Stat 102

Introduction to Business Statistics

Class 13

Richard P. Waterman

Wharton

Table of contents I

Today's module

Topics to be covered in this module:

- Last time
- Motivation for multiple regression
- Marginal v. partial association
- Interpretation of regression coefficients
- The scatterplot matrix and Mahalanobis plot
- The path diagram
- Summary
- Next time

Last time

- Broken regression assumptions
  1. Lack of independence of adjacent residuals
  2. Lack of constant variance of residuals
  3. Lack of normality of residuals
- Outliers and what to do with them

Introduction to multiple regression

- Making more realistic models with many X-variables – multiple regression analysis.
- For example, adding cut, clarity and color to the regression model for diamond prices.
- What is multiple regression?
  - The model:
$$E(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$
  - Here's what the regression looks like in 3-Dimensional space.

The general model

In general, there are k predictors:

- The model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i,$$

- where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$, with ϵ_i independent, constant variance and Normally distributed.

But we are going to have some difficulty viewing this as the dimensionality is very high.

The fundamental differences between simple and multiple regression

- The X-variables may be related (correlated) with one another.
- Consequence: looking at one X-variable at a time may present a misleading picture of the true relationship between Y and the X-variables.
- The difference between **marginal** and **partial** slopes. Marginal: the slope of the regression line for one X-variable **ignoring** the impact all the others. Partial: the slope of the regression line for one X-variable **taking into account** all the others.

Understanding the "partial" interpretation

- Take a two variable multiple regression with X_1 and X_2, then

$$E(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

- Hold X_2 constant and add one to X_1, to get:

$$E(Y|X_1 + 1, X_2) = \beta_0 + \beta_1(X_1 + 1) + \beta_2 X_2.$$

- How much has Y changed by on average (subtract the first equation from the second) ? Answer: β_1.
- *For every one unit change in X_1 holding X_2 constant we expect a β_1 unit change in Y.*

Two useful graphics

Two useful *get started plots*:

1. The scatterplot matrix: reveals marginal associations
2. Mahalanobis multivariate outlier plot

The scatterplot matrix

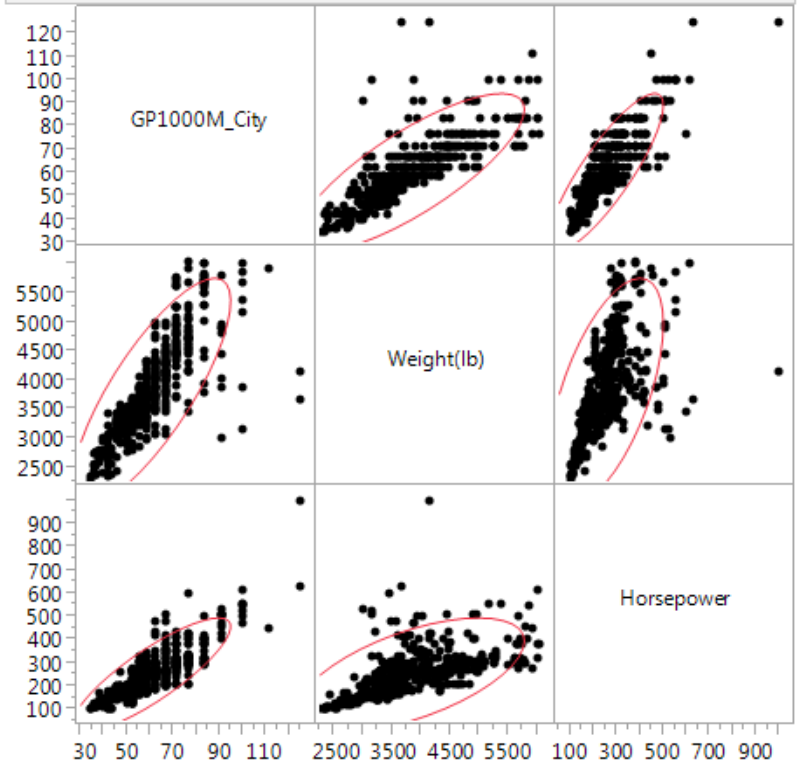
Displays all pairwise **marginal** associations (along with the correlation matrix).

