Multiple Regression

Linear relationship between more than two variables!

We have seen why we care about a regression line when data satisfies a linear correlation between two variables x and y. And, the whole reason is so that we can be **predictive**. However, what if we want to do determine if we can be predictive with more than one independent variable? Can create a **Multiple Regression Equation** describing a linear relationship between $\boldsymbol{y}$ with independent variables $\boldsymbol{x_1, x_2, x_3, \ldots, x_k}$.

Example: Can we predict the weight (pounds) of a man, if we have knowledge of his height (inches) and waist circumference (inches)?

Let $\boldsymbol{y}$ represent the weight in pounds, $\boldsymbol{x_1}$ represent the height in inches, $\boldsymbol{x_2}$ represent the circumference in inches. The following data was collected.

Height inches x_1	**Waist** inches x_2	**Weight** pounds y
70	32	160
72	33	180
71	31	168
77	38	200
72	34	175
78	36	192
72	32	165
77	35	220
70	33	188
74	38	205
70	35	196
68	30	154
72	34	187

(x_1, x_2, y)

$n = 13$

The main question we will need to answer is whether the multivariate sample data fits the **Multiple (Linear) Regression Equation** so that we can be **predictive**.

Population Multiple (Linear) Regression Equation

$$y = \beta_0 + \beta_1 x + \beta_2 x + \cdots + \beta_k x$$

Essentially, we will use technology **(TI-84 Plus CE)** and the added program called ***A2MULREG*** which does not come stock with the **TI-84 PLUS CE** calculator. You will need to install it in your calculator. We will make use of the value of $\boldsymbol{adjusted\ R^2}$ and the $\boldsymbol{p}$**-value** when deterring how well the sample data fits the **Multiple Regression Equation.**

Notation

Sample Multiple Regression Equation

$$\hat{y} = b_0 + b_1 x + b_2 x + \cdots + b_k x$$

Sample Size n; k is the number of independent variables.

Definition
Multiple Coefficient of Determination R^2- Is the measure (number) of how well the Population Multiple Regression Equation fits the sample data.

Definition
We say there is a **perfect fit**, If $R^2 = 1$.

Definition
We say there is a **very good fit**, if R^2 is close to 1.

Definition
We say there is a **very poor fit**, if R^2 is close to 0.

Unfortunately, R^2 will naturally tend to 1 as we add additional variables which is very problematic!

Therefore, we will need to use an adjusted $\boldsymbol{R^2}$ value which is called the **adjusted coefficient of determination**.

Definition
Adjusted Coefficient of Determination R^2- The multiple coefficients of determination R^2 modified to account for the number of variables and the sample size n.

$$\textbf{Adjusted } \boldsymbol{R^2} = \mathbf{1} - \frac{(\boldsymbol{n-1})}{[\boldsymbol{n-(k+1)}]}(\mathbf{1} - \boldsymbol{R^2})$$

n is the sample size
k is the number of predictor variables x

p-value Method

A low p-value suggests that the null H_0 must go!

$H_0: \beta_0 = \beta_1 = \cdots = \beta_k = 0$ (No Linear Correlation)
H_1: at least one $\beta_k \neq 0$ (Linear Correlation)

Unfortunately, when determine the multiple regression equation is very difficult and not always easy to do. We often do not always include the predictor variables x as well. As a result, we must follow **guidelines** to determine the **"Best Multiple Regression Equation".**

Guidelines for Finding the Best Multiple Regression Equation

1. Use **common sense** and practical considerations to **include, or exclude, predictor variables x.** Only use variables that are relevant to what you are seeking to predict. If you are seeking to determine the weight of sons, do not include the weight of physicians that delivered the sons.

2. Consider the **p-value.** Select an equation that has a low p-value found in the technology you use.

3. Consider equations with hive value of **Adjusted R^2 and try to include only a few variables.** You want to try to avoid including every predictor variable and limit it only to a few predictors' variables x.

- Select an equation having an **Adjusted R^2** value: if you are looking to add an additional predictor variable x, make sure the **Adjusted R^2** does not increase very much.

- For a particular number of predictor variables, select the equation with the largest **Adjusted R^2** value.

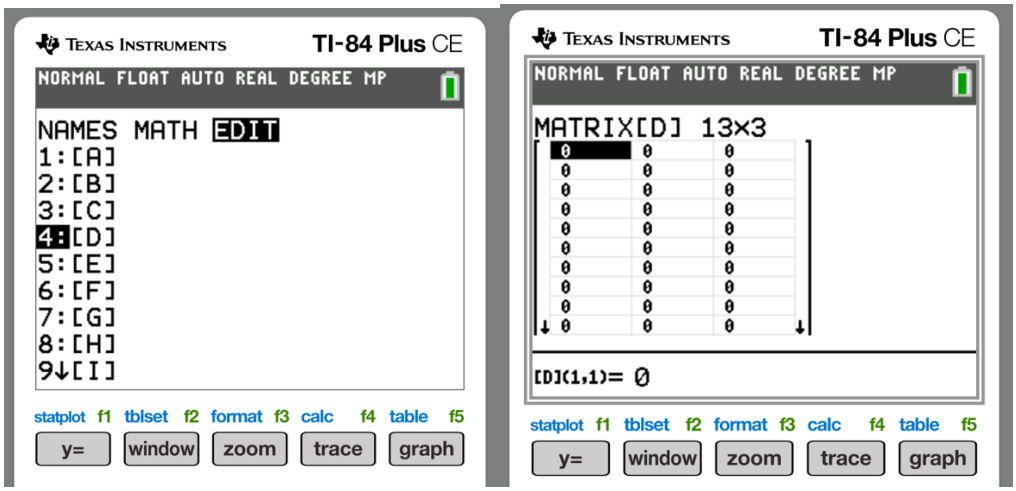
- When excluding predictor variables x that do not have much of an effect on the response variable y, it may be helpful to find the linear correlation coefficient r for each pair of variables being considered. If two predictor variables x have a very high linear correlation **(multicollinearity)**, there is no need to include both of those variables. Exclude the variable with the lowest **Adjusted R^2** value.

Common Sense and critical thinking are essential to for effective use of processes in statistics and this is where these skills come into play.

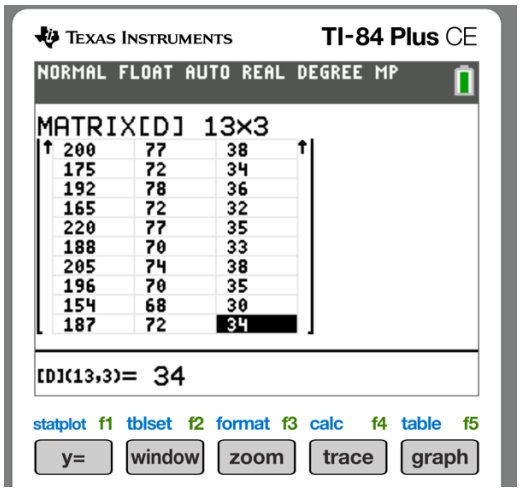
Enter the tri-variate data on the **weight of men** in the **TI-84 PLUS CE** calculator as a **matrix.**

Enter the data and name the matrix D

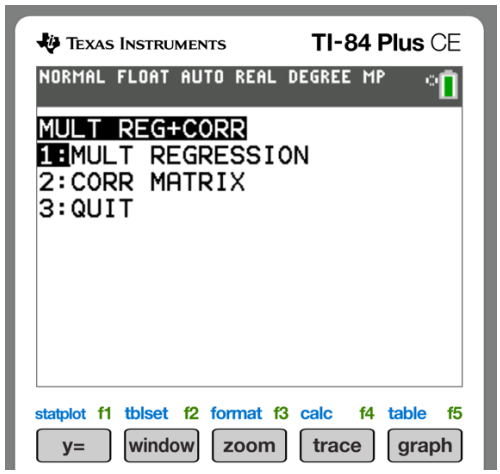
2nd>Matrix> edit>D and enter the dimensions of 13x3



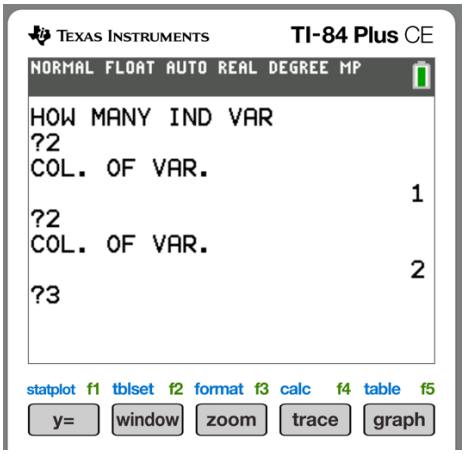
Enter the values for your variables y, x_1, x_2. y must be in the first column of your matrix!



