Dr. Khaing S Htun

Multiple Linear Regression Analysis

Multiple Linear Regression Model

Why multiple linear regression model?

Simple linear regression is a model that assesses the relationship between a dependent variable and an independent variable. The simple linear model is expressed using the following equation:

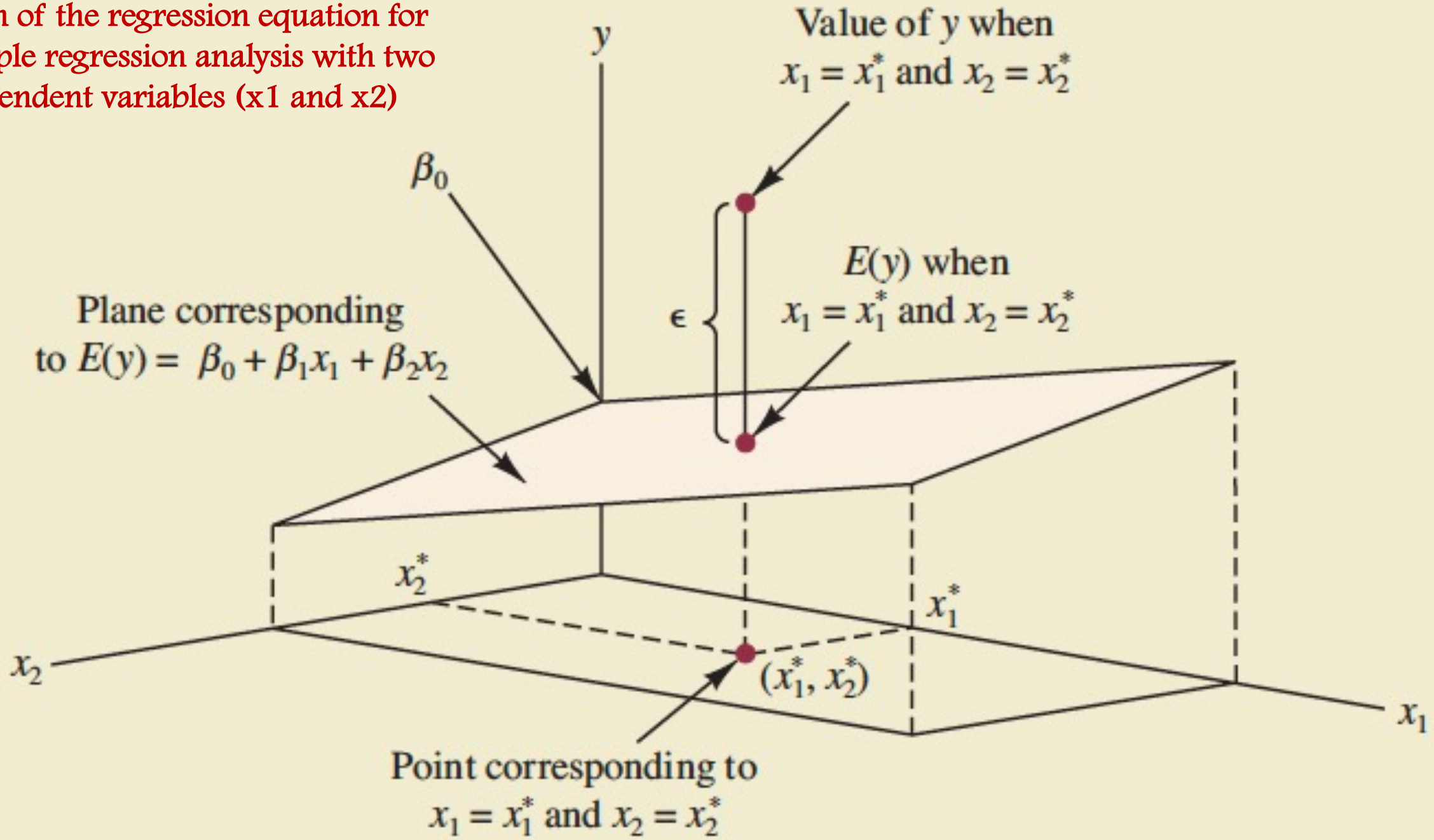
$$y = \beta_0 + \beta_1 x + \varepsilon$$

Multiple Linear Regression Model

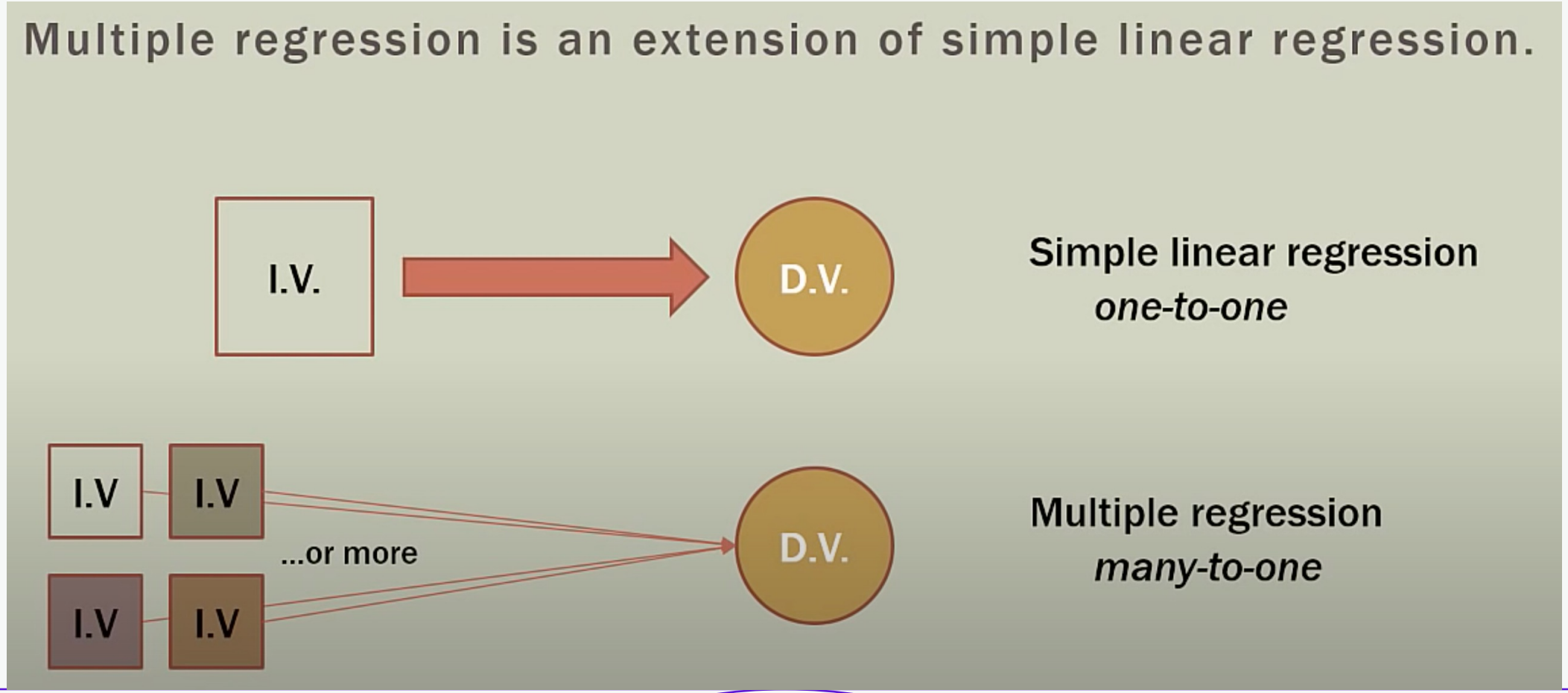
Multiple linear regression analysis is essentially similar to the simple linear model, with the exception that multiple independent variables are used in the model. The mathematical representation of multiple linear regression is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

