Spring 2023 **Firstname M. Lastname**

University of Central Florida
College of Business

QMB 6911
Capstone Project in Business Analytics

Solutions: Problem Set #8

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 will investigate this nonlinear relationship by incorporating a nonparametric specification for the value of horsepower. Similarly, for the other continuous variables engine hours and age, to investigate whether these forms of depreciation are best described with nonlinear forms.

2 Linear Regression Model

A natural staring point is the recommended linear model from Problem Set #7.

2.1 Quadratic Specification for Horsepower

Last week 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 -1.404×10^{-5}, which is nearly ten times as large as the standard error of 2.255×10^{-6}, 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 0.7993, 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.

	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

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.

The models shown in Table 2 illustrate this possibility. Model 1 is the original model in Table 1. Model 2 is a regression omitting the horsepower variables. Model 3 is a regression to predict horsepower with the other explanatory variables in Model 2. Model 4 is a regression to predict squared horsepower with the other explanatory variables in Model 2. Finally, Model 5 shows the coefficients for horsepower from a regression of the residuals of Model 2 on the residuals from Model 3. Notice that these coefficients match those in Model 1. You might notice a slight difference in the standard errors, however, because these are calculated assuming coefficients for two variables, horsepower and squared horsepower, rather than the full suite of ten parameters. This equivalence of the coefficients can be used to fit nonlinear models between a pair of variables by partialing out the effect of the other variables, using a mathematical result called the Frisch-Waugh-Lovell (FWL) theorem, maned after early statisticians and econometricians who used these methods.

	Original (1)	Reduced (2)	H.P. (3)	HP. Sq. (4)	FWL H.P. (5)
(Intercept)	8.72792^{***}	9.01091^{***}	27.05189	1269.75517	
	(0.10602)	(0.13785)	(17.13098)	(8102.93510)	
horsepower	0.01112^{***}				
	(0.00107)				
squared_horsepower	-0.00001^{***}				
	(0.00000)				
age	-0.03233^{***}	-0.03609^{***}	-1.27409^{*}	-741.49174^{**}	
	(0.00358)	(0.00473)	(0.58819)	(278.21478)	
enghours	-0.00004^{***}	-0.00000	0.00735^{***}	2.90344^{***}	
	(0.00001)	(0.00001)	(0.00153)	(0.72137)	
diesel	0.20350^{*}	0.32931^{*}	6.21382	-4041.08422	
	(0.09805)	(0.12991)	(16.14428)	(7636.22715)	
fwd	0.26539^{***}	0.34908^{***}	18.48035	8678.65733	
	(0.05820)	(0.07756)	(9.63839)	(4558.94619)	
manual	-0.15015^{*}	-0.08564	15.32259	7543.06505	
	(0.06189)	(0.08268)	(10.27470)	(4859.92047)	
johndeere	0.31872^{***}	0.37245^{***}	13.28577	6697.80712	
	(0.07186)	(0.09618)	(11.95221)	(5653.38319)	
cab	0.48345^{***}	1.03138^{***}	72.82965^{***}	18662.03208^{***}	
	(0.07003)	(0.07637)	(9.49032)	(4488.91194)	
horsepower_resid					0.01112^{***}
					(0.00105)
horsepower_2_resid					-0.00001^{***}
					(0.00000)
R^2	0.80591	0.64790	0.40004	0.22696	0.44877
Adj. R^2	0.79935	0.63871	0.38437	0.20677	0.44474
Num. obs.	276	276	276	276	276

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

Tab. 2: Quadratic Model for Tractor Prices: FWL Regressions

To illustrate the fit of the model, Figure 1 shows a scatter plot of the residual log prices on horsepower. The observations are shown in blue and the fitted values are shown in red. The variation in the fitted values results from the fact that it is not plotted against the transformed excess horsepower variable used in the regressions. Still, the quadratic pattern is apparent and appears to match the data.

