Department of Statistics
The Wharton School
University of Pennsylvania

Statistics 621 Fall 2003

Module 4

The Multiple Regression Model

Example: Explaining and predicting fuel efficiency

The file car89.jmp contains many characteristics of various makes and models of cars. Variables include:

> MPG City, Make/Model, Weight, Cargo, Seating, Horsepower, Displacement, Number of cylinders, Length, Headroom, Legroom, Price…

Questions of interest

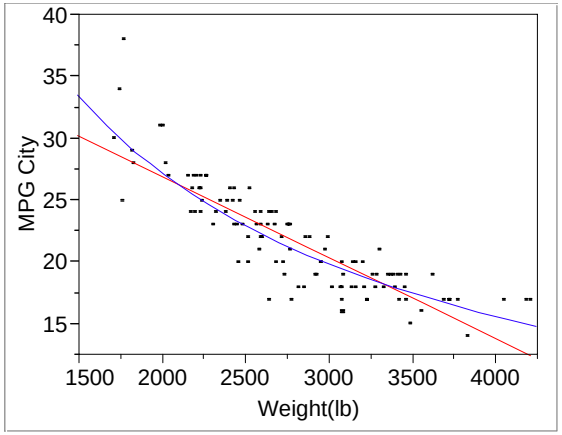
> "What is the predicted mileage for a 4000 lb. new design, and which characteristics of the design are crucial?"[1]
>
> "How much does my 200 pound brother owe me for 3000 miles of riding with me?"

To get started, let's consider using simple regression to model the effect of Weight (lb) on MPG City

[1] Such questions of mileage are important to manufacturers that sell cars in the US. The so-called CAFE standards set requirements for the average fuel efficiency of the fleet of cars produced by a manufacturer.

Applying Fit Y by X, we consider the regression of MPG City on Weight(lb) (using Fit Line) and the regression of (p 110)

(1/MPG City) on Weight(lb)[2]



Linear Fit

MPG City = 40.11 - 0.00655 Weight(lb)

Transformed Fit Recip

Recip(MPG City) = 0.00943 + 0.0000136 Weight(lb)

Which of these regressions seems more reasonable?

Do the signs of the slope coefficients make sense here?

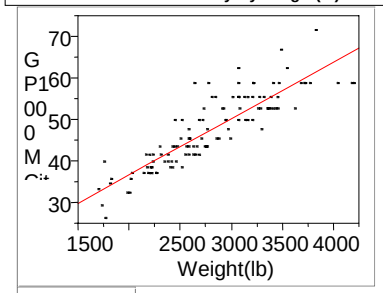
[2] Use the fit special dialog to get the reciprocal 1/Y of the response.

Based on the previous regressions we created a "new", rescaled dependent variable[3]

GP1000M = 1000/MPG

The regression of GP1000M on Weight(lb) yields (p 111)

Bivariate Fit of GP1000M City By Weight(lb)



Linear Fit

Linear Fit

GP1000M City = 9.4323391 + 0.0136159

Summary of Fit

RSquar	0.76526
RSquare	0.76313
Root Mean Square	4.23319
Mean of	47.5946
Observations (or Sum	112

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	9.432339	2.05451	4.59	<.0001
Weight(lb)	0.013615	0.00071	18.94	<.0001

What is the interpretation of the LS regression slope here?

Is the only difference between the BMW 735i and the Suzuki Swift just the 2000-pound different in weight?

[3] Multiplying 1/MPG by 1000 serves only to multiply the intercept and slope estimates by 1000 resulting in "friendlier" (and more impressive) regression output. What other easily motivated change of scales would make the slope 2000 times larger still?

Other factors obviously contribute as well. Let's use the Multivariate command to explore the pairwise relationships between some of these. (p 115-116)

