AAEA Agricultural & Applied Economics Association

Case Study

Hedonic Price Analysis of Used Tractors

Ryan Feuz
Utah State University

JEL Codes: A22, A23, D12, Q13
Keywords: Hedonic regression, misspecification, ordinary least squares, tractor prices

Abstract
This case follows Nate Shepard, a fictionalized data analyst for John Deere, as he is tasked with finding a suitable method for predicting used tractor prices. Nate uses hedonic price theory to specify and estimate a regression equation that can be used to evaluate marginal values of specific tractor attributes and predict out-of-sample tractor prices. Beyond price prediction, Nate must also consider the inflationary environment the used tractor market has been experiencing of late in his regression specification as well as compare the John Deere brand to rival manufacturers. The case allows readers to go along with Nate in the journey as he completes the process of data collection and cleaning, initial model specification based on relevant literature and theory, model estimation, evaluation of the model for misspecification issues, model revision and re-estimation, and model interpretation and use. The case provides an excellent example of empirical regression analysis in an agribusiness setting and gives readers an opportunity to familiarize themselves with hedonic price theory using a data set of actual used tractor auction results from 2020–2022.

1 Introduction

As a recent graduate in agricultural economics from the University of Illinois, Nate Shepard was excited when he landed his first job out of school as a Market Analyst for John Deere. The job was everything he hoped for. It was based in his hometown and location of John Deere company headquarters, Moline, Illinois. As the job title suggests, it was an analytical position with an opportunity for Nate to work on empirical analysis, which played to Nate's strong suite at working quantitatively. Most important, however, the position allowed for Nate to remain connected with the agricultural industry. Nate grew up on a family farm just outside of Moline producing mainly soybeans and corn. Nate enjoyed life on the farm, yet for a variety of reasons continuing to work on the farm post-college was not a viable option for him.

Nate was hired at an interesting time in the agricultural industry as well as within the company. The world was emerging from the global COVID-19 pandemic, which brought a host of challenges and opportunities to the economy at large and the agricultural industry. Notable challenges included the Ukrainian invasion by Russia, farm labor shortages, ongoing and persistent drought, as well as inflation and rising input costs. John Deere was in the middle of a great year. The company had reported net income of just under $6 billion for 2021, and forecasts indicated that 2022 was on pace to increase net income to $6.5–7 billion for 2022 (John Deere 2021). John Deere CEO, John May expected "demand for farm and construction equipment to continue benefiting from positive fundamentals, including favorable crop prices, economic growth, and increased investment in infrastructure" (John Deere 2021). At the onset of Nate's employment with John Deere, his manager, Todd Smith, assigned him a task with an objective that Nate recognized would require the use of much of his quantitative analysis skill set he had acquired through his schooling. Todd walked Nate through the problem and the objective.

> "*Throughout the last few years, the company has seen strong sales and an increase in demand in our agricultural tractor division,"* Todd said*. "However, one issue we face is continued*

> *supply chain disruptions and their associated impact on sales. Pandemic-related disruptions as well as the labor strike the company faced in the fall of 2021 lead to significant reductions in output for the company in the fourth quarter of 2021 and first quarter of 2022* (Tita 2021)*. This reduction as well as simultaneous reductions in supply by our competitors mainly due to COVID-19–related disruptions formed a chain reaction ultimately pushing used tractor prices upward* (Deaux 2021)*.*"

Nate was keenly aware of this increase in used farm equipment prices, having just been through the purchasing process of a used tractor at an auction for the family farm with his dad. Nate relayed this purchasing experience to Todd and described the financial difficulty it created for their farm as they paid approximately 90 percent the original retail price for the used tractor, despite it being three years old with nearly 800 hours of use. Todd sympathized with Nate and then proceeded to lay out the objective of the project he wanted Nate to work on.

> "*Your family's experience unfortunately was not an isolated incident as I have heard several other firsthand accounts similar to yours from friends and family of late. Deere is committed to continual improvement of our supply chain to help ensure these disruptions can be avoided in the future. However, for the time being, we are very interested in an analysis of the used tractor market to help us better understand three things. First, how can we more appropriately advise our dealerships on pricing used tractors that we take in on trade. Second, regarding brand, how has the John Deere brand specifically been fairing relative to our competitors in the used market. Third, how much have used tractor prices been hit with inflation. I would like for you to perform a detailed analysis of the used tractor market over the last three years to help answer these questions. I will present your analysis to our upper management team and distribute the information across our dealer network*."

Todd provided Nate with no specific method of analysis, leaving that to Nate's discretion as to best accomplish the objectives that he had laid out for him. After receiving this directive, Nate reflected on the opportunity that was before him. Nate recognized that this was an ideal project for him to work on because he could leverage many of his newly acquired analytical skills from his time at the university and have a chance to have his work distributed broadly within the company, which could lead to greater opportunities for growth and recognition in his new position. It was time to get to work!

2 Data: Clean Up, Visualization, and Initial Analysis

The beginning of any successful analytical project begins with quality data. Nate wanted to focus his analysis on the last three years because he felt this would capture the years of an increased rate of inflation. After consulting with Todd, Nate also decided to constrain his analysis at least initially to used tractors within the 40–99 horsepower (hp) range that were equipped with a loader. This would be an important constraint as Todd was certain that when comparing price variance between smaller tractors versus large commercial size tractors there would be differences in the values that producers placed on certain attributes of the tractors themselves. For this reason, Nate felt if he didn't constrain his analysis to a specific horsepower class, his predictive accuracy would decline. The aims of the project required that the data include actual sale prices for used tractors in this hp class across the entire United States. Nate decided to rely on publicly available auction result data from TractorHouse.com. Before pulling the data from the site, Nate constrained the results to the appropriate hp range, U.S. sales only in the last three years, model years 2010 or newer, loader included, and only tractors from the top manufacturers, including John Deere, Case IH, Kubota, Mahindra, Massey Ferguson, and New Holland. Altogether, Nate's data set comprised 1,103 observations of tractors sold at auction.

With the data in hand, the next step was to look at summary statistics and scatterplots of the data to identify possible outliers and data entry errors. Nate first created a scatterplot (Figure 1) with sale prices on the *y* axis and hours (usage hours) on the *x* axis.