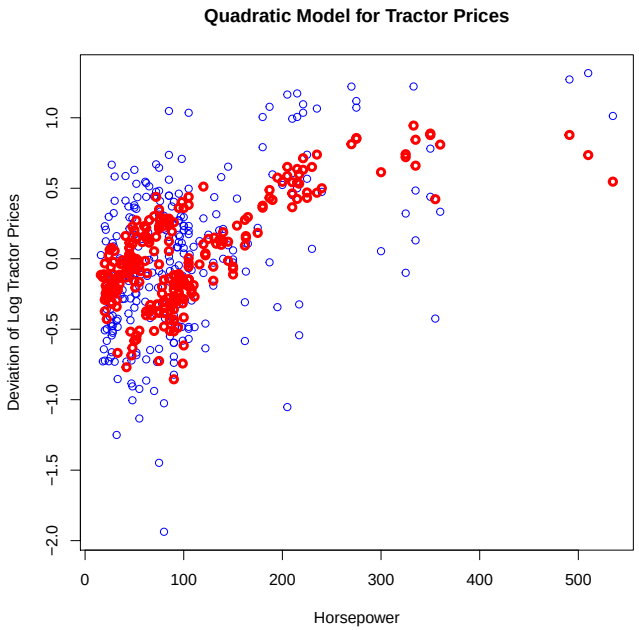

Fig. 1: Linear-Quadratic Model for Tractor Prices

As a comparison, Figure 2 augments the above by showing the plot against the residuals from the regression for horsepower: the "excess horsepower" compared to what would be expected given the other characteristics of a tractor. The quadratic function is more clear from this perspective. This time, the variation in the fitted values results from the two-dimensional nature of the horsepower variable when we consider the quadratic form.

Fig. 2: Linear-Quadratic Model for Tractor Prices: Excess Horsepower

Now consider a nonparametric specification for the relationship between prices and horsepower. Figure 3 overlays the nonparametric estimate (shown in green) with the above in Figure 2. 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. 3: Nonparametric Model for Tractor Prices: Excess Horsepower

Finally, consider a set of nonparametric specifications for the relationship between prices and horsepower. Figure 4 overlays other nonparametric estimates with the above in Figure 3. The points in orange and in magenta represent alternate models with different degrees of smoothing. When we estimated probability densities, we adjusted the bandwidth parameter to fit with different degrees of smoothness. The `loess` method used for the nonparametric method has a span parameter for this function. The default smoother `span` (bandwidth parameter) is 0.75.

In the magenta points, with `span` parameter 0.1, the pattern has more variation in slope but closely follows the prediction from the quadratic model. The smoother curve in orange even more closely represents a quadratic line. Again, it appears that the quadratic form is close enough. Perhaps the result will be different for other continuous variables in the model.

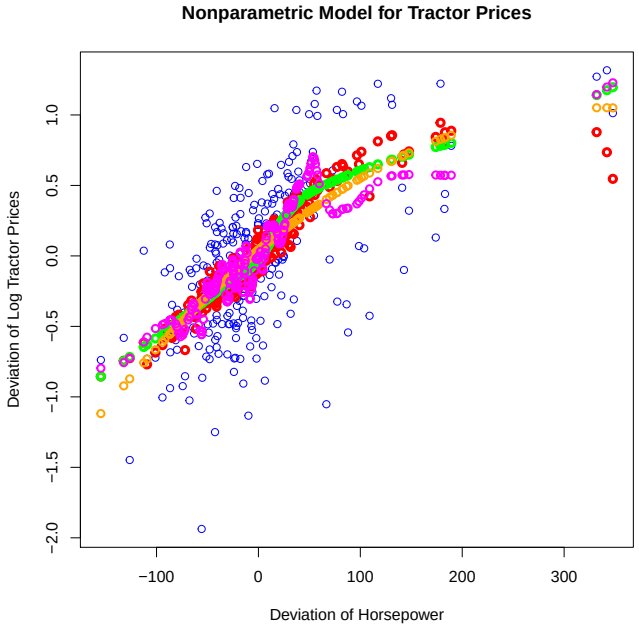

Fig. 4: Nonparametric Model for Tractor Prices: Excess Horsepower

3.2 Nonparametric Specification for Age

As above, first conduct FWL regressions to reduce the problem to two dimensions. The models shown in Table 3 illustrate this possibility. Model 1 is the same original model in Table 1. Model 2 is a regression omitting the age variable. Model 3 is a regression to predict age with the other explanatory variables in Model 2. Finally, Model 4 shows the coefficient for age from a regression of the residuals of Model 2 on the residuals from Model 3. Notice that these coefficients match those in Model 1.

