Module 4

Introduction to Multiple Linear Regression

Contents

- Multiple Linear Regression Model
- Partial Regression Coefficients
- Testing Significance overall significance of Overall fit of the model
- Testing for Individual Regression Coefficients

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

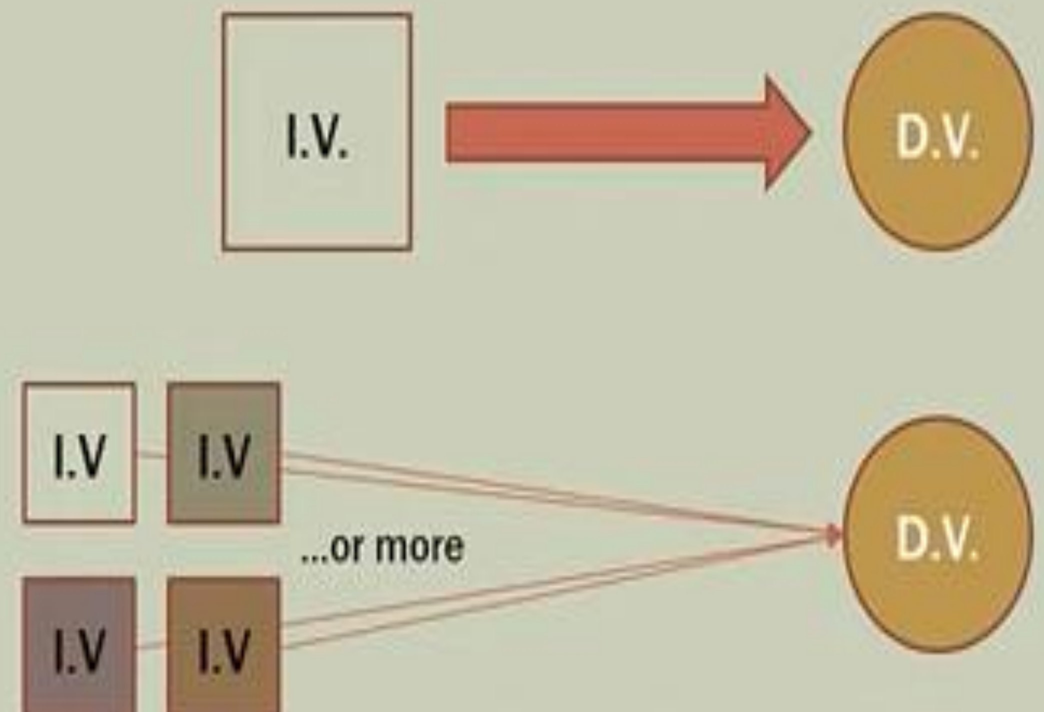
Remember that in this case, you would like to be able to predict the total travel time using both the miles traveled and number of deliveries on each trip.

