Paper 333-2012

The Steps to Follow in a Multiple Regression Analysis

Theresa Hoang Diem Ngo, La Puente, CA

ABSTRACT

Multiple regression analysis is the most powerful tool that is widely used, but also is one of the most abused statistical techniques (Mendenhall and Sincich 339). There are assumptions that need to be satisfied, statistical tests to determine the goodness fit of the data and accuracy of the model, potential problems that may occur in the model, and difficulties of interpreting the results. The first challenge is in the application of the techniques – how well analysts can apply the techniques to formulate appropriate statistical models that are useful to solve real problems. The second challenge is how to use a suitable statistical software package – such as SAS® – to deploy the correct procedures and produce the necessary output for assessing and validating the postulated model.

INTRODUCTION

In order to apply the techniques shown in this paper, analysts must have taken an undergraduate course in applied regression analysis and have well-rounded understanding of the statistical tests and terms. It would help to review the concepts before applying the techniques. Analysts will develop an ability to build appropriate multiple regression models and to interpret the results of their analyses. For other statisticians who have experience in model building, it is still beneficial to explore and practice different procedures. The five steps to follow in a multiple regression analysis are model building, model adequacy, model assumptions – residual tests and diagnostic plots, potential modeling problems and solution, and model validation.

DATA SET

Using a data set called Cars in SASHELP library, the objective is to build a multiple regression model to predict the invoice of a vehicle. The invoice (y) is modeled as a function of cylinders, engine, horsepower, length, MPG city, MPG highway, weight, wheelbase, drive train, make, and type.

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$$

Response variable (y) = invoice

Independent Variables (x_i) = cylinders, engine, horsepower, length, MPG city, MPG highway, weight, wheelbase, drive train, make, and type.

STEP 1. MODEL BUILDING

Building a model is rarely a simple or straightforward process (Mendenhall and Sincich 339). Analysts must have a prior knowledge of the variables to identify as independent variables to be included in the model. The independent variables can be first-order or second-order terms, interaction terms, and dummy variables. The following variable screening methods, stepwise regression and all-possible-regressions selection procedure, can help analysts to select the most important variables that contribute to the response variable.

1) Stepwise Regression determines the independent variable(s) added to the model at each step using t-test.

CODE

```
DATA cars ;
SET sashelp.cars ;
     IF DriveTrain = 'All'   THEN dummy1=1 ; ELSE dummy1=0 ;
     IF DriveTrain = 'Front' THEN dummy2=1 ; ELSE dummy2=0 ;
     IF Make = 'Acura'       THEN dummy3=1 ; ELSE dummy3=0 ;
/* Also create dummy variables for classification variables: Make and Type*/
------------------------------------ OMITTED CODES ----------------------------------------
RUN ;

PROC REG DATA = cars ;
MODEL invoice = Cylinders EngineSize Horsepower Length MPG_City
                MPG_Highway  Weight Wheelbase dummy: / SELECTION=stepwise;
/* The colon after dummy: lists all the dummy variables for MAKE, TYPE, DRIVETRAIN */
RUN ;
```

Note: The REG Procedure does not have a CLASS statement to specify classification or categorical independent variables. Therefore the number of dummy variables is created and specified in the MODEL statement will be one less than the number of levels for each classification independent variable. For example, DRIVETRAIN has three distinct values: All, Front, Rear. Let "Rear" be the base level; therefore only two dummy variables are created for "All" and "Front". Dummy variables are also created for MAKE and TYPE. Ford and SUV are the base levels.

OUTPUT

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	Horsepower		1	0.6791	0.6791	308.950	897.31	<.0001
2	_MercedesBenz		2	0.0412	0.7203	217.174	62.26	<.0001
3	_Porsche		3	0.0334	0.7537	143.014	57.29	<.0001
		OMITTED RESULTS						
16		_Sedan	14	0.0008	0.8157	23.8055	1.88	0.1708
17	_Toyota		15	0.0021	0.8178	20.9379	4.81	0.0289
18	_Honda		16	0.0023	0.8202	17.6483	5.28	0.0221
19	_GMC		17	0.0013	0.8214	16.7245	2.93	0.0875
20	_Audi		18	0.0012	0.8227	15.9221	2.82	0.0936
21	_Acura		19	0.0015	0.8241	14.6178	3.35	0.0680
22	_Cadillac		20	0.0011	0.8253	14.0453	2.62	0.1065
23	_Lincoln		21	0.0013	0.8265	13.1659	2.94	0.0870
24		_GMC	20	0.0008	0.8257	13.0851	1.96	0.1621
		OMITTED RESULTS						
29		_Honda	21	0.0008	0.8280	9.6793	1.81	0.1793
30		_Toyota	20	0.0006	0.8274	9.0773	1.44	0.2305
31	_Sports		21	0.0016	0.8291	7.3594	3.86	0.0502
32		_Infiniti	20	0.0008	0.8283	7.1926	1.90	0.1686
		OMITTED RESULTS						