	Original (1)	Reduced (2)	Age (3)	FWL Age (4)
(Intercept)	8.72792^{***}	8.25387^{***}	14.66329^{***}	
	(0.10602)	(0.10509)	(1.57457)	
horsepower	0.01112^{***}	0.01055^{***}	0.01761	
	(0.00107)	(0.00122)	(0.01821)	
squared_horsepower	-0.00001^{***}	-0.00001^{***}	-0.00007	
	(0.00000)	(0.00000)	(0.00004)	
age	-0.03233^{***}			
	(0.00358)			
enghours	-0.00004^{***}	-0.00008^{***}	0.00131^{***}	
	(0.00001)	(0.00001)	(0.00014)	
diesel	0.20350^{*}	0.32816^{**}	-3.85602^{*}	
	(0.09805)	(0.11075)	(1.65943)	
fwd	0.26539^{***}	0.48826^{***}	-6.89389^{***}	
	(0.05820)	(0.06014)	(0.90112)	
manual	-0.15015^{*}	-0.30537^{***}	4.80111^{***}	
	(0.06189)	(0.06784)	(1.01643)	
johndeere	0.31872^{***}	0.30073^{***}	0.55646	
	(0.07186)	(0.08196)	(1.22801)	
cab	0.48345^{***}	0.48403^{***}	-0.01803	
	(0.07003)	(0.07990)	(1.19717)	
age_resid				-0.03233^{***}
				(0.00352)
R^2	0.80591	0.74641	0.51998	0.23464
Adj. R^2	0.79935	0.73881	0.50560	0.23186
Num. obs.	276	276	276	276

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

Tab. 3: Linear Model for Age: FWL Regressions

To illustrate the fit of the model, Figure 5 shows a scatter plot of the residual log prices on age. The observations are shown in blue and the fitted values are shown in red. The variation in the fitted values results from the fact that it is not plotted against the transformed excess age variable used in the regressions. Still, the linear pattern is apparent and appears to match the data.

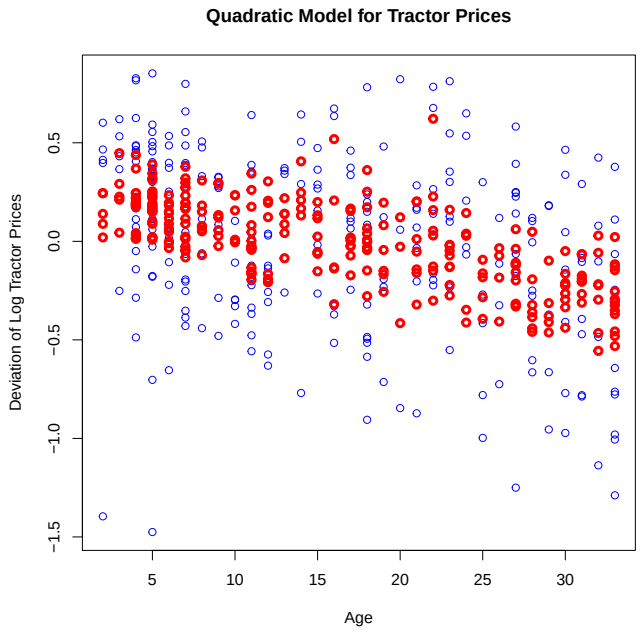

Fig. 5: Linear-Quadratic Model for Tractor Prices

As a comparison, Figure 6 augments the above by showing the plot against 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. Notice that this time the fit follows a straight line, since we have a single variable with no quadratic transformation.