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



Simple linear regression
one-to-one

Multiple regression
many-to-one

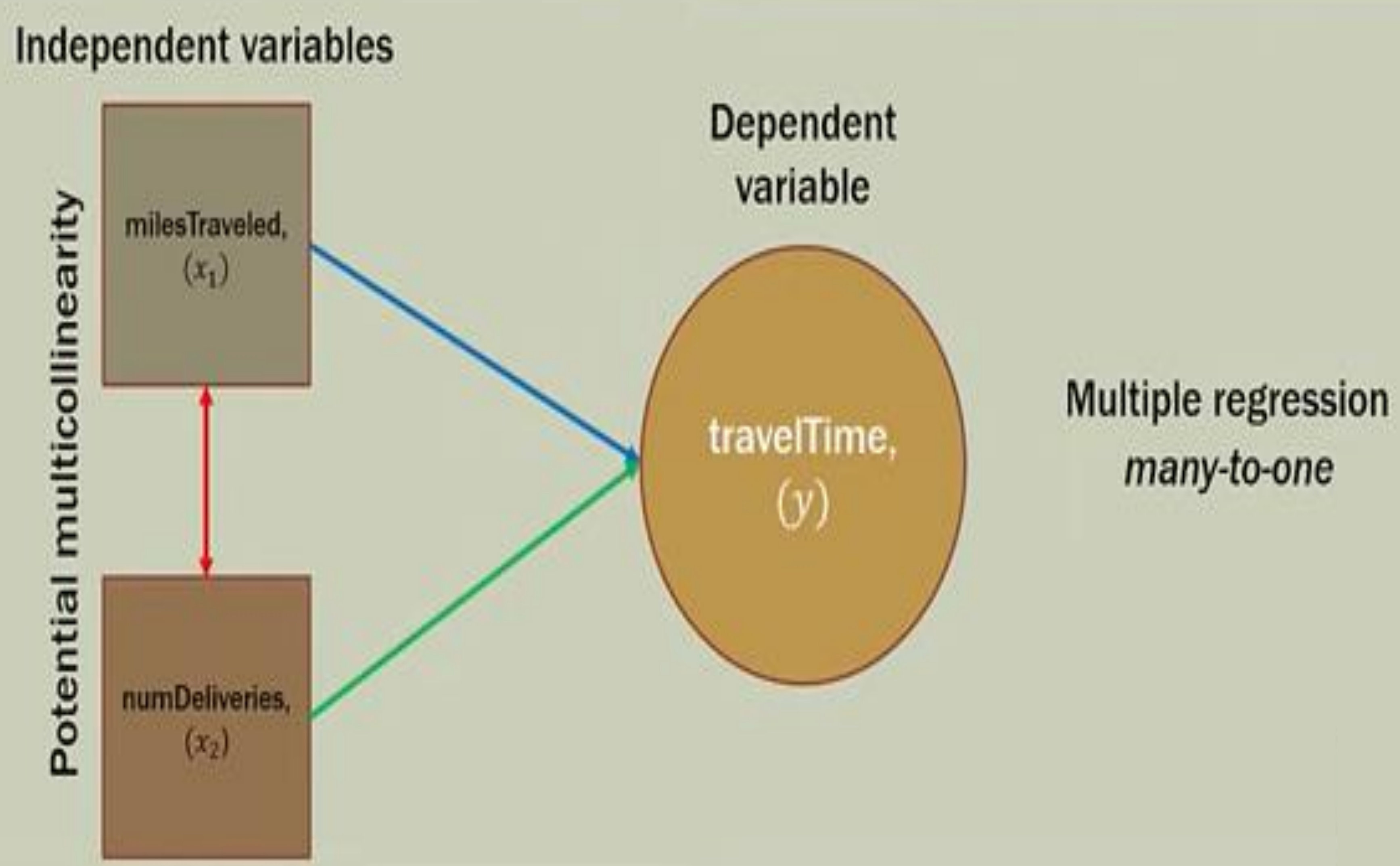
NEW CONSIDERATIONS

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

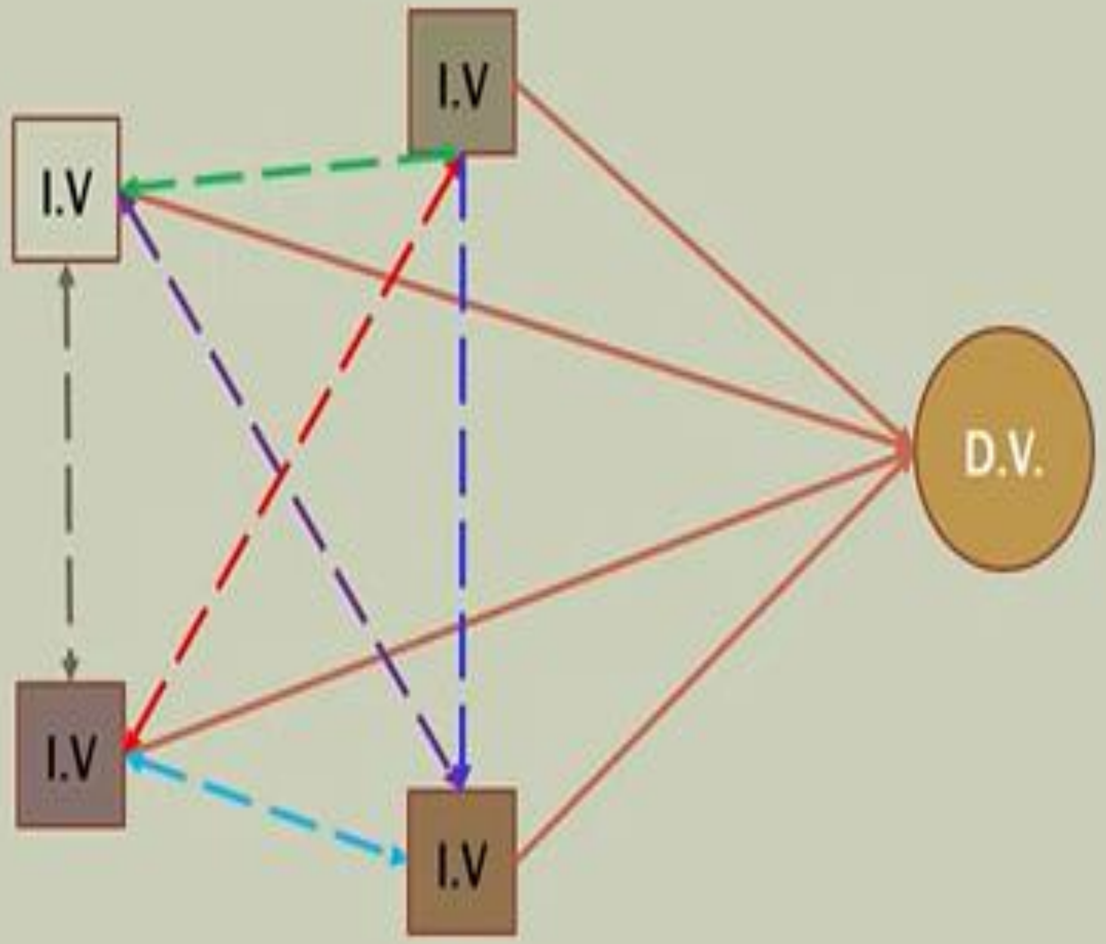
NEW CONSIDERATIONS

- 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

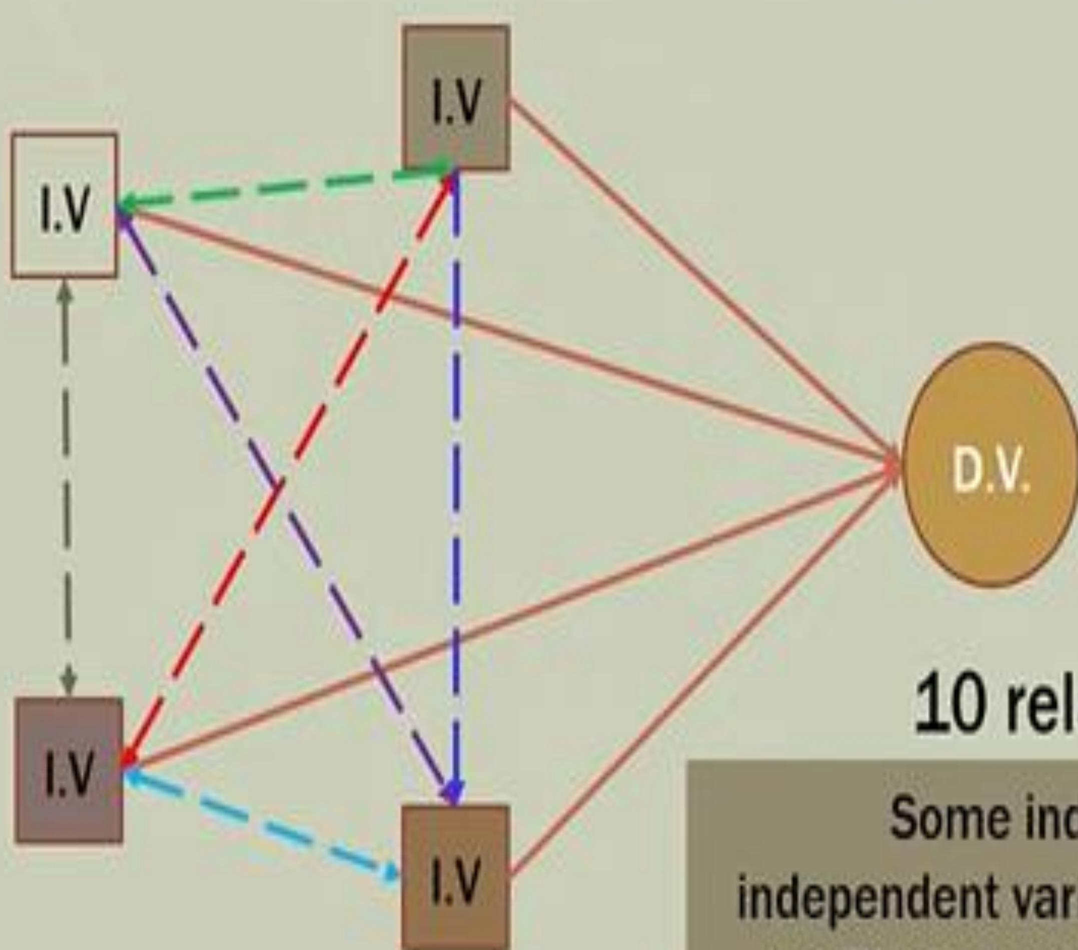


MANY RELATIONSHIPS



Multiple regression
many-to-one

MANY RELATIONSHIPS



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

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

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

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

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.**

MULTIPLE REGRESSION PREP

As we discussed in Part 1, conducting multiple regression analysis requires a fair amount of pre-work before actually running the regression. Here are the steps:

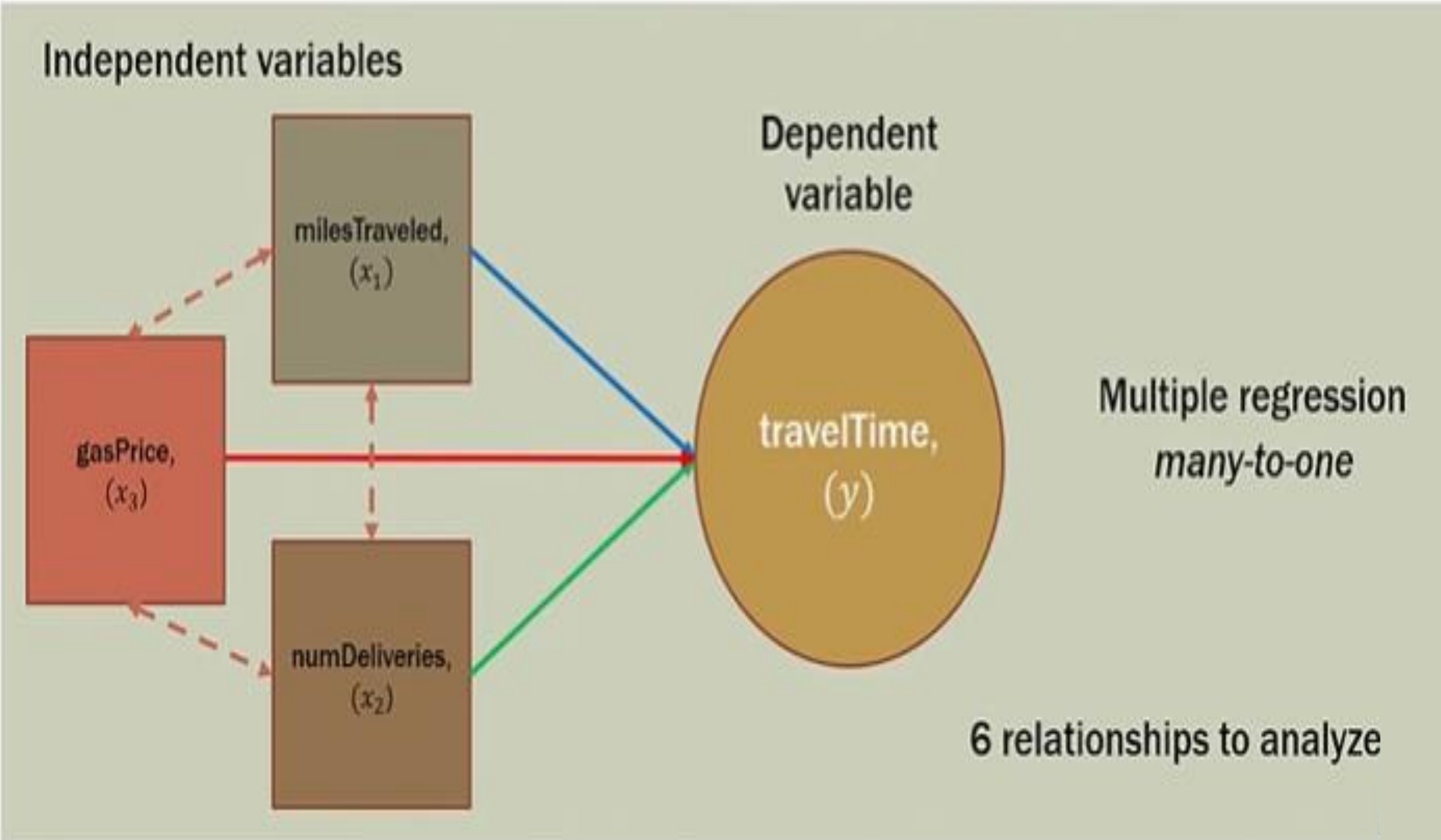
1. Generate a list of potential variables; independent(s) and dependent
2. Collect data on the variables
3. Check the relationships between each independent variable and the dependent variable using scatterplots and correlations
4. Check the relationships among the independent variables using scatterplots and correlations
5. (Optional) Conduct simple linear regressions for each IV/DV pair
6. Use the non-redundant independent variables in the analysis to find the best fitting model
7. Use the best fitting model to make predictions about the dependent variable.

RDS DATA AND VARIABLE NAMING

To conduct your analysis you take a random sample of 10 past trips and recorc four pieces of information for each trip: 1) total miles traveled, 2) number of deliveries, 3) the daily gas price, and 4) total travel time in hours.