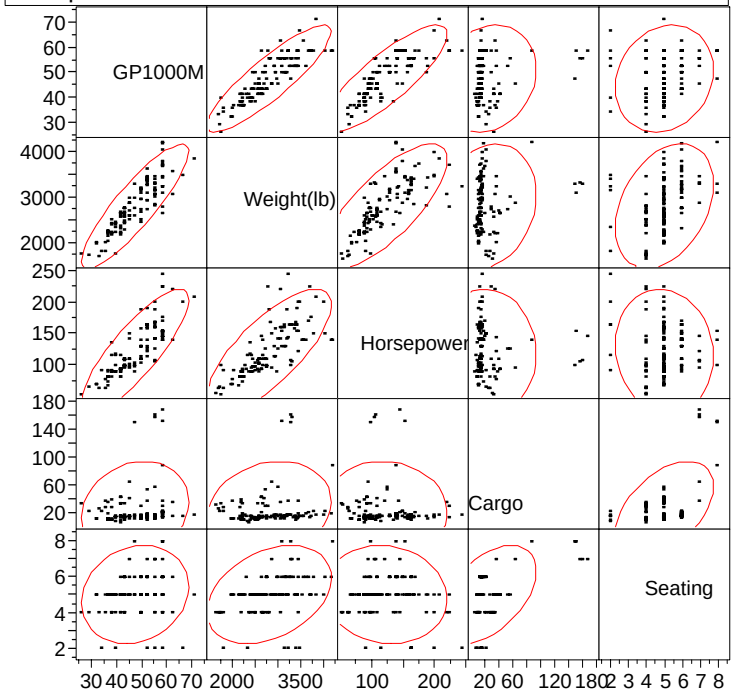
Multivariate

Correlations

	GP1000M City	Weight(lb)	Horsepower	Cargo	Seating
GP1000M City	1.0000	0.8798	0.8334	0.1672	0.1620
Weight(lb)	0.8798	1.0000	0.7509	0.1816	0.3499
Horsepower	0.8334	0.7509	1.0000	-0.0548	-0.0914
Cargo	0.1672	0.1816	-0.0548	1.0000	0.4894
Seating	0.1620	0.3499	-0.0914	0.4894	1.0000

7 rows not used due to missing values.

Scatterplot Matrix



The multivariate command provides a correlation matrix and scatterplot matrix[4] for all pairwise relationships between the five variables GP1000M, Weight, Horsepower, Cargo and Seating.

Besides Weight, which variable is appears most strongly associated with GP1000M?

To consider the *joint* effect of Weight and Horsepower on GP1000M, we apply the Fit Model command[5] to obtain the multiple regression output (p 118)

Response GP1000M

Summary of

RSquare	0.841022
RSquare Adj	0.838105
Root Mean Square Error	3.499726
Mean of Response	47.59468
Observations (or Sum Wgts)	112

Parameter

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	11.684254	1.727038	6.77	<.0001
Weight(lb)	0.0089183	0.000882	10.11	<.0001
Horsepower	0.0883837	0.012264	7.21	<.0001

To interpret this output, let's first describe the underlying multiple regression model.

[4] The density ellipses in each of these plots are estimates of the highest density population regions under the assumption of joint normality. Note how these ellipses guide your eye towards the strongest linear associations.

[5] Fit Y by X in JMP only performs simple regressions. To fit a multiple regression, use Fit Model. Here we select GP1000M as Y and add Weight and Horsepower to the Model Effects box in the dialog used to specify the multiple regression.

The Multiple Regression Model (MRM)

A model for the relationship between

y - a dependent variable or response, and

$x_1,\ldots, x_K$ - a set of independent variables, explanatory variables or predictors

Denote the n observations of the K+1 terms $y, x_1, \ldots, x_K$ by

$$y_i, x_{1i},\ldots, x_{Ki}, \qquad i = 1,\ldots,n$$

Under the MRM, the data is assumed to be a realization of

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \varepsilon_i, \qquad i = 1,\ldots,n$$

$$\varepsilon_1,\ldots,\varepsilon_n \text{ iid} \sim N(0,\sigma_\varepsilon^2)$$

Pictorially

$K = 1$

$K = 2$

$K \geq 3$, hyperplane

Remark: Even for $K > 1$, $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K$ is usually just called the regression line.

Some key interpretations:

$\beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki}$

β_0

β_k for $k = 1,\ldots, K$ (Careful!)

σ_ε

$\beta_0, \beta_1, \ldots, \beta_K$ and σ_ε are the (usually) unknown parameters of the MRM. An objective of regression is to estimate them.

The Least Squares (LS) Regression

In order to estimate the *"true" regression*

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K$$

we use the *least squares (LS) regression*

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_K x_K$$

which has the property of minimizing the sum of squared vertical distances from the plane to the data

The values of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K$ are calculated by computer programs such as JMP which insert the data into formulas, (*which if you must know, we'll tell you during office hours*).