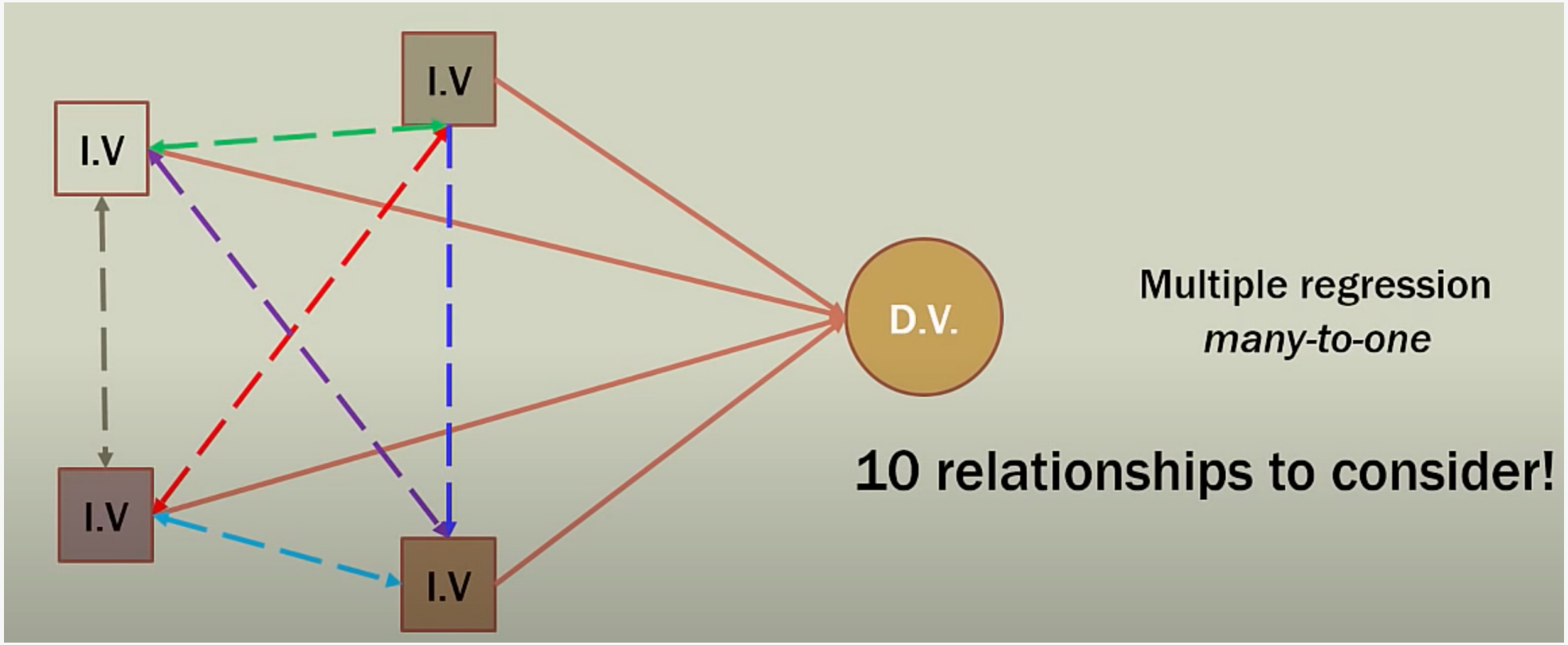
Graph of the regression equation for multiple regression analysis with two independent variables (x1 and x2)



Multiple Linear Regression Model



Multiple Linear Regression Model



Multiple Linear Regression Model

Multiple linear regression model :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Multiple linear regression equation:

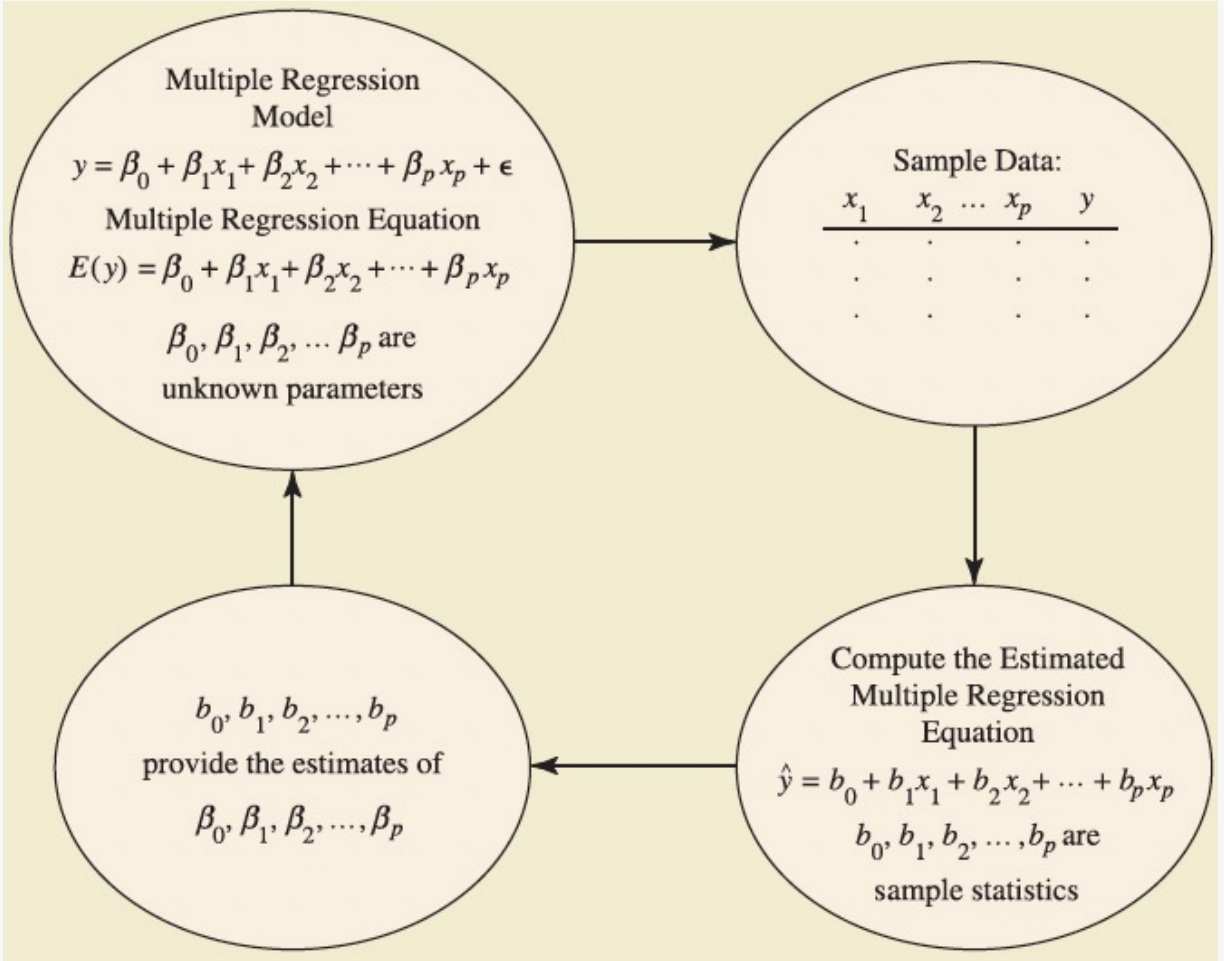
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Estimated multiple linear regression equation :

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

where $b_0, b_1, b_2, \cdots, b_k$ are the least-squares estimators of the parameters $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$

The estimation process for multiple regression



Least Squares Method

Non-collinearity

Multiple linear regression follows the same conditions as the simple linear model. However, since there are several independent variables in multiple linear analysis, there is another mandatory condition for the model:

Non-collinearity

Independent variables ($\mathbf{x_1, x_2, \dots, x_k}$) should show a **minimum correlation** with each other.

If the independent variables are highly correlated with each other, it will be difficult to assess the true relationships between the dependent and independent variables.

Prep-work for Model Building

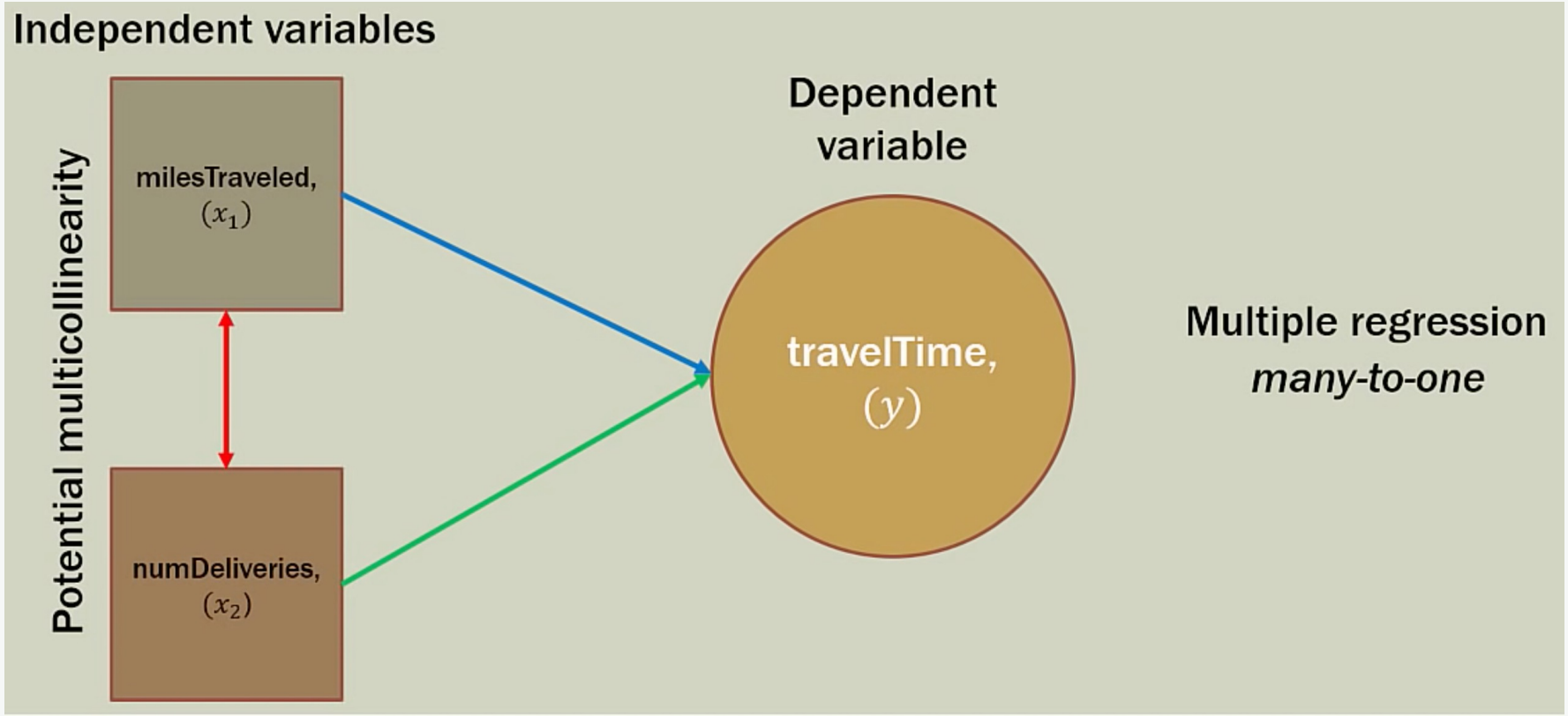
Beacause of **multicollinearity and overfitting**, do these, before conducting multiple regression analysis.

