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.

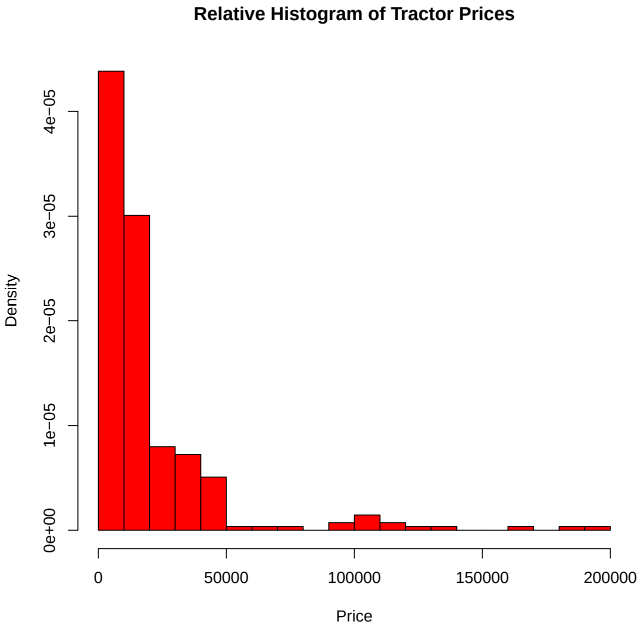

Fig. 1: Histogram of Tractor Prices

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

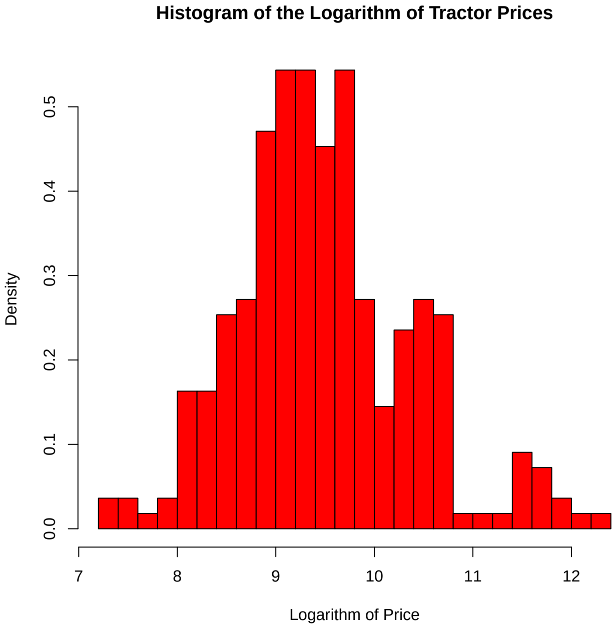

Fig. 2: Histogram of the Logarithm of Tractor Prices

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, even if the pricing difference is not statistically significant in this preliminary model.

	Prices	Log. Prices
(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, so I include it, regardless. I removed the indicator for manual transmission and left in the seasonal indicators; this specification appears in the second column. The next column shows the model without the seasonal indicators. Sometimes we see an improvement in significance of some variables with minimal loss of predictive ability after removing the first few variables. In the fourth column, I remove the variable for the engine hours. The coefficient for the value of the John Deere brand became insignificant as the model was trimmed down. Still, the coefficient is consistently in the neighborhood of 16–18%, which gives an indication of the value of the brand name, if only innaccurately. We can revisit this question after improving the quality of the model. For now, the best model for the full sample of tractors is the fully-reduced Model 5.

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	8.7695^{***}	8.7964^{***}	8.7787^{***}	8.8143^{***}	8.8629^{***}
	(0.1353)	(0.1346)	(0.1279)	(0.1271)	(0.1250)
horsepower	0.0065^{***}	0.0065^{***}	0.0065^{***}	0.0062^{***}	0.0062^{***}
	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
age	-0.0275^{***}	-0.0295^{***}	-0.0297^{***}	-0.0342^{***}	-0.0337^{***}
	(0.0045)	(0.0043)	(0.0043)	(0.0036)	(0.0036)
enghours	-0.0000	-0.0000	-0.0000		
	(0.0000)	(0.0000)	(0.0000)		
diesel	0.4992^{***}	0.4321^{***}	0.4246^{***}	0.4152^{***}	0.3737^{***}
	(0.1196)	(0.1121)	(0.1118)	(0.1122)	(0.1105)
fwd	0.3567^{***}	0.3483^{***}	0.3416^{***}	0.3354^{***}	0.3392^{***}
	(0.0727)	(0.0727)	(0.0721)	(0.0724)	(0.0727)
manual	-0.1217				
	(0.0774)				
johndeere	0.1725	0.1577	0.1707	0.1662	
	(0.0892)	(0.0889)	(0.0885)	(0.0889)	
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	0.6943	0.6907	0.6865	0.6825
Adj. R^2	0.6857	0.6839	0.6838	0.6807	0.6778
Num. obs.	276	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 of the Pricing of Tractors

In the variable redulction exercise of Table 2, the seasonal indicators were not statistically significant individually. A valid approach is to test the joint hypothesis of the three exclusions together with an F-test. This is a more statistically rigorous way 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

$$\frac{(RSS_M - RSS)/M}{RSS/(N-K)} = \frac{(67.17639 - 66.40924)/3}{RSS/263} = 1.024257.$$

since $N = 276$ observations, $K = 10$ 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 we conclude that the seasonal indicators should be left out of the model. The results of the test above coincide with the conclusion drawn from the individual t-statistics: both indicate that tractor prices do not follow a seasonal pattern.