2) All-Possible-Regressions Selection Procedure gives all possible models at each step with the suggested independent variable(s) that are associated with the following criteria. Based on these criteria, the analyst subjectively decides the potential independent variables to be included in the model.

- R^2 Criterion– R^2 represents the fraction of the sample variation of the y values that is explained by the independent variables. One drawback of R^2 is adding more independent variables in the model will increase R^2 eventually to 1.
- Adjusted R^2 or MSE Criterion – R^2_{adj} takes into account the sample size and the number of β parameters in the model. R^2_{adj} increases only if MSE decreases. The largest R^2_{adj} or smallest MSE indicates the best fit of the model.
- C_p Criterion – A small value of C_p indicates that the total mean square error and the regression bias are minimized.
- PRESS Criterion – A small PRESS (small differences $y_i - \hat{y}_{(i)}$) value indicates the model has a well predictive ability (Mendenhall and Sincich 328 – 329).

CODE

```
PROC RSQUARE DATA = cars CP ADJRSQ MSE JP ;
MODEL invoice = Cylinders EngineSize Horsepower Length
                MPG_City MPG_Highway Weight Wheelbase;
RUN ;
```

OUTPUT

```
Number in            Adjusted
  Model    R-Square  R-Square      C(p)        J(p)          MSE  Variables in Model

      1      0.6791    0.6783   91.7749   101008007    100536007  Horsepower
      1      0.4163    0.4149  512.5191   183726183    182867649  Cylinders
------------------------------------ OMITTED RESULTS ----------------------------------------
      2      0.7137    0.7123   38.4397  90553268.6     89920029  Horsepower Wheelbase
      5      0.3316    0.3236  656.2243   214394792    211417086  Length MPG_City MPG_Highway
                                                                  Weight Wheelbase
------------------------------------ OMITTED RESULTS ----------------------------------------
      6      0.7394    0.7357    5.1854  83966859.4     82609427  Cylinders EngineSize Horsepower
------------------------------------ OMITTED RESULTS ----------------------------------------
      8      0.7395    0.7346    9.0000  84721616.0     82968755  Cylinders EngineSize Horsepower
                                                                  Length MPG_City MPG_Highway
                                                                  Weight Wheelbase
```

At each step, there is the number of suggested independent variable(s) that contribute to the response variable based on the criteria (R^2, R^2_{adj}, C_p, PRESS).

Caution: Stepwise regression and all-possible-regressions selection procedure typically do not include interactions and higher-order terms in the model (Mendenhall and Sincich 333). These variable screening methods should only be used to assist in identifying the potentially important independent variables for predicting y.

STEP 2. MODEL ADEQUACY

The following criteria are important for checking the utility of the model:

1) Global *F* test: To test the significance of the independent variables as a group for predicting the response variable.
2) $100(1-\alpha)\%$ Confidence intervals and t-tests: Inferences about the β parameters.
3) R^2_{adj}: The total sample variation of the response variable y that is explained by the model after adjusting for the sample size and the number of parameters. Both R^2 and R^2_{adj} are indicators of how well the prediction equation fits the data.
4) Root MSE or s: The estimated standard deviation of the random error. The interval $\pm 2s$ is an approximation of the accuracy in predicting y based on a specific set of independent variables.
5) Coefficient of variation (CV): The ratio of the estimated standard deviation of ε to the sample mean of the response variable $\bar{y}$. Models with CV values of 10% or smaller usually lead to accurate predictions (Mendenhall and Sincich 108).