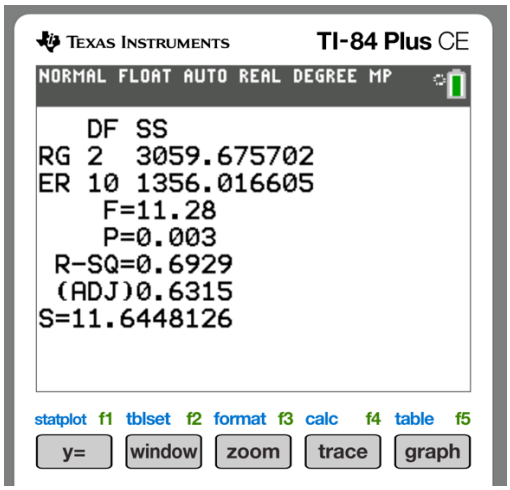
Enter Program>A2MULREG>Mult Regression press enter three times



Enter the number of independent variables x and the column numbers of the variables x

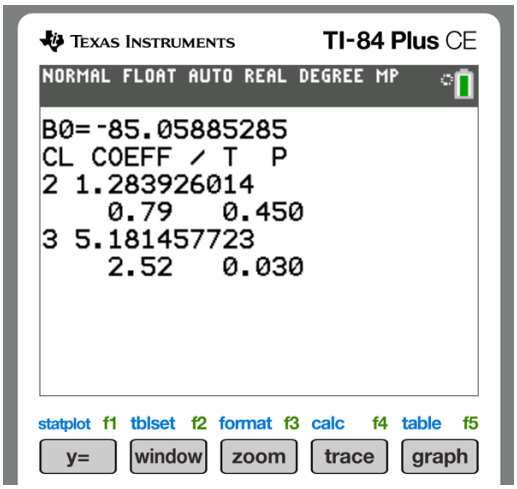


Enter the column number of each variable x from the D matrix and press enter.



$p \approx 0.003$ **which is low, the Null** H_0 **has to go!**
There is a Multilinear Correlation between the variables.
$R^2 \approx 0.6929$**; Poor to Good fit?**
Adjusted $R^2 \approx 0.6315$; **Poor to Good fit?**

Press enter to see the Multiple Regression Equation Coefficients.



In Data Analysis Pack with Microsoft Excel

Regression Statistics								
Multiple R	0.83241193							
R Square	0.69290962							
Adjusted R Square	0.63149155							
Standard Error	11.6448126							
Observations	13							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	3059.6757	1529.83785	11.2818519	0.002731059			
Residual	10	1356.016608	135.601661					
Total	12	4415.692308						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-85.058853	81.00069231	-1.0501003	0.31838695	-265.5396424	95.4219367	-265.53964	95.4219367
X Variable 1	1.28392601	1.633592918	0.78595224	0.45011744	-2.355945834	4.92379786	-2.3559458	4.92379786
X Variable 2	5.18145772	2.053504126	2.52322732	0.03021971	0.605965397	9.75695005	0.6059654	9.75695005

Notes about what values in these tables.

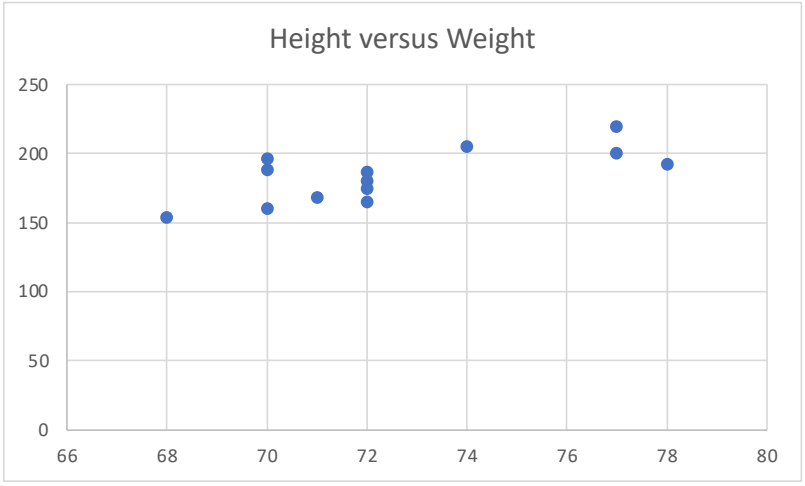
$\boldsymbol{R^2 \approx 0.693}$ is the known coefficient of determination where 69.3% of the variation in weights of men can be explained by the height and waist circumference.

Standard Error $\boldsymbol{\approx 11.645}$ The average distance the observed values fall from the regression line.

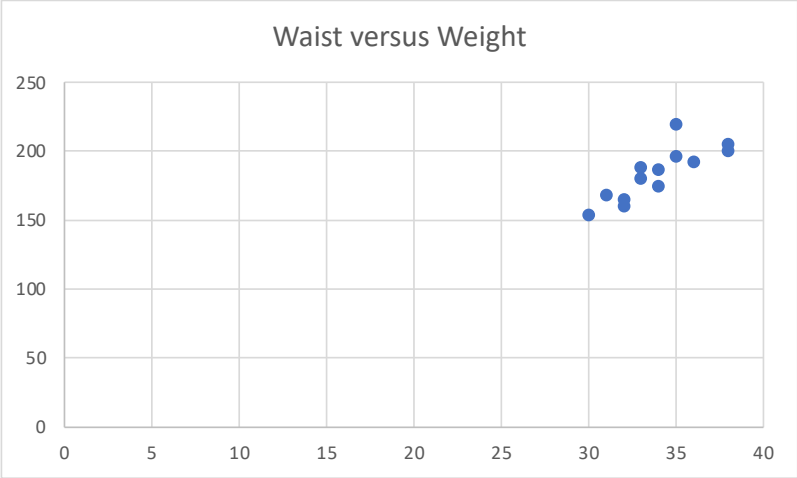
$\boldsymbol{F \approx 11.282}$ The overall F Statistic for the regression model. Calculated as regression MS/Residual MS.

Significance F$\boldsymbol{\approx 0.003}$ is the p-value**.** It tells us whether the regression model is statistically significant. The height and waist circumference variables have a statistically significant association with the weight of men.

The **individual p values** tell is whether each variable is statistically significant. The p value of the x1 variable (height) is approximately 0.450 which tells us this variable not statistically significant at $\boldsymbol{\alpha = 5\%}$ and can be removed. If we consider the scatter plot of the weight versus the height variable, we see there is a **weak linear correlation** between the variables.



The x2 variable (waist) has a p value that's approximately 0.030 and is statistically significant at $\boldsymbol{\alpha = 5\%}$. Consider the Scatter plot for this variable. There is a **stronger linear correlation**.