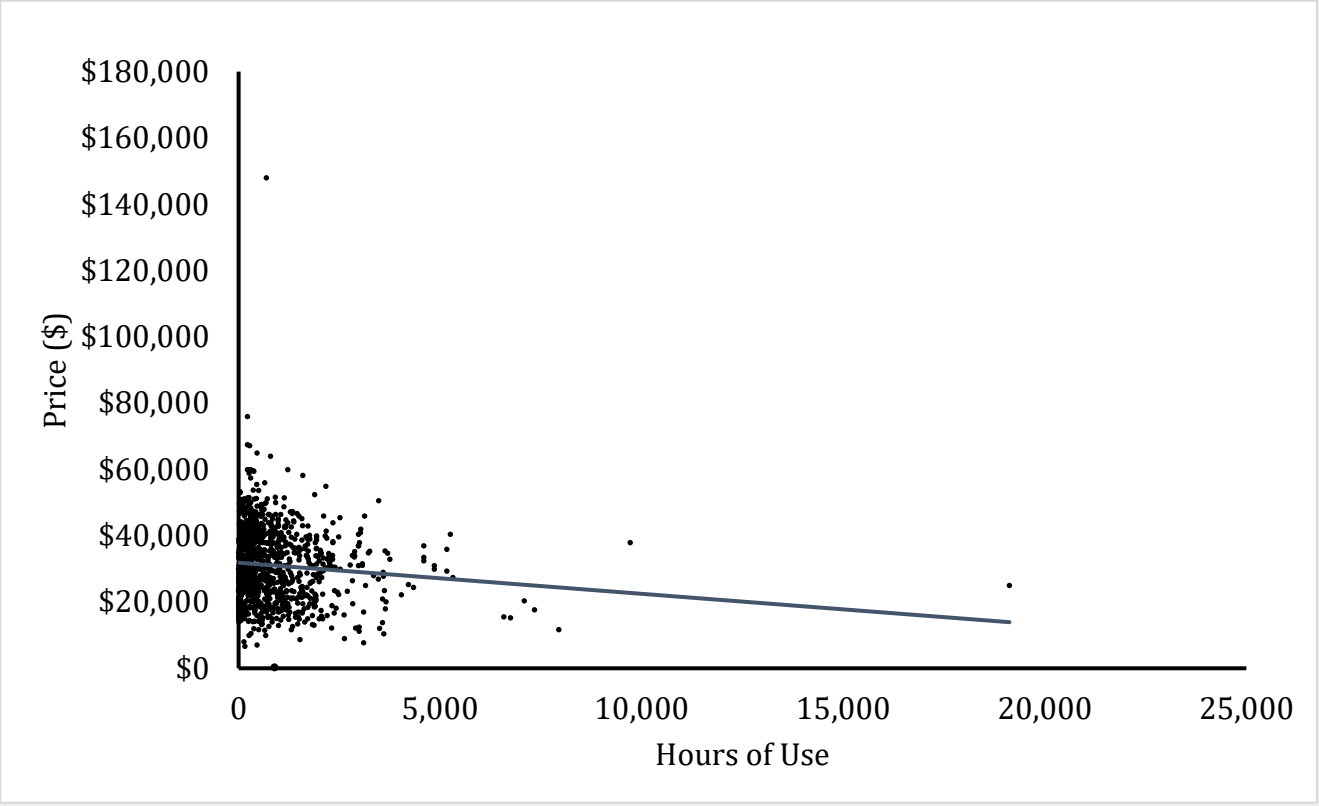

Figure 1. Tractor Sale Price versus Hours of Use

Hours on a tractor is an indicator of total usage, and Nate was confident that hours of use should be negatively correlated with sales price. As Nate looked over the scatterplot, he noted that there was indeed a negative correlation between these variables as indicated by the included trendline. From the graph, Nate was alarmed by three outliers. First, the highest auction price was recorded at nearly $150,000 whereas all other prices were below $80,000. Second, the lowest price was recorded at $227 whereas the next lowest was at $4,730. Third, the highest hours were listed at nearly 20,000 with the next highest at approximately 10,000. As Nate looked closely at the data, the hours outlier was easily identified as a data entry error. The hours were recorded as 19,122 in his data set as well as one location on the specific tractor advertisement. Yet in the written description in the online advertisement, the hours were stated as approximately 9,800. With no way to verify the correct hours, Nate decided it was best to eliminate this tractor from the data set for his analysis. Though not as clear-cut, Nate ultimately decided the high-priced tractor of nearly $150,000 and low-priced tractor of $227 were also presumed data entry errors. Nate determined this by comparing the sales prices of the other tractors of the same makes and models. The average sale price for the high-priced make and model (John Deere 5075E) across 74 total tractor sales was $34,513, with the maximum price at $51,000 and the low price at $11,750. There was only one other tractor of the same make and model as the low-priced tractor (New Holland T5050), and it sold in the same month and year for $33,500. Nate found nothing in the

advertisement of the low-priced tractor that would suggest it had any reason to be sold at such an extremely low price.

For these reasons, Nate concluded that it was unreasonable for tractors of these makes and models to sell at these extreme outlier prices and felt it was best to remove them from the data set, leaving him with 1,100 observations for his analysis. With the outliers removed, Nate knew the average predictive accuracy of his future analytical model would improve. As predictive accuracy was one of the goals that Todd had laid out for him, he felt removing the outliers was the appropriate action.

After finishing a further evaluation of data scatterplots with other variables included, Nate was confident he had addressed outliers adequately. He then summarized all the variables in the data set in a table of summary statistics (Table 1).

Table 1. Data Summary Statistics

	Price	Hours	hp[a]	Rear Remotes[b]	Cab[c]	Air[d]	Heat[e]	Repair/Salvaged[f]
Average	$30,806.29	843	65	0.88	0.47	0.37	0.26	0.05
Standard Deviation	$10,756.50	1022	17	0.94	0.50	0.48	0.44	0.22
Minimum	$4,730.00	0	40	0	0	0	0	0
Maximum	$76,000.00	9717	99	4	1	1	1	1

Notes: Total sample size *n* = 1,100 with six tractor makes: John Deere = 450, Kubota = 203, Mahindra = 155, New Holland = 148, Case IH = 76, and Massey Ferguson = 68.
[a] hp = tractor engine horsepower.
[b] Rear Remotes = the number of rear remote auxiliary hydraulics.
[c] Cab = is an indicator variable equal to 1 if the tractor has a cab and equal to 0 otherwise.
[d] Air = is an indicator variable equal to 1 if the tractor has air conditioning (AC) and equal to 0 otherwise.
[e] Heat = is an indicator variable equal to 1 if the tractor has heat and equal to 0 otherwise.
[f] Repair/Salvaged = is an indicator variable equal to 1 if the tractor requires major repairs or has been categorized as a salvage only vehicle and equal to 0 otherwise.

In addition to the sale price and hours of usage, the data set also included variables for the tractor make/model, engine horsepower, and the number of rear remote hydraulics. From the tractor advertisement descriptions, Nate was also able to create four additional variables that he felt could be useful in his analysis. These variables included *Cab, Air, Heat*, and *Repair/Salvaged* and were all created as indicator (dummy) variables. An indicator variable can be coded in various ways but most often takes on the value of 1 for any observation that includes the specific trait suggested by the variable name and 0 otherwise. Thus, for example, Nate placed a one under the variable *Cab* in his data set for any tractor observation that was described as including a cab. If a tractor did not include a cab, Nate placed a 0 under the *Cab* variable for that observation. He created the variables *Air*, *Heat*, and *Repair/Salvaged* similarly by observing from the ad descriptions if tractors included air conditioning (AC) and heating, or required significant repairs (or classified as salvage only). After creating the summary statistics table, Nate reflected on what he could learn from the table.

