Predicting Mow-Times

Josiah Rottari

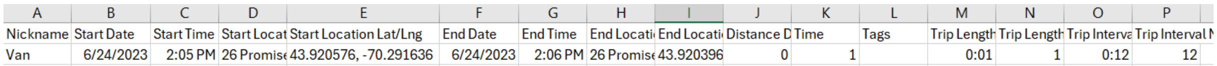
Spring 2024

Overview

This report seeks to answer the questions "To what extent is the size of a lawn associated with mowing and trimming times with a specific selection of equipment?" and "To what extent can we predict service-times using the square footage of a lawn?". Mowing is exactly what it sounds like and trimming is the process of using a string-trimmer or "weed-wacker" to cutback growth around various obstacles on a property (eg. houses, sheds, fences, gardens) that cannot be cutback with a mower. We will call the total time taken to perform these tasks "service time." Our goals here are two-field. We want to make inferences about the extent to which the size of the lawn affects service-time, in-so-far as our inferences have practically meaningful implications. We also want to know how well we can predict service-time which has implications for the service-provider. Our hypothesis is that lawn-size has a practically meaningful effect on mow-times and that useful predictions are possible. We are open to possibility of modeling some statistic based on service time other than the conditional mean and are also open to adding covariates to the model.

Data

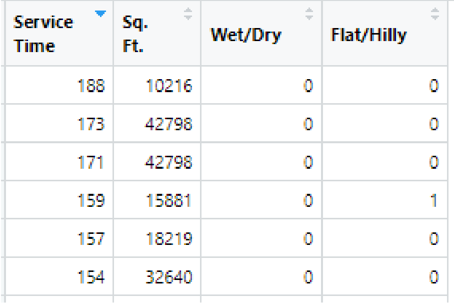
If you are just interested in reporducing results with the already cleaned data please see the Github. The van data that we start with looks like the following where we calculate the service time via a simple formula in Excel. There are 822 entries of this form and not every trip was for mowing, so in R we filter by the addresses of the serviced lawns and put a minimum on the amount of time spent at site since, for example, none of the lawns took less than 20 minutes to service. The data is a bit messy since the tracking software records whenever the van is started. After we filtered/cleaned the data, we added a column for the square footage of a lawn using mapdevelopers.com and added dummy variables to encode whether a lawn was Dry/Wet and Flat/Hilly.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Nickname	Start Date	Start Time	Start Locat	Start Location Lat/Lng	End Date	End Time	End Locati	End Locati	Distance D	Time	Tags	Trip Length	Trip Length	Trip Interva	Trip Interval N
Van	6/24/2023	2:05 PM	26 Promise	43.920576, -70.291636	6/24/2023	2:06 PM	26 Promise	43.920396	0	1		0:01	1	0:12	12

Figure 1: Van data with trip time calculated via simple excel formula.

The data looks like



Service Time	Sq. Ft.	Wet/Dry	Flat/Hilly
188	10216	0	0
173	42798	0	0
171	42798	0	0
159	15881	0	1
157	18219	0	0
154	32640	0	0

Figure 2: Service time with square feet and dummy variables.

where the sample size is $n = 295$. Please see the GitHub for csv file with the summary statistics or view the code used for this project to view these summary statistics. You can also compute them yourself using a csv file for the full sample of size $n = 295$ on GitHub. The R code used to clean the data, and the 822 pre-cleaned entries are available upon request since they contain sensitive information.

