

Multiple Linear Regression

So far, we have seen the concept of simple linear regression where a single predictor variable X was used to model the response variable Y . In many applications, there is more than one factor that influences the response. Multiple regression models thus describe how a single response variable Y depends linearly on a number of predictor variables.

Examples:

- The selling price of a house can depend on the desirability of the location, the number of bedrooms, the number of bathrooms, the year the house was built, the square footage of the lot and a number of other factors.
- The height of a child can depend on the height of the mother, the height of the father, nutrition, and environmental factors.

Note: We will reserve the term MULTIPLE REGRESSION for models with two or more predictors and one response. There are also regression models with two or more response variables. These models are usually called MULTIVARIATE REGRESSION MODELS.

In this chapter, we will introduce a new (linear algebra based) method for computing the parameter estimates of multiple regression models. This more compact method is convenient for models for which the number of unknown parameters is large.

Example: A multiple linear regression model with k predictor variables X_1, X_2, \dots, X_k and a response Y , can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon.$$

As before, the ϵ are the residual terms of the model and the distribution assumption we place on the residuals will allow us later to do inference on the remaining model parameters. Interpret the meaning of the REGRESSION COEFFICIENTS $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ in this model.

More complex models may include higher powers of one or more predictor variables, e.g.,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \quad (1)$$

or interaction effects of two or more variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \quad (2)$$

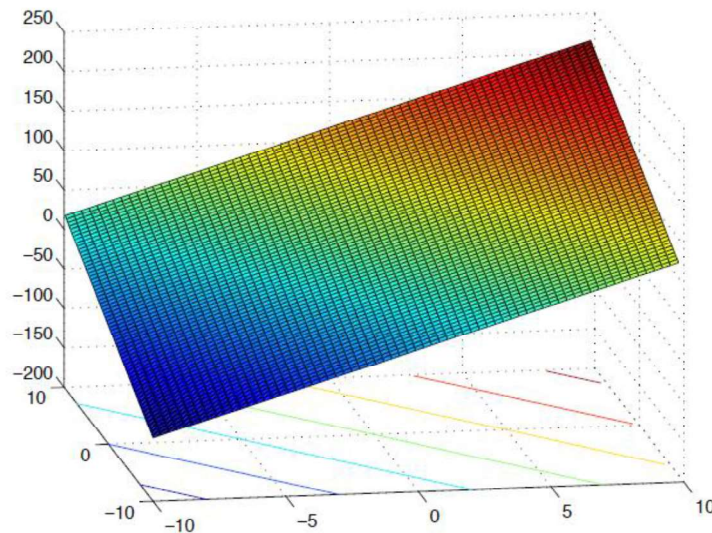
Example: The simplest multiple regression model for two predictor variables is

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

The surface that corresponds to the model

$$y = 50 + 10x_1 + 7x_2$$

looks like this. It is a plane in \mathbb{R}^3 with different slopes in x_1 and x_2 direction.



Estimation of the Model Parameters

While it is possible to estimate the parameters of more complex linear models with methods similar to those we have seen in chapter 2, the computations become very complicated very quickly. Thus, we will employ linear algebra methods to make the computations more efficient.

The setup: Consider a multiple linear regression model with k independent predictor variables x_1, \dots, x_k and one response variable y .

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \epsilon$$

Suppose, we have n observations on the $k + 1$ variables.

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \dots, n$$

n should be bigger than k . Why?

You can think of the observations as points in $(k + 1)$ -dimensional space if you like. Our goal in least-squares regression is to fit a hyper-plane into $(k + 1)$ -dimensional space that minimizes the sum of squared residuals.

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right)^2$$

As before, we could take derivatives with respect to the model parameters β_0, \dots, β_k , set them equal to zero and derive the LEAST-SQUARES NORMAL EQUATIONS that our parameter estimates $\hat{\beta}_0, \dots, \hat{\beta}_k$ would have to fulfill.

$$\begin{array}{cccccc} n\hat{\beta}_0 & +\hat{\beta}_1 \sum_{i=1}^n x_{i1} & +\hat{\beta}_2 \sum_{i=1}^n x_{i2} & +\cdots & +\hat{\beta}_k \sum_{i=1}^n x_{ik} & = \sum_{i=1}^n y_i \\ \hat{\beta}_0 \sum_{i=1}^n x_{i1} & +\hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 & +\hat{\beta}_2 \sum_{i=1}^n x_{i1}x_{i2} & +\cdots & +\hat{\beta}_k \sum_{i=1}^n x_{i1}x_{ik} & = \sum_{i=1}^n x_{i1}y_i \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \hat{\beta}_0 \sum_{i=1}^n x_{ik} & +\hat{\beta}_1 \sum_{i=1}^n x_{ik}x_{i1} & +\hat{\beta}_2 \sum_{i=1}^n x_{ik}x_{i2} & +\cdots & +\hat{\beta}_k \sum_{i=1}^n x_{ik}^2 & = \sum_{i=1}^n x_{ik}y_i \end{array}$$

These equations are much more conveniently formulated with the help of vectors and matrices.