> "*My total sample of 1,100 observations should be adequate for my analysis, and it's good to see that I have many observations for each of the various tractor makes (John Deere = 450, Kubota = 203, Mahindra = 155, New Holland = 148, Case IH = 76, and Massey Ferguson = 68). This should allow me to make good comparisons across tractor makes. I also have good variability within the other variables. The averages of my indicator variables (Cab, Air, Heat, and Repair/Salvaged) let me know the proportion of my sample that have the characteristics*

indicated by the variables, so 47 percent of the tractors in the data set have a cab, 37 percent of those cabs have AC, and 26 percent are heated. I would think that having a cab and AC/heat would be correlated with higher auction prices. Many of the tractors that have AC will also have heat, and if a tractor has either AC or heat, it will naturally also have a cab."

Nate took a mental note that this relationship between *Cab*, *Air*, and *Heat* suggests that they are correlated with each other and could present difficulties in his analysis. He would revisit this topic as he progressed with the statistical analysis. Nate also noted from his *Repair/Salvaged* variable that only 5 percent of the tractors needed significant repairs or were salvaged.

"*I bet these tractors are highly negatively correlated with price,"* he thought*. "Before moving to more advanced statistical methods, I should evaluate the conditional average auction prices for various tractor attributes*."

Nate had worked extensively with Microsoft Excel (Microsoft Corporation 2018) managing data sets during his time at school and felt confident in his ability to quickly create a summary table of conditional average prices and average horsepower using pivot tables. Nate created the table (Table 2) to display the conditional average price and horsepower for each of his indicator variables as well as for each of the tractor makes. These conditional averages helped Nate quickly identify the direction of the relationship between these variables and the auction prices as well as make comparisons between the levels within a variable. Comparing the conditional average prices between the various tractor makes, Nate found that within his data set Case IH tractors sold for the highest average price ($34,473) followed closely by John Deere ($33,953). The Mahindra tractors sold for the lowest average price at $21,980, with all other makes right around $30,000. These findings were not surprising to Nate, as he felt John Deere and Case IH had long been held in high regard as quality brands that command top-dollar prices. Mahindra on the other hand is the top tractor manufacturer in the world (Tractor Junction 2022) and specializes in producing quality tractors at an affordable price. Nate also knew that Mahindra had a

Table 2. Conditional Average Tractor Prices by Variable Levels

Variable	Level	Price	Horsepower
Make	Case IH	$34,473	71
	John Deere	$33,953	67
	Kubota	$29,828	62
	Mahindra	$21,980	56
	Massey Ferguson	$30,064	72
	New Holland	$30,282	69
Cab	Yes	$36,337	62
	No	$25,901	70
Air	Yes	$36,113	63
	No	$27,738	70
Heat	Yes	$35,817	64
	No	$29,021	69
Repair/Salvaged	Yes	$19,074	66
	No	$31,447	59

strong presence in the lower horsepower tractor market in the United States and thought that the lower average price could also be influenced by a lower average horsepower of the Mahindra tractors in the data set. Returning to his pivot table, he found that the average horsepower of the Mahindra tractors was 56 as compared to the average of the other makes ranging from 62 to 72.

This analysis of conditional averages demonstrated an important principle that Nate had learned while in school. The analysis showed that the Mahindra tractors had the lowest average price but also the lowest average horsepower. Thus, it was impossible to tell if the price was lower on average because of the quality effect of the manufacturer or if the price was lower simply due to the lower average horsepower. Nate assumed the answer was some combination of the two possibilities but knew that other variables also had an effect on price that were not included in this analysis of conditional means. For example, Nate suspected that tractors sold without a cab were cheaper than those with a cab, but having a cab may also be correlated with horsepower, make, and so on. Given the objective from Todd, Nate knew it would be important for him to estimate the marginal effects on the auction price of the variables while holding all else constant. The marginal effects would provide John Deere with marginal values of various tractor attributes for the used tractors that the dealer network took in on trade. The marginal values in turn could be used to better price those used tractors for the resale market.

3 Hedonic Price Analysis

To take his analysis to the next level, Nate needed to calculate the marginal values of a used tractor's attributes on the total auction price. Nate recalled the theory and techniques he had learned in his advanced agricultural marketing course at the university concerning hedonic price theory. At its core, hedonic price theory states that the total value of a good is equal to the sum of the values of its individual attributes. Court (1939), in his paper which created a hedonic pricing index for automobiles, is often credited as being the first to demonstrate the basic principles of hedonic analysis (Goodman 1998), though he did not formalize the theory. While several other researchers used a similar approach following Court (1939), it was not until 1966 in Lancaster's seminal paper on consumer theory where hedonic theory began to take form. Lancaster (1966) broke away from the traditional consumer theory at the time wherein it was assumed that goods were the direct objects of utility. Instead, Lancaster suggested that the total utility derived from consumption of a good could be decomposed into the utility provided by the individual characteristics or attributes of the good. Lancaster's work focused on how consumers make decisions given a choice set of goods, each providing utility equal to the sum of their individual attributes. Thus, there was no connection between Lancaster's new consumer theory and the market equilibrium or pricing. This gap was filled in 1974 when Rosen formalized the theory that a good's market value (price) can be decomposed into a sum of the individual values of its utility generating attributes. Rosen (1974) demonstrated how this theory could be applied in a hedonic regression analysis by using a good's price as the dependent variable regressed upon variables indicating the good's attributes to determine the way in which each attribute uniquely contributes to the price. Not only could such a regression analysis provide these marginal values of the attributes of a good, but the resulting estimated regression equation could be used to predict prices for a good based on the sum of its parts. For example, hedonic regression is often applied to housing markets to predict the price of a house based on the individual value of its attributes such as square footage, number of bedrooms, number of bathrooms, and so on. Nate felt hedonic regression analysis would be the ideal method for him to be able to accomplish the objectives given to him by Todd.

Before estimating any regression equation, Nate knew the importance of defining the equation to be estimated as well as forming a hypothesis for the sign of each variable included. Nate had been taught that this was an important step in the research process for a couple of reasons. First, clearly defining the regression equation to be estimated can help reduce p-hacking tendencies, that is, the tendency for researchers to collect and analyze data in such a way to represent statistically significant effects when

there may be no real underlying effect at all (Head et al. 2015). One questionable data practice that contributes to p-hacking is adding or removing variables to regression equations in an attempt to increase the goodness of fit while ignoring theoretical considerations for inclusion or deletion of variables. Second, forming an educated hypothesis of the sign of each variable before estimation can help the researcher evaluate the equation post-estimation and identify possible problems with the model.

