The PAST Model

Prices of Automobiles regressed STatistically

Gian Alix
gian.alix@gmail.com

Celina Landolfi
clandolfi17@gmail.com

Department of Mathematics and Statistics
York University

Project Report
June 22, 2020

Abstract. Automobiles are one of man's extravagant properties. However, when luxury turns into a need, this poses a major problem - indicating a negative impact to man himself, particularly if one is unable to make an acquisition. This is due to the rapid constant increase of auto prices, leading to affordability becoming an alarming issue.

In this report, our team investigates which factors contribute to the price of automobiles given a set of predictors. By becoming aware of these factors, one is able to identify the problem of which aspects or components of a car makes it economical, and which elements to look out for that impose a more expensive price. The methods we have applied in this report involve building a regression model, which our team called **PAST**. Through various testing, it was found that the best model was $y = \beta_0 + \beta_1 x_{7,4wd} + \beta_2 x_{7,rwd} + \beta_3 x_{11} + \beta_4 x_{20} + \beta_5 x_{21} + \beta_6 x_{11} x_{21} + \beta_7 x_{16} + \beta_8 x_{16} x_{21}$, which utilizes type of wheel drive, car width, engine size, compression ratio, horse power, as well as interaction terms between independent variables, in order to accurately predict the price of a car. The model has an adjusted R^2 of 0.8529 and a root mean square error value of 3048.34, with 9 parameters and a C_k value of 9. The model does not seriously violate any inference assumptions and proves to have good predictability, shown through cross validation.

Photo taken by Matthew Henry from `Burst`

1 Introduction

1.1 Problem Definition

Given a set of predictor values, our team intends to build a regression model that is able to predict the price of a *new* automobile. Some of these attributes include `engineSize`, `compressionRatio`, `wheelDrive` among others. Throughout the report, we shall name this desired model ***Prices of Automobiles regressed STatistically***, or **PAST** for short.

1.2 Rationale

In a [1] poll conducted by Gallup in April 2019, it has been reported that an average of 64% of Americans drive everyday, may it be to go to work, to school, or to anywhere that they need to go. As pointed out by [2] CNBC, car prices have been rapidly increasing over the years; and while this is healthy to the economy, it is detrimental to drivers, those who own cars, and people who make use of automobiles as their means of transportation. Data gathered by the [3] Kelley Blue Book suggests that the average price of a new car in September 2018 has increased by 2% from September 2017, and then another 6% hike to September 2019, as suggested in the bar graph of **Fig. 1**. This increase in car price could be problematic especially to those relying on cars mostly as a way to get from Point A to Point B. [2] CNBC has argued that this can push many to take out more auto loans, leading to prolonged monthly payments and higher interest rates. This begs the question then, *is one even able to afford a car?*

In this paper, we examine a dataset called [4] `auto price` (available on Kaggle), where we aim to develop a regression model that would predict the price of a *new* car based off of several features, such as the `engine size`, `horsepower`, `mileage`, to name a few. If we can successfully pinpoint which are the significant predicting factors that determine the price of a car, then this produces a convenient tool for identifying which factors to take into account whenever one decides to make a purchase.

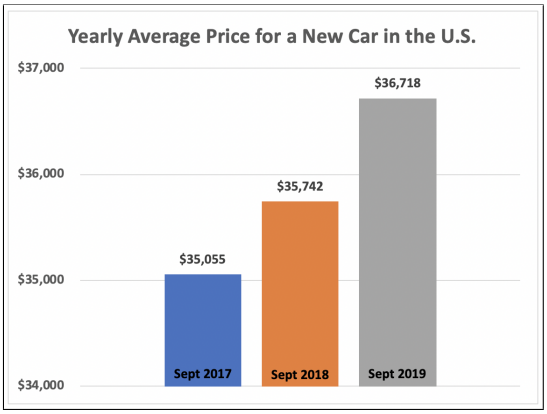


Fig. 1. A bar chart for the yearly average price for a brand new automobile in the U.S. from 2017-2019, based on the Kelley Blue Book [3].

1.3 Proposed Solution

Our team proposes to find a linear regression model **PAST**, with a specifically high predictive ability to determine the price of a new automobile given a set of features. Several methods and techniques will be used in the design of the model, some of which include the ever-popular *Stepwise Regression*, the addition of *Interaction Terms* and model improvement via *Outlier Handling*, to name a few.

2 Data

2.1 The Auto Price Dataset

[4] The `auto price` dataset, found on **Appendix A** and also publicly available on Kaggle, is the dataset our team used in our analyses. This appendix also contains a data dictionary of the dataset's features, with most predictors being self-explanatory. One particular variable, x_1, which denotes `symboling` is a measure given by actuarials that relates to risk. As the source in Kaggle does not fully clarify the details on this particular variable, our team regarded this an unimportant variable in our analysis.

2.2 Data Preprocessing

Data in the real world is known to be noisy, inconsistent, erroneous and can contain missing values. By performing data preprocessing on our dataset, we are able to transform the set into one that is useful - especially for our future data analyses. Preprocessing is therefore crucial in any analytic work on any given data. Fortunately, our dataset is already in good, working condition to begin with, upon careful study and exploration. The only things needed to be done are to deal with categorical values and to split our datasets.

2.2.1 Preliminary Scan of Variables

We examined the 24 independent variables from our dataset and agreed upon removing a few of the features, that by intuition, do not seem to bare any significance towards the price of an automobile. Firstly, `symboling` denoted by x_1, as mentioned earlier is one variable to be dropped.

The `make` variable denoted by x_2, has 22 levels that only has a few observations per level. This low number made us decide to remove this column. A similar argument goes for x_6 (the `carbody`), x_{14} (the `engineType`), x_{15} (the `cylinderNumber`), and the x_{17} (the `fuelSystem`).

`FuelType`, or x_3, is also another variable that we think we believe does not affect car prices. [5] A study on fuel economy has claimed that fuel does not seem to have correlation towards auto prices. Hence, we can drop this too.

[6] On EagleRidge, it has discussed the different pros and cons to using turbo vs a standard `aspiration` engine. As both sides seem to have almost equal number of pros and cons anyway, then our team has made a decision to take this variable out, denoted x_4.

`EngineLocation`, or x_8, only has two possible values: "front" and "rear". Unfortunately, we thought of this as a "biased" variable as only 1% of the datasets (3 in fact) have the "rear" value. This can indicate a skewed dataset towards the "front" value; and should be enough justification to be a variable removed from the dataset.

The `boreRatio` denoted by x_{18} and the `strokeRatio` denoted by x_{19} (often referred together as the *bore-stroke ratio*) are also variables deemed to be dropped. Upon careful research, [7] Qin et al. have shown that the bore-stroke ratio in engines can affect its combustion performance. And while combustion performance can also impact a car's fuel efficiency, this in turn also suggests that fuel economy and fuel price are also effected by this. Unfortunately, [5] a U.S. agency called the Energy Information Administration has reported that there does not seem to suggest a strong correlation between fuel economy with auto prices. On the contrary, vehicle costs vary across vehicle types. And according to [8] Tao that transitivity in correlation does not always hold (unless if all correlations are *very close* to 1). And because fuel economy and auto price are not significantly correlated, then so does the bore and stroke ratios towards the auto price. Hence, our team decided to drop these variables.

To complete this step, x_{22} (or the `peakRPM`) was also removed. The peak RPM (or the highest revolutions per minute in a car), while it may appear to be significant towards the price of a vehicle, has a similar argument for why our team has identified to remove this feature from our dataset. [9] McMahon from AutoGuru has established that there is a direct relationship between the RPM and fuel prices. And again, as fuel prices do not manifest a strong correlation towards auto prices, then so does the feature `peakRPM` as transitivity is highly unlikely to hold. Thus, this concludes our preliminary assessment of the dataset's independent variables.

2.2.2 Categorical Variables

Categorical variables are not always easy to handle especially in a regression analysis. One problem that can be seen in qualitative data is that we cannot apply any sort of statistical calculations and therefore making it problematic to build a model, to perform analyses and to make any interpretations. One solution in dealing with categorical data is through the use of dummy variables.

For instance, x_7 which is the `wheelDrive`, can either be "four-wheel drive", "rear-wheel drive", or "front-wheel drive". We can create dummy variables $\{x_{7,4wd}, x_{7,rwd}, x_{7,fwd}\}$ that indicate "four-wheel drive", "rear-wheel drive", and "front-wheel drive" respectively. If a particular observation has an x_7 value of "four-wheel drive" for instance, then $x_{7,4wd} = 1$ and a value of 0 for both $x_{7,rwd}$ and $x_{7,fwd}$. This applies similarly for observations with an x_7 value of "rear-wheel drive" and "front-wheel drive". Altogether, x_7 is removed from the dataset and is replaced by these three new columns.

2.2.3 Data Split

Part of any machine learning task is being able to validate whether the model found has a high predictive power, that is to say that the model is able to predict well on data that it has never seen before. Therefore, it makes sense to set aside a portion of the dataset just for that sole purpose, and use the remaining to be able to train the model. Hence, our dataset ought to be split into a *training set* and a *testing set*. The method of being able to validate the model is known as *Cross-Validation*, which we'll see later after we build the model. But it is good to understand as early as now the purpose of this process in data preprocessing.

There is generally no rule of thumb to follow when doing a data split. We have applied the popular 80-20 split on our dataset with 205 observations, resulting into 164 of them being in the training set and the remaining 41 in the testing set. To ensure that the data splitting method is unbiased, we apply a simple random sampling method to our dataset. This should complete data preprocessing and in the next section, we discuss the various methods used to design the **PAST** model.

