University of Ferrara

DIPARTIMENTO
DI ECONOMIA
E MANAGEMENT

Statistics for Economics and Business

Stefano Bonnini & Valentina Mini

Multiple Linear Regression Analysis

Lecture 7 – 1rst of March 2019

The state of the art

- Last 2 weeks we talked **about simple linear regression**, that is finding a **best line** through a bunch of data, and that’s good for models where you have an output (or dependent , Y) variable depending on one input (or explanatory , x) variable.

- But the world is a very complex place, so we may have **more explanatory variables** affecting a dependent variable.

- What do you do if you have got a Y that **depends on two or more explanatory (x) variables**?

  How can we model that?

  How can we use linear algebra to find the **best fit**?

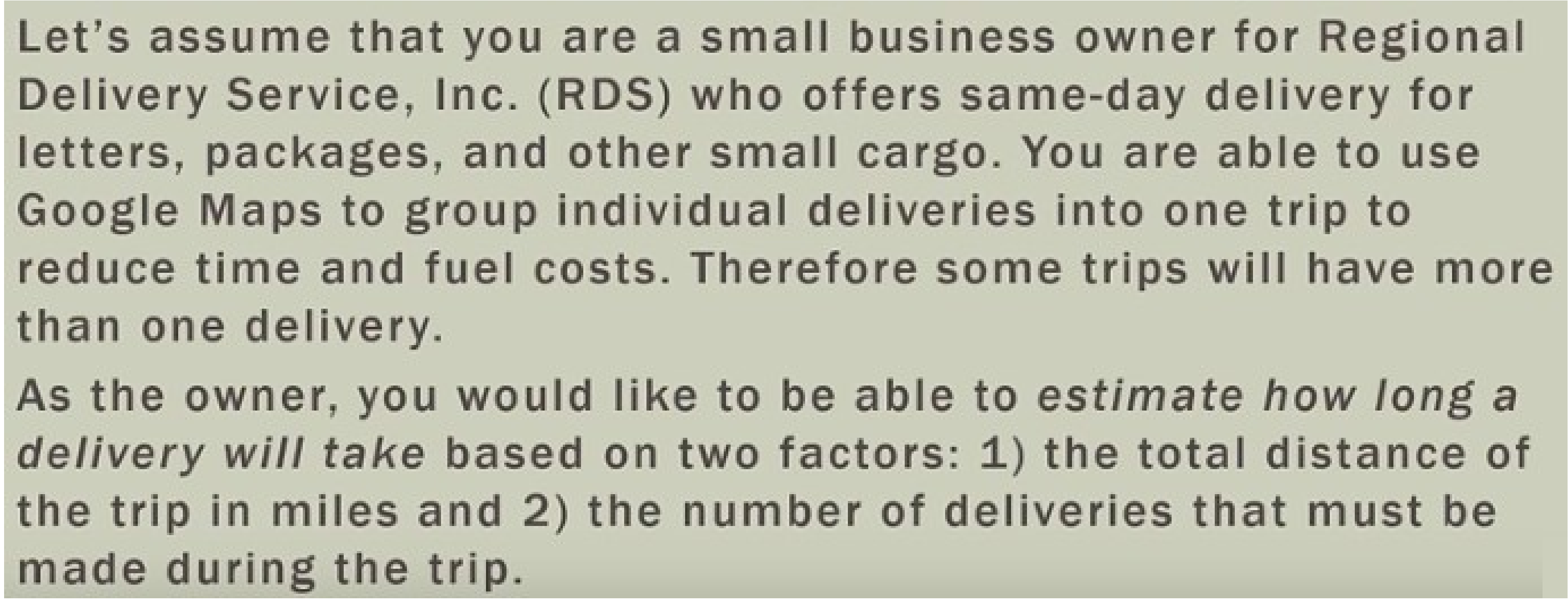
  How we can interpret the obtained results?

contents

1. Introduction
2. The MLR model
3. Using algebra to find the best fit
4. Graphical representation
5. A case study guides our first interpretation
6. Lab - R

Examples of multiple linear regression problems:

EX 1 - REGIONAL DELIVERY SERVICES

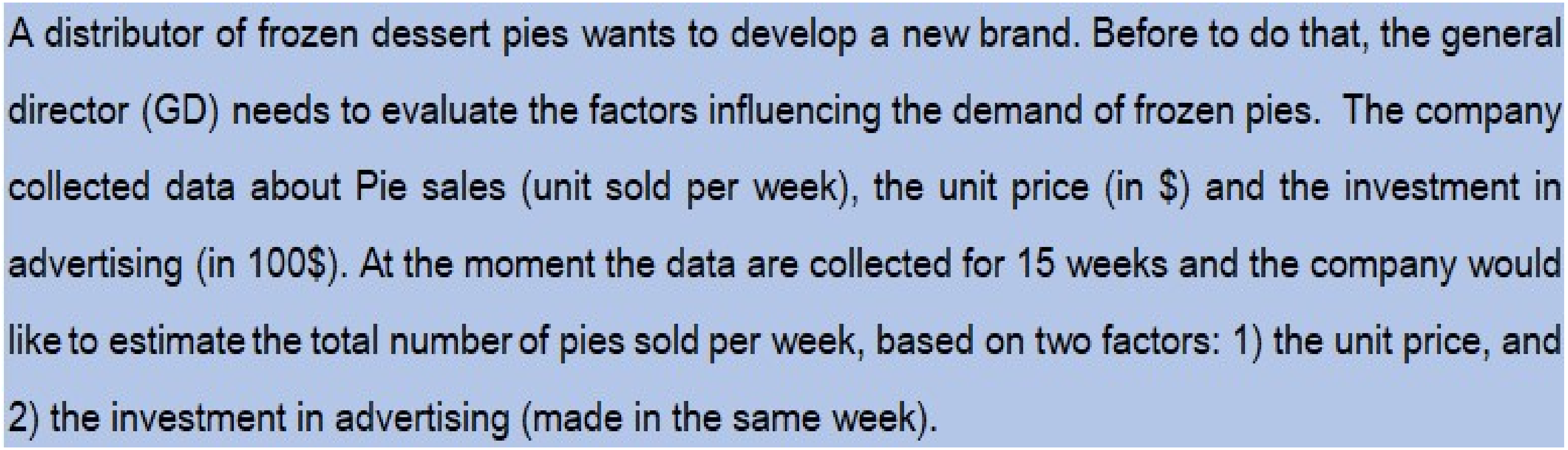


Let's assume that you are a small business owner for Regional Delivery Service, Inc. (RDS) who offers same-day delivery for letters, packages, and other small cargo. You are able to use Google Maps to group individual deliveries into one trip to reduce time and fuel costs. Therefore some trips will have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