CODE

```
PROC GLM DATA = cars ;
CLASS DriveTrain Make Type ;
MODEL invoice = Cylinders EngineSize Horsepower Length MPG_City MPG_ Highway
                Weight Wheelbase DriveTrain Make  Type /   SOLUTION CLPARM;
RUN ;
```

OUTPUT

```
                                      Sum of
      Source                   DF       Squares     Mean Square    F Value    Pr > F   ❶

      Model                    52  111086480371  2136278468.7      36.63    <.0001
      Error                   373   21752475640  58317629.061
      Corrected Total         425  132838956010

                ❷   R-Square     Coeff Var      Root MSE    Invoice Mean
                     0.836249      25.42088      7636.598       30040.65
```

❸ Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		-24522.29832	B	12679.92304	-1.93	0.0539
Cylinders		590.26309		783.01737	0.75	0.4514
EngineSize		-1137.65881		1305.37059	-0.87	0.3840
Horsepower		172.61109		13.86506	12.45	<.0001
Length		202.01271		86.25760	2.34	0.0197
MPG_City		794.44503		354.83407	2.24	0.0258
MPG_Highway		-46.41969		314.19327	-0.15	0.8826
Weight		4.84787		1.69433	2.86	0.0045
Wheelbase		-455.96249		148.61460	-3.07	0.0023
DriveTrain	All	-511.08911	B	1577.12766	-0.32	0.7461
DriveTrain	Front	-271.35939	B	1488.93592	-0.18	0.8555
DriveTrain	Rear	0.00000	B	.	.	.

❹ Parameter	95% Confidence Limits	
Intercept	-49455.39268	410.79605
Cylinders	-949.41864	2129.94482
EngineSize	-3704.46683	1429.14921
Horsepower	145.34761	199.87458

---------- OMITTED RESULTS ----------

❶ Global *F* test (P-value < .0001) indicates that model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$ is significant for predicting invoice based on a group of independent variables in the model.
❷ The value of R-square is 0.836249, which means approximately 84% of the variation of invoice is explained by the independent variables. Given Root MSE (s) is 7636.598, approximately 95% of the sampled invoice values fall within two standard deviations (2s = $15,273.20) of their respective predicted values.
❸Based on t-test with the significant level (α) equals 0.10, the p-values for Horsepower, Length, MPG City, Weight, and Wheelbase are less than 0.10 indicating sufficient evidence for predicting the vehicle invoices. Each β_i parameter represents the mean change in the response variable (y) for every 1-unit increase in the corresponding x_i when all the other x's are held fixed. For example, the invoice of a vehicle increases $172.61 for every 1-horsepower increase.
❹ A 95% confidence interval for Horsepower is (145.34761, 199.87458). This means that we are 95% confident that the invoice increases between $145.35 and $199.87 for every 1-horsepower increase. Note: A zero in the 95% Confidence Intervals can also indicate that the independent variable is insignificant.

STEP 3. MODEL ASSUMPTIONS

- Random error $\varepsilon \sim N(0, \sigma^2)$.
- All pairs of random errors are independent.

Using the data to obtain the least squares estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, the error value can be estimated to detect the deviation between the observed and the predicted value of y:

A regression model: $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$
Solve for the error term: $\varepsilon = y - (\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)$
Estimate the error value: $\hat{\varepsilon} = y - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k) = y - \hat{y}$ for each observation.
The least squares prediction equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$ (Mendenhall and Sincich 366).

When assumptions are violated, certain inferences and predictions from the regression analysis may be unreliable and inaccurate.

RESIDUAL TESTS AND DIAGNOSTIC PLOTS

Residual tests and diagnostic plots are commonly used to detect violations in regression modeling assumptions. Such tests and diagnostic plots can also help indicate when a model transformation or modification is needed. Several graphical tools and statistical tests can be applied to detect model "lack of fit", violation of assumptions, invalidity of the inferences, and outliers and influential observations.

1) Residuals and Partial Residuals Plots: Detect model lack of fit and unequal variances.
   Any trends or patterns in the plots indicate lack of fit and potential problems in the model.
2) Normal Probability Plot: Check the assumption of normality.
   A linear trend in the plot suggests that the normality assumption is nearly satisfied; nonlinear trend indicates that the assumption is likely violated.
3) Standardized Residual: An observation with a standardized residual that is larger than absolute value of 3 is considered to be a potential outlier.
4) Influential Observations: Observations that have high impacts on the response variable.

- Leverage: Measures the influence of y_i on its predicted value $\hat{y}_i$. The observed value y_i is influential if $h_i > \frac{2(k+1)}{n}$ where h_i is the leverage for the i^{th} observation and k is the number of β's in the model (excluding β_0).
- Cook's Distance: An influential observation has a value of at least 50th percentile of the *F* distribution.
- Dffits: If $Dffits_i$ is greater than $2\sqrt{\frac{k+1}{n}}$, the i^{th} observation is influential.

