Spring 2021 **Ali Berra, Alex Georgiopoulus, Caroline Gear**

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 `Trucks_Reg_Model.R` to produce a more accurate model for used trucks cost with the data from `UsedTrucks.dat` in the `Data` folder. The dataset includes the following variables. From

$type_i$	=	sale type
$pauc$	=	price when sold at auction
$pret$	=	price when sold retail
$mileage$	=	odometer
$make$	=	make of vehicule
$year$	=	model year of vehicle
$damage$	=	an index of damage to vehicle, 1 little damage, 10 a lot
$dealer$	=	dealer id
ror	=	rate-of-return
$cost$	=	net amount given to trade in

problem set 7, we build a model that can predict the cost of the vehicule depending on the damage of the trucks. we will add quadratic specification for damage which will increase the relationship between cost and damage.

	Model 1
(Intercept)	8.76488***
	(0.08135)
squared_damage	0.00459*
	(0.00214)
make	−0.00619
	(0.00537)
damage	−0.04677
	(0.02543)
dealer	0.00196
	(0.00417)
mileage	−0.00001***
	(0.00000)
age	−0.07163***
	(0.01114)
type	0.04438*
	(0.02182)
R^2	0.19087
Adj. R^2	0.19029
Num. obs.	9861

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

Tab. 1: Quadratic Model for Trucks cost

2 Linear Regression Model

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

2.1 Quadratic Specification for Damage

we can say that the more a Truck is damaged the less likely it will sell. to impliment this in the model we squared the damage variable to have more impact inthe realtionship.

The results of this regression specification are shown in Table 1.

according to our model milage and age are more significant predictor than damge with -5.053e-06 and -7.163e-02 coeffiecient respectively. the P value for damage 0.06 and for squared damage 0.03 which is slightly significant when squared

3 Nonlinear Specifications

3.1 Nonparametric Specification for Horsepower

The specification in Table 1 assumes a quadratic functional form for the relationship between cost and damage. To consider the damage variable alone, while accounting for the effects of other variables, one can fit a nonparametric model to the residuals from a model of Trucks cost, after regressing Trucks cost on the other variables. This leaves only the variation in trucks cost that is not explained by the other variables. Going one step further, perform the same transformation to the damage variable: take the residuals from a model of damage, after regressing damage 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 **??** illustrate this possibility. Model 1 is the original model in Table 1. Model 2 is a regression omitting the damage variables. Model 3 is a regression to predict damage with the other explanatory variables in Model 2. Finally, Model 4 shows the coefficients for damage 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, damage and squared damage, 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.

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	8.76488***	8.65101***	2.83163***	4.04282***	
	(0.08135)	(0.05049)	(0.06877)	(0.81893)	
squared_damage	0.00459*				
	(0.00214)				
make	−0.00619	−0.00621	−0.00411	−0.04520	
	(0.00537)	(0.00537)	(0.00731)	(0.08708)	
damage	−0.04677				
	(0.02543)				
dealer	0.00196	0.00207	0.00528	0.07682	
	(0.00417)	(0.00417)	(0.00568)	(0.06761)	
mileage	−0.00001***	−0.00001***	0.00000	0.00001	
	(0.00000)	(0.00000)	(0.00000)	(0.00002)	
age	−0.07163***	−0.06985***	0.32609***	3.70576***	
	(0.01114)	(0.01088)	(0.01482)	(0.17644)	
type	0.04438*	0.04590*	0.34444***	3.83624***	
	(0.02182)	(0.02168)	(0.02952)	(0.35157)	
damage_resid					−0.04677
					(0.02542)
damage_2_resid					0.00459*
					(0.00213)
R^2	0.19087	0.19044	0.50020	0.47707	0.00053
Adj. R^2	0.19029	0.19003	0.49994	0.47680	0.00032
Num. obs.	9861	9861	9861	9861	9861

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

Tab. 2: Quadratic Model for Trucks cost: FWL Regressions

To illustrate the fit of the model, Figure 1 shows a scatter plot of the residual log cost on damage. 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 damage variable used in the regressions. Still, the quadratic pattern is apparent and appears to match the data.