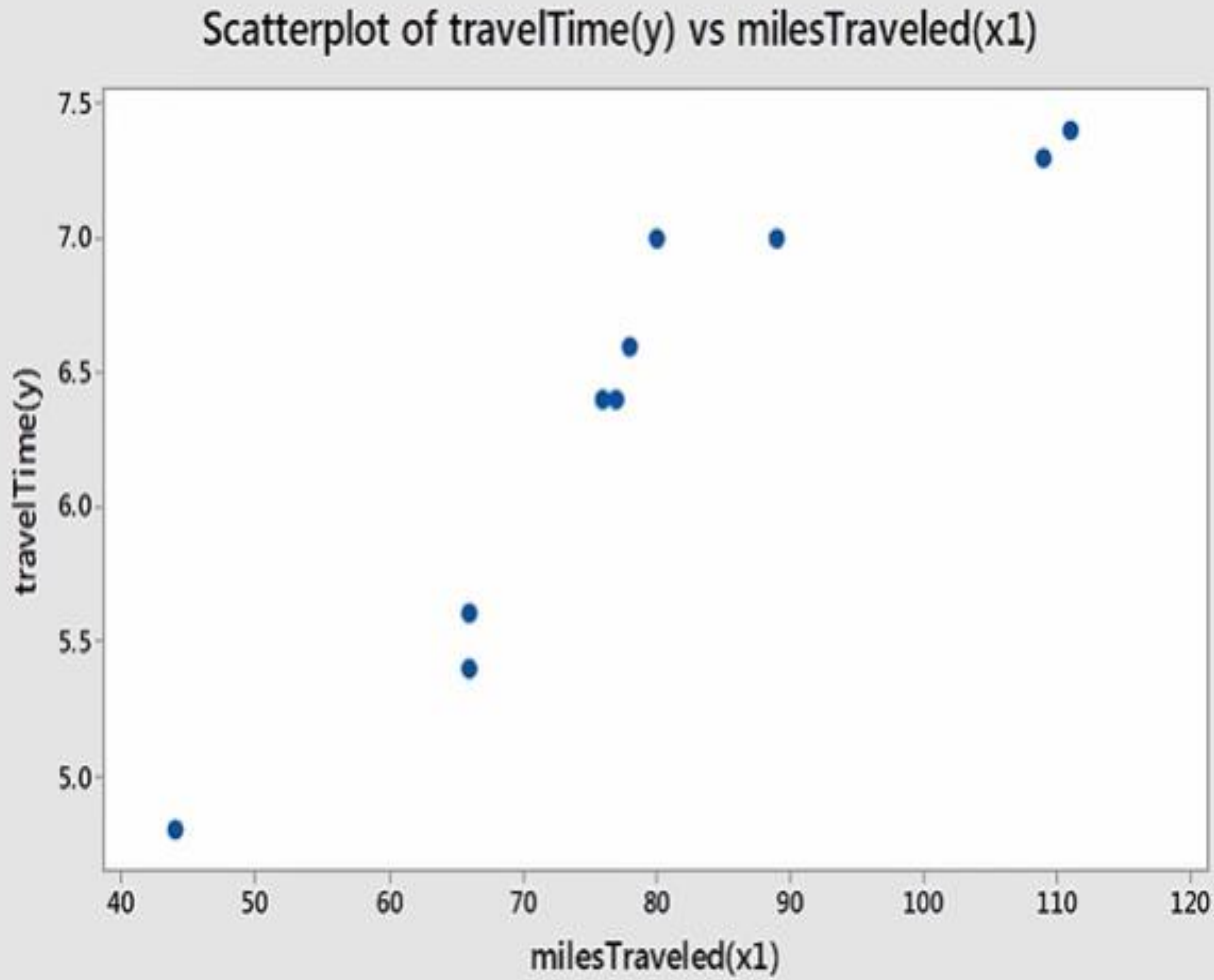
milesTraveled,(x_1)	numDeliveries,(x_2)	gasPrice,(x3)	travelTime(hrs),(y)
89	4	3.84	7
66	1	3.19	5.4
78	3	3.78	6.6
111	6	3.89	7.4
44	1	3.57	4.8
77	3	3.57	6.4
80	3	3.03	7
66	2	3.51	5.6
109	5	3.54	7.3
76	3	3.25	6.4