  Note: Analysts need to check whether the outlier and influential observations are either correct or entry errors. If the observations are entry errors, either correct it or remove it from the data set for modeling. If the observations are correct, run the model again without those observations to see if the parameter coefficients are unstable. If so, analysts need to decide whether to keep or remove those influential observations for modeling.

5) The Durbin – Watson Test (d): Detect residual correlation.
   Properties of the d statistic:

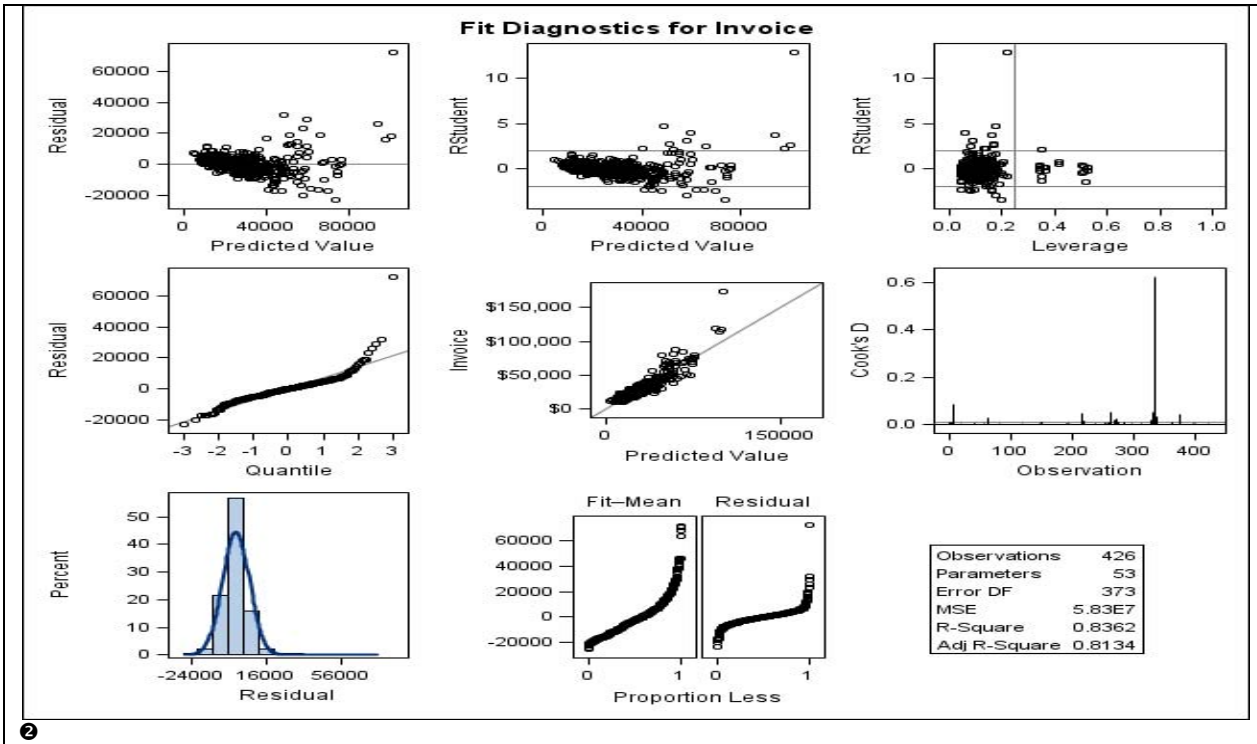
- $0 \leq d \leq 4$
- If $d \approx 2$, residuals are uncorrelated.
- If $d < 2$, $residuals\ are\ postively\ correlated; if\ d \approx 0, the\ correlation\ is\ strong.$
- If $d > 2$, $residuals\ are\ negatively\ correlated; if\ d \approx 4, the\ correlation\ is\ strong.$
  Note: A time series regression model should be considered whenever residuals exhibit strong temporal correlation (Mendenhall and Sincich 369 – 417).