Models

This section has some technical details which can be skipped if the reader is only interested in results, but the notation outlined might be useful. Let y be the service time we'll initially model

$$y = X\beta + \epsilon$$

where $\beta \in \mathbb{R}^p$ for some p, X is the matrix of predictors, and $\epsilon \sim N(0, \sigma^2 I)$. That is, $y|\beta, \sigma^2, X \sim N(X\beta, \sigma^2 I)$. We are open to the possibility that y is linear as a function of some transformation of square feet such as $\sqrt{\text{square feet}}$ or $\log$ square feet. Since we have multiple observations for each lawn we are also open to letting $y* = \mathbb{E}[y|X]$, where when we use this notion we mean that we're conditioning on the size of the lawn, so that $\mathbb{E}[y|X]|\beta, \sigma^2, X \sim N(X\beta, \sigma^2 I)$ or, more generally, letting $y* = f(y)$ for some function f. We acknowledge that this is generally a very bad practice due to the reduction in information. First note that $\mathbb{E}[\mathbb{E}[y|X]|X] = \mathbb{E}[y|X]$ so in taking this approach we are still modeling the same quantity, just at the expense of loss of information which we have ways to account for. We believe the goal of this analysis is to be correct about the time it takes to do a lawn, over the course of an entire season, so that aggregating the data in this manner is justified. Furthermore, there is an incredible amount of noise in the data since sometimes a mower gets stuck or clients ask for additional tasks, so aggregation is a way to uncover the fundamental relationship of interest.

Exploratory Modeling Process

One thing to remain aware of and to clarify is what we mean by "practically useful". The model will be practically useful if a service-provider can make predictions with it to develop estimates that minimize the risk of losing profit over an entire season. In this sense, a model that overshoots actual service times will be better than one that is biased downward. Of course, this comes at the cost of the service-provider losing out on work in the bidding process, but we hope whatever model we develop can be a starting point before addressing the problem of developing some sort of optimal pricing strategy. Shown below are various plots for the exploratory analysis stage. Note that more models were tried than shown here, but we're displaying the ones that seemed possibly useful to motivate how we got to a more refined model. From here on out $x_1 = 1$ for the intercept, $x_2 = g(\text{square feet})$ for some function g and $x_{p-1} = \text{Dry/Wet}$ and $x_p = \text{Flat/Hilly}$ where $x_3, \ldots, x_{p-2}$ might also be functions of x_2 to leave open the possibility of a polynomial or rational function fit.

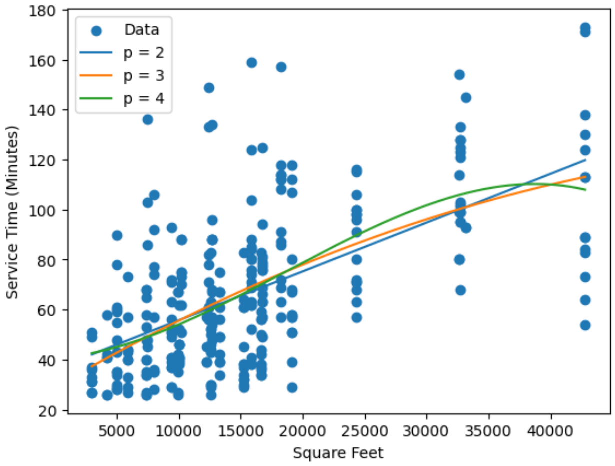


Figure 3: $p = 1 + \deg$ of polynomial fit. Dummy variables not included.

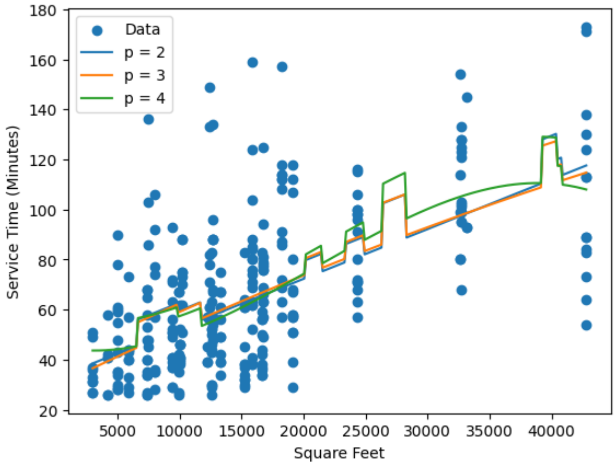


Figure 4: $p = 1 + \deg$ of polynomial fit. Dummy variables included.