3.1 Previous Literature

One of the best methods to establish which variables should be included in a regression equation is to reference published literature with similar objectives to identify the consensus and theoretical implications for inclusion/exclusion of specific variables. Nate scoured the literature to gain better understanding of what variables he should include in his model. One of the earliest studies Nate found was Fettig (1963). Noting Court's 1939 study estimating hedonic price indices in automobiles, Fettig (1963) applied similar methodology to the tractor market. Fettig's (1963) regression equations predicting new tractor prices included horsepower as well as an indicator variable for whether the tractor was diesel powered (as opposed to gasoline). Fettig considered numerous other variables such as fuel efficiency, maximum pounds of pull, miles per hour at maximum drawbar horsepower, and weight of the tractor. However, Fettig ultimately excluded these other variables from the analysis because they were not statistically significant or were highly correlated with horsepower and provided no additional explanatory power. Fettig noted that tractor prices could vary substantially due to tractors various attachments (e.g., fast hitches, remote hydraulics, power steering, independent power takeoffs, etc.). To control for such variables without explicitly including them in the regression, Fettig adjusted the prices to strip tractors of added attachments or add-ons. Using these stripped prices as the dependent variable and the two explanatory variables, horsepower and diesel, the regression equations estimated for years 1950–1962 were able to explain approximately 87–95 percent of the variation in tractor prices. This percentage of the variation of the dependent variable explained by the models was indicated by the range of R-squared values.

Berck (1985) used a data set that could be characterized as mixed time series, cross-sectional. From this data set, he estimated a single hedonic regression equation for tractors for the years 1923, 1930, and 1933. His objective was to determine the value of technical progress over these years by estimating the values attributable to quality (attributes of the tractors) with the remaining difference attributed to technical progress. He included horsepower as well as indicator variables for the year in which the tractors were sold. He considered fuel efficiency but ultimately removed this variable because it was not found to be statistically significant. The resulting regression equation had an R-squared of 0.84, which Berck concluded was quite good considering the time series, cross-sectional nature of the data set.

Nate felt both previous studies provided a good foundation for him to begin his own hedonic tractor analysis, but they only analyzed new tractor prices and they were both very old studies. Nate knew that just as Fettig (1963) noted, tractor prices would undoubtedly vary greatly due to equipped add-on attachments. Stripping the tractor prices of the values of these add-ons allowed Fettig (1963) to control for them. Nate would not be able to make such an adjustment. His analysis was for used tractors that included a loader. However, beyond loaders, each tractor may or may not have been sold with other additional add-ons. Nate needed to find a more recent study with used tractor prices to see how best to control for known add-ons or features. In his research, Nate found just such an article by Diekmann, Roe, and Batte (2008). These researchers used hedonic regression to compare used tractor prices for tractors sold on ebay.com (online auction website) versus at in-person auctions. They pooled the tractor sales data from both ebay.com and in-person auctions and included a dummy variable in the model for type of auction. They then included a host of explanatory variables designed to account for quality differences in the tractors, many of which were associated with add-ons. These variables include horsepower, age,

diesel/gas, implement/no implements included, manual/automatic transmission, four-wheel drive/two-wheel drive, tractor make indicator variables, sold on weekend/weekday, and monthly seasonal dummy variables. In addition to these variables, the researchers also included squared variables for hours, horsepower, and age. Including them in this manner allowed for their effects to take a nonlinear form (increasing/decreasing at an increasing/decreasing rate). These researchers also evaluated additional functional forms of the regression equation and relied upon an endogenous switching regression. Their final model explained 83 percent of the variation in their used tractor auction prices as indicated by the R-squared value.

After considering the literature and theory, Nate specified his initial regression equation to be estimated as:

$$P_i = \alpha_0 + \beta_1 Year_i + \beta_2 Age_i + \beta_3 Age_i^2 + \beta_4 hp_i + \beta_5 hp_i^2 + \beta_6 Hours_i + \beta_7 Hours_i^2 + \sum_{j=8}^{13} \beta_j Make_i + \beta_{14} Air_i + \beta_{15} Heat_i + \beta_{16} Cab_i + \beta_{17} R\_Hyd_i + \beta_{18} Auto_i + \beta_{19} 2WD_i + \beta_{20} Rep\_Sal_i + e_i \quad (1)$$

where P_i is the auction sale price of the ith tractor; *Year* is the year in which the tractor was sold (2020, 2021, or 2022); *Age* is the tractor age in years; *hp* is horsepower; *Hours* is hours of usage; $Make_i$ represents five dummy variables included for manufacturer of the i^{th} tractor (John Deere omitted as the reference base); *Air*, *Heat*, and *Cab* are all dummy variables equal to 1 if the tractor includes AC, heat, or a cab, respectively, and equal to 0 otherwise; R_Hyd the number of rear hydraulic remotes included; $Auto$, $2WD$, and Rep_Sal are dummy variables equal to 1 if the tractor has an automatic transmission (i.e., hydrostatic or continuously variable transmission); two-wheel drive or needs repairs/salvaged, respectively, and equal to 0 otherwise; and e_i is the stochastic error term assumed to be normally distributed with a mean of 0 and constant variance.

Nate ran over the variables quickly in his head to think about his hypothesized signs for each.

> "***Year**, I would expect to be positive as we have seen tractor prices inflating over the last few years. **Age**, on the other hand, I would expect to be negative as older tractor models I would expect to be correlated with lower auction prices. Horsepower should take on a positive sign as higher horsepower tractors should command higher expected prices. **Hours** should be negative since more hours indicates increased usage, which in turn would decrease the life expectancy of a used tractor. Since my dummy variables for make are all going to be relative to 'John Deere,' I believe they will all possess negative signs, with the possible exception of Case IH, which could be positive or negative depending on the effect of the other variables as the average prices of the tractors in the data set are very close for Case IH compared to John Deere. I do believe used John Deere tractors hold their value and sell for higher average prices relative to the other makes I have included in the data set. **Air, Heat, Cab**, and the number of rear hydraulics should all have positive signs as I believe buyers value these attributes positively. A tractor that is only two-wheel drive as compared to a 4×4 should be a lower value, so I would expect a negative sign for that variable. Finally, any tractor in need of repairs or classified as salvaged I would expect to be heavily discounted compared to those in good repair, so I would expect a negative sign on that dummy variable as well*."

To reaffirm his hypotheses, Nate created scatterplots of the various explanatory variables on the *x* axis, with price on the *y* axis. Figures 2 and 3 contain the scatterplots of horsepower and hours, respectively. With a trendline added to these scatterplots, Nate was able to compare his hypothesized signs with the direction of the trendlines.

