The strength of the linear relationship between Y and the set of predictors $X_1, X_2, \cdots, X_p$ can be assessed through the examination of the scatter plot of Y versus $\hat{Y}$ and the correlation coefficient between Y and $\hat{Y}$, which is given by

$$\text{Cor}(Y, \hat{Y}) = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\hat{y}_i - \bar{\hat{y}})^2}}, \quad (3.28)$$

where $\bar{y}$ is the mean of the response variable Y and $\bar{\hat{y}}$ is the mean of the fitted values. As in the simple regression case, the coefficient of determination $R^2 = [\text{Cor}(Y, \hat{Y})]^2$ is also given by

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}, \quad (3.29)$$

Thus, R^2 may be interpreted as the proportion of the total variability in

the response variable Y that can be accounted for by the set of predictor variables $X_1, X_2, \cdots, X_p$. In multiple regression, $R = \sqrt{R^2}$ is called the *multiple correlation coefficient* because it measures the relationship between one variable Y and a set of variables $X_1, X_2, \cdots, X_p$.

$$\hat{b}_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

whereas

$x1 = X1 - \overline{X_1}$

$x2 = X2 - \overline{X_2}$

$y1 = Y - \overline{Y}$

$$\hat{b}_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\hat{b}_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

MULTIPLE LINEAR REGRESSION - EXAMPLE - 03 SEP 2020

SNO	QD (Y)	Price (X1)	Income (X2)	y	x1	x2	y2	(x1)2	(x2)2	y * x1	y * x2	x1 * x2
1	100	5	1000	20	-1	200	400	1	40000	-20	4000	-200
2	75	7	600	-5	1	-200	25	1	40000	-5	1000	-200
3	80	6	1200	0	0	400	0	0	160000	0	0	0
4	70	6	500	-10	0	-300	100	0	90000	0	3000	0
5	50	8	300	-30	2	-500	900	4	250000	-60	15000	-1000
6	65	7	400	-15	1	-400	225	1	160000	-15	6000	-400
7	90	5	1300	10	-1	500	100	1	250000	-10	5000	-500
8	100	4	1100	20	-2	300	400	4	90000	-40	6000	-600
9	110	3	1300	30	-3	500	900	9	250000	-90	15000	-1500
10	60	9	300	-20	3	-500	400	9	250000	-60	10000	-1500
	800	**60**	**8000**	**0**	**0**	**0**	**3450**	**30**	**1580000**	**-300**	**65000**	**-5900**

count(n)	10
Mean(Y)	80
Mean(X1)	6
Mean(X2)	800

Beta1 cap	NR =	-90500000	NR/DR =	-7.188
	DR =	12590000		

Beta2 Cap	NR =	180000	NR/DR	0.0143
	DR =	12590000		

Beta0 Cap	68.56

MULTIPLE LINEAR REGRESSION			
SNO	Miles Traveled (X1)	Num Deliveries (X2)	Travel Time (hrs) (Y)
1	89	4	7
2	66	1	5.4
3	78	3	6.6
4	111	6	7.4
5	44	1	4.8
6	77	3	6.4
7	80	3	7
8	66	2	5.6
9	109	5	7.3
10	76	3	6.4

MULTIPLE LINEAR REGRESSION				
SNO	Miles Traveled (X1)	Num Deliveries (X2)	gasPrice (X3)	Travel Time (hrs) (Y)
1	89	4	3.84	7
2	66	1	3.19	5.4
3	78	3	3.78	6.6
4	111	6	3.89	7.4
5	44	1	3.57	4.8
6	77	3	3.57	6.4
7	80	3	3.03	7
8	66	2	3.51	5.6
9	109	5	3.54	7.3
10	76	3	3.25	6.4

MULTIPLE REGRESSION PROCESS

As we discussed in Parts 1 & 2, conducting multiple regression analysis requires a fair amount of pre-work before actually running the regression. Here are the steps:

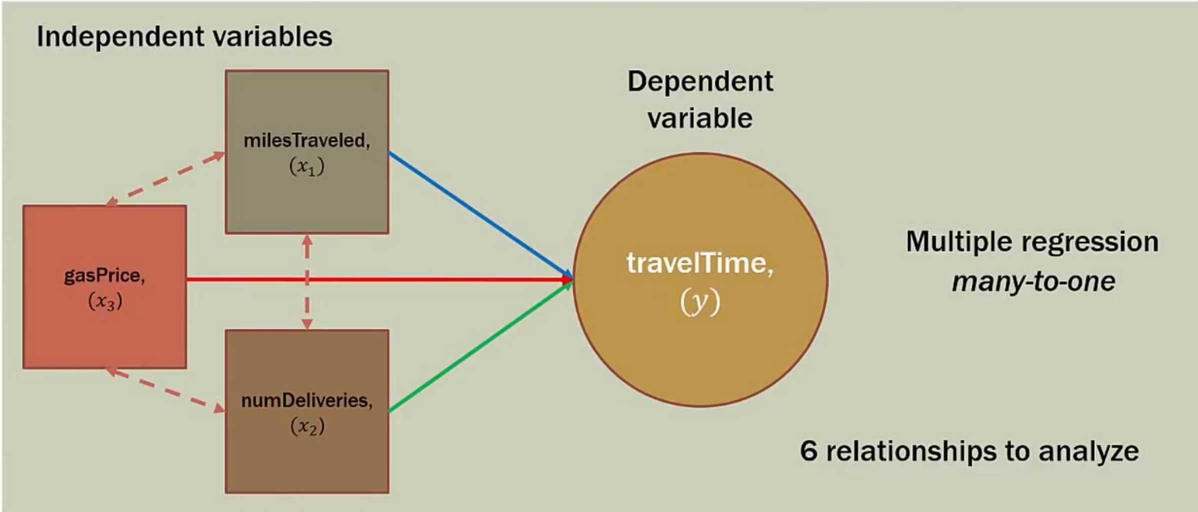
1. Generate a list of potential variables; independent(s) and dependent
2. Collect data on the variables
3. Check the relationships between each independent variable and the dependent variable using scatterplots and correlations
4. Check the relationships among the independent variables using scatterplots 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

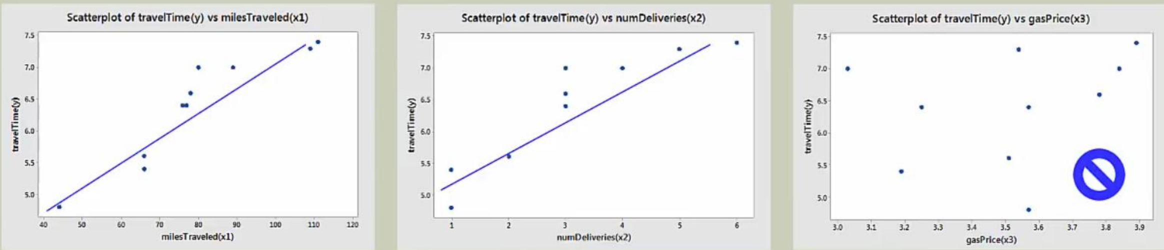
To conduct your analysis you take a random sample of 10 past trips and record four pieces of information for each trip: 1) total miles traveled, 2) number of deliveries, 3) the daily gas price, and 4) total travel time in hours.

