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.

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

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

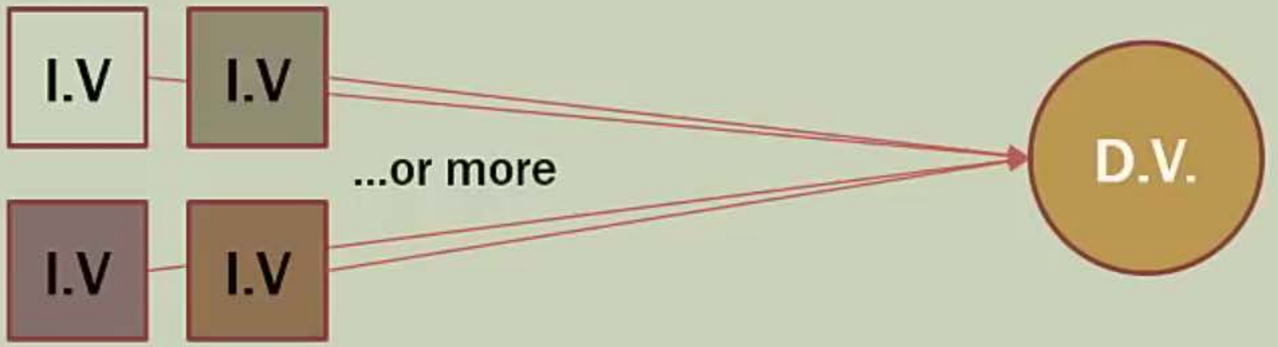
Multiple regression is an extension of simple linear regression.

Simple linear regression
one-to-one

MULTIPLE REGRESSION

Multiple regression is an extension of simple linear regression.