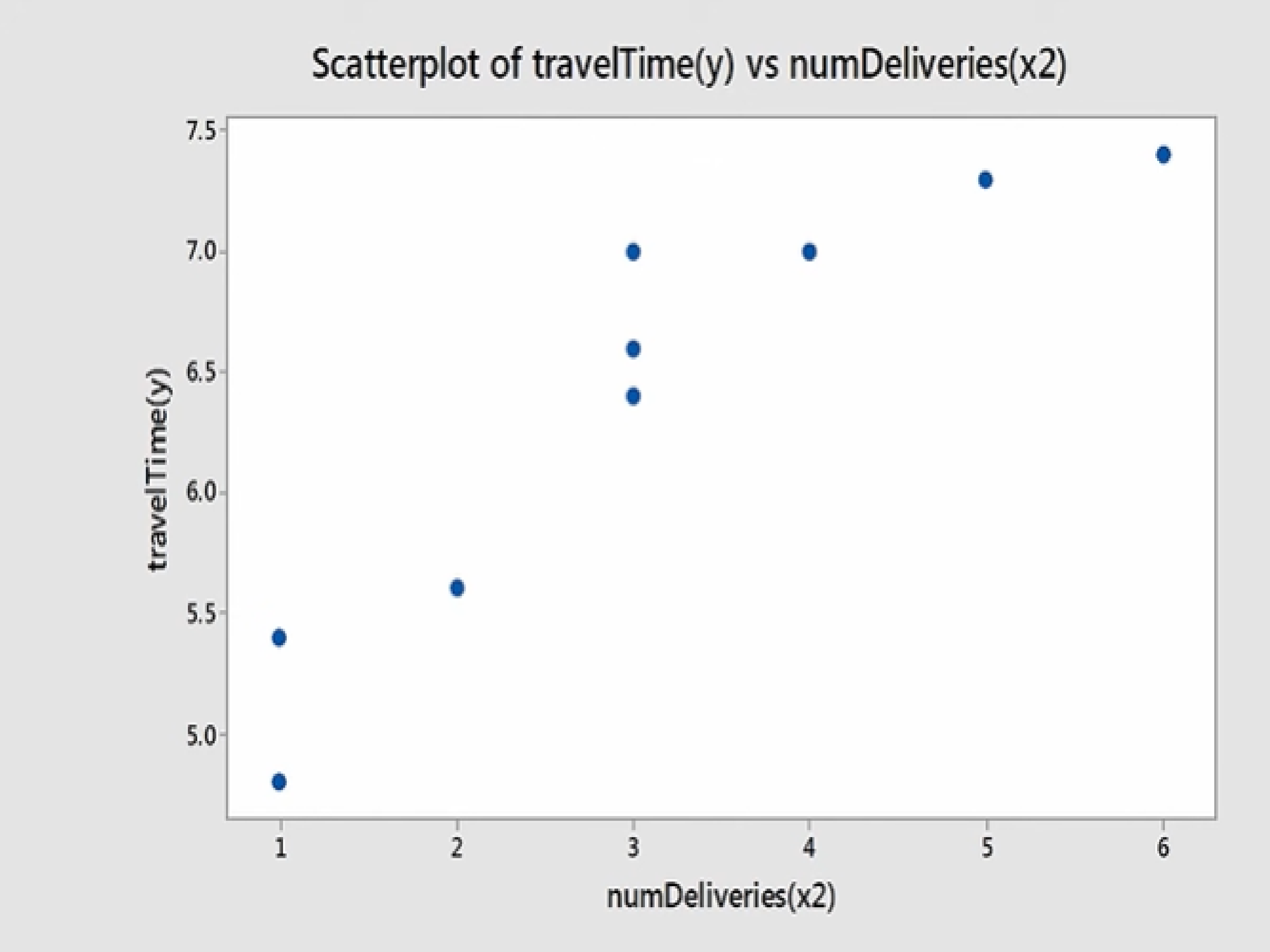
SKETCHING OUT RELATIONSHIPS

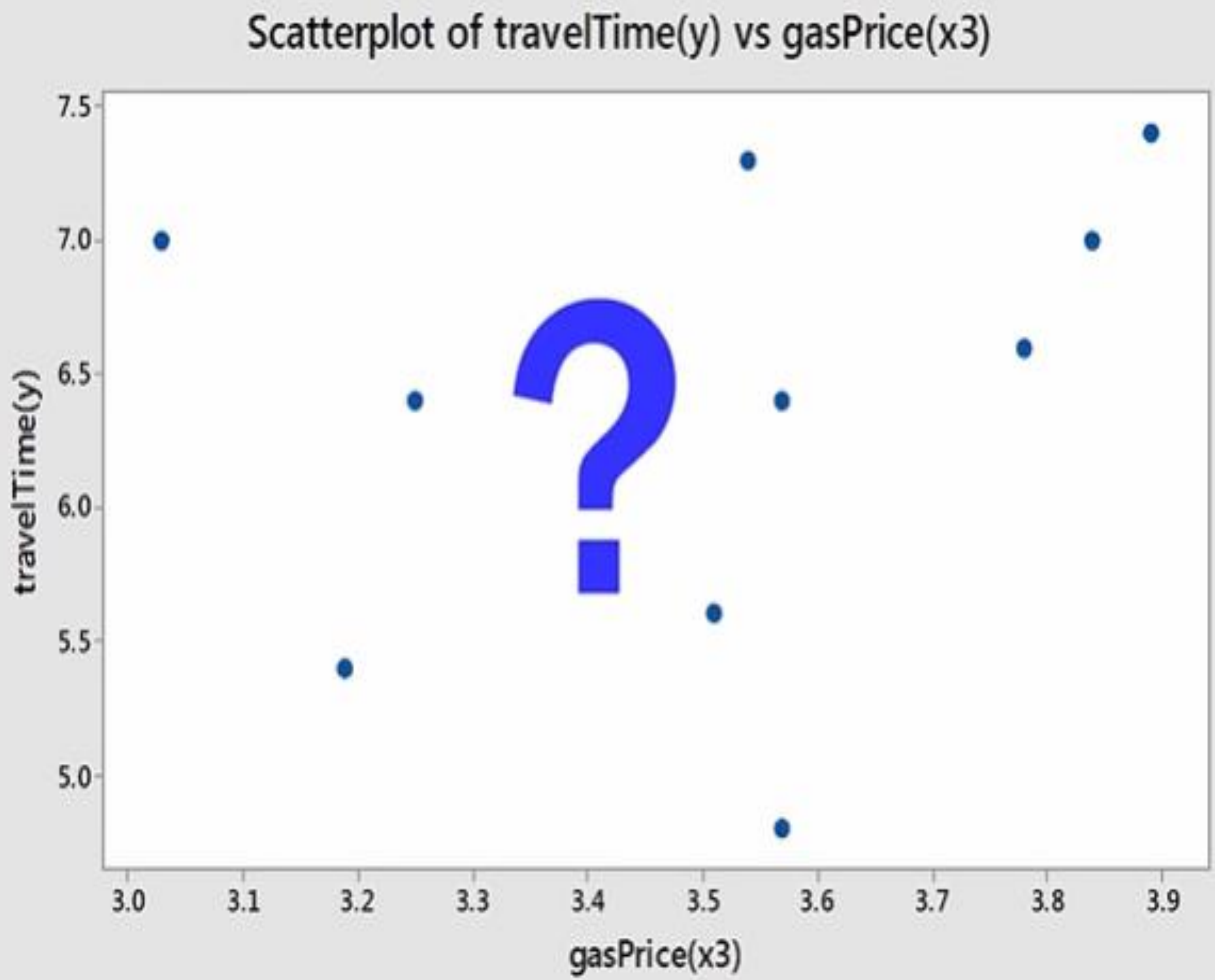


IV TO DV SCATTERPLOTS

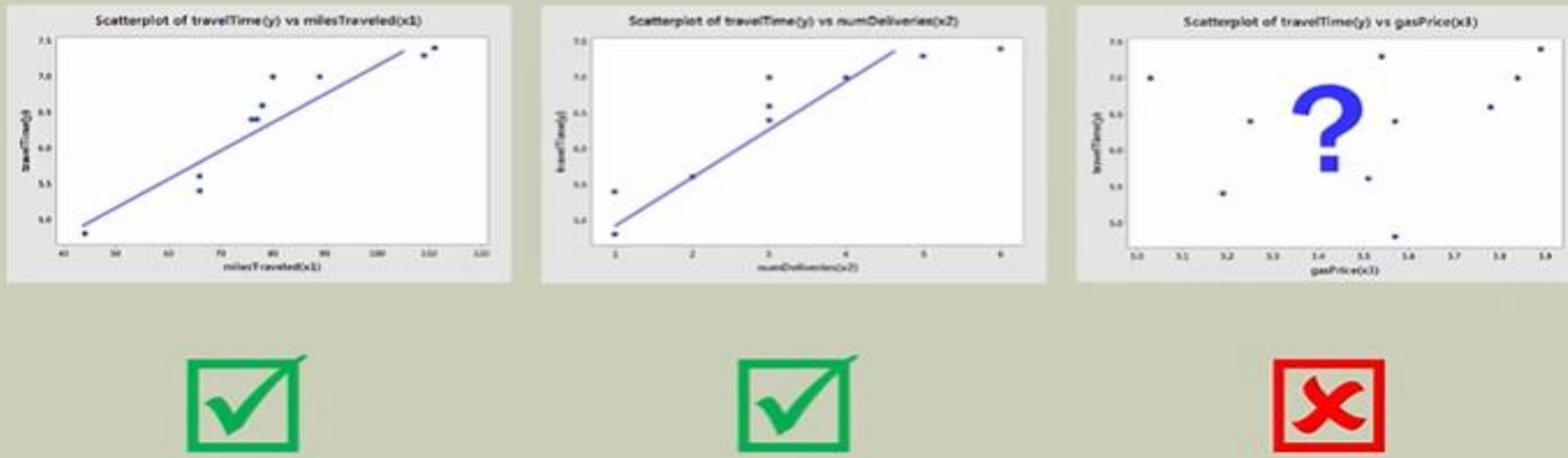
Relevancy
Check


Scatterplot of travelTime(y) vs milesTraveled(x1)
7.5
7.0
6.5
6.0
5.5
5.0
travelTime(y)
40
50
60
70
80
90
100
110
120
milesTraveled(x1)


Scatterplot of travelTime(y) vs numDeliveries(x2)
travelTime(y)
numDeliveries(x2)
7.5
7.0
6.5
6.0
5.5
5.0
1
2
3
4
5
6


Scatterplot of travelTime(y) vs gasPrice(x3)
travelTime(y)
7.5
7.0
6.5
6.0
5.5
5.0
3.0
3.1
3.2
3.3
3.4
3.5
3.6
3.7
3.8
3.9
gasPrice(x3)
?

DV VS IV SCATTERPLOTS

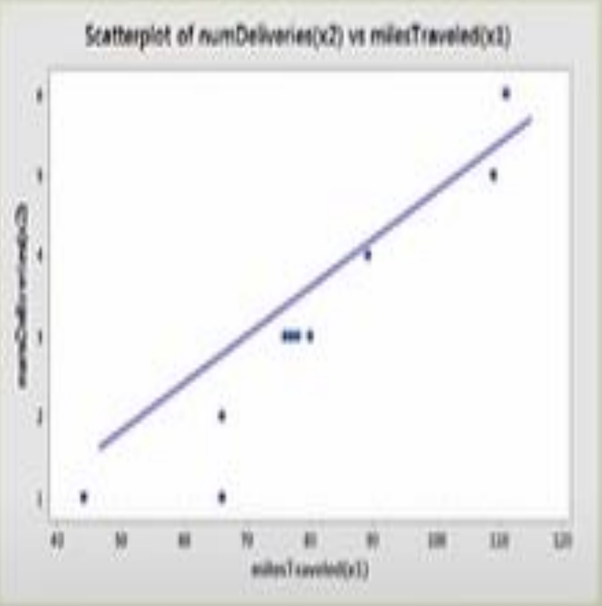


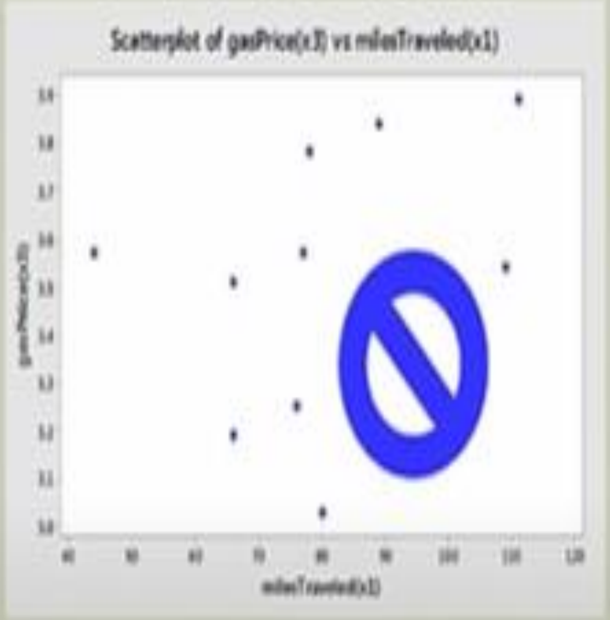
- Dependent variable vs independent variables
  - travelTime(y) appears highly correlated with milesTraveled(x_1)
  - travelTime(y) appears highly correlated with numDeliveries(x_2)
  - travelTime(y) DOES NOT appear highly correlated with gasPrice(x_3)

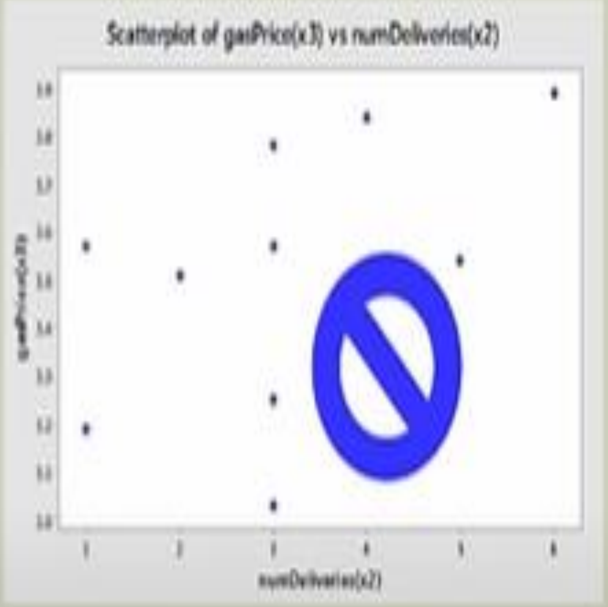
- *Note: for now, we will keep gasPrice in and then take it out later for learning purposes*

IV TO IV SCATTERPLOTS

IV SCATTERPLOTS (MULTICOLLINEARITY)







IV SCATTERPLOT SUMMARY

- Independent variable vs independent variable
  - numDeliveries(x_2) APPEARS highly correlated with milesTraveled(x_1); this is multicollinearity
  - milesTraveled(x_1) does not appear highly correlated with gasPrice(x_3)
  - gasPrices(x_3) does not appear correlated with numDeliveries(x_2)
- Since numDeliveries is HIGHLY CORRELATED with milesTraveled, we would NOT use BOTH in the multiple regression; they are *redundant*