Multivariate

Correlations

	GP1000M_City	Weight(lb)	Horsepower
GP1000M_City	1.0000	0.7992	0.8497
Weight(lb)	0.7992	1.0000	0.5955
Horsepower	0.8497	0.5955	1.0000

Scatterplot Matrix

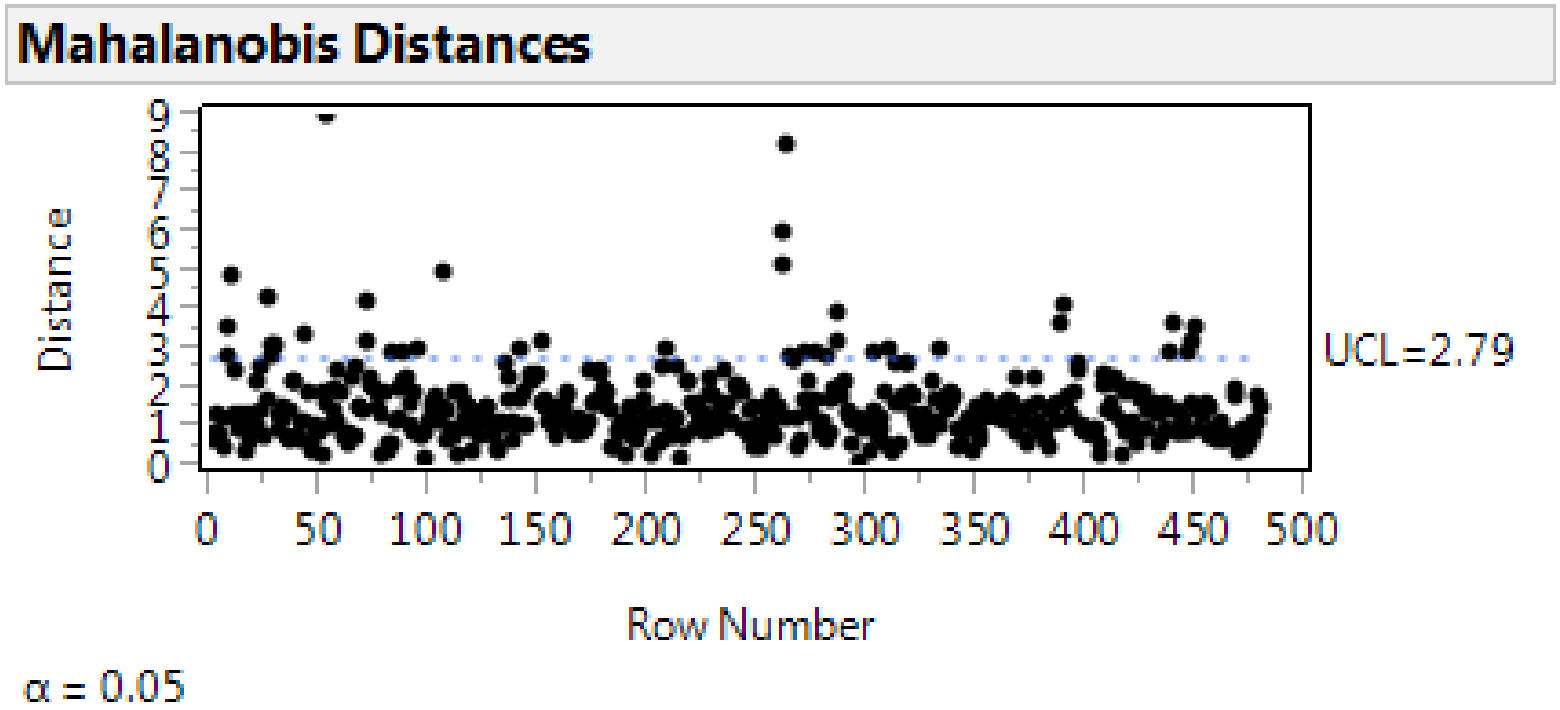


The scatterplot matrix

- A "marginal tool": presents all the two-variable (bivariate) relationships. Potential uses:
  - Identify collinearity (correlation) between X-variables.
  - Identify marginal non-linear relationships between Y and X-variables.
  - Determine which X-variables are marginally most significant (thin ellipses).