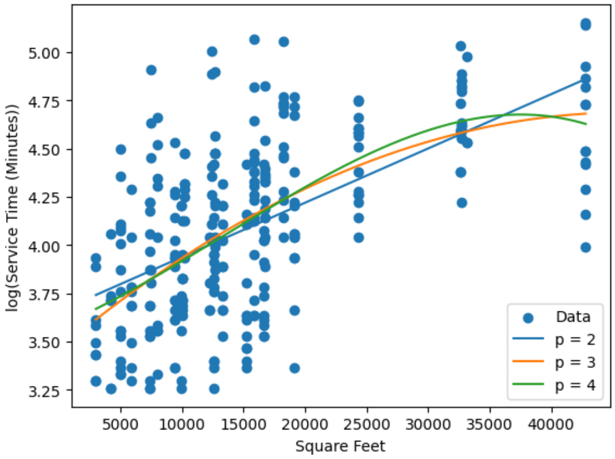


Figure 5: $p = 1 + \deg$ of polynomial fit. Dummy variables not included. Modeling $\log(y)$.

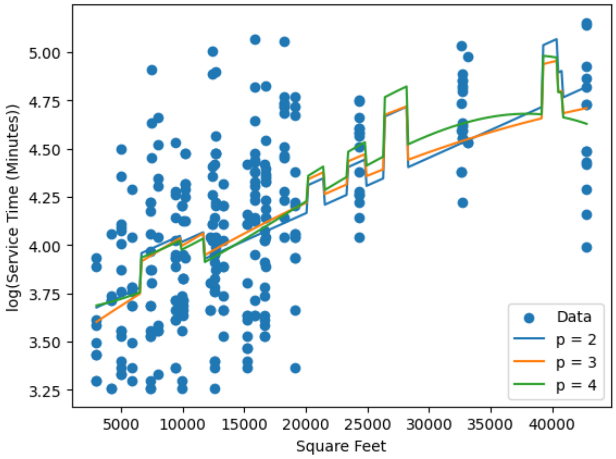
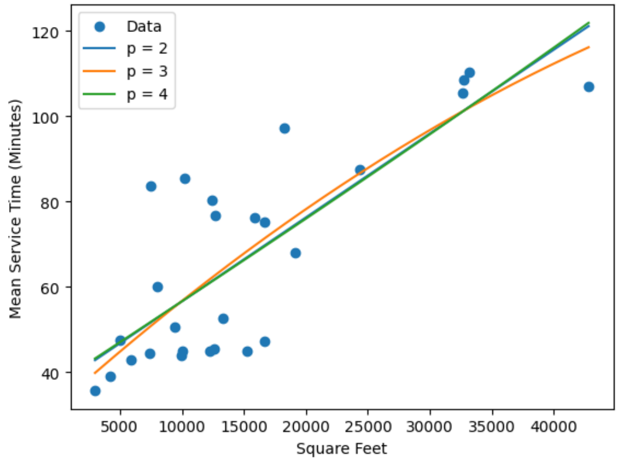


Figure 6: $p = 1 + \deg$ of polynomial fit. Dummy variables included. Modeling $\log(y)$.



