Chapter 3

How Much is Your Car Worth? Introduction to Multiple Regression

Have you ever browsed through a car dealership and observed the sticker prices of the vehicles? If you have ever seriously considered purchasing a vehicle, you can probably relate to the difficulty of determining if that vehicle is a good deal or not. Most dealerships are willing to negotiate on the sale price, so how can you know how much to negotiate? For novices like the author of this lab, it is very helpful to refer to an outside pricing source, such as Kelley Blue Book, before agreeing on a purchase price.

For over 80 years, Kelley Blue Book has been a resource for accurate vehicle pricing. Their website, www.kbb.com, provides a free on-line resource where anyone can input several car characteristics (such as age, mileage, make, model and condition) and quickly receive a good estimate of the retail price.

In this lab, you will use a relatively small subset of the Kelley Blue Book database to describe the association of several explanatory variables (car characteristics) to the retail value of a car. Before developing a complex multiple regression model with several variables, let's start with a quick review of the simple linear regression model by asking a question: "Are cars with lower mileage worth more?" Clearly, it seems reasonable to expect to see a relationship between mileage (number of miles the car has been driven) and retail value. The dataset, `carlab`, contains the make, model, equipment, mileage and Kelley Blue Book suggested retail price of several used 2005 GM cars[1].

[On Your Own:]

(1) Produce a scatterplot from the `carlab` dataset to display the relationship between mileage (`Mileage`) and suggested retail price (`Price`). [In Minitab, `Graph` → `Scatterplot` → `With Regression`. Specify `Price` as the Y variable and `Mileage` as the X variable.]

[1] Collected using tables from the 2005 Central Edition of the Kelley Blue Book.

Does the scatterplot show a relationship between Mileage and Price?

Calculate the least squares regression line. [In Minitab, `Stat` → `Regression` → `Regression`. Specify `Price` as the response and `Mileage` as the predictor.] Report the regression model, the root mean square error (S), the R^2 value, the correlation coefficient, the t-statistics and p-values for the estimated **model coefficients** (the intercept and slope). Based on these statistics, can you conclude that mileage is a strong indicator of price? Why or why not?

The first car in this dataset is a Buick Century with 8221 miles. Calculate the residual value for this car (the observed suggested retail price minus the expected price calculated from the regression line).

The t-statistic for the regression model slope indicates that mileage is an important variable, however, the R^2 value and the scatterplot clearly show that the regression line does not explain much of the variation in retail prices. It is always better to take a few minutes to visualize the data instead of solely focusing on a p-value. This scatterplot and R^2 value suggest that including other explanatory variables in the regression model might help to better explain the variation in retail price.

In this lab, you will build a linear combination of explanatory variables that explain the response variable, retail price. As you work through the lab, you will find that there is not one technique or "recipe" that will give the best model. In fact, you will come to see that there isn't just one "best" model for these data.

Multiple regression is arguably the single most important method in all of statistics, in large part because regression models are so broad in their potential applications, but also because a good understanding of regression is all but essential for understanding so many other, more sophisticated statistical methods.

Unlike an assignment for most mathematics classes, where every student is expected to submit the one right answer, it is expected that each final regression model submitted by various students for this lab will be at least slightly different. This lab focuses on understanding the scientific process of developing a statistical model. It doesn't matter if you are developing a regression model in economics, psychology, sociology or engineering, there are always key questions and processes that should be evaluated before a final model is submitted. While a "best" model may not exist for these data, there are certainly many bad models that should be avoided.

3.1 Multiple Regression Can Serve Multiple Purposes

It is important to note that multiple regression analysis can be used to serve different goals. The most common goals of multiple regression are to:

(A) **Describe:** Develop a model to describe the relationship between the explanatory variables and the response variable.

(B) **Predict:** Use a set of sample data to forecast or make predictions. A regression model can be used to predict future response values from explanatory variables observed within the range of our sample data.

(C) **Confirm:** Theories are often developed about which variables, or combination of variables, need to be included in a model. Regression methods can be used to determine if the contribution of each explanatory variable in a model (e.g., mileage) captures much of the variability in the response variable. This includes determining if the association between the explanatory variables and the response could just be due to chance. For example, should mileage be used to predict retail price?

Theory may also predict the type of relationship that exists, such as "cars with lower mileage are worth more." More specific theories can also be tested, such as "retail price decreases linearly with mileage."

In most circumstances, multiple regression analysis is conducted on data from observational studies, not experiments. A significant test statistic for a coefficient can provide evidence that a proposed theoretical relationship exists. However, without *a priori* theoretical justification, a significant correlation (and thus a significant coefficient) does not imply a causal link between the explanatory variable and the response.