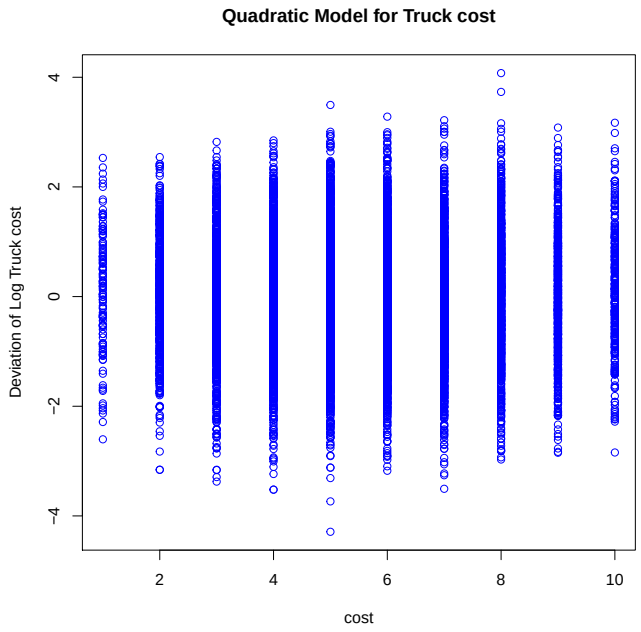

Fig. 1: Linear-Quadratic Model for Truck cost

As a comparison, Figure **??** augments the above by showing the plot against the residuals from the regression for damage: the "excess damage" compared to what would be expected given the other characteristics of a truck.

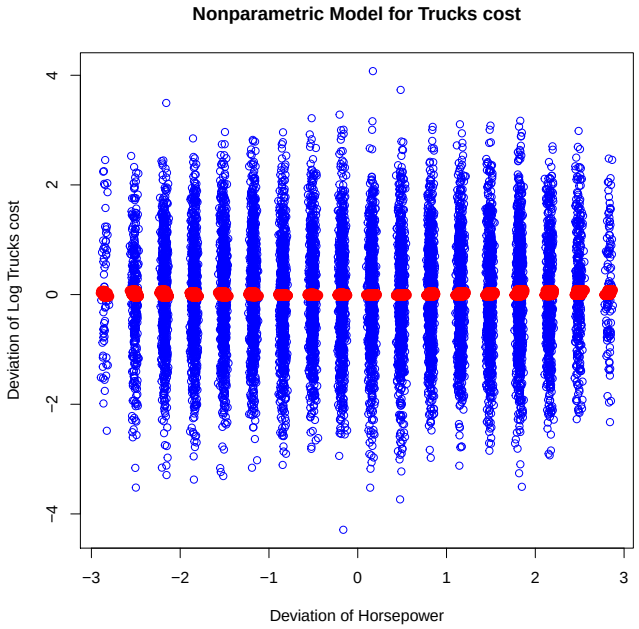

Fig. 2: Linear-Quadratic Model for Tractor Prices: damage

Now consider a nonparametric specification for the relationship between cost and damage. Figure 3 overlays the nonparametric estimate (shown in green) with the above in Figure **??**. 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.

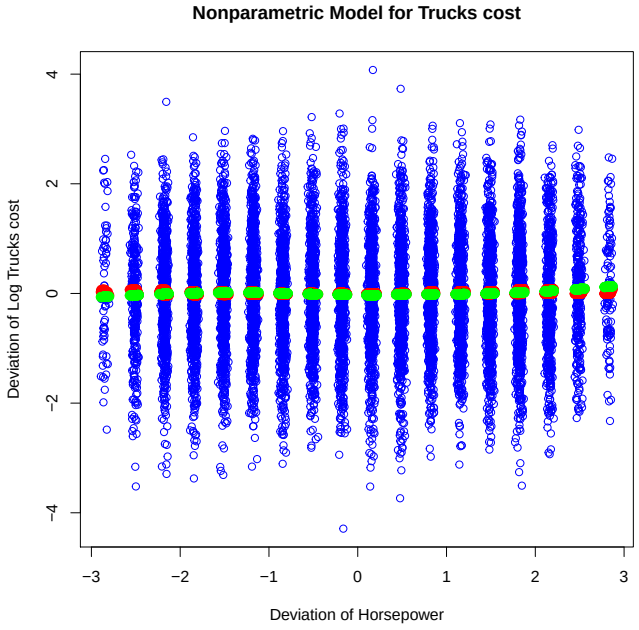

Fig. 3: Nonparametric Model for Tractor Prices: Damage

Finally, consider a set of nonparametric specifications for the relationship between cost and damage. 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.