Note: Bold-faced lower case letters will now denote vectors and bold-faced upper case letters will denote matrices. Greek letters cannot be bold-faced in LaTeX. Whether a Greek letter denotes a random variable or a vector of random variables should be clear from the context, hopefully.

Let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

With this compact notation, the linear regression model can be written in the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

As was the case in the simple linear regression setting, the regression coefficients $\beta_0, \beta_1, \dots, \beta_p$ in (3.19) are unknown, and must be estimated. Given estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, we can make predictions using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p. \quad (3.21)$$

The parameters are estimated using the same least squares approach that we saw in the context of simple linear regression. We choose $\beta_0, \beta_1, \dots, \beta_p$ to minimize the sum of squared residuals

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_p x_{ip})^2. \end{aligned} \quad (3.22)$$

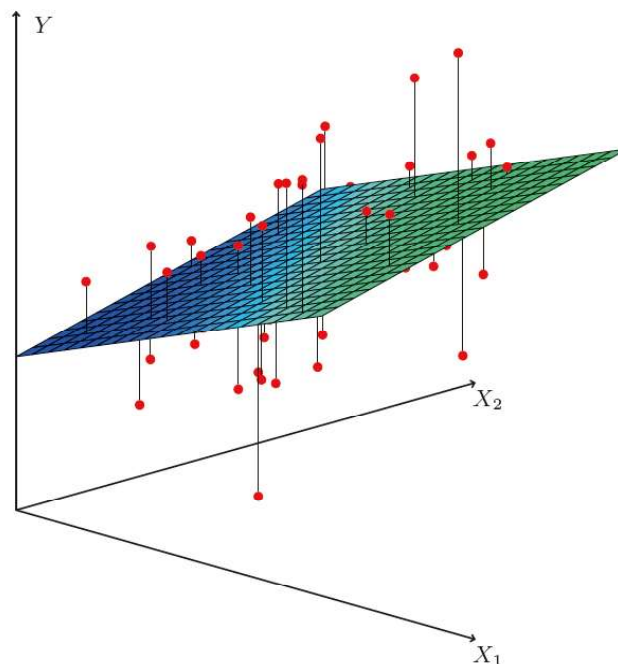


FIGURE 3.4. In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.

When we perform multiple linear regression, we usually are interested in answering a few important questions.

1. *Is at least one of the predictors X_1, X_2, \dots, X_p useful in predicting the response?*
2. *Do all the predictors help to explain Y , or is only a subset of the predictors useful?*
3. *How well does the model fit the data?*
4. *Given a set of predictor values, what response value should we predict, and how accurate is our prediction?*

Assessing the Accuracy of the Model

Once we have rejected the null hypothesis in favor of the alternative hypothesis, it is natural to want to quantify *the extent to which the model fits the data*. The quality of a linear regression fit is typically assessed using two related quantities: the *residual standard error* (RSE) and the R^2 statistic.

Residual Standard Error: Recall from the model that associated with each observation is an error term ε . Due to the presence of these error terms, even if we knew the true regression line (i.e. even if β_0 and β_1 were known), we would not be able to perfectly predict Y from X . The RSE is an estimate of the standard deviation of ε . Roughly speaking, it is the average amount that the response will deviate from the true regression line. It is computed using the formula:

$$\text{RSE} = \sqrt{\frac{1}{n-2} \text{RSS}} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2}.$$

RSS is given by the formula:

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

The RSE is considered a measure of the *lack of fit* of the model to the data. If the predictions obtained using the model are very close to the true outcome values—that is, if

$$\hat{y}_i \approx y_i$$

for $i = 1, \dots, n$, then RSE will be small, and we can conclude that the model fits the data very well.

On the other hand, if \hat{y}_i is very far from y_i for one or more observations, then the RSE may be quite large, indicating that the model doesn't fit the data well.

