Cost Function

The cost function helps us to figure out the best possible values for a_0 and a_1 which would provide the best fit line for the data points. Since we want the best values for a_0 and a_1, we convert this search problem into a minimization problem where we would like to minimize the error between the predicted value and the actual value.

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

Minimization and Cost Function

We choose the above function to minimize.

The difference between the predicted values and ground truth measures the error difference.

We square the error difference and sum over all data points and divide that value by the total number of data points.

This provides the average squared error over all the data points.

Therefore, this cost function is also known as the **Mean Squared Error (MSE) function**.

Now, using this MSE function we are going to change the values of a_0 and a_1 such that the MSE value settles at the minima.

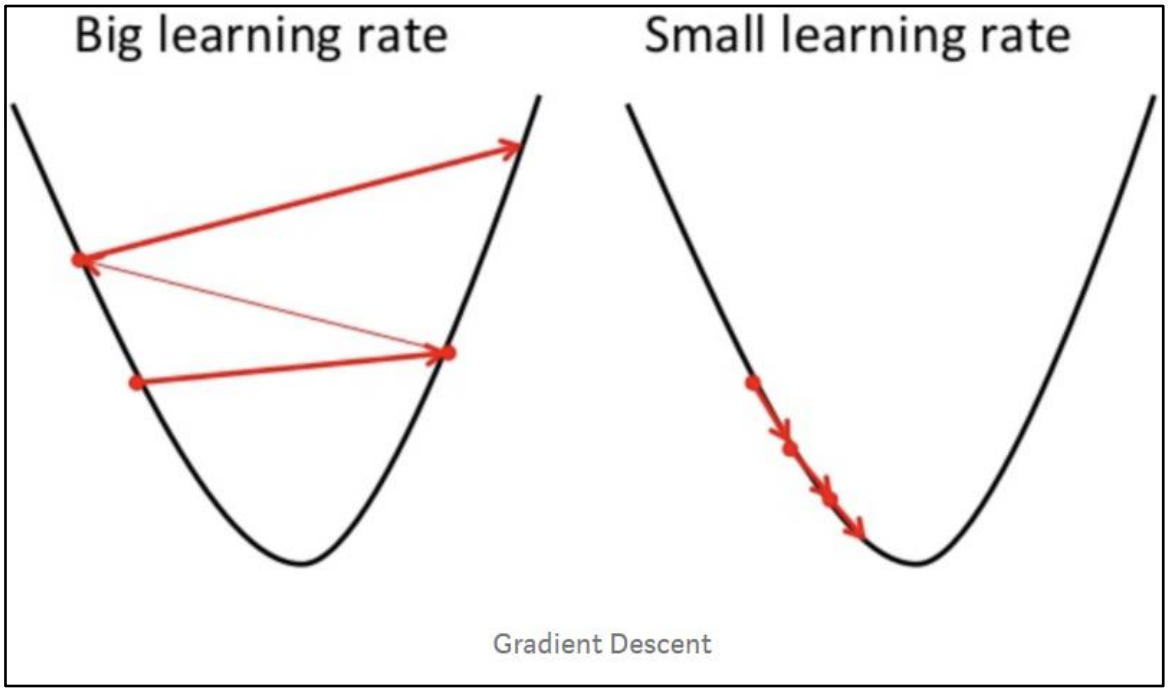
Gradient Descent

The next important concept needed to understand linear regression is gradient descent.

Gradient Descent is a method of updating a_0 and a_1 to reduce the cost function (MSE).

The idea is that we start with some values for a_0 and a_1 and then we change these values iteratively to reduce the cost.

Gradient descent helps us on how to change the values.



To draw an analogy, imagine a pit in the shape of U and you are standing at the topmost point in the pit and your objective is to reach the bottom of the pit.

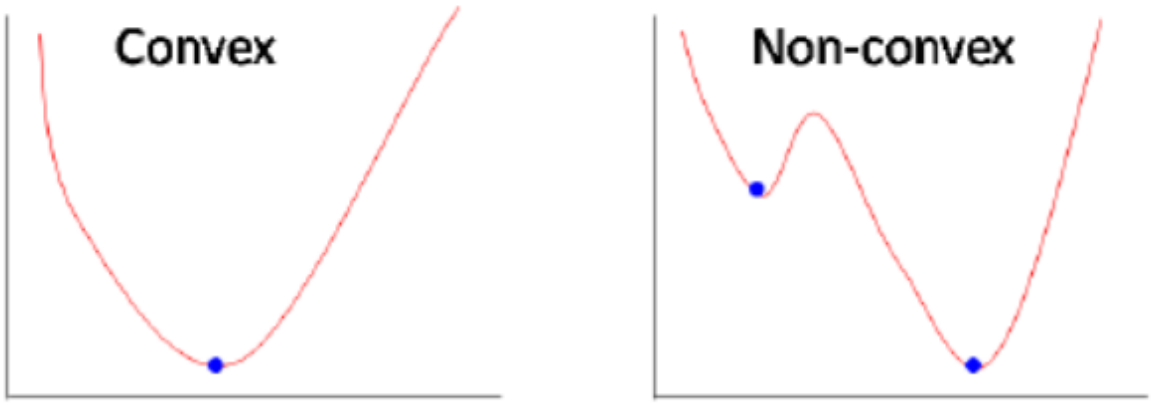
There is a catch, you can only take a discrete number of steps to reach the bottom.

If you decide to take one step at a time you would eventually reach the bottom of the pit but this would take a longer time.

If you choose to take longer steps each time, you would reach sooner but, there is a chance that you could overshoot the bottom of the pit and not exactly at the bottom.

In the gradient descent algorithm, the number of steps you take is the learning rate.

This decides on how fast the algorithm converges to the minima.



Convex vs Non-convex function

Sometimes the cost function can be a non-convex function where you could settle at a local minima but for linear regression, it is always a convex function.

You may be wondering how to use gradient descent to update a_0 and a_1.

To update a_0 and a_1, we take gradients from the cost function.

To find these gradients, we take partial derivatives with respect to a_0 and a_1.

Now, to understand how the partial derivatives are found below you would require some calculus but if you don't, it is alright.

You can take it as it is.

The partial derivates are the gradients and they are used to update the values of a_0 and a_1.

Alpha is the learning rate which is a hyperparameter that you must specify.

A smaller learning rate could get you closer to the minima but takes more time to reach the minima, a larger learning rate

converges sooner but there is a chance that you could overshoot the minima.

Training Set

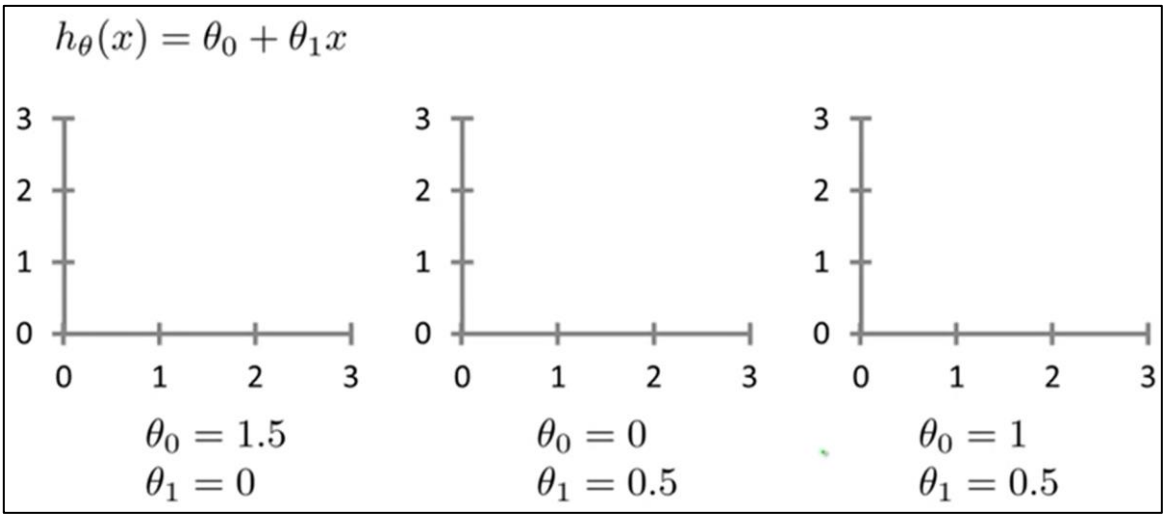
Size in feet² (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
...	...

$m = 47$

Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

θ_i's: Parameters

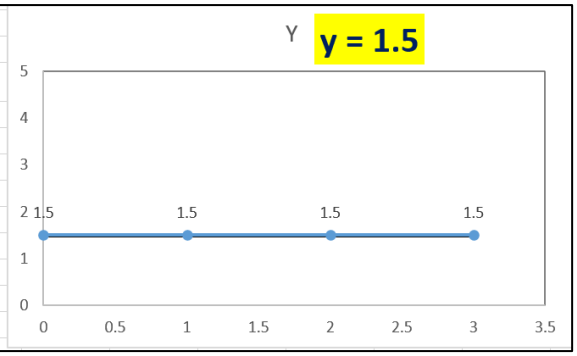
How to choose θ_i's ?



Y=MX + C

CASE1: M= 0; C = 1.5

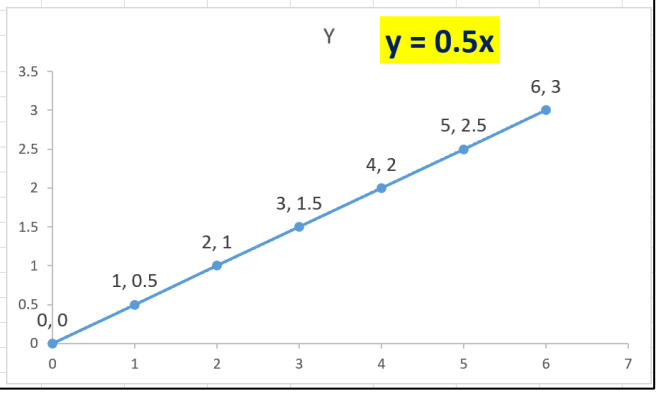
X	Y
0	1.5
1	1.5
2	1.5
3	1.5



Y=MX + C

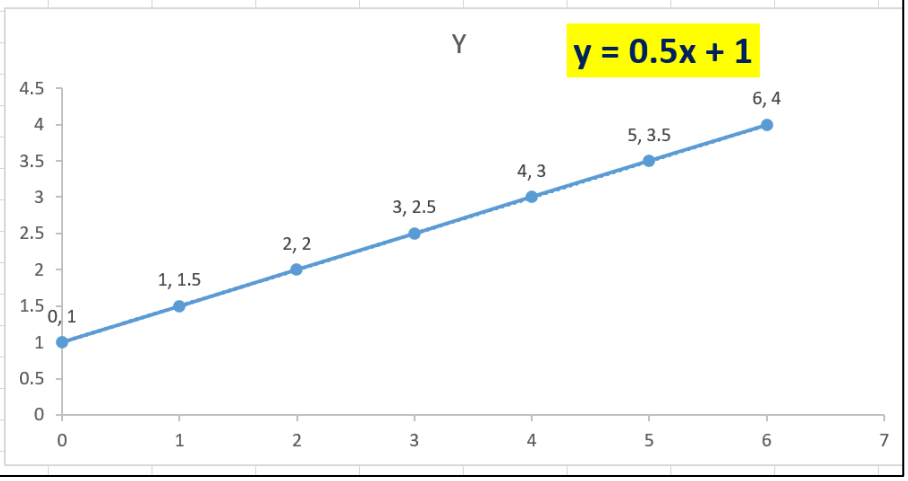
CASE1: M= 0.5; C = 0

M	C	X	Y
0.5	0	0	0
		1	0.5
		2	1
		3	1.5
		4	2
		5	2.5
		6	3



Y=MX + C

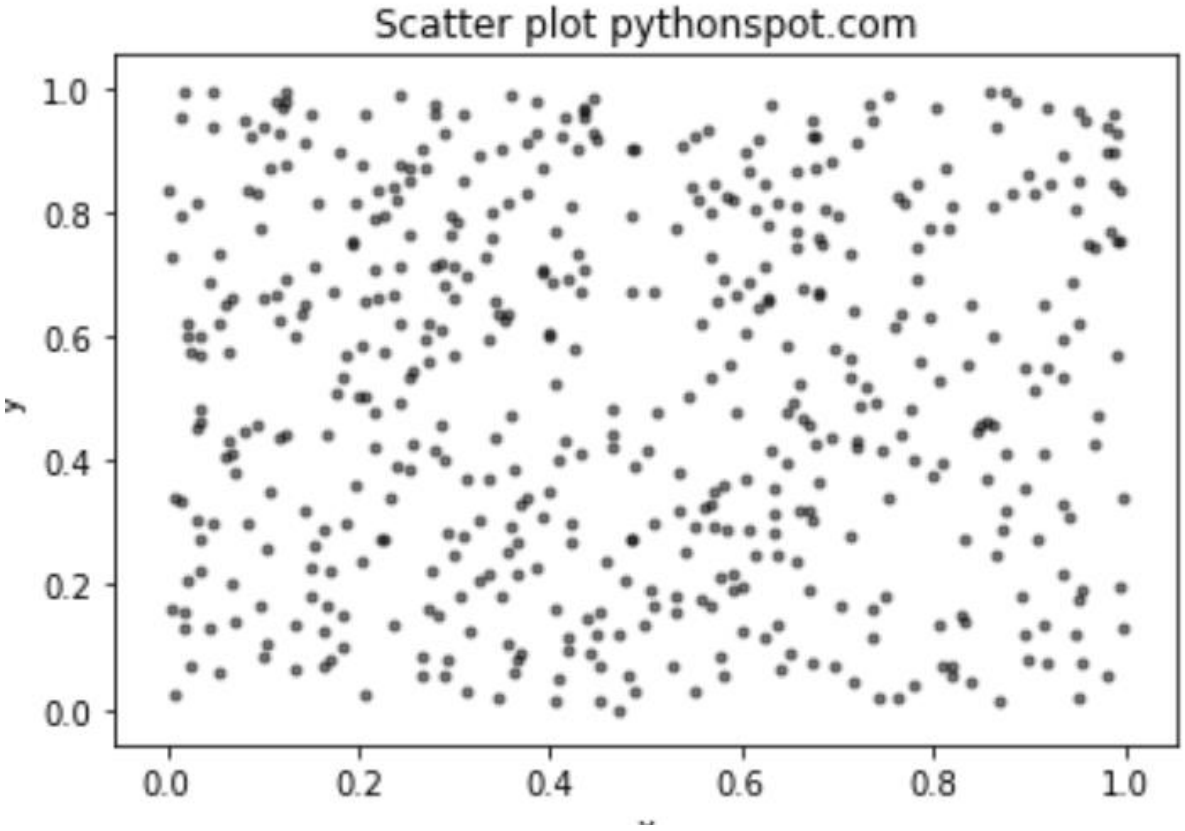
M	C	X	Y
0.5	1	0	1
		1	1.5
		2	2
		3	2.5
		4	3
		5	3.5
		6	4
		8	5



```python
import numpy as np
import matplotlib.pyplot as plt

# Create data
N = 500
x = np.random.rand(N)
y = np.random.rand(N)
colors = (0,0,0)
area = np.pi*3

# Plot
plt.scatter(x, y, s=area, c=colors, alpha=0.5)
plt.title('Scatter plot pythonspot.com')
plt.xlabel('x')
plt.ylabel('y')
plt.show()
```



```
#import necessary modules
import csv
with open('D:\MLIIISEMNOTES-09AUG2020\MYPYTHONPRGS-26JULY2020\TEMP1.csv','rt')as f:
  data = csv.reader(f)
  for row in data:
        print(row)
```