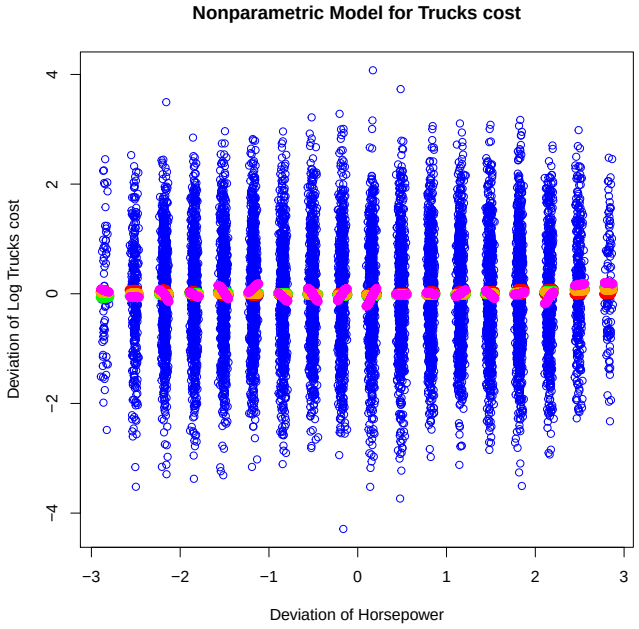

Fig. 4: Nonparametric Model for Tractor Prices: Damage

	Model 1	Model 2	Model 3	Model 4
(Intercept)	8.76488^{***}	8.96193^{***}	-2.75101^{***}	
	(0.08135)	(0.07551)	(0.06814)	
squared_damage	0.00459^{*}	0.00462^{*}	-0.00035	
	(0.00214)	(0.00214)	(0.00193)	
make	-0.00619	-0.00593	-0.00371	
	(0.00537)	(0.00538)	(0.00485)	
damage	-0.04677	-0.05735^{*}	0.14767^{***}	
	(0.02543)	(0.02543)	(0.02295)	
dealer	0.00196	0.00234	-0.00526	
	(0.00417)	(0.00418)	(0.00377)	
mileage	-0.00001^{***}	-0.00001^{***}	0.00009^{***}	
	(0.00000)	(0.00000)	(0.00000)	
age	-0.07163^{***}			
	(0.01114)			
type	0.04438^{*}	0.04478^{*}	-0.00556	
	(0.02182)	(0.02187)	(0.01973)	
age_resid				-0.07163^{***}
				(0.01114)
R^2	0.19087	0.18747	0.94978	0.00418
Adj. R^2	0.19029	0.18698	0.94975	0.00408
Num. obs.	9861	9861	9861	9861

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

Tab. 3: Linear Model for Age: FWL Regressions

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 **??**. 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.

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.

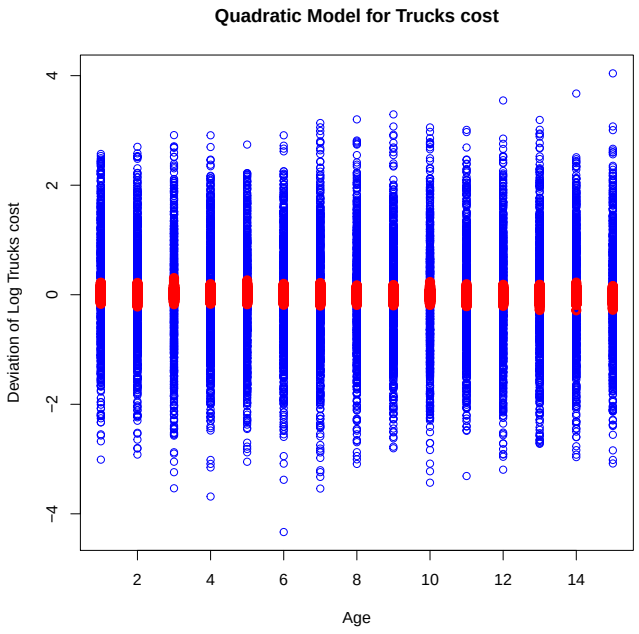

Fig. 5: Linear-Quadratic Model for Tractor Prices

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 truck compared to what would be expected given the other characteristics of the truck. Notice that this time the fit follows a straight line, since we have a single variable with no quadratic transformation.