Simple linear regression
one-to-one



Multiple regression
many-to-one

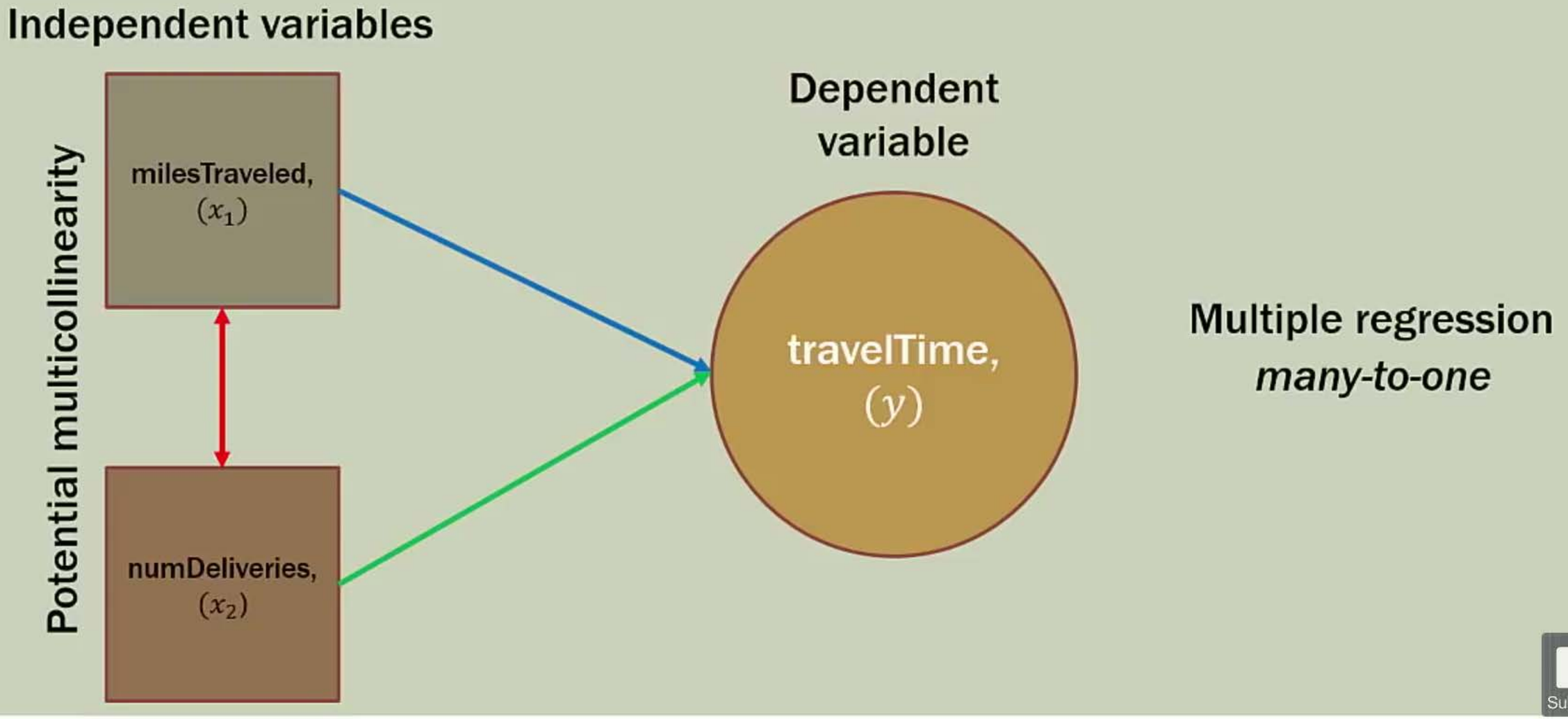
NEW CONSIDERATIONS

- 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.

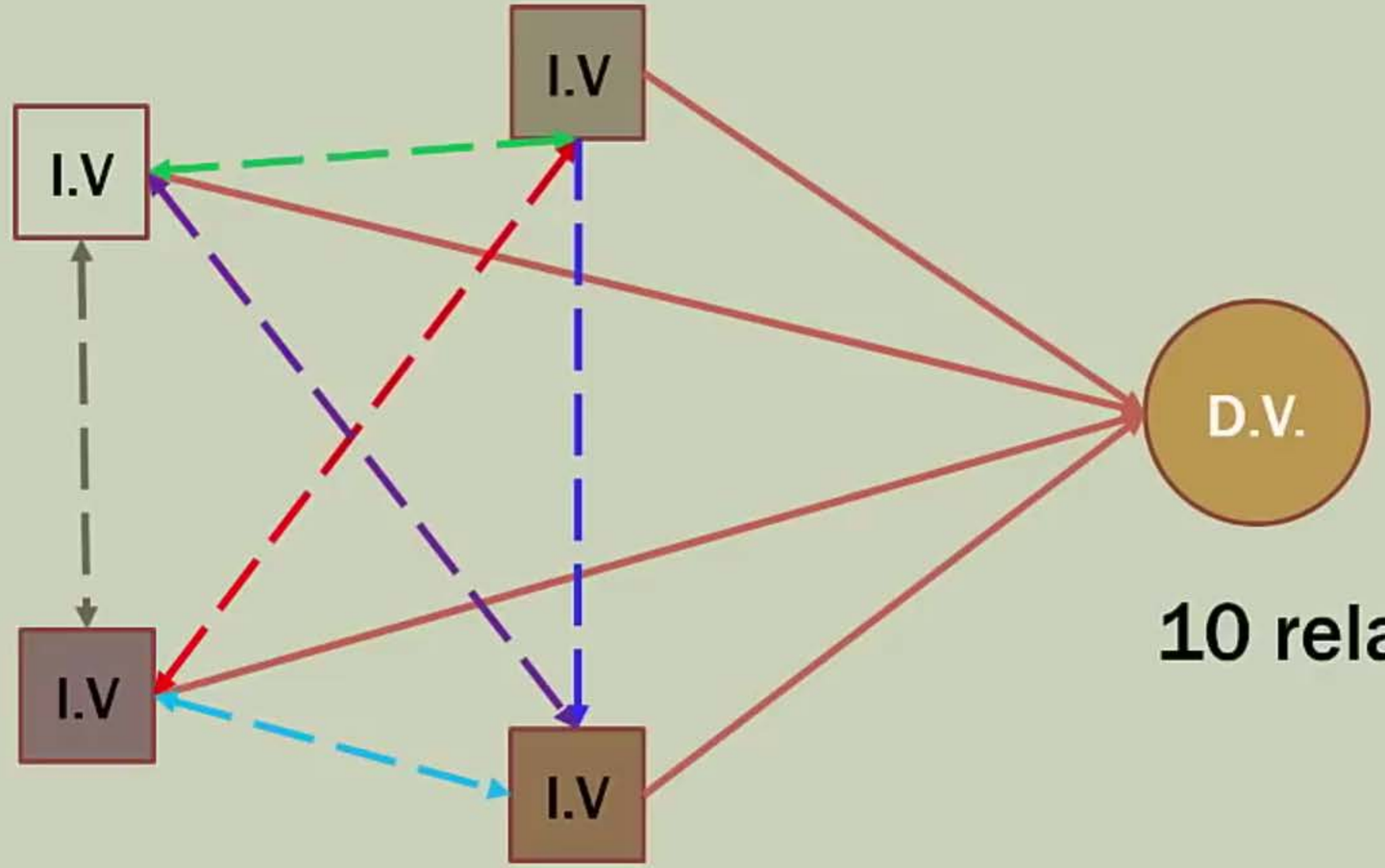
NEW CONSIDERATIONS

- 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.