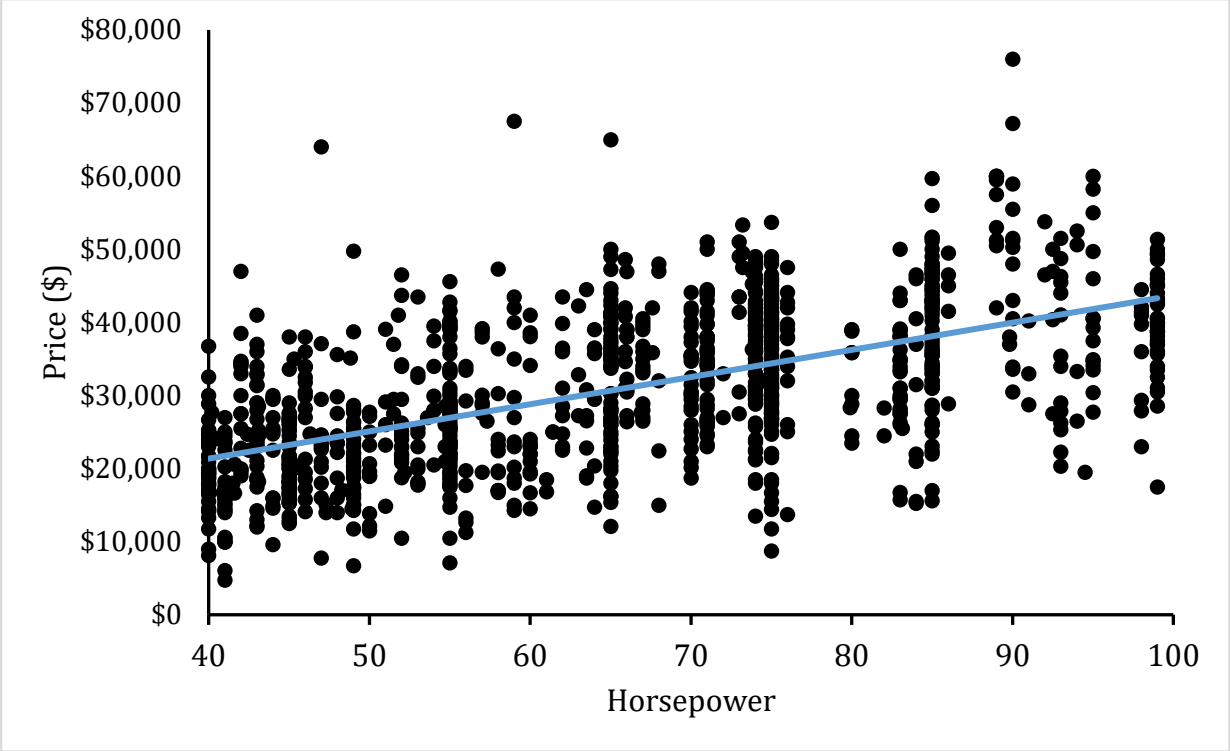

Figure 2. Price versus Horsepower Scatterplot

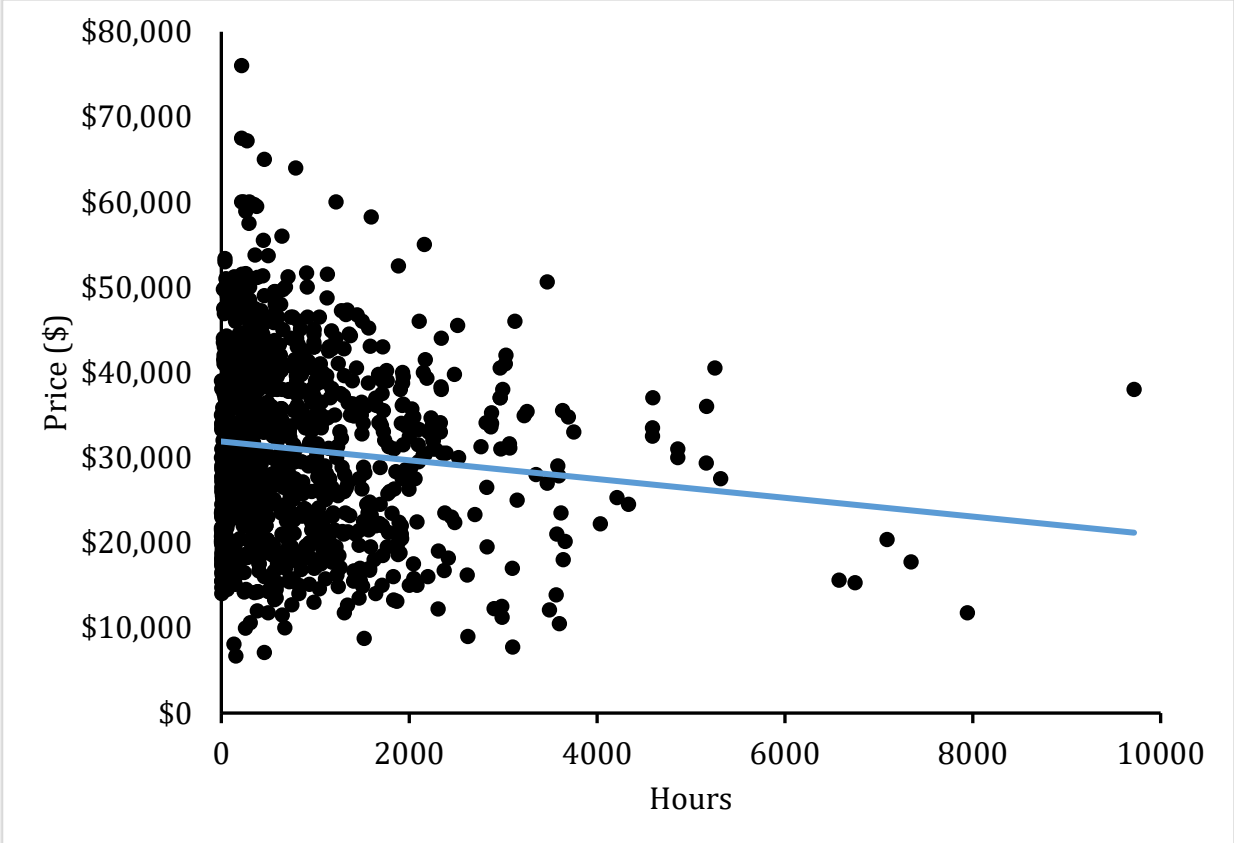

Figure 3. Price versus Hours Scatterplot

After completing the scatterplot analysis, Nate felt the next step was to estimate the regression equation and evaluate the results for potential problems or possible improvements. Nate estimated the regression equation using ordinary least squares (OLS) for Equation 1 on his computer; results are shown in Table 3.

Table 3. Initial Regression Results

Variable	Coefficient	Standard Error	*p*-Value
Year	2671.87	247.84	0.000
Age	-739.78	242.13	0.002
Age^2	19.12	19.68	0.332
hp	127.56	86.76	0.142
hp^2	1.61	0.64	0.012
Hours	-5.21	0.41	0.000
$Hours^2$	0.0004	0.0001	0.000
MAKES			
Case IH	-1856.85	687.34	0.007
Kubota	-2957.47	470.12	0.000
Mahindra	-10221.67	556.16	0.000
Massey Ferguson	-6018.95	731.47	0.000
New Holland	-4123.33	529.74	0.000
Air	466.67	710.51	0.511
Heat	221.57	611.84	0.717
Cab	7025.00	570.79	0.000
R_Hyd.	739.09	195.58	0.000
Auto	701.94	500.57	0.161
2WD	-3764.26	660.75	0.000
Rep_Sal	-4745.79	769.65	0.000
Constant	-5378111.00	500632.50	0.000

Note: All variables are as defined in Equation 1.
Adjusted R-squared = 0.7422
RMSE = 5,464
n = 1,100

Looking at the initial results, Nate had several takeaways—(1) the signs of all variables were as he hypothesized; (2) the negative sign on the *Age* estimated coefficient together with the positive sign on the Age^2 coefficient suggested that the effect of *Age* on price was one that was decreasing at a decreasing rate; (3) there was a similar decreasing at a decreasing rate relationship with *Hours*; (4) the positive signs on the *hp* and hp^2 coefficients suggested that the effect of *hp* on price was one that was increasing at an increasing rate; (5) all variables other than Age^2, *hp*, *Air*, *Heat*, and *Auto* were statistically significant at the 5-percent level (*p* value < 0.05); (6) the goodness of fit of the model as evaluated by the adjusted R-squared suggested that 74.2 percent of the variation in the used tractor auction prices could be explained by the variables included in the model, while 25.8 percent of the variation was left unexplained by variables not included in the model.