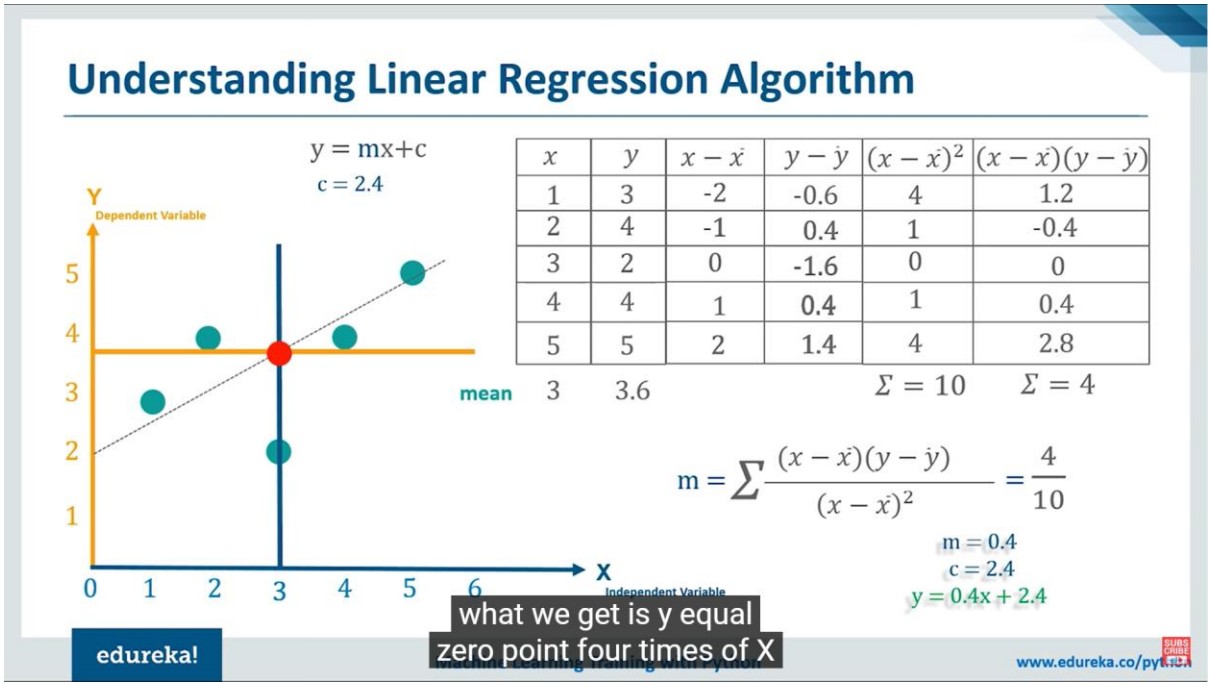
```
['X', 'Y ']
['0', '1.5']
['1', '2']
['2', '2.5']
['3', '3']
['4', '3.5']
['5', '4']
['6', '4.5']
['8', '5']
```

```
#import necessary modules
import pandas
result = pandas.read_csv('D:\MLIIISEMNOTES-09AUG2020\MYPYTHONPRGS-26JULY2020\data.csv')
print(result)
```

```
  Programming language         Designed by  Appeared Extension
0               Python   Guido van Rossum       1991       .py
1                 Java      James Gosling       1995     .java
2                  C++  Bjarne Stroustrup       1983      .cpp
```

```
#import necessary modules
import pandas
result = pandas.read_csv('D:\MLIIISEMNOTES-09AUG2020\MYPYTHONPRGS-26JULY2020\TEMP1.csv')
print(result)
```

```
   X   Y
0  0  1.5
1  1  2.0
2  2  2.5
3  3  3.0
4  4  3.5
5  5  4.0
6  6  4.5
7  8  5.0
```



x	y	$x - \dot{x}$	$y - \dot{y}$	$(x - \dot{x})^2$	$(x - \dot{x})(y - \dot{y})$
1	3	-2	-0.6	4	1.2
2	4	-1	0.4	1	-0.4
3	2	0	-1.6	0	0
4	4	1	0.4	1	0.4
5	5	2	1.4	4	2.8

$$m = \sum \frac{(x - \dot{x})(y - \dot{y})}{(x - \dot{x})^2} = \frac{4}{10}$$

$$m = 0.4$$

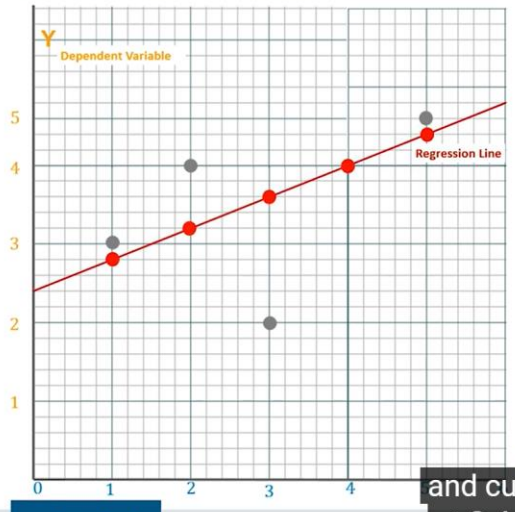
$$c = 2.4$$

$$y = 0.4x + 2.4$$

For given $m = 0.4$ & $c = 2.4$, lets predict values for y for $x = \{1,2,3,4,5\}$

$$y = 0.4 \times 1 + 2.4 = 2.8$$
$$y = 0.4 \times 2 + 2.4 = 3.2$$
$$y = 0.4 \times 3 + 2.4 = 3.6$$
$$y = 0.4 \times 4 + 2.4 = 4.0$$
$$y = 0.4 \times 5 + 2.4 = 4.4$$

Mean Square Error



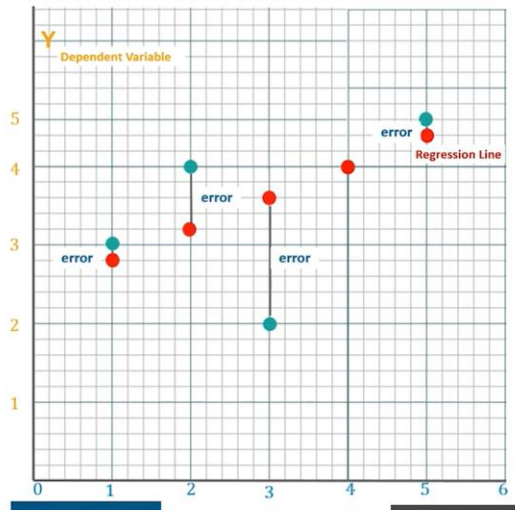
$m = 0.4$
$c = 2.4$
$y = 0.4x + 2.4$

For given $m = 0.4$ & $c = 2.4$, lets predict values for y for $x = \{1,2,3,4,5\}$

$y = 0.4 \times 1 + 2.4 = 2.8$
$y = 0.4 \times 2 + 2.4 = 3.2$
$y = 0.4 \times 3 + 2.4 = 3.6$
$y = 0.4 \times 4 + 2.4 = 4.0$
$y = 0.4 \times 5 + 2.4 = 4.4$

and cutting y-axis at 2.4 as the line

Mean Square Error



Distance between actual & predicted value

Like or in other words,

$$m = 0.4$$
$$c = 2.4$$
$$y = 0.4x + 2.4$$

For given $m = 0.4$ & $c = 2.4$, lets predict values for y for $x = \{1,2,3,4,5\}$

$$y = 0.4 \times 1 + 2.4 = 2.8$$
$$y = 0.4 \times 2 + 2.4 = 3.2$$
$$y = 0.4 \times 3 + 2.4 = 3.6$$
$$y = 0.4 \times 4 + 2.4 = 4.0$$
$$y = 0.4 \times 5 + 2.4 = 4.4$$

```
function gradientDescent() {
  var learning_rate = 0.05;
  for (var i = 0; i < data.length; i++) {
    var x = data[i].x;
    var y = data[i].y;
    var guess = m * x + b;
    var error = y - guess;
    m = m + error * x * learning_rate;
    b = b + error * learning_rate;
  }
```

What Is MLE?

At its simplest, MLE is a method for estimating parameters. Every time we fit a statistical or machine learning model, we are estimating parameters. A single variable linear regression has the equation:

$$Y = B0 + B1^{*}X$$

Our goal when we fit this model is to estimate the parameters B0 and B1 given our observed values of Y and X. We use Ordinary Least Squares (OLS), not MLE, to fit the linear regression model and estimate B0 and B1. But similar to OLS, MLE is a way to estimate the parameters of a model, given what we observe.

MLE asks the question, **"Given the data that we observe (our sample), what are the model parameters that maximize the likelihood of the observed data occurring?"**

A Simple Example

That's quite a mouthful. Let's use a simple example to show what we mean. Say we have a covered box containing an unknown number of red and black balls. If we randomly choose 10 balls from the box with replacement, and we end up with 9 black ones and only 1 red one, what does that tell us about the balls in the box?

Let's say we start out believing there to be an equal number of red and black balls in the box, what's the probability of observing what we observed?

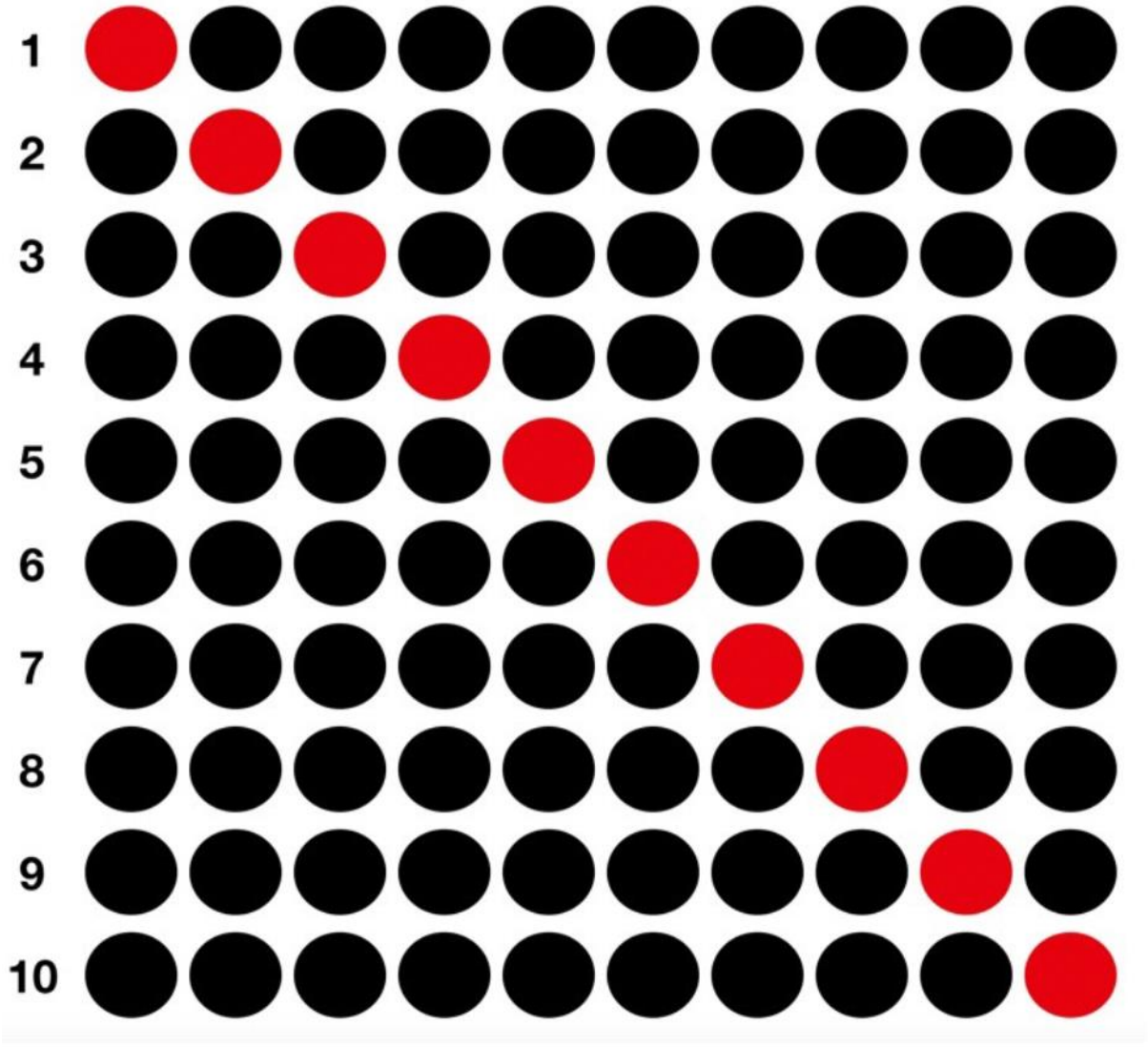
Probability of drawing 9 black and 1 red (assuming 50% are black):

We can do this 10 possible ways (see picture below).

Each of the 10 has probability = 0.5^10 = 0.097%

Since there are 10 possible ways, we multiply by 10:

Probability of 9 black and 1 red = 10 * 0.097% = **0.977%**



10 possible ways to draw 1 red ball and 9 black ones

```
import numpy as np
# Simulate drawing 10 balls 100000 times to see how frequently
# we get 9
trials = [np.random.binomial(10, 0.5) for i in range(1000000)]
print('Probability = ' + str(round(float(sum([1 for i\
                                        in trials if i==9]))\
                                    /len(trials),5)*100) + '%')
```

```
Probability = 0.9769999999999999%
```

The simulated probability is really close to our calculated probability (they're not exact matches because the simulated probability has variance).

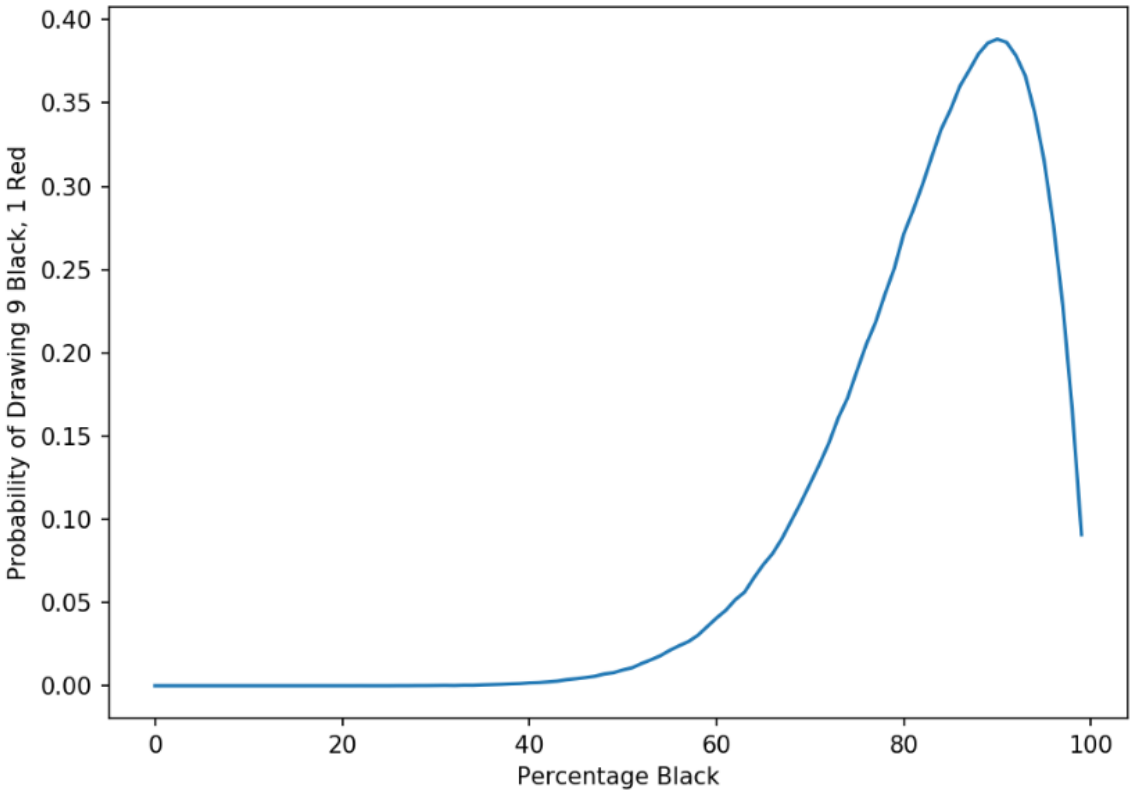
So our takeaway is that the likelihood of picking out as many black balls as we did, assuming that 50% of the balls in the box are black, is extremely low. Being reasonable folks, we would hypothesize that the percentage of balls that are black must not be 50%, but something higher. Then what's the percentage?

This is where MLE comes in. **Recall that MLE is a way for us to estimate parameters. The parameter in question is the percentage of balls in the box that are black colored.**

> MLE asks what should this percentage be to maximize the likelihood of observing what we observed (pulling 9 black balls and 1 red one from the box).

We can use Monte Carlo simulation to explore this. The following block of code loops through a range of probabilities (the percentage of balls in the box that are black). For each probability, we simulate drawing 10 balls 100,000 times in order to see how often we end up with 9 black ones and 1 red one.

We end up with the following plot:



Probabilities of drawing 9 black and 1 red balls

See that peak? That's what we're looking for. The value of percentage black where the probability of drawing 9 black and 1 red ball is maximized is its maximum likelihood estimate — **the estimate of our parameter (percentage black) that most conforms with what we observed**.

So MLE is effectively performing the following:

- Write a probability function that connects the probability of what we observed with the parameter that we are trying to estimate: we can write ours as ***P(9 black, 1 red | percentage black=b)*** — the probability of drawing 9 black and 1 red balls given that the percentage of balls in the box that are black is equal to b.
- Then we find the value of b that maximizes ***P(9 black, 1 red | percentage black=b)***.

It's hard to eyeball from the picture but the value of percentage black that maximizes the probability of observing what we did is 90%. Seems obvious right? And while this result seems obvious to a fault, the underlying fitting methodology that powers MLE is actually very powerful and versatile.

In statistics, **Maximum Likelihood Estimation (MLE)** is a method of estimating the parameters of a probability distribution by maximizing a likelihood function, so that under the assumed statistical model the observed data is most probable.

The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimate.

The logic of maximum likelihood is both intuitive and flexible, and as such the method has become a dominant means of statistical inference.

Maximum Likelihood Estimation

- Let's use a very simple sequence model
  - every position is independent of the others
  - every position generated from the same multinomial distribution
- we want to estimate the parameters Pr(a), Pr(c), Pr(g), Pr(t)
- and we're given the sequences

  accgcgctta
  gcttagtgac
  tagccgttac

$$\Pr(a) = \frac{n_a}{\sum_i n_i}$$

- then the maximum likelihood estimates are the observed frequencies of the bases

$$\Pr(a) = \frac{6}{30} = 0.2 \qquad \Pr(g) = \frac{7}{30} = 0.233$$

$$\Pr(c) = \frac{9}{30} = 0.3 \qquad \Pr(t) = \frac{8}{30} = 0.267$$

What are parameters?

Often in machine learning we use a model to describe the process that results in the data that are observed.

