Spring 2023 **Firstname M. Lastname**

University of Central Florida
College of Business

QMB 6911
Capstone Project in Business Analytics

Solutions: Problem Set #6

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 first estimate a model with our choices of functional form, and then consider exclusions of insignificant variables from the full model. This approach allows for inclusion of possibly irrelevant variables and avoids excluding any relevant variables.

2 Choosing the Dependent Variable

Before we begin, I review the evidence for the suitability of the dependent variable without transformation and compare that with the logarithmic transformation. Although, in this case, this decision is fairly clearly made by plotting the dependent variable alone, in many cases, the decision is not so clear and other forms of evidence can be considered once building a model.

2.1 Univariate Analysis

Figure 1 shows a histogram of tractor prices. The distribution is highly skewed to the right, with most tractors selling for about $10,000 or less, and very few tractors priced above $50,000. This is a highly skewed distribution, which might influence the estimates of parameters in the model.

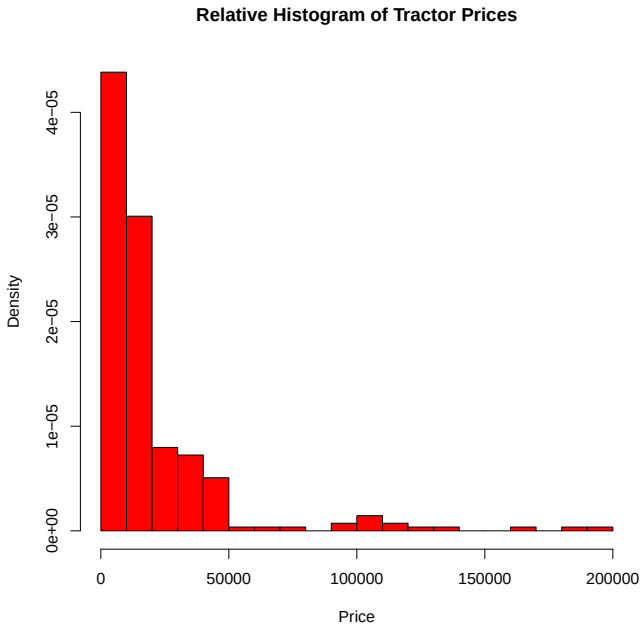

Fig. 1: Histogram of Tractor Prices

As a comparison, Figure 2 shows the histogram of the natural logarithm of price.

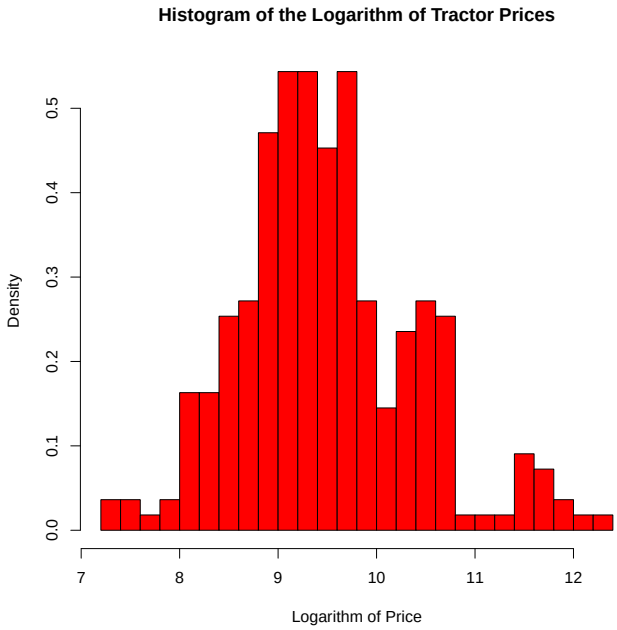

Fig. 2: Histogram of the Logarithm of Tractor Prices

This is much better behaved. The distribution looks almost normal. So far it looks as if the logarithm of the sale price is the more promising variable. Another approach to making this decision is to build a model under each alternative and judge the validity of those results.

2.2 Linear Regression Models of Tractor Prices

2.2.1 Predicting Price Levels

First, I will build a model of the price of a used tractor, ignoring the above evidence that the distribution is highly skewed.

The results of Model 1 in Table 1 shows the effect of the variables on the dollar price of the used tractors.

From the coefficients in the table, it appears that a John Deere tractor sells for $12,200 more than an equivalent tractor of another brand. This prediction applies equally for tractors all across the spectrum, from the To put a finer point on it, a 16 horsepower lawn tractor that would otherwise sell for $2,000 is expected to command $14,200 if it is a John Deere. Clearly, this is an unreasonable expectation and a quick search on your browser will confirm that the John Deere premium is more modest.

2.2.2 Predicting Logarithm of Prices

Next, I will build a model of the logarithm of the price of a used tractor, which is consistent with the univariate analysis we conducted earlier.