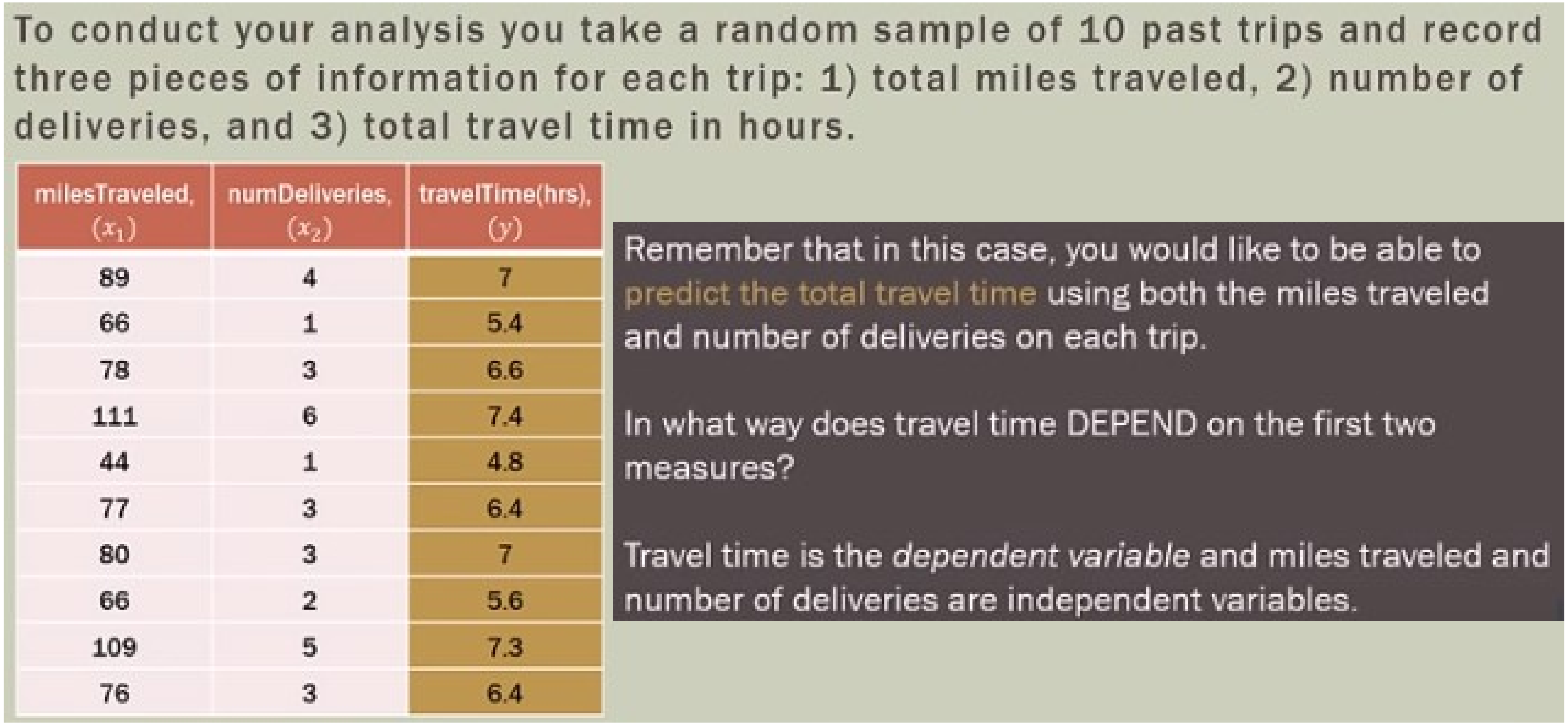
Examples of multiple linear regression problems:

EX 2 - CAKES TRADING COMPANY



A distributor of frozen dessert pies wants to develop a new brand. Before to do that, the general director (GD) needs to evaluate the factors influencing the demand of frozen pies. The company collected data about Pie sales (unit sold per week), the unit price (in $) and the investment in advertising (in 100$). At the moment the data are collected for 15 weeks and the company would like to estimate the total number of pies sold per week, based on two factors: 1) the unit price, and 2) the investment in advertising (made in the same week).

Example 1:
definition of data and data variables



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

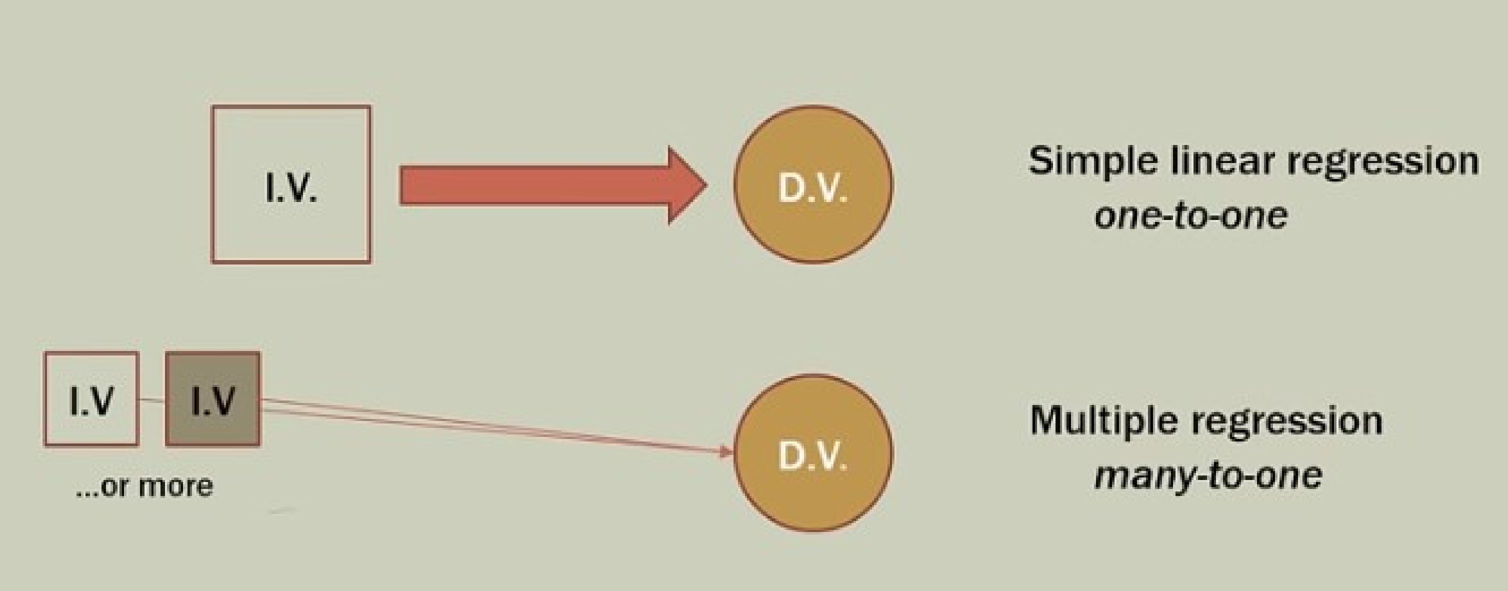
Remember that in this case, you would like to be able to predict the total travel time using both the miles traveled and number of deliveries on each trip.