Example:

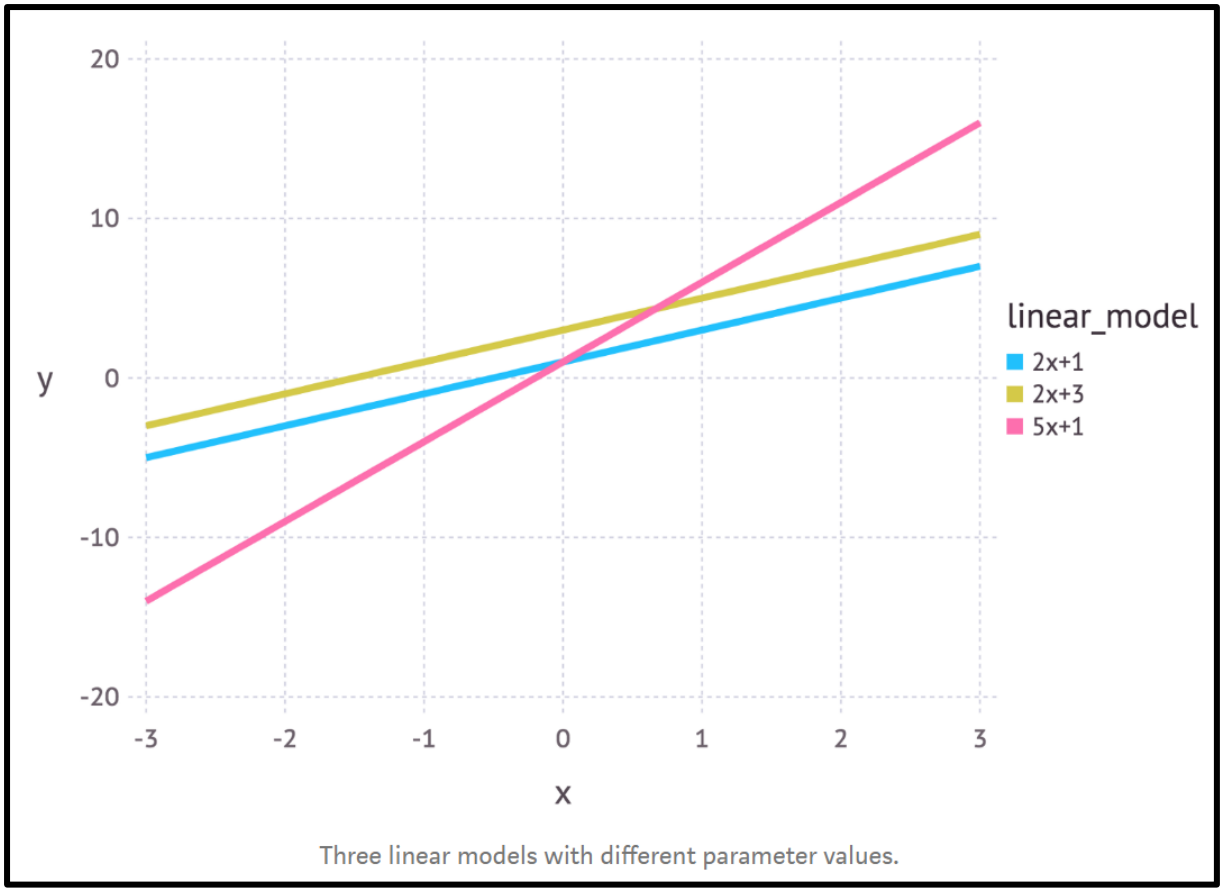
(i) We may ***use a random forest model*** to classify whether customers may cancel a subscription from a service (known as churn modelling) or

(ii) We may ***use a linear model*** to predict the revenue that will be generated for a company depending on how much they may spend on advertising (this would be an example of linear regression).

Each model contains its own set of parameters that ultimately defines what the model looks like.

For a linear model we can write this as $y = mx + c$. In this example x could represent the advertising spend and y might be the revenue generated. ***m* and *c* are parameters for this model**.

Different values for these parameters will give different lines (see figure below).



Three linear models with different parameter values.

Parameters define a blueprint for the model.

It is only when specific values are chosen for the parameters that we get an instantiation for the model that describes a given phenomenon.

Maximum likelihood estimation or otherwise noted as MLE is a popular mechanism which is used to estimate the model parameters of a regression model. Other than regression, it is very often used in statics to estimate the parameters of various distribution models.

Likelihood: The state of being probable; Something that is probable.

The probability of a given sample being randomly drawn regarded as a function of the parameters of the population.

The likelihood ratio is the ratio of this to the maximized likelihood.

Maximum Likelihood:

1. (Statistics) the probability of randomly drawing a given sample from a population maximized over the possible values of the population parameters

2. (Statistics) the non-Bayesian rule that, given an experimental observation, one should utilize as point estimates of parameters of a distribution those values which give the highest conditional probability to that observation, irrespective of the prior probability assigned to the parameters

Normal / Gaussian distribution

In probability the normal or gaussian distribution is a very famous continous probability distribution. It basically depends on two factors — the mean and the standard deviation. The mean of the distribution determines the location of the center of the graph and the standard deviation determines the height and width of the graph. So notation of normal distribution becomes:

$$N(\mu, \sigma^2) \quad ; \ \mu \text{ - mean} \ , \sigma^2 \text{ - variance}$$

PDF of Normal distribution is:

$$f(x \mid \mu, \ \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

2. Multiple linear regression

In multiple linear regression, there are several independent variables or functions of independent variables.

$$y_i=\beta_0+ \beta_1 x_{i1}+\ldots+\beta_p x_{ip}+e_i$$

y_i = the value of the i^{th} case of the dependent variable

p = the number of predictors (independent variables)

β_j = the value of the j^{th} coefficient, $j=0,\ldots,p$

x_{ij} = the value of the i^{th} case of the j^{th} predictor

e_i = the error in the observed value for the i^{th} case, or the difference between the predicted value of the dependent variable and its true value.

For example, adding a term in x_i^2 to the preceding regression gives: parabola:

12

2. Multiple linear regression

For example, adding a term in xi^2 to the preceding regression gives:

parabola:

$$y_i=\beta_0+ \beta_1 x_{i1}+\beta_2 x_i^2+e_i \quad \text{where, } i = 1,\ldots n$$

This is still linear regression; although the expression on the right hand side is quadratic in the independent variable x_i, it is linear in the parameters β_0, β_1 and β_2.

Example 1 – Multiple Linear regression
Main Effect

1. A researcher has conducted 28 experiments with different process parameters as shown below. Obtain the regression equation.

Sl. No.	A Vel mm/min	B Feed mm/min	C DOC mm	Ra µm
1	55	0.04	0.1	438.33
2	55	0.04	0.1	431.3
3	93	0.04	0.1	560.21
4	93	0.04	0.1	567.02
5	55	0.12	0.1	722.63
6	55	0.12	0.1	723.55
7	93	0.12	0.1	745.65
8	93	0.12	0.1	725.18
9	55	0.04	0.2	429.38
10	55	0.04	0.2	459.27
11	93	0.04	0.2	387.77
12	93	0.04	0.2	434.48
13	55	0.12	0.2	836.55
14	55	0.12	0.2	849.7

Sl. No.	A Vel mm/min	B Feed mm/min	C DOC mm	Ra µm
15	93	0.12	0.2	820.37
16	93	0.12	0.2	853.56
17	74	0.08	0.15	566.22
18	74	0.08	0.15	580.79
19	74	0.08	0.15	568.24
20	74	0.08	0.15	544.21
21	55	0.08	0.15	482.28
22	93	0.08	0.15	467.39
23	74	0.04	0.15	364.21
24	74	0.12	0.15	710.37
25	74	0.08	0.1	408.61
26	74	0.08	0.2	416.69
27	74	0.08	0.15	565.67
28	74	0.08	0.15	555.24

Lalvani, Mehtha, Jain, (2008). Experimental investigations of cutting parameters influence on cutting forces and surface roughness in finish hard turning of MDN250 steel, Journal of materials processing technology 2 0 6 (2 0 0 8) 167–179

14

Example – Multiple Linear regression
SPSS Output

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	186.783	105.617		1.768	.090
	A	.552	1.036	.055	.533	.599
	B	4049.431	491.915	.857	8.232	.000
	C	183.656	393.532	.049	.467	.645

a. Dependent Variable: Ra

Equation of the line is **y = 186.783 + 0.552A + 4049.431B + 183.656C**

- Only B has a significant relationship influence on Surface Roughness.

Model Summary[b]

Model	R	R^2	Adjusted R^2	Std. Error of the Estimate	Change Statistics				
					R^2 Change	F Change	df1	df2	Sig. F Change
1	.860[a]	.740	.707	83.48075	.740	22.756	3	24	.000

a. Predictors: (Constant), C, B, A

b. Dependent Variable: Ra

Summary

- R, the **multiple correlation coefficient**, is the linear correlation between the observed and model-predicted values of the dependent variable. Its large value (0.860) indicates a strong relationship.
- R^2, the coefficient of **determination**, is the squared value of the multiple correlation coefficient. It shows that almost three fourth (74%) the variation in Ra is explained by the model.
- With the linear regression model, the **error** of your estimate is about **83.48**.

16

SPSS output:

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	475753.938	3	158584.646	22.756	.000[a]
	Residual	167256.850	24	6969.035		
	Total	643010.788	27			

a. Predictors: (Constant), C, B, A

b. Dependent Variable: Ra

Summary

- The regression and residual sums of squares indicates that the variation in Ra is explained by the model to about **74%** (474754/643011).
- The significance value of the F statistic is less than 0.05, which means that the variation **explained by the model is not due to chance**. Also, null hypothesis stands rejected.

3.2 DESCRIPTION OF THE DATA AND MODEL

The data consist of n observations on a dependent or response variable Y and p predictor or explanatory variables, $X_1, X_2, \cdots, X_p$. The observations are usually represented as in Table 3.1. The relationship between Y and $X_1, X_2, \cdots, X_p$ is formulated as a linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon, \tag{3.1}$$

where $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ are constants referred to as the model *partial* regression coefficients (or simply as the *regression coefficients*) and ε is a random disturbance or error. It is assumed that for any set of fixed values of $X_1, X_2, \cdots, X_p$ that fall within the range of the data, the linear equation (3.1) provides an acceptable approximation of the true relationship between Y and the X's (Y is approximately a linear function of the X's, and ε measures the discrepancy in that approximation). In particular, ε contains no systematic information for determining Y that is not already captured by the X's.

Table 3.1 Notation for Data Used in Multiple Regression Analysis

Observation	Response	Predictors			
Number	Y	X_1	X_2	$\cdots$	X_p
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
3	y_3	x_{31}	x_{32}	$\cdots$	x_{3p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

According to (3.1), each observation in Table 3.1 can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \cdots, n, \tag{3.2}$$

where y_i represents the ith value of the response variable Y, $x_{i1}, x_{i2}, \cdots, x_{ip}$ represent values of the predictor variables for the ith unit (the ith row in Table 3.1), and ε_i represents the error in the approximation of y_i.

Multiple Linear Regression is an extension (or Generalization) of Simple Linear Regression.

We can similarly think that Simple Linear Regression is a special case of Multiple Linear Regression because all Simple Linear Regression results can be obtained using the Multiple Regression results when the number of predictor variables p = 1.

3.3 EXAMPLE: SUPERVISOR PERFORMANCE DATA

Throughout this chapter we use data from a study in *industrial psychology* (management) to illustrate some of the standard regression results. A recent survey of the clerical employees of a large financial organization included questions related to employee satisfaction with their supervisors. There was a question designed to measure the overall performance of a supervisor, as well as questions that were related to specific activities involving interaction between supervisor and employee. An exploratory study was undertaken to try to explain the relationship between specific supervisor characteristics and overall satisfaction with supervisors as perceived by the employees. Initially, six questionnaire items were chosen as possible explanatory variables. Table 3.2 gives the description of the variables in the study.

Table 3.2 Description of Variables in Supervisor Performance Data

Variable	Description
Y	Overall rating of job being done by supervisor
X_1	Handles employee complaints
X_2	Does not allow special privileges
X_3	Opportunity to learn new things
X_4	Raises based on performance
X_5	Too critical of poor performance
X_6	Rate of advancing to better jobs

As can be seen from the list, there are two broad types of variables included in the study. Variables X_1, X_2, and X_5 relate to direct interpersonal relationships between employee and supervisor, whereas variables X_3 and X_4 are of a less personal nature and relate to the job as a whole. Variable X_6 is not a direct evaluation of the supervisor but serves more as a general measure of how the employee perceives his or her own progress in the company.

The data for the analysis were generated from the individual employee response to the items on the survey questionnaire. The response on any item ranged from 1 through 5, indicating very satisfactory to very unsatisfactory, respectively. A dichotomous index was created to each item by collapsing the response scale to two categories: {1,2}, to be interpreted as a favorable response, and {3,4,5}, representing an unfavorable response. The data were collected in 30 departments selected at random from the organization. Each department had approximately 35 employees and one supervisor. The data to be used in the analysis, given in Table 3.3, were obtained by aggregating responses for departments to get the proportion of favorable responses for each item for each department. The resulting data therefore consist of 30 observations on seven variables, one observation for each department. We refer to this data set as the Supervisor Performance data. The data set can also be found at the book's Website.[1]

A linear model of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_6 X_6 + \varepsilon,$$

relating Y and the six explanatory variables, is assumed.

3.4 PARAMETER ESTIMATION

Based on the available data, we wish to estimate the parameters $\beta_0, \beta_1, \cdots, \beta_p$.

We use the method of Least Squares to minimize the sum of squares errors.

The error term can be written as:

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}, \quad i = 1, 2, \cdots, n.$$

The sum of squares of these errors is

$$S(\beta_0, \beta_1, \cdots, \beta_p) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2.$$

By a direct application of calculus, it can be shown that the least squares estimates $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$, which minimize $S(\beta_0, \beta_1, \cdots, \beta_p)$, are given by the solution of a system of linear equations known as the *normal equations*.[2] The estimate $\hat{\beta}_0$ is usually referred to as the *intercept* or *constant*, and $\hat{\beta}_j$ as the *estimate* of the (partial) regression coefficient of the predictor X_j.

We assume that the system of equations is solvable and has a unique solution.

Using the estimated regression coefficients $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$, we write the fitted least squares regression equation as

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p. \tag{3.6}$$

For each observation in our data we can compute

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}, \quad i = 1, 2, \cdots, n. \tag{3.7}$$

These are called the *fitted* values. The corresponding *ordinary* least squares residuals are given by

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \cdots, n. \tag{3.8}$$

An unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n - p - 1}, \tag{3.9}$$

where

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2, \tag{3.10}$$

is the *sum of squared residuals*. The number $n - p - 1$ in the denominator of (3.9) is called the *degrees of freedom* (df). It is equal to the number of observations minus the number of estimated regression coefficients.

3.5 INTERPRETATIONS OF REGRESSION COEFFICIENTS

The interpretation of the regression coefficients in a multiple regression equation is a source of common confusion. The simple regression equation represents a line, while the multiple regression equation represents a plane (in cases of two predictors) or a hyperplane (in cases of more than two predictors). In multiple regression, the coefficient β_0, called the *constant coefficient*, is the value of Y when $X_1 = X_2 = \cdots = X_p = 0$, as in simple regression. The regression coefficient $\beta_j, j = 1, 2, \cdots, p$, has several interpretations. It may be interpreted as the change in Y corresponding to a unit change in X_j when all other predictor variables are held constant. Magnitude of the change is not dependent on the values at which the other predictor variables are fixed. In practice, however, the predictor variables may be inherently related, and holding some of them constant while varying the others may not be possible.

The regression coefficient β_j is also called the *partial regression coefficient* because β_j represents the contribution of X_j to the response variable Y after it has been adjusted for the other predictor variables. What does "adjusted for" mean in multiple regression? Without loss of any generality, we address this question using the simplest multiple regression case where we have two predictor variables. When $p = 2$, the model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon. \tag{3.11}$$

We use the variables X_1 and X_2 from the Supervisor data to illustrate the concepts. A statistical package gives the estimated regression equation as

$$\hat{Y} = 15.3276 + 0.7803 X_1 - 0.0502 X_2. \tag{3.12}$$

The coefficient of X_1 suggests that each unit of X_1 adds 0.7803 to Y when the value of X_2 is held fixed. As we show below, this is also the effect of X_1 after adjusting for X_2. Similarly, the coefficient of X_2 suggests that each unit of X_2 subtracts about 0.0502 from Y when the value of X_1 is held fixed. This is also the effect of X_2 after adjusting for X_1.

This interpretation can be easily understood when we consider the fact that the multiple regression equation can be obtained from a series of simple regression equations. For example, the coefficient of X_2 in (3.12) can be obtained as follows:

1. Fit the simple regression model that relates Y to X_1. Let the residuals from this regression be denoted by $\mathbf{e_{Y \cdot X_1}}$. This notation indicates that the variable that comes before the dot is treated as a response variable and the variable that comes after the dot is considered as a predictor. The fitted regression equation is

$$\hat{Y} = 14.3763 + 0.754610 X_1. \tag{3.13}$$

2. Fit the simple regression model that relates X_2 (considered temporarily here as a response variable) to X_1. Let the residuals from this regression be denoted by $\mathbf{e_{X_2 \cdot X_1}}$. The fitted regression equation is

$$\hat{X}_2 = 18.9654 + 0.513032 X_1. \tag{3.14}$$

3. Fit the simple regression model that relates the above two residuals. In this regression, the response variable is $\mathbf{e_{Y \cdot X_1}}$ and the predictor variable is $\mathbf{e_{X_2 \cdot X_1}}$. The fitted regression equation is

$$\hat{\mathbf{e}}_{Y \cdot X_1} = 0 - 0.0502 \mathbf{e_{X_2 \cdot X_1}}. \tag{3.15}$$

The tow coeffieceints are equal to -0.0502. In fact their standard errors are also same.

are equal to -0.0502. In fact, their standard errors are also the same. What's the intuition here? In the first step, we found the linear relationship between Y and X_1. The residual from this regression is Y after taking or partialling out the linear effects of X_1. In other words, the residual is that part of Y that is not linearly related to X_1. In the second step we do the same thing, replacing Y by X_2, so the residual is the part of X_2 that is not linearly related to X_1. In the third step we look for the linear relationship between the Y residual and the X_2 residual. The resultant regression coefficient represents the effect of X_2 on Y after taking out the effects of X_1 from both Y and X_2.