To calculate R^2 , we use the formula

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where $\text{TSS} = \sum (y_i - \bar{y})^2$ is the *total sum of squares*.

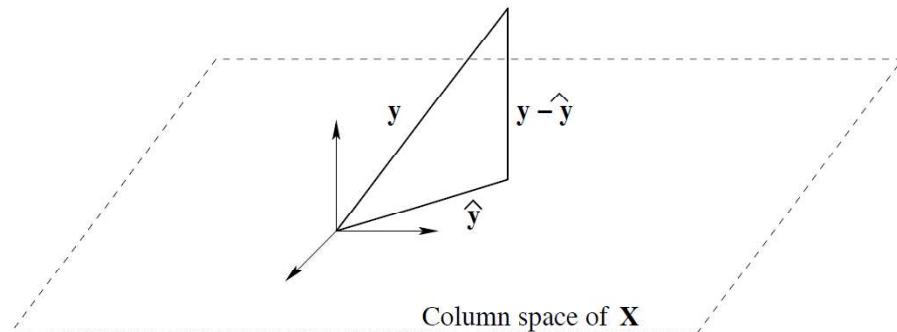
In linear algebra terms, the least-squares parameter estimates β are the vectors that minimize

$$\sum_{i=1}^n \epsilon_i^2 = \epsilon' \epsilon = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

Any expression of the form $\mathbf{X}\beta$ is an element of a (at most) $(k+1)$ -dimensional hyperspace in \mathbb{R}^n spanned by the $(k+1)$ columns of \mathbf{X} . Imagine the columns of \mathbf{X} to be fixed, they are the data for a specific problem, and imagine β to be variable. We want to find the “best” β in the sense that the sum of squared residuals is minimized. The smallest that the sum of squares could be is zero. If all ϵ_i were zero, then

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

Here $\hat{\mathbf{y}}$ is the projection of the n -dimensional data vector \mathbf{y} onto the hyperplane spanned by \mathbf{X} .



The $\hat{\mathbf{y}}$ are the predicted values in our regression model that all lie on the regression hyper-plane. Suppose further that $\hat{\beta}$ satisfies the equation above. Then the residuals $\mathbf{y} - \hat{\mathbf{y}}$ are orthogonal to the columns of \mathbf{X} (by the Orthogonal Decomposition Theorem) and thus

$$\begin{aligned}
\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\beta}) &= 0 \\
\Leftrightarrow \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\hat{\beta} &= 0 \\
\Leftrightarrow \mathbf{X}'\mathbf{X}\hat{\beta} &= \mathbf{X}'\mathbf{y}
\end{aligned}$$

These vector normal equations are the same normal equations that one could obtain from taking derivatives. To solve the normal equations (i.e., to find the parameter estimates $\hat{\beta}$), multiply both sides with the inverse of $\mathbf{X}'\mathbf{X}$. Thus, the least-squares estimator of β is (in vector form)

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

This of course works only if the inverse exists. If the inverse does not exist, the normal equations can still be solved, but the solution may not be unique. The inverse of $\mathbf{X}'\mathbf{X}$ exists, if the columns of \mathbf{X} are linearly independent. That means that no column can be written as a linear combination of the other columns.

This of course works only if the inverse exists. If the inverse does not exist, the normal equations can still be solved, but the solution may not be unique. The inverse of $\mathbf{X}'\mathbf{X}$ exists, if the columns of \mathbf{X} are linearly independent. That means that no column can be written as a linear combination of the other columns.

The vector of fitted values $\hat{\mathbf{y}}$ in a linear regression model can be expressed as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

The $n \times n$ matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is often called the HAT-MATRIX. It maps the vector of observed values \mathbf{y} onto the vector of fitted values $\hat{\mathbf{y}}$ that lie on the regression hyper-plane. The regression residuals can be written in different ways as

$$\epsilon = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

MULTIPLE REGRESSION: PART 1: THE VERY BASICS

REGIONAL DELIVERY SERVICE

Let's assume that you are a small business owner for Regional Delivery Service, Inc. (RDS) who offers same-day delivery for letters, packages, and other small cargo. You are able to use Google Maps to group individual deliveries into one trip to reduce time and fuel costs. Therefore some trips will have more than one delivery.

As the owner, you would like to be able to *estimate how long a delivery will take* based on two factors: 1) the total distance of the trip in miles and 2) the number of deliveries that must be made during the trip.