CORRELATION SUMMARY

- Correlation analysis confirms the conclusions reached by visual examination of the scatterplots
- Redundant multicollinear variables
  - milesTraveled and numDeliveries are both highly correlated with each other and therefore are redundant; only one should used in the multiple regression analysis
- Non-contributing variables
  - gasPrice is NOT correlated with the depended variable and should be excluded

VARIABLE REGRESSIONS

- In this first step, we will perform a simple regression for each independent variable individually. The first will be conducted in Excel then the rest in Minitab (SPSS, SAS, JMP, R, etc. are all fine as well)
- We will discuss interpretations of results
- We will note how our results change:
  - Coefficients
    - Values, t-statistic, p-value
  - Analysis of Variance (ANOVA)
    - F-value, p-value
  - R-squared, R-squared(adjusted), R-squared(predicted)
  - VIF (Variance Inflation Factor)
  - Mallows C_p

MLR Equation

- $X_1 = F(X_2, X_{3,\ldots})$
- Regression equation of X_1, on X_2 and X_3
- $X_{1.23} = a_{1.23} + b_{12.3} X_2 + b_{13.2} X_3$
- Partial regression coefficients-$b_{12.3}$, $b_{13.2}$
- $b_{12.3}$ – amount by which a unit change in X_2 is expected to affect X_1 when X_3 is held constant
- X_1 varies partially because of variation in X_2 and partially because of X_3

Normal Equations

$y=a+bx$

$$\sum y = Na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

$y=a+b_1x_1+b_2x_2$

$$\sum y = Na + b_1\sum \mathrm{x}_1 + b_2\sum \mathrm{x}_2$$

$$\sum x_1 y = a\sum x_1 + b_1\sum {x_1}^2 + b_2\sum \mathrm{x}_1\mathrm{x}_2$$

$$\sum x_2 y = a\sum x_2 + b_1\sum \mathrm{x}_1\mathrm{x}_2 + b_2\sum {\mathrm{x}_2}^2$$

Multiple Regression

General Parametric Equation:

$$y = \underline{f(\boldsymbol{X})} + \epsilon$$

Depends on Statistical Method

$$f(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \cdots + \widehat{\beta_p} X_p$$

For n samples, number of operations = $n \times (p-1)^2$

Multiple Regression

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & . & . & X_{1,p} \\ 1 & X_{2,1} & X_{2,2} & . & . & . \\ 1 & X_{3,1} & X_{3,2} & . & . & . \\ 1 & . & . & . & . & . \\ 1 & . & . & . & . & . \\ 1 & X_{n,1} & X_{n,2} & . & . & X_{n,p} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \qquad \hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

$$
\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_{1,1} + \beta_2 X_{1,2} + \cdots + \beta_p X_{1,p} + \epsilon_1 \\ \beta_0 + \beta_1 X_{2,1} + \beta_2 X_{2,2} + \cdots + \beta_p X_{2,p} + \epsilon_2 \\ \beta_0 + \beta_1 X_{3,1} + \beta_2 X_{3,2} + \cdots + \beta_p X_{3,p} + \epsilon_3 \\ \cdot \\ \cdot \\ \cdot \\ \beta_0 + \beta_1 X_{n,1} + \beta_2 X_{n,2} + \cdots + \beta_p X_{n,p} + \epsilon_n \end{bmatrix}
$$

Multiple Regression

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

$$\boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ . \\ . \\ . \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 - \widehat{y_1} \\ y_2 - \widehat{y_2} \\ y_3 - \widehat{y_3} \\ . \\ . \\ . \\ y_n - \widehat{y_n} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{bmatrix} - \begin{bmatrix} \widehat{y_1} \\ \widehat{y_2} \\ \widehat{y_3} \\ . \\ . \\ . \\ \widehat{y_n} \end{bmatrix} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$$

$$RSS = \sum_{i=1}^{n} e_i^2 \longrightarrow RSS = \boldsymbol{e}^T \boldsymbol{e}$$

$$RSS = \boldsymbol{e}^T \boldsymbol{e}$$

$$RSS = (\boldsymbol{y} - \widehat{\boldsymbol{y}})^T (\boldsymbol{y} - \widehat{\boldsymbol{y}})$$

$$RSS = \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$$

$$RSS = \left(\boldsymbol{y}^T - \widehat{\boldsymbol{\beta}^T}\boldsymbol{X}^T\right)(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$$

$$RSS = \boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\boldsymbol{X}^T \boldsymbol{y} + \widehat{\boldsymbol{\beta}^T}\boldsymbol{X}^T \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

Matrix Differentiation

x = m x 1 matrix
A = n x m matrix; $A \perp x$

$$y = A \rightarrow \frac{\delta y}{\delta x} = 0$$

$$y = Ax \rightarrow \frac{\delta y}{\delta x} = A$$

$$y = xA \rightarrow \frac{\delta y}{\delta x} = A^T$$

$$y = x^T A x \rightarrow \frac{\delta y}{\delta x} = 2x^T A$$

$$RSS = y^T y - y^T X\hat{\beta} - \hat{\beta} X^T y + \widehat{\beta^T} X^T X\hat{\beta}$$

$$\frac{\delta(RSS)}{\delta\hat{\beta}} = \frac{\delta\left(y^T y - y^T X\hat{\beta} - \hat{\beta} X^T y + \widehat{\beta^T} X^T X\hat{\beta}\right)}{\delta\hat{\beta}} = 0$$

$$\frac{\delta(y^T y)}{\delta\hat{\beta}} - \frac{\delta(y^T X\hat{\beta})}{\delta\hat{\beta}} - \frac{\delta(\hat{\beta} X^T y)}{\delta\hat{\beta}} + \frac{\delta(\widehat{\beta^T} X^T X\hat{\beta})}{\delta\hat{\beta}} = 0$$

$$0 - y^T X - (X^T y)^T + 2\hat{\beta}^T X^T X = 0$$

$$0 - y^T X - y^T X + 2\hat{\beta}^T X^T X = 0$$

$$2\hat{\beta}^T X^T X = 2y^T X \qquad \hat{\beta}^T X^T X = y^T X$$

$$\hat{\beta}^T = y^T X (X^T X)^{-1} \qquad \hat{\beta} = (X^T X)^{-1} X^T y$$

Multiple Regression

$$\{ (X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n) \}$$

$$\Big\downarrow \text{ Least Squares Criteria}$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

$$\Big\downarrow$$

Predict House Price $\longrightarrow$ $\widehat{\boldsymbol{y}} = \widehat{f(\boldsymbol{X})} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$

Example

Method 1- Normal Equations

y	x_1	x_2
4	15	30
6	12	24
7	8	20
9	6	14
13	4	10
15	3	4
54	**48**	**102**

x_1y	x_2y	x_1x_2	x_1^2	X_2^2
60	120	450	225	900
72	144	288	144	576
56	140	160	64	400
54	126	84	36	196
52	130	40	16	100
45	60	12	9	16
339	**720**	**1034**	**494**	**2188**

$$54 = 6a + 48b_1 + 102b_2$$

$$339 = 48a + 494b_1 + 1034b_2$$

$$720 = 102a + 1034b_1 + 2188b_2$$

$y=16.47+0.38x_1-0.62x_2$

Example

y	x_1	x_2
2	3	4
4	5	6
6	7	8
8	9	10
20	24	28

x_1y	x_2y	x_1x_2	x_1^2	X_2^2
6	8	12	9	16
20	36	30	25	36
42	48	56	49	64
72	80	90	81	100
140	172	188	164	216

$$20 = 4a + 24b_1 + 28b_2$$

$$140 = 24a + 164b_1 + 188b_2$$

$$172 = 28a + 188b_1 + 216b_2$$

$y=0+2x_1-1x_2$

MLR

Method 2-

Deviations taken from mean

$$b_1 = \frac{[(\sum \mathrm{x_2}^2)(\sum x_1 y) - (\sum \mathrm{x_1 x_2})(\sum x_2 y)]}{(\sum x_1{}^2)(\sum x_2{}^2) - \sum \mathrm{x_1 x_2}^2}$$

$$b_2 = \frac{[(\sum \mathrm{x_1}^2)(\sum x_2 y) - (\sum \mathrm{x_1 x_2})(\sum x_1 y)]}{(\sum x_1{}^2)(\sum x_2{}^2) - \sum \mathrm{x_1 x_2}^2}$$

$$b_0 = \bar{y} - b_1 \overline{x_1} - b_2 \overline{x_2}$$

MLR - two independent variables

y	X_1	X_2
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11

	y	X_1	X_2
	140	60	22
	155	62	25
	159	67	24
	179	70	20
	192	71	15
	200	72	14
	212	75	14
	215	78	11
Mean	181.5	69.375	18.125
Sum	1452	555	145

	X_1^2	X_2^2	X_1y	X_2y	X_1X_2
	3600	484	8400	3080	1320
	3844	625	9610	3875	1550
	4489	576	10653	3816	1608
	4900	400	12530	3580	1400
	5041	225	13632	2880	1065
	5184	196	14400	2800	1008
	5625	196	15900	2968	1050
	6084	121	16770	2365	858
Sum	38767	2823	101895	25364	9859

$$\Sigma x_1^2 = \Sigma X_1^2 - (\Sigma X_1)^2 / n = 38{,}767 - (555)^2 / 8 = \mathbf{263.875}$$
$$\Sigma x_2^2 = \Sigma X_2^2 - (\Sigma X_2)^2 / n = 2{,}823 - (145)^2 / 8 = \mathbf{194.875}$$
$$\Sigma x_1 y = \Sigma X_1 y - (\Sigma X_1 \Sigma y) / n = 101{,}895 - (555*1{,}452) / 8 = \mathbf{1{,}162.5}$$
$$\Sigma x_2 y = \Sigma X_2 y - (\Sigma X_2 \Sigma y) / n = 25{,}364 - (145*1{,}452) / 8 = \mathbf{-953.5}$$
$$\Sigma x_1 x_2 = \Sigma X_1 X_2 - (\Sigma X_1 \Sigma X_2) / n = 9{,}859 - (555*145) / 8 = \mathbf{-200.375}$$

$$\mathbf{\hat{y} = -6.867 + 3.148x_1 - 1.656x_2}$$

∴

Example 16.3. A modulation study on *R. Trifoli* yields the following data.