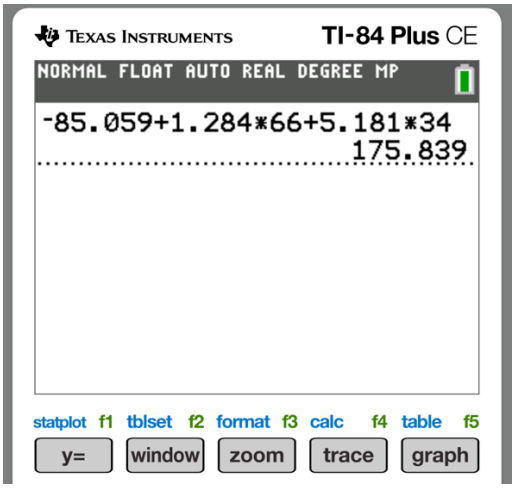


It may be that in removing the x1 variable (height of men) we will have a stronger correlation between variables and are better able to be **predictive** as a linear correlation model with only the x2 (waist) variable. Nevertheless, the **Multi Regression Equation** is illustrated below.

$$\hat{y} = b_0 + b_1 x + b_2 x$$

$$\boldsymbol{\hat{y} = -85.059 + 1.284 \cdot height + 5.181 \cdot waist}$$

Let's try to be predictive. What is the weight of a man whose weight is 66 inches and waist is 34 inches?



$\boldsymbol{\hat{y} \approx 176}$ **pounds**

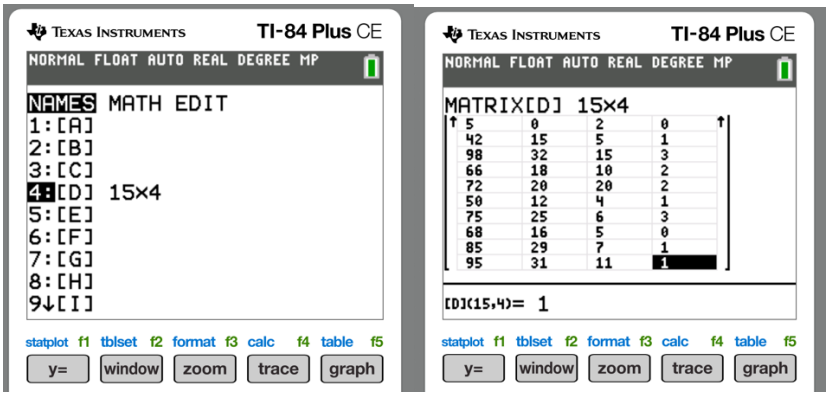
Example: Consider the following set of multivariate data that relates the percentage earned on a **final exam** based on the preparation students engaged in for their final exam. There are **three independent variables** in this experiment.

HW hours x_1	Tutoring hours x_2	Practice exams x_3	Score percent y
25	12	2	95
30	5	1	88
22	8	0	78
4	0	0	12
8	6	0	15
0	2	0	5
15	5	1	42
32	15	3	98
18	10	2	66
20	20	2	72
12	4	1	50
25	6	3	75
16	5	0	68
29	7	1	85
31	11	1	95

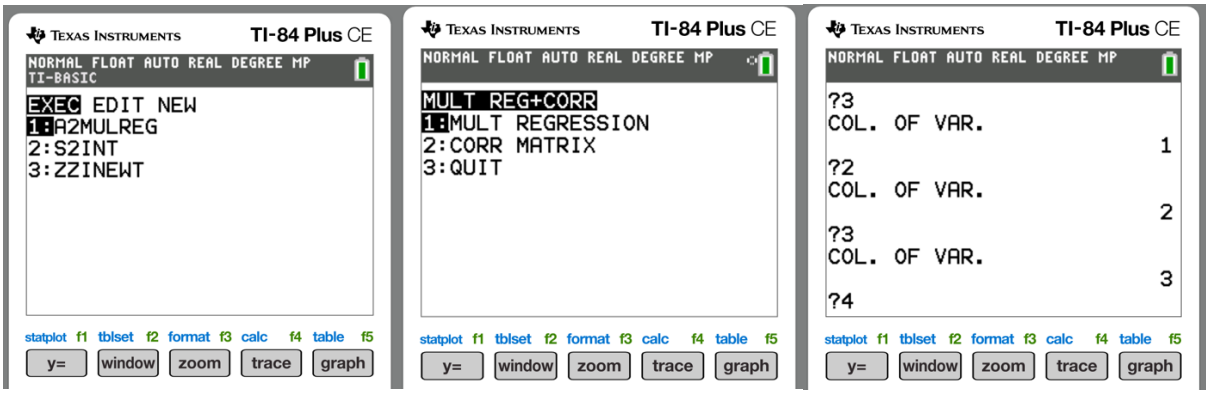
(x_1, x_2, x_3, y)

$n = 15$

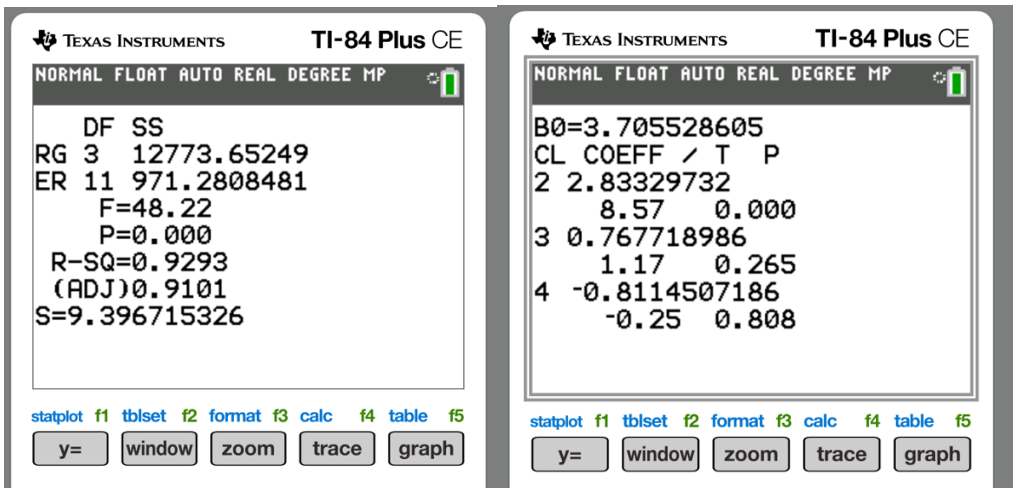
Enter the data and name the matrix D as a 15x4 Matrix.
Enter the values in Matrix D
Be sure to enter the y variables in the first row.



Enter Program>A2MULREG>Multi Regression press enter three times.



Results



In Data Analysis Pack with Microsoft Excel

Regression Statistics	
Multiple R	0.964020412
R Square	0.929335354
Adjusted R Square	0.910063178
Standard Error	9.396715326
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	12773.65249	4257.884162	48.22160951	1.2817E-06
Residual	11	971.280848	88.29825891		
Total	14	13744.93333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3.705528605	5.592832408	0.662549552	0.521259329	-8.604212528	16.01526974	-8.604212528	16.01526974
X Variable 1	2.83329732	0.330666126	8.56845348	3.38119E-06	2.105506083	3.561088556	2.105506083	3.561088556
X Variable 2	0.767718986	0.654078375	1.173741581	0.265285798	-0.671897811	2.207335783	-0.671897811	2.207335783
X Variable 3	-0.811450719	3.265600792	-0.248484359	0.808337708	-7.998989601	6.376088164	-7.998989601	6.376088164

Notes about what values in these tables.

