2	0.77001	0.79884	0.79935
Num. obs.	276	276	276

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

Tab. 2: Quadratic Models for Tractor Prices

3.2 Seasonality with the Quadratic Specification for Horsepower

The seasonal indicators in Model 3 of Table 2 are not statistically significant individually. It is possible, however, that jointly, they offer an improvement in prediction. This can be tested with an F-test to test the joint hypothesis that the time of year has no effect on the sale of tractors. The null hypothesis is the joint hypothesis that all coefficients on spring, summer and winter are equal to zero. The alternative hypothesis is that one of these coefficients is nonzero.

From the script `Tractor_Reg_Models.R`, the Residual Sum of Squares from the unconstrained model (the model which includes the seasonal indicators) is 41.78944. The constrained model is the one that excludes seasonal indicators and it has a Residual Sum of Squares of 42.15882.

The F-statistic has a value of

$$\frac{(RSS_M - RSS)/M}{RSS/(N-K)} = \frac{(42.15882 - 41.78944)/3}{RSS/263} = 0.7748937.$$

since $N = 276$ observations, $K = 13$ variables (including the intercept) and $M = 3$ restrictions, one for each seasonal indicator excluded. This is a low value compared to the critical value of 2.60 for the F-statistic with 3 degrees of freedom in the numerator and $263(> 120)$ degrees of freedom in the denominator. There is no evidence to reject the null that all seasonal indicators have coefficients of zero and conclude that the seasonal indicators should be left out of the model. The results of the test above indicate that tractor prices do not follow a seasonal pattern.

Since this restriction reduces our model until all coefficients are significant, we can consider adding higher-order features to the model.

3.3 Interaction Terms

3.3.1 Durability of Engine Types

Now consider a higher-order modification to the model. Diesel engines tend to be more durable than gasoline engines. This raises the question of whether an additional hour of use affects the value of a diesel tractor differently than for a gasoline tractor. This is tested in Model 1 of Table 3.

This hypothesis is a test of the *interaction* of the diesel indicator and the slope on engine hours. Given the above result, this test should be conducted with the model that excludes the seasonal indicators. The coefficient on `enghours:diesel` is $4.116e-06$ with a standard error of $2.736e-05$, resulting in a t-statistic of 0.150. Since this is a very low value, we cannot reject the null hypothesis that an additional hour of use affects the value of a diesel tractor the same as that for a gasoline tractor. Note that this conclusion does not change if you test a one-sided hypothesis.

Furthermore, the $\bar{R}^2$ statistic decreases with the inclusion of this variable. The F-statistic is high and statistically significant, indicating that this model is better than the simple average but so is the model without this new variable. Finally, the estimates of the other coefficients change very little when this variable is omitted. The theory may be sound but there is nothing else to support the inclusion of this new variable.

3.3.2 Differences in Value of Characteristics by Brand

The remaining columns of Table 3 show the results of tests for interactions between the John Deere indicator variable on the effects of age, engine hours and horsepower. There seems to be no evidence for relationships that differ by brand name. In this table, we have investigated several individual types of differences by brand.

	Model 1	Model 2	Model 3	Model 4
(Intercept)	8.73714^{***}	8.72858^{***}	8.73093^{***}	8.74493^{***}
	(0.12263)	(0.10630)	(0.10718)	(0.10691)
horsepower	0.01111^{***}	0.01111^{***}	0.01113^{***}	0.01100^{***}
	(0.00107)	(0.00107)	(0.00107)	(0.00107)
squared_horsepower	-0.00001^{***}	-0.00001^{***}	-0.00001^{***}	-0.00001^{***}
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
age	-0.03232^{***}	-0.03234^{***}	-0.03257^{***}	-0.03209^{***}
	(0.00359)	(0.00359)	(0.00377)	(0.00358)
enghours	-0.00005	-0.00004^{***}	-0.00004^{***}	-0.00004^{***}
	(0.00003)	(0.00001)	(0.00001)	(0.00001)
diesel	0.19339	0.20462^{*}	0.20203^{*}	0.19327
	(0.11903)	(0.09849)	(0.09847)	(0.09835)
fwd	0.26589^{***}	0.26499^{***}	0.26529^{***}	0.26801^{***}
	(0.05840)	(0.05837)	(0.05831)	(0.05820)
manual	-0.15065^{*}	-0.14954^{*}	-0.14825^{*}	-0.15257^{*}
	(0.06209)	(0.06213)	(0.06267)	(0.06188)
johndeere	0.32106^{***}	0.30641^{**}	0.29296^{*}	0.23158^{*}
	(0.07367)	(0.10705)	(0.14255)	(0.10282)
cab	0.48311^{***}	0.48420^{***}	0.48386^{***}	0.48270^{***}
	(0.07019)	(0.07033)	(0.07018)	(0.06998)
enghours:diesel	0.00000			
	(0.00003)			
enghours:johndeere		0.00000		
		(0.00002)		
age:johndeere			0.00146	
			(0.00695)	
horsepower:johndeere				0.00085
				(0.00072)
R^2	0.80593	0.80593	0.80594	0.80693
Adj. R^2	0.79861	0.79861	0.79862	0.79965
Num. obs.	276	276	276	276

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

Tab. 3: Regression Models for Tractor Prices with Interactions

3.3.3 Separate Models by Brand

To test for many possible differences in models by brand of tractor, Table 4 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 the type of fuel. 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. Notice the much-improved fit of the model for John Deere tractors, at the expense of some performance for the larger sample of other tractors.

We can also test for all of the differences at the same time by using an F-test. In the textbook case, when both models include all variables, the full, unrestricted model has $K = 2 \times 10 = 20$ parameters, one for each of the ten variables in two models. In this case, with a reduced model for each subsample, the full, unrestricted model has $K = 6 + 8 = 14$ parameters, one for each variable in each of the two models. The test that all of the coefficients are the same has $M = 14 - 10 = 4$ 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

$$\frac{(RSS_M - RSS)/M}{RSS/(N-K)} = \frac{(42.15882 - 41.70574)/4}{41.70574/262} = 0.7115771.$$

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.

Notice the the conclusion has changed from our specification that excluded all forms of nonlinearity. This can occur if the relationship is nonlinear and each subsample has a different distribution of an explanatory variable, with mass over different intervals, where the slope is different. In this example, more of the John Deere tractors have high horsepower, for which the added benefit is lower than that for tractors with lower horsepower. The separate nonlinear models compensate for the lack of fit of a linear model for modeling the nonlinear relationship between prices and horsepower.

	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. 4: Separate Partially Nonlinear Models by Brand