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



DV VS IV SCATTERPLOTS

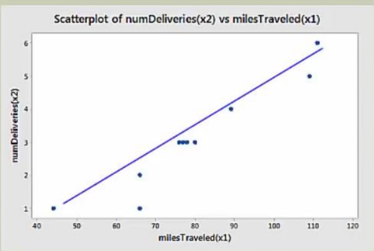


$r = 0.928$, p value=.000

$r = 0.916$, p value=.000

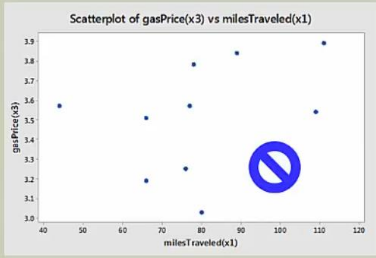
$r = 0.267$, p value=.455

IV SCATTERPLOTS (MULTICOLLINEARITY)



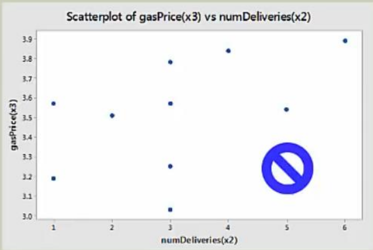
$r = 0.956$

p value=.000



$r = 0.356$

p value=.313



$r = 0.498$

p value=.143

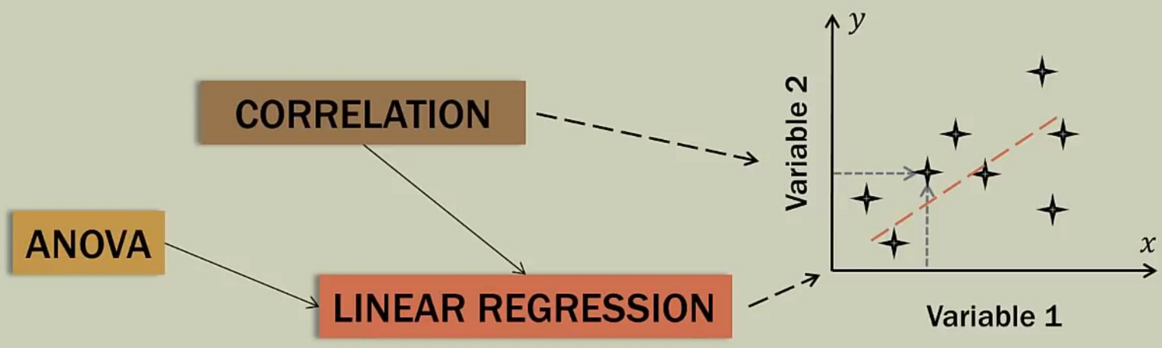
CORRELATION SUMMARY

- 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

BIVARIATE STATISTICS



The value of one variable, is a function of the other variable.

The value of y, is a function of x; $y = f(x)$.

The value of the dependent variable, is a function of the independent variable.

ALGEBRA REVIEW: LINES

slope–intercept form of a line

$$y = mx + b$$

x = random variable

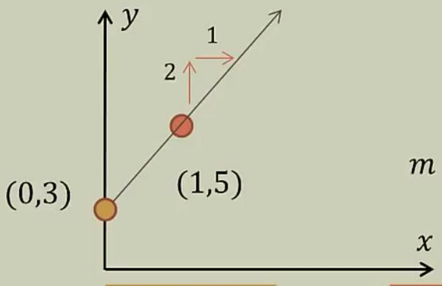
m = slope of the line $\frac{rise}{run}$

b = y–intercept (crosses y–axis)

y–intercept is where $x = 0$

Coordinate of $(0, y)$

$$y = 2x + 3$$



$$m = slope = \frac{2}{1}$$

$y = 2(0) + 3$

$y = 3$

$(0,3)$

where $x = 1$;