The values $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K$ are called the *least squares (LS) estimates* of $\beta_0, \beta_1, \ldots, \beta_K$

Partial versus Marginal Regression Coefficients

Returning to the previous regressions, let

y = GP1000M, x_1 = Weight and x_2 = Horsepower

From the output on p 4-5, we can see that the LS regression

GP1000M = 11.68 + 0.00891 Weight(lb) + 0.0883 Horsepower

estimates the "true" regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

In this model, β_1 is called a *partial regression coefficient.*

The interpretation of $\hat{\beta}_1 = 0.00891$ here is

In contrast, the LS regression line on p 4-3,

GP1000M = 9.43 + 0.01362 Weight(lb)

is an estimate of the "true" regression line

$$y = \beta_0 + \beta_1 x_1$$

In this model, β_1 is called a *marginal regression coefficient*.

The interpretation of $\hat{\beta}_1 = 0.01362$ here is

What is the essential difference between partial and marginal regression coefficients?

To get some insight into what is going on, we note that a simple regression of Horsepower on Weight yields (p 120)

Horsepower = –26.10 + 0.0533 Weight

Substituting this expression for Horsepower into the multiple regression yields

GP1000M = 11.68 + 0.00891 Weight + 0.0883 Horsepower

= 11.68 + 0.00891 Weight + 0.0883 (–26.10 + 0.0533 Weight)

= 9.43 + 0.01362 Weight

which is just the previous simple regression!

A "graphical view with nodes and edges" provides a convenient representation of what's going on.

"How much does my 200 pound brother owe me for 3000 miles of riding with me?"[6]

[6] Is there ever a context in which you would rather have the marginal coefficient? Yes. Suppose you only know the weight of a car. Which slope would help you estimate its fuel consumption?

Inference about $\beta_0, \beta_1, \ldots, \beta_K$

Tests and confidence intervals used in simple regression generalize naturally to multiple regression.

Yet another "astonishing fact" (which is probably not so surprising at this point)

Under the MRM, the sampling distributions of $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_K$ are normal with means $\beta_0, \beta_1, \ldots, \beta_K$

Along with the estimates $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_K$, programs such as JMP report their standard errors SE($\hat{\beta}_0$), SE($\hat{\beta}_1$),..., SE($\hat{\beta}_K$)

Confidence Intervals for β_k

Approximate 95% CI's for $\beta_0, \beta_1, \ldots, \beta_K$ are given by

Hypothesis Tests for β_k

For testing the null hypothesis H_0: [$\beta_k = c$ in the fitted model] vs

H_1: [$\beta_k \neq c$ in the fitted model], t ratio = $\dfrac{\hat{\beta}_k - c}{SE(\hat{\beta}_k)}$

Hypotheses of the form H_0: [$\beta_k = 0$ in the fitted model] are usually of most interest. Why?

If $|t \text{ ratio}| > 2$ *or* p-value < .05 *or* 95% CI does not contain c, reject H_0 at the .05 level of significance.

Example

JMP provides t ratios and p-values for testing

Suppose we consider adding the variables Cargo and Seating to the car89 regression. What would you conclude about the effect of either addition from the following output? (p 125)

Response GP1000M

Summary of

RSquare	0.852239
RSquare Adj	0.846556
Root Mean Square Error	3.411697
Mean of Response	47.67511
Observations (or Sum Wgts)	109

Parameter

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	12.930547	2.020835	6.40	<.0001
Weight(lb)	0.0091318	0.001159	7.88	<.0001
Horsepower	0.0857712	0.01509	5.68	<.0001
Cargo	0.0346363	0.013277	2.61	0.0104
Seating	-0.476467	0.412437	-1.16	0.2506

Note that if Seating is removed here, the other t ratios and p-values will change.

WARNING! Used properly, the t ratios justify removing *at most one variable at a time*. Regression must then be rerun to get a new set of t ratios.

The Fitted Values and the Residuals

As in simple regression, the LS regression line again serves to decompose the data into the fitted values and the residuals

$$y_i = \hat{y}_i + e_i$$

where

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_K x_{Ki} \quad \text{and} \quad e_i = y_i - \hat{y}_i$$

Root Mean Squared Error (*RMSE*) – An Estimate of σ_ε

When the MRM holds, σ_ε is estimated by[7]

$$RMSE = \sqrt{\frac{1}{n-K-1}\sum (y_i - \hat{y}_i)^2}$$

For example, in the car89 regression output on p 4-12, *RMSE* is given by Root Mean Square Error = 3.41.

As in a simple regression, *RMSE* is also called the *standard deviation of the residuals* and measures the dispersion of the residuals about the LS regression line.

It again measures the predictive accuracy of the model used to forecast values for new cases.