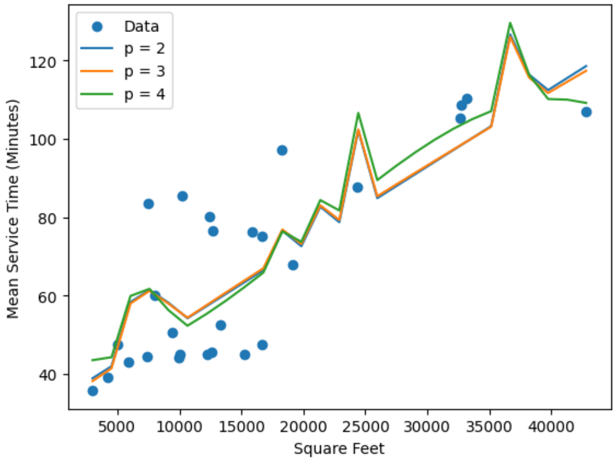
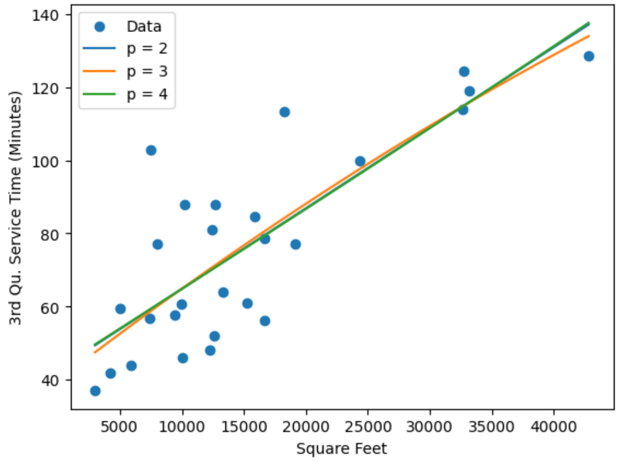


Figure 7: $p = 1 + \deg$ of polynomial fit. Dummy variables not included, for $\mathbb{E}[y|X]$.

Figure 8: $p = 1 + \deg$ of polynomial fit. Dummy variables included, for $\mathbb{E}[y|X]$



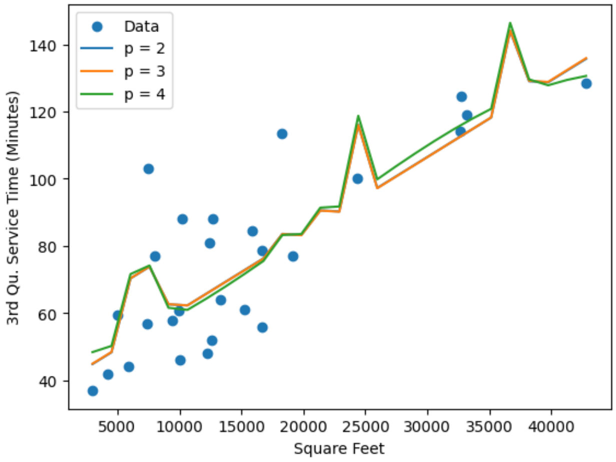


Figure 9: $p = 1 + \deg$ of polynomial fit. Dummy variables not included, for $3rdQu.y|X$.

Figure 10: $p = 1 + \deg$ of polynomial fit. Dummy variables included, for $3rdQu.y|X$.

Table 1: Exploratory Results for Various Models Tested

Quant. Modeled	Predictors	Norm. Resid.	β and Resid. Ind./ Homoscedasticity	LOO CV MSE Est.	5−Fold CV MSE Est.
y	x_1, x_2	T	F	623	650
y	x_1, x_2, x_2^2	T	F	622	651
y	x_1, x_2, x_{p-1}, x_p	T	F	604	681
y	$x_1, x_2, x_2^2, x_{p-1}, x_p$	T	F	609	709
$\log(y)$	x_1, x_2	T	F	.138	.147
$\log(y)$	x_1, x_2, x_2^2	T	F	.134	.138
$\log(y)$	x_1, x_2, x_{p-1}, x_p	T	F	.130	.153
$\log(y)$	$x_1, x_2, x_2^2, x_{p-1}, x_p$	T	F	.129	.150
$\mathbb{E}[y\|X]$	x_1, x_2	T	T	239	230
$\mathbb{E}[y\|X]$	x_1, x_2, x_2^2	T	T	263	255
$\mathbb{E}[y\|X]$	x_1, x_2, x_{p-1}, x_p	T	T	235	280
$\mathbb{E}[y\|X]$	x_1, x_2, x_{p-1}, x_p	T	T	275	308
$3rdQu.y\|X$	x_1, x_2	T	T	286	267
$3rdQu.y\|X$	x_1, x_2, x_2^2	T	T	304	279
$3rdQu.y\|X$	x_1, x_2, x_{p-1}, x_p	T	T	291	306
$3rdQu.y\|X$	$x_1, x_2, x_2^2, x_{p-1}, x_p$	T	T	322	325