Nate felt that the results were reasonable. Although the adjusted R-squared was lower than those found in previous literature he had read, he was not surprised by this result. The studies he had read were quite dated. Tractors have seen large technological and mechanical advancements through recent

decades, which would suggest that capturing the variability in used prices would be much more difficult as the tractors could differ substantially in the attributes they now possessed. The R-squared was also lower than the more recent study he had reviewed of Diekmann, Roe, and Batte (2008). However, this too did not come as a surprise to Nate. These researchers had used regression techniques more sophisticated than OLS to help improve the goodness of fit. Nate was unfamiliar with the methods these researchers used and felt his simple model could still be useful to accomplish his objectives.

Before going further into the interpretation of the coefficients, Nate wanted to evaluate the model for common potential problems that arise in regression analysis.

3.2 Issues of Scale

Nate was initially concerned by the estimated value of the constant of -5,378,111. However, then he remembered how a constant can easily be manipulated to a more interpretable number by adjusting the scale of key explanatory variables. In this case, Nate recognized the *Year* variable as the one needing to be rescaled. Currently, it was not scaled at all, meaning if the tractor was sold in 2020, the value of the *Year* variable would be equal to 2020. However, because there were only three years contained in the data set (2020, 2021, and 2022), Nate recognized that if he rescaled the variable to be equal to the year the tractor was sold less 2019, the constant value could take on a more meaningful value while the marginal value for year would be left unaffected.

3.3 Multicollinearity

Multicollinearity refers to a problem that arises when two or more variables are highly correlated with each other. Independent (left hand side) variables as the name suggests should be independent of one another. When independent variables are highly correlated within a regression equation, it can reduce the precision of the coefficients of the correlated variables. Nate recalled his concern about the variables *Cab, Air*, and *Heat* being correlated. To evaluate the degree of multicollinearity within his variables, Nate calculated the variance-covariance matrix (Table 4). Nate had been taught a rule-of-thumb that any variables with a correlation coefficient of >0.7 could cause multicollinearity problems in a regression equation, and specification changes should be considered. Looking at the variance-covariance matrix, Nate identified, just as he expected, *Cab, Air*, and *Heat* as the only variables with a correlation coefficient >0.7. Nate considered what to do about his multicollinearity issue he had identified.

> "*I could simply omit variables **Air** and **Heat**. They are collinear with **Cab**, and judging by the initial parameter estimate for Cab, it appears to be much more influential toward price as having AC or heat,"* he thought.
>
> "*However, I do feel like people value AC and heat to some degree. Perhaps, if I combined the **Air** and **Heat** variables into one dummy variable that is equal to 1 when a tractor includes AC, heat, or both and equal to 0, otherwise this might resolve my multicollinearity issue*."

3.4 Heteroskedasticity

As Nate thought about other potential problems his model could have, he considered his error term, e_i as specified in Equation 1. Nate was using OLS as his estimator for Equation 1. The OLS estimator can have many desirable properties but only if the standard set of assumptions for this estimator are met. One such assumption is that the error term be normally and independently distributed with a zero mean and constant variance. Constant variance is said to be homoskedastic, whereas if the errors exhibit unequal variance the error term is said to be heteroskedastic (Kaufman 2013). One way to quickly evaluate

Table 4. Variance-Covariance Matrix for Variables Included in Initial Regression

Variable	Price	Year	Age	hp	Hours	Air	Heat	Cab	Auto	R_Hyd.	2WD	Rep_Sal
Price	1											
Year	0.21	1										
Age	-0.25	-0.07	1									
hp	0.58	0.01	0.07	1								
Hours	-0.10	0.02	0.44	0.31	1							
Air	0.38	-0.03	0.03	0.19	0.07	1						
Heat	0.28	-0.04	0.05	0.13	0.00	0.78	1					
Cab	0.48	0.00	-0.02	0.25	0.07	0.80	0.63	1				
Auto	-0.15	0.00	-0.02	-0.42	-0.08	0.00	0.02	0.02	1			
R_Hyd.	0.31	-0.02	0.07	0.38	0.11	0.22	0.23	0.19	-0.11	1		
2WD	-0.12	0.01	0.04	-0.03	-0.04	-0.10	-0.10	-0.11	-0.06	-0.05	1	
Rep_Sal	-0.26	-0.07	0.14	-0.08	0.03	-0.02	0.01	-0.05	-0.01	-0.06	0.03	1

whether a model suffers from heteroskedasticity is to plot the residuals against the fitted (predicted) values. Residuals are the difference between the actual values of the dependent variable and the predicted values. Nate created and evaluated the residual versus fitted values plot (Figure 4) for his hedonic regression equation. Nate recognized an undeniable pattern in his residuals right away. The cone-shaped pattern, as indicated by the blue lines Nate included in Figure 4, was a classic sign of heteroskedasticity within a model. The residuals exhibit a pattern of increasing in variance as the predicted prices increase in magnitude.