3 Methodology

3.1 Multicollinearity

Once we complete the data preprocessing step, the remaining independent variables were checked for multicollinearity. *Multicollinearity*, the strong linear relationship between two independent variables, occurs when the independent variables are interrelated or dependent on each other. They must be eliminated in order to prevent several problems from arising, including the computation of *conditional* least squares point estimates, which could lead to problems when interpreting the results of the model. We computed the **Pearson Correlation Coefficients** matrix (seen in **Appendix B.1**), as well as the VIF **Table** (shown in **Appendix B.2**) in order to determine if multicollinearity exists among our remaining independent variables. The results from the tables show that $r_{x_{23},x_{24}}$, which is the simple correlation coefficient between x_{23} and x_{24}, was 0.96871, suggesting that the `city mpg` and `highway mpg` are dependent on each other. In addition, the VIF table showed that VIF_{23} and VIF_{24} were 28.77 and 26.00 respectively, which is another indicator of multicollinearity since both values are larger than 10. Thus, it was decided that x_{23} should be dropped, as it has the largest VIF value. In addition, VIF_{13} was 18.50, which is also larger than 10, so x_{13} was dropped as well in order to avoid multicollinearity between x_{13} and any other independent variable. Finally, it was shown that $x_{7,fwd}$ is a linear combination of $x_{7,4wd}$ and $x_{7,rwd}$, so the model will be written with $x_{7,fwd}$ as the reference variable, so it will not appear in the model. After dropping these variables, the Pearson Correlation Coefficients matrix did not suggest any more multicollinearity and each VIF value was less than 10 **(Appendix B.3)**, with the mean $\overline{VIF} = 3.92$, which suggests that multicollinearity no longer exists.

Number	Model
1	$\hat{y} = \beta_0 + \beta_1 x_{7,4wd} + \beta_2 x_{7,rwd} + \beta_3 x_{11} + \beta_4 x_{20} + \beta_5 x_{21}$
2	$\hat{y} = \beta_0 + \beta_1 x_{7,4wd} + \beta_2 x_{7,rwd} + \beta_3 x_{11} + \beta_4 x_{20} + \beta_5 x_{21} + \beta_6 x_{11} x_{21}$
3	$\hat{y} = \beta_0 + \beta_1 x_{7,4wd} + \beta_2 x_{7,rwd} + \beta_3 x_{11} + \beta_4 x_{20} + \beta_5 x_{21} + \beta_6 x_{21} x_{7,rwd}$
4	$\hat{y} = \beta_0 + \beta_1 x_{7,4wd} + \beta_2 x_{7,rwd} + \beta_3 x_{11} + \beta_4 x_{20} + \beta_5 x_{21} + \beta_6 x_{21} x_{7,rwd} + \beta_7 x_{11} x_{21}$
5	$\hat{y} = \beta_0 + \beta_1 x_{7,4wd} + \beta_2 x_{7,rwd} + \beta_3 x_{11} + \beta_4 x_{21} + \beta_5 x_{21} x_{7,rwd}$
6	$\hat{y} = \beta_0 + \beta_1 x_{7,4wd} + \beta_2 x_{7,rwd} + \beta_3 x_{11} + \beta_4 x_{20} + \beta_5 x_{21} + \beta_6 x_{21} x_{7,rwd} + \beta_7 x_{11} x_{21} + \beta_8 x_{16}$
7	$\hat{y} = \beta_0 + \beta_1 x_{7,4wd} + \beta_2 x_{7,rwd} + \beta_3 x_{11} + \beta_4 x_{20} + \beta_5 x_{21} + \beta_6 x_{21} x_{7,rwd} + \beta_7 x_{11} x_{21} + \beta_8 x_{16} + \beta_9 x_{16} x_{21}$

Table 1: These are the seven "best" models that we are using for comparison.

3.2 Stepwise Regression

After a quick examination on the multicollinearity between independent variables, we perform a preliminary screening of our predictor variables, to test each one's significance towards our target variable. This task can be done in a multiple ways; [10] Frost claims that some of the most popular ones include the *Best-Subsets Selection* and the *Stepwise Regression*. The Best-Subsets Selection method tests on all possible [combination of] models; thus is unbiased. Meanwhile, [11, p.58-59] Hastie et al. have argued that Stepwise Regression is a more greedy algorithm that looks for the best model by adding in significant and taking out insignificant terms to the model one step at a time. This tends to be more biased as it does not inspect all possible models. However, due to this algorithm's constrained statistical search, it has lead to become a method that computes for low variance, while maintaining efficient computational performance. [10] This results into a choice that is better, more simple for data analysts, as suggested by Frost. And this will be the prime method of use for screening which among the features have high (or have some) significance towards the target.

There will be two passes of Stepwise Regressions applied to our dataset, both passes of which can be seen in **Appendix B.4** and in **Appendix B.5**. The first step will be an initial stepwise regression to filter out any variable that has weak contributions towards the dependent variable `price`. After the first pass, we end up with the following predictors: x_7 or the `wheelDrive` (which includes the two dummy variables $x_{7,4wd}$ and $x_{7,rwd}$), x_{11} denoting the `carWidth`, x_{16} or the `engineSize`, x_{20} which is the `compressionRatio` and the x_{21} denoting the `horsepower`. Right off the bat, these variables do sound like they have some bearing towards the price of a vehicle.

The next step in building the model is through what is known as the *Interaction and Higher Order Terms*. These are additional terms we will be adding to the model to improve the its performance. A second pass of the Stepwise Regression will take out any of those unneccessary interaction variables, which we will examine in the next section.

3.3 Interaction Term Integration (including Higher Order Terms)

As mentioned in the previous section, we take a look into adding interaction terms to our current model. A systematic way of doing this is to list combinations of independent variables pairwise from the list of variables given to us from the initial Stepwise Regression: $\{x_{7,4wd}, x_{7,rwd}, x_{11}, x_{16}, x_{20}, x_{21}\}$. This means that there are a total of $\binom{6}{2} = 15$ interaction terms that can be added into the model, since there are 6 available predictors to choose from and each interaction term is composed of 2 distinct variables. We perform a second pass of the Stepwise Regression as done in **Sec. 3.2**, and we are left with the following: $\{x_{7,4wd}, x_{7,rwd}, x_{11}, x_{16}, x_{20}, x_{21}, x_{11}x_{21}, x_{16}x_{21}\}$.

Usually, higher order terms can also be added to the model. For instance, by inspection of the residual plots for some independent variable, say x_1, a quadratic term x_1^2 may be needed to validate the *Fundamental Assumption* of the regression model (see **Sec. 3.6**) if points on this plot happen to form a parabolic-like pattern. None of the independent variables at hand seemed to have a residual plot that resembled the pattern to that of a parabola, as discussed in **Sec. 3.6**. Hence, it does not make sense to include higher order terms to the model.

As a result of all the preliminary steps we took, we are able to produce 7 different models for comparison. A summary of those 7 "best" models can be seen in **Table 1**.

3.4 Model Comparison

We want to use the independent variables and interaction terms that we have in order to build a model with the best balance of precision and simplicity. Seven models with different combinations of independent variables were compared on the basis of their R^2, $\bar{R}^2$ (adjusted R^2), C_k, s (root mean square error) and $PRESS$ statistic.

The correlation coefficient of determination, R^2, is the proportion of the variation explained by the regression model over the total variation. The R^2 is often not the best indicator of the goodness of fit, as the addition of any independent variable to the model, even an unimportant one, will increase the R^2 value. A better indicator is the adjusted R^2 value, $\bar{R}^2$, as $\bar{R}^2$ can be compared across models with a different number of parameters. We say that a model with a larger $\bar{R}^2$ value is better. Additionally, a model with a small root mean square error (s) is desirable, as well as a model with a small C_k statistic that is approximately equal to the number of parameters. Finally, a model with a small $PRESS$ statistic, which the sum of squared deleted residu-

als, is also more desirable. The summary table can be found below:

Model	k	C_k	R^2	$\bar{R}^2$	s	$PRESS$
1	6	6	0.7637	0.7563	3960.65	2,787,102,686
2	7	7	0.7646	0.7556	3966.35	3,841,614,615
3	7	7	0.7795	0.7710	3838.75	2,782,421,119
4	7	7	0.7820	0.7722	3829.10	3,547,983,422
5	6	6	0.7728	0.7657	3883.58	2,817,907,210
6	9	9	0.8373	0.8289	3318.50	2,271,013,952
7	10	10	0.8622	0.8542	3063.75	1,828,314,522

After comparing all the statistics, it is evident that model 7 has the lowest root mean square error value (s), the lowest $PRESS$ statistic, the highest adjusted R square value, $\bar{R}^2$, and has a C_k value that is equal to its number of parameters (k). All of these outcomes suggest that model 7 is the superior model when compared to the other 6. Thus, we will continue using model 7 for the next tests.

3.5 Model Improvement

3.5.1 Outliers and Influential Points

In order to improve the model, various testing was done on the training set in order to determine which observations contained outliers and/or influential points, which can be seen in **Appendix B.6**. Outliers with respect to the independent variables x are often unimportant, while outliers with respect to the dependent variable y are often important. This is because the regression model will be pulled towards y_i when the i^{th} observation, given that it is an outlier with respect to y, is used to calculate the model. This in turn will give an inaccurate regression line; thus, these outlying points can be removed in order to build a more representative model. Although many observations were found to be outliers with respect to x and y, only the most serious ones will be included for the purposes of this report. These include the observations that are not only outliers with respect to x and y, but also pass other tests, indicating that they may be appropriate for removal. See the results in **Table 2**.

From the table, it is evident that the observations of particular interest are $i = 14, 16, 89, 109$. Although observation 109 is an outlier with respect to x and y, is an influential point and has a large Difference of Fits, it is determined that dropping the observation would not be beneficial. This is because dropping this particular observation would substantially change the least square point estimates of x_{21} and $x_{11}x_{21}$ and would damage the precision of the model. Thus, it was decided to keep observation 109 in the training set. However, observations 14, 16 and 89 were all outliers with respect to x and y and dropping these observations would enhance the precision of the model. Hence, it was decided to drop these three observations from the training set. After dropping these observations, the new statistics were as follows:

C_k	k	R^2	R^2	s
10	10	0.8603	0.8520	3057.83921

The R^2 and $\bar{R}^2$ values did not change substantially (a 0.0019 and 0.0022 decrease respectively); however, the s value was reduced by 5.91 and thus, it can be said that the precision of the model has increased after dropping those three observations.

3.5.2 Further Hypothesis Testing

Further hypothesis testing was conducted in order to verify that the independent variables in the model are important and to see if any should be dropped in order to simplify the model, while maintaining accuracy.

Test 1: F-Test for Overall Model
The F-test for the overall model was used in order to determine whether at least some of the independent variables are significant or not.