1. Correlations,
2. Scatter plots,
3. Simple regressions

Example: Multiple Linear Regression Model

Driving Assignment	Travel time (hours) (y)	Miles Traveled(x1)	Number of Deliveries(x2)
1	7	89	4
2	5.4	66	1
3	6.6	78	3
4	7.4	111	6
5	4.8	44	1
6	6.4	77	3
7	7	80	3
8	5.6	66	2
9	7.3	109	5
10	6.4	76	3

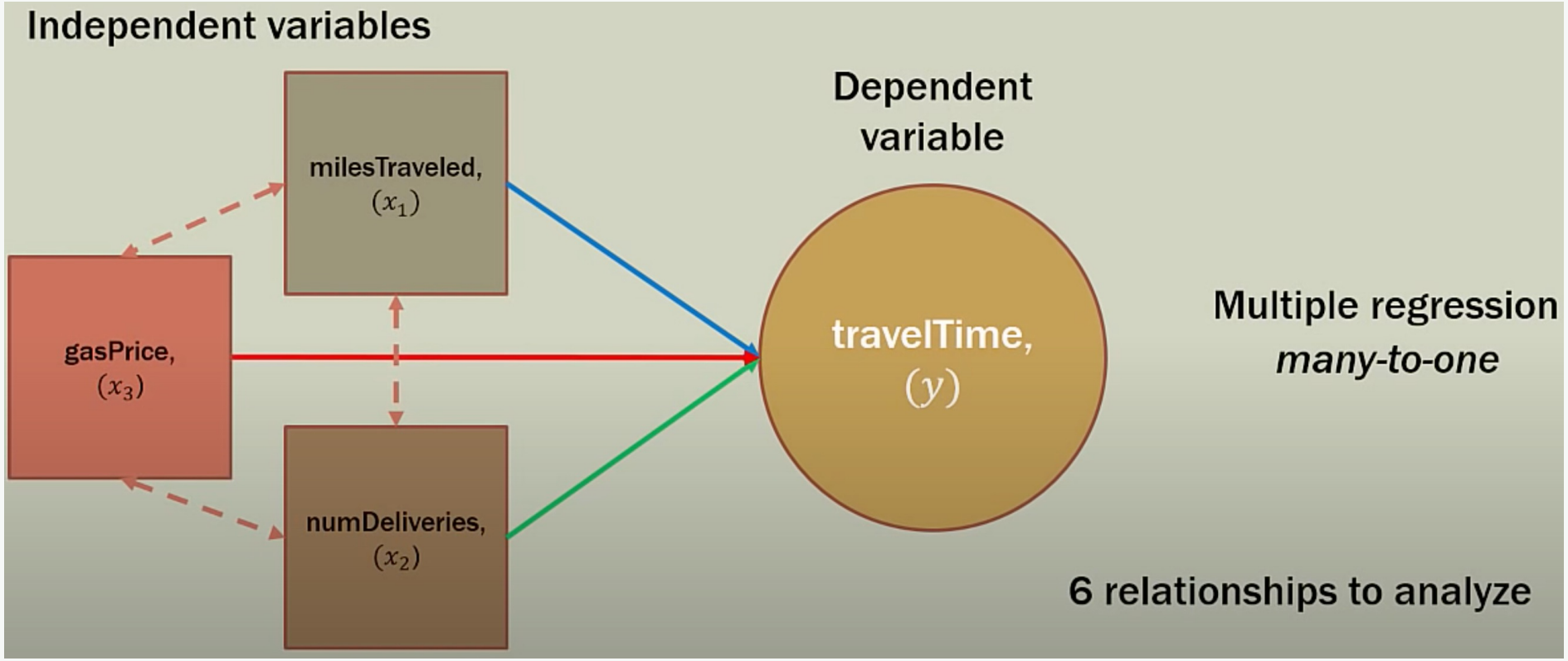
Example


Independent variables
Dependent variable
Potential multicollinearity
milesTraveled, (x_1)
numDeliveries, (x_2)
travelTime, (y)
Multiple regression
many-to-one

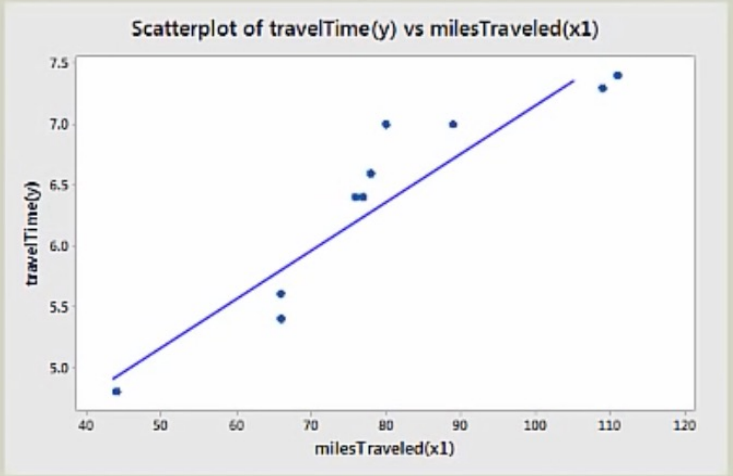
Example

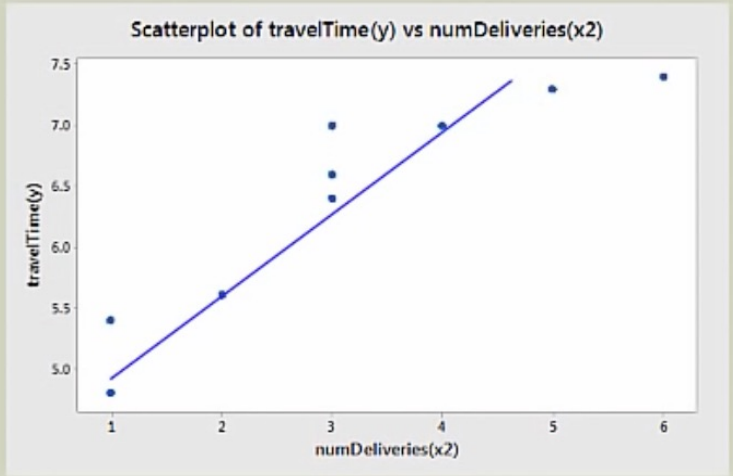
Driving Assignment	Travel time (hours) (y)	Miles Traveled(x1)	Number of Deliveries(x2)	Gas price (x3)
1	7	89	4	3.84
2	5.4	66	1	3.19
3	6.6	78	3	3.78
4	7.4	111	6	3.89
5	4.8	44	1	3.57
6	6.4	77	3	3.57
7	7	80	3	3.03
8	5.6	66	2	3.51
9	7.3	109	5	3.54
10	6.4	76	3	3.25

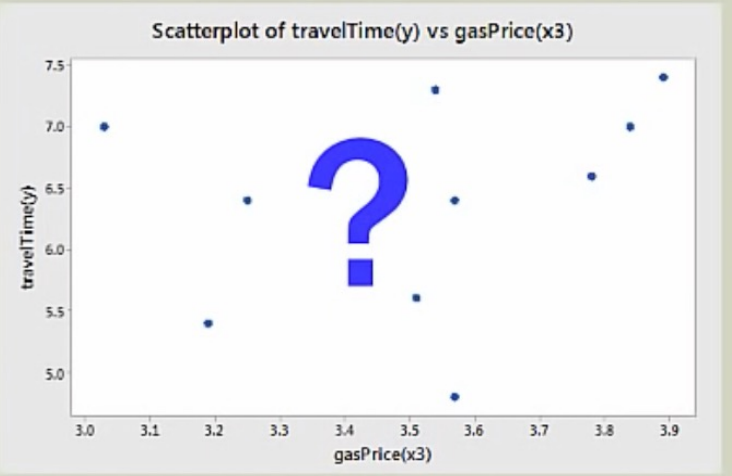
Example


Independent variables
milesTraveled, (x_1)
gasPrice, (x_3)
numDeliveries, (x_2)
Dependent variable
travelTime, (y)
Multiple regression
many-to-one
6 relationships to analyze

