Spring 2021 **Firstname M. Lastname**

University of Central Florida
College of Business

QMB 6911
Capstone Project in Business Analytics

Solutions: Problem Set #9

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 included 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. I augmented this model in the demonstration for Problem Set #8 by considering semiparametric specifications within a Generalized Additive Model.

Then I will further investigate this nonlinear relationship by incorporating a 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 #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, as we did for Problem Set #8.

3 Nonlinear Specifications

3.1 Nonparametric Specification for Horsepower

The specification in Table 1 assumes a quadratic functional form for the relationship between price and horsepower. To consider the horsepower variable alone, while accounting for the effects of other variables, one can fit a nonparametric model to the residuals from a model of tractor prices, after regressing tractor prices on the other variables. This leaves only the variation in tractor prices that is not explained by the other variables. Going one step further, perform the same transformation to the horsepower variable: take the residuals from a model of horsepower, after regressing horsepower on the other variables. This allows a model that would fit exactly the same as if it were estimated within a full model with all variables included.

I first conducted FWL regressions to reduce the problem to two dimensions. The results are not shown here, since the comparison only verifies the conclusion of the FWL theorem.

To illustrate the fit of the model, **??** shows a scatter plot of the residual log prices on residuals from the regression for horsepower: the "excess horsepower" compared to what would be expected given the other characteristics of a tractor. The observations are shown in blue and the fitted values are shown in red. The quadratic function is clear from this perspective, except that we observe variation in the fitted values results from the two-dimensional nature of the horsepower variable when we consider the quadratic form.

I next considered a nonparametric specification for the relationship between prices and horsepower. Figure 1 overlays the nonparametric estimate (shown in green) compared to the linear model. The pattern has more variation in slope but closely follows the prediction from the quadratic model. So far, it appears that the quadratic form is close enough.

Fig. 1: Nonparametric Model for Tractor Prices: Excess Horsepower

3.2 Nonparametric Specification for Age

As above, first conduct FWL regressions to reduce the problem to two dimensions. To illustrate the fit of the model, Figure 2 shows a scatter plot of the residual log prices on the residuals from the regression for age: the "excess age" of a tractor compared to what would be expected given the other characteristics of the tractor. The observations are shown in blue and the fitted values are shown in red.

Next we considered a nonparametric specification for the relationship between prices and age. Figure 2 overlays the nonparametric estimate (shown in green) with the linear model. The pattern has more variation in slope but closely follows the prediction from the linear model. Although the nonparametric estimate varies around the linear estimate, it appears that the linear form is a close enough approximation without the added complexity. Next, I will revisit the remaining continuous variable.

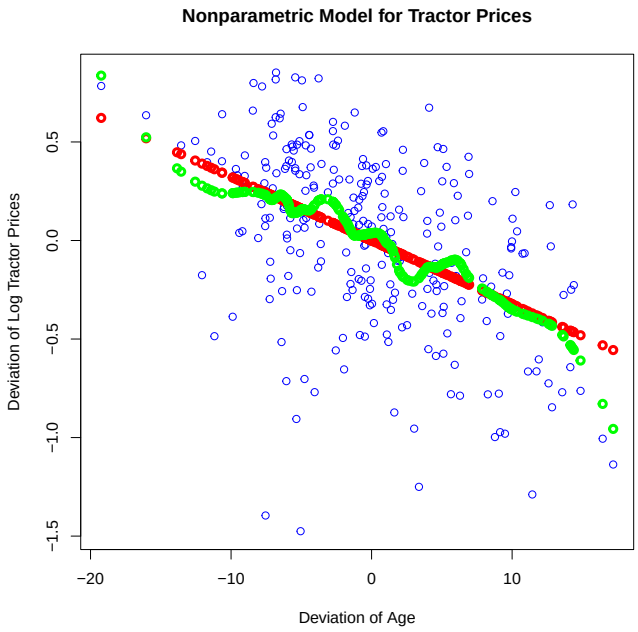

Fig. 2: Nonparametric Model for Tractor Prices: Excess Age

3.3 Nonparametric Specification for Engine Hours

As above, first conduct FWL regressions to reduce the problem to two dimensions. To illustrate the fit of the model, Figure 3 shows a scatter plot of the residual log prices on residuals from the regression for engine hours: the "excess engine hours" of a tractor compared to what would be expected given the other characteristics of the tractor. The observations are shown in blue and the fitted values are shown in red. As with age, the linear fit follows a straight line, since we have a single variable with no quadratic transformation. I moved directly to the nonparametric specification for the relationship between prices and engine hours. Figure 3 overlays the nonparametric estimate, shown in green. The pattern has more variation in slope but closely follows the prediction from the linear model. Although the nonparametric estimate varies around the linear estimate, it appears that the linear form is also a close enough approximation, just as was found for the age variable.