[Key Concept] Correlation does not imply causation.

3.2 Variable Selection Techniques for Description or Prediction Models

Often, the primary issue in multiple regression is determining which variables to include in the model (and which to leave out). Clearly, all potential explanatory variables could be included in a regression model, but that often results in a cumbersome model that is difficult to understand. On the other hand, a model that includes only one or two of the measured explanatory variables may provide substantially different predictions than the complex model. This tension between finding a simple model and finding a model that best explains the response is what makes it difficult to find a "best" model. Finding just the right mix that provides a relatively simple linear combination of explanatory variables often resembles an exploratory artistic process much more than a formulaic recipe.

In addition to avoiding an unwieldy model, including redundant or unnecessary variables can result in an inefficient regression model. Inefficiently estimated coefficients have larger variability and thus their t-statistics may be lower than they should be. Remember that with

any multiple comparison problem, an α-level of 0.05 means there is a 5% chance that an irrelevant variable will be found significant and may inappropriately be determined important for the model.

Failing to include a relevant variable is usually more serious. It can cause biased estimates of the regression coefficients and invalid t-statistics, especially when the excluded variable is highly significant or if the excluded variable is correlated with other variables also in the model.

For this lab, we will consider the response to be the suggested retail price from Kelley Blue Book (the `Price` variable in the data). The following variables are considered relevant potential explanatory variables:

- `Make` (Buick, Cadillac, Chevrolet, Pontiac, SAAB, Saturn)
- `Model` (specific car for each previously listed `Make`)
- `Trim` (specific type of `Model`)
- `Type` (Sedan, Coupe, Hatchback, Convertible or Wagon)
- `Cyl` (number of cylinders: 4, 6, or 8)
- `Liter` (a measure of engine size)
- `Doors` (number of doors: 2, 4)
- `Cruise` (1= cruise control, 0 = no cruise control)
- `Sound` (1= upgraded speakers, 0 = standard speakers)
- `Leather` (1= leather seats, 0 = not leather seats)
- `Mileage` (number of miles the car has been driven)

Several statistical techniques are available for choosing between models if the objective of your regression model is to describe a relationship or predict new response variables.

For example, a larger R^2 value indicates that more of the variation in retail price is explained by the model. However, R^2 increases whenever another predictor is added (this is just an algebraic fact not proven here). R^2 is most useful when comparing models with the same number of predictors, but when comparing models with the different numbers of terms (explanatory variables), other techniques are suggested.

When a large number of variables are available, **stepwise regression** is a sequential variable selection technique that can automate the process of building a model. Stepwise regression is an iterative technique that can be used to identify key variables to include in a regression

model. For example, **forward stepwise regression** begins by fitting several single-predictor regression models for the response, one for each individual explanatory variable. The single explanatory variable with the highest R^2 value is then determined to be in the model. In the next step, the first explanatory variable (call it X_1) is already in the model, and all possible regression models using X_1 and exactly one other explanatory variable are determined. From among these, the regression model with the highest R^2 value is again selected to identify X_2. After the first and second explanatory variables, X_1 and X_2, are selected, the process is repeated to find X_3. This continues until including additional variables in the model no longer greatly improves the R^2 value.

Sequential procedures have a tendency to include too many variables and at the same time they sometimes eliminate important variables. With improvements in technology, most statisticians prefer to use more "global" techniques that compare all possible subsets of the explanatory variables, such as Mallows' Cp, Akaike's or the Bayes' Information Criteria.

[On Your Own:]

(2) Conduct a stepwise regression analysis. [In Minitab, use `Stat` → `Regression` → `Stepwise`. In the response use `Price,` in `Predictors` use all quantitative explanatory variables available.]

(3) Use Mallows' Cp to develop a model. [In Minitab, use `Stat` → `Regression` → `Best Subsets`. In the response use `Price,` in `Free predictors` use all quantitative explanatory variables available.] In general, models are selected where Cp is small and close to the number of terms (explanatory variables) in the model. Select a model that has a relatively low Cp, a large R^2, and a relatively small number of explanatory variables.

(4) Compare the regression models in Questions (2) and (3). Are different explanatory variables considered important? Are the R^2 values similar? Explain.

If your goal is to develop a model to describe or predict, it is common to evaluate the strength of the relationship described by the model. Here, we are not concerned about the significance of each explanatory variable, but how well the overall model fits. Iterative techniques are useful in providing a high R^2 value while limiting the number of variables.