$\boldsymbol{R^2 \approx 0.929}$**; Very Good Fit**

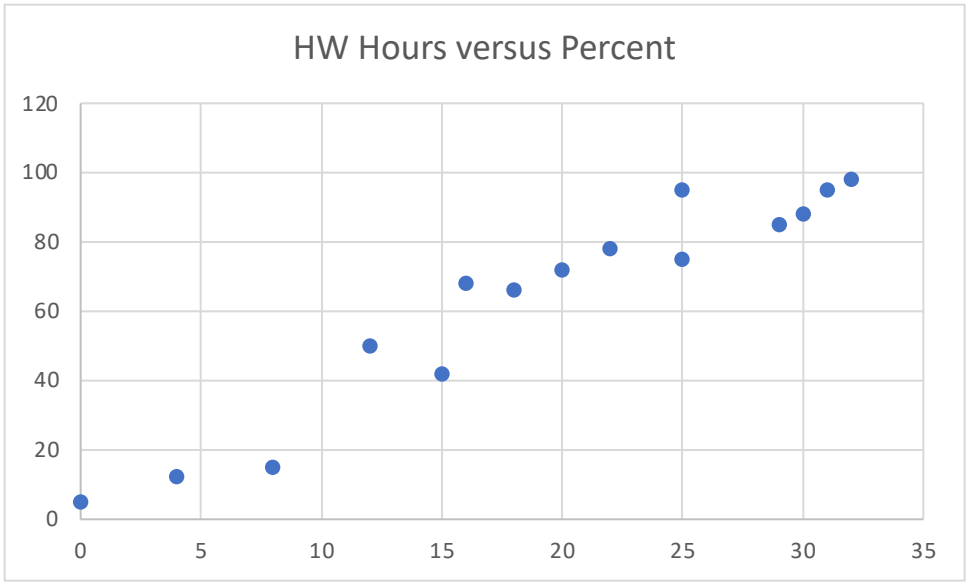
Adjusted $\boldsymbol{R^2 \approx 0.0.910}$**; Very Good Fit**

Standard Error $\boldsymbol{\approx 9.397}$ The average distance the observed values fall from the regression line.

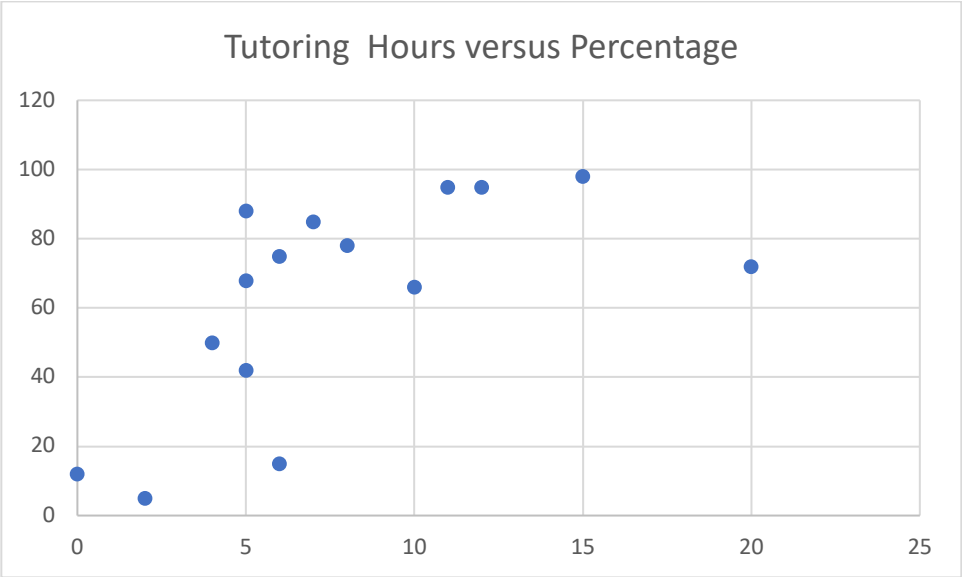
$\boldsymbol{F \approx 48.222}$ The overall F Statistic for the regression model. Calculated as regression MS/Residual MS.

Significance F$\boldsymbol{\approx 0.000}$ is the p-value**.** There is statistically significant association with independent variables.

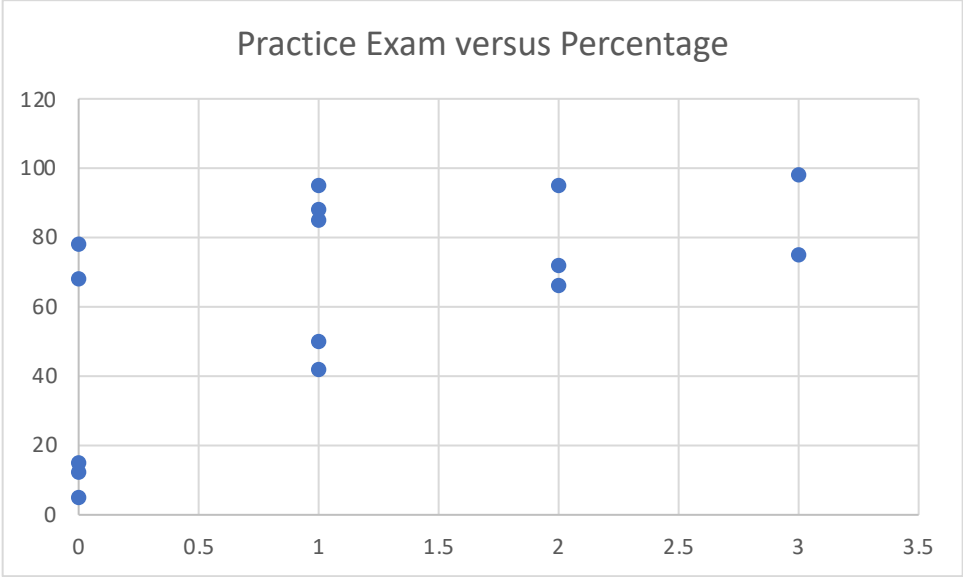
Individual p values
The p value of the x1 variable (Study Hours) is approximately 0.000 which tells us this variable is statistically significant at $\boldsymbol{\alpha = 5\%}$. If we consider the scatter plot **HW versus the percentage**, we see there is a **strong linear correlation** between the variables.



The p value of the x2 variable (Tutoring Hours) is approximately 0.265 which tells us this variable is not statistically significant at $\alpha = 5\%$. If we consider the scatter plot of **Tutoring Hours variable versus percentage**, we see there is a at best a **weak linear correlation** between the variables. We can possibly remove this variable from the model.



The p value of the x3 variable (practice exams) is approximately 0.808 which tells us this variable is not statistically significant at $\alpha = 5\%$. We consider the scatter plot of **Practice Exams versus Percentage**; we see there is a **no linear correlation** between the variables. We can remove this variable from the model.



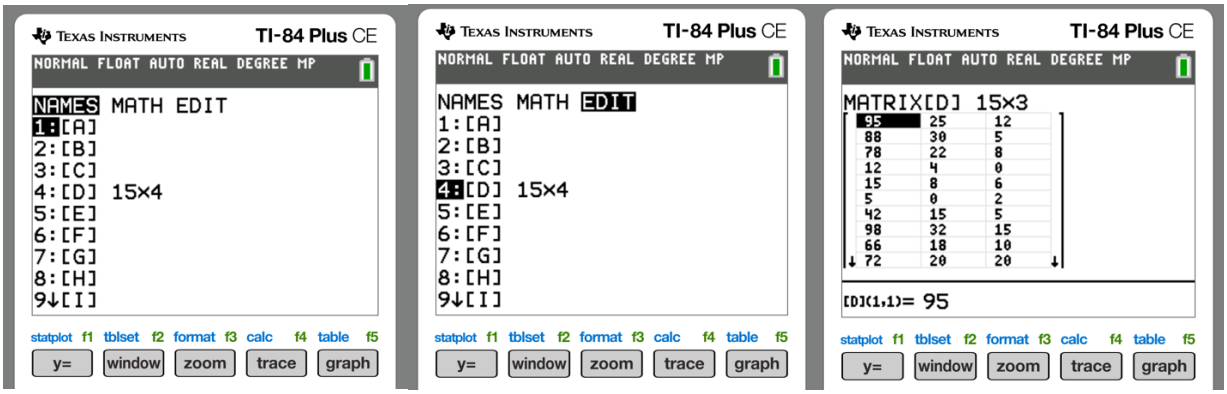
Now we can rerun the experiment by excluding the practice exam variable x_3.