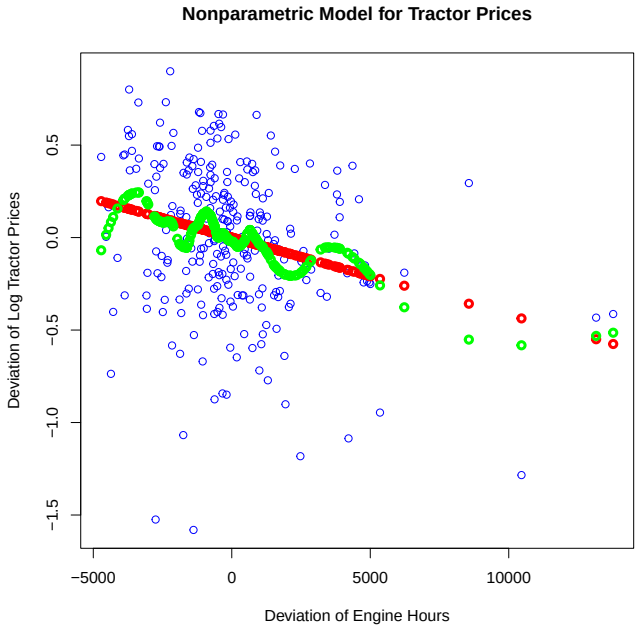

Fig. 3: Nonparametric Model for Tractor Prices: Excess Engine Hours

4 Semiparametric Estimates

As I was building the above nonparametric models, I stored the predictions and will now use them as variables in linear models. Table 2 shows the estimates from a set of models. Model 1 is the benchmark linear model in Table 1. Model 2 is a semi-parametric model with a nonparametric fit on horsepower substituted in for the horsepower variables. Models 3 and 4 are semi-parametric models with nonparametric fits on age and engine hours, respectively. Model 5 is a maximally semiparametric model, with nonparametric fits for all continuous variables. For each of the single-variable semiparametric models, the coefficients are near one and the fits are similar to the linear model. Even with maximal flexibility, the fit of Model 5 is not much better than the benchmark linear model. Across all models, the adjusted $\bar{R}^2$ values are all hovering around 0.80. All things considered, these are excellent models and the linear model is sufficient.

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	8.72792***	8.97543***	8.26683***	8.79186***	8.28804***
	(0.10602)	(0.10479)	(0.09102)	(0.10295)	(0.08340)
horsepower	0.01112***		0.01087***	0.01020***	
	(0.00107)		(0.00105)	(0.00103)	
squared_horsepower	−0.00001***		−0.00001***	−0.00001***	
	(0.00000)		(0.00000)	(0.00000)	
age	−0.03233***	−0.03813***		−0.04025***	
	(0.00358)	(0.00360)		(0.00306)	
enghours	−0.00004***	0.00000	−0.00009***		
	(0.00001)	(0.00001)	(0.00001)		
diesel	0.20350*	0.31981**	0.31266**	0.22271*	0.49492***
	(0.09805)	(0.09872)	(0.09593)	(0.09617)	(0.09408)
fwd	0.26539***	0.39101***	0.46564***	0.24905***	0.69747***
	(0.05820)	(0.05901)	(0.05214)	(0.05709)	(0.04973)
manual	−0.15015*	−0.06208	−0.29946***	−0.15841**	−0.31240***
	(0.06189)	(0.06285)	(0.05875)	(0.06067)	(0.05689)
johndeere	0.31872***	0.40778***	0.29792***	0.30143***	0.35156***
	(0.07186)	(0.07313)	(0.07098)	(0.07039)	(0.07014)
cab	0.48345***	1.05513***	0.47953***	0.46718***	0.96339***
	(0.07003)	(0.05806)	(0.06920)	(0.06842)	(0.05248)
horsepower_np		0.96671***			0.97150***
		(0.06886)			(0.06845)
age_np			0.98471***		1.50214***
			(0.10383)		(0.11402)
eng_np				1.05037***	1.46063***
				(0.19026)	(0.21432)
R^2	0.80591	0.79743	0.81049	0.81338	0.81228
Adj. R^2	0.79935	0.79136	0.80408	0.80707	0.80665
Num. obs.	276	276	276	276	276

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

Tab. 2: Semiparametric Models for Tractor Prices

5 Generalized Additive Model

5.1 Linear Model

As an example of the output from the GAM specification, I first estimated the model with no nonlinear terms, which is essentially a linear regression.

```
Family: gaussian
Link function: identity

Formula:
log_saleprice ~ horsepower + squared_horsepower + age + enghours +
    diesel + fwd + manual + johndeere + cab

Parametric coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.728e+00  1.060e-01  82.327  < 2e-16 ***
horsepower          1.112e-02  1.067e-03  10.423  < 2e-16 ***
squared_horsepower -1.404e-05  2.255e-06  -6.223 1.89e-09 ***
age                -3.233e-02  3.580e-03  -9.031  < 2e-16 ***
enghours           -4.178e-05  9.569e-06  -4.367 1.81e-05 ***
diesel              2.035e-01  9.805e-02   2.076   0.0389 *
fwd                 2.654e-01  5.820e-02   4.560 7.82e-06 ***
manual             -1.502e-01  6.189e-02  -2.426   0.0159 *
johndeere           3.187e-01  7.186e-02   4.435 1.35e-05 ***
cab                 4.834e-01  7.003e-02   6.903 3.72e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1


R-sq.(adj) =  0.799   Deviance explained = 80.6%
GCV = 0.16445  Scale est. = 0.15849   n = 276
```