CODE

```
ODS GRAPHICS ON ;
    PROC GLM DATA = cars PLOTS=all ;
    CLASS DriveTrain Make Type ;
    MODEL invoice = Cylinders EngineSize Horsepower Length MPG_City
    MPG_Highway Weight Wheelbase DriveTrain Make Type / P ;
    OUTPUT OUT = stat
    P=pred R=Residual RSTUDENT=r1 DFFITS=dffits
    COOKD=cookd H=hatvalue PRESS=res_del ;
    RUN ;
ODS GRAPHICS OFF ;
```

OUTPUT

```
                    Sum of Residuals                             -0
                    Sum of Squared Residuals           21752475640
                    Sum of Squared Residuals - Error SS           -0
                    First Order Autocorrelation                    0
                ❶   Durbin-Watson D                                2
```



❷

❶ The Durbin-Watson Test (d) = 2 indicating that the residuals are uncorrelated and the independent error assumption is satisfied.
❷ The SAS graphs are a quick way to check the assumptions and to look for outliers and influential observations. The residuals plotted against the predicted values (Row 1, Col 1) show no trends or patterns. If there are any patterns such as the "cone" or "sphere" shapes, this indicates the lack of model fit and unequal variances. We will explore this problem further in **STEP 4. Potential Modeling Problems and Solutions**. Of all the assumptions, the normality assumption is the least restrictive. The Q-Q plot (Row2, Col 1) shows a linear trend with a slight deviation at the tail, which suggests that the normality assumption is satisfied. The histogram (Row 3, Col 1) shows the distribution is mound-shaped with a slightly skewed right tail. Studentized Residual vs. Levarage graph (Row 1, Col 3) shows some potential outliers and influential observations outside of the reference lines. Cook's D graph (Row 2, Col 3) also shows an influential observation that is at least above 50th percentile.

STEP 4. POTENTIAL MODELING PROBLEMS AND SOLUTIONS

When building a multiple regression model, analysts should be cautious of potential problems, many of which are caused by the violation of assumptions. Some of these problems can only be minimized, while others can be fixed to improve the accuracy of the model.

1) Assumptions Violation
   Problem: There are patterns and trends in your residual diagnostics.
   Solution: You may need to transform the response variable to satisfy the assumptions. Transformations are typically used to either (i) help induce the homogeneous (constant) variance assumption, (ii) transform a nonlinear model into an approximately linear model, and/or (iii) change multiplicative effects into additive effects using natural log transformation. The Box-Cox method shown below is helpful in identifying an appropriate transformation for the response variable based on a set of independent variables.

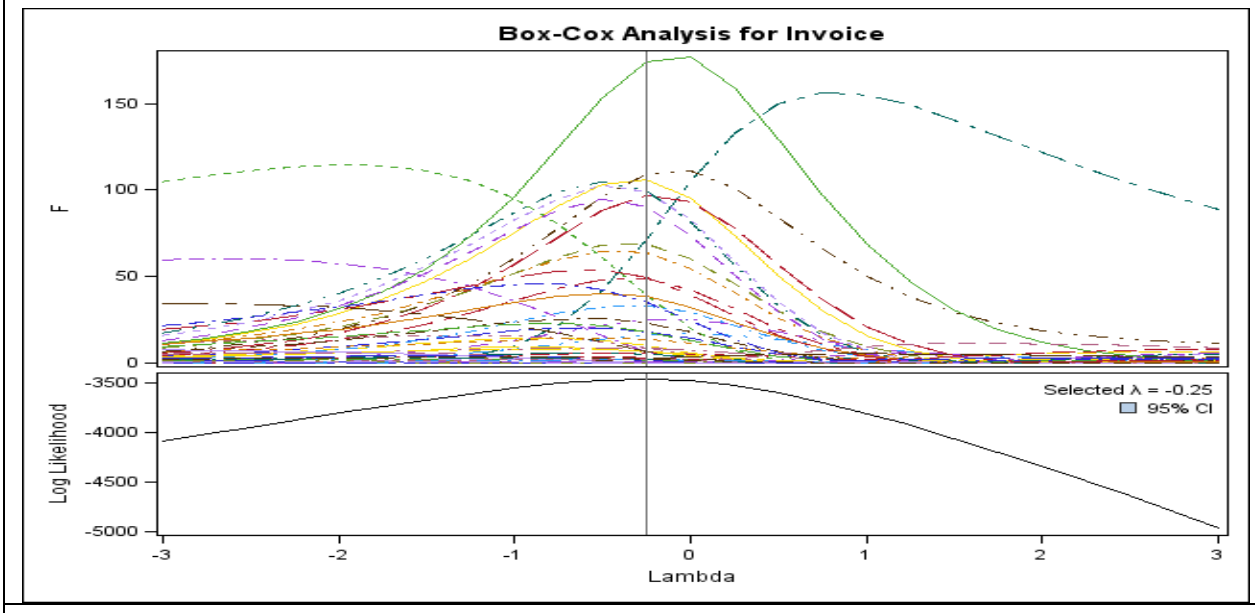
CODE

```
ODS GRAPHICS ON;
  PROC TRANSREG DATA = cars TEST;
  MODEL BOXCOX(invoice) = IDENTITY(Cylinders EngineSize Horsepower Weight
                        Length MPG_City MPG_Highway Wheelbase dummy: );
  RUN;
ODS GRAPHICS OFF;
```