- H_0: $\beta_1 = \beta_2 = ... = \beta_9 = 0$
  (no relation between y and the independent variables, i.e. no significant independent variables in the model)
- H_a: At least one in $\{\beta_1, \beta_2, ..., \beta_9\}$ is non-zero
  (at least one independent variable has significant relation with y)

Test Statistic	**Description**	**Threshold**	**Applicable Observations** (i)
Leverage Point (h_{ii})	Outlier with respect to x test	$h_{ii} > 0.06097$	i=8, 60, 62, 85, 87, 92, 109
Studentized Residual ($\frac{d_i}{s_{di}}$)	Outlier with respect to y test	$\lvert\frac{d_i}{s_{di}}\rvert$>1.97559	i=14, 16, 60, 62, 85, 87, 89, 92, 109
Cook's Distance (D_i)	Influential point test	D_i>0.938263	i=109
Difference of Betas ($g_j^{(i)}/s_{g_j}^{(i)}$)	a test for whether or not removing observation i will substantially change the parameter estimates	$\lvert\frac{f_i}{s_{d_i}}\rvert$>2	i=109 for x_{21} and $x_{11}x_{21}$
Difference in Fits Statistic (f_i/s_{d_i})	difference between the point predictions of y_i made with and without using the ith observation	$\lvert\frac{f_i}{s_{d_i}}\rvert$>2	i=109
(a) Covariance Ratio (CVR_i)	removing obs i enhances model precision	CVR_i<0.817	i=8, 14, 16, 89
(b) Covariance Ratio (CVR_i)	removing obs i damages model precision	$CVR_i > 1.1829$	i=109

Table 2: Outlying and Influential Observations

After conducting this test (the results shown in **Appendix B.7**), the p-value was found to be < 0.0001, which is less than our significance level of $\alpha = 0.05$. Thus, we reject H_0, meaning that there is not enough evidence to suggest that none of our independent variables are significant.

Test 2: Hypothesis Test for Parameters
Now that it has been confirmed that at least some independent variables are significant to the model, hypothesis testing using the t-test for the parameter estimates was used in order to determine which specific independent variables in the model are significant.

- H_0: $\beta_i = 0$, for $i = 1, 2, ..., 9$
- H_a: $\beta_i \neq 0$, for $i = 1, 2, ..., 9$

From the output of **Appendix B.8**, it is evident that all variables have a p-value of less than our $\alpha = 0.05$ significance level, which means that we reject H_0 and therefore each $\beta_i \neq 0$, meaning that each independent variable has a significant relation to the y and should be kept in the model. The only exception to this was $x_{21}x_{7,rwd}$, which has a p-value of 0.8117. This means that we do not have enough evidence to reject H_0: $\beta_4 \neq 0$ (Note: β_4 is the parameter associated with $x_{21}x_{7,rwd}$), meaning that $x_{21}x_{7,rwd}$ is insignificant and can be dropped from the model.

Test 3: Partial F-Test
Using the previous tests, we have confirmed that $x_{7,4wd}, x_{7,rwd}, x_{11}, x_{20}, x_{21}, x_{11}x_{21}, x_{16}$ and $x_{16}x_{21}$ are all significant independent variables to the model. Our team questioned whether the type of wheel drive really influences car price and thus, whether it would be significant to the model. In order to determine this, a partial F-test was used where $x_{7,4wd}$ and $x_{7,rwd}$ were dropped, in order to see the effect this action had.

- H_0: $\beta_1 = \beta_2 = 0$
  (the independent variables to be dropped, $x_{7,4wd}$ and $x_{7,rwd}$ are not significant to y)
- H_a: At least one of β_1, β_2 is non-zero
  (at least one of the independent variables to be dropped, $x_{7,4wd}$ and $x_{7,rwd}$ are significant to y)

After conducting this test (results shown below), the p-value was found to be 0.0408, which is less than our significance level of $\alpha = 0.05$. Thus, we reject H_0, meaning that there is not enough evidence to suggest that the type of wheel drive is insignificant to y and thus, $x_{7,4wd}$ and $x_{7,rwd}$ are kept in the model.

Partial F-Testing - drop x74wd and x7rwd?

The REG Procedure
Model: MODEL1

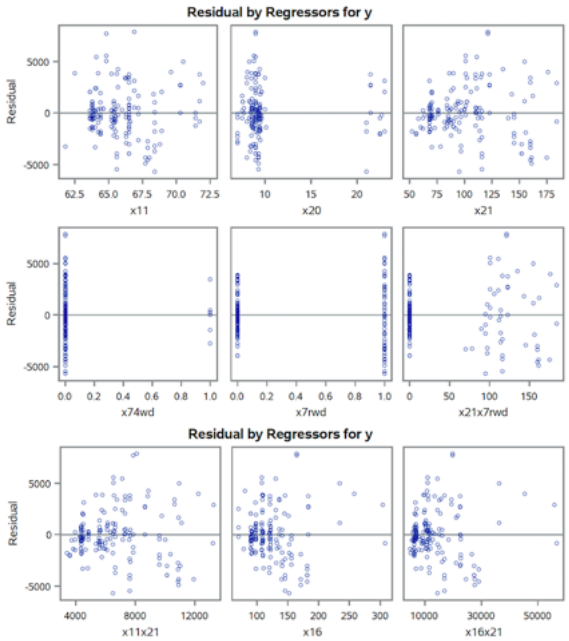
Test pft Results for Dependent Variable y

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	85562895	9.21	0.0002
Denominator	152	9292368		

3.6 Model Assumptions

In order to determine if our model is appropriate, we must confirm the fundamental assumption (A0), as well as the three inference assumptions (A1, A2 and A3).

A0: The Fundamental Assumption and A1: The Constant Variance Inference Assumption
A0, the assumption of correct functional form, states that the straight line equation $\mu = \beta_0 + \beta_1 x_{7,4wd} + \beta_2 x_{7,rwd} + \beta_3 x_{11} + \beta_4 x_{20} + \beta_5 x_{21} + \beta_6 x_{11}x_{21} + \beta_7 x_{16} + \beta_8 x_{16}x_{21}$ appropriately relates μ to the independent variables being utilized. A1, the constant variance inference assumption, states that the populations of possible values of y_i based on each x_i has the same variance. In order to check if these inference assumptions hold, we must check the residual plots against each independent variable, which can be seen below:



It is evident that each residual plot displays a random pattern with no obvious or serious trend, which means that both A0 and A1 hold.

A2: The Independence Inference Assumption
A2 states that any one value of the dependent variable y is statistically independent of any other value of y. To check if this assumption holds, we can use the Durbin-Watson Hypothesis test to check for autocorrelation. The presence of autocorrelation would indicate that A2 does not hold:

- H_0: Error terms are not autocorrelated.
- H_a: Error terms are positively or negatively autocorrelated.

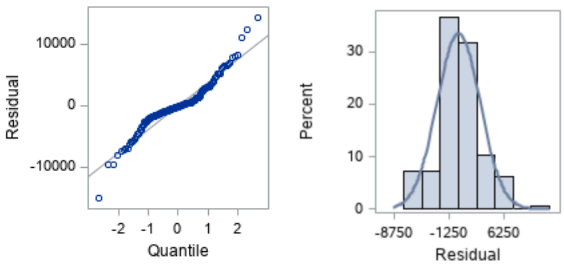
The results are shown below:

Durbin-Watson D	0.900
Pr < DW	<.0001
Pr > DW	1.0000
Number of Observations	164
1st Order Autocorrelation	0.548

It can be seen that the p-value for positive correlation is < 0.0001, which is less than our significance level of $\alpha = 0.05$, meaning that we reject H_0 and that positive autocorrelation does exist. This is a violation of A2; however, this violation can be explained. Although it is not explicit, the data collected is dependent on time, as the prices of cars in this dataset were collected within the same time frame. However, as previously discussed, the prices of new cars have been rapidly increasing over the years and if this trend continues, the prices of cars years from now will be very different than the prices of cars that appear in this dataset. What this means for our model is that although it has good predictive power for now, the predictability will diminish as time progresses.

A3: The Normality Inference Assumption

A3 states that for any independent variable x_i, the corresponding population of potential values of the dependent variable is normally distributed. To check if this assumption holds, we look at the histogram and quantile vs residual plot, as seen below:



It is evident that A3 holds, as the histogram shows a normal distribution and the points of the quantile vs residual plot are clustered towards the 45 degree line.

3.7 Final Model and Interpretation

After extensive testing, the best model was found to be:

$$\hat{y} = \beta_0 + \beta_1 x_{7,4wd} + \beta_2 x_{7,rwd} + \beta_3 x_{11} + \beta_4 x_{20} + \beta_5 x_{21} + \beta_6 x_{11} x_{21} + \beta_7 x_{16} + \beta_8 x_{16} x_{21}$$

Using the parameter estimates as shown in **Appendix B.9**, the model is:

$$\hat{y} = -170238 + 2680.67 x_{7,4wd} + 2634.04 x_{7,rwd} + 2886.72 x_{11} + 177.98 x_{20} + 918.89 x_{21} - 15.98 x_{11} x_{21} - 126.73 x_{16} + 1.47 x_{16} x_{21}$$

The model can be interpreted as follows. The mean value of car prices when all independent variables take a value of 0 is $-\$170,238$. The value is negative, which indicates an impossibility: it is impossible to have a car whose features all have a value of zero. All other variables remaining fixed,

- A one unit increase in car width (x_{11}) will increase the mean of the car price by \$2,886.72.
- A one unit increase in compression ratio (x_{20}) will increase the mean of the car price by \$177.98.
- A one unit increase in horsepower (x_{21}) will increase the mean of the car price by \$918.89.
- A one unit increase in engine size (x_{16}) will increase the mean of the car price by \$1.47.
- Since the reference type of drive wheel is front wheel drive, if a car is four wheel drive ($x_{7,4wd}$), it will increase the mean of the car price by \$2,680.67, whereas if the type of drive wheel is rear wheel drive ($x_{7,rwd}$), it will increase the mean of the car price by \$2,634.04.
- The interaction term $x_{11}x_{21}$ shows that the effect on price by car width (x_{11}) depends on the horsepower (x_{21}) and the interaction term $x_{16}x_{21}$ shows that the effect on price by engine size (x_{16}) depends on the horsepower (x_{21}).