RDS DATA AND VARIABLE NAMING

To conduct your analysis you take a random sample of 10 past trips and record three pieces of information for each trip: 1) total miles traveled, 2) number of deliveries, and 3) total travel time in hours.

milesTraveled, (x_1)	numDeliveries, (x_2)	travelTime(hrs), (y)
89	4	7
66	1	5.4
78	3	6.6
111	6	7.4
44	1	4.8
77	3	6.4
80	3	7
66	2	5.6
109	5	7.3
76	3	6.4

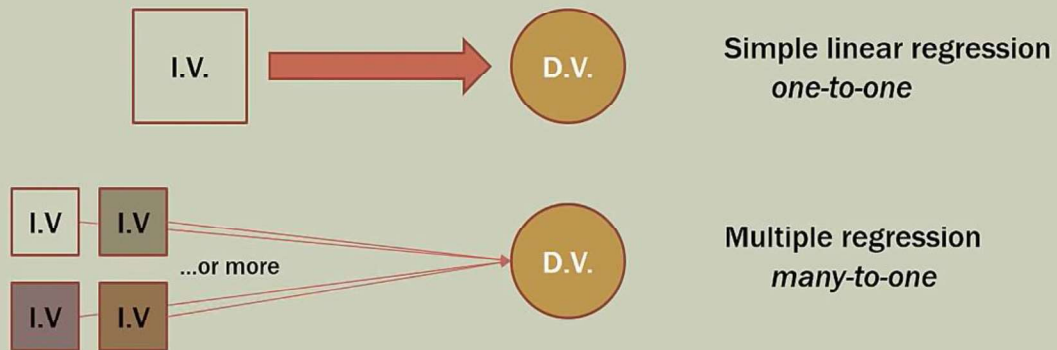
Remember that in this case, you would like to be able to **predict the total travel time** using both the miles traveled and number of deliveries on each trip.

In what way does travel time **DEPEND** on the first two measures?

Travel time is the *dependent variable* and miles traveled and number of deliveries are independent variables.

MULTIPLE REGRESSION

Multiple regression is an extension of simple linear regression.



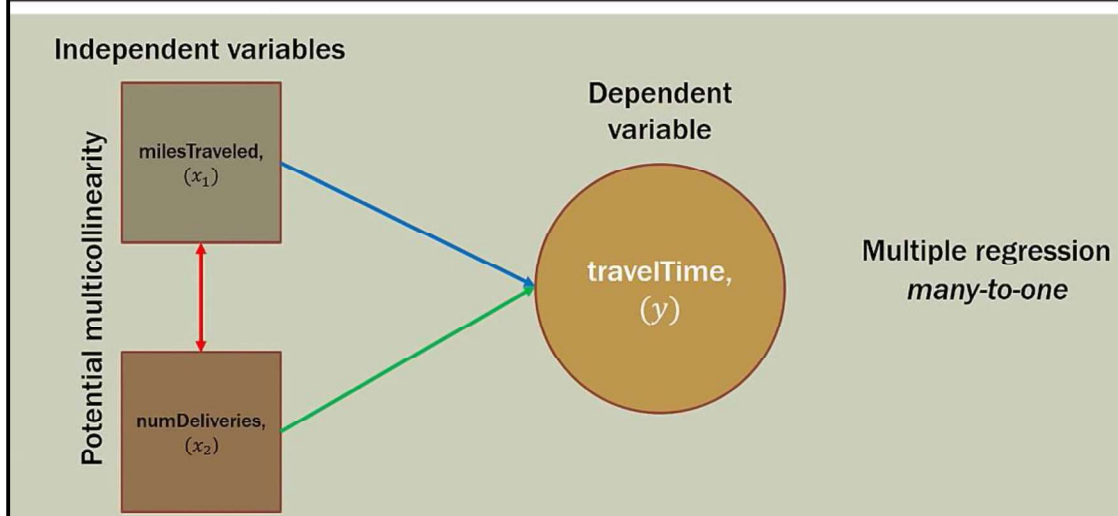
NEW CONSIDERATIONS

- Adding more independent variables to a multiple regression procedure does not mean the regression will be “better” or offer better predictions; in fact it can make things worse. This is called **OVERFITTING**.
- The addition of more independent variables creates more relationships among them. So not only are the independent variables potentially related to the dependent variable, they are also potentially *related to each other*. When this happens, it is called **MULTICOLLINEARITY**.
- The ideal is for all of the independent variables to be correlated with the dependent variable but **NOT** with each other.