The Mahalanobis plot

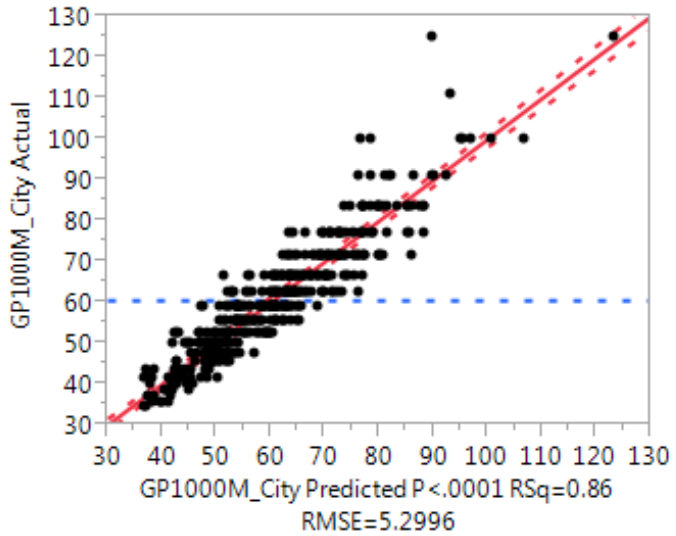
Displays the multivariate outlier metric. Note the blue line as a reference line. Points above this blue line are flagged as being unusual (but that doesn't imply they must be removed).



Running the multiple regression

In JMP use the "Fit model platform".

Actual by Predicted Plot



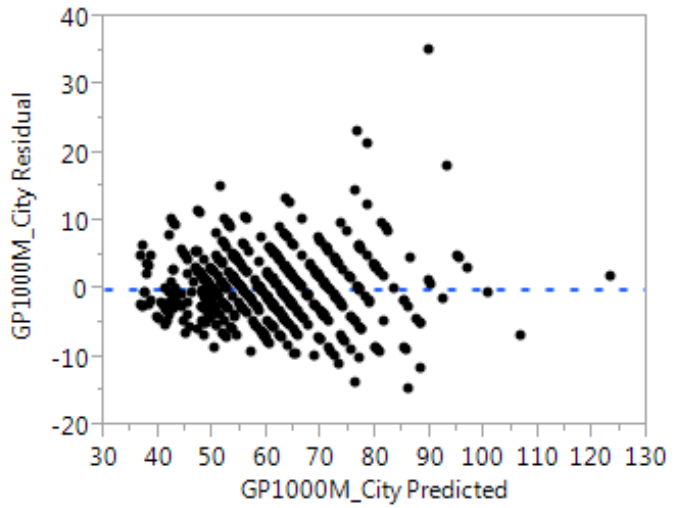
Summary of Fit

RSquare	0.855146
RSquare Adj	0.854538
Root Mean Square Error	5.299623
Mean of Response	60.40038
Observations (or Sum Wgts)	480

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	10.079054	1.190001	8.47	<.0001*
Weight(lb)	0.0079262	0.000378	20.94	<.0001*
Horsepower	0.0800873	0.003	26.70	<.0001*

Residual by Predicted Plot



Running the multiple regression

Note the many familiar features in the output: R^2, RMSE, residual plot etc, but we will focus on the *partial slopes* today.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	10.079054	1.190001	8.47	<.0001*
Weight(lb)	0.0079262	0.000378	20.94	<.0001*
Horsepower	0.0800873	0.003	26.70	<.0001*