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.

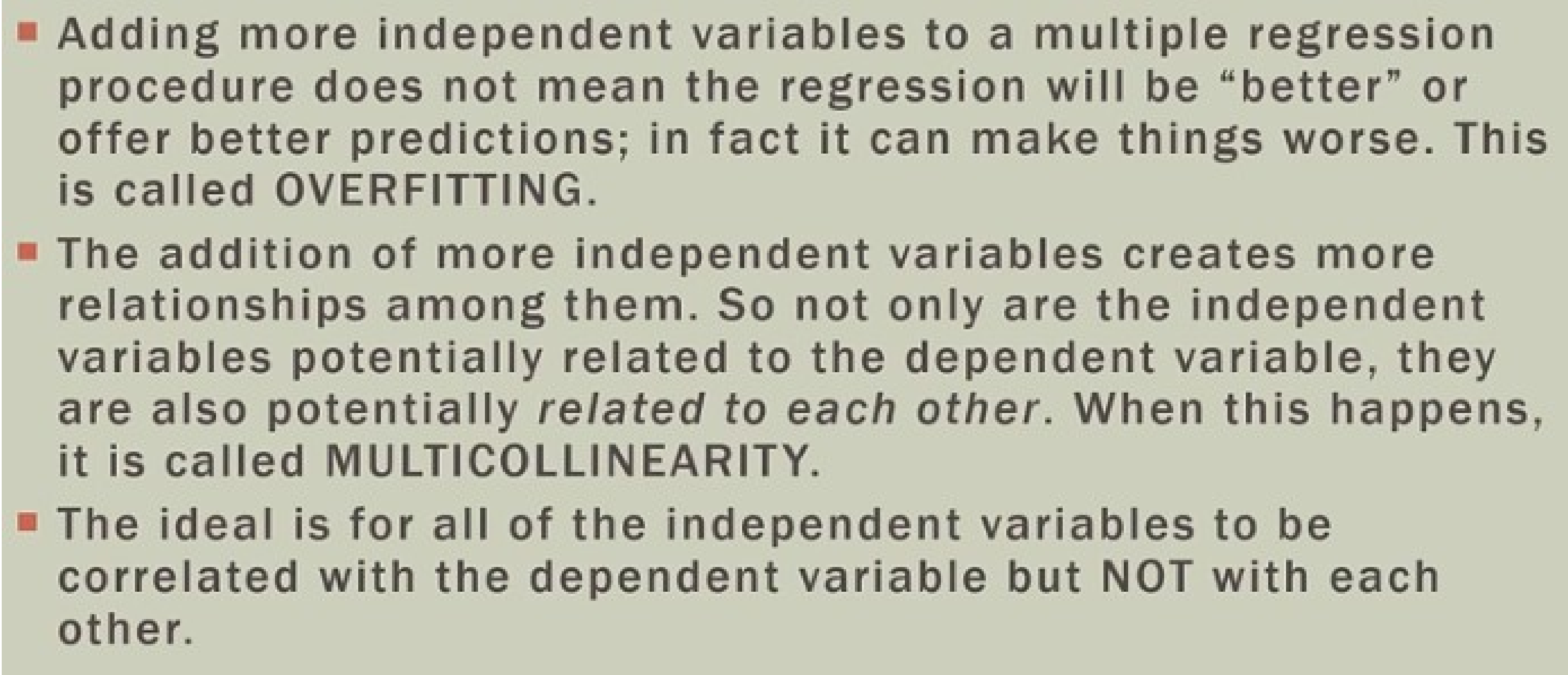
Note: *Y = dependent variable; X_i = independent variables OR*
Y= response variable; X_i = predictor variables OR
Y= output variable; X_i = input variables.

Multiple regression is an extension of simple linear regression



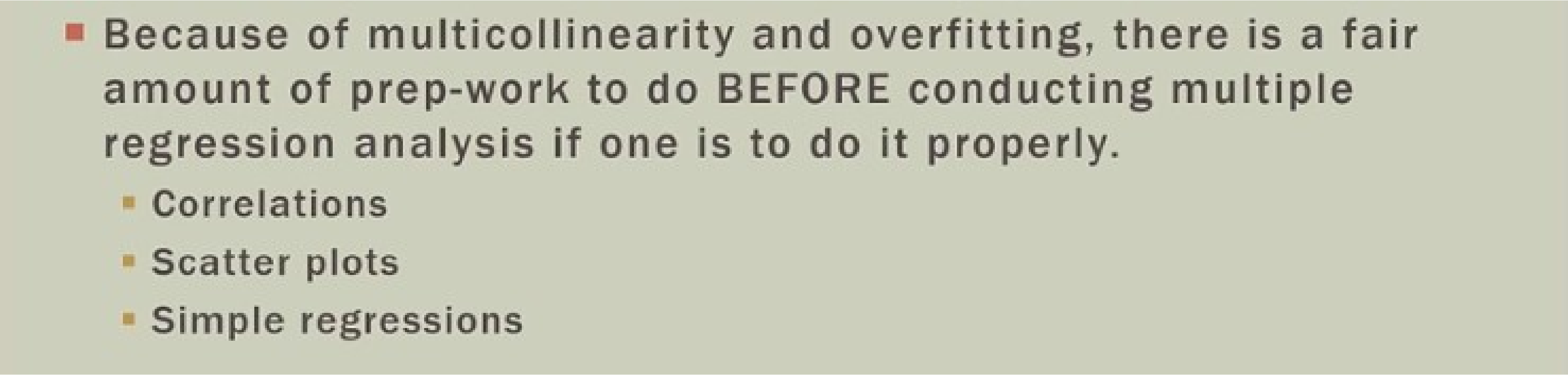
Having more independent variables complicates things a bit…
Thus we need to make new considerations:

New considerations (1/2)



- 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.