HW hours x1	Tutoring hours x2	Score percent y
25	12	95
30	5	88
22	8	78
4	0	12
8	6	15
0	2	5
15	5	42
32	15	98
18	10	66
20	20	72
12	4	50
25	6	75
16	5	68
29	7	85
31	11	95

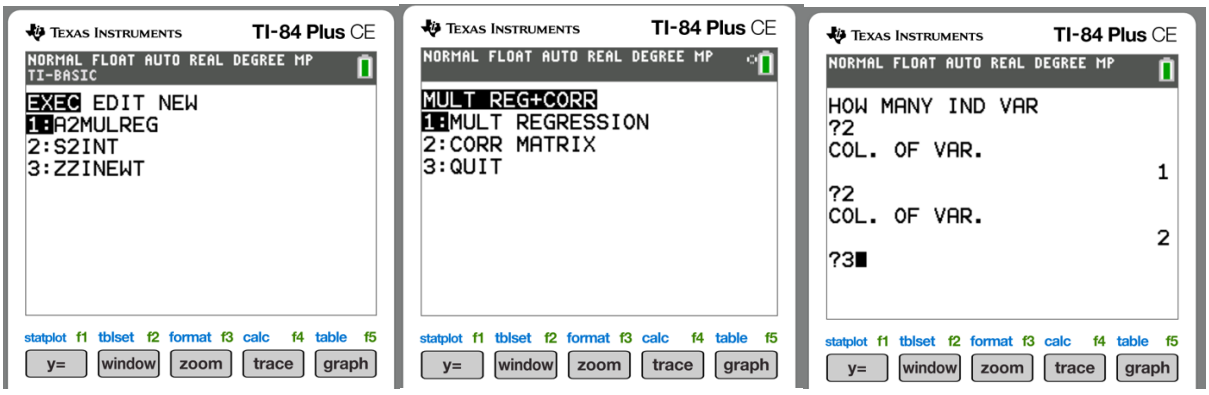
(x_1, x_2, y)

$n = 15$

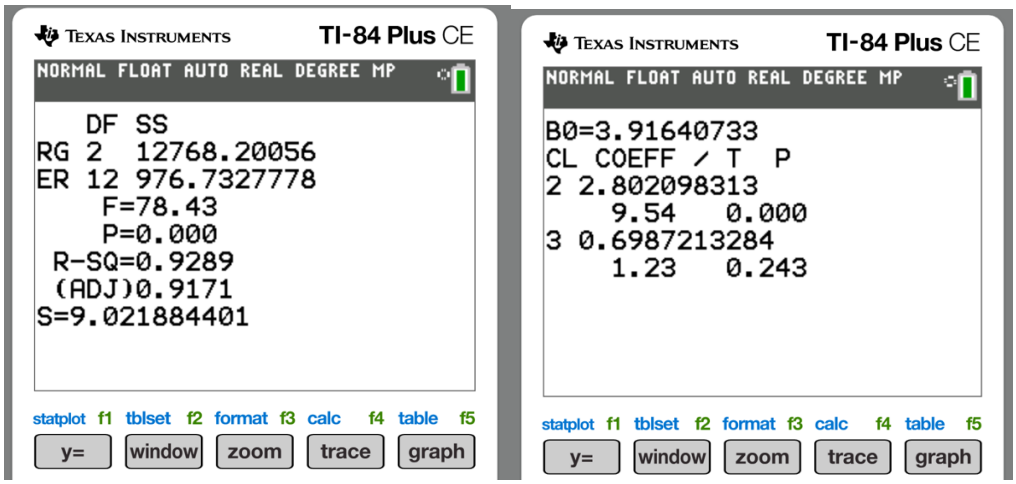
Enter the data and name the matrix D as a 15x3 Matrix.
Enter the values in Matrix D
Be sure to enter the y variables in the first row.



Notice by changing the dimensions to 15x3 the last column x_3 was naturally deleted.



Results



In Data Analysis Pack with Microsoft Excel

Regression Statistics								
Multiple R	0.963814663							
R Square	0.928938704							
Adjusted R Square	0.917095155							
Standard Error	9.021884401							
Observations	15							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	12768.20056	6384.100278	78.43414808	1.28765E-07			
Residual	12	976.7327778	81.39439815					
Total	14	13744.93333						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3.91640733	5.307557036	0.737892651	0.474762882	-7.647766035	15.48058069	-7.647766035	15.48058069
X Variable 1	2.802098313	0.293698817	9.540720452	5.93766E-07	2.162183562	3.442013065	2.162183562	3.442013065
X Variable 2	0.698721328	0.568589398	1.228868021	0.242675293	-0.540128547	1.937571203	-0.540128547	1.937571203

Notes about what values in these tables.

$R^2 \approx 0.929$; **Very Good Fit**

Adjusted $R^2 \approx 0.0.917$; **Very Good Fit**

Standard Error ≈ 9.022 The average distance the observed values fall from the regression line.

$F \approx 48.434$ The overall F Statistic for the regression model. Calculated as regression MS/Residual MS.

Significance F≈ 0.000 is the p-value**.** There is **statistically significant association with independent variables**.

Individual p values

The p value of the x1 variable (Study Hours) is approximately 0.000 which tells us this variable is **statistically significant** at $\alpha = 5\%$.

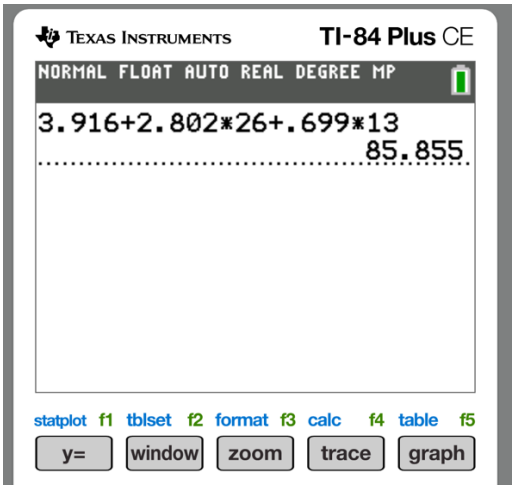
The **p value** of the x2 variable (Tutoring Hours) is approximately 0.243 which tells us this variable is **not statistically significant** at $\alpha = 5\%$.

Multi Regression Equation

$$\hat{y} = b_0 + b_1 x + b_2 x$$

$$\hat{\boldsymbol{y}} = \mathbf{3.916} + \mathbf{2.802} \cdot \boldsymbol{HW} + \mathbf{0.699} \cdot \boldsymbol{Tutoring}$$

If a student did **homework** for **26 hours** and did **tutoring** for **13 hours**, what percent will they earn on the test?



$$\hat{\boldsymbol{y}} \approx \mathbf{86}\%$$

Example: Instructor Classroom Enrollment (First Day) versus Perceived Instructor Effectiveness Survey. Is there a multilinear relationship between the class size and student perceptions based on a survey. Instructors were rated on a scale of 1 to 10 on various student perceived attributes. The following is a summary of the multivariate data that was collected.

Instructor Organization- Syllabus provided, grades assignments on time, lectures are clear and concise. 1 is the lowest rating and 10 is the highest rating.

Instructor Communication Skills- Students can clearly understand the oral communication and written communication of an instructor. 1 is the lowest rating and 10 is the highest rating.

Instructor Attitude- Students feel comfortable with the instructors' responses over all questions and interaction. 1 is the lowest rating (snarky) and 10 is the highest rating (friendly).

Instructor Punctuality- The instructor shows up to lecture and starts class on time on a consistent basis. 1 is the lowest rating (rarely) and 10 is the highest rating (consistently).