In addition, the final statistics for the **PAST** model is:

C_k	k	R^2	R^2	s
9	9	0.8603	0.8529	3048.34

4 Testing and Cross Validation

Cross-Validation is the process of validating whether our model has high predictive power especially on newly, unseen data. This is why we initially set aside 20% of the original dataset for testing. Here we would like to produce 95% prediction intervals for each observation in the testing set. Then, we determine how many of those observations in the testing set have an actual y_i auto price value that falls in their respective prediction interval. It can be seen in **Appendix B.10** the prediction interval for each of the observations in the testing set, then out of those 41, we have discovered that 39 of them has a y_i that lies within their prediction interval. This is $39/41 \approx 95.12\%$, which is a good indication that our **PAST** model has a high predictive performance. For each observation in the test set, it can be interpreted that the respective price of the car cannot be greater than the upper bound of the prediction interval.

5 Conclusion

Our team was successful in designing a linear regression model, called **PAST**, that accurately predicts auto prices given a set of features. In building our model, we have applied various methods such as Stepwise Regression and Interaction Term Integration. We have further improved the design of our model by removing any outliers in the dataset and by applying further hypothesis tests to distinguish significantly-possessing terms vs those that have little to no correlation towards the price of automobiles. Model assumptions have also been checked to verify whether the model on the data is reliable and if it is valid. Finally, cross-validation was applied on the testing set using this **PAST** model to validate its high predictive power.

We can extend our study on the **PAST** model by exploring other models that use datasets that also attempt to predict the cost or the price of a particular commodity given some attributes, for instance predicting the price of a computer given predictors such as `relativePerformanceMeasure`, `monitorSize`, `cpuClockCycles` among others.

References

[1] M. Brenan. "83% of U.S. Adults Drive Frequently; Fewer Enjoy It a Lot." Internet: https://news.gallup.com/poll/236813/adulAccessed:tAccessed:s-drive-frequently-fewer-enjoy-lot.aspx, July 9, 2018 [Accessed: June 16, 2020].

[2] A. Hecht. "Car prices are increasing—here's how that can hurt Americans." Internet: https://www.cnbc.com/2019/10/22/car-prices-are-rapidly-increasing-heres-why-thats-bad-for-americans.html, Oct. 22, 2019 [Accessed: June 16, 2020].

[3] "Average New-Car Prices Rise 2 Percent Year-Over-Year According to Kelley Blue Book." Internet: https://mediaroom.kbb.com/2018-10-02-Average-New-Car-Prices-Rise-2-Percent-Year-Over-Year-According-to-Kelley-Blue-Book, Oct. 2, 2018 [Accessed: June 16, 2020].

[4] T. Thunder. *Auto Data Car Price Prediction Regression*, vol. 1, Kaggle, 2020. [Online]. Available: https://www.kaggle.com/thorgodofthunder/auto-data-car-price-prediction-regression. [Accessed: June 12, 2020].

[5] "Fuel economy and average vehicle cost vary significantly across vehicle types." Internet: https://www.eia.gov/todayinenergy/detail.php?id=17211, July 22, 2014 [Accessed: June 16, 2020].

[6] "Tubrocharged vs Naturally Aspirated Engines." Internet: https://www.eagleridgegm.com/turbocharged-vs-naturally-aspirated-engines/, Feb. 18, 2017 [Accessed: June 16, 2020].

[7] X. Qin, F. Ntone, L. LaPointe, E. Lyford-Pike, "The Effect of Stroke-to-Bore Ratio on Combustion Performance of a Lean Burn Heavy-Duty Gaseous SI Engine," in *ASME 2010 Internal Combustion Engine Division Fall Technical Conference*, San Antonio, Texas, USA, 2010, pp.801-810. Accessed on: June 16, 2020. [Online]. Available: https://doi.org/10.1115/ICEF2010-35108

[8] T. Tao. "When is correlation transitive?" Internet: https://terrytao.wordpress.com/2014/06/05/when-is-correlation-transitive/, June 5, 2014 [Accessed: June 16, 2020].

[9] B. McMahon. "What does RPM mean in a car?" Internet: https://www.autoguru.com.au/car-advice/articles/what-does-rpm-mean-in-a-car, Oct. 11, 2019 [Accessed: June 16, 2020].

[10] J. Frost. "Guide to Stepwise Regression and Best Subsets Regression." Internet: https://statisticsbyjim.com/regression/guide-stepwise-best-subsets-regression/, n.d. [Accessed: June 16, 2020].

[11] T. Hastie, R. Tibshirani, J. Friedman *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY: Springer, 2017.

Appendix A: Dataset and Dictionary

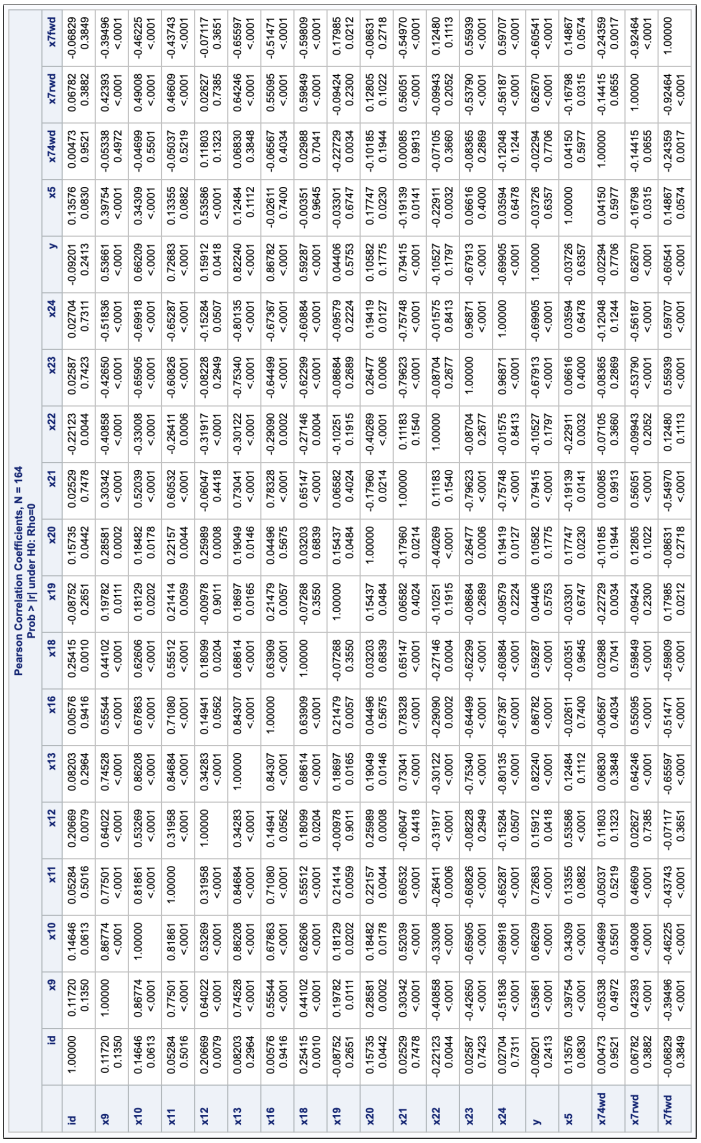
Variable	Feature/Target	Type	Range of Values
x_1	symboling	numerical discrete	$\{-2, -1, 0, 1, 2, 3\}$
x_2	make	categorical	{alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo}
x_3	fueltype	categorical	{diesel, gas}
x_4	aspiration	categorical	{turbo, std}
x_5	doornumber	numerical discrete	$\{2, 4\}$
x_6	carbody	categorical	{convertible, hardtop, hatchback, sedan, wagon}
x_7	drivewheel	categorical	{fwd, rwd, 4wd}
x_8	enginelocation	categorical	{front, rear}
x_9	wheelbase	numerical continuous	$[86.6, 120.9]$
x_{10}	carlength	numerical continuous	$[141.1, 208.1]$
x_{11}	carwidth	numerical continuous	$[60.3, 72.3]$
x_{12}	carheight	numerical continuous	$[47.8, 59.8]$
x_{13}	curbweight	numerical continuous	$[1488, 4066]$
x_{14}	enginetype	categorical	{dohc, dohcv, ohc, ohcf, ohcv, rotor}
x_{15}	cylindernumber	numerical discrete	$\{2, 3, 4, 5, 6, 8, 12\}$
x_{16}	enginesize	numerical continuous	$[61, 326]$
x_{17}	fuelsystem	categorical	{1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi}
x_{18}	boreratio	numerical continuous	$[2.54, 3.94]$
x_{19}	stroke	numerical continuous	$[2.07, 4.17]$
x_{20}	compressionratio	numerical continuous	$[7, 23]$
x_{21}	horsepower	numerical continuous	$[48, 288]$
x_{22}	peakrpm	numerical continuous	$[4150, 6600]$
x_{23}	citympg	numerical continuous	$[13, 49]$
x_{24}	highwaympg	numerical continuous	$[16, 54]$
y	price	numerical continuous	$[5118, 45400]$