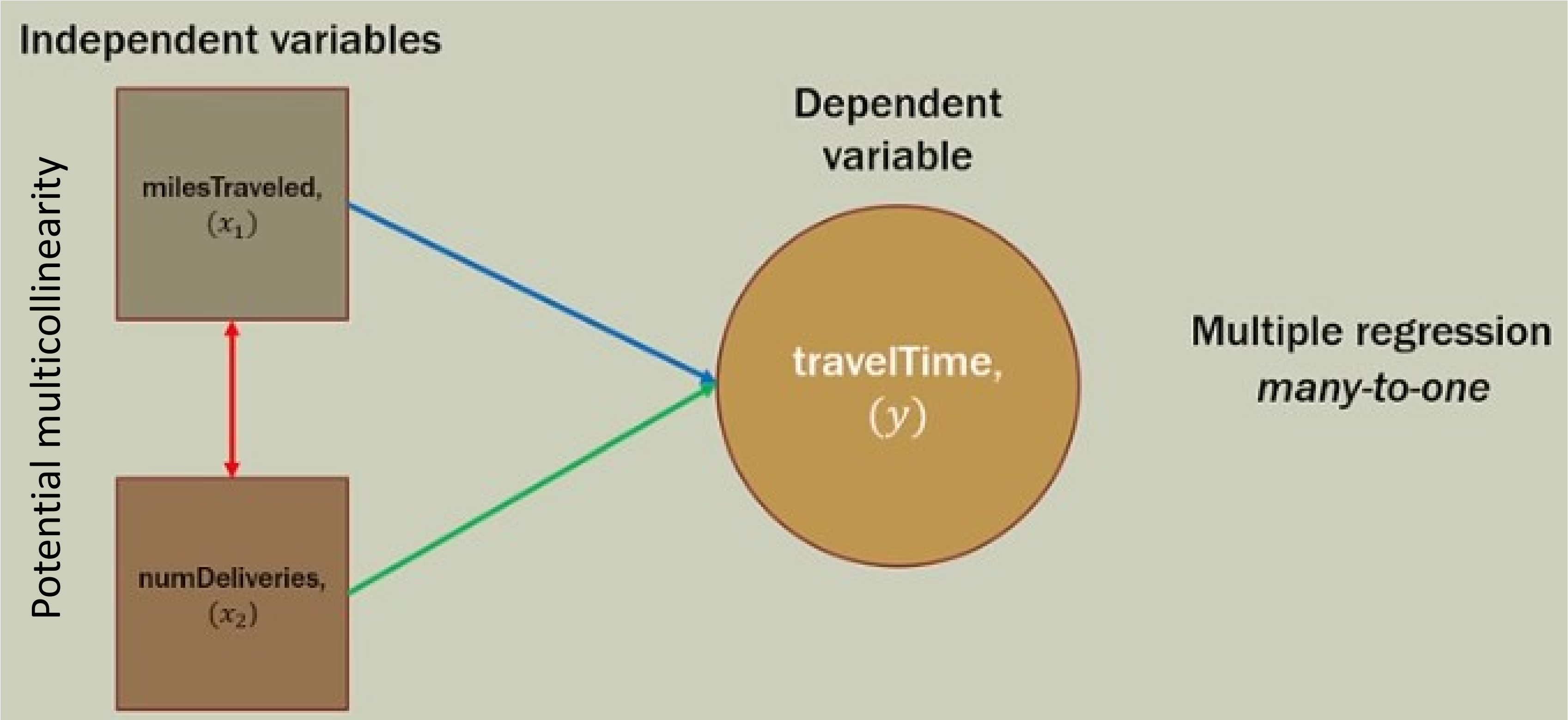
New considerations (2/2)



- 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

Running the Multiple Regression is the very last step

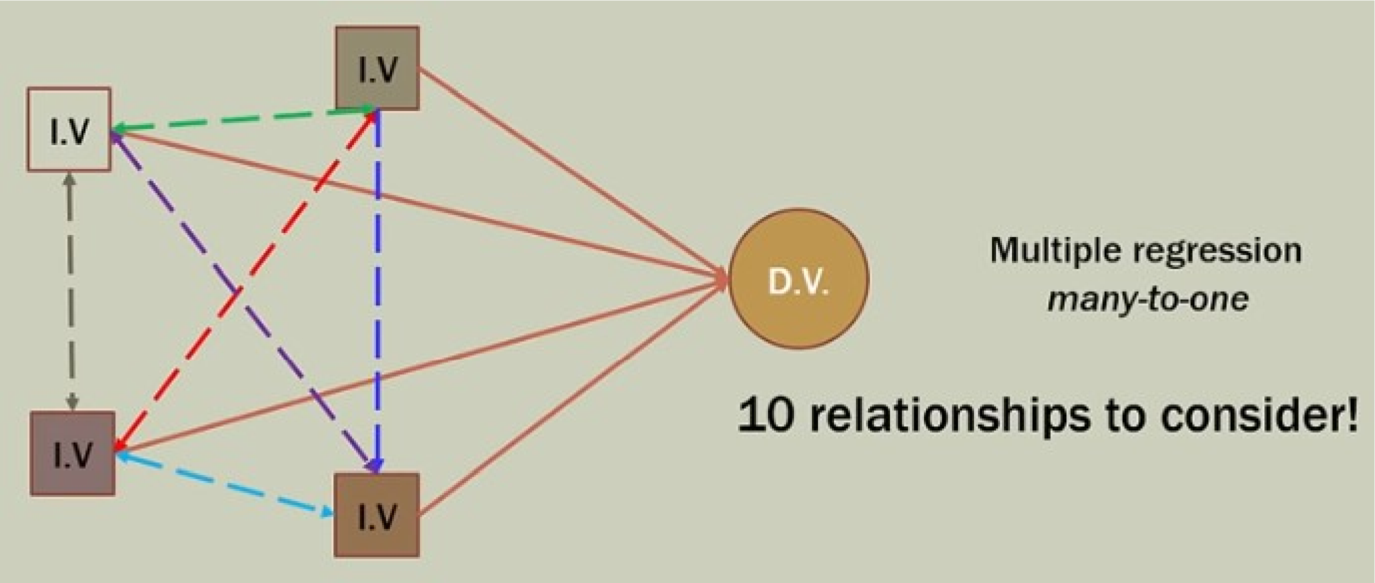
Multiple regression analysis: many relationship to deal with



We don’t have just 2 relationship (independents – dependent)

We may discover also the relationship between the independents: we need to check for that to avoid the multicollinearity → we need to guarantee that they are independent to each other

Multiple regression analysis: many relationship to deal with



- As each **independent** variable **is added**, the relationships may become very numerous.
- The ART of doing multiple regression **is deciding which independent variables make the cut** and which do not.
- Some independents variables, or set of independent variables**, are better at predicting the dependent variable** than other.
- Some contributes nothing.