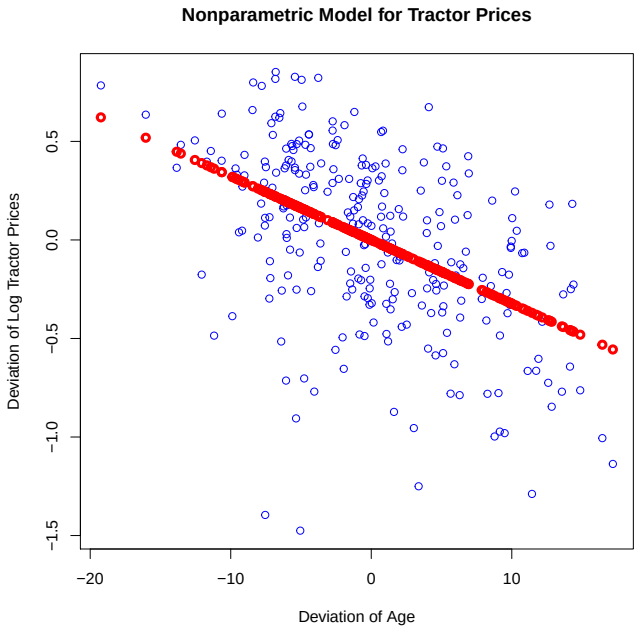

Fig. 6: Linear-Quadratic Model for Tractor Prices: Excess Age

Now consider a nonparametric specification for the relationship between prices and age. Figure 7 overlays the nonparametric estimate (shown in green) with the above in Figure 6. 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 explore the remaining continuous variable.

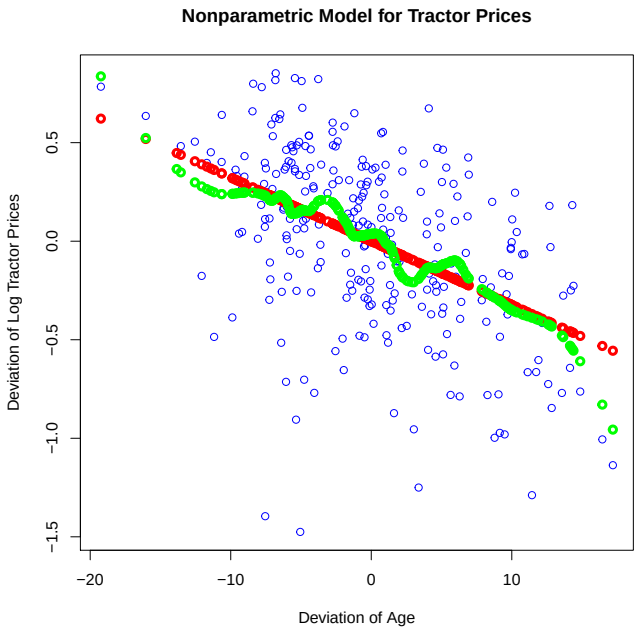

Fig. 7: Nonparametric Model for Tractor Prices: Excess Age

3.3 Nonparametric Specification for Engine Hours

As above, first conduct FWL regressions to reduce the problem to two dimensions. The models shown in Table 4 illustrate this possibility. Model 1 is the same original model in Table 1. Model 2 is a regression omitting the age variable. Model 3 is a regression to predict engine hours with the other explanatory variables in Model 2. Finally, Model 4 shows the coefficient for engine hours from a regression of the residuals of Model 2 on the residuals from Model 3. Notice that these coefficients match those in Model 1.

	Original (1)	Reduced (2)	Eng. Hrs. (3)	FWL Eng. Hrs. (4)
(Intercept)	8.72792^{***}	8.79201^{***}	-1533.86328^{*}	
	(0.10602)	(0.10849)	(671.52168)	
horsepower	0.01112^{***}	0.01030^{***}	19.70996^{**}	
	(0.00107)	(0.00108)	(6.71576)	
squared_horsepower	-0.00001^{***}	-0.00001^{***}	-0.02012	
	(0.00000)	(0.00000)	(0.01437)	
age	-0.03233^{***}	-0.04000^{***}	183.63531^{***}	
	(0.00358)	(0.00322)	(19.94860)	
enghours	-0.00004^{***}			
	(0.00001)			
diesel	0.20350^{*}	0.20799^{*}	-107.43464	
	(0.09805)	(0.10130)	(627.05359)	
fwd	0.26539^{***}	0.25967^{***}	136.84215	
	(0.05820)	(0.06013)	(372.16237)	
manual	-0.15015^{*}	-0.15684^{*}	160.15539	
	(0.06189)	(0.06393)	(395.72599)	
johndeere	0.31872^{***}	0.30410^{***}	349.77262	
	(0.07186)	(0.07417)	(459.11233)	
cab	0.48345^{***}	0.45588^{***}	659.90018	
	(0.07003)	(0.07207)	(446.07457)	
eng_resid				-0.00004^{***}
				(0.00001)
R^2	0.80591	0.79200	0.45819	0.06689
Adj. R^2	0.79935	0.78577	0.44195	0.06350
Num. obs.	276	276	276	276