Figures 3-10 show the 16 models tried. The models tried also includes higher degree polynomials than have results in Table 1. Some simulated and observational results are shown in the table. At this stage we were looking to find a model that gives a decent enough fit and that satisfies the model assumptions of a linear model. Our indication of the residuals being normal and the residuals and β being independent should be interpreted as "sufficiently true" versus not sufficiently true to make inferences based on the model. Normality was determined by observing the histograms and empirical CDFs of the residuals and by looking at the Shapiro test, the Anderson test, and the Kolmogorov-Smirnov test. Independence/homoscedasticity was determined by looking at permuted residual plots and residuals versus fitted plots. In the case of modeling y using the full data, we were never satisfied with the distribution of the residuals, hence the "F"s in the table. The hypothesis tests rejected normality, but we found their distributed to be normal enough. The lack of homoscedasticty was a major problem that we could not remedy. We tried multiple transformations of the predictors and the output variable as well as weighted-least-squares regression, but found now way to remedy this issue. Note that models with covariates of the form $1 + x_1 + x_1^{1/2} + \ldots$ were also tried, but in no case were models of this form significantly different from the polynomial models based on an eye test of the plots and the anlytical results reported in the table. Also note that all the plots with the dummy variables indicating whether a lawn is dry verse wet and flat verse hilly are actually the trace plots of the output versus the size of lawn. We also looked at the effects of regularization on these different results, but it did not produce any significant effects one way or the other, so we omit reporting the effects of any regularization in estimation and refrain from using regularization for the rest of the analysis.

In the case of modeling $\log(y)$ the assumptions of the linear model were all sufficiently met to proceed onto doing inference, but at the loss of interpretability and the ability to get results about future observations of y. The lower MSE estimate results in the case of $\log(y)$ are also an artifact of the change in scale and are of a similar magnitude of the other MSE when transforming $\log(y)$ back to the original scale. Hence, we find that the results when modeling just y are unsatisfactory when compared to the results we got modeling $\mathbb{E}[y|X]$ and $3rdQu.y|X$. Although these first results, might still be practically useful we find them to be less useful than the latter results.

For the cases where we model $\mathbb{E}[y|X]$ and $3rdQu.y|X$ we first describe the choice to look at the $3rd$ quartile. The data includes service times done with varying amounts of people. Some of the services were done using 2 people and some were done by a single person and we do not know which ones were done by a single person versus 2 people. Hence, for the purpose of making predictions for services done with one person we find it reasonable to look at the third quartile for two reasons. First, the third-quartile is more conservative, by definition, for the purpose of estimation. Second, for the reasons described about the number of people working, the third quartile may better track service times done by a single person.

With this out of the way, using knowledge based on industry experience and the data shown in the table, we consider it worth while to proceed to inference making around modeling $\mathbb{E}[y|X]$ as a function of x_1 and x_2 and as a function of x_1, x_2, x_{p-1}, and x_p. Likewise, for the third quartile model. That is, we'll proceed to inference making and further model validation and analysis for four different models. At this point we should note that there are concerns with including information about a lawn being wet vs dry since this knowledge isn't necessarily available a priori when making estimates. It's something that gets determined over time as one can make multiple observations about how a particular lawn retains moisture.

Classical Inference

In this section we stick to classical statistical methods. We report simultaneous intervals for the regression coefficients, the Scheffe intervals, for each model besides m5 in the table below along with point estimates. We also plot analytical, report their widths, and report MLE point estimaates. The models for $\mathbb{E}[y|X]$ and $3rdQu.y|X$ with just two parameters are called m1 and m3, respectively. We call the models that include the dummy variables for $\mathbb{E}[y|X]$ and for the third quartile m2 and m4 respectively. We call a weighted-least-squares model for the third quartile m5 and do not include the dummy variables for this model nor investigate a weighted-least-squares model with the dummy variables.