Auto Price Dataset

Obs	id	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18	x19	x20	x21	x22	x23	x24	y
1	1	3	alfa-rom	gas	std	two	converti	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130	mpfi	3.47	2.680	9.00	111	5000	21	27	13495.00
2	2	3	alfa-rom	gas	std	two	converti	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130	mpfi	3.47	2.680	9.00	111	5000	21	27	16500.00
3	3	1	alfa-rom	gas	std	two	hatchbac	rwd	front	94.5	171.2	65.5	52.4	2823	ohcv	six	152	mpfi	2.68	3.470	9.00	154	5000	19	26	16500.00
4	4	2	audi 100	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3	2337	ohc	four	109	mpfi	3.19	3.400	10.00	102	5500	24	30	13950.00
5	5	2	audi 100	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3	2824	ohc	five	136	mpfi	3.19	3.400	8.00	115	5500	18	22	17450.00
6	6	2	audi fox	gas	std	two	sedan	fwd	front	99.8	177.3	66.3	53.1	2507	ohc	five	136	mpfi	3.19	3.400	8.50	110	5500	19	25	15250.00
7	7	1	audi 100	gas	std	four	sedan	fwd	front	105.8	192.7	71.4	55.7	2844	ohc	five	136	mpfi	3.19	3.400	8.50	110	5500	19	25	17710.00
8	8	1	audi 500	gas	std	four	wagon	fwd	front	105.8	192.7	71.4	55.7	2954	ohc	five	136	mpfi	3.19	3.400	8.50	110	5500	19	25	18920.00
9	9	1	audi 400	gas	turbo	four	sedan	fwd	front	105.8	192.7	71.4	55.9	3086	ohc	five	131	mpfi	3.13	3.400	8.30	140	5500	17	20	23875.00
10	10	0	audi 500	gas	turbo	two	hatchbac	4wd	front	99.5	178.2	67.9	52.0	3053	ohc	five	131	mpfi	3.13	3.400	7.00	160	5500	16	22	17859.17
11	11	2	bmw 320i	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc	four	108	mpfi	3.50	2.800	8.80	101	5800	23	29	16430.00
12	12	0	bmw 320i	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc	four	108	mpfi	3.50	2.800	8.80	101	5800	23	29	16925.00
13	13	0	bmw x1	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2710	ohc	six	164	mpfi	3.31	3.190	9.00	121	4250	21	28	20970.00
14	14	0	bmw x3	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2765	ohc	six	164	mpfi	3.31	3.190	9.00	121	4250	21	28	21105.00
15	15	1	bmw z4	gas	std	four	sedan	rwd	front	103.5	189.0	66.9	55.7	3055	ohc	six	164	mpfi	3.31	3.190	9.00	121	4250	20	25	24565.00
16	16	0	bmw x4	gas	std	four	sedan	rwd	front	103.5	189.0	66.9	55.7	3230	ohc	six	209	mpfi	3.62	3.390	8.00	182	5400	16	22	30760.00
17	17	0	bmw x5	gas	std	two	sedan	rwd	front	103.5	193.8	67.9	53.7	3380	ohc	six	209	mpfi	3.62	3.390	8.00	182	5400	16	22	41315.00
18	18	0	bmw x3	gas	std	four	sedan	rwd	front	110.0	197.0	70.9	56.3	3505	ohc	six	209	mpfi	3.62	3.390	8.00	182	5400	15	20	36880.00
19	19	2	chevrole	gas	std	two	hatchbac	fwd	front	88.4	141.1	60.3	53.2	1488	l	three	61	2bbl	2.91	3.030	9.50	48	5100	47	53	5151.00
20	20	1	chevrole	gas	std	two	hatchbac	fwd	front	94.5	155.9	63.6	52.0	1874	ohc	four	90	2bbl	3.03	3.110	9.60	70	5400	38	43	6295.00

Dataset contains 205 observations; the first 20 rows shown above. It has 24 predictors and one target variable `price`