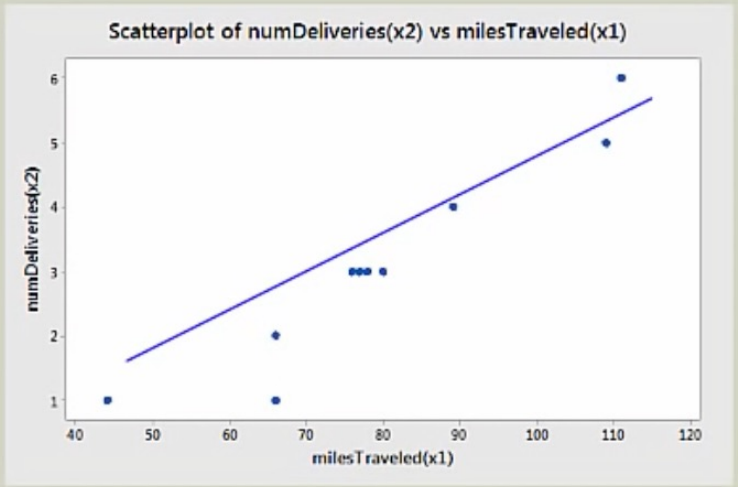
Example: relevancy check DV and IVs

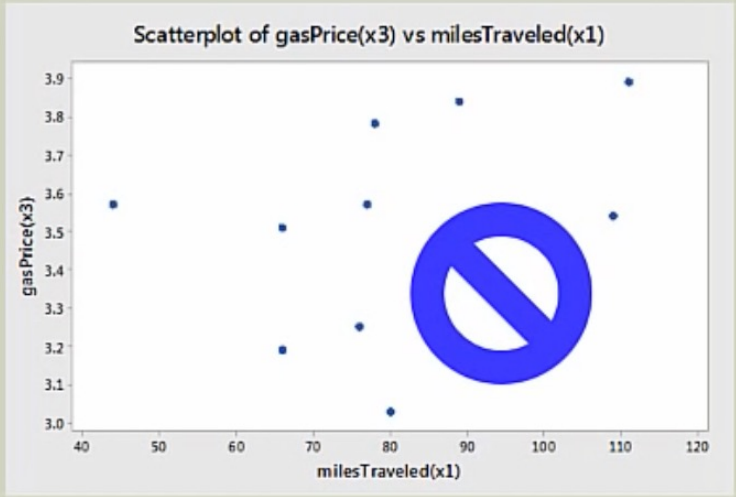


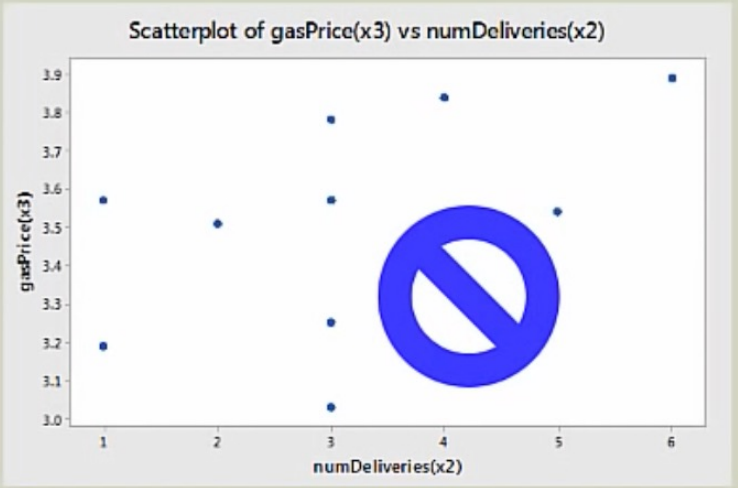




Example: multicollinearity check IVs







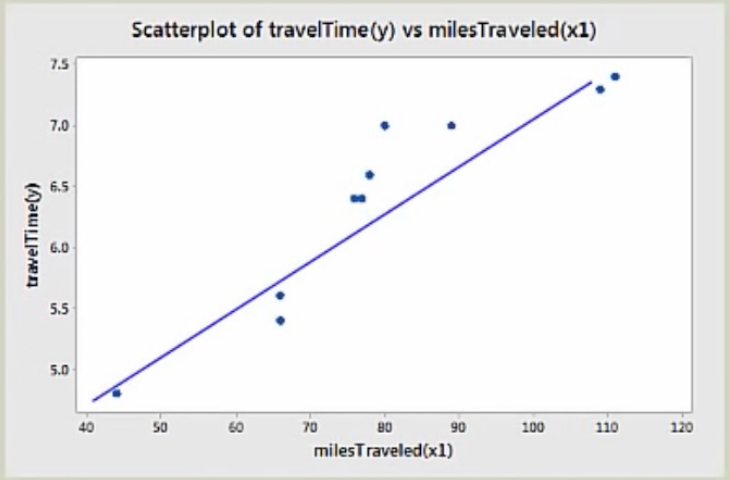
Example: correlation, DV and IVs

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*

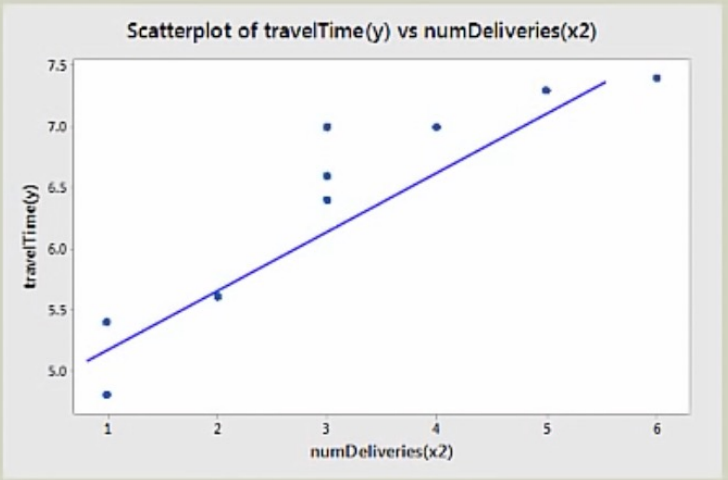
Cell Contents: Pearson correlation
P-Value

Example



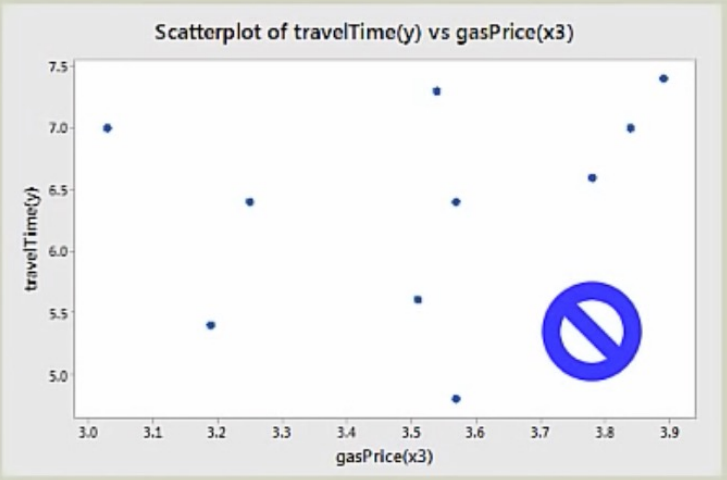
$r = 0.928$

p value=.000



$r = 0.916$

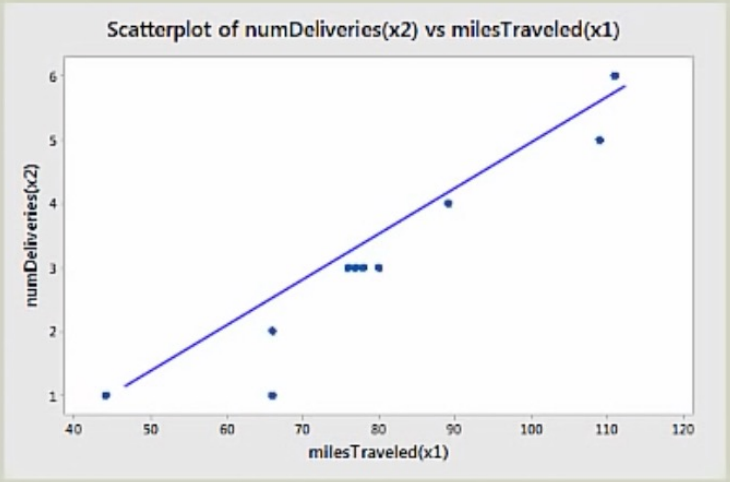
p value=.000



$r = 0.267$

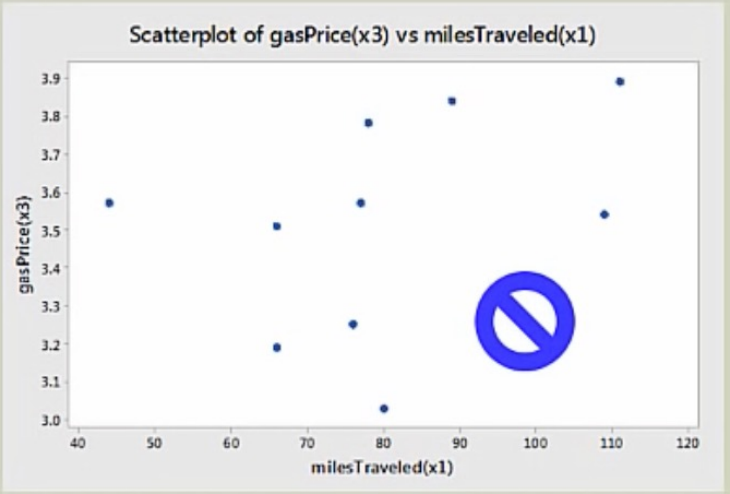
p value=.455

Example



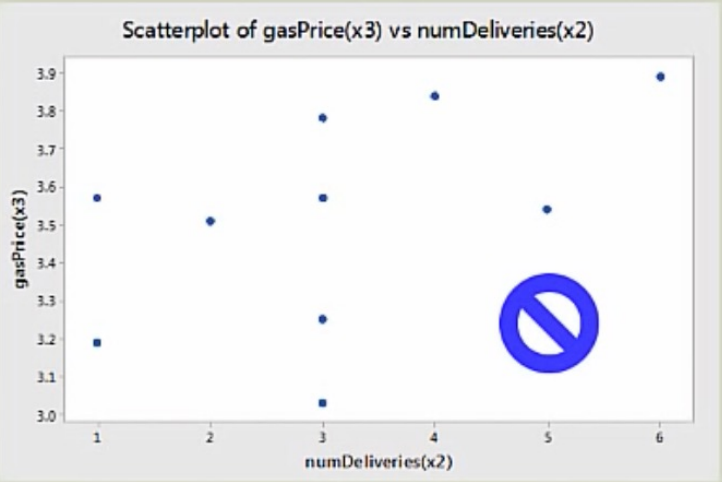
$r = 0.956$

p value=.000



$r = 0.356$

p value=.313



$r = 0.498$

p value=.143

Example

- **Correlation analysis** confirms the conclusions reached by visual examination of the scatterplots
- **Redundant multicollinear variables**
  - Miles traveled (x1) and number of deliveries (x2) are both **highly correlated** with each other and therefore are redundant; only one should used in the multiple regression analysis
- **Non-contributing variables**
  - Gas price is **NOT correlated** with the depended variables and should be excluded