MORE RELATIONSHIPS



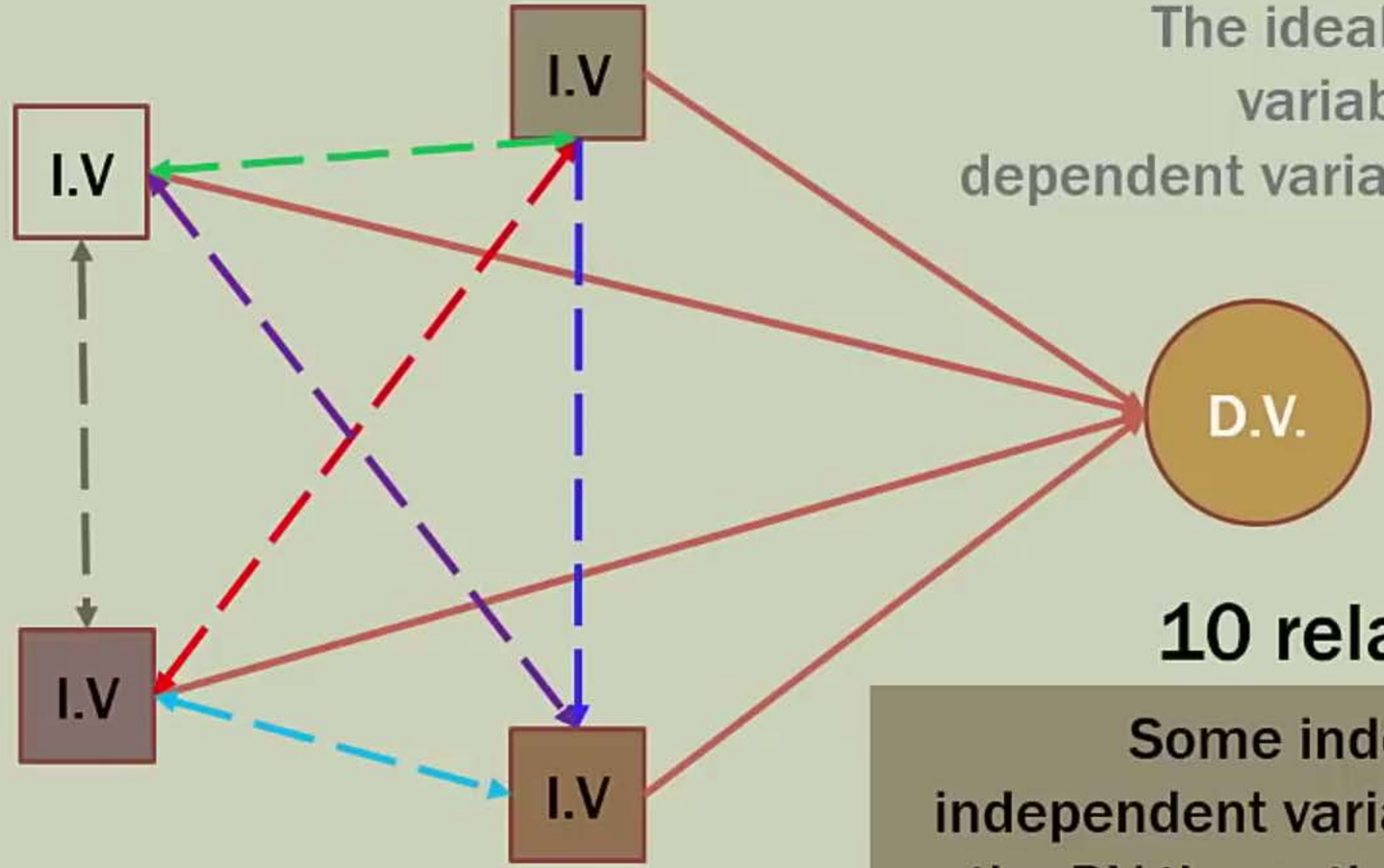
MANY RELATIONSHIPS



Multiple regression
many-to-one

10 relationships to consider!

MANY RELATIONSHIPS



The ideal is for all of the independent variables to be correlated with the dependent variable but NOT with each other.

Multiple regression
many-to-one

10 relationships to consider!

Some independent variables, or sets of independent variables, are better at predicting the DV than others. Some contribute nothing.

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

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

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

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

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

ESTIMATED MULTIPLE REGRESSION EQUATION

variables

Example

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

intercept

coefficients

Estimated Multiple Regression Equation

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