Multiple regression analysis: many relationship to deal with

The ideal is for all the
INDEPENDENT VARIABLES to be correlated
with the dependent variable,
but NOT WITH EACH OTHER

The model

Idea: Examine the linear relationship between
1 dependent (Y) & 2 or more independent (or explanatory) variables (X_i)

Multiple Regression Model with k Independent Variables:

Population Y-intercept

Population slopes

Random Error

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

The coefficients of the multiple regression model

are estimated using sample data

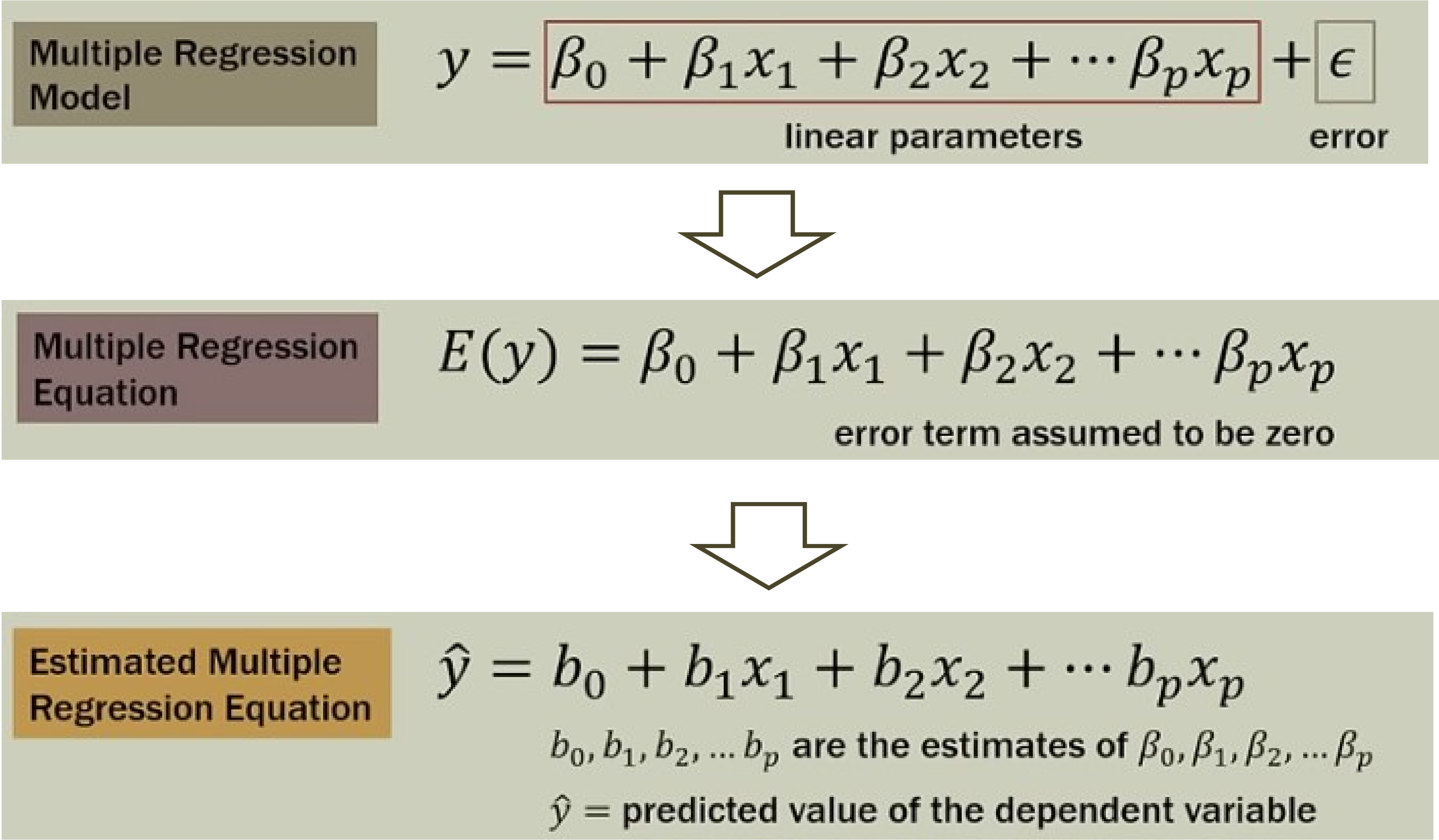
Multiple regression equation with k independent variables:

Estimated (or predicted) value of Y

Estimated intercept

Estimated slope coefficients

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki}$$



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

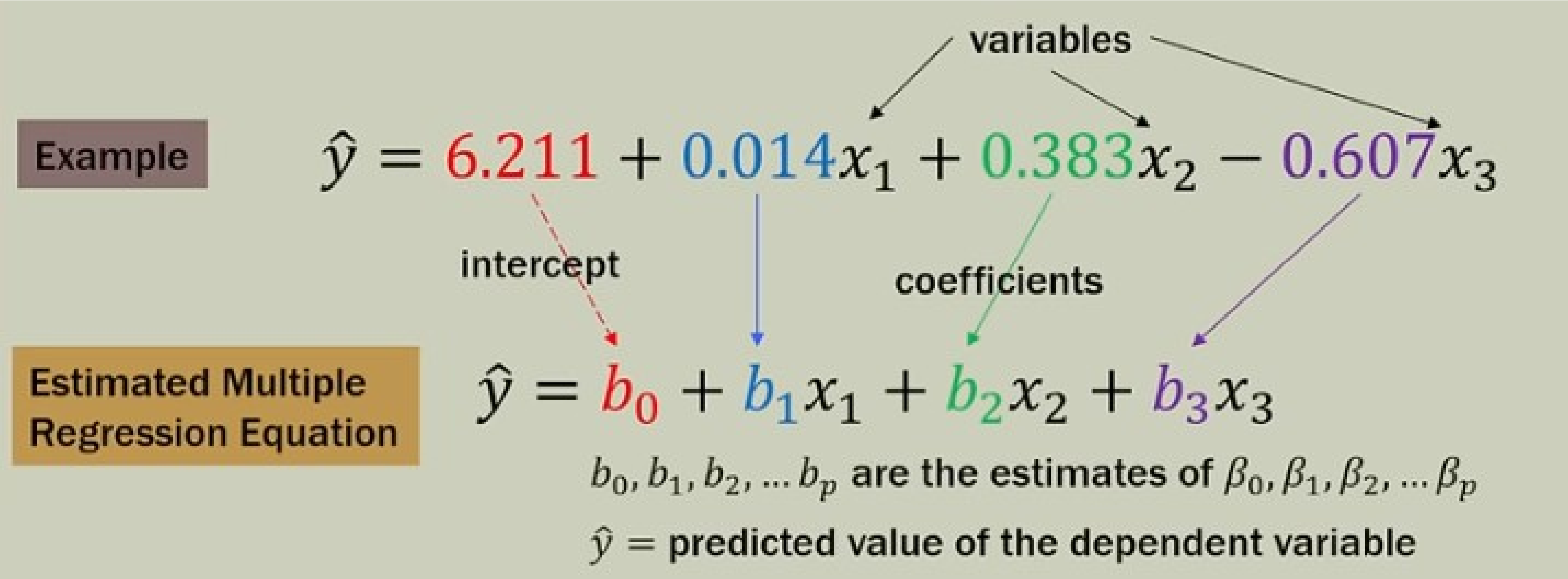
Estimated Multiple Regression Equation

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

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

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

Example using numbers



variables

Example

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

intercept

coefficients

Estimated Multiple Regression Equation

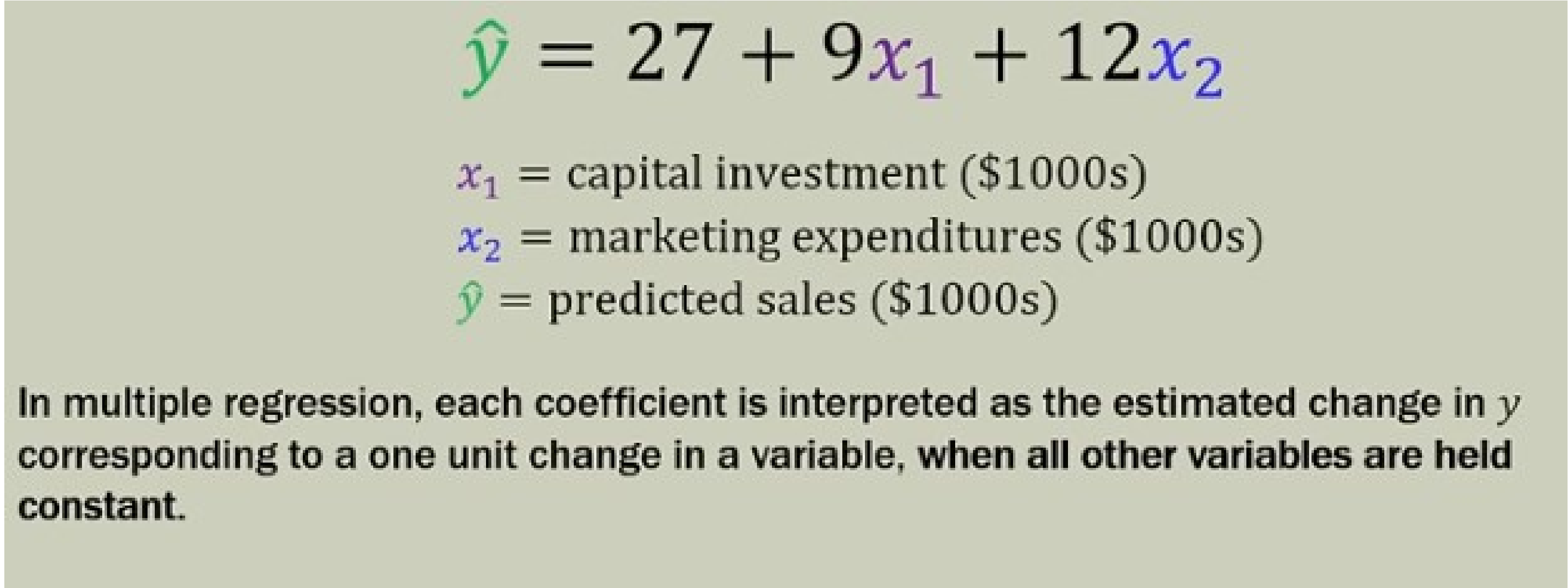
$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

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

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

Interpreting coefficients in MLRM

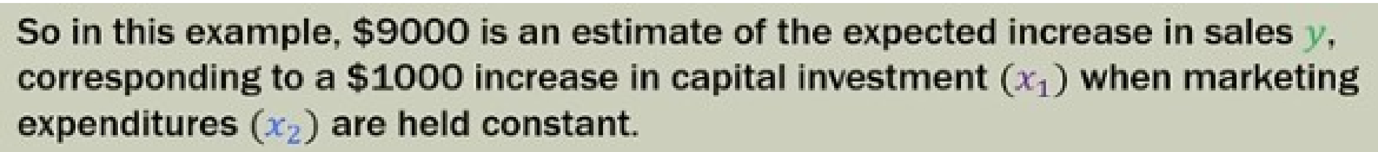
A numerical example



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

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.

Or: for each increase of 1 unit (\$1000) in capital investment, we expect an increase of \$9’000 in sales, when marketing expenditures are held constant.

Using linear algebra to find the best fit

General Parametric Equation:

$$y = \underbrace{f(\boldsymbol{X})}_{\text{Depends on Statistical Method}} + \epsilon$$

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

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

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

To deal with a high number of variables and thus a high number of operations, *computer tends to perform much easily using matrixes*

Using matrixes:

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

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

Using matrixes:

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

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

Now, using this matrix form, how we can compute the coefficients?

We use exactly the same principle we used for LRM:
minimizing the least square,
but for this time using matrixes

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

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

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

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

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

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

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

$$\left. RSS = \boldsymbol{y}^{T}\boldsymbol{y} - \boldsymbol{y}^{T}\boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\boldsymbol{X}^{T}\boldsymbol{y} + \widehat{\boldsymbol{\beta}^{T}}\boldsymbol{X}^{T}\boldsymbol{X}\hat{\boldsymbol{\beta}} \quad \right\}$$

The aim of the Least Square method is to find out the Beta values that minimize this RSS (Error)

3. Using linear algebra to find the best fit

Matrix Differentiation

$x = m \text{ x } 1$ matrix
$A = n \text{ x } m$ matrix; $A \perp x$

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

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

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

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

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

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

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

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

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

$$2\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X} = 2\boldsymbol{y}^T\boldsymbol{X}$$

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

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

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

Matrix Approach

We wish to find the vector of least squares estimators that minimizes:

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

The resulting least squares estimate is

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

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

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} =$$

$$= \left[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]\boldsymbol{y}$$

- The least squares function is given by

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

- The least squares estimates must satisfy

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0$$

and

$$\left.\frac{\partial L}{\partial \beta_j}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0 \quad j = 1, 2, \ldots, k$$