Example

Technique:

- Scatter plots
- Correlation analysis
- **Individual/group** regression

Example

- Perform simple regression for each IV individually (conduct with excel)
- Interpretation of result
- Take note of how results change:
  - Coefficients
    - Values, t-statistic, p-value
  - Analysis of Variance (ANOVA)
    - F-value, p-value (significance F)
  - R-squared, adjusted R-squared, predicted R-squared
  - VIR (Variance Inflation Factor)
  - Mallows $C_p = \frac{RSS_k}{S^2} - n + 2(k+1)$

RSS_k: The residual sum of squares for a model with *p* predictor variables
S^2: The residual mean square for the model (estimated by MSE)
n: The sample size
k: The number of predictor variables

Example

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.928178501
R Square	0.86151533
Adjusted R Square	0.844204746
Standard Error	0.34230884
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	5.831597265	5.831597265	49.76812677	0.000106676
Residual	8	0.937402735	0.117175342		
Total	9	6.769			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3.185560249	0.466950794	6.822046973	0.000134762	2.108769788	4.2623507
milesTraveled	0.040256781	0.005706416	7.054652845	0.000106676	0.027097763	0.05341579

$$\hat{Y} = 3.1856 + 0.0403x_1$$

An increase in 1 mile will increase delivery time by .0403 hours.

84 mile trip estimate

$\hat{y} = 3.1856 + 0.0403(84)$

$\hat{y} = 6.5708$ hours (6:34)

$\hat{y} = 6.570$

$\hat{y} = 5.776$

$\hat{y} = 5{:}47$ t

df	0.10	0.05	0.025	0.01
2	2.9200	4.3027	6.2054	9.9250
3	2.3534	3.1824	4.1765	5.8408
4	2.1318	2.7765	3.4954	4.6041
5	2.0150	2.5706	3.1634	4.0321
6	1.9432	2.4469	2.9687	3.7074
7	1.8946	2.3646	2.8412	3.4995
8	1.8595	2.3060	2.7515	3.3554
9	1.8331	2.2622	2.6850	3.2498
10	1.8125	2.2281	2.6338	3.1693
11	1.7959	2.2010	2.5931	3.1058
12	1.7823	2.1788	2.5600	3.0545
13	1.7709	2.1604	2.5326	3.0123
14	1.7613	2.1448	2.5096	2.9768
15	1.7531	2.1315	2.4899	2.9467
16	1.7459	2.1199	2.4729	2.9208

Example

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	5.6851	5.6851	41.96	0.000
numDeliveries(x2)	1	5.6851	5.6851	41.96	0.000
Error	8	1.0839	0.1355		
Lack-of-Fit	4	0.6639	0.1660	1.58	0.334
Pure Error	4	0.4200	0.1050		
Total	9	6.7690			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.368091	83.99%	81.99%	70.27%

4 delivery estimate

$$\hat{y} = 4.845 + 0.4983(4)$$

$$\hat{y} = 6.838 \text{ hours } (6{:}50)$$

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4.845	0.265	18.26	0.000	
numDeliveries(x2)	0.4983	0.0769	6.48	0.000	1.00

Regression Equation

travelTime(y) = 4.845 + 0.4983 numDeliveries(x2)

An increase in 1 delivery will increase delivery time by .4983 hours.

Example

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	0.4833	0.4833	0.62	0.455
gasPrice(x3)	1	0.4833	0.4833	0.62	0.455
Error	8	6.2857	0.7857		
Lack-of-Fit	7	5.0057	0.7151	0.56	0.777
Pure Error	1	1.2800	1.2800		
Total	9	6.7690			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.886403	7.14%	0.00%	0.00%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.54	3.65	0.97	0.361	
gasPrice(x3)	0.81	1.03	0.78	0.455	1.00

Regression Equation

travelTime(y) = 3.54 + 0.81 gasPrice(x3)

Example

Model options summary

F	p—value	S	$R^2(adj)$	$R^2(pred)$	x_1	x_2	x_3
49.77	< 0.001	0.34230	84.42%	79.07%	X		
41.96	< 0.001	0.36809	81.99%	70.27%		X	
0.62	0.455	0.88640	0.00%	0.00%			X

Example

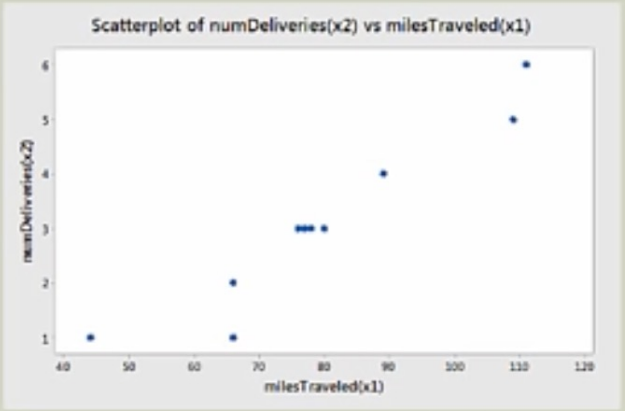
Regression Analysis: travelTime(y) versus milesTraveled(x1), numDeliveries(x2)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	5.89850	2.94925	23.72	0.001
milesTraveled(x1)	1	0.21343	0.21343	1.72	0.232
numDeliveries(x2)	1	0.06691	0.06691	0.54	0.487
Error	7	0.87050	0.12436		
Total	9	6.76900			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.352642	87.14%	83.47%	59.95%



$r = 0.956$

p value=.000

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.732	0.887	4.21	0.004	
milesTraveled(x1)	0.0262	0.0200	1.31	0.232	11.59
numDeliveries(x2)	0.184	0.251	0.73	0.487	11.59

Regression Equation

travelTime(y) = 3.732 + 0.0262 milesTraveled(x1) + 0.184 numDeliveries(x2)

VIF
Variance
Inflation
Factor

Example

Regression Analysis: travelTime(y) versus milesTraveled(x1), gasPrice(x3)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	5.86239	2.93119	22.63	0.001
milesTraveled(x1)	1	5.37907	5.37907	41.53	0.000
gasPrice(x3)	1	0.03079	0.03079	0.24	0.641
Error	7	0.90661	0.12952		
Total	9	6.76900			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.359883	86.61%	82.78%	68.11%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.87	1.48	2.61	0.035	
milesTraveled(x1)	0.04137	0.00642	6.44	0.000	1.14
gasPrice(x3)	-0.219	0.449	-0.49	0.641	1.14

Regression Equation

travelTime(y) = 3.87 + 0.04137 milesTraveled(x1) - 0.219 gasPrice(x3)

If gasPrice is held constant, then travelTime is expected to increase by 0.04137 hours for each additional mile traveled.

If milesTraveled is held constant, then travelTime is expected to decrease by 0.219 hours for each additional dollar increase in gasPrice.

Example

Regression Analysis: travelTime(y) versus numDeliveries(x2), gasPrice(x3)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	6.0081	3.0040	27.63	0.000
numDeliveries(x2)	1	5.5248	5.5248	50.82	0.000
gasPrice(x3)	1	0.3230	0.3230	2.97	0.128
Error	7	0.7609	0.1087		
Total	9	6.7690			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.329703	88.76%	85.55%	71.76%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	7.32	1.46	5.03	0.002	
numDeliveries(x2)	0.5665	0.0795	7.13	0.000	1.33
gasPrice(x3)	-0.765	0.444	-1.72	0.128	1.33

Regression Equation

travelTime(y) = 7.32 + 0.5665 numDeliveries(x2) - 0.765 gasPrice(x3)

If gasPrice is held constant, then travelTime is expected to increase by 0.5665 hours for each additional delivery.

If numDeliveries is held constant, then travelTime is expected to decrease by 0.765 hours for each additional dollar increase in gasPrice.