If your goal involves confirming a relationship, iterative techniques are not suggested. Confirming a theory is similar to hypothesis testing. Iterative variable selection techniques test each variable, or combination of variables, several times and thus the p-values are not reliable. The stated significance level for a t-statistic is only valid if the data are used for a single test. If multiple tests are conducted to find the best equation, the actual significance level for each test for an individual component is invalid.

[Key Concept] Iterative techniques should not be used to evaluate the importance of each explanatory variable. The t-statistics in the final model may be

inflated.

3.3 Checking Model Assumptions

3.3.1 Assessing Shapes and Patterns in Residual Plots

Whenever a regression model is created, it is necessary to check the following model assumptions:

- The variance of the error population is constant at all levels of the explanatory variables. In other words, the error terms in the regression model (also called residuals) are assumed to come from a single population with variance σ^2.
- The error terms are independent and identically distributed.
- The error terms follow a normal probability distribution: denoted as $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1, \ldots, n$ (n = the total number of observational units in the data).

In regression, these assumptions are generally checked by looking at the **residuals** from the data: $y_i - \widehat{y}_i$. Here, y_i are the observed responses and $\widehat{y}_i$ are the estimated responses calculated by the regression model. Multi-variable regression equations are difficult to visualize and even for single-variable (simple) regression lines, residual plots often emphasize violations of model assumptions better than a plot of the regression line on a scatterplot.

Instead of conducting formal hypothesis tests about error terms (i.e., residual values), plots are used to visually determine if the assumptions hold. The theory and methods are simplest when any scatterplot of residuals resembles a single, horizontal, oval balloon, but real data may not cooperate by conforming to the ideal pattern. An ornery plot may show a wedge, a curve, or multiple clusters. Any of these plot patterns reveal that our error terms are violating at least one model assumption and it is likely that we have inefficient (more variable than necessary) and biased estimates of our model coefficients. The following section illustrates strategies for dealing with one of these unwanted shapes, a wedge-shaped (or right-opening megaphone) pattern.

Heteroskedasticity is a term used to describe the situation where the variance of the error term is not constant for all levels of the explanatory variables. For example, in the regression equation, $\texttt{Price} = 24,765 - 0.173\ \texttt{Mileage}$, the spread of the suggested retail price values around the regression line should be about the same whether mileage is 0 or mileage is 50,000. If heteroskedasticity exists in the model, the most common remedy is to transform either the explanatory variable, the response variable, or both in the hope that the transformed relationship will exhibit **homoskedasticity** (equal variances) in the error terms.

[On Your Own:]

(5) Create residual plots for the regression equation calculated in Question (3). [In Minitab, `Stat` → `Regression` → `Regression` select `Price` as response and select the appropriate predictors. In `Graphs` select `Residuals versus fits` and select all the explanatory variables in your model for `Residuals versus the other variables`. `Fits` is another term for predicted or estimated retail price (the $\widehat{y}_i$ values).] Describe any pattern you see in the residual plots. Does the size of the residual change as mileage changes? Does the size of the residual change as the predicted retail price changes? You should see patterns indicating heteroskedasticity (non-constant variance).

Another pattern that may not be immediately obvious from these residual plots is the right skewness seen in the residuals versus mileage plot. Often economic data, such as price, is right skewed. To see the pattern, look at just one vertical slice of this plot. For example, with a pencil sketch an oval around all points with mileage close to 8,000. Describe any skewness seen in these points.

(6) Transforming data using roots, logarithms, or reciprocals can often reduce heteroskedasticity and right skewness. Transform the suggested retail price to log(price) and $\sqrt{\text{price}}$. [Use Minitab's `logten` and `sqrt` functions of `Price` to form two new transformed response variables. In Minitab, `Calc` → `Calculator`, use the function key to select `logten` of `Price, LOGT(Price)`. Do the same for `sqrt`.] Create regression models and residual plots for these transformed response variables using the explanatory variables selected in Question (3).

Which transformation had the best impact on the residual plots? Give the R^2 values of both new models. Do the best residual plots correspond to the best R^2 values? While other transformations could be tried, use the better of the two you have tried as the response variable of choice in the remainder of this lab. You can label your transformed data `TPrice` (Or you can call it something else, but this lab will refer to the transformed response variable as `TPrice` from here on.)

Note that in single variable regression models, residual plots show the same information as the initial fitted line plot. However, the residual plots often emphasize violations of model assumptions better than the fitted line plot. In addition, multivariate regression lines are very difficult to visualize. Thus residual plots are essential when multiple explanatory variables are used.