The regression coefficient β_j is the partial regression coefficient because it represents the contribution of X_j to the response variable Y after both variables have been linearly adjusted for the other predictor variables (see also Exercise 3.5).

Note that the estimated intercept in the regression equation in (3.15) is zero because the two sets of residuals have a mean of zero (they sum up to zero). The same procedures can be applied to obtain the multiple regression coefficient of X_1 in (3.12). Simply interchange X_2 by X_1 in the above three steps. This is left as an exercise for the reader.

From the above discussion we see that the simple and the multiple regression coefficients are not the same unless the predictor variables are uncorrelated. In non-experimental or observational data, the predictor variables are rarely uncorrelated. In an experimental setting, in contrast, the experimental design is often set up to produce uncorrelated explanatory variables because in an experiment the researcher sets the values of the predictor variables. So in samples derived from experiments it may be the case that the explanatory variables are uncorrelated and hence the simple and multiple regression coefficients in that sample would be the same.

3.7 PROPERTIES OF THE LEAST SQUARES ESTIMATORS

Under certain standard regression assumptions (to be stated in Chapter 4), the least squares estimators have the properties listed below. A reader familiar with matrix algebra will find concise statements of these properties employing matrix notation in the Appendix at the end of the chapter.

1. The estimator $\hat{\beta}_j, j = 0, 1, \cdots, p$, is an unbiased estimate of β_j and has a variance of $\sigma^2 c_{jj}$, where c_{jj} is the jth diagonal element of the inverse of a matrix known as the *corrected sums of squares and products* matrix. The covariance between $\hat{\beta}_i$ and $\hat{\beta}_j$ is $\sigma^2 c_{ij}$, where c_{ij} is the element in the ith row and jth column of the inverse of the corrected sums of squares and products matrix. For all unbiased estimates that are linear in the observations the least squares estimators have the smallest variance. Thus, the least squares estimators are said to be BLUE (*best linear unbiased estimators*).

2. The estimator $\hat{\beta}_j, j = 0, 1, \cdots, p$, is normally distributed with mean β_j and variance $\sigma^2 c_{jj}$.

3. $W = \text{SSE}/\sigma^2$ has a χ^2 distribution with $n - p - 1$ degrees of freedom, and $\hat{\beta}_j$'s and $\hat{\sigma}^2$ are distributed independently of each other.

4. The vector $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)$ has a $(p+1)$-dimensional normal distribution with mean vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)$ and variance-covariance matrix with elements $\sigma^2 c_{ij}$.

The results above enable us to test various hypotheses about individual regression parameters and to construct confidence intervals.

After fitting the linear model to a given data set, an assessment is made of the adequacy of fit.

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

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

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

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

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 $X_1, X_2, \cdots, X_p$.

The value of R^2 for the Supervisor Performance data is 0.73, showing that about 73% of the total variation in the overall rating of the job being done by the supervisor can be accounted for by the six variables.

When the model fits the data well, it is clear that the value of R^2 is close to unity. With a good fit, the observed and predicted values will be close to each other, and $\sum(y_i - \hat{y}_i)^2$ will be small. Then R^2 will be near unity. On the other hand, if there is no linear relationship between Y and the predictor variables, $X_1, \cdots, X_p$, the linear model gives a poor fit, the best predicted value for an observation y_i would be $\bar{y}$; that is, in the absence of any relationship with the predictors, the best estimate of any value of Y is the sample mean, because the sample mean minimizes the sum of squared deviations. So in the absence of any linear relationship between Y and the X's, R^2 will be near zero. The value of R^2 is used as a summary measure to judge the fit of the linear model to a given body of data. As pointed out in Chapter 2, a large value of R^2 does not necessarily mean that the model fits the data well. As we outline in Section 3.10, a more detailed analysis is needed to ensure that the model adequately describes the data.

A quantity related to R^2, known as the *adjusted R-squared*, R_a^2, is also used for judging the goodness of fit. It is defined as

$$R_a^2 = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}, \tag{3.30}$$

which is obtained from R^2 in (3.29) after dividing SSE and SST by their respective degrees of freedom. From (3.30) and (3.29) it follows that

$$R_a^2 = 1 - \frac{n-1}{n-p-1}(1-R^2). \tag{3.31}$$

R_a^2 is sometimes used to compare models having different numbers of predictor variables. (This is described in Chapter 11.) In comparing the goodness of fit of models with different numbers of explanatory variables, R_a^2 tries to "adjust" for the unequal number of variables in the different models. Unlike R^2, R_a^2 cannot be interpreted as the proportion of total variation in Y accounted for by the predictors. Many regression packages provide values for both R^2 and R_a^2.

3.10 TESTS OF HYPOTHESES IN A LINEAR MODEL

In addition to looking at hypotheses about individual β's, several different hypotheses are considered in connection with the analysis of linear models. The most commonly investigated hypotheses are

1. All the regression coefficients associated with the predictor variables are zero.
2. Some of the regression coefficients are zero.
3. Some of the regression coefficients are equal to each other.
4. The regression parameters satisfy certain specified constraints.

The different hypotheses about the regression coefficients can all be tested in the same way by a unified approach. Rather than describing the individual tests, we first describe the general unified approach, then illustrate specific tests using the Supervisor Performance data.

The model given in (3.1) will be referred to as the *full model* (FM). The null hypothesis to be tested specifies values for some of the regression coefficients. When these values are substituted in the full model, the resulting model is called the *reduced model* (RM). The number of *distinct* parameters to be estimated in the reduced model is smaller than the number of parameters to be estimated in the full model. Accordingly, we wish to test

H_0 : Reduced model is adequate against H_1 : Full model is adequate.

Note that the reduced model is *nested.* A set of models are said to be nested if they can be obtained from a larger model as special cases. The test for these nested hypotheses involves a comparison of the goodness of fit that is obtained when using the full model, to the goodness of fit that results using the reduced model specified by the null hypothesis. If the reduced model gives as good a fit as the full model,

the null hypothesis, which defines the reduced model (by specifying some values of β_j), is not rejected. This procedure is described formally as follows.

Let $\hat{y}_i$ and $\hat{y}_i^*$ be the values predicted for y_i by the full model and the reduced model, respectively. The lack of fit in the data associated with the full model is the sum of the squared residuals obtained when fitting the full model to the data. We denote this by SSE(FM), the sum of squares due to error associated with the full model,

$$\text{SSE(FM)} = \sum (y_i - \hat{y}_i)^2. \tag{3.38}$$

Similarly, the lack of fit in the data associated with the reduced model is the sum of the squared residuals obtained when fitting the reduced model to the data. This quantity is denoted by SSE(RM), the sum of squares due to error associated with the reduced model,

$$\text{SSE(RM)} = \sum (y_i - \hat{y}_i^*)^2. \tag{3.39}$$

In the full model there are $p + 1$ regression parameters $(\beta_0, \beta_1, \beta_2, \cdots, \beta_p)$ to be estimated. Let us suppose that for the reduced model there are k distinct parameters. Note that SSE(RM) $\geq$ SSE(FM) because the additional parameters (variables) in the full model cannot increase the residual sum of squares. Note also that the difference SSE(RM) – SSE(FM) represents the increase in the residual sum of squares due to fitting the reduced model. If this difference is large, the reduced model is inadequate. To see whether the reduced model is adequate, we use the ratio

$$F = \frac{[\text{SSE(RM)} - \text{SSE(FM)}]/(p + 1 - k)}{\text{SSE(FM)}/(n - p - 1)}. \tag{3.40}$$

This ratio is referred to as the F-Test. Note that we divide SSE(RM) – SSE(FM) and SSE(FM) in the above ratio by their respective degrees of freedom to compensate for the different number of parameters involved in the two models as well as to ensure that the resulting test statistic has a standard statistical distribution. The full model has $p+1$ parameters, hence SSE(FM) has $n-p-1$ degrees of freedom. Similarly, the reduced model has k parameters and SSE(RM) has $n - k$ degrees of freedom. Consequently, the difference SSE(RM) – SSE(FM) has $(n - k) - (n - p - 1) = p + 1 - k$ degrees of freedom. Therefore, the observed F-ratio in (3.40) has F-distribution with $p + 1 - k$ and $n - p - 1$ degrees of freedom.

What Is the Coefficient of Determination?

The coefficient of determination is a statistical measurement that examines how differences in one variable can be explained by the difference in a second variable, when predicting the outcome of a given event. In other words, this coefficient, which is more commonly known as R-squared (or R^2), assesses how strong the linear relationship is between two variables, and is heavily relied on by researchers when conducting trend analysis. To cite an example of its application, this coefficient may contemplate the following question: if a woman becomes pregnant on a certain day, what is the likelihood that she would deliver her baby on a particular date in the future? In this scenario, this metric aims to calculate the correlation between two related events: conception and birth.

KEY TAKEAWAYS

- The coefficient of determination is a complex idea centered on the statistical analysis of models for data.
- The coefficient of determination is used to explain how much variability of one factor can be caused by its relationship to another factor.
- This coefficient is commonly known as R-squared (or R^2), and is sometimes referred to as the "goodness of fit."
- This measure is represented as a value between 0.0 and 1.0, where a value of 1.0 indicates a perfect fit, and is thus a highly reliable model for future forecasts, while a value of 0.0 would indicate that the model fails to accurately model the data at all.

Understanding the Coefficient of Determination

The coefficient of determination is a measurement used to explain how much variability of one factor can be caused by its relationship to another related factor. This correlation, known as the "goodness of fit," is represented as a value between 0.0 and 1.0. A value of 1.0 indicates a perfect fit, and is thus a highly reliable model for future forecasts, while a value of 0.0 would indicate that the calculation fails to accurately model the data at all. But a value of 0.20, for example, suggests that 20% of the dependent variable is predicted by the independent variable, while a value of 0.50 suggests that 50% of the dependent variable is predicted by the independent variable, and so forth.

Graphing the Coefficient of Determination

On a graph, the goodness of fit measures the distance between a fitted line and all of the data points that are scattered throughout the diagram. The tight set of data will have a regression line that's close to the points and have a high level of fit, meaning that the distance between the line and the data is small. Although a good fit has an R^2 close to 1.0, this number alone cannot determine whether the data points or predictions are biased. It also doesn't tell analysts whether the coefficient of determination value is intrinsically good or bad. It is at the discretion of the user to evaluate the meaning of this correlation, and how it may be applied in the context of future trend analyses.

What Does R-Squared Tell You?

R-squared values range from 0 to 1 and are commonly stated as percentages from 0% to 100%. An R-squared of 100% means that all movements of a security (or another dependent variable) are completely explained by movements in the index (or the independent variable(s) you are interested in).

In investing, a high R-squared, between 85% and 100%, indicates the stock or fund's performance moves relatively in line with the index. A fund with a low R-squared, at 70% or less, indicates the security does not generally follow the movements of the index. A higher R-squared value will indicate a more useful beta figure. For example, if a stock or fund has an R-squared value of close to 100%, but has a beta below 1, it is most likely offering higher risk-adjusted returns.

The Difference Between R-Squared and Adjusted R-Squared

R-Squared only works as intended in a simple linear regression model with one explanatory variable. With a multiple regression made up of several independent variables, the R-Squared must be adjusted. The adjusted R-squared compares the descriptive power of regression models that include diverse numbers of predictors. Every predictor added to a model increases R-squared and never decreases it. Thus, a model with more terms may seem to have a better fit just for the fact that it has more terms, while the adjusted R-squared compensates for the addition of variables and only increases if the new term enhances the model above what would be obtained by probability and decreases when a predictor enhances the model less than what is predicted by chance. In an overfitting condition, an incorrectly high value of R-squared is obtained, even when the model actually has a decreased ability to predict. This is not the case with the adjusted R-squared.

Limitations of R-Squared

R-squared will give you an estimate of the relationship between movements of a dependent variable based on an independent variable's movements. It doesn't tell you whether your chosen model is good or bad, nor will it tell you whether the data and predictions are biased. A high or low R-square isn't necessarily good or bad, as it doesn't convey the reliability of the model, nor whether you've chosen the right regression. You can get a low R-squared for a good model, or a high R-square for a poorly fitted model, and vice versa.

What Is Multiple Linear Regression (MLR)?

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable.

In essence, multiple regression is the extension of ordinary least-squares (OLS) regression that involves more than one explanatory variable.

Formula and Calcualtion of Multiple Linear Regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

$y_i = \text{dependent variable}$

$x_i = \text{expanatory variables}$

$\beta_0 = \text{y-intercept (constant term)}$

$\beta_p = \text{slope coefficients for each explanatory variable}$

$\epsilon = \text{the model's error term (also known as the residuals)}$

KEY TAKEAWAYS

- Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.
- Multiple regression is an extension of linear (OLS) regression that uses just one explanatory variable.
- MLR is used extensively in econometrics and financial inference.

What Multiple Linear Regression (MLR) Can Tell You

Simple linear regression is a function that allows an analyst or statistician to make predictions about one variable based on the information that is known about another variable. Linear regression can only be used when one has two continuous variables—an independent variable and a dependent variable. The independent variable is the parameter that is used to calculate the dependent variable or outcome. A multiple regression model extends to several explanatory variables.

The multiple regression model is based on the following assumptions:

- There is a linear relationship between the dependent variables and the independent variables.
- The independent variables are not too highly correlated with each other.
- y_i observations are selected independently and randomly from the population.
- Residuals should be normally distributed with a mean of 0 and variance σ.

The coefficient of determination (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. R^2 always increases as more predictors are added to the MLR model even though the predictors may not be related to the outcome variable.

R^2 by itself can't thus be used to identify which predictors should be included in a model and which should be excluded. R^2 can only be between 0 and 1, where 0 indicates that the outcome cannot be predicted by any of the independent variables and 1 indicates that the outcome can be predicted without error from the independent variables.[1]

When interpreting the results of multiple regression, beta coefficients are valid while holding all other variables constant ("all else equal"). The output from a multiple regression can be displayed horizontally as an equation, or vertically in table form.[2]

Example How to Use Multiple Linear Regression (MLR)

As an example, an analyst may want to know how the movement of the market affects the price of ExxonMobil (XOM). In this case, their linear equation will have the value of the S&P 500 index as the independent variable, or predictor, and the price of XOM as the dependent variable.

In reality, there are multiple factors that predict the outcome of an event. The price movement of ExxonMobil, for example, depends on more than just the performance of the overall market. Other predictors such as the price of oil, interest rates, and the price movement of oil futures can affect the price of XOM and stock prices of other oil companies. To understand a relationship in which more than two variables are present, multiple linear regression is used.

Multiple linear regression (MLR) is used to determine a mathematical relationship among a number of random variables. In other terms, MLR examines how multiple independent variables are related to one dependent variable. Once each of the independent factors has been determined to predict the dependent variable, the information on the multiple variables can be used to create an accurate prediction on the level of effect they have on the outcome variable. The model creates a relationship in the form of a straight line (linear) that best approximates all the individual data points.[3]

Referring to the MLR equation above, in our example:

- y_i = dependent variable—the price of XOM
- x_{i1} = interest rates
- x_{i2} = oil price
- x_{i3} = value of S&P 500 index
- x_{i4}= price of oil futures
- B_0 = y-intercept at time zero
- B_1 = regression coefficient that measures a unit change in the dependent variable when x_{i1} changes - the change in XOM price when interest rates change
- B_2 = coefficient value that measures a unit change in the dependent variable when x_{i2} changes—the change in XOM price when oil prices change

The least-squares estimates, B_0, B_1, B_2…B_p, are usually computed by statistical software. As many variables can be included in the regression model in which each independent variable is differentiated with a number—1,2, 3, 4...p. The multiple regression model allows an analyst to predict an outcome based on information provided on multiple explanatory variables.

Still, the model is not always perfectly accurate as each data point can differ slightly from the outcome predicted by the model. The residual value, E, which is the difference between the actual outcome and the predicted outcome, is included in the model to account for such slight variations.

Assuming we run our XOM price regression model through a statistics computation software, that returns this output:

```
XOM Price = 75 - 1.5 interest rates + 7.8 oil price + 3.2 S&P 500 + 5.7 oil futures
R-Sq = 86.5%
```

An analyst would interpret this output to mean if other variables are held constant, the price of XOM will increase by 7.8% if the price of oil in the markets increases by 1%. The model also shows that the price of XOM will decrease by 1.5% following a 1% rise in interest rates. R^2 indicates that 86.5% of the variations in the stock price of Exxon Mobil can be explained by changes in the interest rate, oil price, oil futures, and S&P 500 index.