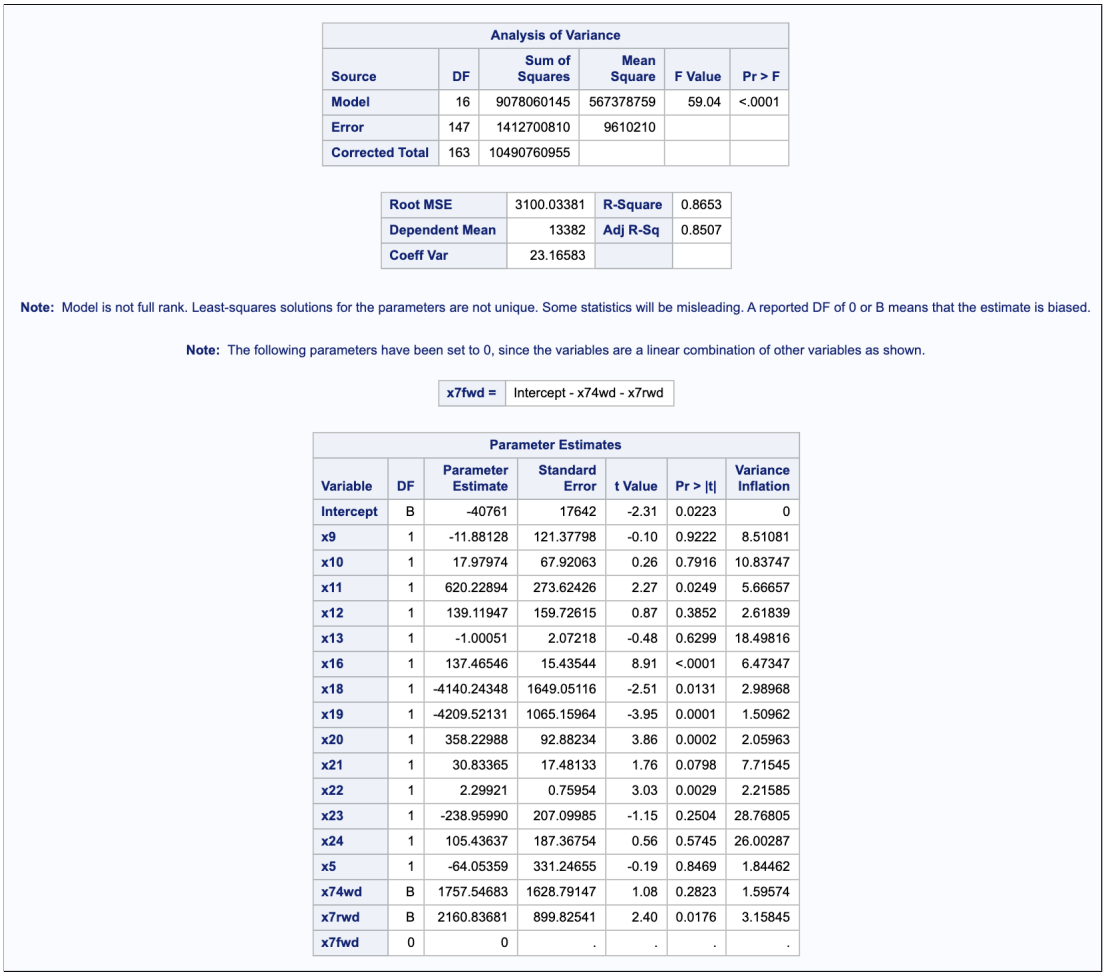
Appendix B: SAS Outputs



Pearson Correlation Coefficients, N = 164
Prob > |r| under H0: Rho=0

	id	x9	x10	x11	x12	x13	x16	x18	x19	x20	x21	x22	x23	x24	y	x5	x74wd	x7rwd	x7fwd
id	1.00000	0.11720 0.1350	0.14646 0.0613	0.05284 0.5016	0.20669 0.0079	0.08203 0.2964	0.00576 0.9416	0.25415 0.0010	-0.08752 0.2651	0.15735 0.0442	0.02529 0.7478	-0.22123 0.0044	0.02587 0.7423	0.02704 0.7311	-0.09201 0.2413	0.13576 0.0830	0.00473 0.9521	0.06782 0.3882	-0.06829 0.3849
x9	0.11720 0.1350	1.00000	0.86774 <.0001	0.77501 <.0001	0.84022 <.0001	0.74528 <.0001	0.55544 <.0001	0.44102 <.0001	0.19782 0.0111	0.28581 0.0002	0.30342 <.0001	-0.40858 <.0001	-0.42650 <.0001	-0.51836 <.0001	0.53661 <.0001	0.39754 <.0001	-0.05338 0.4972	0.42393 <.0001	-0.39496 <.0001
x10	0.14646 0.0613	0.86774 <.0001	1.00000	0.81861 <.0001	0.53269 <.0001	0.86208 <.0001	0.67863 <.0001	0.62606 <.0001	0.18129 0.0202	0.18482 0.0178	0.52039 <.0001	-0.33008 <.0001	-0.65905 <.0001	-0.69918 <.0001	0.66209 <.0001	0.34309 <.0001	-0.04699 0.5501	0.49008 <.0001	-0.46225 <.0001
x11	0.05284 0.5016	0.77501 <.0001	0.81861 <.0001	1.00000	0.31958 <.0001	0.84684 <.0001	0.71080 <.0001	0.55512 <.0001	0.21414 0.0059	0.22157 0.0044	0.60532 <.0001	-0.26411 0.0006	-0.60826 <.0001	-0.65287 <.0001	0.72683 <.0001	0.13355 0.0882	-0.05037 0.5219	0.46609 <.0001	-0.43743 <.0001
x12	0.20669 0.0079	0.84022 <.0001	0.53269 <.0001	0.31958 <.0001	1.00000	0.34283 <.0001	0.14941 0.0562	0.18099 0.0204	-0.00978 0.9011	0.25989 0.0008	-0.06047 0.4418	-0.31917 <.0001	-0.08228 0.2949	-0.15284 0.0507	0.15912 0.0418	0.53586 <.0001	0.11803 0.1323	0.02627 0.7385	-0.07117 0.3651
x13	0.08203 0.2964	0.74528 <.0001	0.86208 <.0001	0.84684 <.0001	0.34283 <.0001	1.00000	0.84307 <.0001	0.68614 <.0001	0.18697 0.0165	0.19049 0.0146	0.73041 <.0001	-0.30122 <.0001	-0.75340 <.0001	-0.80135 <.0001	0.82240 <.0001	0.12484 0.1112	0.06830 0.3848	0.64246 <.0001	-0.65597 <.0001
x16	0.00576 0.9416	0.55544 <.0001	0.67863 <.0001	0.71080 <.0001	0.14941 0.0562	0.84307 <.0001	1.00000	0.63909 <.0001	0.21479 0.0057	0.04496 0.5675	0.78328 <.0001	-0.29090 0.0002	-0.64499 <.0001	-0.67367 <.0001	0.86782 <.0001	-0.02611 0.7400	-0.06567 0.4034	0.55095 <.0001	-0.51471 <.0001
x18	0.25415 0.0010	0.44102 <.0001	0.62606 <.0001	0.55512 <.0001	0.18099 0.0204	0.68614 <.0001	0.63909 <.0001	1.00000	-0.07268 0.3550	0.03203 0.6839	0.65147 <.0001	-0.27146 0.0004	-0.62299 <.0001	-0.60884 <.0001	0.59287 <.0001	-0.00351 0.9645	0.02988 0.7041	0.59849 <.0001	-0.59809 <.0001
x19	-0.08752 0.2651	0.19782 0.0111	0.18129 0.0202	0.21414 0.0059	-0.00978 0.9011	0.18697 0.0165	0.21479 0.0057	-0.07268 0.3550	1.00000	0.15437 0.0484	0.06582 0.4024	-0.10251 0.1915	-0.08684 0.2689	-0.09579 0.2224	0.04406 0.5753	-0.03301 0.6747	-0.22729 0.0034	-0.09424 0.2300	0.17985 0.0212
x20	0.15735 0.0442	0.28581 0.0002	0.18482 0.0178	0.22157 0.0044	0.25989 0.0008	0.19049 0.0146	0.04496 0.5675	0.03203 0.6839	0.15437 0.0484	1.00000	-0.17960 0.0214	-0.40269 <.0001	-0.64477 0.0006	0.19419 0.0127	0.10582 0.1775	0.17747 0.0230	-0.10185 0.1944	0.12805 0.1022	-0.08631 0.2718
x21	0.02529 0.7478	0.30342 <.0001	0.52039 <.0001	0.60532 <.0001	-0.06047 0.4418	0.73041 <.0001	0.78328 <.0001	0.65147 <.0001	0.06582 0.4024	-0.17960 0.0214	1.00000	0.11183 0.1540	-0.79623 <.0001	-0.75748 <.0001	0.79415 <.0001	-0.19139 0.0141	0.00085 0.9913	0.56051 <.0001	-0.54970 <.0001
x22	-0.22123 0.0044	-0.40858 <.0001	-0.33008 <.0001	-0.26411 0.0006	-0.31917 <.0001	-0.30122 <.0001	-0.29090 0.0002	-0.27146 0.0004	-0.10251 0.1915	-0.40269 <.0001	0.11183 0.1540	1.00000	-0.08704 0.2677	-0.01575 0.8413	-0.10527 0.1797	-0.22911 0.0032	-0.07105 0.3660	-0.09943 0.2052	0.12480 0.1113
x23	0.02587 0.7423	-0.42650 <.0001	-0.65905 <.0001	-0.60826 <.0001	-0.08228 0.2949	-0.75340 <.0001	-0.64499 <.0001	-0.62299 <.0001	-0.08684 0.2689	-0.26477 0.0006	-0.79623 <.0001	-0.08704 0.2677	1.00000	0.96871 <.0001	-0.67913 <.0001	0.06616 0.4000	-0.08365 0.2869	-0.53790 <.0001	0.55939 <.0001
x24	0.02704 0.7311	-0.51836 <.0001	-0.69918 <.0001	-0.65287 <.0001	-0.15284 0.0507	-0.80135 <.0001	-0.67367 <.0001	-0.60884 <.0001	-0.09579 0.2224	0.19419 0.0127	-0.75748 <.0001	-0.01575 0.8413	0.96871 <.0001	1.00000	-0.69905 <.0001	0.03594 0.6478	-0.12048 0.1244	-0.56187 <.0001	0.59707 <.0001
y	-0.09201 0.2413	0.53661 <.0001	0.66209 <.0001	0.72683 <.0001	0.15912 0.0418	0.82240 <.0001	0.86782 <.0001	0.58287 <.0001	0.04406 0.5753	0.10582 0.1775	0.79415 <.0001	-0.10527 0.1797	-0.67913 <.0001	-0.69905 <.0001	1.00000	-0.03726 0.6357	-0.02294 0.7706	0.62670 <.0001	-0.60541 <.0001
x5	0.13576 0.0830	0.39754 <.0001	0.34309 <.0001	0.13355 0.0882	0.53586 <.0001	0.12484 0.1112	-0.02611 0.7400	-0.00351 0.9645	-0.03301 0.6747	0.17747 0.0230	-0.19139 0.0141	-0.22911 0.0032	0.06616 0.4000	0.03594 0.6478	-0.03726 0.6357	1.00000	0.04150 0.5977	-0.16798 0.0315	0.14867 0.0574
x74wd	0.00473 0.9521	-0.05338 0.4972	-0.04699 0.5501	-0.05037 0.5219	0.11803 0.1323	0.06830 0.3848	-0.06567 0.4034	0.02988 0.7041	-0.22729 0.0034	-0.10185 0.1944	0.00085 0.9913	-0.07105 0.3660	-0.08365 0.2869	-0.12048 0.1244	-0.02294 0.7706	0.04150 0.5977	1.00000	-0.14415 0.0655	-0.24359 0.0017
x7rwd	0.06782 0.3882	0.42393 <.0001	0.49008 <.0001	0.46609 <.0001	0.02627 0.7385	0.64246 <.0001	0.55095 <.0001	0.59849 <.0001	-0.09424 0.2300	0.12805 0.1022	0.56051 <.0001	0.12480 0.1113	0.55939 <.0001	0.59707 <.0001	0.62670 <.0001	0.04150 0.5977	-0.14415 0.0655	1.00000	-0.92464 <.0001
x7fwd	-0.06829 0.3849	-0.39496 <.0001	-0.46225 <.0001	-0.43743 <.0001	-0.07117 0.3651	-0.65597 <.0001	-0.51471 <.0001	-0.59809 <.0001	0.17985 0.0212	-0.08631 0.2718	-0.54970 <.0001	0.12480 0.1113	0.55939 <.0001	0.59707 <.0001	-0.60541 <.0001	0.14867 0.0574	-0.24359 0.0017	-0.92464 <.0001	1.00000

Fig. 1: Correlation Matrix



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	9078060145	567378759	59.04	<.0001
Error	147	1412700810	9610210		
Corrected Total	163	10490760955			

Root MSE	3100.03381	R-Square	0.8653
Dependent Mean	13382	Adj R-Sq	0.8507
Coeff Var	23.16583		

Note: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

x7fwd =	Intercept - x74wd - x7rwd

Parameter Estimates

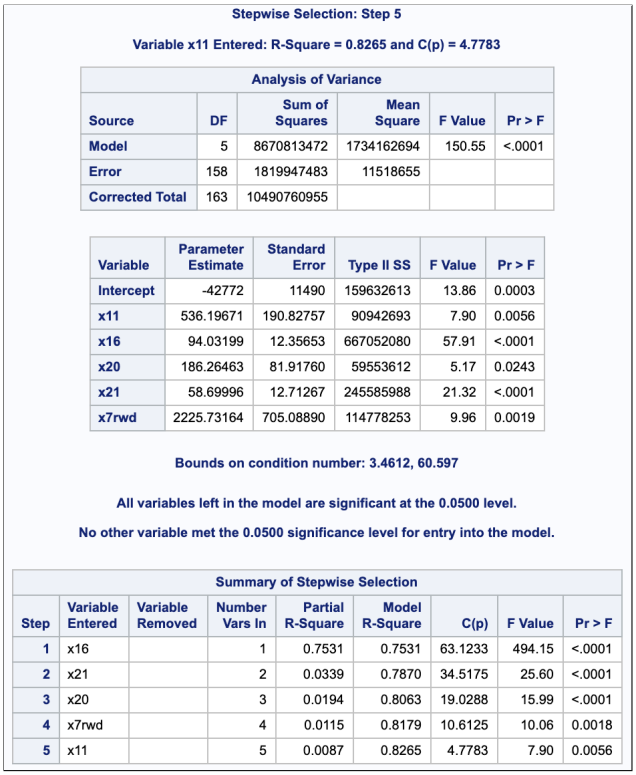
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	B	-40761	17642	-2.31	0.0223	0
x9	1	-11.88128	121.37798	-0.10	0.9222	8.51081
x10	1	17.97974	67.92063	0.26	0.7916	10.83747
x11	1	620.22894	273.62426	2.27	0.0249	5.66657
x12	1	139.11947	159.72615	0.87	0.3852	2.61839
x13	1	-1.00051	2.07218	-0.48	0.6299	18.49816
x16	1	137.46546	15.43544	8.91	<.0001	6.47347
x18	1	-4140.24348	1649.05116	-2.51	0.0131	2.98968
x19	1	-4209.52131	1065.15964	-3.95	0.0001	1.50962
x20	1	358.22988	92.88234	3.86	0.0002	2.05963
x21	1	30.83365	17.48133	1.76	0.0798	7.71545
x22	1	2.29921	0.75954	3.03	0.0029	2.21585
x23	1	-238.95990	207.09985	-1.15	0.2504	28.76805
x24	1	105.43637	187.36754	0.56	0.5745	26.00287
x5	1	-64.05359	331.24655	-0.19	0.8469	1.84462
x74wd	B	1757.54683	1628.79147	1.08	0.2823	1.59574
x7rwd	B	2160.83681	899.82541	2.40	0.0176	3.15845
x7fwd	0	0	.	.	.	.

Fig. 2: VIF Table Before Dropping Multicollinear Variables

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-44257	15916	-2.78	0.0061	0
x9	1	-90.15524	123.67704	-0.73	0.4671	7.31655
x10	1	-50.15394	64.55542	-0.78	0.4384	8.10638
x11	1	667.64102	285.74975	2.34	0.0208	5.11705
x12	1	230.83833	170.90337	1.35	0.1788	2.48211
x16	1	98.07887	13.07958	7.50	<.0001	3.84879
x20	1	205.87058	89.41857	2.30	0.0227	1.58058
x21	1	51.76525	15.61819	3.31	0.0011	5.09931
x24	1	-86.78054	85.62539	-1.01	0.3124	4.49650
x5	1	186.64274	355.06868	0.53	0.5999	1.75495
x74wd	1	1054.65618	1542.50068	0.68	0.4952	1.18500
x7rwd	1	2552.95632	809.06671	3.16	0.0019	2.11428

Fig. 3: VIF Table After Dropping Multicollinear Variables



Stepwise Selection: Step 5

Variable x11 Entered: R-Square = 0.8265 and C(p) = 4.7783

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	8670813472	1734162694	150.55	<.0001
Error	158	1819947483	11518655		
Corrected Total	163	10490760955			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-42772	11490	159632613	13.86	0.0003
x11	536.19671	190.82757	90942693	7.90	0.0056
x16	94.03199	12.35653	667052080	57.91	<.0001
x20	186.26463	81.91760	59553612	5.17	0.0243
x21	58.69996	12.71267	245585988	21.32	<.0001
x7rwd	2225.73164	705.08890	114778253	9.96	0.0019

Bounds on condition number: 3.4612, 60.597

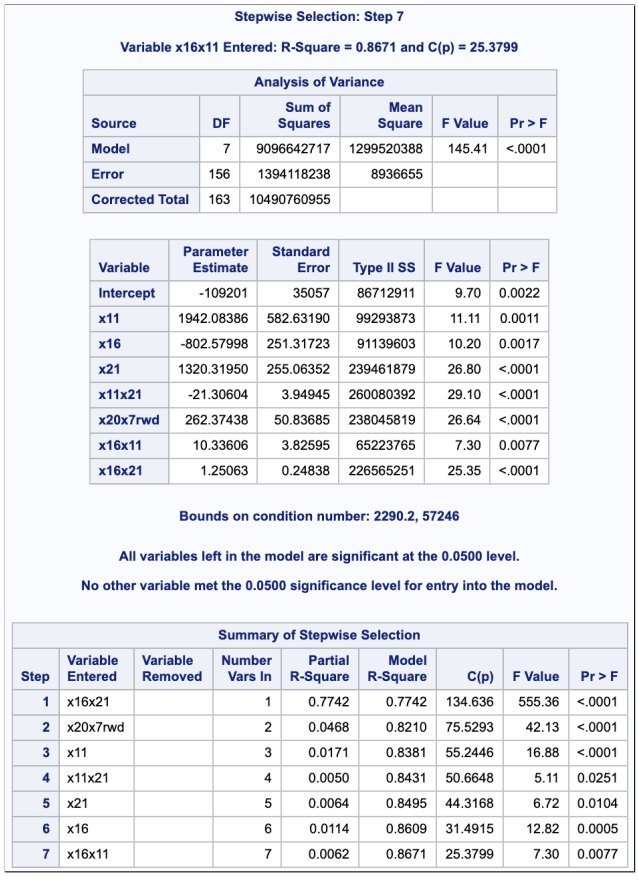
All variables left in the model are significant at the 0.0500 level.