Understanding the dangers of simple regression

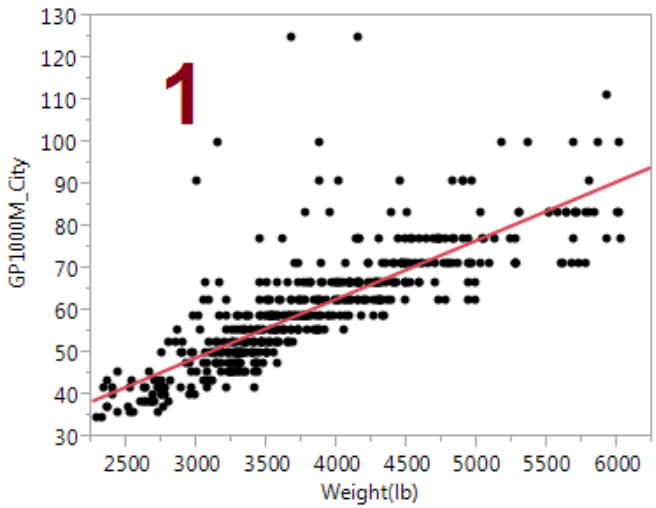
Beware the marginal regression coefficient ...
We will relate three sets of regression coefficients to gain some insight into the difference between marginal and partial:

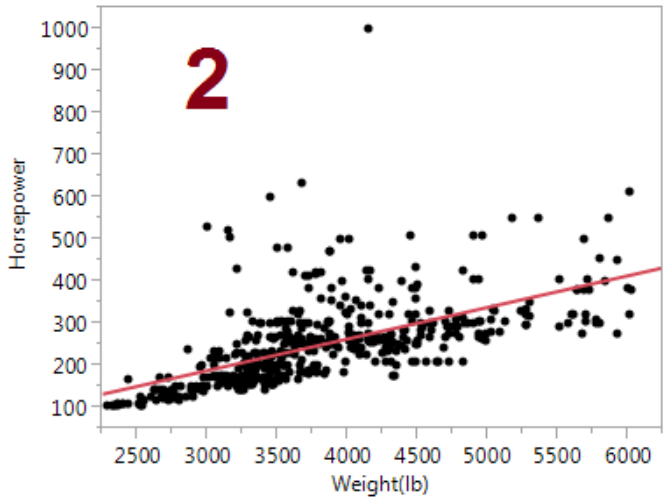
1. The marginal regression coefficient of `GPM1000` against `weight`.
2. The marginal regression coefficient of `horesepower` against `weight`.
3. The partial regression coefficients of `GPM1000` against both `weight` and `horsepower`.

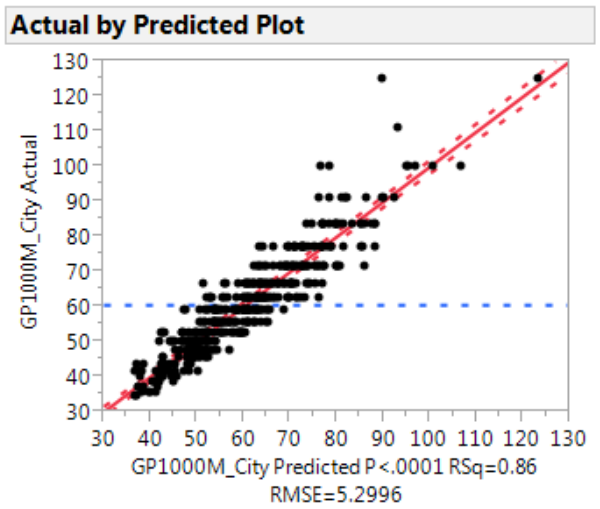
By relating these coefficients we will see that marginal regression coefficients can be seriously misleading because they are **confounded** by correlated explanatory variables not yet in the model.

Understanding the dangers of simple regression

Beware the marginal regression coefficient ...