The Difference Between Linear and Multiple Regression

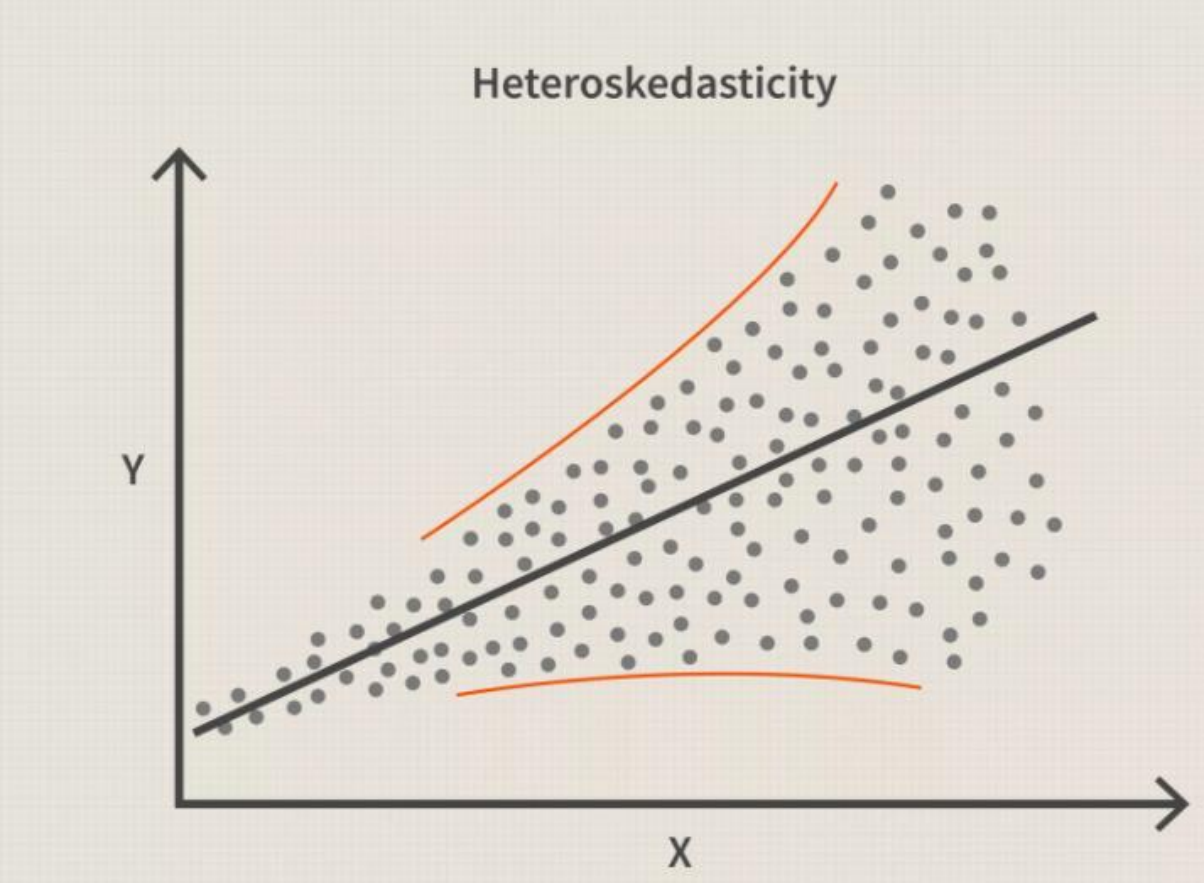
Ordinary linear squares (OLS) regression compares the response of a dependent variable given a change in some explanatory variables. However, it is rare that a dependent variable is explained by only one variable. In this case, an analyst uses multiple regression, which attempts to explain a dependent variable using more than one independent variable. Multiple regressions can be linear and nonlinear.

Multiple regressions are based on the assumption that there is a linear relationship between both the dependent and independent variables. It also assumes no major correlation between the independent variables.

What Is Heteroskedasticity?

In statistics, heteroskedasticity (or heteroscedasticity) happens when the standard deviations of a predicted variable, monitored over different values of an independent variable or as related to prior time periods, are non-constant. With heteroskedasticity, the tell-tale sign upon visual inspection of the residual errors is that they will tend to fan out over time, as depicted in the image below.

Heteroskedasticity often arises in two forms: conditional and unconditional. Conditional heteroskedasticity identifies nonconstant volatility related to prior period's (e.g., daily) volatility. Unconditional heteroskedasticity refers to general structural changes in volatility that are not related to prior period volatility. Unconditional heteroskedasticity is used when future periods of high and low volatility can be identified.



KEY TAKEAWAYS

- In statistics, heteroskedasticity (or heteroscedasticity) happens when the standard errors of a variable, monitored over a specific amount of time, are non-constant.
- With heteroskedasticity, the tell-tale sign upon visual inspection of the residual errors is that they will tend to fan out over time, as depicted in the image above.
- Heteroskedasticity is a violation of the assumptions for linear regression modeling, and so it can impact the validity of econometric analysis or financial models like CAPM.

Important: While heteroskedasticity does not cause bias in the coefficient estimates, it does make them less precise; lower precision increases the likelihood that the coefficient estimates are further from the correct population value.

The opposite of heteroskedastic is homoskedastic. Homoskedasticity refers to a condition in which the variance of the residual term is constant or nearly so. Homoskedasticity is one assumption of linear regression modeling. It is needed to ensure that the estimates are accurate, that the prediction limits for the dependent variable are valid, and that confidence intervals and p-values for the parameters are valid.

Linear Regression is one of the simplest and most widely used algorithms for Supervised machine learning problems where the output is a numerical quantitative variable and the input is a bunch of independent variables or single variable.

The math behind it is easy to understand and that's what makes Linear Regression one of my most favorite algorithms to work with. But this simplicity comes with a price.

When we decide to fit a Linear Regression model, we have to make sure that some conditions are satisfied or else our model will perform poorly or will give us incorrect interpretations. So what are some of these conditions that have to be met?

1. **Linearity**: X and the mean of Y have a Linear Relationship

2. **Homoscedasticity**: variance of the error terms is the same for all values of X.

3. **No collinearity:** independent variables are not highly correlated with each other

4.**Normality**: Y is normally distributed for any value of X.

If the above four conditions are satisfied, we can expect our Linear Regression model to perform well.

So how do we ensure the above conditions are met? Well, If I start going into the depth of all of the above conditions, it might result in a very long article. So for this article, I will go over the third condition of No collinearity meaning I will explain what Multicollinearity and how it is a problem in the first place and what can be done to overcome it.

When we have a Supervised Machine Learning Regression problem, We know we have a bunch of Independent variables and an Output variable which will be used to train our model and make predictions and interpretations.

In a Multivariate Linear Regression problem, we make predictions based off of the model trained and use the coefficients to make interpretations of the model for example:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

Multivariate Linear Regression

The above equation states that a unit increase in X1, will result in a B1 increase in the value of Y and a unit increase in X2 will result in a B2 increase in the value of Y.

The coefficients are mandatory in order to understand which variable has the highest influence on the model.

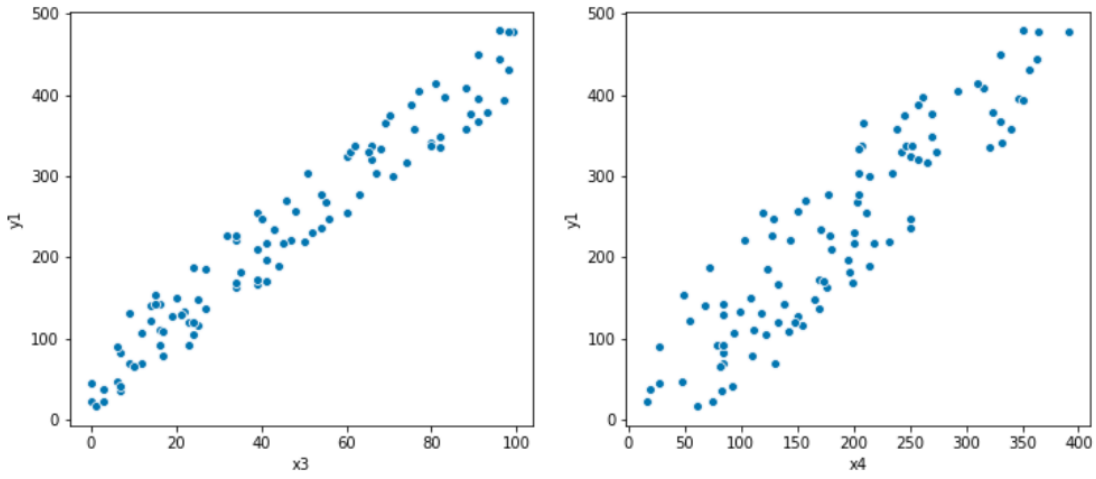
So how is multicollinearity a problem? Well, When we have independent variables that are highly related to each other, our coefficients won't be reliable and we cannot make accurate interpretations based on their values.

To explain this point further, I created two dummy input variables in Python and one dependent output variable.

```
x3 = np.random.randint(0,100,100)
x4 = 3*x3 + np.random.randint(0,100,100)
y1 = (4*x3) + np.random.randint(0,100,100)
```

Creating the scatterplot for the variables gives us:

```
plt.figure(figsize = (12,5))
plt.subplot(1,2,1)
plt.xlabel('x3')
sns.scatterplot(x3,y1)
plt.subplot(1,2,2)
plt.xlabel('x4')
sns.scatterplot(x4,y1)
```



Scatterplots for the Input variables against y1

The scatterplot shows that both x3 and x4 are have a linear relationship with y1. Lets look at the correlation matrix for the variables and see what else can we interpret.
I put my variables into a DataFrame by the name of S2 and created a correlation matrix.

```
S2.corr()
```

	X3	X4	y1
X3	1.000000	0.947306	0.970036
X4	0.947306	1.000000	0.908995
y1	0.970036	0.908995	1.000000

Correlation matrix

By the looks of the correlation matrix, it seems that both X3 and X4 not only have a high positive correlation with y1 but also are highly correlated with each other. Let's see how this will affect our results.

Before I fit a Linear Regression model to my variables, we have to understand the concept of **P-values and the Null hypothesis**.

The P-value is used to either reject or accept the Null Hypothesis.

The Null Hypothesis in our case is that '**The variable does not have a significant relation with y**".

If the P-value is less than the threshold of 0.005, **then we have to reject the Null hypothesis**, otherwise, **we have to accept it**. So let's move forward

I import the stats model from the scipy library and use it to fit an Ordinary Least Squares model to my variables.

The independent variables being X3 and X4 and the dependent variable being y1.

```
X = S2[['X3','X4']]
y = S2['y1']


import statsmodels.api as sm
from scipy import stats


X = sm.add_constant(X3)
est = sm.OLS(y,X)
est2 = est.fit()
print(est2.summary())
```

The results we get are:

OLS Regression Results

Dep. Variable:	y1	R-squared:	0.942
Model:	OLS	Adj. R-squared:	0.941
Method:	Least Squares	F-statistic:	786.7
Date:	Thu, 13 Aug 2020	Prob (F-statistic):	1.13e-60
Time:	22:45:47	Log-Likelihood:	-481.33
No. Observations:	100	AIC:	968.7
Df Residuals:	97	BIC:	976.5
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	50.0344	7.514	6.659	0.000	35.121	64.948
X3	4.4721	0.322	13.900	0.000	3.833	5.111
X4	-0.1289	0.102	-1.266	0.208	-0.331	0.073

Omnibus:	37.237	Durbin-Watson:	2.407
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6.794
Skew:	0.192	Prob(JB):	0.0335
Kurtosis:	1.782	Cond. No.	533.

Summary for the OLS method

We get a very high R2 score which shows that our model explains the variance in the model quite well. The coefficients on the other hand, tell an entirely different story.

The P-value for our X4 variable shows that we cannot reject the Null-hypothesis meaning **X4 does not have a significant relation with y.**

Furthermore, the coefficient is negative as well which cannot be possible as the scatterplots showed that y had a positive relationship will the independent variable.

So to sum it up, **Our coefficients are not reliable and our P-values cannot be trusted.**

Regression with one variable only

In the previous multivariate example, our results showed that X4 did not have a significant relation with y1.
So let us try to analyze y1 and X4 alone and see what we get.

```
X3 = S2['X4']
y1 = S2['y1']
import statsmodels.api as sm
from scipy import stats


X = sm.add_constant(X3)
est = sm.OLS(y1,X)
est2 = est.fit()
print(est2.summary())
```