OUTPUT

The TRANSREG Procedure Hypothesis Tests for BoxCox(Invoice)

Root MSE	0.01002	R-Square	0.9350
Dependent Mean	3.68434	Adj R-Sq	0.9260
Coeff Var	0.27188	Lambda	-0.2500 ❶



❶ Based on the Recommended Transformation Chart below, Lambda equals -0.2500 suggesting natural log transformation for invoice (y).

Recommended Transformation	Equation	Lambda
Square	Y^2	1.5 to 2.5
None	Y	0.75 to 1.5
Square-root	$Y^{1/2}$	0.25 to 0.75
Natural log	Ln(Y)	-0.25 to 0.25
Inverse square-root	$1/Y^{1/2}$	-0.75 to -0.25
Reciprocal	1/Y	-1.5 to -0.75
Inverse square	$1/Y^2$	-2.5 to -1.5

("Box-Cox Method")

2) Parameter Estimability
Problem: Highly correlated classification independent variables.
Solution: Remove one of the correlated classification independent variables in the model.

CODE

```
PROC GLM DATA = sashelp.cars  ;
CLASS DriveTrain Make Type ;
MODEL invoice = Cylinders EngineSize Horsepower Length MPG_City
                MPG_Highway Weight Wheelbase DriveTrain Make  Type / E ;
RUN ;
```

OUTPUT

General Form of Estimable Functions

	Case 1	Case 2
Effect	Coefficients	Coefficients
Intercept	L1	L1
Cylinders	L2	L2
EngineSize	L3	L3
Horsepower	L4	L4
Length	L5	L5
MPG_City	L6	L6
MPG_Highway	L7	L7
Weight	L8	L8
Wheelbase	L9	L9
DriveTrain All	L10	L2-L10
DriveTrain Front	L11	L2-L11
DriveTrain Rear	L1-L10-L11	L2-L10-L11

In PROC GLM, A letter "B" always appears next to each parameter estimate that is associated to a categorical variable indicating the estimates are not uniquely estimable. The analyst needs to look at the estimable functions to determine whether or not a linear dependence exists between the categorical variables. In CASE 1, all parameters, including the linear combination of parameters, are estimable. In CASE 2, the parameters for Drivetrain (All, Front, Rear) and Cylinders are not jointly estimable. Either Cylinders or Drivetrain needs to be removed from the model.

3) Multicollinearity

Problem: When independent variables are highly correlated in the model, the results from t-test and *F* test may contradict each other and the parameter estimates may have opposite signs from what are expected.

Solution: Calculate the coefficient of correlation between each pair of numeric independent variables in the model. If one or more correlation coefficients are close to 1 or -1, the variables are highly correlated and a severe multicollinearity problem may exist; remove one of the correlated independent variables in the model.

CODE

```
PROC CORR DATA = cars ;
VAR Cylinders EngineSize Horsepower Length MPG_City MPG_Highway Weight Wheelbase ;
RUN ;
```

OUTPUT

	Cylinders	Engine Size	Horse power	Length	MPG City	MPG Highway	Weight	Wheel base
Cylinders	1	0.908	0.81	0.548	-0.684	-0.676	0.74	0.547
Engine Size	0.908	1	0.787	0.637	-0.709	-0.717	0.81	0.637

Horse power	0.81	0.787	1	0.382	-0.677	-0.647	0.63	0.387
Length	0.548	0.637	0.382	1	-0.502	-0.466	0.69	0.889
MPG City	-0.684	-0.709	-0.677	-0.502	1	0.941	-0.74	-0.507
MPG Highway	-0.676	-0.717	-0.647	-0.466	0.941	1	-0.79	-0.525
Weight	0.742	0.808	0.631	0.69	-0.738	-0.791	1	0.761
Wheel base	0.547	0.637	0.387	0.889	-0.507	-0.525	0.76	1

In Step 2. Model Adequacy, the parameter estimate for MPG Highway is -46.41969. It should be positive because MPG supposedly adds values to a vehicle. In the correlation matrix, it shows that MPG City and MPG Highway are highly correlated (0.941) because both variables measure mileage per gallon (MPG).