$y = 2(1) + 3$

$y = 5$

SIMPLE LINEAR REGRESSION MODEL

$$y = mx + b$$

$$y = \beta_0 + \beta_1 x + \epsilon$$

β_0 = y–intercept **population** parameter

β_1 = slope **population** parameter

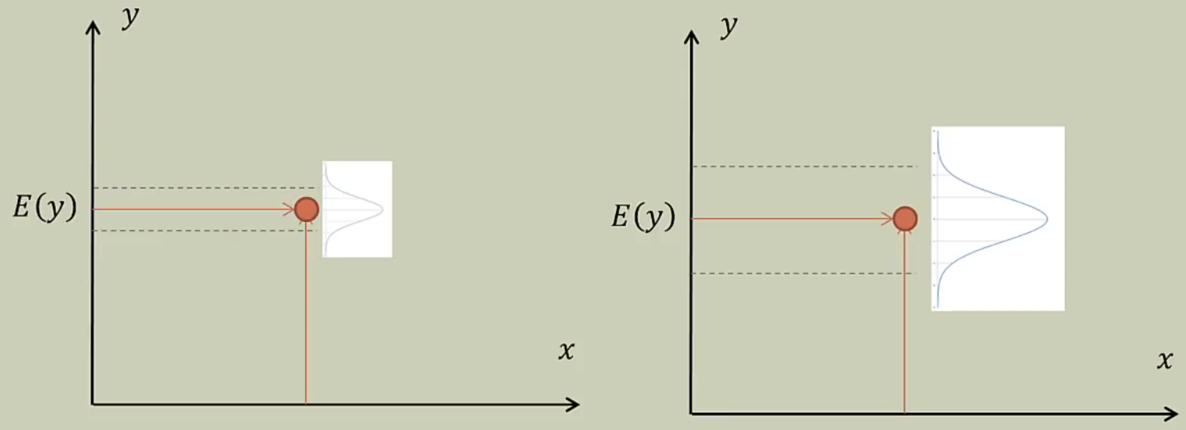
ϵ = error term, unexplained variation in y

Simple Linear Regression Equation

$$E(y) = \beta_0 + \beta_1 x$$

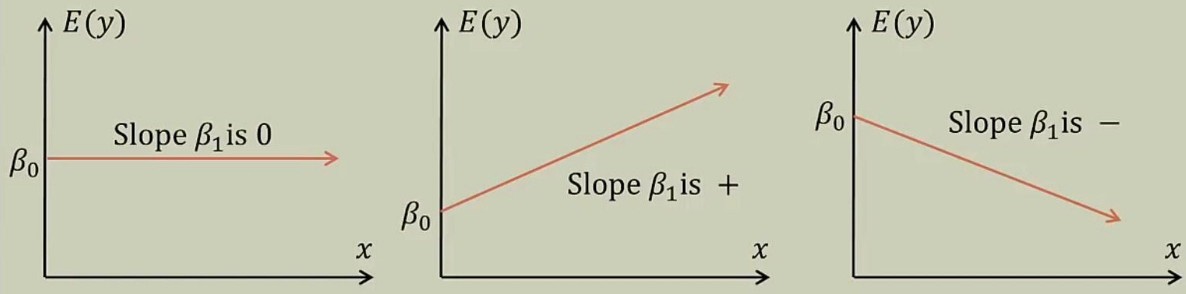
$E(y)$ is the **mean or expected value** of y, for a given value of x

DISTRIBUTION OF y-VALUES



GENERAL REGRESSION LINES

$$E(y) = \beta_0 + \beta_1 x$$



$$E(y) = \beta_0 + 0(x) \qquad E(y) = \beta_0 + \beta_1 x \qquad E(y) = \beta_0 - \beta_1 x$$

REGRESSION EQUATION WITH ESTIMATES

If we actually knew the population parameters, β_0 and β_1, we could use the Simple Linear Regression Equation.