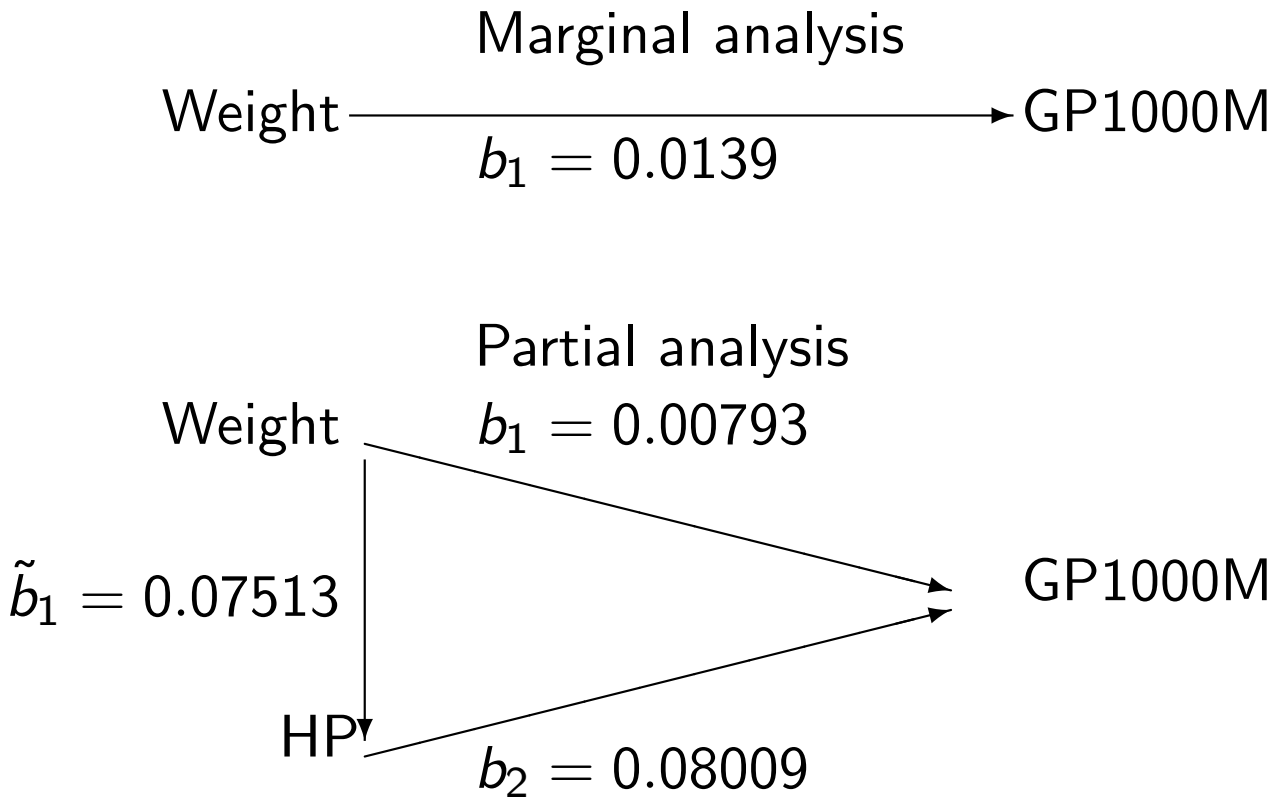
3

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	10.079054	1.190001	8.47	<.0001*
Weight(lb)	0.0079262	0.000378	20.94	<.0001*
Horsepower	0.0800873	0.003	26.70	<.0001*

The path diagram for a multiple regression slope

The goal: understand why marginal regression coefficients can be so misleading.



Math for the path diagram

- Note that $\tilde{b}_1 = 0.07513$ is regression coefficient for horsepower against weight.
- Call $b_1 = 0.00793$ the $\boxed{\text{DIRECT}}$ effect of weight on fuel economy.
- Call $\tilde{b}_1 \times b_2 = 0.07513 \times 0.08009 = 0.00602$ the $\boxed{\text{INDIRECT}}$ effect of weight on fuel economy, via horsepower.
  - $\boxed{\text{DIRECT EFFECT}} + \boxed{\text{INDIRECT EFFECT}} = \boxed{\text{MARGINAL COEFFICIENT}}$.

$$\boxed{0.00793} + \boxed{0.00602} = \boxed{0.0139}.$$

- Summary: marginal coefficients confound direct and indirect effects, partial coefficients isolate the direct effects.

Module summary

- Intro to multiple regression
- Marginal v. partial
- Useful graphics
- Beware marginal regression coefficients

Next time

- The usual suspects: R^2, RMSE
- Adjusted-R^2
- Checking assumptions in multiple regression
- Inference and the model building process