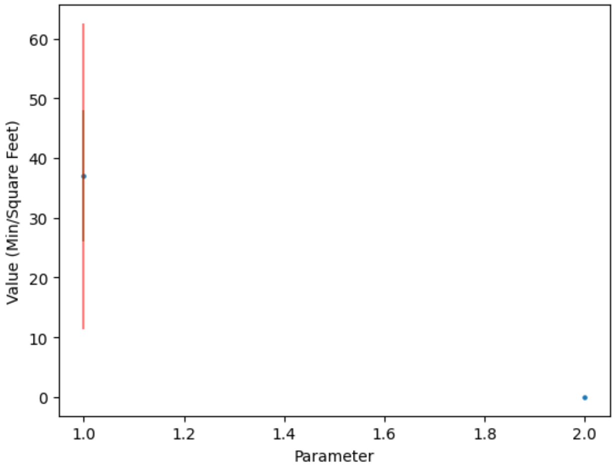


Figure 11: Model 1: Scheffe Red, t Green

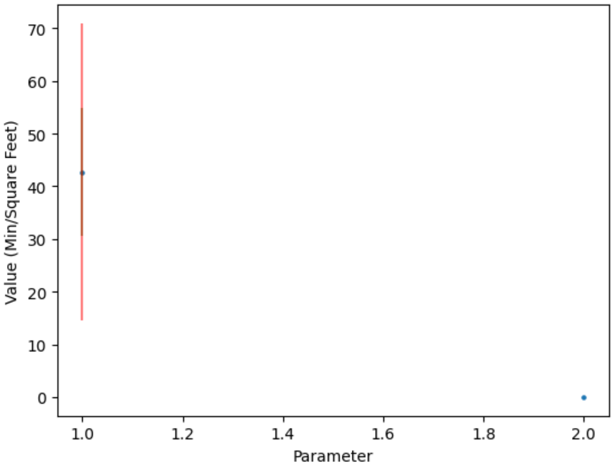


Figure 12: Model 3: Scheffe Red, t Green

Table 2: Confidence Interval Results for Models 1 and 3

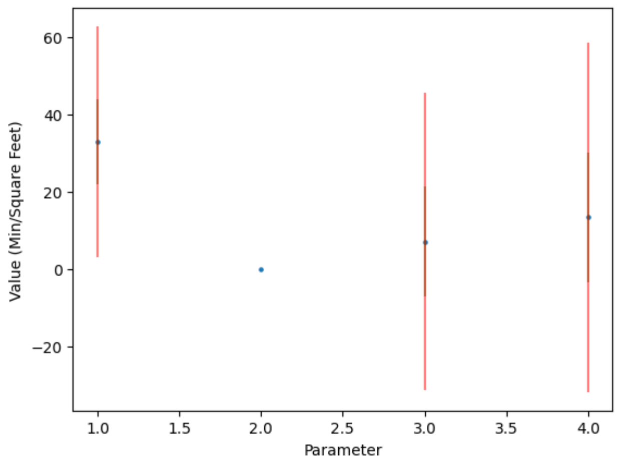


Figure 13: Model 2: Scheffe Red, t Green

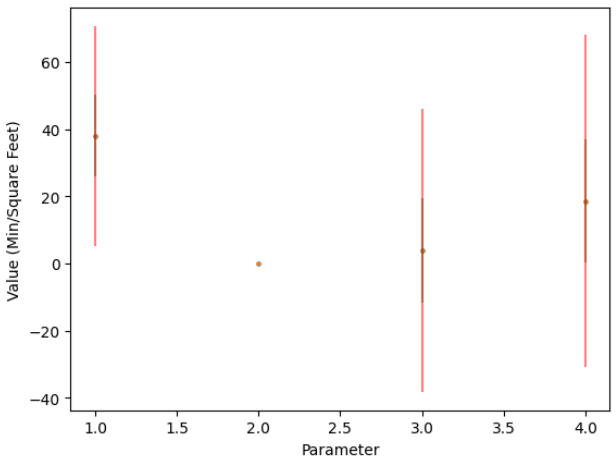


Figure 14: Model 4: Scheffe Red, t Green

CI Type (95%)	Model	Parameter	Lower	Upper	Point Est.	One-Side Width
t	m1	β_1	25.7	46.5	36.1	10.4
t	m1	β_2	0.0014	0.0026	0.0019	0.00058
Scheffe	m1	β_1	11.9	63	60.3	24.2
Scheffe	m1	β_2	0.00065	0.00333	0.0019	0.00134
t	m3	β_1	30.2	54	42.1	11.8
t	m3	β_2	0.0016	0.0029	0.0022	0.00065
Scheffe	m3	β_1	14.6	69.6	42.1	27.5
Scheffe	m3	β_2	0.0007	0.00374	0.0022	0.0015
t	m5	β_1	19	42	30.5	11.6
t	m5	β_2	0.002	0.0035	0.0028	0.00074

Table 3: Confidence Interval Results for Models 2 and 4

Model	Parameter	Point Est.	One-Side t Width	One-Side Scheffe Width
m2	β_1	32.8	11	29.9
m2	β_2	0.002	.00059	0.0016
m2	β_3	7	14.2	38.5
m2	β_4	13	16.7	45.2
m4	β_1	38	12.1	32.7
m4	β_2	0.0023	0.00064	0.0017
m4	β_3	3.80	15.6	42
m4	β_4	18.49	18.3	49

From the intervals, the only case in which we may reject $H_0 : \beta_4 = 0$ at the $\alpha = 0.05$ level using a t-test is for $m4$ and the Scheffe interval is massively inflated in this case as well. This evidence, in combination with the results shown in Table 1, lead us to the conclusion that the more complicated models which include all four parameters are not worth analyzing further, and so we narrow our analysis to models m1, m3, and m5 which only include an intercept and a coefficient for the square footage of a lawn. We now produce confidence bands around $\mathbb{E}[y|X]$ and the third quartile of y, at a variety of lawn sizes Figures 15-19.