$$E(y) = \beta_0 + \beta_1 x$$

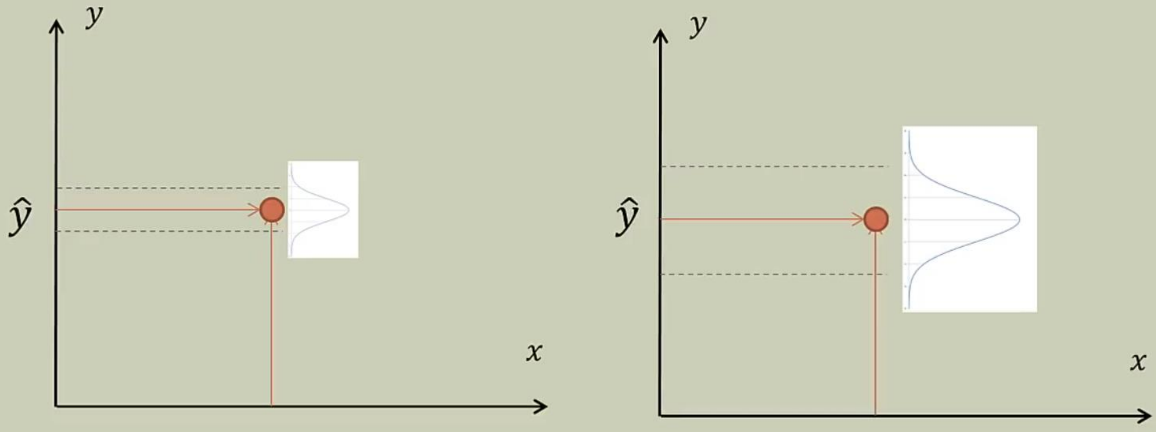
In reality we almost never have the population parameters. Therefore we will estimate them using sample data. When using sample data, we have to change our equation a little bit.

$\hat{y}$, pronounced "y-hat" is the point estimator of $E(y)$

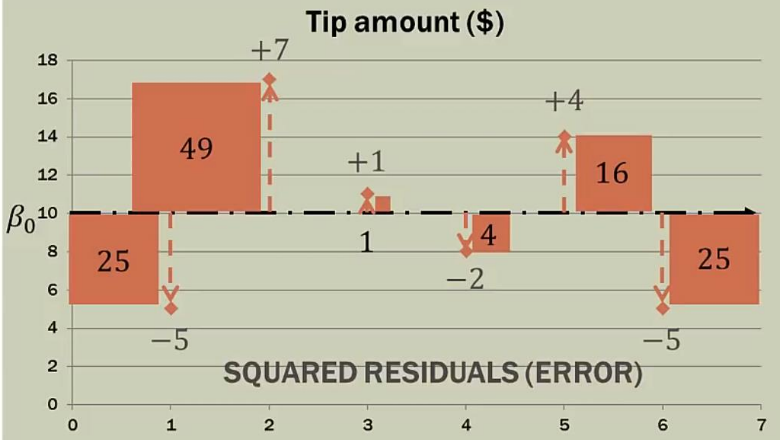
$$\hat{y} = b_0 + b_1 x$$

$\hat{y}$, is the **mean value** of y for a given value of x.

DISTRIBUTION OF SAMPLE y-VALUES



WHEN THE SLOPE, $\beta_1 = 0$



$Sum\ of\ squared\ errors\ (SSE) = 120$

When conducting simple linear regression with TWO variables, we will determine how good the regression line "fits" the data by comparing it to THIS TYPE; where we pretend the second variable does not even exist; the slope, $\boldsymbol{\beta_1 = 0}$.

In this situation, the value of $\hat{y}$ is 10 for every value of x.