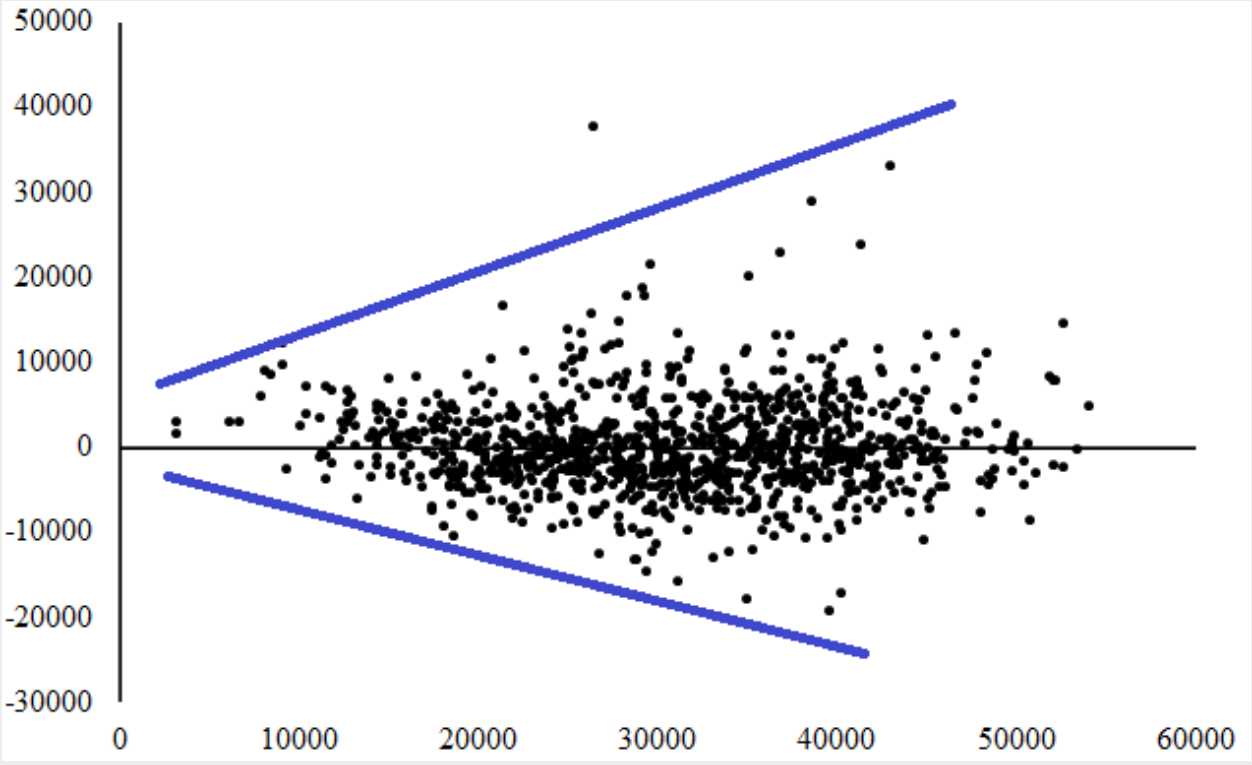

Figure 4. Residuals versus Fitted Values (Evaluation of Heteroskedasticity)

Nate recalled what he had learned about the problems with heteroskedasticity.

> "*For a model that meets all other assumptions of OLS, one with heteroskedastic errors will still produce unbiased coefficient estimates. This means I can still make good predictions with my model and can rely on the marginal values of the tractor attributes. But with heteroskedasticity, the standard errors of the coefficients will be biased. This means that unless I correct for the heteroskedasticity, I will not be able to make reliable statistical inferences from my results*." (Kaufman 2013)

Nate ran through the prescribed methods for addressing heteroskedasticity.

1) Transforming the dependent variable. This requires that a transformation be identified that is variance-stabilizing and has the downside of changing the scale of the dependent variable and complicates the interpretation of the marginal affects.
2) Use weighted least squares (WLS) in place of OLS. WLS is the optimal estimator for heteroskedastic data but requires the researcher to know or estimate the structure of the unequal variance.
3) Leave the heteroskedasticity in place but re-estimate the standard errors of the coefficients using a method that is robust to heteroskedasticity. The upside of this method is that no knowledge of what is causing the heteroskedastic errors is required, but a downside is that it is only suitable when working with large sample sizes because the OLS estimator will still be inefficient (Kaufman 2013).

Nate thought that some of the unequal variance could be proportionate to the horsepower variable. However, because Nate was unsure of the structure of the variance and his sample size was large (n = 1,100), he felt the best solution was the third method: use OLS with heteroskedastic robust standard errors estimated. This would allow for him to make correct statistical inferences about the significance of the explanatory variables without changing the parameter values estimated with OLS.

3.5 Irrelevant Variables

As Nate continued to look at his initial model results, he considered variables that may be irrelevant. Irrelevant variables can often be identified as those not statistically significant and not backed by theoretical reasoning. Inclusion of such variables has similar consequences as heteroskedasticity in that the coefficients estimated remain unbiased, but their variances are increased. This tends to understate statistical significance of the relevant variables included in the model (increases p values) and can lead to incorrect statistical inferences. As Nate considered the variables in his model, he felt that Age^2 fit the description of an irrelevant variable and determined that he would drop it from his final specification.

3.6 Variable Misspecification

One final problem that Nate considered was the possibility of variables being misspecified. He reflected on the specification of the *R_Hyd* (the number of rear hydraulic remotes) variable. Nate had specified this variable as a continuous variable. When he considered the values this variable could take on, he felt a change was in order. Looking over the summary statistics of his variables (Table 1), Nate noted that *R_Hyd* had a minimum of 0, a maximum of 4, and an average of 0.88. Although the variables average could be computed as any real number, the variable was discrete in that it only took on values from 0–4 in whole numbers (integers). Nate recalled from his schooling that discrete variables are often better represented through a series of dummy variables. Therefore, Nate determined he would remove the continuous *R_Hyd* variable and instead include three dummy variables *Rear1, Rear2, and Rear3.* These

variables would be equal to one if the tractor contained one, two, or at least three rear hydraulic remotes, respectively, and equal to zero otherwise. Including them in this manner would mean that the reference group would be tractors without rear remotes and the marginal values of these attributes would be interpreted relative to the reference group. Nate felt it was best to include tractors that had four rear remotes in the variable *Rear3* because upon inspection of the data, he found only three observations with four rear remotes included.

3.7 Final Model Results and Discussion

Nate made the changes to rescale the *Year* variable, account for the multicollinearity problem (combine AC and Heat), correct for heteroskedasticity (robust standard errors), drop irrelevant variables (Age^2), fix the misspecification of the rear remote hydraulic variable (change from continuous to discrete), and then re-estimated the regression equation (results in Table 5).

Table 5. Final Regression Results

Variable	Coefficient	Standard Error[1]	*p* Value
Year	2650.51	224.83	0.000
Age	-510.36	76.61	0.000
Age^2	172.30	89.68	0.055
hp	1.20	0.68	0.076
hp^2	-5.30	0.40	0.000
Hours	0.0004	0.00006	0.000
MAKES			
Case IH	-2103.92	683.99	0.002
Kubota	-3071.34	458.97	0.000
Mahindra	-10405.18	482.29	0.000
Massey Ferguson	-6086.24	714.00	0.000
New Holland	-4316.69	525.68	0.000
Auto	710.31	424.67	0.095
Air_Heat	766.89	619.29	0.216
Cab	6813.91	665.59	0.00
Rear1	-682.10	398.81	0.087
Rear2	1728.32	437.71	0.000
Rear3	2793.32	1188.62	0.019
2WD	-3761.50	915.73	0.000
Rep_Sal	-4661.08	851.67	0.000
Constant	15229.17	2893.00	0.000

[1] Robust standard errors calculated to correct for heteroskedasticity of the error term.
Adjusted R-squared = 0.7498
RMSE = 5430.1
n = 1,100

Nate reflected on his new results.