	Dry weight of plants (mg) Y	Root length (cm) X_1	Shoot length (cm) X_2
	412	28.7	21.5
	226	13.4	11.7
	292	14.6	12.9
	323	18.0	14.8
	233	12.1	11.0
	368	23.4	19.2
	239	12.6	11.4
	382	30.2	22.6
	218	11.6	10.8
	222	12.0	10.2
	214	12.4	10.1
Total	3129	189.0	156.2

$$\hat{Y} = 40.96 - 6.30X_1 + 24.77X_2$$

SUBJECT	Y	X_1	X_2	X_1X_1	X_2X_2	X_1X_2	X_1Y	X_2Y
1	-3.7	3	8	9	64	24	-11.1	-29.6
2	3.5	4	5	16	25	20	14	17.5
3	2.5	5	7	25	49	35	12.5	17.5
4	11.5	6	3	36	9	18	69	34.5
5	5.7	2	1	4	1	2	11.4	5.7
Σ	19.5	20	24	90	148	99	95.8	45.6

Final Regression equation or Model is:

$$\boldsymbol{Y = 2.796 + 2.28\, x_1 - 1.67 x_2}$$

$Now\ given \quad x_1 = 3 \ \ and \quad x_2 = 2 \qquad Y = ?$

$$Y = 2.796 + 2.28 * 3 - 1.67 * 2$$
$$= \mathbf{6.296}$$

Method 3
(Yule's Notation)

Finding regression coefficient from correlation coefficients (deviation from mean)

(i) X_1 on X_2 and X_3

$$X_1 = b_{12.3} X_2 + b_{13.2} X_3$$

$$b_{12.3} = \frac{\sigma_1}{\sigma_2} \times \frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2}$$

$$b_{13.2} = \frac{\sigma_1}{\sigma_3} \times \frac{r_{13} - r_{23} r_{12}}{1 - r_{23}^2}$$

$b_{12.3}$ = Partial Regression coefficient of X_1 on X_2

$b_{13.2}$ = Partial Regression coefficient of X_1 on X_3

Example

Given the following, determine the regression equation of :
(i) X_1 on X_2 and X_3. (ii) X_2 on X_1 and X_3.
$r_{12} = 0.8, r_{13} = 0.6, r_{23} = 0.5, \sigma_1 = 10, \sigma_2 = 8, \sigma_3 = 5.$

(i) X_1 on X_2 and X_3

1- 0.25

$$X_1 = b_{12.3} X_2 + b_{13.2} X_3$$

b12.3

$$b_{12.3} = \frac{\sigma_1}{\sigma_2} \times \frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2} = \frac{10}{8} \times \frac{0.8 - (0.6 \times 0.5)}{1 - (0.5)^2} = \frac{1.25 \times 0.50}{0.75} = 0.833$$

$$b_{13.2} = \frac{\sigma_1}{\sigma_3} \times \frac{r_{13} - r_{23} r_{12}}{1 - r_{23}^2} = \frac{10}{5} \times \frac{0.6 - (0.8 \times 0.5)}{1 - (0.5)^2} = 2 \times \frac{0.20}{0.75} = 0.533$$

Therefore required regression equation is

$$X_1 = 0.833 X_2 + 0.533 X_3$$

(ii) Regression equation of X_2 on X_1 and X_3.

$$X_2 = b_{21.3} X_1 + b_{23.1} X_3$$

$$b_{21.3} = \frac{\sigma_2}{\sigma_1} \times \frac{r_{12} - r_{13} \cdot r_{23}}{1 - r_{13}^2} = \frac{8}{10} \times \frac{0{\cdot}8 - (0{\cdot}6 \times 0{\cdot}5)}{1 - (0{\cdot}6)^2} = 0{\cdot}8 \times \frac{0{\cdot}50}{0{\cdot}64} = 0{\cdot}625$$

$$b_{23.1} = \frac{\sigma_2}{\sigma_3} \times \frac{r_{23} - r_{13} r_{12}}{1 - r_{13}^2} = \boxed{\frac{8}{5}} \times \frac{0{\cdot}5 - (0{\cdot}6 \times 0{\cdot}8)}{1 - (0{\cdot}6)^2} = 1{\cdot}6 \times \frac{0{\cdot}02}{0{\cdot}64} = 0{\cdot}05$$

$$X_2 = 0{\cdot}625\, X_1 + 0{\cdot}05\, X_3$$

Example

Given $r_{12} = 0{\cdot}28$, $r_{23} = 0{\cdot}49$, $r_{31} = 0{\cdot}51$, $\sigma_1 = 2{\cdot}7$, $\sigma_2 = 2{\cdot}4$, $\sigma_3 = 2{\cdot}7$. Find the regression equation of X_3 on X_1 and X_2. [B.A. Bombay Uni

Given $r_{12} = 0{\cdot}28$, $r_{23} = 0{\cdot}49$, $r_{31} = 0{\cdot}51$, $\sigma_1 = 2{\cdot}7$, $\sigma_2 = 2{\cdot}4$, $\sigma_3 = 2{\cdot}7$. Find the regression equation of X_3 on X_1 and X_2. [B.A. Bombay Univ. 19

Solution :

The regression equation of X_3 on X_1 and X_2 is

$$X_3 = b_{31.2}\, X_1 + b_{32.1}\, X_2$$

$$b_{31.2} = \frac{\sigma_3}{\sigma_1} \times \frac{r_{13} - (r_{12} \times r_{23})}{1 - r_{12}^2} = \frac{2{\cdot}7}{2{\cdot}7} \times \frac{0{\cdot}51 - (0{\cdot}28 \times 0{\cdot}49)}{1 - (0{\cdot}28)^2}$$

$$= 1 \times \frac{0{\cdot}3728}{0{\cdot}9216} = 1 \times 0{\cdot}4045 = 0{\cdot}4045$$

$$b_{32.1} = \frac{\sigma_3}{\sigma_2} \times \frac{r_{23} - (r_{12} \times r_{13})}{1 - r_{12}^2} = \frac{2{\cdot}7}{2{\cdot}4} \times \frac{0{\cdot}49 - (0{\cdot}28 \times 0{\cdot}51)}{1 - (0{\cdot}28)^2}$$

$$= 1{\cdot}125 \times \frac{0{\cdot}49 - 0{\cdot}1428}{0{\cdot}9216} = 1{\cdot}125 \times 0{\cdot}3767 = 0{\cdot}4238$$

$$X_3 = 0{\cdot}4045\, X_1 + 0{\cdot}4238\, X_2$$

Example

Similarly, ... $1 - (2 \times 5) - 6 = 4$.

Example 41. *The following table shows the corresponding values of three variables. X_1, X_2 and X_3. Find the least square regression equation of X_3 on X_1 and X_2. Estimate X_3 when X_1 … 0 and $X_2 = 6$.*

X_1 :	3	5	6	8	12	14
X_2 :	16	10	7	4	3	2
X_3 :	90	72	54	42	30	12.

Solution: The regression equation of X_3 on X_2 and X_1 can be written as follows :

$$X_3 - \bar{X}_3 = \left(\frac{r_{23} - r_{13}r_{12}}{1 - r_{12}^2}\right)\left(\frac{S_3}{S_2}\right)(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{23}r_{12}}{1 - r_{13}^2}\right)\left(\frac{S_3}{S_1}\right)(X_1 - \bar{X}_1).$$

$$S_1 = \sqrt{\frac{\Sigma (X_1 - \bar{X}_1)^2}{N}}$$

$$S_2 = \sqrt{\frac{\Sigma (X_2 - \bar{X}_2)^2}{N}}$$

$$S_3 = \sqrt{\frac{\Sigma (X_3 - \bar{X}_3)^2}{N}}$$

$$r_{12} = \frac{\Sigma x_1 x_2}{\sqrt{\Sigma x_1^2 \times \Sigma x_2^2}}$$

$$r_{13} = \frac{\Sigma x_1 x_3}{\sqrt{\Sigma x_1^2 \times \Sigma x_3^2}}$$

$$r_{23} = \frac{\Sigma x_2 x_3}{\sqrt{\Sigma x_2^2 \times \Sigma x_3^2}}$$

Example

Table : Computation of $\bar{X}_1, \bar{X}_2, \bar{X}_3; S_1, S_2, S_3; r_{12}, r_{13}, r_{23}$