3.3.2 Examining Residual Patterns Across Time/Order

Recall that independence of the error terms is one of the assumptions of our model. **Autocorrelation** exists when consecutive error terms are related. To identify autocorrelation, we plot the residuals versus the order of the data entries. If the ordered plot shows a pattern, then we conclude that autocorrelation exists. When autocorrelation exists, the variable

responsible for creating the pattern in the ordered residual plot should be included in the model.

Autocorrelation can come from various sources, time (order in which the data were collected) is perhaps the most common but spatial autocorrelations can also be present. If time is indeed a variable that should be included in the model, a specific type of regression model, called a time series model, should be used.

[On Your Own:]

(7) Create a `Residual vs. Order` plot from the `TPrice` vs. `Mileage` regression line. [In Minitab, use `Stat` → `Regression` → `Regression`, in `Graphs` select `Residuals versus order`.] Describe any pattern you see in the ordered residual plot. Apparently something in our table of data is affecting the residuals in terms of the order of the data. Clearly, time is not the influential factor in our dataset (all of the data are from 2005). Can you suggest a variable in our table that may be causing this pattern?

Create a second `Residual vs. Order` plot using `TPrice` as the response and using the explanatory variables selected in Question (3). Notice that there are still strong patterns.

While ordered plots only make sense in model checking when there is a meaningful order to the data, the residuals versus order plots could demonstrate the need for including additional explanatory variables in the regression model.

We do not have a time variable in this dataset, so reordering the data would not change the meaning of the data. It is expected that reordering the data could eliminate the pattern. However, the clear pattern seen in the residuals vs. order plots should not be ignored because it indicates that we could create a model with a much higher R^2 value if we can account for this pattern in our model. This type of autocorrelation is called **taxonomic autocorrelation** meaning that the relationship seen in this residual plot is due to how the items in the dataset are classified. Suggestions on how to address this issue are discussed in later sections of this lab.

3.3.3 Outliers and Influential Observations

Any data values that don't seem to fit the general pattern of the dataset are called **outliers**. If the outlier has an extreme value in the direction of the response variable the R^2 value may be influenced. If the outlier has an extreme value in the direction of an explanatory variable, the model coefficients may be impacted.

[On Your Own:]

(8) Using the regression equation calculated in Question (3), identify any residuals (or cluster of residuals) that don't seem to fit the overall pattern in the `Residual vs. Fits` and `Residual vs. Mileage` plots. [In Minitab, click on a graph, then select `Editor` → `Brush`, move the cursor over any outliers]. Look at these specific rows of data, are there any consistencies that you can find? Does this help identify the patterns that were found in the ordered residual plots? Why or Why not?

(9) Run the analysis with and without these potential outliers. If the coefficients change dramatically between the regression models, these points are considered **influential**. If any observations are influential, great care should be taken to verify their accuracy. Explain why the outliers in the `carlab` dataset have higher residuals than other observations.

In some situations, clearly understanding outliers can be more time consuming (and possibly more interesting) than working with the rest of the data. In can be quite difficult to determine if an outlier was accurately recorded or to know whether the outliers should be included in the analysis.

The simplest approach to this problem is to run the analysis twice: once with the outliers included and once without. If the results are similar, then it doesn't matter if the outliers are included or not. If the results do change, it is much more difficult to know what to do. Most statisticians tend to err on the side of keeping the outliers in the sample dataset unless there is clear evidence that they were mistakenly recorded. Whatever final model is selected, it is important to clearly state if you are aware that your results are sensitive to outliers.

3.3.4 Normally Distributed Residuals

To determine if the residuals are normally distributed two graphs are often created, a histogram of the residuals and a normal probability plot.

Normal probablity plots are created by sorting the data (the residuals in this case) from smallest to largest. Then the sorted residual are plotted against a theoretical normal distribution. If the plot forms a straight line, the actual data and the theoretical data have the same shape (i.e. the same distribution).

[On Your Own:]

(10) Create a regression line to predict TPrice from Mileage and create a histogram of the residuals as well as a normal probability plot. [In Minitab, use `Stat` → `Regression` → `Regression`, in `Graphs` select `Four in one`.] Does the data look normal? Are the ten outliers visible on the normal probability plot and the histogram?

Checking for normality is most important when using hypothesis tests (t-tests) to determine the significance of individual variables. It is not needed to create a regression line or R^2

value. The first four sub-sections above on checking model assumptions are all based on simply looking for patterns in various residual plots. Before a final model is selected, the residuals should be plotted against fitted (estimated) values, observation order, and each explanatory variable in the model. If any patterns exist, it is likely that another model exists that better explains the response variable.