> "*The signs of all variables fit my original hypotheses. All variables are significant at the 5 percent level other than* ***hp***, ***hp²***, ***Auto***, ***Air_Heat***, *and* ***Rear1***, *and a couple of those variables are very close to significant (especially* ***hp*** *with a p value of 0.055), and all possess signs and magnitudes that would coincide with theory*."

He then considered the goodness of fit and predictive accuracy.

> "*The adjusted R-squared improved marginally from my original model as well as my RMSE. Since the RMSE represents the square root of the variance of my residuals, it has the useful property of being in the same unit as my dependent variable (price) and gives me an idea of how closely predictions using my model would be expected to match actual values*."

Nate evaluated the magnitude of the coefficients estimated and performed a mental interpretation for a few of them.

> "*For any variable not included as a squared term, I can interpret the coefficient itself as the variable's marginal effect. This is because a marginal effect of any variable can be calculated as the partial derivative of the price equation with respect to that variable. This suggests that for a variable included as both a linear and squared term the marginal effect is not constant. If I take, for example, the partial derivative of the price equation with respect to* ***hp****, I find its marginal effect to be equal to* **172.3 + 2.4** ***hp***."

Nate took mental note that this suggests that the marginal effects of variables included as squared terms depend on the level of the variable themselves.

> "*For variables included only linearly, the marginal effects are constant. The coefficient for Year of 2,650 suggests that holding all other variables constant, used tractor prices have been increasing by $2,650 each year over the years 2020–2022. Todd will be keen to see this result as it addresses his third objective concerning inflation of used tractor prices*," Nate thought.

> "*A coefficient of -510 for* ***Age*** *suggests that while holding all other variables constant, for each additional year in age of a tractor, its value would be expected to decrease on average by $510. All my "Make" variables included are relative to the reference make of John Deere. Since the coefficients for the other makes are all negative and statistically significant, I can conclude that I would expect all other makes to be discounted relative to a John Deere tractor by the value of their coefficient holding all other variables constant. This should help to address Todd's second objective. Switching my* ***R_Hyd*** *variable to discrete dummy variables was a good idea. I can now get an idea of the marginal differences between various quantities of remotes. It appears that based on the lack of statistical significance for Rear1, buyers don't really value having only one rear remote as compared to none. However, increasing to two rear remotes suggests the tractor value would increase by $1,728 with another $1,065 added to those tractors that have three or more rear remotes. These marginal values are exactly what Todd is looking for and should work great to begin helping us set our prices on our used tractors*."

3.8 Making Predictions

Nate wanted to test the model to be sure he felt it could make reasonable predictions of price to accomplish Todd's first objective. Nate called one of the local dealerships and spoke with the manager. He asked the manager to provide him with an example of a tractor recently taken in on trade that had been sold. The manager told him they had just sold a 2017 Massey Ferguson 60 horsepower tractor last week for $32,000. It had 250 hours on it, was an automatic transmission, came equipped with two rear remote hydraulics, and included a cab. Nate quickly input the tractor's information into a spreadsheet and then using the marginal values of the attributes calculated with his hedonic regression equation, he estimated the tractor value (as in Table 6) to be $37,153. Nate thought the prediction was reasonable and suggested to him that the dealership had undervalued the tractor by about $5,000. Of course this was only one observation, and given the RMSE of the model was approximately $5,500, Nate felt the manager was not too far off with his pricing. Nate felt the next steps were to collect additional samples from other dealerships across the country and evaluate the performance of the model using out-of-sample data. After that, the only thing left to do was write up his results into a report that he could provide to Todd and the management team.

Table 6. Used Tractor Price Prediction Example

Variable	Coefficient	Attributes	Marginal Values
Year	2650.51	3	$7,952
Age	-510.36	5	-$2,552
Age^2	172.30	60	$10,338
hp	1.20	3600	$4,320
hp^2	-5.30	250	-$1,325
Hours	0.0004	62500	$25
MAKES			
Case IH	-2103.92	0	$0
Kubota	-3071.34	0	$0
Mahindra	-10405.18	0	$0
Massey Ferguson	-6086.24	1	-$6,086
New Holland	-4316.69	0	$0
Auto	710.31	1	$710
Air_Heat	766.89	0	$0
Cab	6813.91	1	$6,814
Rear1	-682.10	0	$0
Rear2	1728.32	1	$1,728
Rear3	2793.32	0	$0
2WD	-3761.50	0	$0
Rep_Sal	-4661.08	0	$0
Constant	15229.17	1	$15,229
		Total Expected Value	**$37,153**

About the Authors: Ryan Feuz is an Assistant Professor at Utah State University (Corresponding author: ryan.feuz@usu.edu).

References

Berck, P. 1985. "A Note on the Real Cost of Tractors in the 1920s and 1930s." *Agricultural History* 59(1):66–71.

Court, A.T. 1939. "Hedonic Price Indexes with Automotive Examples." *The Dynamics of Automobile Demand*. New York: General Motors.

Deaux, J. 2021. "Wild Bidding Wars Erupt at Used-Tractor Auctions Across the U.S." *Bloomberg*, November 13.

Diekmann, F., B.E. Roe, and M.T. Batte. 2008. "Tractors on eBay: Differences between Internet and In-Person Auctions." *American Journal of Agricultural Economics* 90:306–320.

Fettig, L.P. 1963. "Adjusting Farm Tractor Prices for Quality Changes, 1950–1962." *Journal of Farm Economics* 45(3):599–611.

Goodman, A.C. 1998. "Andrew Court and the Invention of Hedonic Price Analysis." *Journal of Urban Economics* 44(2):291–298.

Head, M.L., L. Holman, R. Lanfear, A.T. Kahn, and M.D. Jennions. 2015. "The Extent and Consequences of P-Hacking in Science." *PLOS Biology* 13(3):e1002106.

John Deere. 2021. "Deere Reports Net Income of $1.283 Billion for Fourth Quarter, $5.963 Billion for Fiscal Year [Press release]." https://www.deere.com/en/news/all-news/fy21-fourth-quarter-earnings/.

Kaufman, R.L. 2013. "Heteroskedasticity in Regression: Detection and Correction." SAGE Publications, Inc.

Lancaster, K.J. 1966. "A New Approach to Consumer Theory." *Journal of Political Economy* 74:132–157.

Microsoft Corporation. 2018. Microsoft Excel. https://office.microsoft.com/excel.

Rosen, S. 1974. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *Journal of Political Economy* 82(1):34–55.

Tita, B. 2021. "Deere's Strike Is Over, but Order Backlog, Higher Costs Remain." *Wall Street Journal*, November 22.

Tractor Junction. 2022. "Top 10 Tractor Companies in the World—Tractor List 2022," May 7. https://www.tractorjunction.com/blog/top-10-tractor-companies-in-the-world-tractor-list/.

©2022 All Authors. Copyright is governed under Creative Commons BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/). Articles may be reproduced or electronically distributed as long as attribution to the authors, Applied Economics Teaching Resources and the Agricultural & Applied Economics Association is maintained. Applied Economics Teaching Resources submissions and other information can be found at: https://www.aaea.org/publications/applied-economics-teaching-resources.