After fitting our OLS model, we get

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     y1   R-squared:                       0.826
Model:                            OLS   Adj. R-squared:                  0.824
Method:                 Least Squares   F-statistic:                     466.1
Date:                Fri, 14 Aug 2020   Prob (F-statistic):           5.00e-39
Time:                        01:47:17   Log-Likelihood:                -536.12
No. Observations:                 100   AIC:                             1076.
Df Residuals:                      98   BIC:                             1081.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.7634     11.652      0.409      0.684     -18.361      27.887
X4             1.2109      0.056     21.589      0.000       1.100       1.322
==============================================================================
Omnibus:                       15.086   Durbin-Watson:                   2.319
Prob(Omnibus):                  0.001   Jarque-Bera (JB):                4.819
Skew:                           0.187   Prob(JB):                       0.0898
Kurtosis:                       1.992   Cond. No.                         465.
==============================================================================
```

The coefficient is now positive and **we can reject the Null Hypothesis** that X4 is not related to y1.

But one more thing we can take from this model is that our R-squared value has reduced significantly from 0.942 to 0.826.

So what does that tell us? Well, if our goal is prediction, we might need to think before removing variables but if our goal is an interpretation of each coefficient, then collinearity can be troublesome and we have to consider which variables to keep and which to remove.

How to Interpret a Correlation Coefficient *r*

By **Deborah J. Rumsey**

In **statistics**, the correlation coefficient *r* measures the strength and direction of a linear relationship between two variables on a **scatterplot**. The value of *r* is always between +1 and –1. To interpret its value, see which of the following values your correlation *r* is closest to:

- **Exactly –1.** A perfect downhill (negative) linear relationship
- **–0.70.** A strong downhill (negative) linear relationship
- **–0.50.** A moderate downhill (negative) relationship
- **–0.30.** A weak downhill (negative) linear relationship
- **0.** No linear relationship

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

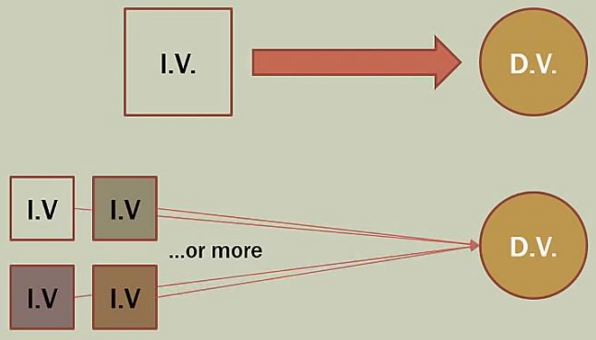
Remember that in this case, you would like to be able to predict the total travel time using both the miles traveled and number of deliveries on each trip.

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.



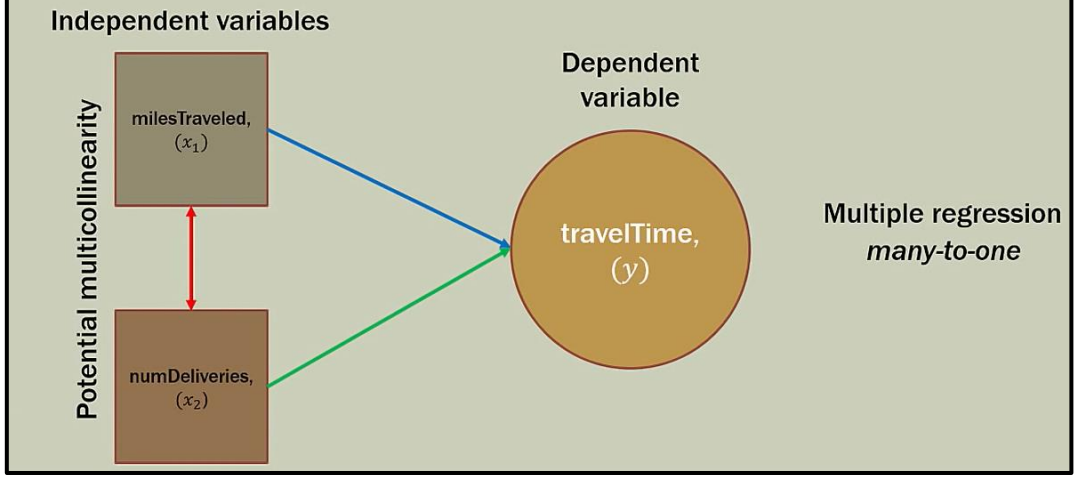
NEW CONSIDERATIONS

- 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.

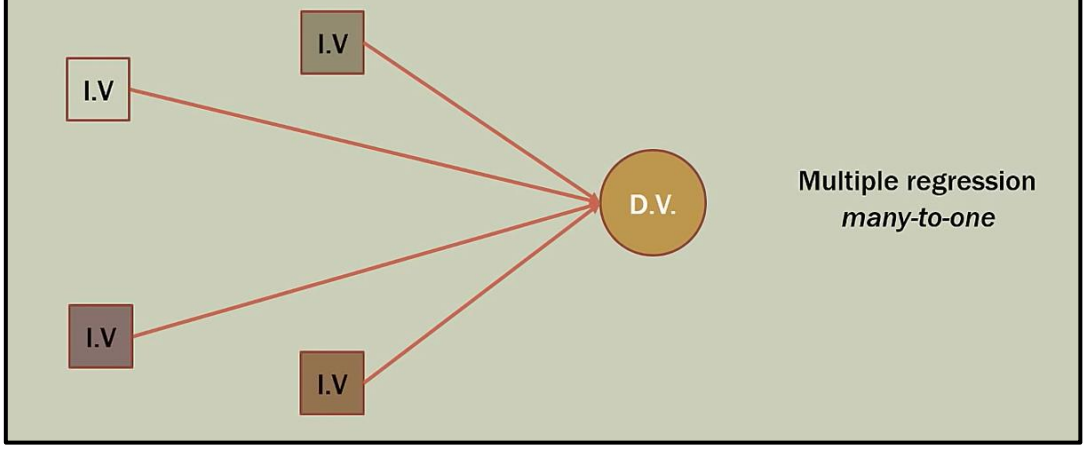
NEW CONSIDERATIONS

- 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

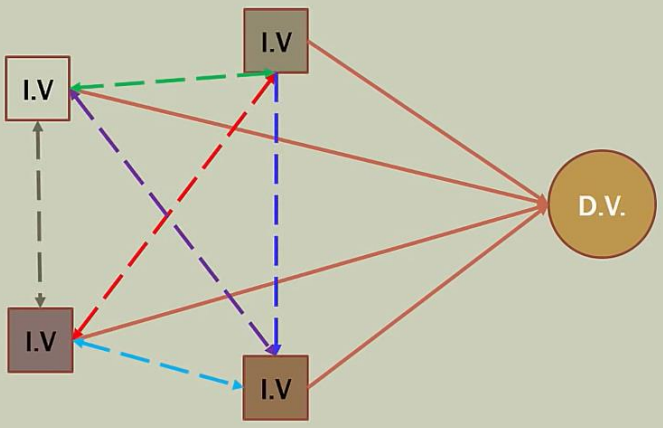
MORE RELATIONSHIPS



MANY RELATIONSHIPS

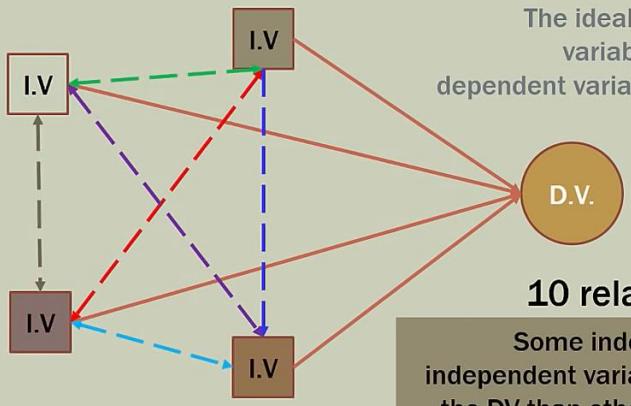


MANY RELATIONSHIPS



Multiple regression
many-to-one

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

$$y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p}_{\text{linear parameters}} + \underbrace{\epsilon}_{\text{error}}$$

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

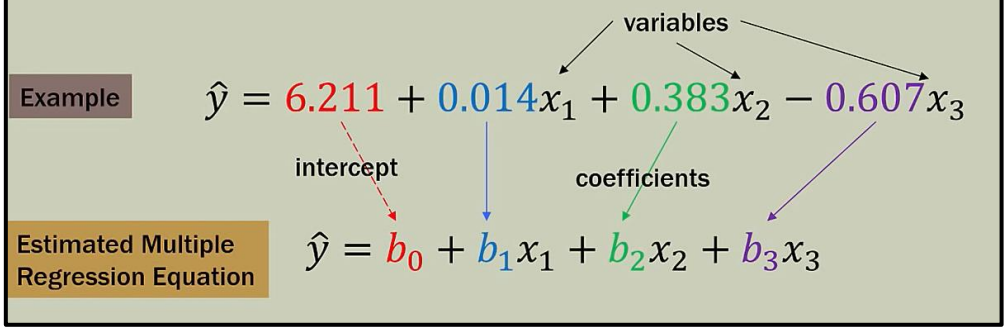
Estimated Multiple Regression Equation

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

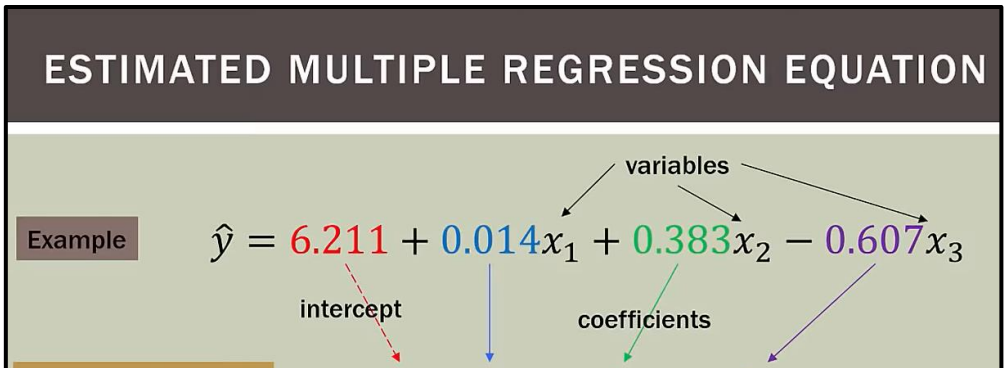
$b_0, b_1, b_2, \ldots b_p$ are the estimates of $\beta_0, \beta_1, \beta_2, \ldots \beta_p$

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

ESTIMATED MULTIPLE REGRESSION EQUATION



ESTIMATED MULTIPLE REGRESSION EQUATION



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.**

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 three factors: 1) the total distance of the trip in miles, 2) the number of deliveries that must be made during the trip, and 3) the daily price of gas/petrol in U.S. dollars.

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:

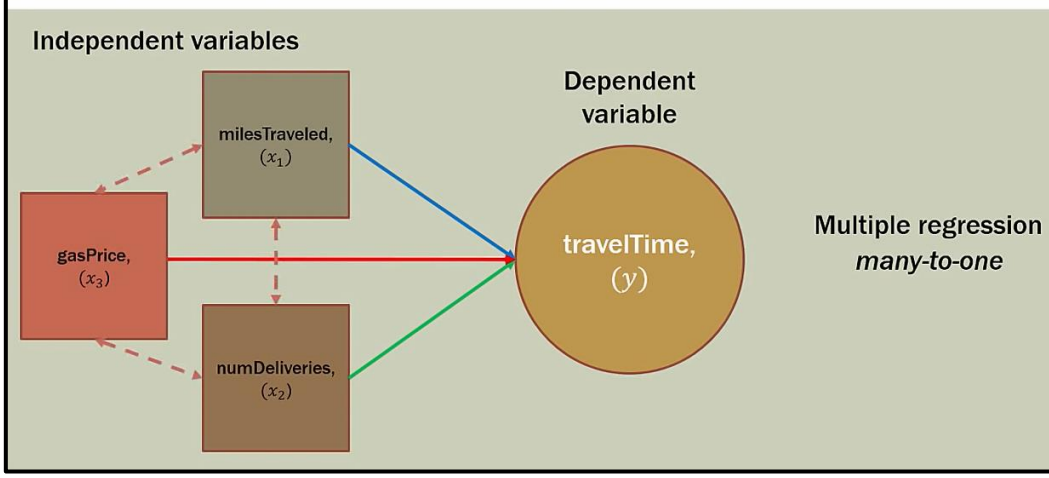
1. Generate a list of potential variables; independent(s) and dependent
2. Collect data on the variables
3. Check the relationships between each independent variable and the dependent variable using scatterplots and correlations
4. Check the relationships among the independent variables using scatterplots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

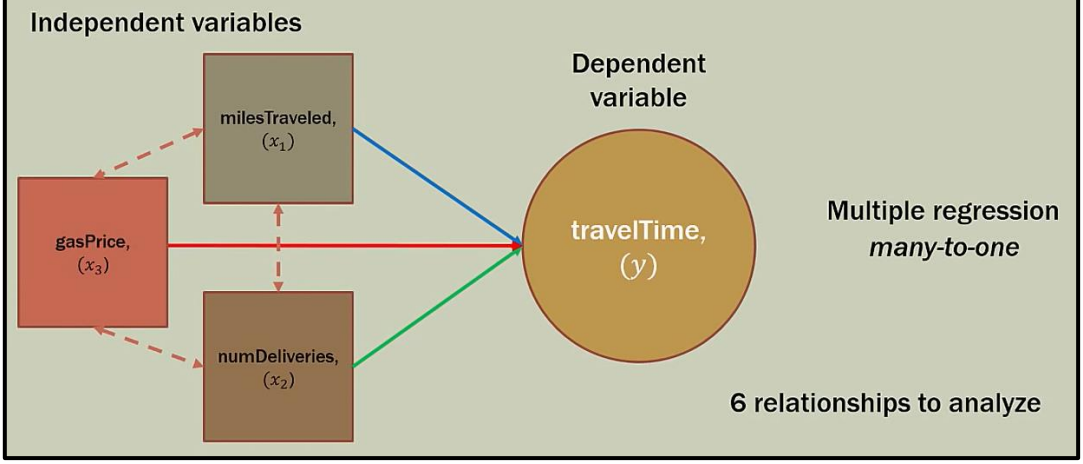