The results of Model 2 in Table 1 shows the effect of the variables on the logarithm of the dollar price of the used tractors. This specification calculates coefficients that approximately represent percentage changes in tractor prices.

From the coefficients in the table, it appears that a John Deere tractor sells for 17% more than an equivalent tractor of another brand. That is, a tractor worth $1,700 would sell for $2,000 if it is a John Deere, which is clearly more reasonable. This more sensible interpretation supports the strategy of modeling the logarithm of the tractor price.

	Model 1	Model 2
(Intercept)	11670.40884^{*}	8.76953^{***}
	(4519.28608)	(0.13528)
horsepower	246.39828^{***}	0.00654^{***}
	(13.98454)	(0.00042)
age	-674.63576^{***}	-0.02754^{***}
	(148.99958)	(0.00446)
enghours	-1.75752^{***}	-0.00002
	(0.39558)	(0.00001)
diesel	2731.97348	0.49917^{***}
	(3996.04648)	(0.11962)
fwd	2570.56963	0.35672^{***}
	(2427.51242)	(0.07266)
manual	-3713.28280	-0.12167
	(2586.95528)	(0.07744)
johndeere	12194.22591^{***}	0.17253
	(2979.90884)	(0.08920)
spring	-1721.00588	-0.03210
	(2716.03441)	(0.08130)
summer	-5569.45586^{*}	-0.11876
	(2654.93625)	(0.07947)
winter	-1541.98750	0.04009
	(2981.31036)	(0.08924)
R^2	0.64748	0.69709
Adj. R^2	0.63418	0.68566
Num. obs.	276	276

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

Tab. 1: Linear and Logarithmic Models of Tractor Prices

3 Model Specification

3.1 Variable Reduction

Next, I can refine the model by removing some explanatory variables that do not have string predictive value. The first candidates are those with coefficients that are not statistically significant. The results in Table 2

The first column of Table 2 shows the results from the original model of the logarithm of tractor prices in Table 1. The coefficients for seasonal indicators, engine hours and manual transmission are not significant. The John Deere indicator is not significant but since it is a key empirical question, I include it, regardless. The second column shows the model without the seasonal indicators. We see an improvement in significance of some variables with minimal loss of predictive ability. Removing the variables in the other order, I removed the indicator for manual transmission and left in the seasonal indicators; this specification appears in the next column. In the last column, I remove the indicator for manual transmission. All these changes have a similar effect on the quality of the model.

With the squared horsepower variable, the $\bar{R}^2$ has increased substantially to 0.764, 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. Models 2 and 3 both include an indicator that the tractor has an enclosed cab, which is also statistically significant. The seasonal indicators in Model 2 are not statistically significant under this specification neither.

	Model 1	Model 2	Model 3	Model 4
(Intercept)	8.7695^{***}	8.7559^{***}	8.7964^{***}	8.7787^{***}
	(0.1353)	(0.1284)	(0.1346)	(0.1279)
horsepower	0.0065^{***}	0.0066^{***}	0.0065^{***}	0.0065^{***}
	(0.0004)	(0.0004)	(0.0004)	(0.0004)
age	-0.0275^{***}	-0.0278^{***}	-0.0295^{***}	-0.0297^{***}
	(0.0045)	(0.0044)	(0.0043)	(0.0043)
enghours	-0.0000	-0.0000	-0.0000	-0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
diesel	0.4992^{***}	0.4884^{***}	0.4321^{***}	0.4246^{***}
	(0.1196)	(0.1192)	(0.1121)	(0.1118)
fwd	0.3567^{***}	0.3492^{***}	0.3483^{***}	0.3416^{***}
	(0.0727)	(0.0721)	(0.0727)	(0.0721)
manual	-0.1217	-0.1169		
	(0.0774)	(0.0773)		
johndeere	0.1725	0.1858^{*}	0.1577	0.1707
	(0.0892)	(0.0888)	(0.0889)	(0.0885)
spring	-0.0321		-0.0356	
	(0.0813)		(0.0815)	
summer	-0.1188		-0.1190	
	(0.0795)		(0.0797)	
winter	0.0401		0.0316	
	(0.0892)		(0.0893)	
R^2	0.6971	0.6934	0.6943	0.6907
Adj. R^2	0.6857	0.6853	0.6839	0.6838
Num. obs.	276	276	276	276

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

Tab. 2: Models for the Log. of Tractor Prices

3.2 Seasonality with the Quadratic Specification for Horsepower

The seasonal indicators in Model 2 of Table **??** 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

$$\frac{(RSS_M - RSS)/M}{RSS/(N-K-1)} = \frac{(42.15882 - 41.78944)/3}{RSS/263} = 0.7748937.$$

since $N = 276$ observations, $K = 12$ variables 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.

3.3 Interaction Terms

3.3.1 Durability of Engine Types

Finally, I consider another modification to your 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 Depreciation 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.

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 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
(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

	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 Models by Brand