At this time, it should be clear that simply plugging data into a software package and using an iterative variable selection technique will not reliably create a "best" model. Even though the calculations of the regression model and R^2 do not depend on model assumptions, identifying patterns in residual plots can often lead to another model that better explains the response variable.

[Key Concept] Always check residual plots when developing a regression model. If a pattern exists in any of the residual plots, the R^2 value is likely to improve if additional terms or transformations are included in the model.

3.3.5 Multicollinearity: Correlation Between Explanatory Variables

Multicollinearity exists when the explanatory variables are highly correlated with each other. If two explanatory variables, X_1 and X_2, are highly correlated, it can be very difficult to identify whether X_1, or X_2, or both variables are actually responsible for influencing the response variable, Y.

[On Your Own:]

(11) Create three regression models using `TPrice` as the response variable. In all three models, provide the R^2 value, the coefficients, t-statistics and p-values.

- In the first model, use only `Mileage` and `Liter` as the explanatory variables.
- In the second model, use only `Mileage` and number of cylinders (`Cyl`) as the explanatory variables.
- In the third model, use `Mileage, Liter` and number of cylinders (`Cyl`) as the explanatory variables.

Note that the R^2 values are essentially the same. Look at each coefficient, t-statistic and p-value in each model. How does your interpretation of the importance of `Cyl` change depending on the model used? Plot number of cylinders (`Cyl`) vs. `Liter` and calculate the correlation between these two variables. Create four residual plots for the third model, `residuals vs. Liter, Mileage, Fits, and Order`

If multicollinearity exists in a regression model, then the coefficients are not reliable. This is clearly evident for the explanatory variables `Liter` and `Cyl` after observing the changes

in the estimated regression coefficients ocurring from model to model in Question (11). If multicollinearity exists, try one of the following approaches:

(A) **Get more information:** If it is possible, expanding the data collection may lead to samples where the variables are not so correlated. Consider whether the data could be collected or measured differently so the variables are not correlated. For example, the data here are only for GM cars. Perhaps the relationship between engine size in liters and the number of cylinders is not so strong for data from across a wider variety of manufacturers.

(B) **Re-evaluate the model:** When two explanatory variables are highly correlated, deleting one variable will not significantly impact the R^2 value. However, if there are theoretical reasons to include both variables in the model, keep both terms. In our example, `Liter` and number of cylinders (`Cyl`) are measuring essentially the same quantity. `Liter` represents the volume displaced during one complete engine cycle. The number of cylinders (`Cyl`) also is a measure of the volume that can be displaced.

(C) **Combine the variables:** Using other statistical techniques such as *principal components*, it is possible to combine the correlated variables "optimally" into a single variable that can be used in the model. There may be theoretical reasons to combine variables in a certain way. For example, the volume (size) and weight of a car are likely highly positively correlated. Perhaps a new variable defined as density = weight/volume could be used in a model predicting price rather than either of these individual variables.

In this example, we have theoretical reasons to re-evaluate our model. `Liter` and number of cylinders (`Cyl`) are both measuring displacement (engine size). We will keep `Liter` and throw out number of cylinders, since `Liter` is a more specific variable taking on several values (only 4, 6, or 8 cylinder cars appear in the dataset).

In general, it may not be possible to "fix" a multicollinearity problem. If the goal is to simply describe or predict retail prices, multicollinearity is not a critical issue. Redundant variables should be eliminated from the model, but highly correlated variables that both contribute to the model would be acceptable if you are not interpreting the coefficients. However, if your goal is to confirm whether an explanatory variable is associated with a response (test a theory), then it is essential to identify the presence of multicollinearity and to recognize that the coefficients are unreliable when it exists.

[Key Concept] Coefficients (and their p-values) are unreliable when multicollinearity exists.

3.4 Categorical Explanatory Variables in a Regression Model

As we saw in Question (7), there is a clear pattern/structure in the residuals vs. order plot for the Kelley Blue Book car pricing data. It is likely that one of the categorical variables (`Make, Model, Trim or Type`) could explain this pattern.