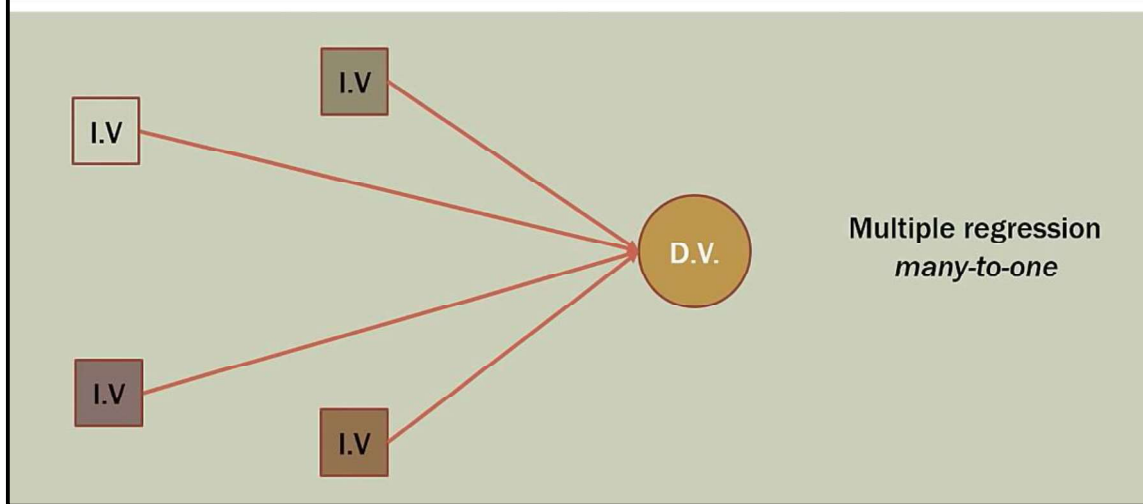
NEW CONSIDERATIONS

- Because of multicollinearity and overfitting, there is a fair amount of prep-work to do BEFORE conducting multiple regression analysis if one is to do it properly.
 - Correlations
 - Scatter plots
 - Simple regressions