Instructor Organuzation x_1	**Instructor** Communication Skills x_2	**Instructor** Attitude x_3	**Instructor** Punctuality x_4	**Enrollment** 1st Day Class Size y
10	10	10	10	50
10	10	9	8	45
6	5	5	9	12
1	2	2	8	10
1	1	2	8	8
6	6	7	10	22
1	2	2	8	2
8	8	8	10	40
10	10	10	10	48
10	10	10	10	50
8	10	10	8	45
10	10	10	8	49
2	2	2	8	10
2	2	4	8	7
1	1	1	1	4
7	7	7	10	28
7	7	7	10	30
3	3	4	8	16

(x_1, x_2, x_3, x_4, y)

$n = 18$

In Data Analysis Pack with Microsoft Excel

Regression Statistics								
Multiple R	0.976506089							
R Square	0.953564142							
Adjusted R Square	0.939276186							
Standard Error	4.499735731							
Observations	18							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	4	5405.225363	1351.306341	66.73901578	1.55514E-08			
Residual	13	263.2190814	20.24762165					
Total	17	5668.444444						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.334304549	4.643939188	-0.07198728	0.943707773	-10.36692521	9.698316116	-10.36692521	9.698316116
X Variable 1	-0.567007279	1.786701434	-0.317348645	0.75601841	-4.426941056	3.292926498	-4.426941056	3.292926498
X Variable 2	4.260276919	2.490691538	1.710479541	0.110919304	-1.120535014	9.641088851	-1.120535014	9.641088851
X Variable 3	1.452226845	2.065715524	0.703013957	0.494445106	-3.010480226	5.914933916	-3.010480226	5.914933916
X Variable 4	-0.466546544	0.65460271	-0.712717098	0.48861692	-1.880729722	0.947636634	-1.880729722	0.947636634

Notes about what values in these tables.

$\boldsymbol{R^2 \approx 0.954}$; **Very Good Fit**

Adjusted $\boldsymbol{R^2 \approx 0.939}$; **Very Good Fit**

Standard Error $\boldsymbol{\approx 4.500}$ The average distance the observed values fall from the regression line.

$\boldsymbol{F \approx 66.739}$ The overall F Statistic for the regression model. Calculated as regression MS/Residual MS.

Significance F$\boldsymbol{\approx 0.000}$ is the p-value**.** There is **statistically significant association with independent variables**.

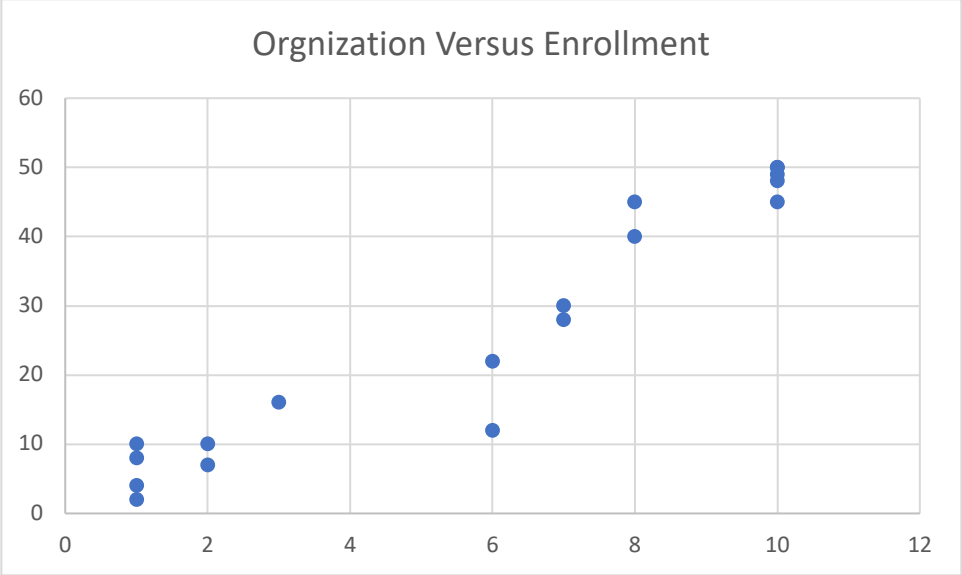
Individual p values
The p value of the x1 variable (Instructor Organization) is approximately 0.756 which tells us this variable is **not statistically significant** at $\boldsymbol{\alpha = 5\%}$ and may be removed.

The **p value** of the x2 variable (Instructor Communication Skills) is approximately 0.111 which tells us this variable is **not statistically significant** at $\boldsymbol{\alpha = 5\%}$. However, it is much closer to the level of significance than all the other variables. We should keep this variable.

The **p value** of the x3 variable (Instructor Attitude) is approximately 0.494 which tells us this variable is **not statistically significant** at $\boldsymbol{\alpha = 5\%}$. However, we may keep the variable when we rerun the experiment.

The **p value** of the x4 variable (Instructor Punctuality) is approximately 0.489 which tells us this variable is **not statistically significant** at $\boldsymbol{\alpha = 5\%}$. However, we may keep the variable when we rerun the experiment.

The overall experiment is **statistically significant**, but let's look at the **Scatter Plot** of **Instructor Organization** versus **Enrollment** on 1st day.



Looks like there is a **Strong Linear Correlation** between the Instructor **Organization** versus 1st Day **Enrollment.** We should seriously consider keeping the variable. However, let's remove the variable and rerun the experiment to see, if we get a better R^2 **and Adjusted** R^2 **values.**

Rerun

Instructor Communication Skills x1	Instructor Attitude x2	Instructor Punctuality x3	Enrollment 1st Day Class Size y
10	10	10	50
10	9	8	45
5	5	9	12
2	2	8	10
1	2	8	8
6	7	10	22
2	2	8	2
8	8	10	40
10	10	10	48
10	10	10	50
10	10	8	45
10	10	8	49
2	2	8	10
2	4	8	7
1	1	1	4
7	7	10	28
7	7	10	30
3	4	8	16

(x_1, x_2, x_3, y)

$n = 18$

In Data Analysis Pack with Microsoft Excel

Regression Statistics								
Multiple R	0.976321876							
R Square	0.953204406							
Adjusted R Square	0.943176779							
Standard Error	4.352817008							
Observations	18							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	3	5403.186222	1801.062074	95.05782245	1.51207E-09			
Residual	14	265.2582227	18.94701591					
Total	17	5668.444444						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.069957257	4.419453775	-0.01582939	0.987593876	-9.548742881	9.408828367	-9.548742881	9.408828367
X Variable 1	3.7313417	1.790393235	2.084090594	0.05595428	-0.108669876	7.571353276	-0.108669876	7.571353276
X Variable 2	1.422327816	1.996189376	0.712521484	0.487847977	-2.859072585	5.703728217	-2.859072585	5.703728217
X Variable 3	-0.491572693	0.628617744	-0.781989846	0.44724508	-1.839823663	0.856678277	-1.839823663	0.856678277

Notes about what values in these tables.

$\boldsymbol{R^2 \approx 0.953}$**; Very Good Fit and is slightly better than the previous experiment.**

Adjusted $\boldsymbol{R^2 \approx 0.943}$**; Very Good Fit and is slightly better than the previous experiment.**

Standard Error $\boldsymbol{\approx 4.353}$ The average distance the observed values fall from the regression line.

$\boldsymbol{F \approx 95.058}$ The overall F Statistic for the regression model. Calculated as regression MS/Residual MS.

Significance F$\approx \mathbf{0.000}$ is the p-value**.** There is **statistically significant association with independent variables**.

Individual p values
The p value of the x1 variable (Instructor Organization) is approximately 0.006 which tells us this variable is **statistically significant** at $\boldsymbol{\alpha = 5\%}$.

The **p value** of the x2 variable (Instructor Communication Skills) is approximately 0.488 which tells us this variable is **not statistically significant** at $\boldsymbol{\alpha = 5\%}$.