Example

F	p–value	S	$R^2(adj)$	$R^2(pred)$	x_1	x_2	x_3	VIF
49.77	< 0.001	0.34230	84.42%	79.07%	X			1.00
41.96	< 0.001	0.36809	81.99%	70.27%		X		1.00
0.62	0.455	0.88640	0.00%	0.00%			X	1.00
23.72	0.001	0.35264	83.47%	59.95%	X	X		11.59
22.63	0.001	0.35988	82.78%	68.11%	X		X	1.14
27.63	< 0.001	0.32970	85.55%	71.76%		X	X	1.33

Example

Regression Analysis: travelTime(y) versus milesTraveled(x1), numDeliveries(x2), gasPrice(x3)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	6.05612	2.01871	16.99	0.002
milesTraveled(x1)	1	0.04805	0.04805	0.40	0.548
numDeliveries(x2)	1	0.19373	0.19373	1.63	0.249
gasPrice(x3)	1	0.15761	0.15761	1.33	0.293
Error	6	0.71288	0.11881		
Total	9	6.76900			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.344694	89.47%	84.20%	57.49%

Example

Regression Analysis: travelTime(y) versus milesTraveled(x1), numDeliveries(x2), gasPrice(x3)

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	6.21	2.32	2.68	0.037	
milesTraveled(x1)	0.0141	0.0222	0.64	0.548	14.94
numDeliveries(x2)	0.383	0.300	1.28	0.249	17.35
gasPrice(x3)	-0.607	0.527	-1.15	0.293	1.71

Regression Equation

travelTime(y) = 6.21 + 0.0141 milesTraveled(x1) + 0.383 numDeliveries(x2)
- 0.607 gasPrice(x3)

Example

F	p–value	S	$R^2(adj)$	$R^2(pred)$	x_1	x_2	x_3	VIF
49.77	< 0.001	0.34230	84.42%	79.07%	X			1.00
41.96	< 0.001	0.36809	81.99%	70.27%		X		1.00
0.62	0.455	0.88640	0.00%	0.00%			X	1.00
23.72	0.001	0.35264	83.47%	59.95%	X	X		11.59
22.63	0.001	0.35988	82.78%	68.11%	X		X	1.14
27.63	< 0.001	0.32970	85.55%	71.76%		X	X	1.33
16.99	0.002	0.34469	84.20%	57.49%	X	X	X	below
					14.94	17.35	1.71	

Example

Best Subsets Regression: y **vs.** x_1(milesTraveled) an x_2 (numDeliverires)

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	x1	x2	x3
1	86.2	84.4	79.1	1.9	0.34231	X		
1	84.0	82.0	70.3	3.1	0.36809		X	
2	88.8	85.5	71.8	2.4	0.32970		X	X
2	87.1	83.5	59.9	3.3	0.35264	X	X	
3	89.5	84.2	57.5	4.0	0.34469	X	X	X

1. R^2 (adj) – the highest values
2. R^2 (pred) - the highest values
3. the difference between R^2 (adj) and R^2 (pred) ***[A large drop-off indicated overfitting]***
4. Mallows C_p closwer to low and approximately equals the number of Xs + 1 (that is k+1)
5. Using all of the above, choose the best model.

Exercise

Shipping Cost

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Let's see the multiple regression output

Shipment	Distance (in miles), X1	Number of items, X2	Cost per Unit, Y
1	67	32	28
2	59	12	30
3	50	32	16
4	58	31	24
5	45	20	14
6	63	19	37
7	58	28	32
8	60	27	36
9	63	30	29
10	62	25	30
11	51	20	17
12	48	27	18
13	63	27	45
14	64	10	48
15	56	27	41
16	49	23	20
17	55	31	26
18	57	24	27
19	59	26	34
20	59	12	50

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Coefficient of Deterrmination

A measure of **"explained variarion"**

Result shows that about 63% of the total variation in shipment cost (Y) is explained by the regression.

$$r^2 = \frac{SSR}{SST}$$

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.793690296
R Square	0.629944285
Adjusted R Square	0.586408319
Standard Error	6.770451909
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	k = 2	1326.536676	663.268338	14.46951422	0.000213929
Residual	n - k - 2 = 17	779.263324	45.83901906		
Total	n - k - 1 = 19	2105.8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-25.67584713	16.18042417	-1.586846356	0.130971754	-59.81355809	8.46186383
Distance (in miles), X1	1.212501152	0.259470084	4.672990175	0.000218376	0.665067127	1.759935177
Number of items, X2	-0.5673072	0.231022934	-2.455631526	0.025123479	-1.054722985	-0.079891415

First , we firm perform F test.

F statistic is a test of significance for the entire regression. **At α=0.05**, this regression is statistically significant because **p-value (significance F) < 0.05**.

P-value

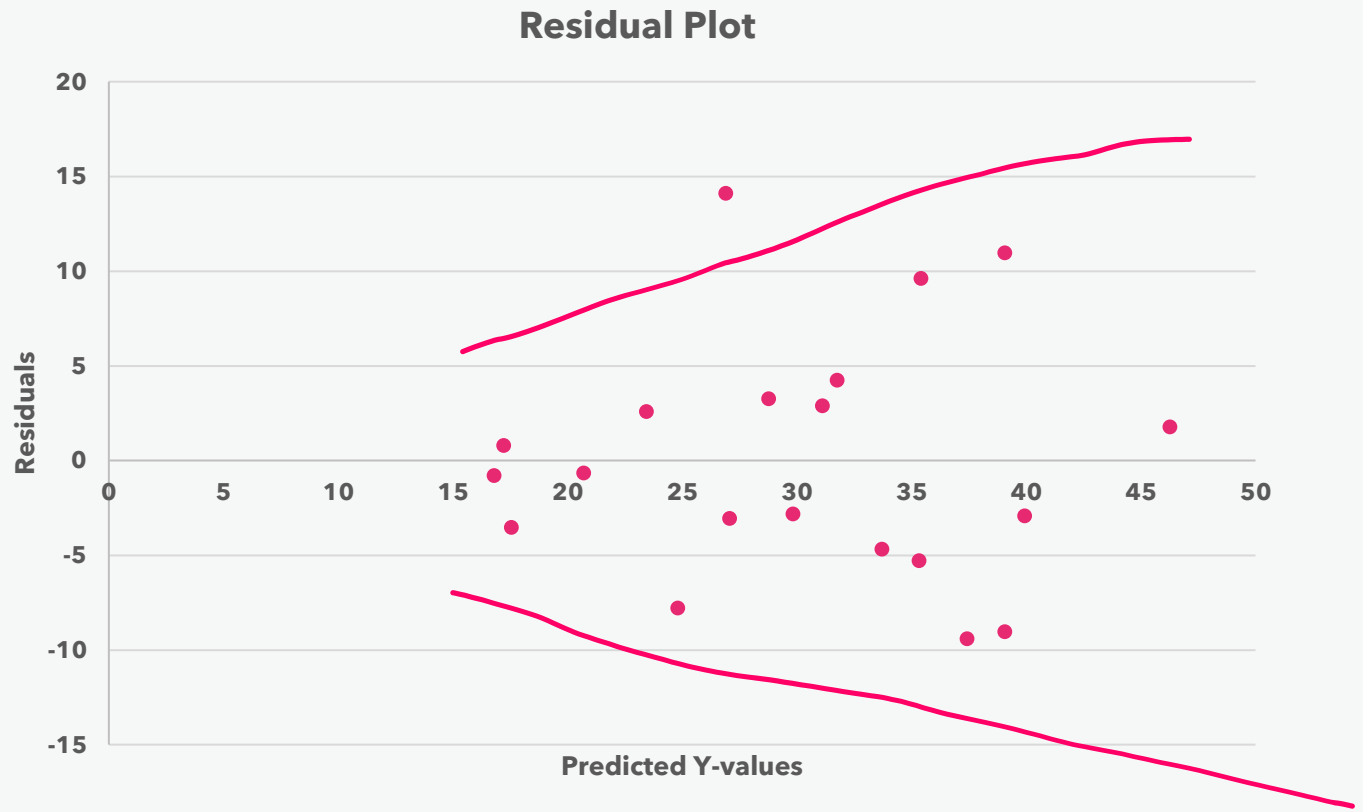
Prediction interval for E(Y)

Confidence interval for β_1

Confidence interval for β_2

At α=0.05, both t-values are **statistically significant** because their corresponding **p-values < 0.05**. Therefore, both X_1 and X_2 are individually useful in the prediction of Y.

Exercise


Residual Plot
Residuals
Predicted Y-values
20
15
10
5
0
-5
-10
-15
0
5
10
15
20
25
30
35
40
45
50

Exercise

Question 1. Define the prediction equation

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

Substituting:

$$\hat{Y} = -25.6758 + 1.2125 X_1 - 0.5673 X_2$$

for X_1=52 and X_2=17,

$\hat{Y} \approx -25.6758 + 1.2125\ (\mathbf{52}) - 0.5673(\mathbf{17}) = \mathbf{27.73}$

Exercise

Question 2. What propotion of the total variation in cost shipment (Y) is explained by the regression?

coefficient of determination, $\mathbf{R^2}$ =0.6299 = 62.99%

About **~63%** of the varriation in shipment cost is explained by changes in both travel distance (X_1) and number of items in each shipment (X_2)

> About **~37%** of the varriation varriation in shipment cost **CAN NOT BE explained** by changes in both travel distance (X_1) and number of items in each shipment (X_2)

The composite effect of this reflected in the adjusted coefficient of determination (**adjusted $\mathbf{R^2}$**). In this regression, **adj $\mathbf{R^2}$** = 59%

Exercise

Question 3: Is there a relationship between shipping cost (Y) and the two independent variables combined?