- A severe multicollinearity problem exists if the variance inflation factors (VIF) for the $\beta's$ are greater than 10.

CODE

```
PROC REG DATA = cars ;
MODEL invoice = Cylinders EngineSize Horsepower Length MPG_City
                MPG_Highway  Weight Wheelbase dummy: / VIF ;
RUN ;
```

OUTPUT

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-32811	12311	-2.67	0.0080	0
Cylinders		1	590.26309	783.01737	0.75	0.4514	10.85209
EngineSize	Engine Size (L)	1	-1137.65881	1305.37059	-0.87	0.3840	15.12228
Horsepower		1	172.61109	13.86506	12.45	<.0001	7.26089
Length	Length (IN)	1	202.01271	86.25760	2.34	0.0197	11.19161
MPG_City	MPG (City)	1	794.44503	354.83407	2.24	0.0258	25.27722
MPG_Highway	MPG (Highway)	1	-46.41969	314.19327	-0.15	0.8826	23.80516
Weight	Weight (LBS)	1	4.84787	1.69433	2.86	0.0045	12.07993
Wheelbase	Wheelbase (IN)	1	-455.96249	148.61460	-3.07	0.0023	11.16877
_Acura		1	9236.02746	3464.72512	2.67	0.0080	1.41723

---------- OMITTED RESULTS ----------

In addition to the correlation matrix, MPG City and MPG Highway have the highest variance inflation (25.27722 & 23.80516) indicating a severe multicollinearity. One of the MPG variables should be removed from the model.

4) Extrapolation
Problem: Predicting y outside of the range of the independent variables may give inaccurate results.

5) Data issues
   Problem: Missing and/or invalid data values for certain days.
   Solution: Use dummy variables in the model to account for those days.

STEP 5. MODEL VALIDATION

Models that fit the sample data well may not be statistically useful when applied to a new data set because of changes or unexpected events that may occur in the future. In addition to checking the model adequacy, it is important to validate the model's performance in practice. The following techniques have been proposed for the model validation.

1) Examine the predicted values: If the predicted values seem unreasonable such that the values are extremely outside of the range of the response variable, this indicates that either the model is incorrect or the parameter coefficients are poorly estimated. If the predicted values seem reasonable, continue to check the model validity.
2) Examine the model parameters: Coefficients are poorly estimated and/or multicollinearity exists if they are opposite signs to what are expected, have unusual large or small values, and/or are inconsistent when applied to new data.
3) Apply the model to the new data for prediction: Use $\mathrm{R}^2_{\mathrm{prediction}}$ and MSE to measure the model validity.
4) Perform data-splitting: The sample data can be split into two parts with one part used to estimate the model parameters and other part used to validate the predictions.
5) Perform Jackknifing for data sets with small sample sizes: Use $\mathrm{R}^2_{\mathrm{jackknife}}$ and MSE to measure the model validity (Mendenhall and Sincich 307 – 309).

CONCLUSION

A multiple regression model is commonly used because it is not as complex as other statistical models. Yet it is the most abused model because analysts overlook the assumptions, fail to minimize or fix potential problems in the model, and do not validate the model's predictions. It is crucial to follow all these steps as a check list when building a multiple regression model. However, even following all these steps may still not produce the best most useful regression model when the underlying data does not meet the conditions that make linear regression the appropriate model to fit the data. For example, a time series regression equation is often more appropriate for modeling seasonal data. Likewise, a logistic regression model should be used for modeling binary, ordinal, or nominal response variables. The rule of thumb is analysts need to check the assumptions, look for potential problems, and validate the model accuracy and prediction reliability for all statistical models.

REFERENCES

Mendenhall, William and Terry Sincich. A Second Course in Statistics Regression Analysis. New Jersey: Pearson Education, Inc., 2003.

"The TRANSREG Procedure: Box-Cox Transformations". 2011. SAS/STAT® 9.2 User's Guide, Second Edition. October 2011
<http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_transreg_sect015.htm>

"Box-Cox Method". 2008. Chapter 6 – Analysis of Experiments. October 2011
<http://www.weibull.com/DOEWeb/box_cox_method.htm>.