[On Your Own:]

(12) Plot the response variable `TPrice` vs. the categorical variables. [In Minitab, `Graph` → `Individual Value Plots` → `One Y, With Groups`, in `Graphs variables` select `TPrice`, in `Categorical variables for grouping`, select `Make`.] Describe any pattern you see. Repeat the process for `Model, Trim, and Type.`

If any of these categorical variables are related to the response variable, then we want to add these variables to our regression model. A common procedure used to incorporate categorical explanatory variables into a regression model is to define **dummy variables**, also called **indicator variables**. Creating dummy variables is a process of mapping the one column (variable) of categorical data into several columns (dummy variables) of 0 and 1 data. Using the variable `Make` as an example, the 5 possible values (`Buick, Cadillac, Chevrolet, Pontiac, SAAB, Saturn`) can be recoded using 6 dummy variables: one for each of the 6 makes of car. For example, the dummy variable for Buick will have the value 1 for every car that is a Buick and 0 for each car that is not a Buick. Most statistical software packages have a command for creating the dummy variables automatically.

[On Your Own:]

(13) Create dummy variables for `Make` [In Minitab, `Calc` → `Make Indicator Variables` → `Indicator Variable for Make`. `Store Results into c18-c23` (select any 6 unused columns here).] Name the columns, in order, `Buick, Cadillac, Chevrolet, Pontiac, SAAB, Saturn`. Look at the new data columns to understand how the dummy variables are defined.

Now, any of these dummy variables can be incorporated into a regression model. However if you want to include `Make` in its entirety into the model, do not include all 6 dummy variables; 5 will suffice. This is because there is complete redundancy in the sixth dummy variable: knowing the values on the other five variables are all 0 for a particular car automatically tells us that this car belongs to the sixth category. Below, we will leave the `Saturn` dummy variable out of our model. The coefficient for a dummy variable is an estimate of the average amount (of the response variable) by which a "1" for that dummy variable will exceed a "0." For example, this will mean that the estimated coefficient for the `Buick` variable is an estimate of the average change in `TPrice` when the car is a Buick rather than a Saturn (while all other explanatory variables in the model remain unchanged).

[On Your Own:]

(14) Build a new regression model using `TPrice` as the response and `Mileage, Liter, Buick, Cadillac, Chevrolet, Pontiac,` and `SAAB` as the explanatory variables.

(15) Create dummy variables for `Type.` Include the `Make` and `Type` dummy variables, plus the variables `Liter, Doors, Cruise, Sound, Leather and Mileage` in a model to predict `TPrice.` Remember to leave the last category out of the model for the `Make` and `Type` dummy variables (i.e., leave `Saturn` and `Wagon` out of the model). Compare this regression model to the other models that you have fit to the data in this lab. Do the residual plots look more random than they did in earlier problems? Explain why or why not.

The additional categorical variables are important in connecting the residual plots and the model assumptions. When additional variables, such as `Make` and `Type` were included in the model, the residual plots look much better.

[On Your Own:]

(16) Create a "best" regression model. Validate the model assumptions. Look at residual plots and check for heteroskedasticity, multicollinearity, autocorrelation and outliers. Submit your suggested least squares regression formula along with a limited number of appropriate graphs that provide justification for your model. Your final modal should not have significant clusters, skewness, outliers or heteroskedasticity appearing in the residual plots. Be prepared to defend your model in class.

3.5 Lab Summary

For this dataset, cars were randomly selected within each make-model-type of 2005 GM cars produced, then suggested retail prices were determined from Kelley Blue Book. While this is not a simple random sample of all 2005 GM cars actually on the road, there is still reason to believe that your final model will provide an accurate description or prediction of retail price for GM cars. This is also an observational study, not an experiment. Therefore, even though we may have high correlation between our explanatory variables and the response, we certainly can not use this model as proof of causation. There may be theoretical or practical reasons to believe that higher mileage causes higher prices, but the final model can only be used to show that there is an association.

In multi-variable datasets, it is possible to obtain seemingly conflicting results. Such as having a p-value indicate that mileage is significant predictor of price, but an R^2 value that indicates that mileage does not explain much of the variability in price. Attempts to improve

a regression model, such as finding the appropriate data transformation or addressing multicollinearity is often more of an art than automated process. Even highly technical statistical software packages can not automatically develop a "best" model. By visually interpreting residual plots, a final model can be developed that is much better than either of the models developed by iterative (algorithmic) statistical software techniques.

Chapter 4

A Closer Look at Multiple Regression

4.1 Interaction and Terms for Curvature

In addition to using the variables provided in a dataset, it is often beneficial to create new variables that are functions of the existing explanatory variables. These new explanatory variables terms can be quadratic (X_2), cubic (X_3), or products of two explanatory variables ($X_1 * X_2$), called **interaction terms**.

If a plot of residuals versus an explanatory variable shows curvature, the model may be improved by including a quadratic term. Is the relationship between mileage and retail price linear or quadratic for the Kelley Blue Book data? To test this, a quadratic term, `Mileage*Mileage` can be created and included in a regression model.