The **p value** of the x3 variable (Instructor Attitude) is approximately 0.494 which tells us this variable is **not statistically significant** at $\boldsymbol{\alpha = 5\%}$. However, we may keep the variable when we rerun the experiment.

The **p value** of the x4 variable (Instructor Punctuality) is approximately 0.447 which tells us this variable is **not statistically significant** at $\boldsymbol{\alpha = 5\%}$.

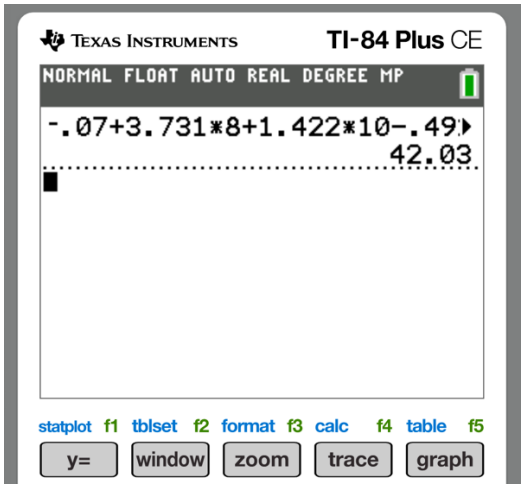
The overall experiment is **statistically significant**, so let's not exclude any more variables and proceed with our **Multi Regression Equation.**

Multi Regression Equation

$$\hat{y} = b_0 + b_1x + b_2x + b_3x$$

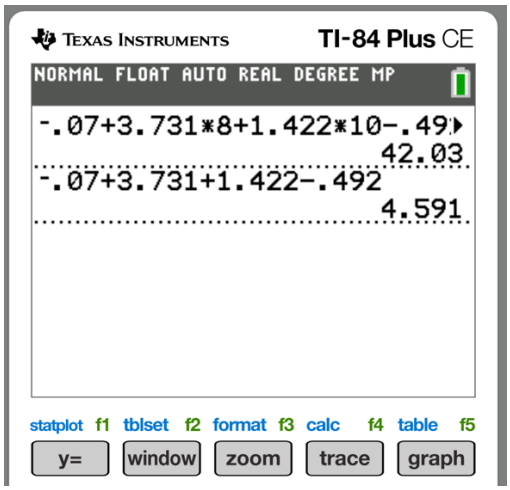
$$\boldsymbol{\hat{y} = -.070 + 3.731 \cdot Communication + 1.422 \cdot Attitude - 0.492 \cdot Punctuality}$$

If an instructor is rated an 8 for communication, 10 for attitude, and a 4 for punctuality, determine the 1st day enrollment class size.



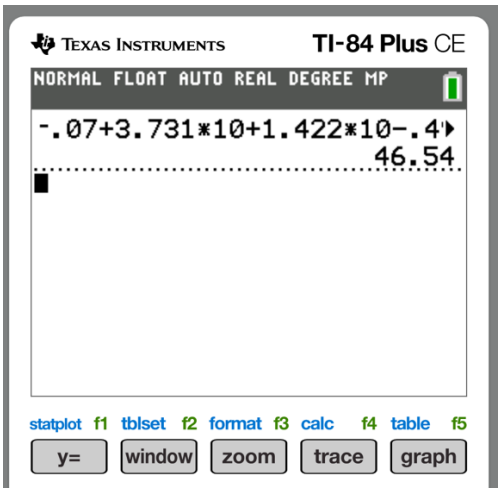
$\boldsymbol{\hat{y} \approx 42}$ **students**

If an instructor is rated a 1 for communication, 1 for attitude, and a 1 for punctuality, determine the 1st day enrollment class size. Worst Rating!



$\hat{y} \approx 5$ **students**

If an instructor is rated a 10 for communication, 10 for attitude, and a 10 for punctuality, determine the 1st day enrollment class size. Best Rating!



$\hat{y} \approx 47$ **students**

Annual Income after 10 Years of Education and Other Variables

The following data was collected regarding the incomes after 10 years of formal education for people with carious traits.

Formal Education
Below High School = 0
High School=1
2-year degree or certificate=2
4- year degree=3
Graduate Degree=4

Sex
Male=1
Female=2

IQ Score
Intelligence Quotient

Birth Month
January=1
February=2
March=3
April=4
May=5
June=6
July=7
August=8
September=9
October=10
November=11
December=12

$$(x_1, x_2, x_3, x_4, y)$$

$$n = 18$$

Education x_1	**Sex** x_2	**IQ Score** x_3	**Birth Month** x_4	**Income** *Thousandths* y
0	1	102	10	45
2	2	110	5	68
4	2	125	11	152
3	1	118	8	138
1	1	90	4	52
0	2	82	1	38
1	2	92	9	65
3	2	122	6	145
2	1	96	12	74
2	2	112	3	82
4	1	130	9	165
4	2	128	8	172
0	2	80	7	45
0	1	75	5	48
3	1	110	2	128
3	2	118	1	138
1	1	96	12	75
2	2	112	6	100
0	1	68	4	33
4	1	136	10	188

In Data Analysis Pack with Microsoft Excel

Regression Statistics								
Multiple R	0.96546105							
R Square	0.932115039							
Adjusted R Square	0.914012383							
Standard Error	14.84570939							
Observations	20							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	4	45393.02369	11348.25592	51.49051213	1.38323E-08			
Residual	15	3305.926311	220.3950874					
Total	19	48698.95						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-3.186459163	38.73830743	-0.082256024	0.935530723	-85.75520693	79.38228861	-85.75520693	79.38228861
X Variable 1	25.82805071	6.183724181	4.176779229	0.000810034	12.64775462	39.00834681	12.64775462	39.00834681
X Variable 2	-5.343961989	7.209642814	-0.741224236	0.470003649	-20.71095189	10.02302791	-20.71095189	10.02302791
X Variable 3	0.560048509	0.487568282	1.148656566	0.268685747	-0.479178685	1.599275703	-0.479178685	1.599275703
X Variable 4	-0.071187222	1.072711422	-0.06636195	0.947966096	-2.357617494	2.21524305	-2.357617494	2.21524305

Notes about what values in these tables.

$\boldsymbol{R^2 \approx 0.932}$; **Very Good Fit.**

Adjusted $\boldsymbol{R^2 \approx 0.914}$; **Very Good Fit.**

Standard Error $\boldsymbol{\approx 18.846}$ The average distance the observed values fall from the regression line.

$\boldsymbol{F \approx 51.491}$ The overall F Statistic for the regression model. Calculated as regression MS/Residual MS.

Significance F$\boldsymbol{\approx 0.000}$ is the p-value**.** There is **statistically significant association with independent variables**.

Individual p values
The p value of the x1 variable (Education) is approximately 0.000 which tells us this variable is **statistically significant** at $\boldsymbol{\alpha = 5\%}$.

The **p value** of the x2 variable (Sex) is approximately 0.470 which tells us this variable is **not statistically significant** at $\boldsymbol{\alpha = 5\%}$.

The **p value** of the x3 variable (Q Score) is approximately 0.269 which tells us this variable is **not statistically significant** at $\boldsymbol{\alpha = 5\%}$. However, we may keep the variable when we rerun the experiment.

The **p value** of the x4 variable (Instructor Punctuality) is approximately 0.948 which tells us this variable is **not statistically significant** at $\boldsymbol{\alpha = 5\%}$.

The overall experiment is **statistically significant**, But let's exclude the x_4 variable due to the high p value and rerun the experiment.