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

Tab. 4: Linear Model for Engine Hours: FWL Regressions

To illustrate the fit of the model, Figure 8 shows a scatter plot of the residual log prices on engine hours. The observations are shown in blue and the fitted values are shown in red. The variation in the fitted values results from the fact that it is not plotted against the transformed excess engine hours variable used in the regressions. Still, the linear pattern is apparent and appears to match the data.

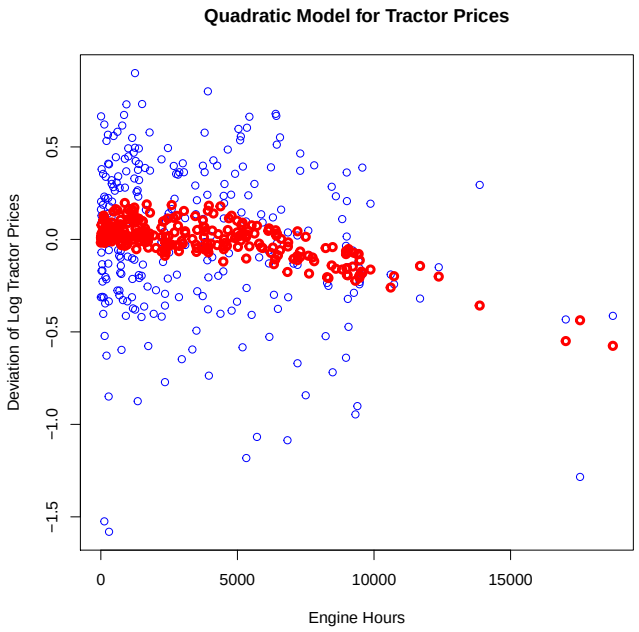

Fig. 8: Linear-Quadratic Model for Tractor Prices

As a comparison, Figure 9 augments the above by showing the plot against the 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. As with age, the fit follows a straight line, since we have a single variable with no quadratic transformation. I move directly to the nonparametric specification for the relationship between prices and engine hours. Figure 9 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.

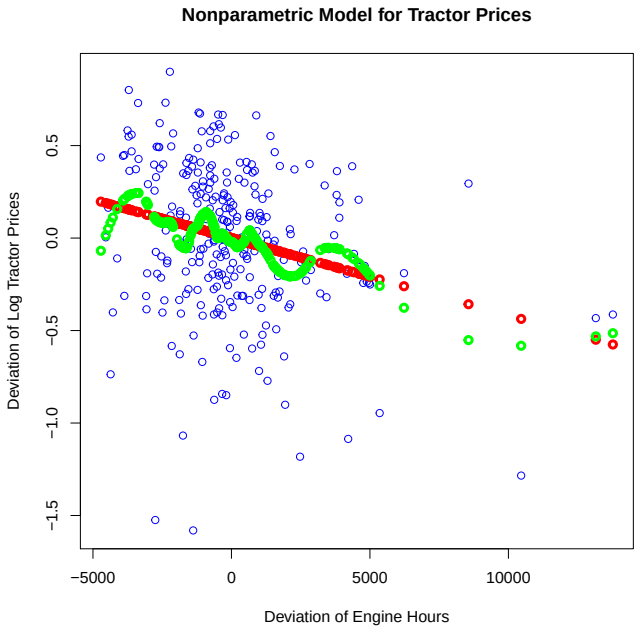

Fig. 9: Nonparametric Model for Tractor Prices: Excess Engine Hours

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 5 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. 5: 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 5, 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
```