3.3 Interaction Terms

3.3.1 Differences in Parameters by Brand

The columns of Table 3 show the results of tests for interactions between the John Deere indicator variable on the effects of horsepower, engine hours, age and whether the tractor has a manual transmission. There seems to be no evidence for relationships that differ by brand name.

	Model 1	Model 2	Model 3	Model 4
(Intercept)	8.89127^{***}	8.88028^{***}	8.87926^{***}	8.92856^{***}
	(0.11165)	(0.11042)	(0.11163)	(0.11314)
horsepower	0.00476^{***}	0.00490^{***}	0.00489^{***}	0.00484^{***}
	(0.00043)	(0.00039)	(0.00039)	(0.00039)
age	-0.02969^{***}	-0.02994^{***}	-0.02990^{***}	-0.02948^{***}
	(0.00381)	(0.00381)	(0.00400)	(0.00379)
enghours	-0.00004^{***}	-0.00004^{***}	-0.00004^{***}	-0.00004^{***}
	(0.00001)	(0.00001)	(0.00001)	(0.00001)
diesel	0.29293^{**}	0.30271^{**}	0.29874^{**}	0.26263^{*}
	(0.10387)	(0.10378)	(0.10389)	(0.10492)
fwd	0.26050^{***}	0.25727^{***}	0.25879^{***}	0.26183^{***}
	(0.06226)	(0.06230)	(0.06229)	(0.06192)
manual	-0.16217^{*}	-0.15793^{*}	-0.16018^{*}	-0.19452^{**}
	(0.06619)	(0.06632)	(0.06692)	(0.06841)
johndeere	0.24957^{*}	0.25824^{*}	0.30345^{*}	0.05400
	(0.10998)	(0.11399)	(0.15228)	(0.15882)
cab	0.67665^{***}	0.67717^{***}	0.67571^{***}	0.67563^{***}
	(0.06722)	(0.06728)	(0.06734)	(0.06686)
horsepower:johndeere	0.00057			
	(0.00077)			
enghours:johndeere		0.00001		
		(0.00002)		
age:johndeere			0.00023	
			(0.00742)	
manual:johndeere				0.32327
				(0.17752)
R^2	0.77811	0.77795	0.77766	0.78040
Adj. R^2	0.77060	0.77043	0.77014	0.77297
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	Model 6
(Intercept)	8.86286^{***}	9.10365^{***}	9.43342^{***}	9.45837^{***}	8.89978^{***}	9.11582^{***}
	(0.12501)	(0.26487)	(0.18213)	(0.18093)	(0.12806)	(0.08270)
horsepower	0.00619^{***}	0.00536^{***}	0.00572^{***}	0.00553^{***}	0.00475^{***}	0.00473^{***}
	(0.00038)	(0.00094)	(0.00095)	(0.00093)	(0.00044)	(0.00044)
age	-0.03368^{***}	-0.03723^{**}	-0.03696^{***}	-0.04338^{***}	-0.02847^{***}	-0.03007^{***}
	(0.00360)	(0.01099)	(0.00941)	(0.00722)	(0.00413)	(0.00410)
diesel	0.37369^{***}	0.26933			0.27337^{*}	
	(0.11051)	(0.21625)			(0.12445)	
fwd	0.33924^{***}	0.09401			0.28380^{***}	0.29761^{***}
	(0.07268)	(0.18980)			(0.06626)	(0.06651)
enghours		-0.00002	-0.00003		-0.00004^{***}	-0.00004^{**}
		(0.00003)	(0.00003)		(0.00001)	(0.00001)
manual		0.17233			-0.19998^{**}	-0.15152^{*}
		(0.20925)			(0.07072)	(0.06775)
cab		0.53380^{*}	0.69193^{**}	0.67384^{**}	0.69669^{***}	0.72527^{***}
		(0.20531)	(0.19312)	(0.19272)	(0.07147)	(0.07086)
R^2	0.68248	0.87602	0.85776	0.85305	0.75596	0.75082
Adj. R^2	0.67779	0.84802	0.84103	0.84046	0.74850	0.74432
Num. obs.	276	39	39	39	237	237

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

Tab. 4: Separate Models by Brand

3.3.2 Separate Models by Brand

In Table 3, we 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, Models 2, 3 and 4 show the estimates from two models for John Deere tractors and Models 5 and 6 represent all other brands. Models 3 and 5 show the estimates from the best, reduced version of each model, in which all coefficients are statistically significant. The coefficients appear similar across the two subsamples, when the coefficients are significant. 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. We do so using a comparison between the best models for each subsample and the best model for the full sample. In this case, the full, unrestricted model has $K = 4 + 8 = 12$ parameters, one for each variable in the two models, including the intercepts. The test that all of the coefficients are the same has $M = 12 - 5 = 7$ restrictions, since the reduced model on the full sample has 5 parameters. 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

$$\frac{(RSS_M - RSS)/M}{RSS/(N-K)} = \frac{(68.97084 - 48.38915)/7}{48.38915/264} = 16.04128.$$

This is a very high value for the F-statistic. This is evidence to reject the null that all coefficients are equal across both samples and conclude that several parameters should be estimated separately depending on whether the tractor was made by John Deere .