To conduct your analysis you take a random sample of 10 past trips and record four pieces of information for each trip: 1) total miles traveled, 2) number of deliveries, 3) the daily gas price, and 4) total travel time in hours.

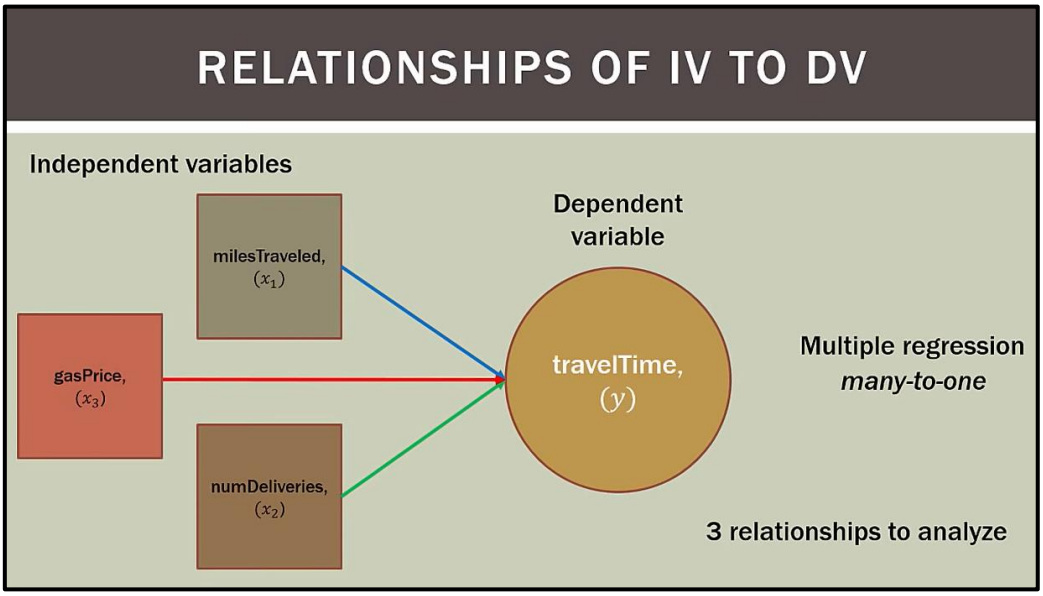
milesTraveled,(x_1)	numDeliveries,(x_2)	gasPrice,(x3)	travelTime(hrs),(y)
89	4	3.84	7
66	1	3.19	5.4
78	3	3.78	6.6
111	6	3.89	7.4
44	1	3.57	4.8
77	3	3.57	6.4
80	3	3.03	7
66	2	3.51	5.6
109	5	3.54	7.3
76	3	3.25	6.4

SKETCHING OUT RELATIONSHIPS

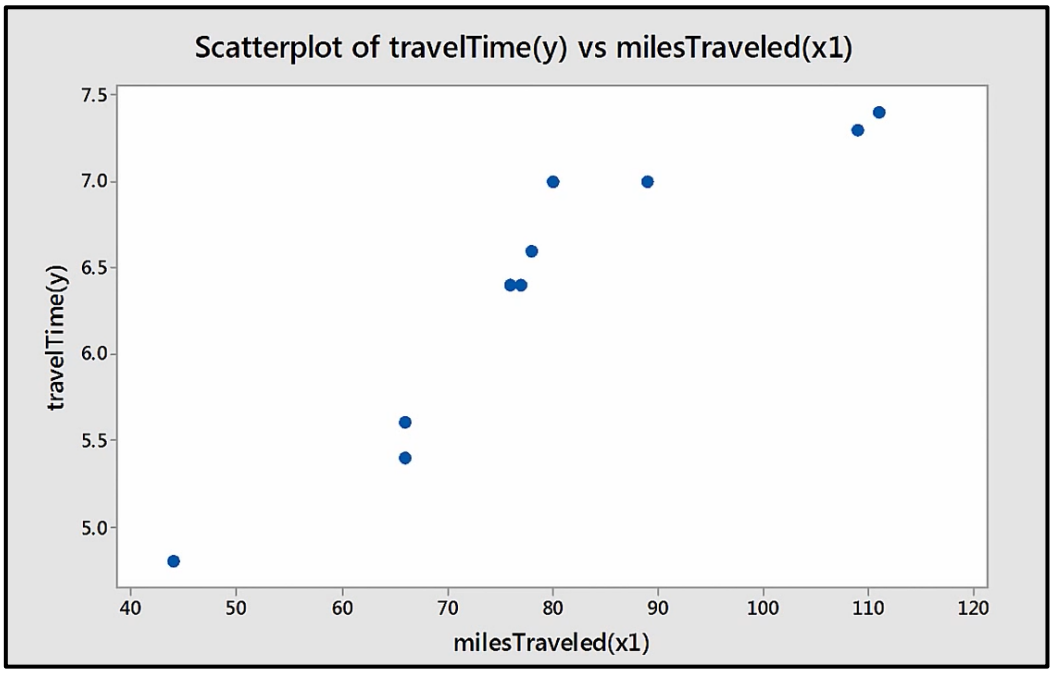


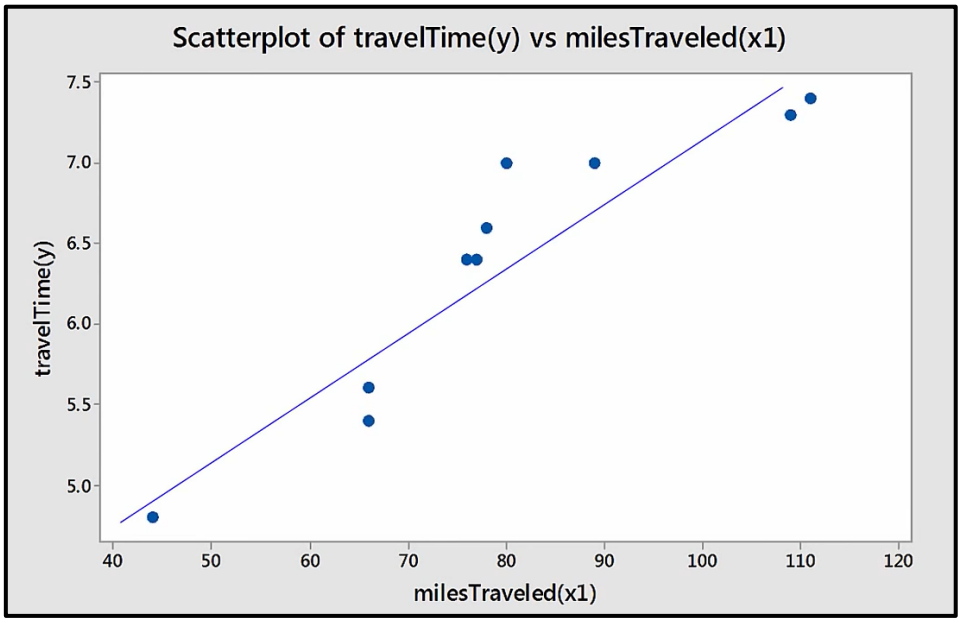
SKETCHING OUT RELATIONSHIPS

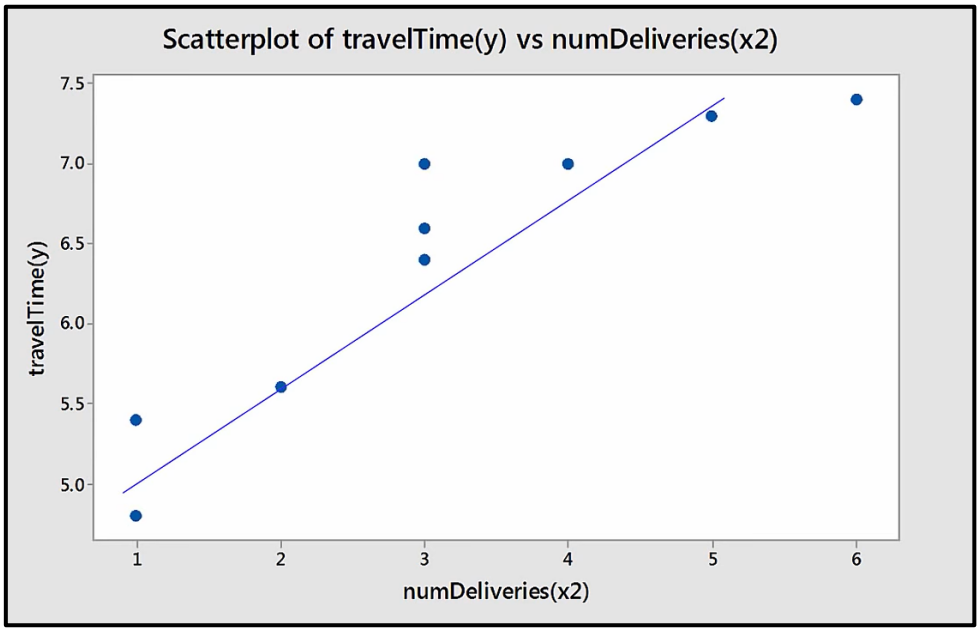


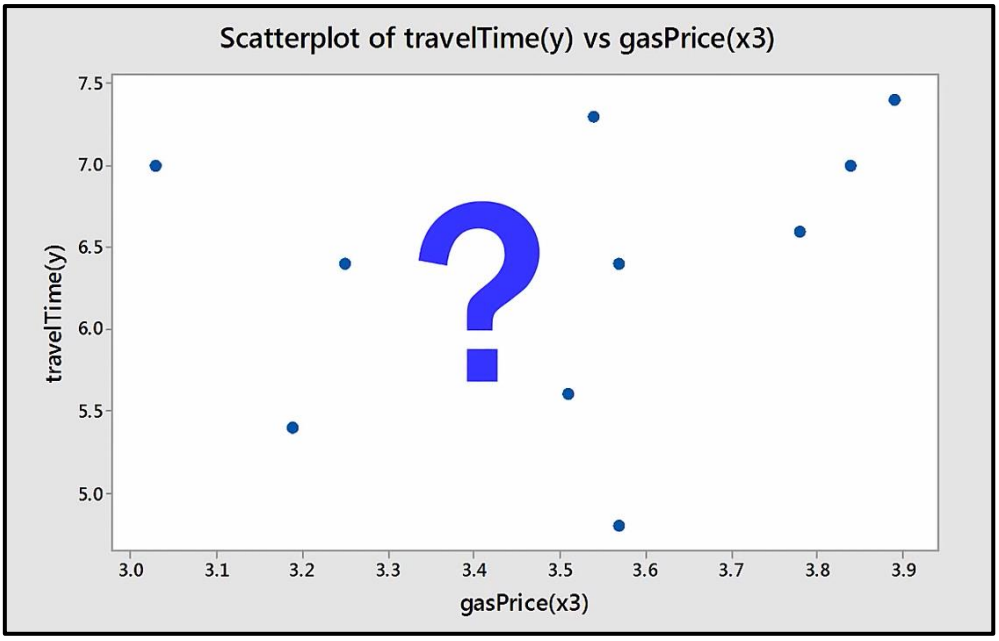


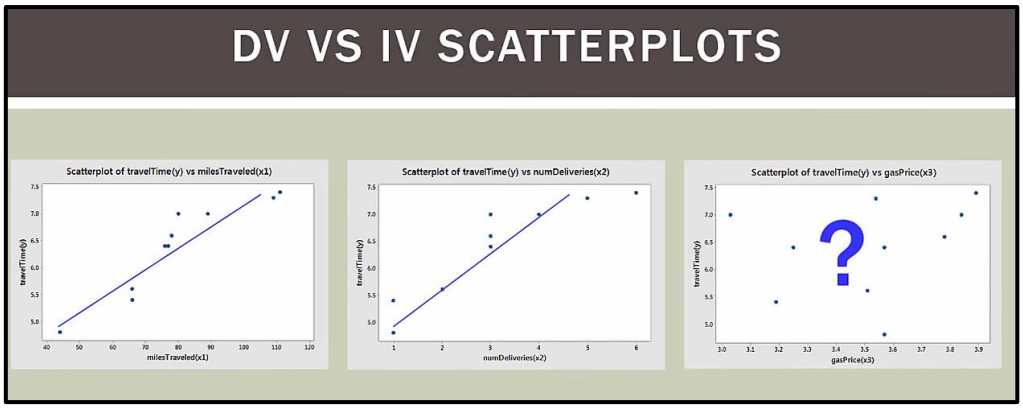
IV TO DV SCATTERPLOTS

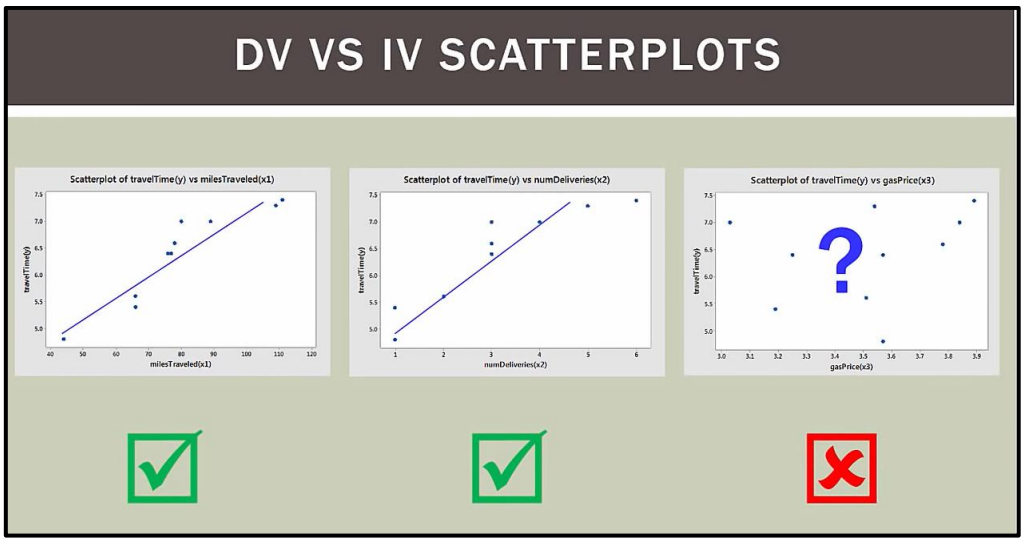



Scatterplot of travelTime(y) vs milesTraveled(x1)
travelTime(y)
7.5
7.0
6.5
6.0
5.5
5.0
40
50
60
70
80
90
100
110
120
milesTraveled(x1)


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


Scatterplot of travelTime(y) vs gasPrice(x3)
travelTime(y)
gasPrice(x3)
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
?

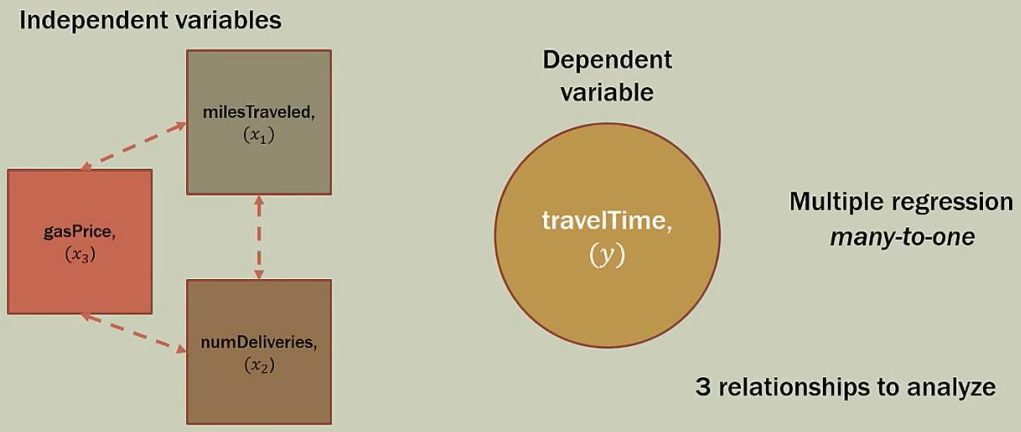

DV VS IV SCATTERPLOTS
Scatterplot of travelTime(y) vs milesTraveled(x1)
Scatterplot of travelTime(y) vs numDeliveries(x2)
Scatterplot of travelTime(y) vs gasPrice(x3)
?


DV VS IV SCATTERPLOTS
Scatterplot of travelTime(y) vs milesTraveled(x1)
Scatterplot of travelTime(y) vs numDeliveries(x2)
Scatterplot of travelTime(y) vs gasPrice(x3)
?

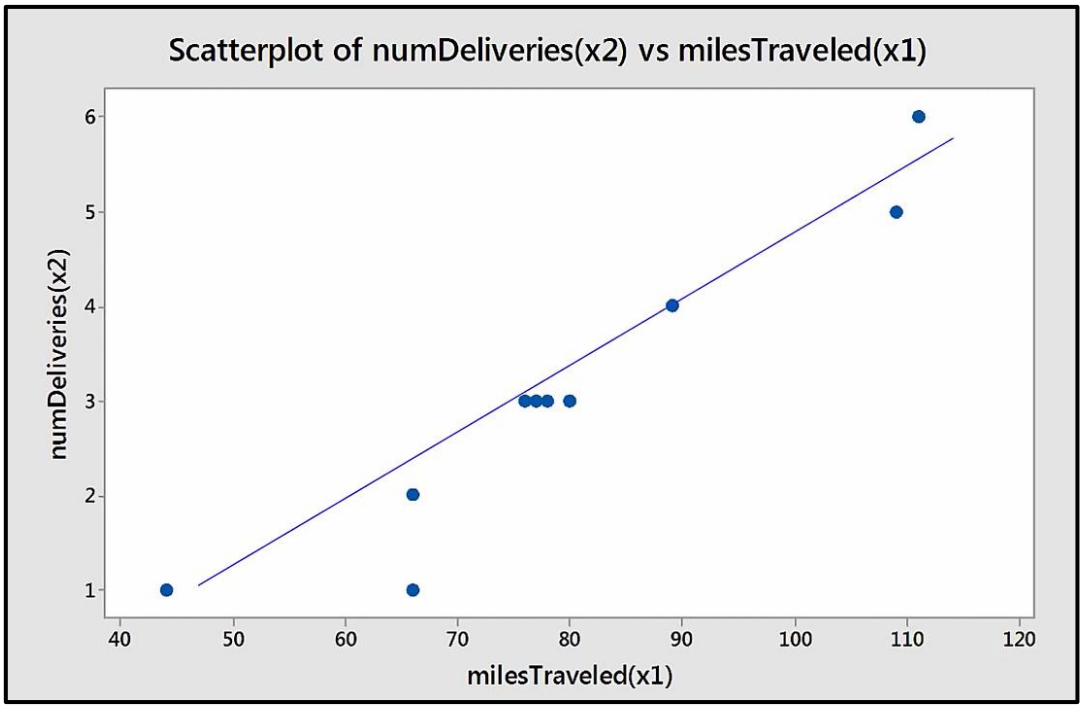
SCATTERPLOT SUMMARY

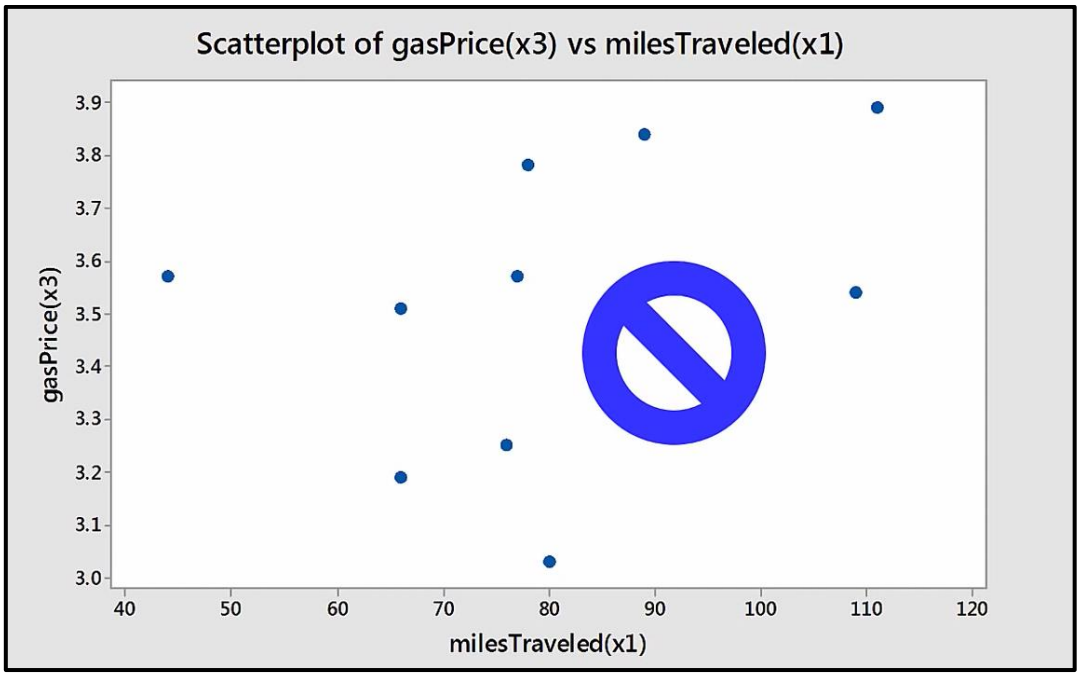
- Dependent variable vs independent variables
  - travelTime(y) appears highly correlated with milesTraveled(x_1)
  - travelTime(y) appears highly correlated with numDeliveries(x_2)
  - travelTime(y) DOES NOT appear highly correlated with gasPrice(x_3)
- Since gasPrice(x_3) does NOT APPEAR CORRELATED with the dependent variable we would NOT use that variable in the multiple regression
- *Note: for now, we will keep gasPrice in and then take it out later for learning purposes*

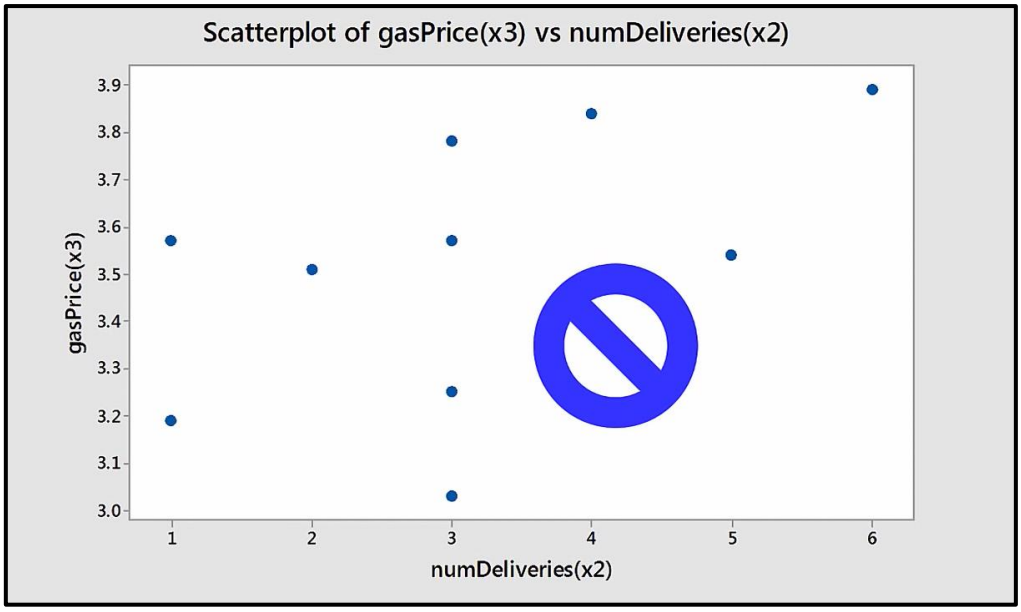
SKETCHING OUT RELATIONSHIPS



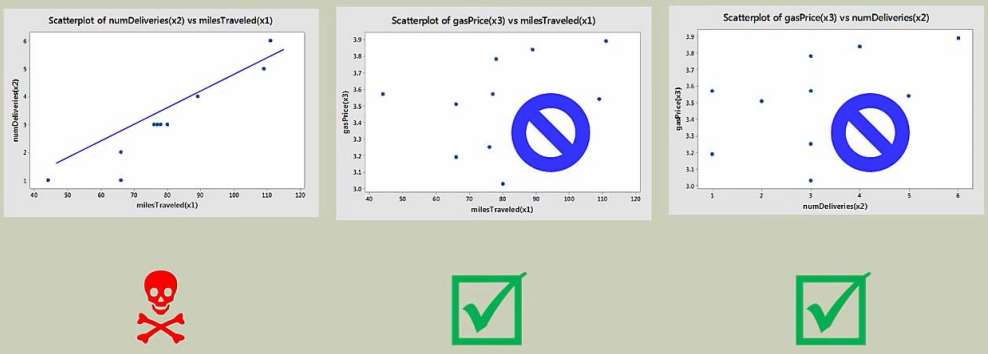
IV TO IV SCATTERPLOTS







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

CORRELATIONS

Correlation: milesTraveled(x1), numDeliveries(x2), gasPrice(x3), travelTime(y)

	milesTraveled(x1)	numDeliveries(x2)	gasPrice(x3)
numDeliveries(x2)	0.956		
	0.000		
gasPrice(x3)	0.356	0.498	
	0.313	*0.143*	
travelTime(y)	0.928	0.916	0.267
	0.000	*0.000*	*0.455*

Cell Contents: Pearson correlation
P-Value

CORRELATIONS

Correlation: milesTraveled(x1), numDeliveries(x2), gasPrice(x3), travelTime(y)

	milesTraveled(x1)	numDeliveries(x2)	gasPrice(x3)
numDeliveries(x2)	0.956		
	0.000		
gasPrice(x3)	0.356	0.498	
	0.313	*0.143*	
travelTime(y)	0.928	0.916	0.267
	0.000	*0.000*	*0.455*

Cell Contents: Pearson correlation
P-Value

CORRELATIONS

Correlation: milesTraveled(x1), numDeliveries(x2), gasPrice(x3), travelTime(y)

	milesTraveled(x1)	numDeliveries(x2)	gasPrice(x3)
numDeliveries(x2)	0.956		
	0.000		
gasPrice(x3)	0.356	0.498	
	0.313	*0.143*	
travelTime(y)	0.928	0.916	0.267
	0.000	*0.000*	*0.455*

Cell Contents: Pearson correlation
P-Value

CORRELATIONS

Correlation: milesTraveled(x1), numDeliveries(x2), gasPrice(x3), travelTime(y)

	milesTraveled(x1)	numDeliveries(x2)	gasPrice(x3)
numDeliveries(x2)	0.956 *0.000*		
gasPrice(x3)	0.356 *0.313*	0.498 *0.143*	
travelTime(y)	0.928 *0.000*	0.916 *0.000*	0.267 *0.455*

Cell Contents: Pearson correlation
P-Value

CORRELATIONS

Correlation: milesTraveled(x1), numDeliveries(x2), gasPrice(x3), travelTime(y)

	milesTraveled(x1)	numDeliveries(x2)	gasPrice(x3)
numDeliveries(x2)	0.956 *0.000*		
gasPrice(x3)	0.356 *0.313*	0.498 *0.143*	
travelTime(y)	0.928 *0.000*	0.916 *0.000*	0.267 *0.455*

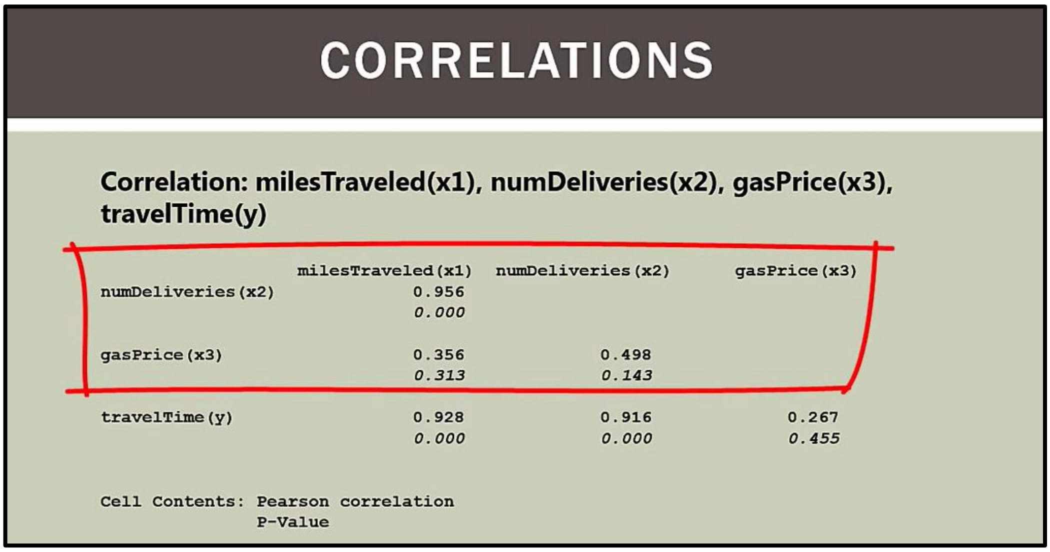
Cell Contents: Pearson correlation
P-Value

CORRELATIONS

Correlation: milesTraveled(x1), numDeliveries(x2), gasPrice(x3), travelTime(y)

	milesTraveled(x1)	numDeliveries(x2)	gasPrice(x3)
numDeliveries(x2)	0.956 *0.000*		
gasPrice(x3)	0.356 *0.313*	0.498 *0.143*	
travelTime(y)	0.928 *0.000*	0.916 *0.000*	0.267 *0.455*

Cell Contents: Pearson correlation
P-Value

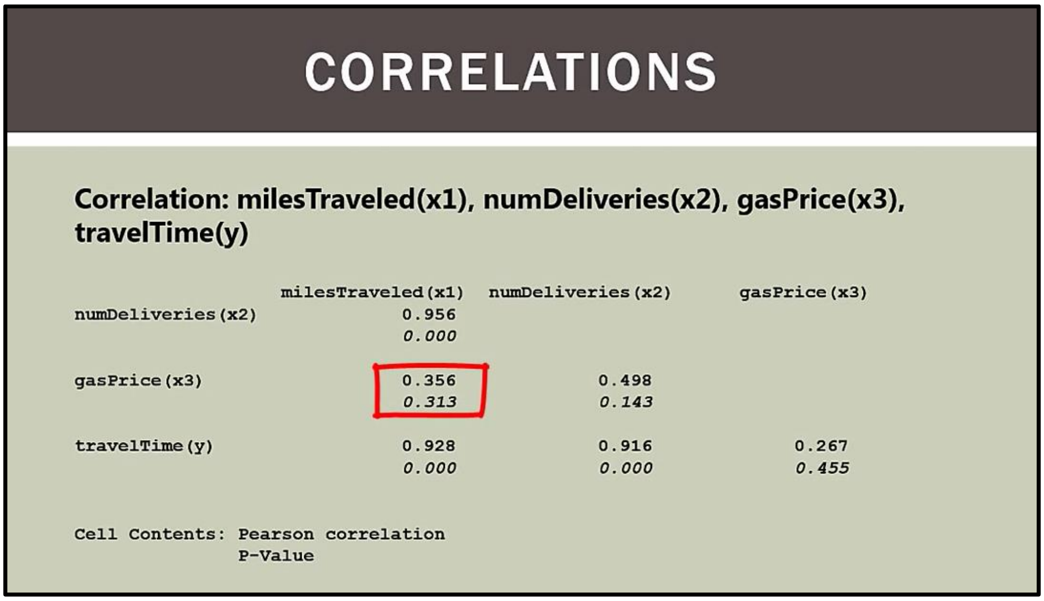


CORRELATIONS

Correlation: milesTraveled(x1), numDeliveries(x2), gasPrice(x3), travelTime(y)

	milesTraveled(x1)	numDeliveries(x2)	gasPrice(x3)
numDeliveries(x2)	0.956 *0.000*		
gasPrice(x3)	0.356 *0.313*	0.498 *0.143*	
travelTime(y)	0.928 *0.000*	0.916 *0.000*	0.267 *0.455*

Cell Contents: Pearson correlation
P-Value