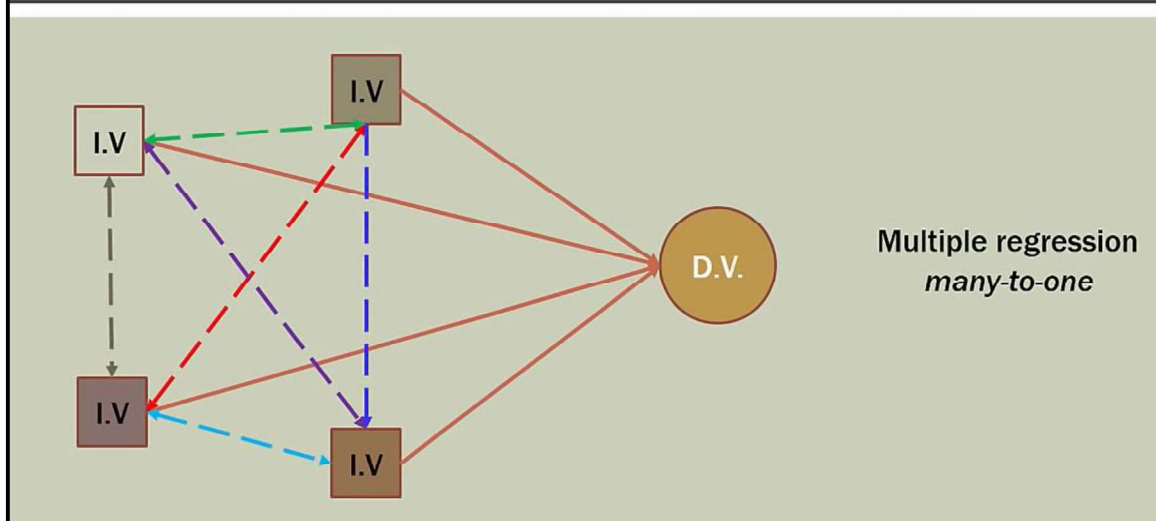
MORE RELATIONSHIPS



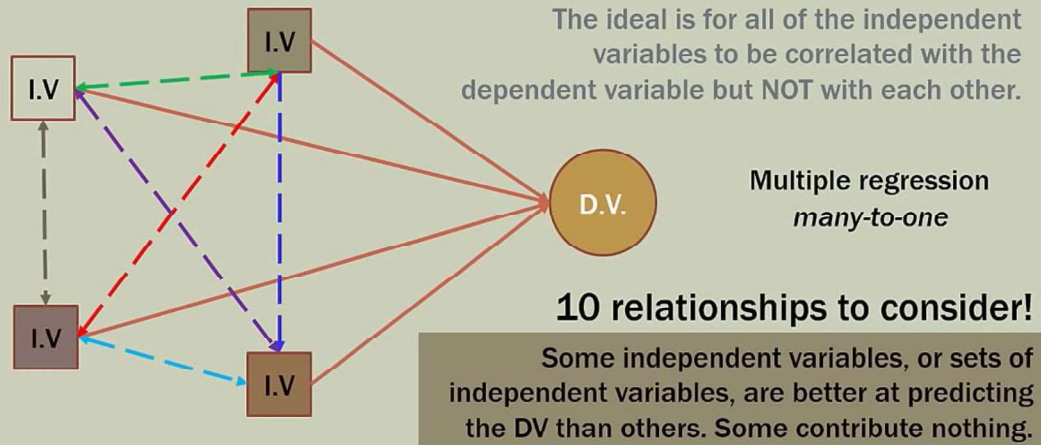
MANY RELATIONSHIPS



MANY RELATIONSHIPS



MANY RELATIONSHIPS



MULTIPLE REGRESSION MODEL

Multiple Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p + \epsilon$$

linear parameters error

Multiple Regression Equation

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p$$

error term assumed to be zero

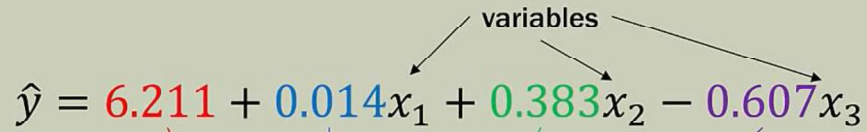
Estimated Multiple Regression Equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots b_p x_p$$

$b_0, b_1, b_2, \dots b_p$ are the estimates of $\beta_0, \beta_1, \beta_2, \dots \beta_p$
 \hat{y} = predicted value of the dependent variable

ESTIMATED MULTIPLE REGRESSION EQUATION

Example

$$\hat{y} = 6.211 + 0.014x_1 + 0.383x_2 - 0.607x_3$$


intercept

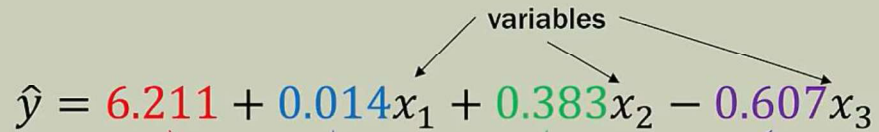
coefficients

Estimated Multiple
Regression Equation

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

ESTIMATED MULTIPLE REGRESSION EQUATION

Example

$$\hat{y} = 6.211 + 0.014x_1 + 0.383x_2 - 0.607x_3$$


intercept

coefficients

Estimated Multiple
Regression Equation

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

$b_0, b_1, b_2, \dots, b_p$ are the estimates of $\beta_0, \beta_1, \beta_2, \dots, \beta_p$

\hat{y} = predicted value of the dependent variable

INTERPRETING COEFFICIENTS

$$\hat{y} = 27 + 9x_1 + 12x_2$$

x_1 = capital investment (\$1000s)

x_2 = marketing expenditures (\$1000s)

\hat{y} = predicted sales (\$1000s)

In multiple regression, each coefficient is interpreted as the estimated change in y corresponding to a one unit change in a variable, when all other variables are held constant.

So in this example, \$9000 is an estimate of the expected increase in sales y , corresponding to a \$1000 increase in capital investment (x_1) when marketing expenditures (x_2) are held constant.

REVIEW

- Multiple regression is an extension of simple linear regression
- Two or more independent variables are used to predict / explain the variance in one dependent variable
- Two problems may arise:
 - Overfitting
 - Multicollinearity
- **Overfitting** is caused by adding too many independent variables; they account for more variance but add nothing to the model
- **Multicollinearity** happens when some/all of the independent variables are correlated with each other
- In multiple regression, each coefficient is interpreted as the estimated change in y corresponding to a one unit change in a variable, when all other variables are held constant.