No other variable met the 0.0500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x16		1	0.7531	0.7531	63.1233	494.15	<.0001
2	x21		2	0.0339	0.7870	34.5175	25.60	<.0001
3	x20		3	0.0194	0.8063	19.0288	15.99	<.0001
4	x7rwd		4	0.0115	0.8179	10.6125	10.06	0.0018
5	x11		5	0.0087	0.8265	4.7783	7.90	0.0056

Fig. 4: Initial Stepwise Regression Output



Stepwise Selection: Step 7

Variable x16x11 Entered: R-Square = 0.8671 and C(p) = 25.3799

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	9096642717	1299520388	145.41	<.0001
Error	156	1394118238	8936655		
Corrected Total	163	10490760955			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-109201	35057	86712911	9.70	0.0022
x11	1942.08386	582.63190	99293873	11.11	0.0011
x16	-802.57998	251.31723	91139603	10.20	0.0017
x21	1320.31950	255.06352	239461879	26.80	<.0001
x11x21	-21.30604	3.94945	260080392	29.10	<.0001
x20x7rwd	262.37438	50.83685	238045819	26.64	<.0001
x16x11	10.33606	3.82595	65223765	7.30	0.0077
x16x21	1.25063	0.24838	226565251	25.35	<.0001

Bounds on condition number: 2290.2, 57246

All variables left in the model are significant at the 0.0500 level.

No other variable met the 0.0500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x16x21		1	0.7742	0.7742	134.636	555.36	<.0001
2	x20x7rwd		2	0.0468	0.8210	75.5293	42.13	<.0001
3	x11		3	0.0171	0.8381	55.2446	16.88	<.0001
4	x11x21		4	0.0050	0.8431	50.6648	5.11	0.0251
5	x21		5	0.0064	0.8495	44.3168	6.72	0.0104
6	x16		6	0.0114	0.8609	31.4915	12.82	0.0005
7	x16x11		7	0.0062	0.8671	25.3799	7.30	0.0077

Fig. 5: Second Stepwise Regression

Checking for Outliers and Influential Points

The REG Procedure
Model: MODEL1
Dependent Variable: y

	Output Statistics											DFBETAS								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	Intercept	x11	x20	x21	x74wd	x7rwd	x21x7rwd	x11x21	x16	x16x21
1	13495	12427	660.1210	1068	2991.8	0.357	0.001	0.3559	0.0464	1.1101	0.0785	0.0370	-0.0379	-0.0004	-0.0204	-0.0032	0.0214	-0.0107	0.0227	0.0273	-0.0243
2	16500	12427	660.1210	4073	2991.8	1.361	0.009	1.3652	0.0464	0.9916	0.3012	0.1420	-0.1453	-0.0014	-0.0782	-0.0124	0.0820	-0.0409	0.0872	0.1046	-0.0933
3	13950	10966	407.7181	2984	3036.5	0.983	0.002	0.9828	0.0177	1.0203	0.1320	-0.0507	0.0496	0.0123	0.0431	-0.0289	-0.0030	-0.0147	-0.0394	-0.0351	0.0214
4	17450	15036	1277	2414	2784.9	0.867	0.016	0.8662	0.1738	1.2301	0.3973	-0.0203	0.0134	-0.0143	0.0188	0.3704	-0.0236	0.0167	-0.0138	0.0415	-0.0317
5	15250	12059	463.4649	3191	3028.5	1.054	0.003	1.0540	0.0229	1.0161	0.1613	0.0113	-0.0202	-0.0239	-0.0085	-0.0376	-0.0265	0.0018	0.0187	0.0746	-0.0670
6	17710	17466	1057	244.1218	2875.5	0.085	0.000	0.0846	0.1191	1.2111	0.0311	-0.0189	0.0186	-0.0112	0.0098	-0.0021	-0.0051	0.0031	-0.0099	-0.0060	0.0038
7	18920	17466	1057	1454	2875.5	0.506	0.003	0.5045	0.1191	1.1917	0.1855	-0.1127	0.1108	-0.0668	0.0583	-0.0125	-0.0303	0.0187	-0.0589	-0.0357	0.0229
8	23875	16430	1089	7445	2863.6	2.600	0.098	2.6502	0.1264	0.7799	1.0081	-0.0946	0.0965	-0.1256	-0.1794	-0.1577	0.1764	-0.2706	0.1851	-0.1301	0.0110
9	17859	17361	1450	497.9626	2699.1	0.184	0.001	0.1839	0.2239	1.3722	0.0988	0.0040	-0.0040	0.0093	-0.0073	0.0741	0.0265	-0.0315	0.0085	-0.0114	0.0071
10	16430	11814	631.2846	4616	2998.0	1.540	0.011	1.5467	0.0425	0.9544	0.3257	0.0508	-0.0376	-0.0231	-0.0237	-0.0113	0.2042	-0.1489	0.0175	-0.0806	0.0680
11	16925	11814	631.2846	5111	2998.0	1.705	0.013	1.7156	0.0425	0.9213	0.3613	0.0563	-0.0417	-0.0257	-0.0263	-0.0126	0.2265	-0.1651	0.0194	-0.0894	0.0755
12	20970	15772	772.2063	5198	2964.8	1.753	0.021	1.7652	0.0635	0.9316	0.4598	0.2016	-0.2217	-0.0399	-0.1103	-0.0093	-0.0169	0.0587	0.1315	0.2972	-0.2454
13	21105	15772	772.2063	5333	2964.8	1.799	0.022	1.8120	0.0635	0.9217	0.4719	0.2070	-0.2276	-0.0409	-0.1132	-0.0096	-0.0174	0.0602	0.1350	0.3051	-0.2519
14	24565	17636	599.5911	6929	3004.5	2.306	0.021	2.3393	0.0383	0.7810	0.4668	0.1071	-0.1287	-0.1479	-0.0935	-0.0047	-0.0495	0.1057	0.1112	0.2923	-0.2561
15	30760	29343	803.6357	1417	2956.5	0.479	0.002	0.4782	0.0688	1.1292	0.1300	-0.0402	0.0400	-0.0086	0.0599	0.0043	-0.0259	0.0300	-0.0603	-0.0218	0.0352
16	41315	29274	688.9524	12041	2985.3	4.033	0.087	4.2510	0.0506	0.3681	0.9811	-0.2012	0.2035	-0.1386	0.2637	0.0230	-0.2350	0.2730	-0.2743	-0.0941	0.1786
17	6295	6122	432.9755	172.5867	3033.0	0.057	0.000	0.0567	0.0200	1.0888	0.0081	0.0025	-0.0019	0.0004	-0.0019	-0.0010	0.0002	-0.0002	0.0015	-0.0014	0.0013
18	6575	6122	432.9755	452.5867	3033.0	0.149	0.000	0.1487	0.0200	1.0875	0.0212	0.0064	-0.0051	0.0011	-0.0049	-0.0026	0.0004	-0.0006	0.0039	-0.0036	0.0034
19	5572	6360	432.4537	-788.0174	3033.1	-0.260	0.000	-0.2590	0.0199	1.0842	-0.0369	-0.0065	0.0040	0.0015	0.0054	0.0040	0.0006	-0.0007	-0.0032	0.0089	-0.0085
20	6377	6358	432.4856	18.7630	3033.1	0.006	0.000	0.006166	0.0199	1.0890	0.0009	0.0002	-0.0001	-0.0000	-0.0001	-0.0001	-0.0000	0.0000	0.0001	-0.0002	0.0002

Fig. 6: Outlying and Influential Observations (First 20 rows)

Complete F-Test

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	161
Number of Observations Used	161

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	8697637432	966404159	103.35	<.0001
Error	151	1411907476	9350381		
Corrected Total	160	10109544907			

Root MSE	3057.83921	R-Square	0.8603
Dependent Mean	13252	Adj R-Sq	0.8520
Coeff Var	23.07542		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-169318	30000	-5.64	<.0001
x11	1	2876.88967	506.16133	5.68	<.0001
x20	1	181.76961	75.69650	2.40	0.0176
x21	1	913.89498	217.09589	4.21	<.0001
x74wd	1	2655.83332	1307.74064	2.03	0.0440
x7rwd	1	3231.48238	2587.91389	1.25	0.2137
x21x7rwd	1	-5.37010	22.50410	-0.24	0.8117
x11x21	1	-15.92851	3.59548	-4.43	<.0001
x16	1	-131.38162	41.71614	-3.15	0.0020
x16x21	1	1.50720	0.27730	5.44	<.0001

Fig. 7: Overall F-Test

Hypothesis Testing for b_j

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	161
Number of Observations Used	161

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	8697637432	966404159	103.35	<.0001
Error	151	1411907476	9350381		
Corrected Total	160	10109544907			