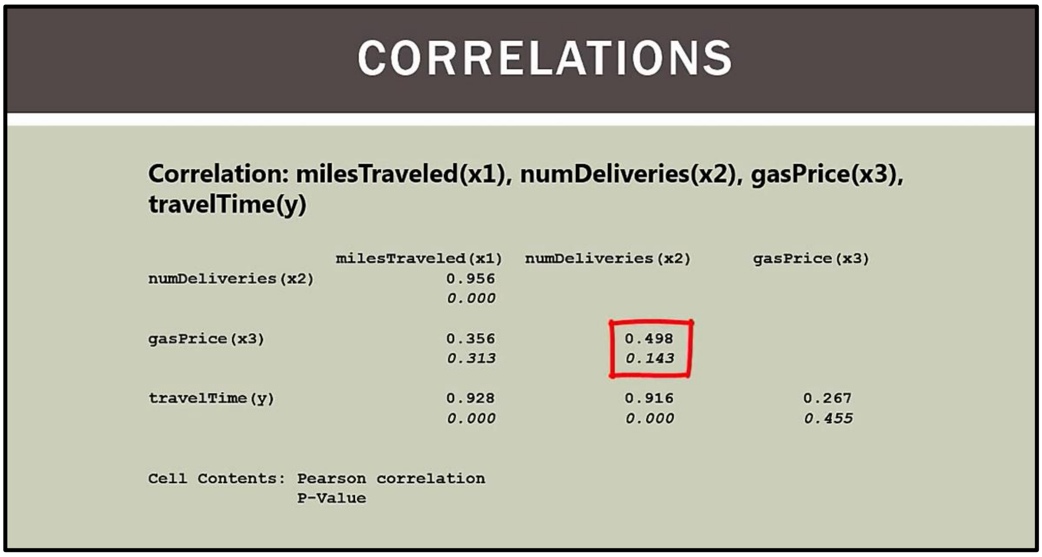


CORRELATIONS

Correlation: milesTraveled(x1), numDeliveries(x2), gasPrice(x3), travelTime(y)

	milesTraveled(x1)	numDeliveries(x2)	gasPrice(x3)
numDeliveries(x2)	0.956 *0.000*		
gasPrice(x3)	0.356 *0.313*	0.498 *0.143*	
travelTime(y)	0.928 *0.000*	0.916 *0.000*	0.267 *0.455*

Cell Contents: Pearson correlation
P-Value



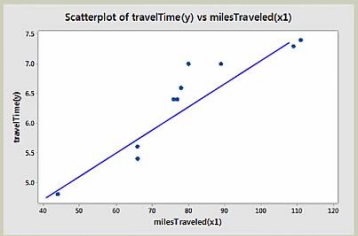
CORRELATIONS

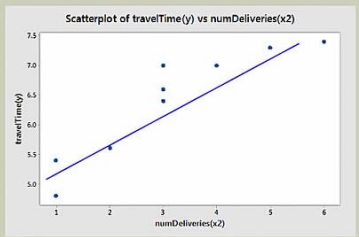
Correlation: milesTraveled(x1), numDeliveries(x2), gasPrice(x3), travelTime(y)

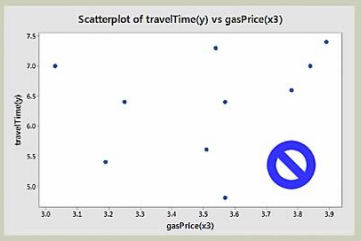
	milesTraveled(x1)	numDeliveries(x2)	gasPrice(x3)
numDeliveries(x2)	0.956 *0.000*		
gasPrice(x3)	0.356 *0.313*	0.498 *0.143*	
travelTime(y)	0.928 *0.000*	0.916 *0.000*	0.267 *0.455*

Cell Contents: Pearson correlation
P-Value

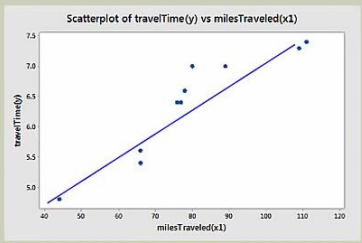
DV VS IV SCATTERPLOTS

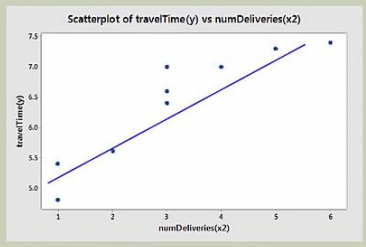


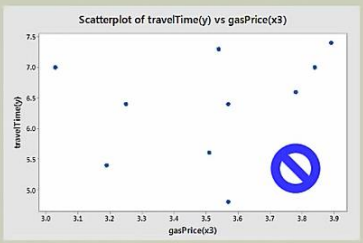




DV VS IV SCATTERPLOTS





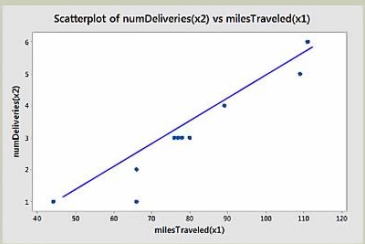


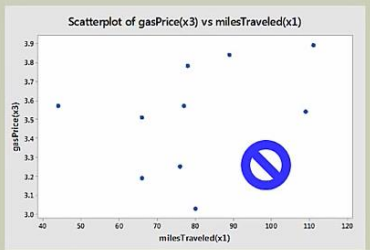
$r = 0.928$
p value=.000

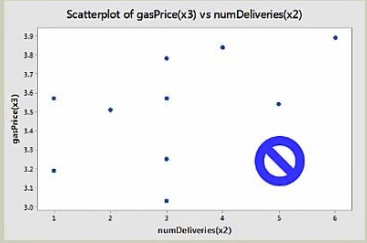
$r = 0.916$
p value=.000

$r = 0.267$
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

REVIEW AND CONCLUSION

- In multiple regression, a lot of prep work must be done before ever clicking the "Run" button in your software.
- Do not blindly mash buttons in stats software!
- Techniques:
  - Scatterplots
  - Correlation analysis
  - Individual / group regressions

NEXT STEPS

- For the sake of learning, we are going to break the rules and include all three independent variables in the regression at first
- Then we will remove the problematic independent variables as we should and then watch what happens to the regression results
- We will also perform simple regressions with the dependent variable to use as a baseline (again for the sake of learning)
- In the end, we will end up with the best model

MULTIPLE REGRESSION:
PART 3: EVALUATING BASIC MODELS

Brandon Foltz, M.Ed.
education / statistics / business / tech / math / opinion
http://www.bcfoltz.com/blog
Twitter: @BCFoltz
YouTube: BCFoltz

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.

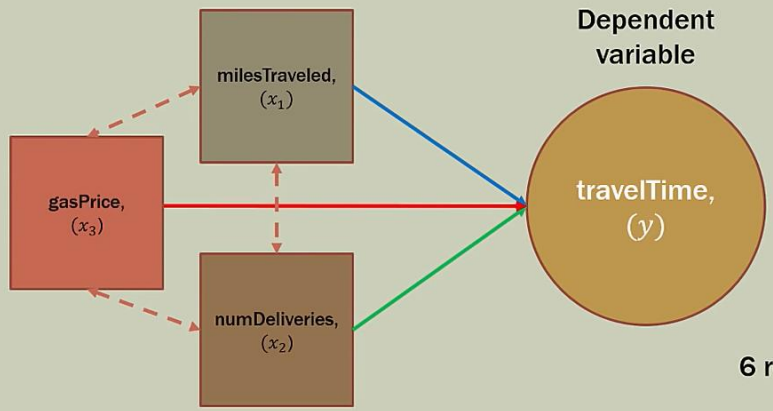
RDS DATA AND VARIABLE NAMING

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

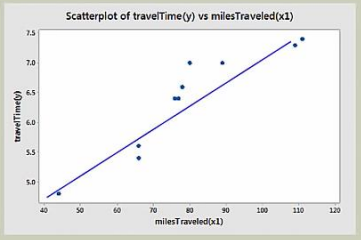
Independent variables



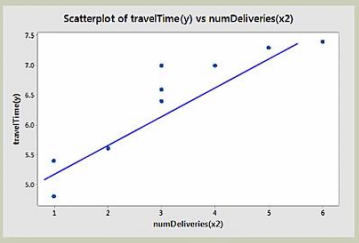
Multiple regression
many-to-one

6 relationships to analyze

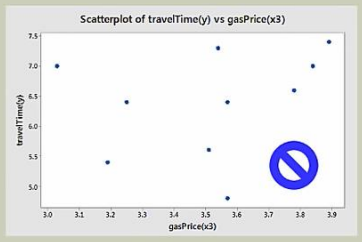
DV VS IV SCATTERPLOTS



$r = 0.928$
p value=.000

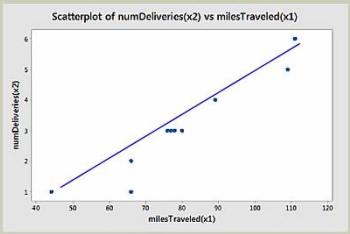


$r = 0.916$
p value=.000

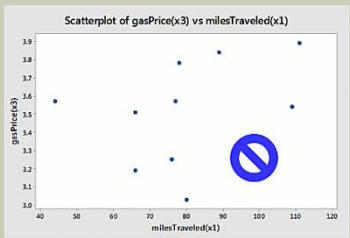


$r = 0.267$
p value=.455

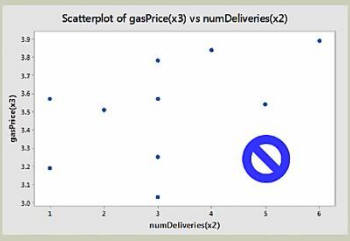
IV SCATTERPLOTS (MULTICOLLINEARITY)



$r = 0.956$
p value=.000



$r = 0.356$
p value=.313



$r = 0.498$
p value=.143

URLS:

https://pythonspot.com/matplotlib-scatterplot/#:~:text=Matplot%20has%20a%20built%2Din,the%20horizontal%20or%20vertical%20dimension.

Gradient Descent Step-by-Step (Numerical Problem):

https://www.youtube.com/watch?v=sDv4f4s2SB8

Stochastic Gradient Descent: https://www.youtube.com/watch?v=vMh0zPT0tLI

Linear Regression Algorithm | Linear Regression in Python | Machine Learning Algorithm | Edureka

https://www.youtube.com/watch?v=E5RjzSK0fvY

1) http://faculty.marshall.usc.edu/gareth-james/ISL/
2) https://www.analyzemath.com/statistics/linear_regression.html
3) https://medium.com/analytics-vidhya/how-multicollinearity-is-a-problem-in-linear-regression-dbb76e25cd80
4) https://www.dummies.com/education/math/statistics/how-to-find-right-tail-values-and-confidence-intervals-using-the-t-table/
5) https://datascienceplus.com/analytical-and-numerical-solutions-to-linear-regression-problems/
6) https://www.investopedia.com/terms/r/regression.asp
7) https://www.dummies.com/education/math/statistics/how-to-determine-the-confidence-interval-for-a-population-proportion/

http://cs229.stanford.edu/livenotes2020spring/cs229-livenotes-lecture2.pdf

http://cs229.stanford.edu/notes2020spring/lecture1_slide.pdf

http://www.holehouse.org/mlclass/

https://www.coursera.org/learn/machine-learning#syllabus

https://www.youtube.com/watch?v=wPJ1_Z8b0wk

https://machinelearningmedium.com/2017/08/11/cost-function-of-linear-regression/

https://towardsdatascience.com/machine-learning-fundamentals-via-linear-regression-1a5d11f5220

https://medium.com/@lachlanmiller_52885/machine-learning-week-1-cost-function-gradient-descent-and-univariate-linear-regression-8f5fe69815fd

https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a