[7] $RMSE^2$ is the "average" sum of squared deviations from the regression line. We divide by $(n - K - 1)$ instead of n to compensate for the fact that the LS line always obtains a smaller sum of squared deviations than the true regression line.

R-square and Adjusted R-square

As in simple regression, the multiple regression decomposition of the response into "signal" plus "noise" ($y_i = \hat{y}_i + \hat{e}_i$) satisfies the amazing identity

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

namely

$$Total\ SS = Model\ SS + Residual\ SS^8$$

Here, too, R^2 gives "the proportion of the total variation explained by the regression", namely

$$R^2 = \frac{Model\ SS}{Total\ SS} = 1 - \frac{Residual\ SS}{Total\ SS}$$

In the simple regression of GP1000M on Weight, R^2 = 76.5%. When Horsepower is added, R^2 increases to 84.1%.

Fact: R^2 can never decrease when another independent variable x is added to a regression. Why?

To avoid this limitation, people sometimes use adjusted R^2 which is essentially R^2 penalized for the number of x's in the regression.

In the previous two regressions, adjusted R^2 goes from 76.3% to 83.8% when Horsepower is added.

[8] As in simple regression, JMP labels the *Residual SS* as the *Error SS*.

Prediction Intervals for a Future Observation

"Where will a future value of the response *y* lie?"

"What GP1000M will I get with a car of a given weight and horsepower?"

After running a multiple regression, each point on the LS regression

$$\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_K x_K$$

is an estimate of the corresponding future point generated by the MRM

$$y_x = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \varepsilon_x$$

For example, suppose we wanted to predict GP1000M for a new design where Weight = 4000 and Horsepower = 200.

Using the regression of GP1000M on Weight and Horsepower, the predicted value of GP1000M for this new design is (p125)

GP1000M = 11.68 + 0.00891 Weight(lb) + 0.0883 Horsepower

= 11.6 + 0.00891 (4000) + 0.0883 (200) = 65.03

JMP provides this calculation for you.[9]

[9]At least, it will if you figure out how to ask it. The trick is to add an extra row to the data. The last row of car89.jump contains the *x* values for the new design. After running Fit Model to obtain the regression output, right-click on one of the title bars , and select Save Columns > Predicted Values from the Pop-up menu. The predicted values for all of the rows, including the new one, will be placed in a column to the right of the data. This can also be done by selecting Save Columns > Prediction Formula which also includes the prediction formula in the calculator window.

JMP also provides[10] [57.9, 72.1] as a 95% prediction interval (PI) for y_x when Weight = 4000 and Horsepower = 200

What is the interpretation of this interval?

These results can also easily be used to predict MPG for the new design. By using the transformation 1000/GP1000M, the prediction of MPG is (1000/65.03) = 15.8 and the 95% PI is [1000/72.1, 1000/57.9] = [13.9, 17.27]. (p 131)

The prediction of GP1000M for the simple regression on Weight is

$$GP1000M = 9.43 + 0.01362\ Weight(lb)$$

$$= 9.43 + 0.01362\ (\ 4000) = \ 63.9$$

and the 95% PI is [55.3, 72.5]. How do these compare with the above?

As in simple regression, extrapolate with caution! If $x_1, \ldots, x_K$ are not in the range of the data, predicting y_x is dangerous and the PIs are unreliable.

Often the intercept in a regression is an extrapolation itself. The intercept *is* the prediction when all of the predictors are set to zero. For data like these cars, we don't see any cases like that, and so the intercept is quite far from the data.

[10]Follow the steps in the previous footnote and select Save Columns > Indiv Confidence Interval.

Some New Graphical Model Diagnostics

In addition to the model checking methods we saw for simple regression, a variety of graphical methods are especially useful for multiple regression.

Plots of the Raw Data: Although it is not possible to plot all the data when $K > 3$, it may be useful to look at scatterplot matrices (pg 4-4) or 3-D spinning plots[11].

The following plots: Actual by Predicted, Residual by Predicted, and Leverage Plots were produced by the Fit Model platform[12] for the regression of GP1000M on Weight, Horsepower, Cargo and Seating.[13]

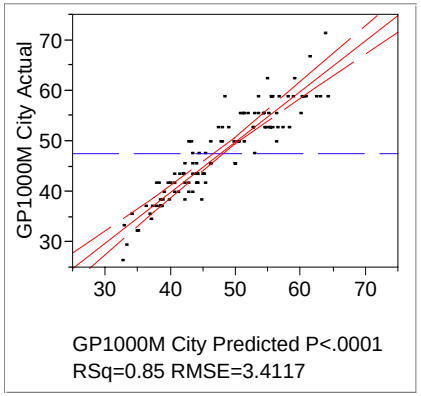
How should each of these be used? The key feature shared by all of these is that each offers you a "simple regression view" of the multiple regression.