	$(X_1 - \bar{X}_1) = x_1$			$(X_2 - \bar{X}_2) = x_2$			$(X_3 - \bar{X}_3) = x_3$				
X_1	x_1	x_1^2	X_2	x_2	x_2^2	X_3	x_3	x_3^2	x_1x_2	x_1x_3	x_2x_3
3	−5	25	16	+9	81	90	+40	1600	−45	−200	+360
5	−3	9	10	+3	9	72	+22	484	−9	−66	+66
6	−2	4	7	0	0	54	+4	16	0	−8	0
8	0	0	4	−3	9	42	−8	64	0	0	+24
12	+4	16	3	−4	16	30	−20	400	−16	−80	+80
14	+6	36	2	−5	25	12	−38	1444	−30	−228	+190
$\Sigma X_1 = 48$	$\Sigma x_1 = 0$	$\Sigma x_1^2 = 90$	$\Sigma X_2 = 42$	$\Sigma x_2 = 0$	$\Sigma x_2^2 = 140$	$\Sigma X_3 = 300$	$\Sigma x_3 = 0$	$\Sigma x_3^2 = 4008$	$\Sigma x_1x_2 = -100$	$\Sigma x_1x_3 = -582$	$\Sigma x_2x_3 = 720$

$$\bar{X}_1 = \frac{48}{6} = 8; \quad \bar{X}_2 = \frac{42}{6} = 7; \quad \bar{X}_3 = \frac{300}{6} = 50.$$

$$S_1 = \sqrt{\frac{\Sigma (X_1 - \bar{X}_1)^2}{N}} = \sqrt{\frac{90}{6}} = \sqrt{15} = 3.87.$$

$$S_2 = \sqrt{\frac{\Sigma (X_2 - \bar{X}_2)^2}{N}} = \sqrt{\frac{140}{6}} = \sqrt{23.33} = 4.83.$$

$$S_3 = \sqrt{\frac{\Sigma (X_3 - \bar{X}_3)^2}{N}} = \sqrt{\frac{4008}{6}} = \sqrt{668} = 25.85.$$

$$r_{12} = \frac{\Sigma x_1 x_2}{\sqrt{\Sigma x_1^2 \times \Sigma x_2^2}} = \frac{-100}{\sqrt{90 \times 140}} = \frac{-100}{112.25} = -0.891.$$

$$r_{13} = \frac{\Sigma x_1 x_3}{\sqrt{\Sigma x_1^2 \times \Sigma x_3^2}} = \frac{-582}{\sqrt{90 \times 4008}} = \frac{-582}{600.599} = -0.969.$$

$$r_{23} = \frac{\Sigma x_2 x_3}{\sqrt{\Sigma x_2^2 \times \Sigma x_3^2}} = \frac{720}{\sqrt{140 \times 4008}} = \frac{720}{749.08} = 0.961.$$

$$X_3 - 50 = \left[\frac{0.961 - (-0.969 \times 0.891)}{1 - (-0.9)^2}\right]\left(\frac{25.85}{4.83}\right)(X_2 - 7)$$

$$+ \left[\frac{-0.969 - (0.961 \times -0.891)}{1 - (-0.9)^2}\right]\left(\frac{25.85}{3.87}\right)(X_1 - 8)$$

$$x_3 - 50 = 2.546\,(X_2 - 7) - 3.664\,(X_1 - 8)$$

$$x_3 - 50 = 2.546\,X_2 - 17.822 - 3.664\,X_1 + 29.312$$

$$X_3 = 2.546\,X_2 - 3.664\,X_1 + 61.49,$$

equired regression equation of X_3 on X_1 and X_2.

F-test of overall significance in regression analysis

The F-Test of overall significance in regression is a test of whether or not your linear regression model provides a better fit to a dataset than a model with no predictor variables.

F-test of overall significance in regression analysis

- The F-statistic is calculated as regression MS/residual MS.
- This statistic indicates whether the regression model provides a better fit to the data than a model that contains no independent variables.
- In essence, it tests if the regression model as a whole is useful.
- If the $P <$ the significance level, there is sufficient evidence to conclude that the regression model fits the data better than the model with no predictor variables.
- This finding is good because it means that the predictor variables in the model actually improve the fit of the model.
- In general, if none of the predictor variables in the model are statistically significant, the overall F statistic is also not statistically significant.

Example

Estimating output (Y) of physiotherapist from a knowledge of his/her test score on the aptitude test (X_1) and years of experience (X_2) in a hospital

X_1	X_2	Y
160	5.5	32
80	6.0	15
112	9.5	30
185	5.0	34
152	8.0	35
90	3.0	10
170	9.0	39
140	5.0	26
115	0.5	11
150	1.5	23

Test the following hypotheses at α=0.05

H_0: $Y = b_0$

H_1: $Y = b_0 + b_1X_1 + b_2X_2$

Y	X_1	X_2	X_1Y	X_2Y	X_1X_2	X_1^2	X_2^2
32	160	5.5	5120	176	880	25600	30.25
15	80	6.0	1200	90	480	6400	36
30	112	9.5	3360	285	1064	12544	90.25
34	185	5.0	6290	170	925	34225	25
35	152	8.0	5320	280	1216	23104	64
10	90	3.0	900	30	270	8100	9
39	170	9.0	6630	351	1530	28900	81
26	140	5.0	3640	130	700	19600	25
11	115	0.5	1265	5.5	57.5	13225	0.25
23	150	1.5	3450	34.5	225	22500	2.25
255	1354	53	37175	1552	7347.5	194128	363

Accordingly, the three equations are:

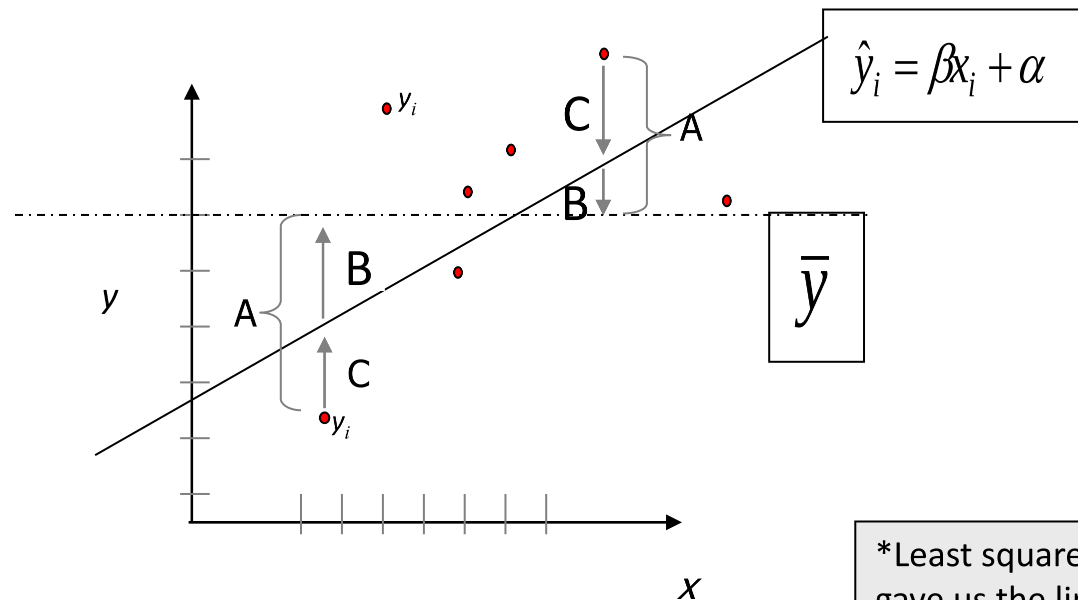
$255 = 10\ b_0 + 1354\ b_1 + 53\ b_2$

$37175 = 1354\ b_0 + 194128\ b_1 + 7374.5\ b_2$

$1552 = 53\ b_0 + 7347.5\ b_1 + 363\ b_2$

Solving the three equations simultaneously, we obtain $b_0 = -13.824567$, $b_1 = 0.212167$, and $b_2 = 1.999461$. Thus, the regression equation of Y on X_1 and X_2 is: $Y_C = -13.824567 + 0.212167\ X_2 + 1.999461\ X_2$.

Regression Picture



*Least squares estimation gave us the line (β) that minimized C^2

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

$R^2 = SS_{reg}/SS_{total}$

A^2

SS_{total}

Total squared distance of observations from naïve mean of y

Total variation

B^2

SS_{reg}

Distance from regression line to naïve mean of y

Variability due to x (regression)

C^2

$SS_{residual}$

Variance around the regression line

Additional variability not explained by x—what least squares method aims to minimize

Y	X_1	X_2	Y_c	$(Y-\bar{Y})^2$	$Y-Y_c$	$(Y-Y_c)^2$	$(Y_c-\bar{Y})^2$	Std residual
32	160	5.5	31.119	42.25	0.881	0.776	31.575	0.780
15	80	6.0	15.146	110.25	−0.146	0.022	107.214	−0129
30	112	9.5	28.933	20.25	1.067	1.138	11.786	0.945
34	185	5.0	35.424	72.25	−1.424	2.027	98.479	−1.260
35	152	8.0	34.421	90.25	0.579	0.336	79.576	0.513
10	90	3.0	11.269	240.25	−1.269	1.610	202.526	−1.123
39	170	9.0	40.239	182.25	−1.239	1.536	217.238	−1.097
26	140	5.0	25.876	0.25	0.123	0.0153	0.141	0.110
11	115	0.5	11.574	210.25	−0.574	0.330	193.922	−0.509
23	150	1.5	21.000	6.25	2.000	4.001	20.253	1.771
255	1354	53		974.5		11.791	962.710	

Y_c: Predicted Y, $Y-Y_c$: Residual

Total variation (sum of squares total, SST) $\sum(Y-\bar{Y})^2$ =974.5.