Rerun after excluding the Birth Month x_4

Education x1	Sex x2	IQ Score x3	Income *Thousandths* y
0	1	102	45
2	2	110	68
4	2	125	152
3	1	118	138
1	1	90	52
0	2	82	38
1	2	92	65
3	2	122	145
2	1	96	74
2	2	112	82
4	1	130	165
4	2	128	172
0	2	80	45
0	1	75	48
3	1	110	128
3	2	118	138
1	1	96	75
2	2	112	100
0	1	68	33
4	1	136	188

(x_1, x_2, x_3, y)

$n = 18$

In Data Analysis Pack with Microsoft Excel

Regression Statistics								
Multiple R	0.965450728							
R Square	0.932095109							
Adjusted R Square	0.919362941							
Standard Error	14.37640626							
Observations	20							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	3	45392.05309	15130.68436	73.20789137	1.46297E-09			
Residual	16	3306.896911	206.6810569					
Total	19	48698.95						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-3.241450564	37.50512677	-0.086426866	0.932199438	-82.74876755	76.26586642	-82.74876755	76.26586642
X Variable 1	25.87509749	5.948759721	4.349662568	0.000496565	13.26429023	38.48590474	13.26429023	38.48590474
X Variable 2	-5.17962421	6.556956515	-0.789943352	0.44111178	-19.07975107	8.720502654	-19.07975107	8.720502654
X Variable 3	0.552849161	0.460319119	1.201012814	0.247224577	-0.42298378	1.528682101	-0.42298378	1.528682101

Notes about what values in these tables.

$\boldsymbol{R^2 \approx 0.932}$; **Very Good Fit.**

Adjusted $\boldsymbol{R^2 \approx 0.919}$; **Very Good Fit- Better than the previous results.**

Standard Error $\boldsymbol{\approx 18.846}$ The average distance the observed values fall from the regression line.

$\boldsymbol{F \approx 73.208}$ The overall F Statistic for the regression model. Calculated as regression MS/Residual MS.

Significance F$\boldsymbol{\approx 0.000}$ is the p-value**.** There is **statistically significant association with independent variables**.

Individual p values
The p value of the x1 variable (Education) is approximately 0.000 which tells us this variable is **statistically significant** at $\boldsymbol{\alpha = 5\%}$.

The **p value** of the x2 variable (Sex) is approximately 0.441 which tells us this variable is **not statistically significant** at $\boldsymbol{\alpha = 5\%}$.

The **p value** of the x3 variable (Q Score) is approximately 0.247 which tells us this variable is **not statistically significant** at $\boldsymbol{\alpha = 5\%}$. However, we may keep the variable when we rerun the experiment.

The overall experiment is **statistically significant**, But let's exclude the x_2 Sex variable due to the high p value and rerun the experiment.

Rerun after excluding the x_2 Sex Variable

Education x1	IQ Score x2	Income *Thousandths* y
0	102	45
2	110	68
4	125	152
3	118	138
1	90	52
0	82	38
1	92	65
3	122	145
2	96	74
2	112	82
4	130	165
4	128	172
0	80	45
0	75	48
3	110	128
3	118	138
1	96	75
2	112	100
0	68	33
4	136	188

(x_1, x_2, y)

$n = 18$

In Data Analysis Pack with Microsoft Excel

Regression Statistics								
Multiple R	0.964078199							
R Square	0.929446773							
Adjusted R Square	0.921146394							
Standard Error	14.21653565							
Observations	20							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	45263.08194	22631.54097	111.9764175	1.63077E-10			
Residual	17	3435.86806	202.1098859					
Total	19	48698.95						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-5.664431647	36.9638175	-0.153242604	0.88001125	-83.65126962	72.32240632	-83.65126962	72.32240632
X Variable 1	26.43446043	5.840783112	4.525841813	0.000298782	14.11148523	38.75743563	14.11148523	38.75743563
X Variable 2	0.491600702	0.448696421	1.095619843	0.288524704	-0.455065997	1.438267401	-0.455065997	1.438267401

Notes about what values in these tables.

$\boldsymbol{R^2 \approx 0.964}$; **Very Good Fit.**

Adjusted $\boldsymbol{R^2 \approx 0.929}$; **Very Good Fit- Better than the previous results.**

Standard Error $\boldsymbol{\approx 14.217}$ The average distance the observed values fall from the regression line.

$\boldsymbol{F \approx 111.976}$ The overall F Statistic for the regression model. Calculated as regression MS/Residual MS.

Significance F$\boldsymbol{\approx 0.000}$ is the p-value**.** There is **statistically significant association with independent variables**.

Individual p values
The p value of the x1 variable (Education) is approximately 0.000 which tells us this variable is **statistically significant** at $\boldsymbol{\alpha = 5\%}$.

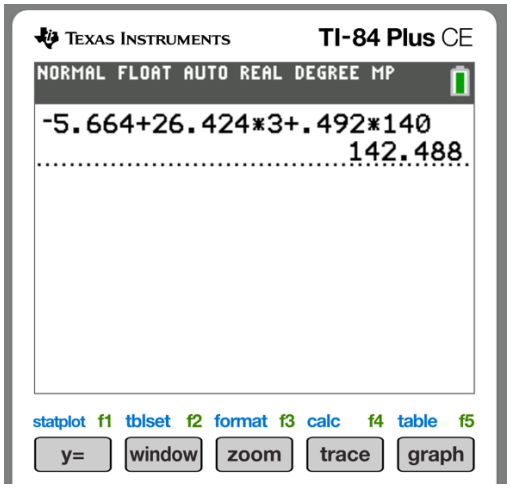
The **p value** of the x2 variable (IQ Score) is approximately 0.289 which tells us this variable is **not statistically significant** at $\boldsymbol{\alpha = 5\%}$.

The overall experiment is **statistically significant**, We are now able to determine the **Multi Regression Equation**.

Multi Regression Equation

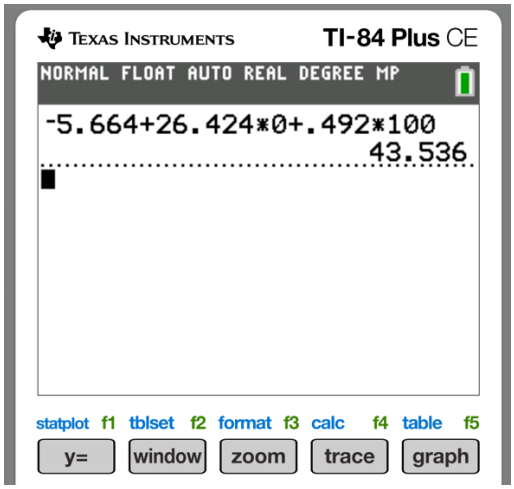
$$\hat{y} = b_0 + b_1 x + b_2 x$$
$$\boldsymbol{\hat{y} = -5.664 + 26.424 \cdot Education + 0.492 \cdot IQ\ Score}$$

If a person has a 4-year college degree and an IQ Score of 140, how much will they make after 10 years of graduation?



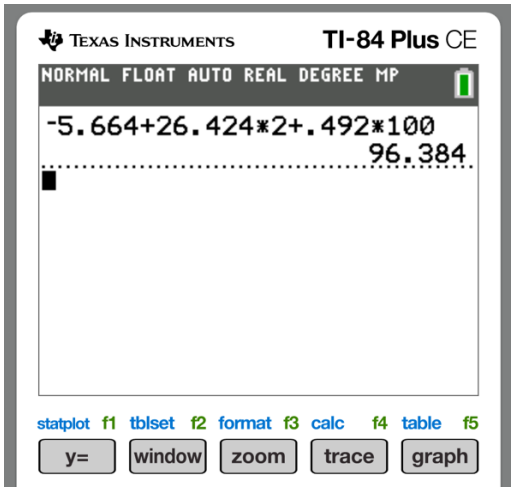
$\hat{y} \approx \$142,000$

If a person has less than a high school degree and an average IQ Score of 100, how much will they make after 10 years of graduation?



$\hat{y} \approx \$44,000$

If a person has less than a 2-year degree and an average IQ Score of 100, how much will they make after 10 years of graduation?



$\hat{y} \approx \$96,000$