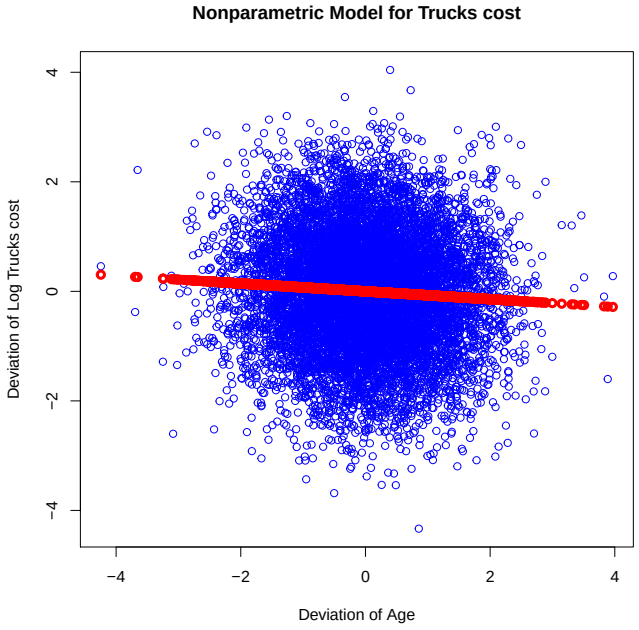

Fig. 6: Linear-Quadratic Model for trucks costs: Excess Age

Now consider a nonparametric specification for the relationship between cost 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.

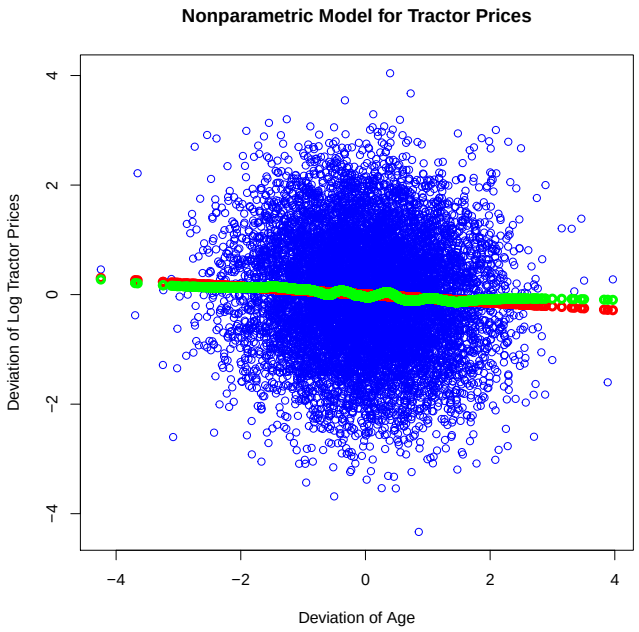

Fig. 7: Nonparametric Model for truck cost: Excess Age

	Model 1	Model 2	Model 3	Model 4
(Intercept)	8.76488***	8.77479***	5.04979***	
	(0.08135)	(0.07858)	(0.18991)	
squared_damage	0.00459*	0.00460*	0.00438	
	(0.00214)	(0.00214)	(0.00516)	
make	−0.00619	−0.00621	−0.01070	
	(0.00537)	(0.00537)	(0.01297)	
damage	−0.04677	−0.04684	−0.03332	
	(0.02543)	(0.02543)	(0.06145)	
dealer	0.00196			
	(0.00417)			
mileage	−0.00001***	−0.00001***	0.00000	
	(0.00000)	(0.00000)	(0.00000)	
age	−0.07163***	−0.07170***	−0.03759	
	(0.01114)	(0.01114)	(0.02692)	
type	0.04438*	0.04446*	0.04225	
	(0.02182)	(0.02182)	(0.05274)	
type_resid				0.00196
				(0.00417)
R^2	0.19087	0.19085	0.00047	0.00002
Adj. R^2	0.19029	0.19036	−0.00014	−0.00008
Num. obs.	9861	9861	9861	9861

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

Tab. 4: Linear Model for type : FWL Regressions

3.3 Nonparametric Specification for dealer

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 **??**. Model 2 is a regression omitting the age variable. Model 3 is a regression to predict dealer 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.

To illustrate the fit of the model, Figure 8 shows a scatter plot of the residual log cost on dealer. The observations are shown in blue and the fitted values are shown in red.

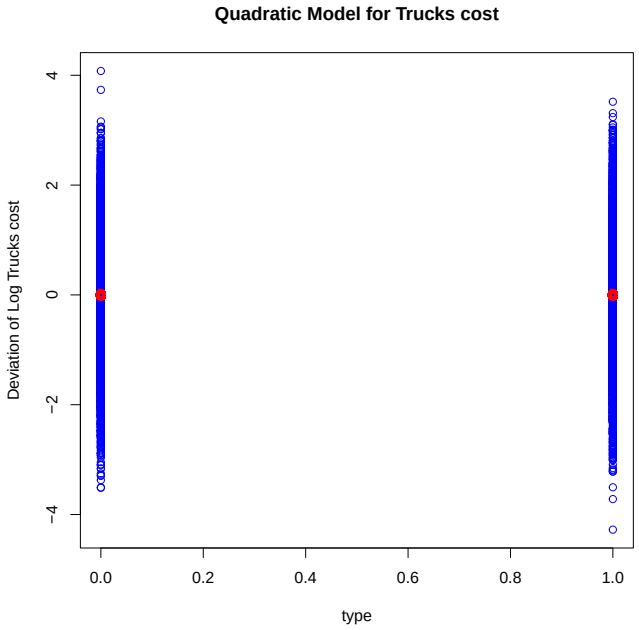

Fig. 8: Linear-Quadratic Model for truck cost

As a comparison, Figure 9 augments the above by showing the plot against the residuals from the regression for dealer. I move directly to the nonparametric specification for the relationship between cost and dealer. 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: dealer

