

Develop a pricing model: Price of cars in terms of performance variables found in Road & Track Magazine

Source of Data: October 1988 issue of Road & Track magazine p. 26-28

Reprinted with permission from Road & Track magazine. Copyright © 1988 by Hachette

Filipacchi Magazines, Inc.

Solution Type: Tactical

Solution Model: Pricing

Otis B. Smith (osmith@appliedmethods.net)

Applied Methods Inc.

1101 East 33rd St. 3rd Floor, Suite B-304

Baltimore, MD 21218

O: 410-730-9661

Abstract

Occasionally autos are provided for testing, but prices have not been set. The pricing model will be used to provide estimated prices for auto dealers. In this project, I focus on the interactions between the price as tested variable and several performance variables. To this end, I begin the modeling process by focusing on price as a function of eight controllable variables that represent the performance results of cars. The goal of this project is to not only produce a pricing model but to produce the optimal pricing model from a set of models. During the modeling process, I show that some predictor variables are related to other predictor variables and therefore they are not included in the final model. I also show that the final model explains 66% of the variance in the prices of cars.

Important Terms and Definitions

Variance: The variance is the measure of dispersion of a given random variable about its mean. $\text{Var}(X) = \sigma^2 = E(X-\mu)^2$

Standard Deviation: The standard deviation is the square root of the variance.

$SD(X) = \sigma = \text{Square Root}(\text{Var}(X))$

Coefficient of Variance (CV): A relative measure of dispersion that expresses the sample Standard Deviation in terms of the sample mean. $CV = \text{sample SD (s)} / \text{Sample Mean (x)}$

Skewness: The lack of symmetry of the distribution $\beta_3 = E(X - \mu)^3 / \sigma^3$. Skewness is 0 for symmetric distributions and is positive or negative depending on whether the distribution is negatively skewed or positively skewed.

Kurtosis: The tail-heaviness (amount of probability in the tails) of the distribution. $\beta_4 = E(X - \mu)^4 / \sigma^4$. A normal distribution (in non-mathematical parlance – a bell curve) has a $\beta_4 = 3$.

Covariance: The covariance of two random variables, X and Y, measures the joint dispersion from their respective means. If X and Y are independent, then their covariance is zero. $Cov(X,Y) = \sigma_{xy} = E [(X - \mu_x) (Y - \mu_y)]$

Correlation Coefficient: A measure of the linear relationship between two random variables X & Y. $\rho_{xy} = Corr(X,Y) = Cov (X, Y) / \text{Square Root} (\text{Var}(X)*\text{Var}(Y))$. The Correlation Coefficient range is [-1, +1]

Multicollinearity: A condition that exists when the columns of the predictor variables are exactly or approximately linearly dependent. Multicollinearity can lead to the following problems: 1. The coefficient estimates are subject to numerical errors and are unreliable. 2. Most of the coefficients have very large standard errors.

Variable Inflation Factor (VIF): A measure of multicollinearity. A large VIF suggests multicollinearity. $VIF_j = 1 / (1 - r_j^2)$ r squared is the coefficient of multiple determination when regression x_j on the other $k-1$ predictor variables.

Cooks Distance (Cooks D): A method of detecting influential observations. It measures the distance between the predicted coefficient and the Least Squares estimate of coefficient. The betas are the estimated regression coefficients. $d_i = (\beta_i - \hat{\beta})' X'X (\beta_i - \hat{\beta}) / (k + 1)s^2$ ($i = 1, 2, \dots, n$)

Stepwise Regression: A method of determining the marginal contribution of each variable by entering or removing variables one at a time.

Backward Regression: A method of selecting the optimal model that begins with the maximum model and deletes variables of no value. Each step corresponds to examining a set of added-last tests for the current model.

Cp Statistic: The measure of the predictive ability of a fitted model. The subscript p corresponds with the number of predictor variables in the model. $C_p = \text{Sum of Squared Error (SSE}_p) / \text{estimated } \sigma^2 + 2(p + 1) - n$. Since a full model is assumed to have zero bias in its predictive capacity, then the optimal model will have the smallest C_p .

P-value: An observed level of significance. The smaller the P-value, the more significant is the test result. It is the probability under the null hypothesis of obtaining a test statistic at least as extreme as the observed value.

R-Square: A ratio that represents the proportion of variation in y that is accounted for by regression on x. $r^2 = \text{Regression sum of squares (SSR)} / \text{Total sum of squares (SST)}$

Adjusted R-square: A criterion used to measure the maximized predictive power of the model $r^2 = 1 - \text{SSE}_p / (n - (p + 1)) / \text{SST} / (n-1)$

Problem Definition

Given the performance data from Road & Track magazine, develop a predictive pricing model. The following information reveals the number observations, the variables, and parameters.