$\hat{y} = b_0 + b_1 x$

$\hat{y} = b_0 + (0)x$

$\hat{y} = b_0$

$b_0 = 10$

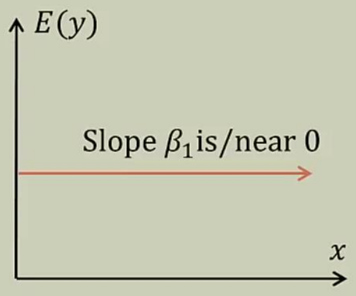
$\hat{y} = 10$

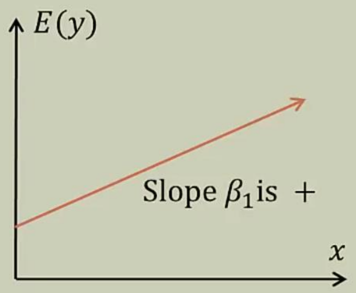
PATTERN MATCHING TO GENERAL REGRESSION MODELS

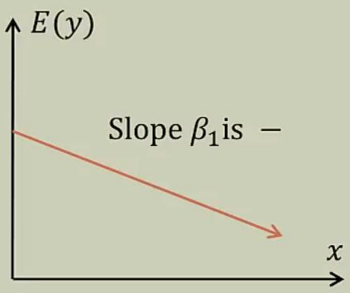
$\hat{y} = 0.3 - 3.3x$

$\hat{y} = 48 + 7.8x$

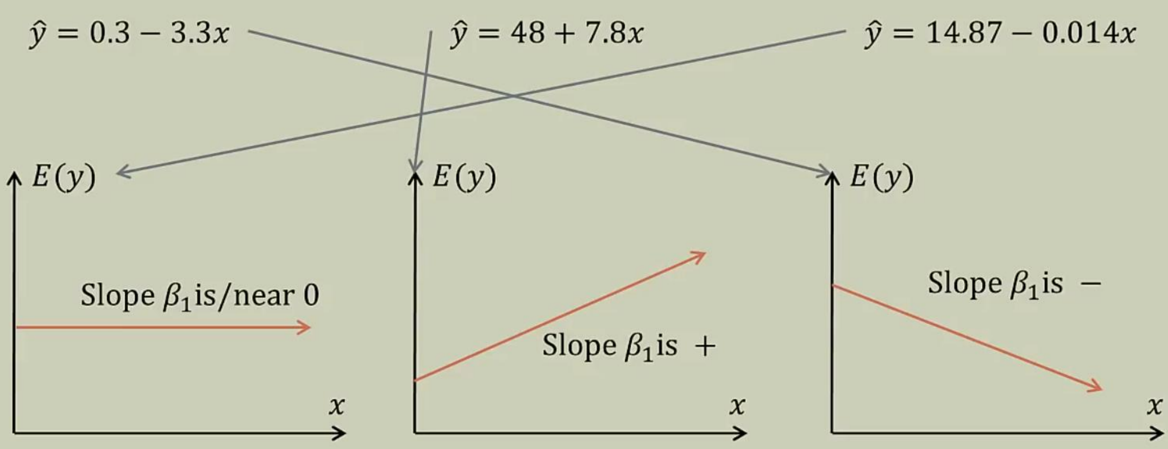
$\hat{y} = 14.87 - 0.014x$



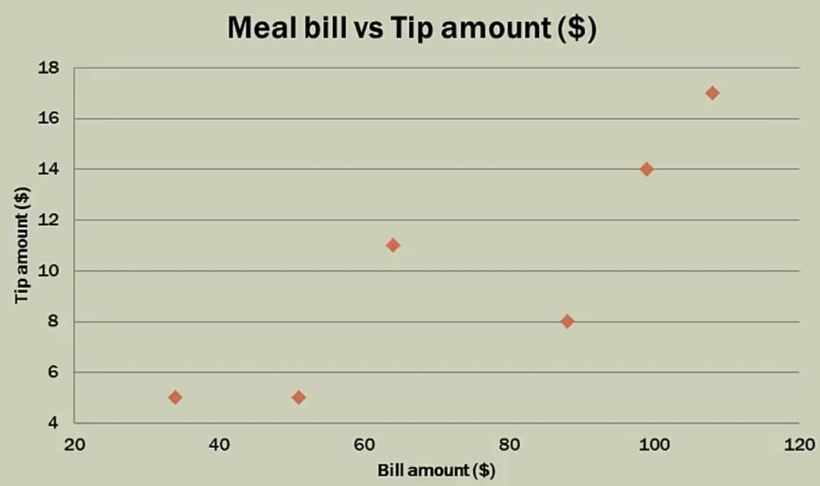




PATTERN MATCHING TO GENERAL REGRESSION MODELS



GETTING READY FOR LEAST SQUARES



Bill ($)	Tip ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00