Root MSE	3057.83921	R-Square	0.8603
Dependent Mean	13252	Adj R-Sq	0.8520
Coeff Var	23.07542		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-169318	30000	-5.64	<.0001
x11	1	2876.88967	506.16133	5.68	<.0001
x20	1	181.76961	75.69650	2.40	0.0176
x21	1	913.89498	217.09589	4.21	<.0001
x74wd	1	2655.83332	1307.74064	2.03	0.0440
x7rwd	1	3231.48238	2587.91389	1.25	0.2137
x21x7rwd	1	-5.37010	22.50410	-0.24	0.8117
x11x21	1	-15.92851	3.59548	-4.43	<.0001
x16	1	-131.38162	41.71614	-3.15	0.0020
x16x21	1	1.50720	0.27730	5.44	<.0001

Fig. 8: Hypothesis Testing for b_j

Confidence Intervals for Parameters

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	161
Number of Observations Used	161

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	8697104991	1087138124	116.99	<.0001
Error	152	1412439917	9292368		
Corrected Total	160	10109544907			

Root MSE	3048.33854	R-Square	0.8603
Dependent Mean	13252	Adj R-Sq	0.8529
Coeff Var	23.00372		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-170238	29659	-5.74	<.0001	-228834	-111642
x11	1	2886.71652	502.91592	5.74	<.0001	1893.10858	3880.32445
x20	1	177.97822	73.78034	2.41	0.0170	32.21086	323.74558
x21	1	918.88711	215.41419	4.27	<.0001	493.29459	1344.47962
x74wd	1	2680.66760	1299.54277	2.06	0.0408	113.16883	5248.16636
x7rwd	1	2634.03701	652.89469	4.03	<.0001	1344.11693	3923.95708
x11x21	1	-15.97842	3.57824	-4.47	<.0001	-23.04793	-8.90891
x16	1	-126.72889	36.76454	-3.45	0.0007	-199.36437	-54.09340
x16x21	1	1.47270	0.23589	6.24	<.0001	1.00664	1.93875

Fig. 9: Parameter Estimates

Prediction Intervals

The REG Procedure
Model: MODEL1
Dependent Variable: y

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	18538	711.3975	17133	19944	12335	24741	.
2	.	29025	858.5085	27328	30721	22749	35300	.
3	.	47.7111	1219	-2361	2457	-6457	6552	.
4	.	7361	461.7648	6448	8273	1250	13471	.
5	.	6906	380.4661	6155	7658	818.1008	12995	.
6	.	8978	330.7515	8325	9632	2901	15055	.
7	.	34942	1020	32927	36958	28573	41311	.
8	.	68415	4557	59412	77419	57573	79258	.
9	.	6993	430.0375	6143	7842	891.4605	13094	.
10	.	15136	536.5173	14076	16196	9002	21270	.
11	.	14618	1135	12376	16860	8174	21062	.
12	.	7395	427.0560	6551	8239	1295	13495	.
13	.	7361	461.7648	6448	8273	1250	13471	.
14	.	7586	1005	5599	9572	1226	13945	.
15	.	6218	418.4994	5391	7045	120.3015	12316	.
16	.	6218	418.4994	5391	7045	120.3015	12316	.
17	.	6218	418.4994	5391	7045	120.3015	12316	.
18	.	19110	854.9251	17421	20800	12837	25384	.
19	.	19110	854.9251	17421	20800	12837	25384	.
20	.	16442	851.2178	14761	18124	10171	22714	.
21	.	17278	725.1280	15845	18710	11068	23487	.
22	.	11861	424.2840	11023	12699	5761	17960	.
23	.	11861	424.2840	11023	12699	5761	17960	.
24	.	5434	474.0620	4498	6371	-679.3889	11548	.
25	.	5689	536.3952	4629	6749	-445.1897	11823	.
26	.	11799	1298	9234	14364	5235	18363	.
27	.	9143	353.6704	8444	9842	3061	15225	.
28	.	11799	1298	9234	14364	5235	18363	.
29	.	5650	484.0919	4694	6607	-466.4950	11767	.
30	.	5650	484.0919	4694	6607	-466.4950	11767	.
31	.	8306	1356	5627	10985	1697	14915	.
32	.	9372	1123	7152	11591	2935	15808	.
33	.	10990	679.8378	9647	12333	4801	17179	.
34	.	14852	585.5228	13695	16008	8700	21003	.
35	.	20089	628.4541	18847	21331	13921	26257	.
36	.	11427	1013	9425	13430	5063	17792	.
37	.	11696	1008	9704	13688	5334	18057	.
38	.	16216	501.5951	15225	17207	10094	22338	.
39	.	17043	878.7472	15307	18779	10757	23329	.
40	.	20623	627.5936	19383	21863	14455	26790	.
41	.	18020	628.1006	16779	19261	11852	24188	.

Fig. 10: Prediction Intervals

Appendix C: SAS Code

```
title "Auto Price Dataset";         /* Import the Dataset */
data auto;
infile '/folders/myfolders/data/auto.csv' dlm=',' firstobs=2;
input id x1$ x2$ x3$ x4$ x5$ x6$ x7$ x8$ x9 x10 x11 x12 x13 x14$ x15$ x16 x17$ x18 x19 x20 x21 x22 x23 x24 y;
drop x1--x4 x6 x8 x14 x15 x17;        /*        Initially drop some categorical variables        */

/* Create dummy variables for the remaining categorical variables */
title "Dummy Variables";
data auto_dummy;
        set auto;
        /* Changing "four" and "two" to numerical */
        if x5 = "four" then do; x5temp = 4; end; else if x5 = "two" then do; x5temp = 2; end;
        drop x5; rename x5temp = x5;

        /* Creating dummy variables for x7 */
        /* x7,4wd = 4-wheel drive, x7,rwd = rear-wheel drive, x7fwd = front-wheel drive */
        if x7 = "4wd" then do; x7,4wd = 1; x7,rwd = 0; x7fwd = 0; end;
        else if x7 = "rwd" then do; x7,4wd = 0; x7,rwd = 1; x7fwd = 0; end;
        else if x7 = "fwd" then do; x7,4wd = 0; x7,rwd = 0; x7fwd = 1; end;
        drop x7;

title "Dropping Variables";
data auto_drop;
        set auto_dummy;
        /* Drop the variables here */
        drop x7fwd x24 x23 x13 x18 x19 x22;
        /* Move the dependent variable to the end */
        tempY = y; drop y; rename tempY = y;

/* Create pairwise interaction terms based on the remaining variables; also higher order terms. */
title "Interaction Terms and Higher Order";
data auto_function;
        set auto_drop;
        x16x11 = x16*x11; x16x20 = x16*x20; x16x21 = x16*x21; x16x7,4wd = x16*x7,4wd; x16x7,rwd = x16*x7,rwd;
        x11x20 = x11*x20; x11x21 = x11*x21; x11x7,4wd = x11*x7,4wd; x11x7,rwd = x11*x7,rwd;
        x20x21 = x20*x21; x20x7,4wd = x20*x7,4wd; x20x7,rwd = x20*x7,rwd;
        x21x7,4wd = x21*x7,4wd; x21x7,rwd = x21*x7,rwd;

        /* Want the dependent y variable to be in the last column. */
        tempY = y; drop y; rename tempY = y;

/* Split the data set into training and testing sets */
title "Data Split";
proc surveyselect data=auto_function rate=0.8
        out= auto_select seed = 12345 outall
        method=srs;
data auto_train auto_test;
        set auto_select;
        if selected =1 then output auto_train; else output auto_test;
        drop selected;
run;

/* Print the correlation matrix to check for any multicollinearities */
title "Correlation Matrix";
proc corr data=auto_train;

/* Check for the Variance Inflation Factor to determine serious multicollinearities. */
title "Regression for VIFs";
proc reg data = auto_train;
        model y = x9--x7,rwd / p vif;
run;
```

```
/* Perform a preliminary screening on the importance of the IV's. */
title "Stepwise Regression";
proc stepwise data = auto_train;
        model y = x9--x7,rwd  / stepwise sle = 0.05 sls = 0.05;
run;

/* Perform a second screening to know which of the interaction terms are significant. */
title "Stepwise Regression";
proc stepwise data = auto_train;
        model y = x11 x16 x20 x21 x7,4wd x7,rwd x11x20 x11x21 x11x7,4wd x11x7,rwd x20x21 x20x7,4wd x20x7,rwd
                    x21x7,4wd x21x7,rwd x16x11 x16x20 x16x21 x16x7,4wd x16x7,rwd
                    / stepwise sle = 0.05 sls = 0.05;

/* Perform a Regression on the model obtained from Stepwise Regression */
title "Regression on Stepwise";
proc reg data = auto_train;
        model y =  x11 x20 x21 x7,4wd x7,rwd x21x7,rwd x11x21 x16 x16x21;
run;

/* Get the R^2 Estimates */
title "R-Squares";
proc rsquare data=auto_train cp mse sse adjrsquare;
        model y = x11 x20 x21 x7,4wd x7,rwd x21x7,rwd x11x21 x16 x16x21;
run;

/* Check for Independence using the DW Testing */
title 'Dubrin-Watson Test';
proc reg data = auto_train;
        model y=x11 x20 x21 x7,4wd x7,rwd x21x7,rwd x11x21 x16 x16x21/ p dw dwprob;
run;

/* Check for any outliers an influential points */
title 'Checking for Outliers and Influential Points';
proc reg data=auto_train outest=r;
        model y = x11 x20 x21 x7,4wd x7,rwd x21x7,rwd x11x21 x16 x16x21 / r influence;
run;

/* Remove any outliers and influential points */
title "Remove Observations";
data auto_removed_train;
        set auto_train;
        if id=14 then delete; if id=16 then delete; if id=89 then delete;
        drop id;
run;

/* Perform a complete F-test on the overall model. */
title "Complete F-Test";
proc reg data = auto_removed_train;
        model y = x11 x20 x21 x7,4wd x7,rwd x21x7,rwd x11x21 x16 x16x21;
run;

/* Do a Hypothesis Testing */
title "Hypothesis Testing for b_j";
proc reg data=auto_removed_train;
        model y=x11 x20 x21 x7,4wd x7,rwd x21x7,rwd x11x21 x16 x16x21 /alpha=0.05 p clm cli;
run;

/* Based on the hypothesis testing, we check for the significance of x7 dummies. */
/* Use the partial F-test on the reduced model. */
title "Partial F-Testing - drop x7,4wd and x7,rwd?";
proc reg data=auto_removed_train;
        model y=x11 x20 x21 x7,4wd x7,rwd x11x21 x16 x16x21; pft: test x7,rwd=0, x7,4wd=0;
run;
```