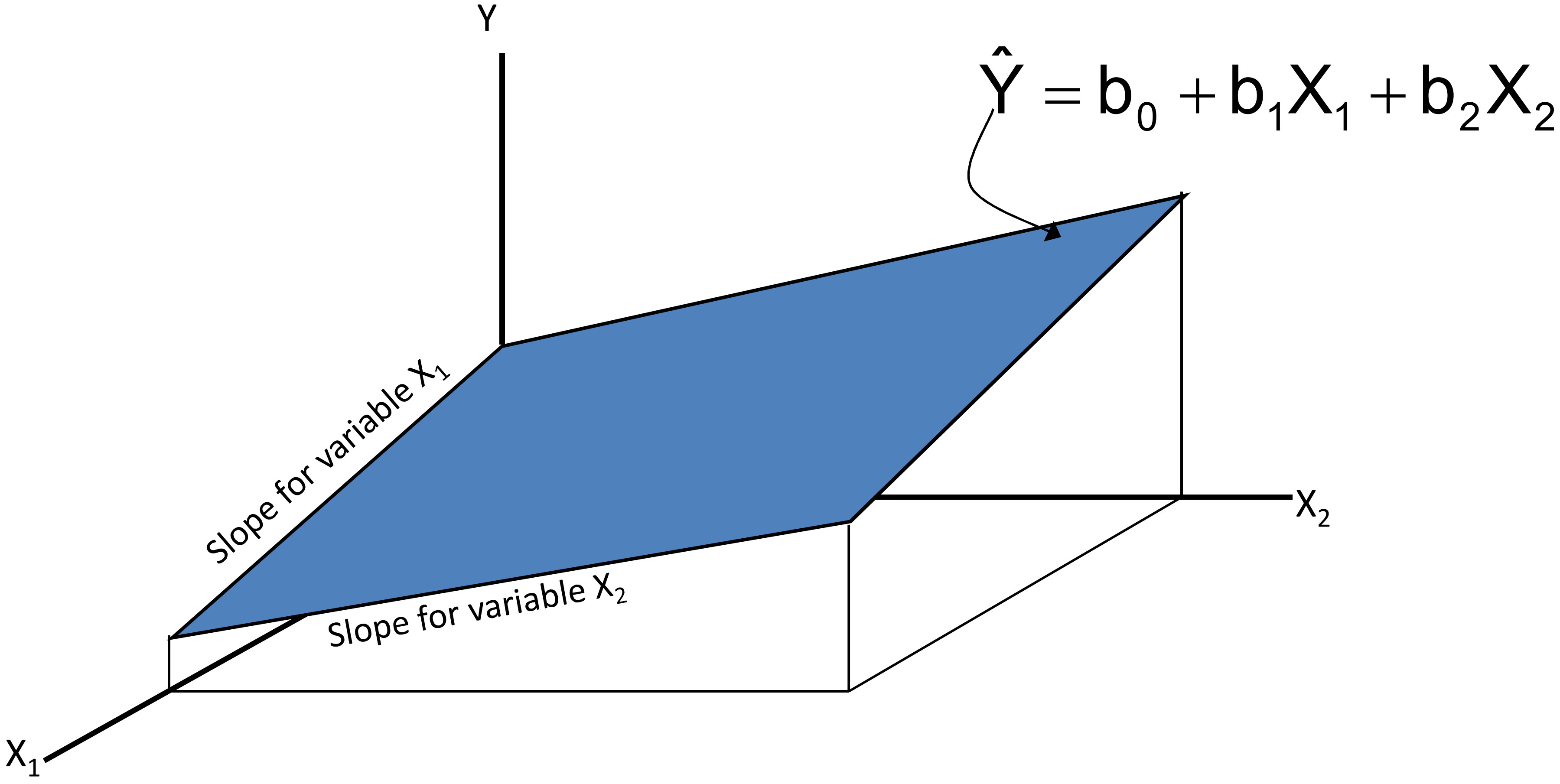
- The **least squares normal Equations** are

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

- The solution to the normal Equations are the **least squares estimators** of the regression coefficients.

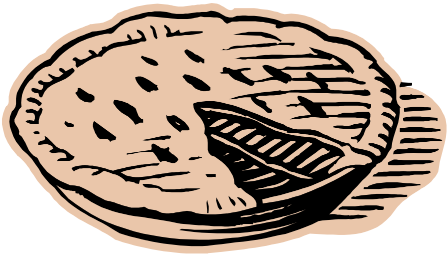
Graphical representation

Two variable model



A case study guides our first interpretation

- A distributor of frozen dessert pies wants to evaluate factors thought to influence demand
  - Dependent variable: Pie sales (units per week)
  - Independent variables: $\begin{cases}\text{Price (in \$)}\\ \text{Advertising (\$100's)}\end{cases}$
- Data are collected for 15 weeks



Week	Pie Sales	Price ($)	Advertising ($100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

Multiple regression equation:

$$\widehat{\text{Sales}} = b_0 + b_1 (\text{Price}) + b_2 (\text{Advertising})$$

Regression Statistics	
Multiple R	**0.72213**
R Square	**0.52148**
Adjusted R Square	**0.44172**
Standard Error	**47.46341**
Observations	**15**

b_1 = -24.975: sales will decrease, on average, by 24.975 pies per week for each $1 increase in selling price, net of the effects of changes due to advertising

b_2 = 74.131: sales will increase, on average, by 74.131 pies per week for each $100 increase in advertising, net of the effects of changes due to price

$$\text{Sales} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

ANOVA	***df***	***SS***	***MS***	***F***	***Significance F***
Regression	**2**	**29460.027**	**14730.013**	**6.53861**	**0.01201**
Residual	**12**	**27033.306**	**2252.776**		
Total	**14**	**56493.333**			

	Coefficients	***Standard Error***	***t Stat***	***P-value***	***Lower 95%***	***Upper 95%***
Intercept	**306.52619**	**114.25389**	**2.68285**	**0.01993**	**57.58835**	**555.46404**
Price	**-24.97509**	**10.83213**	**-2.30565**	**0.03979**	**-48.57626**	**-1.37392**
Advertising	**74.13096**	**25.96732**	**2.85478**	**0.01449**	**17.55303**	**130.70888**

Predict sales for a week in which the selling price is $5.50 and advertising is $350:

$$
\begin{aligned}
\widehat{\text{Sales}} &= 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising}) \\
&= 306.526 - 24.975\,(5.50) + 74.131(3.5) \\
&= 428.62
\end{aligned}
$$