Does the mileage effect the retail price differently depending on the liter size of the car? An **interaction** occurs if the effect of one variable, such a mileage, varies depending on a second variable, such as liter. To test this, an interaction term, `Mileage*Liter` can be created and included in a regression model.

[On Your Own:]

(17) Create a quadratic mileage term. [In Minitab, `Calc` → `Calculator` , store result in variable:`MileMile`, expression `Mileage*Mileage`]. Use both `Mileage` and `MileMile` to predict `TPrice`. Does this quadratic term appear to improve the model?

(18) Create an interaction term. [In Minitab, `Calc` → `Calculator` , store result in variable:`MileLiter`, expression `Mileage*Liter`]. Use `Mileage, Liter`, and `MileLiter` to predict `TPrice`. Does this interaction term appear to improve the model?

The potential outliers identified in Question (8) can provide an interesting demonstration of interaction terms. The fitted slope to predict price from mileage is -0.48 (see Question (1)).

For the 10 Cadillac XLR-V8s, much steeper than the slope of -0.17 found when using the dataset as a whole. This shows that depreciation for these high-end cars is almost 50 cents a mile, as opposed to 17 cents a mile on average for all car types combined. Also, note that many final models included the `Cadillac` make, the `Convertible` type and the engine size (`Liter`) in the regression model. High values for each of these variables tends to represent high-end cars (`Cadillac` $= 1$, `Convertible` $= 1$, and `Liter` $= 4.6$). These comments suggest that an interaction term for `Mileage` and `Liter` and possibly for `Mileage` with `Make` or even `Type` may be helpful additions to the model.

[On Your Own:]

(19) Use the several outliers observed in Question (8) to develop additional potential interaction terms. Determine if they improve the regression model.

Brain and Body Weights

In the process of studying the effect of sleep on mammals, T. Allison and C. Cicchetti collected the brain weights (in grams) and body weights (in kilograms) of 62 mammals. In the dataset `weights`, you can see the measurements for these two variables for each of the 62 animals studied.

[On Your Own:]

(20) Create a scatterplot and regression line to predict `BrainWt` from `BodyWt`. Even though the R^2 value is reasonable, it is clear that there are extreme outliers in both the X and Y axis. Often taking the logarithm of both the X and Y variables, can significantly improve the regression model.

(21) Create a scatterplot and regression line to predict Log(`BrainWt`) from Log(`BodyWt`). Create the appropriate residual plots and evaluate the validity of the model.

Arsenic in Toenails

Karagas, Morris, Weiss, Spate, Baskett, and Greenberg collected toenail clippings of several individuals along with samples of each person's regular drinking water. In this pilot study, the researchers attempted to determine if arsenic concentrations in the toenails could be used determine the level of arsenic in their water. High arsenic concentrations are related to cancer and several studies have found a positive correlation between arsenic concentrations of drinking water and toenail samples.

The data can be found in the dataset `arsenic`.

[On Your Own:]

(22) Create a scatterplot and regression line to predict `ArsWater` from `ArcToe`. There are extreme outliers along both the X and Y axes. However, 6 of the 21 water samples contained no arsenic and taking the logarithm of `ArsWater` is not possible.

(23) Create a scatterplot and regression line to predict Log(`ArsWater`+1) from Log(`ArcToe`). Create the appropriate residual plots and evaluate the validity of the model. While the transformations are helpful, model assumptions are still violated.

(24) Attempt creating a few more variables or transformations to create a better model. This is an example were you are unlikely to find a satisfactory simple regression model.

4.2 Developing a Model to Confirm a Theory

If the goal is to confirm a theoretical relationship, statisticians tend to go through the following steps to identify an appropriate theoretical model. Then, regression analysis is conducted one time to determine if the data support their theories.

[Key Concept] The same data should not be looked at both to develop a model and to test it.

(A) Verify that the response variable provides the information needed to address the question of interest. What is the range and variability of responses you expect to observe? Is the response measurement precise enough to address your question of interest?

(B) Investigate all explanatory variables that may be of importance or could potentially influence your results. Note that some terms in the model will be included even though the coefficients may not be significant. "In some fields there is a large body of physical theory on which to draw in explaining relationships between [explanatory variables] and responses. This type of non-statistical knowledge is invaluable in choosing [explanatory variables,] ..., interpreting the results of the analysis, and so forth. Using statistics is no substitute for thinking about the problem (Montgomery, 2003, p. 21).

(C) For each of the explanatory variables that are planned to be included in the model, describe whether you would expect a positive or negative correlation between the variable and the response variable.