- H_0: $\beta_1 = \beta_2 = 0$ [No relationship between Y and the two variables combined]
- H_1: At least one of the β coefficients is not equal to 0

Test statistics, F = 14.4695 (with p-value =0.0002)

At $\boldsymbol{\alpha = 0.05}$, we **reject H_0**

- F statistic is **significant** (because **p-value is < 0.05**)
- There is evidence of a regression relationship between shipping cost (Y) and the two independent variables.

Exercise

Question 4: Does distance traveled (X_1), contribute information in the prediction of Y?

- H_0: $\beta_1 = 0$ [Distance traveled X_1, is not useful predictor of shipment cost]
- H_1: $\beta_1 \neq 0$

Test statistics, t-test = 4.6730 (with p-value =0.0002)

At $\boldsymbol{\alpha = 0.05}$, we **reject H_0**

- t statistic is **significant** (because **p-value is < 0.05**)
- There is evidence that X1 contributes information in the prediction of Y.
- The value of b1 is 1.2125, suggesting that for every 1 mile of shipping distance, the cost per unit rises by about $1.21, with X2 held constant.

Exercise

Question 5: Does X_2, the number of items in a shipment, contribute information in the prediction of Y?

- H_0: $\beta_2 = 0$ [#items in a shipment X_2, is not useful predictor of shipment cost]
- H_1: $\beta_2 \neq 0$

Test statistics, t-test = -2.4556 (with p-value =0.0251)

At $\boldsymbol{\alpha = 0.05}$, we **reject H_0**

- t statistic is **significant** (because **p-value is < 0.05**)
- There is evidence that X_2 contributes information in the prediction of Y.
- The value of b_2 is -0.5673, suggesting that every additional item conatined in a shipping, the cost per unit of shipment drops by about $0.57, with X_1 held constant.

Steps of Model Building

Step 1- Construct the scatter plot for each y and x_i (all variables)

Step 2 - Compute the correlation coefficients $r_{yx_i}, r_{x_ix_j}$, and then conduct the hypothesis testing of those population correlation coefficients. The hypotheses and test statistic are

Hypotheses	Test statistic	Rejection rule
1. $H_0 : \rho_{yx_i} = 0$ vs $H_1 : \rho_{yx_i} \neq 0$	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ $d.f. = n - 2$	Reject H_0 at α, if the value of test statistic $t \leq -t_{\alpha/2}$ or $t \geq t_{\alpha/2}$.
2. $H_0 : \rho_{x_ix_j} = 0$ vs $H_1 : \rho_{x_ix_j} \neq 0$		

Steps of Model Building

Step 3 - Diagnosing the **effects of multicollinearity** and correct them.

- **Multicollinearity** is the condition where among the independent variables are correlated with each other
- **Detection**
  - Correlation matrix
  - Variance inflation factors
- **Impact of Multicollinearity**
  - *Makes the value of R^2 to increase. This means that it is not the real value of R^2 of our model*

Steps of Model Building

Step 4 - Compute the **regression coefficients** (b_i) and then conduct the hypothesis testing of the **population regression coefficients** (β_i) and the regression model.

Hypotheses concerning the significance of the regression coefficients

Hypotheses	Test statistic	Rejection rule
$H_0 : \beta_i = 0$ $H_1 : \beta_i \neq 0$	$T_i = \dfrac{b_i - \beta_i}{S(b_i)}$	Reject H_0 at α, if $p-value < \alpha$. It implies that x_j has the effect on y

Hypotheses concerning the significance of the regression model as a whole

Hypotheses	Test statistic	Rejection rule
$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ H_1 : At least one of $\beta_1, \beta_2, \cdots, \beta_k$ does not equal 0	$F = \dfrac{MSR}{MSE}$	Reject H_0 at α, if $p-value < \alpha$. If H_0 is rejected, it implies that at least one of the independent variables $x_1, x_2, x_3, \cdots, x_k$ contributes significantly to the model.

Analysis of Variance for Testing

Source of variation	d.f.	SS.	MS.	Test statistic
Regression	k	SSR	MSR	$F = \dfrac{MSR}{MSE}$
Error	$n-k-1$	SSE	MSE	
Total	n - 1	SST		

Steps of Model Building

Step 5 - Diagnosing the **outlier(s) and influential observation(s)** and correct them.

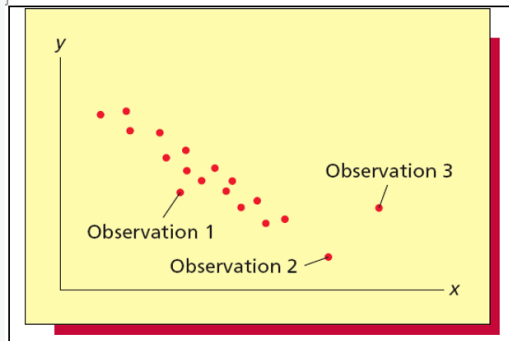


Figure 3.1 outlier(s) and influential observation(s)

- Observation 1: Outlying with respect to y value
- Observation 2: Outlying with respect to x value
- Observation 3: Outlying with respect to x value and y value not consistent with regression relationship (Influential)

- **Cook's Distance Measure** can use to identify influential observations
  - An observation will be the influential observations, if the value of $D_i > 1$
- **Leverage values** can help us identify outliers

The leverage value for an observation is the distance value. This value is a measure of the distance between the x value and the center of the experimental region
If the leverage value for an observation is large, it is an outlier with respect to its x value

- Large means greater than twice the average of all the leverage values
- An observation will be the outliers, if the leverage value is greater than $\frac{2(k+1)}{n}$

Steps of Model Building

Step 6 - Developing the **best estimated regression model**

Step 7 - Conduct the **residual analysis** of the model is obtained from step 6 with the following assumptions:

- **Constant Variance Assumptions**, by examining residual plots against the predicted *y* values.
- **Normality Assumption and E(e) = 0**, by using the Anderson Darling test statistic.
- **Independence Assumption**, by using the Durbin-Watson test statistic

Transforming the Dependent and Independent Variables

- A possible remedy for violations of the constant variance, correct functional form and normality assumptions is to transform the dependent variable
- Possible transformations include: Square root , Quartic root, Logarithmic
- The appropriate transformation will depend on the specific problem with the original data set

Step 8 - Indicate the best estimated regression model

Categorical Independent Variables

such as gender (male, female), method of payment (cash, credit card, check), and so on

Representing categorical variables

- **dummy** variables

  A dummy variable always has a value of either 0 or 1.

  For example, to model sales at two locations, would code the first location as a zero and the second as a 1. Operationally, it does not matter which is coded 0 and which is coded 1.

Categorical Independent Variables

Johnson Filtration, Inc., provides maintenance service for water-filtration systems throughout southern Florida. Customers contact Johnson with requests for maintenance service on their water-filtration systems. To estimate the service time and the service cost, Johnson's managers want to predict the repair time necessary for each maintenance request. Hence, repair time in hours is the dependent variable. Repair time is believed to be related to two factors, the number of months since the last maintenance service and the type of repair problem (mechanical or electrical). Data for a sample of 10 service calls are reported in Table

Service Call	Months Since Last Service	Type of Repair	Repair Time in Hours
1	2	electrical	2.9
2	6	mechanical	3.0
3	8	electrical	4.8
4	3	mechanical	1.8
5	2	electrical	2.9
6	7	electrical	4.9
7	9	mechanical	4.2
8	8	mechanical	4.8
9	4	electrical	4.4
10	6	electrical	4.5

Categorical Independent Variables

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

- the estimated regression equation:

$$\hat{y} = 2.15 + 0.304 x_1$$

Regression Statistics	
Multiple R	0.730873795
R Square	0.534176504
Adjusted R Square	0.475948567
Standard Error	0.781022322
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	5.596033058	5.596033058	9.17388683	0.016338159
Residual	8	4.879966942	0.609995868		
Total	9	10.476			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2.147272727	0.604977289	3.549344356	0.007516627	0.752192596	3.542352858	0.752192596	3.542352858
Months Since Last Service, x1	0.304132231	0.100412033	3.02884249	0.016338159	0.072581669	0.535682794	0.072581669	0.535682794

Categorical Independent Variables

Type of repair (categorical independent variables), x_2 - **dummy** or *indicator* **variable**

$$x_2 = \begin{cases} 0 & \text{if the type of repair is mechanical} \\ 1 & \text{if the type repair is electrical} \end{cases}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Service Call	Months Since Last Service	Type of Repair	Repair Time (hours)
1	2	1	2.9
2	6	0	3.0
3	8	1	4.8
4	3	0	1.8
5	2	1	2.9
6	7	1	4.9
7	9	0	4.2
8	8	0	4.8
9	4	1	4.4
10	6	1	4.5