5.2 Semiparametric Model

Further investigating the results of the full semiparametric specification in Model 5 of Table 2, I estimated the model with all three continuous variables specified as nonparametric functions. The result was that almost all the variables—both linear and nonlinear—were statistically significant. The only exception was a loss in significance of the diesel indicator.

```
Family: gaussian
Link function: identity

Formula:
log_saleprice ~ s(horsepower) + s(age) + s(enghours) + diesel +
    fwd + manual + johndeere + cab

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.04516    0.09366  96.575  < 2e-16 ***
diesel       0.13440    0.09499   1.415  0.15830
fwd          0.29899    0.05754   5.196 4.11e-07 ***
manual      -0.16938    0.05965  -2.839  0.00487 **
johndeere    0.33067    0.06890   4.799 2.68e-06 ***
cab          0.40439    0.07151   5.655 4.08e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Approximate significance of smooth terms:
                edf Ref.df     F  p-value
s(horsepower) 4.387  5.321 44.89  < 2e-16 ***
s(age)        3.264  4.057 21.59  < 2e-16 ***
s(enghours)   1.000  1.000 23.39 2.64e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

R-sq.(adj) =  0.819   Deviance explained = 82.8%
GCV = 0.15063  Scale est. = 0.14263   n = 276
```

On the other hand, the adjusted R-squared has not increased very much, from 0.799 to 0.819 under this specification, which may not justify the added complexity of the model. Perhaps more importantly, the coefficients on the linear terms are very similar across models, indicating that the models support similar conclusions relating to any business decision involving the John Deere premium. With this second model, we have even more support for those conclusions and are certain that the conclusions are not coincidental results of the functional form decisions for previous models.

Perhaps as a middle ground, we can estimate a model with a nonparametric specification for the horsepower variable alone, since it seems to have a nonlinear relationship with value in either case. This retains most of the predictive value of the maximally semiparametric model and accommodates the nonlinear relationship with value of horsepower.

```
Family: gaussian
Link function: identity

Formula:
log_saleprice ~ s(horsepower) + age + enghours + diesel + fwd +
    manual + johndeere + cab

Parametric coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.697e+00  1.120e-01  86.607  < 2e-16 ***
age         -3.114e-02  3.539e-03  -8.799  < 2e-16 ***
enghours    -4.354e-05  9.342e-06  -4.660 5.02e-06 ***
diesel       1.372e-01  9.590e-02   1.431  0.15361
fwd          3.134e-01  5.773e-02   5.428 1.29e-07 ***
manual      -1.650e-01  6.041e-02  -2.732  0.00673 **
johndeere    3.189e-01  6.933e-02   4.599 6.59e-06 ***
cab          3.770e-01  7.202e-02   5.235 3.38e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Approximate significance of smooth terms:
                edf Ref.df     F p-value
s(horsepower) 4.758  5.751 44.49  <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

R-sq.(adj) =  0.815   Deviance explained = 82.3%
GCV = 0.15323  Scale est. = 0.14615   n = 276
```

6 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.

6.1 Transformation of Horsepower

Performing the transformation on the horsepower variable produces a modified from 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.

6.2 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.

6.3 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.

6.4 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.

7 Final Comparison of Candidate Models

I created one more 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 3 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 is the semiparametric model with a nonparametric form for horsepower. Model 3 has the same specification as the other two, 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.

	Model 1	Model 2	Model 3
(Intercept)	8.72792^{***}	8.97543^{***}	3.09024^{***}
	(0.10602)	(0.10479)	(0.39174)
horsepower	0.01112^{***}		
	(0.00107)		
squared_horsepower	-0.00001^{***}		
	(0.00000)		
age	-0.03233^{***}	-0.03813^{***}	-0.02927^{***}
	(0.00358)	(0.00360)	(0.00345)
enghours	-0.00004^{***}	0.00000	-0.00005^{***}
	(0.00001)	(0.00001)	(0.00001)
diesel	0.20350^{*}	0.31981^{**}	0.12070
	(0.09805)	(0.09872)	(0.09500)
fwd	0.26539^{***}	0.39101^{***}	0.32602^{***}
	(0.05820)	(0.05901)	(0.05617)
manual	-0.15015^{*}	-0.06208	-0.20053^{**}
	(0.06189)	(0.06285)	(0.06031)
johndeere	0.31872^{***}	0.40778^{***}	0.33386^{***}
	(0.07186)	(0.07313)	(0.06967)
cab	0.48345^{***}	1.05513^{***}	0.42139^{***}
	(0.07003)	(0.05806)	(0.06768)
horsepower_np		0.96671^{***}	
		(0.06886)	
horsepower_bt			3.99759^{***}
			(0.25577)
R^2	0.80591	0.79743	0.81613
Adj. R^2	0.79935	0.79136	0.81062
Num. obs.	276	276	276

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

Tab. 3: Alternate Models for Tractor Prices