Explained variation (sum of square regression, SSR) $\sum(Y_c-\bar{y})^2$ =962 710

Unexplained variation (sum of squares error, SSE) $\sum(Y-\bar{Y}_c)^2$ =11.791

R square (R^2) $= SSR/SST = 962.710/974.5$ =0.988, R= 0.984

Mean square regression (MS_R) $= SSR/df = 962.710/2$ =481.355

Mean square error (MS_E) $SSE/df = 11.791/7$ =1.684

F = $MS_R/MS_E = 481.355/1.684$ =285 775

The degrees of freedom in a multiple regression equals *N-k-1*, where *k* is the number of variables.

n_2	$P=0.05$				
	n_1				
	1	2	3	4	6
6	5.99	5.14	4.76	4.53	4.28
7	5.59	4.74	4.35	4.12	3.97
8	5.32	4.46	4.07	3.84	3.69
9	5.12	4.26	3.86	3.63	3.48

Since 285.75>4.74 we reject null hypothesis and conclude that model is significant with predictor variables.

H_0: $Y = b_0$

H_1: $Y = b_0 + b_1X_1 + b_2X_2$

F-statistics

The *F* statistic represents the ratio of the variance explained by the regression model (regression mean square) to the not explained variance (residuals mean square). It can be calculated easily using an online calculator in comparison to the manual approach. The F-test of overall significance tests whether all of the predictor variables are jointly significant while the *t*-test of significance for each individual predictor variable merely tests whether each predictor variable is individually significant. Thus, the F-test determines whether or not all of the predictor variables are jointly significant. It is possible that each predictor variable is not significant and yet the F-test says that all of the predictor variables combined are jointly significant.

Hypothesis Testing in Multiple Linear Regression

$$H_o \; : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1 \; : \beta_j \neq 0 \text{ for at least one j}$$

Rejection of H_0 implies that at least one of the regressor variables X_1, X_2, …,X_k contributes significantly to the linear regression model.

The F-test

- For a multiple regression model with intercept, we want to test the following null hypothesis and alternative hypothesis:

  H_0: $\beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$

  H_1: $\beta_j \neq 0$, for at least one value of j

  This test is known as the overall **F-test for regression**.

- Here are the five steps of the **overall F-test for regression**

  1. State the null and alternative hypotheses:

     H_0: $\beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$

     H_1: $\beta_j \neq 0$, for at least one value of j

  2. Compute the test statistic assuming that the null hypothesis is true:

     F = ~~MSM / MSE~~ = (explained variance) / (unexplained variance)

  3. Find a $(1 - \alpha)100\%$ confidence interval I for (DFM, DFE) degrees of freedom using an F-table or statistical software.

  4. Accept the null hypothesis if $F \in I$; reject it if $F \notin I$.

- **Practice Problem:** For a multiple regression model with 35 observations and 9 independent variables (10 parameters), SSE = 134 and SSM = 289, test the null hypothesis that all of the regression parameters are zero at the 0.05 level.

Solution: DFE = n - p = 35 - 10 = 25 and DFM = p - 1 = 10 - 1 = 9. Here are the five steps of the test of hypothesis:

1. State the null and alternative hypothesis:

   H_0: $\beta_1 = \beta_2 = , \ldots , = \beta_{p-1} = 0$

   H_1: $\beta_j \neq 0$ for some j

2. Compute the test statistic:

   F = MSM/MSE = (SSM/DFM) / (SSE/DFE) = (289/9) / (134/25) = 32.111 / 5.360 = 5.991

3. Find a (1 - 0.05)×100% confidence interval for the test statistic. Look in the F-table at the 0.05 entry for 9 df in the numerator and 25 df in the denominator. This entry is 2.28, so the 95% confidence interval is [0, 2.34]. This confidence interval can also be found using the R function call qf(0.95, 9, 25).

4. Decide whether to accept or reject the null hypothesis: 5.991 $\notin$ [0, 2.28], so reject H_0.

F - Distribution ($\alpha = 0.05$ in the Right Tail)

Numerator Degrees of Freedom (df_1); Denominator Degrees of Freedom (df_2)

df_2 \ df_1	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123
4	7.7086	9.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	6.9988
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	2.0868	2.0164	1.9588
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799

- The t tests are used to conduct hypothesis tests on the regression coefficients obtained in simple linear regression. A statistic based on the t distribution is used to test the two-sided hypothesis that the true slope, β1, equals some constant value, $\beta_{1,0}$. The statements for the hypothesis test are expressed as:

- $H0{:}\beta 1 = \beta_{1,0}$
- $H1{:}\beta 1 \neq \beta_{1,0}$

Test for Significance of the Regression Slope Coefficient

- Hypotheses:

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

A slope of 0 implies there is NO LINEAR RELATIONSHIP between x and y, and that x in its linear form is of no use in explaining the variation in y.

- Testing Approaches: Critical Value and p-value

$$t = \frac{b_1 - \beta_1}{s_{b_1}} \quad df = n - 2$$

b_1 - Sample regression slope coefficient
β_1 - Hypothesized slope (usually $\beta_1 = 0$)
s_{b_1} - Estimator of the standard error of the slope

If p-value < $\alpha / 2$, reject H_0

Simple Linear Regression Analysis - Let's Practice

b) $H_0: B_1 = 0.0$ (no linear relationship; not significant)

$H_a: B_1 \neq 0.0$ (there is a linear relationship; significant)

$\alpha = 0.05$ Degrees of Freedom = n – 2 = 19

=T.INV.2T(α, df) Critical t (Appendix F) = $\pm$ 2.093

		Coefficients	*Standard Error*	*t Stat*	*P-value*
b_0	Intercept	171205.8279	59846.1252	2.8608	0.0100
b_1	Store Size (Sq. Ft.)	25.3160	3.5767	7.0780	0.0000

$$t = \frac{b_1 - B_1}{s_{b_1}} = \frac{25.316 - 0}{3.5767} = 7.08$$

Conclusion: Since 7.08 is greater than CV 2.093, Reject H_0 and conclude that the population slope coefficient is significant, there is a linear relationship.

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745

The Coefficient of Determination R^2

- Portion of the total variation in the dependent variable that is explained by its relationship with the independent variable (between 0 and 1.0)

$$\boxed{R^2 = \frac{SSR}{SST}}$$

SSR - Sum of squares regression
SST - Total sum of squares

- Coefficient of Determination for the Single Independent Variable Case

$$\boxed{R^2 = r^2}$$

r - Sample correlation coefficient

Simple Linear Regression Analysis - Let's Practice

c) R^2 = Coefficient of Determination

$$\frac{SSR}{SST} = \frac{82230575305}{113416868002} = 0.7250$$

	df	SS	MS	F	Significance F
Regression	1	82230575305.3160	82230575305.3160	50.0983	0.0000
Residual	19	31186292697.2554	1641383826.1713		
Total	20	113416868002.5710			

Regression Statistics	
Multiple R	0.8515
R Square	0.7250
Adjusted R Square	0.7106
Standard Error	40513.9954
Observations	21

Approximately 72.5% of the variation in average monthly sales can be explained by store size.

Test Statistic for Significance of the Coefficient of Determination

$H_o: \rho^2 = 0.0$

$H_A: \rho^2 > 0.0$

$\alpha = 0.05$

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.8515					
5	R Square	0.7250					
6	Adjusted R Square	0.7106					
7	Standard Error	40513.9954					
8	Observations	21					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	82230575305	82230575305	50.0983	0.0000	
13	Residual	19	31186292697	1641383826			
14	Total	20	113416868003				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	171205.8279	59846.1252	2.8608	0.0100	45946.4483	296465.2075
18	Store Size (Sq. Ft.)	25.3160	3.5767	7.0780	0.0000	17.8299	32.8022

The F-ratio and p-value for testing whether the regression slope = 0.0

Test Statistic $F = \frac{MSR}{MSE}$ $\qquad \frac{82230575305}{1641383826} = 50.0983$

Since F = 50.0983 > $F_{critical.0.05}$ = 4.381 (Appendix H), reject the H_0

- Coefficient of Multiple correlation =sqrt(Coefficient of determination)

  Conditions: intercept is included and best possible linear predictors are used.

- Coefficient of determination is more general case including non-linear predictions and predicted values not derived from model fitting approach

$R_{1.23}$ = Multiple Correlation Coefficient coefficient of X_1 on X_2 and X3

- $R^2{}_{1.23} = \frac{r^2{}_{12+}r^2{}_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2{}_{23}}$

$$R^2{}_{2.13} = \frac{r^2{}_{21+}r^2{}_{23} - 2r_{12}r_{13}r_{23}}{1 - r^2{}_{13}}$$

$$R^2{}_{3.12} = \frac{r^2{}_{21+}r^2{}_{31} - 2r_{12}r_{13}r_{23}}{1 - r^2{}_{21}}$$

- In a trivariate distribution, if $r_{12} = 0.7, r_{13} = 0.61\ and\ r_{23} = 0.4$

  Find all multiple correlation coefficients.

$$R^2{}_{1.23} = \frac{r^2{}_{12} + r^2{}_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2{}_{23}}$$

$R_{1.23} = 0.6196$

$R_{2.13} = 0.4912$

$R_{1.23} = 0.6111$