Number of Observations (N): $N = 105$

Modeling/Target Variable (Y): $Y = \$\text{Price}$ as tested of car

Number of initial Predictor Variables (X): $X = 8$

X1: Zero to 60 mph (sec)

X2: Quartet (Time for standing start $\frac{1}{4}$ mi (sec))

X3: Top speed (mph)

X4: Brake 80 (Distance to brake from 80 mph (ft))

X5: Slalom (Speed through slalom course)

X6: Skidpad (Cornering ability (G))

X7: Noise (Interior noise at 70 mph (dba))

X8: MPG (Mile per gallon (fuel efficiency))

Three Diagnostics describe the status of the data in standard statistical terms and form the essential part of the Problem Definition phase.

Level 0 diagnostics: This diagnostic provides a list of the X variables, the Y variable, and the observations. The purpose of this diagnostic is to reveal the units of measurement for each X variable and to reveal each observation. I did not notice any unusual observations from the Level 0 diagnostics.

Level 1: This diagnostic reveals the minimum, maximum, and any missing values (Nmiss) for each variable. Level 1 diagnostic shows that seven variables have missing values.

Price as tested (Y): $N_{\text{miss}} = 3$

Top speed (X): $N_{\text{miss}} = 18$

Brake Distance (X): $N_{\text{miss}} = 4$

Slalom: $N_{\text{miss}} = 7$

Skidpad: $N_{\text{miss}} = 5$

Noise: $N_{\text{miss}} = 15$

MPG: $N_{\text{miss}} = 3$

Level 2: This diagnostic reveals the distribution (Variance, Standard Deviation, Coefficient of Variance) and a description (skewness & kurtosis) of the distribution of the data for each variable. The following is a summary of the distributions and descriptions for each variable. Variables “price as tested”, “zero to 60”, & “top speed”, have the most variance with respect to their means.

X/Y	Skewness	Kurtosis	CV (variance about mean)
Y (Price as tested)	Positive	Tail Heavy >3	84.19

X1: Zeroto60	Positive	Light Tail <3	23.7
X2: Quartert	Negative	Light Tail <3	8.33
X3: Topspeed	Positive	Light tail <3	13.69
X4: Brake80	Positive	Light Tail <3	8.07
X5: Slalom	Negative	Light tail <3	3.62
X6: Skidpad	Negative	Light tail <3	5.89
X7: Noise	Positive	Light tail <3	4.07
X8: MPG	Positive	Light tail <3	24.4

Level 3: This diagnostic reveals whether there is a relationship between each X variable and the Y variable. The Pearson Correlation Coefficients for “Price as tested” and the X variables reveal that “Top Speed” has the strongest relationship with Price as tested” (corr =.7). “Slalom” appears to have the weakest relationship with “Price as tested” (corr=.023). Predictor variable “Quartert” appears to have a strong relationship with predictor variable “Skidpad” and this relationship is an inverse relationship (corr =-.71) as well as “Skidpad” and “Topspeed” (corr=.67), “Skidpad” and “Slalom” (corr=.65), and “Skidpad” and “Zeroto60” (corr=.68).

The following information reveals the results of the Diagnostic Plots:

Y – Price as Tested

Box Plot: reveals outliers

Normal: reveals a linear pattern and thus supports the normality assumption

Stem & Leaf: reveals three outliers

X1 – Zero to 60

Box Plot: reveals outliers

Normal: reveals a linear pattern and thus supports the normality assumption

Stem & Leaf: reveals three outliers

Y vs X: reveals a logarithmic curve pattern

X2 – Quartert

Box Plot: reveals outliers

Normal: reveals a linear pattern and thus supports the normality assumption

Stem & Leaf: reveals a distribution that is close to normal

Y vs X: reveals a logarithmic curve pattern

X3 – Top Speed

Box Plot: reveals outliers

Normal: reveals a linear pattern and thus supports the normality assumption

Stem & Leaf: reveals two outliers

Y vs X: reveals a linear pattern

X4 – Distance to brake from 80 feet

Box Plot: reveals outliers

Normal: reveals a linear pattern and thus supports the normality assumption

Stem & Leaf: reveals one outlier

Y vs X: reveals a logarithmic pattern

X5 – Speed through the slalom

Box Plot: no outliers

Normal: reveals a linear pattern and thus supports the normality assumption

Stem & Leaf: reveals one outlier

Y vs X: reveals a possible linear pattern

X6 – skidpad

Box Plot: reveals outliers

Normal: reveals a linear pattern and thus supports the normality assumption

Stem & Leaf: reveals three outliers

Y vs X: reveals a linear pattern

X7 – Noise

Box Plot: reveals outliers

Normal: reveals a linear pattern and thus supports the normality assumption

Stem & Leaf: reveals one outlier

Y vs X: reveals a linear pattern

X8 – MPG

Box Plot: reveal outliers

Normal: reveals a linear pattern and thus supports the normality assumption

Stem & Leaf: reveals three outliers

Y vs X: reveals a logarithmic pattern

Pricing Model Design

The design phase involves analyzing the different variables to determine whether the data follows a linear or non-linear pattern. In this regard, I regressed “Price as tested” on “zero to 60”, “Quartert”, “Topspeed”, “Brake80”, “Speed through slalom”, “skidpad”, “Noise”, and “MPG”. The residual versus predicted plot did not reveal a pattern. The randomness of the plot indicates that the model has accounted for linearity. Because the data conforms to a linear pattern as demonstrated by the lack of patterns in the residual versus predicted plot, I added neither quadratic terms nor performed transformations. During the next phase, Synthesis of Variables, I check all eight variables to determine significance with respect to price and to determine whether relationships exist between the predictor variables. The final model must have variables that are mutually exclusive predictors of price.

Synthesis of Variables

The Synthesis of Variables phase includes checks for Multicollinearity (two are more predictor variables are acting together to influence price) and Influence. In order to check for Multicollinearity, I computed the Variable Inflation Factor (VIF) for each variable. In order to check for influential observations, I computed CooksD and plotted the studentized residual for each variable. With respect to collinearity, predictor variables “Zeroto60” and “Quartert” both produced VIFs >10. This result suggests that a

Multicollinearity probably exist. With respect to influence, both CooksD and studentized residuals revealed that observation #43 (CooksD =.7 & studentized residual = approx. 2) might influence the estimate of regression coefficients. Because observation #43 was below thresh holds for both the CooksD and studentized residual results, I did not remove observation #43. Because of the high VIF for both “Zero60” and “Quarter” predictor variables, I decided to conduct an added in order test to identify the predictor that does not add value.

The added in order test helped me to select the following predictor variables based on p-values less than .05 with the exception of one predictor variable with a p-value of .13. The following is a list of the predictor variables for the final model and their p-values from the added in order test:

Predictor variables	Added in order p-values
Top speed	.008
Brake80	.02
Slalom	.002
MPG	.009
Noise	.13

With a clear understanding of the variables that are both influential and mutually exclusive predictors of price as tested, I am ready to build multiple models and compare these models to determine the optimal pricing model.

Modeling and Optimization

Both Stepwise and Backward regression at a significance level of .10 agree that the best model includes the following predictors: Top Speed, Brake80, Slalom, MPG, and Noise. The Cp statistic for this model is 6. The runner-up model at a significance level of .10 with Stepwise regression includes the following predictors: Top Speed, Slalom, MPG, & Noise. The runner-up has a Cp statistic of 9.68. I selected the best model after completing both the Stepwise and Backward regression computations. Thus, the best model has the smallest Cp statistic from these computations.

Decision Making

After I selected the best model and the runner-up model, I plotted the residual vs. predicted for both models and examined the VIF for each parameter. Both models produced similar results.

Customized Pricing Model

The best model: $Price = 166537 + 625.37 \text{ topspeed} - 226.59 \text{ brake80} - 3482.80 \text{ slalom} - 1529 \text{ MPG} + 1215.6 \text{ noise}$.

The following is an interpretation of the beta parameters:

For every 1mph increase in topspeed, price as tested increases on average by \$625.37

For every foot increase in breaking distance, price as tested decreases on average by \$226.59

For every sec increase in slalom course performance, price as tested decreases on average by \$3482.80.

For every increase in miles per gallon, price as tested decreases on average by \$1529
 For every decibal increase in interior noise at 70mph, price as tested increases on average by \$1215.6.

Given an R-square of .6661, the model accounts for 66.6% of the variance in the response variable, price as tested.

The SAS System
The REG Procedure
Model: MODEL1
Dependent Variable: Price Price as tested (\$)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	18780265343	3756053069	25.93	<.0001
Error	65	9415980479	144861238		
Corrected Total	70	28196245822			

Root MSE	12036	R-Square	0.6661
Dependent Mean	29105	Adj R-Sq	0.6404
Coeff Var	41.35354		

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	Intercept	1	166537	58922	2.83	0.0062	0
Topspeed	Top speed	1	625.37153	155.27794	4.03	0.0002	2.83375

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
brake80	Distance to brake from 80mph (feet)	1	-226.58849	95.04621	-2.38	0.0201	1.66426
Slalom	Speed through slalom course	1	-3482.80971	796.99610	-4.37	<.0001	1.49284
MPG	Miles per gallon, fuel efficiency	1	-1529.02068	546.57098	-2.80	0.0068	1.94315
Noise	Interior noise at 70 mph (dba)	1	1215.62044	554.69585	2.19	0.0320	1.37528

The REG Procedure

Price = 166537 +625.37 topspeed -226.59 brake80 -3482.8 slalom -1529 MPG +1215.6 noise