"The Four Types of Estimable Functions." SAS/STAT® 9.2 User's Guide. Page 277.
<http://support.sas.com/documentation/cdl/en/statugestimable/61763/PDF/default/statugestimable.pdf>.

TRADEMARKS

ACKNOWLEGMENTS

I am thankful to my mentor, Art Carpenter, for taking time out of his busy schedule to read my paper and provide guidance and suggestion. I owe a special debt of appreciation to Dr. Scott Lesch and Dr. Daniel Jeske who are always there to educate me on Statistics.

AUTHOR CONTACT

Please contact the author for any questions and comments.

Theresa Hoang Diem Ngo

theresa.ngo1120@gmail.com

APPENDIX

The **CORR Procedure** computes Pearson correlation coefficients, which measure a linear relationship between two numeric variables.

```
PROC CORR DATA = dataset1 ;
VAR numeric variables ;
RUN ;
```

The **GLM Procedure** uses the method of least square to fit general linear models.

```
ODS GRAPHICS ON ;
    PROC GLM DATA = dataset1 PLOTS=all ;
    CLASS classification independent variables;
    MODEL dependent variable = independent variable(s)/ SOLUTION E CLPARM P ;
    OUTPUT OUT = dataset2
    P=pred R=Residual RSTUDENT=r1
    DFFITS=dffits COOKD=cookd H=hatvalue PRESS=res_del ;
    RUN ;
ODS GRAPHICS OFF ;
```

PROC GLM <options> ;

DATA = *SAS data set*
Specifies the SAS data set used for GLM procedure.

PLOTS = *all*
Plots all of the default plots such as residual, histogram, and normality plots. ODS graphics must be enabled to produce the plots. For example,

```
ODS GRAPHICS ON ;
    PROC GLM  PLOTS = ALL ;
    MODEL dependent variable = independent variables ;
    RUN ;
ODS GRAPHICS OFF;
```

MODEL *dependent variable* = *independent variables* **</ options>** ;

CLPARM	Produces confidence intervals for parameter estimates.
E	Displays the general form of all estimable functions.
P	Displays observed, predicted, and residual values for each observation given that the independent variables do not have missing values. Most importantly, the Durbin-Watson statistic is also displayed.
SOLUTION	Produces parameter estimates.

Output statement creates a new SAS data set that contains all of the variable in the original data set and new variables for the specified statistics.

OUTPUT < OUT = *SAS data set* **> <statistics>** ;

COOKD	Cook's D influence statistic
DFFITS	Standard influence of observation on predicted value
H	Leverage
P	Predicted values
PRESS	Residual for the ith observation that results from dropping it and predicting it based on all other observations.
R	Residuals
RSTUDENT	A studentized residual with the current observation deleted.

The **REG Procedure** is used for regression analysis.

```
PROC REG DATA = dataset1 ;
MODEL dependent variable = independent variable(s)/ VIF SELECTION=stepwise ;
RUN ;
```

SELECTION = method
Specifies the method used to select the model; a selection method can be FORWARD, BACKWARD, STEPWISE, MAXR, MINR, RSQUARE, ADJRSQ, or CP.

VIF Computes variance-inflation factors.

The **RSQUARE Procedure** computes the statistics for each model selection.

```
PROC RSQUARE DATA = dataset1 CP ADJRSQ MSE JP;
MODEL dependent variable = independent variable(s) ;
RUN ;
```

CP	Computes Mallows' C_p statistic for each model selected.
JP	Computes the mean square error of prediction for each model selected.
MSE	Computes the mean square error for each model selected.
ADJRSQ	Computes adjusted r-square for each model selected.

The **TRANSREG Procedure** fits many types of linear models with many options of transformations.

```
ODS GRAPHICS ON;
  PROC TRANSREG DATA = dataset1 TEST;
  MODEL BOXCOX(dependent variable) = IDENTITY(numeric independent variables);
  RUN;
ODS GRAPHICS OFF;
```

The **MODEL** statement specifies the dependent and independent variables along with the transformation applied to each variable. In this case, a **Box-Cox** transformation is only applied to the numeric response variable based on the numeric independent variables specified in the **IDENTITY** transformation. Note that **IDENTITY** is used for the purpose of no transformations applied to the independent variables. **TEST** option displays the ANOVA table.