Predicted sales is 428.62 pies

Note that Advertising is in $100's, so $350 means that X_2 = 3.5

- Reports the proportion of total variation in Y explained by all X variables taken together

$$\boxed{r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}}$$

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$r^2 = \frac{SSR}{SST} = \frac{29460.0}{56493.3} = .52148$$

52.1% of the variation in pie sales is explained by the variation in price and advertising

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

(continued)

- Shows the proportion of variation in Y explained by all X variables adjusted for the number of X variables used and sample size

$$\boxed{r_{adj}^2 = 1 - \left[(1 - r^2) \left(\frac{n-1}{n-k-1} \right) \right]}$$

(where n = sample size, k = number of independent variables)

  - Penalize excessive use of unimportant independent variables
  - Smaller than r^2
  - Useful in comparing among models

- F Test for Overall Significance of the Model
- Shows if there is a linear relationship between all of the X variables considered together and Y
- Use F-test statistic
- Hypotheses:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (no linear relationship)

H_1: at least one $\beta_i \neq 0$ (at least one independent variable affects Y)

- Test statistic:

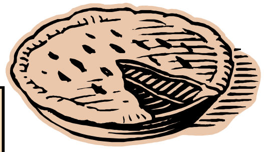
$$F_{STAT} = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}}$$

where F_{STAT} has numerator d.f. = k and

denominator d.f. = (n – k - 1)

(continued)

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15



$$F_{STAT} = \frac{MSR}{MSE} = \frac{14730.0}{2252.8} = 6.5386$$

With 2 and 12 degrees of freedom

P-value for the F Test

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Errors (residuals) from the regression model:

$$\boxed{e_i = (Y_i - \hat{Y}_i)}$$

Assumptions:

- Independence of errors
  - Error values are statistically independent
- Normality of errors
  - Error values are normally distributed for any given set of X values
- Equal Variance (also called Homoscedasticity)
  - The probability distribution of the errors has constant variance

- These residual plots are used in multiple regression:
  - Residuals vs. $\hat{Y}_i$
  - Residuals vs. X_{1i}
  - Residuals vs. X_{2i}
  - *Residuals vs. time (if time series data)*

Use the residual plots to check for violations of regression assumptions

- Use t tests of individual variable slopes
- Shows if there is a linear relationship between the variable X_j and Y holding constant the effects of other X variables
- Hypotheses:
  - H_0: $\beta_j = 0$ (no linear relationship)
  - H_1: $\beta_j \neq 0$ (linear relationship does exist between X_j and Y)

(continued)

H_0: β_j = 0 (no linear relationship)

H_1: $\beta_j \neq 0$ (linear relationship does exist between X_j and Y)

Test Statistic:

$$\boxed{t_{STAT} = \frac{b_j - 0}{S_{b_j}}}$$

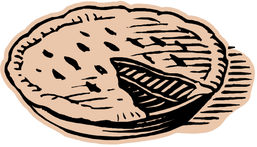
(df = n – k – 1)

(continued)

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

t Stat for Price is t_{STAT} = -2.306, with p-value .0398

t Stat for Advertising is t_{STAT} = 2.855, with p-value .0145



ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

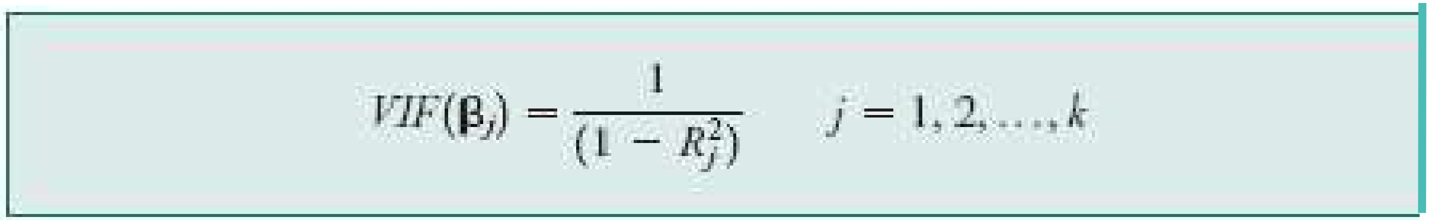
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Multicollinearity (also collinearity) occurs when two or more explanatory variables of the multiple regression model are highly correlated

In the presence of multicollinearity the coefficients estimates can change with high variability as a consequence of small changes in the data (low efficiency).

Perfect multicollinearity $\Rightarrow$ $\boldsymbol{X}$ matrix is singular and cannot be inverted $\Rightarrow$ least square estimates cannot be computed

One way to detect multicollinearity is by computing the **variance inflaction factors**



$$VIF(\beta_j) = \frac{1}{(1 - R_j^2)} \qquad j = 1, 2, \ldots, k$$

R_j^2: coefficient of determination of the regression of X_j on all the other explanatory variables

A VIF greater than or equal to 5 indicates
a multicollinearity problem

In the presence of multicollinearity one or more explanatory variables
should be removed by the model

Example: **VIF**(Price)=VIF(Advertising)= *$1/(1-R_1^2)= 1/(1-0.0009264^2) \simeq 1$*
Price and *Advertising* are almost uncorrelated $\Rightarrow$ absence of collinearity

Regression analysis general procedure

- Specification of the multiple regression model
- Test the significance of the multiple regression model
- Test the significance of the regression coefficents
- Discuss adjusted r^2
- Use residual plots to check model assumptions

Exercises using R

In class exercise

Open R

Using the dataset “torta”, perform a MLRM

Central question:

Are the total sales affected by price and advertising?

Note:

Price is expressed in $

Advertise is expressed in 100$

R exercises

Problem 1 - Passito

- Perform a multiple regression analysis for predicting LIKE_PAS as function of LIKE_AROMA, LIKE_SWEET, LIKE_ALCOHOL and LIKE_TASTE
- Predict the value of LIKE_PAS when LIKE_AROMA=LIKE_ALCOHOL=5

LIKE_TASTE=LIKE_SWEET=6

R exercises

Problem 2 - Hotel

- Perform a multiple regression analysis for predicting *Price* as function of *Cleanliness* and *Courtesy*
- Predict the value of *Price* when *Cleanliness*=80 and *Courtesy*=40

R exercises

Problem 3 - Mall

- Perform a multiple regression analysis for predicting *Product_assortment* as function of *Temp_Level*, *Brightness*, *Salesman* and *Music_volume*
- Predict the value of *Product_assortment* when *Temp_Level*=-50, *Brightness=20, Salesman=30* and *Music_volume=-70*

R exercises

Problem 4 - Students

- Perform a multiple regression analysis for predicting *Econometrics* as function of *Statistics* and *Mathematics*
- Predict the value of *Econometrics* when *Statistics=8* and *Mathematics=7*