4 Semiparametric Estimates

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 damage 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)	8.76488^{***}	8.65711^{***}	8.95824^{***}	8.76488^{***}	8.90648^{***}
	(0.08135)	(0.05515)	(0.07530)	(0.08135)	(0.08167)
squared_damage	0.00459^{*}		0.00454^{*}	0.00459^{*}	
	(0.00214)		(0.00213)	(0.00214)	
make	-0.00619	-0.00619	-0.00580	-0.00619	-0.00595
	(0.00537)	(0.00537)	(0.00536)	(0.00537)	(0.00536)
damage	-0.04677	-0.00394	-0.05654^{*}	-0.04677	-0.01698
	(0.02543)	(0.00808)	(0.02536)	(0.02543)	(0.00866)
dealer	0.00196	0.00232	0.00245	0.00196	0.00096
	(0.00417)	(0.00417)	(0.00416)	(0.00417)	(0.00419)
mileage	-0.00001^{***}	-0.00001^{***}	-0.00001^{***}	-0.00001^{***}	-0.00001^{***}
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
age	-0.07163^{***}	-0.06884^{***}		-0.07163^{***}	0.01873
	(0.01114)	(0.01118)		(0.01114)	(0.02479)
type	0.04438^{*}	0.05465^{*}	0.04512^{*}	0.04438^{*}	0.05681^{*}
	(0.02182)	(0.02212)	(0.02181)	(0.02182)	(0.02209)
damage_np		1.28304^{**}			1.26312^{**}
		(0.43922)			(0.43862)
age_np			1.01862^{***}		1.19224^{***}
			(0.13621)		(0.30329)
dealer_np					1.11233^{***}
					(0.26977)
R^2	0.19087	0.19119	0.19206	0.19087	0.19378
Adj. R^2	0.19029	0.19061	0.19149	0.19029	0.19304
Num. obs.	9861	9861	9861	9861	9861

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

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_cost ~ damage + squared_damage + age + make + damage + dealer +
    mileage + age + type

Parametric coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     8.765e+00  8.135e-02 107.740  < 2e-16 ***
damage         -4.677e-02  2.543e-02  -1.839   0.0659 .
squared_damage  4.595e-03  2.135e-03   2.152   0.0314 *
age            -7.163e-02  1.114e-02  -6.430 1.34e-10 ***
make           -6.194e-03  5.368e-03  -1.154   0.2486
dealer          1.964e-03  4.168e-03   0.471   0.6376
mileage        -5.053e-06  1.062e-06  -4.756 2.01e-06 ***
type            4.438e-02  2.182e-02   2.034   0.0420 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1


R-sq.(adj) =   0.19   Deviance explained = 19.1%
GCV = 1.1444  Scale est. = 1.1435    n = 9861
```

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_cost ~ s(damage) + s(age)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.53777    0.01077   700.2   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Approximate significance of smooth terms:
            edf Ref.df       F p-value
s(damage) 1.000  1.000   0.908   0.341
s(age)    4.149  5.111 243.022  <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

R-sq.(adj) =  0.191   Deviance explained = 19.1%
GCV = 1.1436  Scale est. = 1.1429    n = 9861
```

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_cost ~ s(damage) + age + make + damage + dealer + mileage +
    age + type

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.209e+00  3.007e-02  40.192  < 2e-16 ***
age         -7.148e-02  1.114e-02  -6.417 1.46e-10 ***
make        -6.227e-03  5.368e-03  -1.160   0.2461
damage       1.310e+00  9.977e-03 131.253  < 2e-16 ***
dealer       1.870e-03  4.169e-03   0.448   0.6538
mileage     -5.072e-06  1.062e-06  -4.774 1.84e-06 ***
type         4.453e-02  2.182e-02   2.041   0.0413 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Approximate significance of smooth terms:
           edf Ref.df    F p-value
s(damage) 4.75  5.914 3432  <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Rank: 15/16
R-sq.(adj) =  0.191   Deviance explained = 19.1%
GCV = 1.1444  Scale est. = 1.1431    n = 9861
```