Simple regression is simple because you can easily plot the data and see what is happening. These diagnostic plots shown by JMP with a multiple regression present various scatterplot views of a multiple regression.

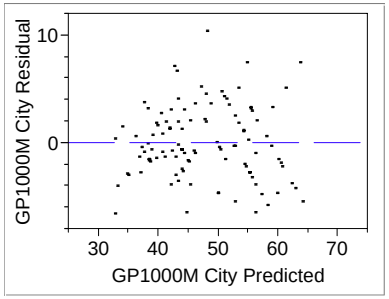
[11] Obtained with Graph > Spinning Plot In JMP.
[12] When Emphasis Select Leverage (the default) is selected.

The first two plots resemble the Fit Y by X plots of the data and residuals. For a simple regression, the one predictor supplied the x-axis. For multiple regression, these use a mixture of the predictors for the x-axis – namely the predicted values.

Actual By Predicted Plot [14]



Residual by Predicted Plot



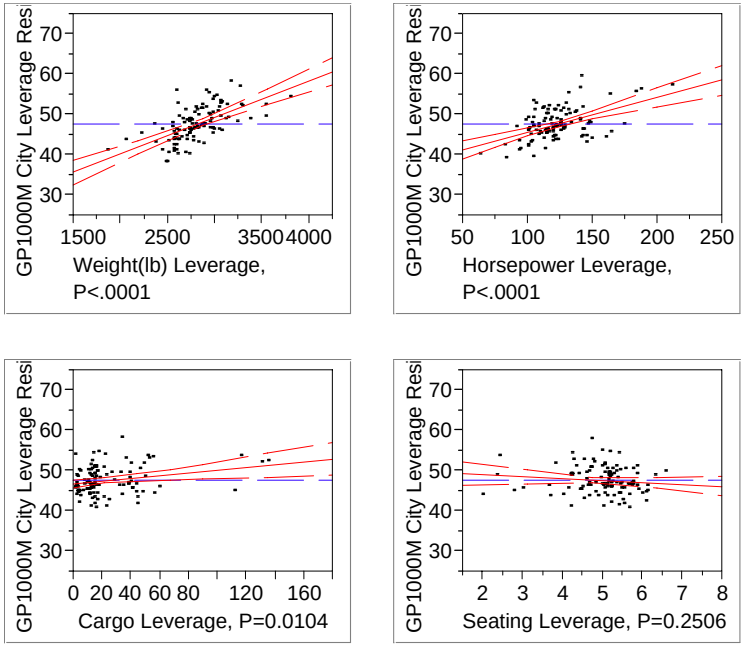
[14] JMP tries to show you a plot for every summary statistic. This time, its showing a plot that goes with the R^2 summary. The higher the R^2, the more the points in this plot cluster along the diagonal.

Leverage Plots

Leverage plots show you a simple regression view of the partial regression coefficient

Scatterplot: marginal coefficient
Leverage plot: partial coefficient

The slope of the fitted line in each leverage plot is the slope for the indicated predictor in the multiple regression.[15]



[15] So why are these called leverage plots? They excel at revealing leverage points in the multiple regression that are hard to spot in the marginal views of the fit. BAR (p 63) describes the calculation of leverage in a simple regression. To make a version of these plots by hand is straightforward, but tedious. To make the leverage plot for Weight, regress fuel consumption on the other predictors (HP, cargo, seating); save the residuals. Now regress Weight on these three other predictors; save these residuals. Finally (whew), plot the residuals of fuel consumption on the residuals of Weight. Though tedious, you can see how the leverage plot removes the effects of the other predictors. It uses regression!

Take-Away Review

Multiple regression extends the ideas of simple regression, allowing one to use several predictors to model simultaneously the variation in the response.

The addition of other variables changes the interpretation of the slope: the slope in a multiple regression is a "partial" effect, adjusted for the other predictors.

The underlying MRM is a natural extension of the SRM, allowing for more predictors. Under these assumptions, we can again use standard errors to form confidence intervals and test hypotheses.

To assess the assumptions of the MRM, new diagnostic plots include plots of fitted value on actual values of the response, residuals on fitted values, and leverage plots.

Next Module

More on multiple regression, with an emphasis on the effects of correlation among the predictors (i.e., collinearity).