Categorical Independent Variables

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.926927504
R Square	0.859194597
Adjusted R Sq	0.818964482
Standard Error	0.459048301
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	9.000922601	4.5004613	21.35700073	0.00104753
Residual	7	1.475077399	0.210725343		
Total	9	10.476			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.930495356	0.46697414	1.992605748	0.086558043	-0.173723	2.03471373	-0.173723	2.03471373
Months Since	0.387616099	0.062565187	6.195395815	0.000447255	0.23967294	0.53555926	0.23967294	0.53555926
Type of Repair	1.262693498	0.314126673	4.019695257	0.005061565	0.51990195	2.00548505	0.51990195	2.00548505

$$\hat{y} = 0.93 + 0.388x_1 + 1.26x_2$$

$\alpha = 0.05$

Categorical Independent Variables

Interpreting the Parameters

The multiple regression equation for the Johnson Filtration example

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\hat{y} = 0.93 + 0.388x_1 + 1.26x_2$$

$$\hat{y} = 0.93 + 0.388x_1$$

$$x_2 = \begin{cases} 0 & \text{if the type of repair is mechanical} \\ 1 & \text{if the type repair is electrical} \end{cases}$$

$$E(y \mid mechanical) = \beta_0 + \beta_1 x_1 + \beta_2 (0) = \beta_0 + \beta_1 x_1$$

$$E(y \mid electrical) = \beta_0 + \beta_1 x_1 + \beta_2 (1) = \beta_0 + \beta_1 x_1 + \beta_2$$

$$= (\beta_0 + \beta_2) + \beta_1 x_1$$

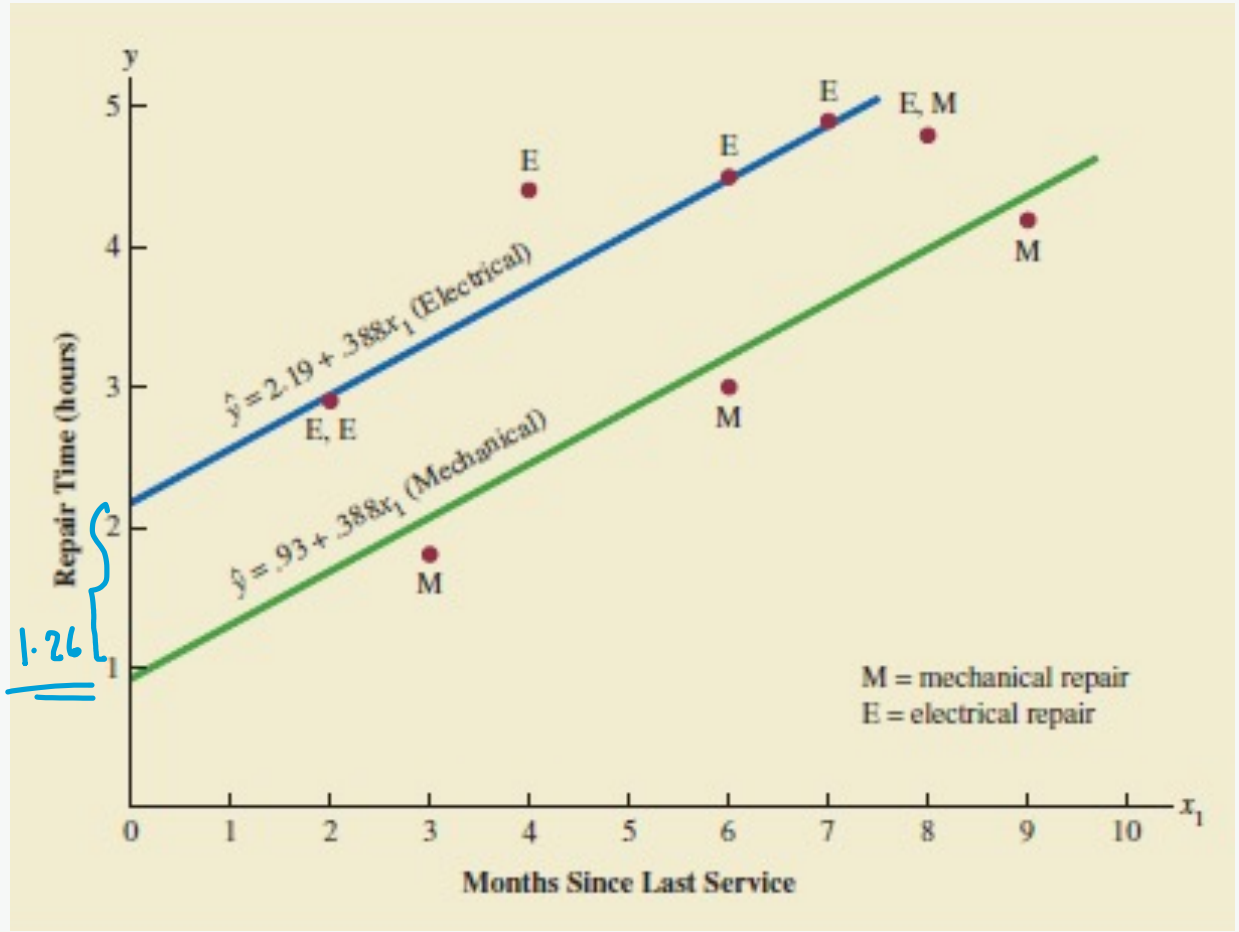
$$\hat{y} = 0.93 + 0.388x_1 + 1.26(1) = 2.19 + 0.388x_1$$

The slope of both equations is β_1, but the *y*-intercept differs

electrical repairs require 1.26 hours longer than mechanical repairs

Categorical Independent Variables

electrical repairs require 1.26 hours longer than mechanical repairs



More Complex Categorical Independent Variables

For example, suppose a manufacturer of copy machines organized the sales territories for a particular state into **three regions: A, B, and C**.

Because sales region is a categorical variable with three levels, A, B and C, we will need 3 – 1 = 2 dummy variables to represent the sales region. Each variable can be coded 0 or 1.

$$x_1 = \begin{cases} 1 & \text{if sales region B} \\ 0 & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if sales region C} \\ 0 & \text{otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if sales region D} \\ 0 & \text{otherwise} \end{cases}$$

Region	x_1	x_2
A	0	0
B	1	0
C	0	1

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$E(y \mid region\ A) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$$

$$E(y \mid region\ B) = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$$

$$E(y \mid region\ C) = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$$

More Complex Categorical Independent Variables

Let

- Type = 0 if a mechanical repair
- Type = 1 if an electrical repair

- Person = 0 if *Bob Jones* performed the service and
- Person = 1 if *Dave Newton* performed the service

Repair Time (hours)	Months Since Last Service	Type of Repair	Repairperson
2.9	2	electrical	Dave Newton
3.0	6	mechanical	Dave Newton
4.8	8	electrical	Bob Jones
1.8	3	mechanical	Dave Newton
2.9	2	electrical	Dave Newton
4.9	7	electrical	Bob Jones
4.2	9	mechanical	Bob Jones
4.8	8	mechanical	Bob Jones
4.4	4	electrical	Bob Jones
4.5	6	electrical	Dave Newton

Review of Multipe Linear Regression

- Multiple regression is an extension od simple linear regression
- Two or more independent variables are ised to predict/explainn the variance in one dependent variable
- Two problems may arise: Overfitting and 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 estomated change in y corresponding to a one unit change in a variable, **when all other variables are held constant.**

Summary (Simple vs. Multiple)

- The models have similar "LINE" assumptions. The only real difference is that whereas in simple linear regression we think of the distribution of errors at a fixed value of the single predictor, with multiple linear regression we have to think of the distribution of errors at a fixed set of values for all the predictors. All of the model checking procedures we learned earlier are useful in the multiple linear regression framework, although the process becomes more involved since we now have multiple predictors.
- The use and interpretation of r2 (which we'll denote R2 in the context of multiple linear regression) remains the same. However, with multiple linear regression we can also make use of an "adjusted" R2 value, which is useful for model building purposes.
- With a minor generalization of the degrees of freedom, we use *t*-tests and *t*-intervals for the regression slope coefficients to assess whether a predictor is significantly linearly related to the response, after controlling for the effects of all the other predictors in the model.
- With a minor generalization of the degrees of freedom, we use prediction intervals for predicting an individual response and confidence intervals for estimating the mean response.

Multiple Linear Regression Model *(Exercise)*

Driving Assignment	Travel time (hours) (y)	Miles Traveled(x1)	Number of Deliveries(x2)
1	9.3	100	4
2	4.8	50	3
3	8.9	100	4
4	6.5	100	2
5	4.2	50	2
6	6.2	80	2
7	7.4	75	3
8	6.0	65	4
9	7.6	90	3
10	6.1	90	2

Multiple Linear Regression Model *(example)*

Question 1: Define the prediction equation

Question 2: What propotion of the total variation in Y is explained by the regression?

Question 3: Is there a relationship between Y and the two independent variables combined?

Question 4: Does miles traveled (X_1), contribute information in the prediction of Y?

Question 5: Does X_2, the number of deliveries, contribute information in the prediction of Y?

ASSIGNMENT

Assignment

A 10-year study conducted by the American Heart Association provided data on how age, blood pressure, and smoking relate to the risk of strokes. Assume that the following data are from a portion of this study. Risk is interpreted as the probability (times 100) that the patient will have a stroke over the next 10-year period. For the smoking variable, define a dummy variable with 1 indicating a smoker and 0 indicating a nonsmoker.

Risk	Age	Pressure	Smoker
12	57	152	No
24	67	163	No
13	58	155	No
56	86	177	Yes
28	59	196	No
51	76	189	Yes
18	56	155	Yes
31	78	120	No
37	80	135	Yes
15	78	98	No
22	71	152	No
36	70	173	Yes
15	67	135	Yes
48	77	209	Yes
15	60	199	No
36	82	119	Yes
8	66	166	No
34	80	125	Yes
3	62	117	No
37	59	207	Yes

Assignment (cont.)

a) Develop an estimated regression equation that relates risk of a stroke to the person's age, blood pressure, and whether the person is a smoker.

b) Is smoking a significant factor in the risk of a stroke? Explain. Use $\alpha=0.05$.

c) What is the probability of a stroke over the next 10 years for Art Speen, a 68-year-old smoker who has blood pressure of 175? What action might the physician recommend for this patient?