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_Box_Tidwell.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 #6, augmented with a quadratic specification for horsepower. This allowed for an increasing relationship between price and horsepower, for tractors with low horsepower, but a decreasing relationship for the tractors with high horsepower. In doing so, I will further investigate nonlinear relationships by incorporating another nonlinear but parametric specification for the value of horsepower. This parametric analysis will be performed using the Box-Tidwell framework to investigate whether the value of these characteristics are best described with parametric nonlinear forms.

	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 #6, augmented with the quadratic specification for horsepower.

2.1 Quadratic Specification for Horsepower

In the demo for Problem Set #6, 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 Nonlinear Specifications

3.1 The Box–Tidwell Transformation

The Box–Tidwell function tests for non-linear relationships to the mean of the dependent variable. The nonlinearity is in the form of an exponential transformation in the form of the Box-Cox transformation, except that the transformation is taken on the explanatory variables.

3.1.1 Transformation of Horsepower

Performing the transformation on the horsepower variable produces a modified form of the linear model. This specification allows a single exponential transformation on horsepower, rather than a quadratic form.

```
 MLE of lambda Score Statistic (z)  Pr(>|z|)
       0.11437             -7.3864 1.509e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

iterations =  5
```

The R output is the statistics for a test of nonlinearity: that the exponent λ in the Box–Tidwell transformation is zero. The ″`MLE of lambda`″ statistic is the optimal exponent on horsepower. Similar to the Box-Cox transformation, with Box-Tidwell, the exponents are on the explanatory variables and are all called lambda, in contrast to the parameter τ in our class notes. The exponent is significantly different from 0, although it is a small positive value, which suggests an increasing relationship for the value of horsepower with a slope that is sharply declining. Next I consider the possibility of a changing relationship for the next continuous variable.

3.1.2 Transformation of Age

```
 MLE of lambda Score Statistic (z) Pr(>|z|)
        0.9815              0.0421   0.9664

iterations =  3
```

This coefficient is effectively 1, which is more evidence of a purely linear relationship between `log_saleprice` and age: the percentage depreciation rate is constant. Next, I will consider the possibility of nonlinearity in depreciation from hours of use.

3.1.3 Transformation of Engine Hours

```
 MLE of lambda Score Statistic (z) Pr(>|z|)
        1.3578             -0.9646   0.3348

iterations =  3
```

Although $\hat{\lambda}$ is not statistically significant, this suggests a moderately increasing relationship between the log of tractor prices and engine hours, which means that tractors with high hours of use depreciate more quickly with each additional hour of use.

Since a nonlinear relationship was detected with horsepower, I will next estimate a model with nonlinearity in all three continuous variables.

3.1.4 Transformation of All Three Continuous Variables

```
           MLE of lambda Score Statistic (z)  Pr(>|z|)
horsepower        0.1153             -7.1510 8.615e-13 ***
age               1.1183             -0.0489    0.9610
enghours          1.1043             -0.5379    0.5907
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

iterations =  6
```

The performance is similar to the other models with forms of nonlinearity for the value of horsepower. Now consider the full set of such models in a table for a final comparison.

	Model 1	Model 2
(Intercept)	8.72792^{***}	3.09024^{***}
	(0.10602)	(0.39174)
horsepower	0.01112^{***}	
	(0.00107)	
squared_horsepower	-0.00001^{***}	
	(0.00000)	
age	-0.03233^{***}	-0.02927^{***}
	(0.00358)	(0.00345)
enghours	-0.00004^{***}	-0.00005^{***}
	(0.00001)	(0.00001)
diesel	0.20350^{*}	0.12070
	(0.09805)	(0.09500)
fwd	0.26539^{***}	0.32602^{***}
	(0.05820)	(0.05617)
manual	-0.15015^{*}	-0.20053^{**}
	(0.06189)	(0.06031)
johndeere	0.31872^{***}	0.33386^{***}
	(0.07186)	(0.06967)
cab	0.48345^{***}	0.42139^{***}
	(0.07003)	(0.06768)
horsepower_bt		3.99759^{***}
		(0.25577)
R^2	0.80591	0.81613
Adj. R^2	0.79935	0.81062
Num. obs.	276	276

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

Tab. 2: Alternate Models for Tractor Prices

4 Final Comparison of Candidate Models

I created a variable `horsepower_bt` by raising horsepower to the optimal exponent $\hat{\lambda} = 0.1143693$. Then, I included this variable in the place of the horsepower variables a the linear regression model. Table 2 collects the results of the set of models from the three forms of nonlinearity. Model 1 is the linear regression model with a quadratic form for horsepower. Model 2 has the same specification as the other one, except that the horsepower variable is transformed using the optimal exponent for the Box-Tidwell transformation. The last model has the highest R-squared among the ones we have estimated. Again, the differences are marginal, so the practical recommendation is the model with the quadratic relationship for horsepower, which has a simpler interpretation. In either case, we conclude that John Deere tractors are worth approximately thirty percent more valuable than an equivalent tractor of another brand. Compare this with the lower premium of 17%, which was not even statistically significant, when we estimated a simpler, linear specification in which we ignored the nonlinearity in the model.