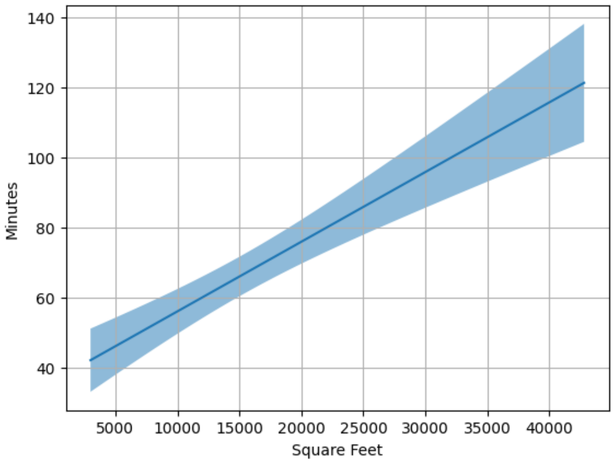


Figure 15: Model 1: 95% t-Confidence Bands

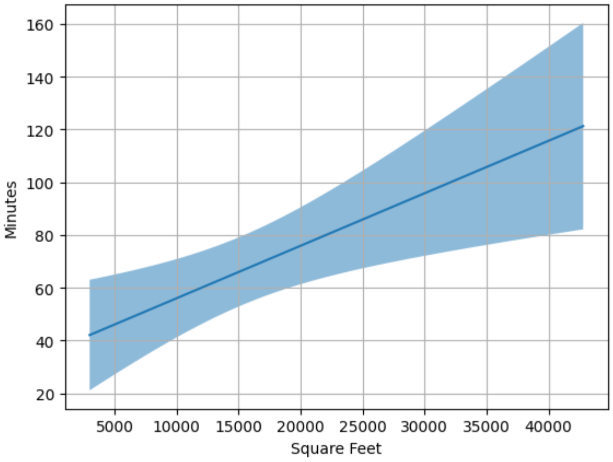


Figure 16: Model 1: 95% Scheffe Confidence Bands

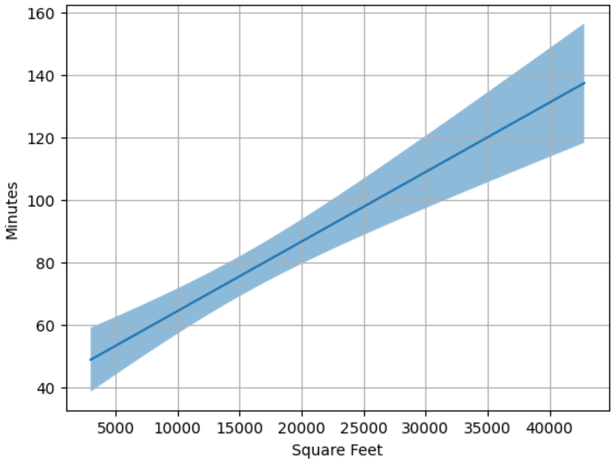


Figure 17: Model 3: 95% t-Confidence Bands

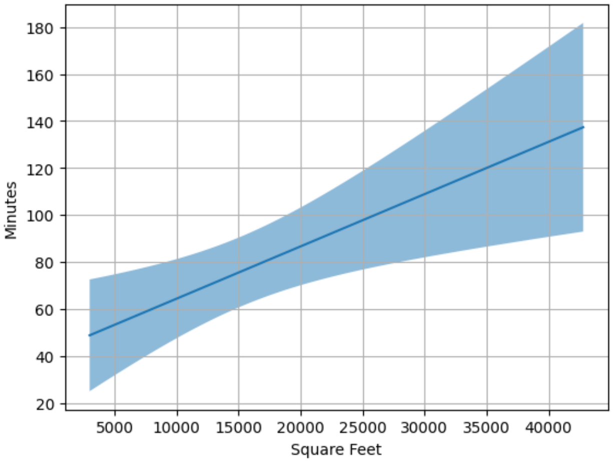


Figure 18: Model 3: 95% Scheffe Confidence Bands

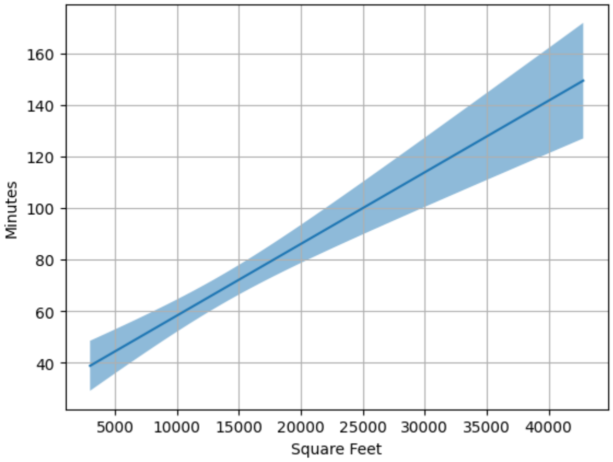


Figure 19: Model 5: 95% t-Confidence Bands

Conclusion

The variability in the data at our disposal makes this a difficult problem in general. The variability comes from equipment failures, weather dealys, the mower getting stuck, and varying numbers of people working. Nonetheless, we feel comfortable recomending the deployment of models m3 and m5 due to their conservative nature and feel comfortable with the times that they recommend. Moreover, in comparison with mowing time recommendations found on the internet such as 0.0011 found here and an estimate from a construction estimation book of 0.00102 minutes per square foot, which we've been personally burned by, we recommend our models as they at least somewhat account for time spent loading and unloading equipment, equipment failures, and getting stuck.