(D) Use any background information available to identify what other factors are assumed to be controlled within the model. Could measurements, materials, and the process of data collection create unwanted variability? Identify any explanatory variables that may influence the response, determine if information on these variables can be collected, and whether the variables can be controlled throughout the study. For example, in the Kelley Blue Book dataset, the condition of the car was assumed to be the same for all cars. The data was collected in 2005 for GM cars with model year 2005. Since these cars are relatively new, we assumed that these cars are in excellent condition. Any model we create for these data would not be relevant for cars that had been in any type of accident.

(E) What conditions would be considered normal for this type of study? Are these conditions controllable? If a condition changed during the study, how might it impact the results?

For the following exercises assume that you have been asked to determine if there is an association between each of the explanatory variables and the response in the `carlab` dataset.

[On Your Own:]

(25) Use any background information you may have (not the `carlab` dataset) to predict how each explanatory variable (except `Model` and `Trim`) will influence `TPrice`. For example will increasing `Liter` or `Mileage` have a positive or negative association with `TPrice`? List each `Make` and identify which will impact `TPrice` most and in which direction?

(26) Identify which factors are controlled in this dataset. Can you suggest any factors outside the provided dataset that should have been included? If coefficients are found to be significant (have small p-values), will these relationship hold for all areas in the US? Will the relationships hold for 2004 or 2001 cars?

(27) Run a regression analysis to test your hypothesized model. Which variables are important in your model? Did you accurately estimate the direction of each relationship? Note that even if a variable is not significant, it is typically kept in the model if there is a theoretical justification for it.

Partisan Politics

Do the population characteristics of each state, (e.g. unemployment level, education level, age, income level, religious preferences, health insurance and voter turnout) influence how people vote? Do these characteristics have an impact on the composition of state legislatures. They collected data from 50 states on each of these factors from the US Census Bureau[1]. The state government compositions are as of 2001.

[On Your Own:]

(28) Formulate hypotheses about how each of the explanatory variables (the population characteristics) will influence voting patterns. Then form a regression model using the composition of the state governments in 2001 to test your theories. Plot the residuals and check the model assumptions. State your conclusions about each hypothesis.

Iowa Caucuses

[1]Data from Nebraska should not be included in the analysis since the state government is composed completely of non-partisans.

The `Caucuses` dataset contains the 2008 democratic and republican caucus results. Note that republicans count actual votes. Democrats don't record individual votes, but each precinct allocates its delegates based on the number of people in the precinct supporting each candidate during the caucus. Senator Clinton was expected to win in most polls, but ended up a surprising third after Senators Obama and Edwards. Many political analysts have found that females, less educated, and older people voted for Senator Clinton while younger, more educated people and African Americans prefer Senator Obama.

[On Your Own:]

(29) Use the `Caucuses` data to test the hypothesis that education level is correlated to a higher percentage of votes for Senator Obama.

(30) Use the `Caucuses` data to create a multivariate model that "best" predicts the percentage of delegates Senator Obama would win.

Socio-Economic Factors and HIV/AIDS

Is there any combination of socio-economic factors that can be used to explain the prevalence of HIV/AIDS in the world's least-developed countries? Least Developed Countries (or fourth-world countries) are countries classified by the United Nations as meeting all three of the following criteria: a low income criteria, a human resource weakness criteria, and an economic vulnerability criteria.

The World Health Organization (WHO) attempts to measure the number of people living with HIV/AIDS. Regrettably, it is often difficult to get accurate counts.

[On Your Own:]

(31) The `AIDS` dataset shows the prevalence of HIV/AIDS in adults aged 15-49 in least-developed countries along with country-level data on several socio-economic factors. Use this dataset to create a multivariate regression model to predict the prevalence of HIV/AIDS from socio-economic factors.

Are you Satisfied?

In recent decades, happiness and satisfaction have become areas of interest for economists on an international level. The World Values Survey (http://www.worldvaluessurvey.org/) attempts to capture the attitudes and values of countries all over the world. Surveys were initially conducted in 1981, and then again around 1990, 1995, 2000 and 2005, enabling social scientists to compare responses within countries and over time. This dataset relies on national surveys, translated into local languages and conducted by independent polling partners in each country. In the year 2000 battery of surveys, the most recent set that is publicly available, more than 60,000 people from nearly 100 different countries participated

in the survey, making it a singular resource that Frey and Stutzer call "the best source available today for international comparisons of life satisfaction".

[On Your Own:]

(32) Create a regression model to predict self-reported levels of personal satisfaction with one's own life using any of the other variables in the World Values survey data.