

provided.

[2 points] 1 of 11 ID: MST.SLR.AV.06.0010a



MoonDollars is a coffee franchise that has many coffee stores all over the world. MoonDollars coffee is investigating a linear regression model to predict the annual income for a store based on the population of the city where the store is located. In the proposed regression model, annual store income is the response variable and population of the city is the explanatory variable.

A random sample of 20 stores is selected and measurements are observed.

Point, Population, Annual income,
1, 1460, 618,
2, 410, 88,
3, 740, 136,
4, 2830, 1941,
5, 1290, 1022,
6, 1800, 2148,
7, 2680, 2232,
8, 460, 872,
9, 1910, 1531,
10, 2620, 2999,
11, 520, -535,
12, 2020, 1520,
13, 1070, 1675,
14, 1990, 1455,
15, 1150, 1094,
16, 1110, 1528,
17, 1730, 544,
18, 2670, 2519,
19, 2600, 2274,
20, 2740, 1713

[Download the data](#)

Population ('000s)	Annual income ('000s \$)
1,460	618
410	88
740	136

2,830	1,941
1,290	1,022
1,800	2,148
2,680	2,232
460	872
1,910	1,531
2,620	2,999
520	-535
2,020	1,520
1,070	1,675
1,990	1,455
1,150	1,094
1,110	1,528
1,730	544
2,670	2,519
2,600	2,274
2,740	1,713

a) Calculate the point prediction for the value $x = 1,000$. Give your answer as a whole number.

$\hat{y} =$

b) Give the 95% prediction interval for the value $x = 1,000$. Give your answers as whole numbers.

$\leq \hat{y} \leq$

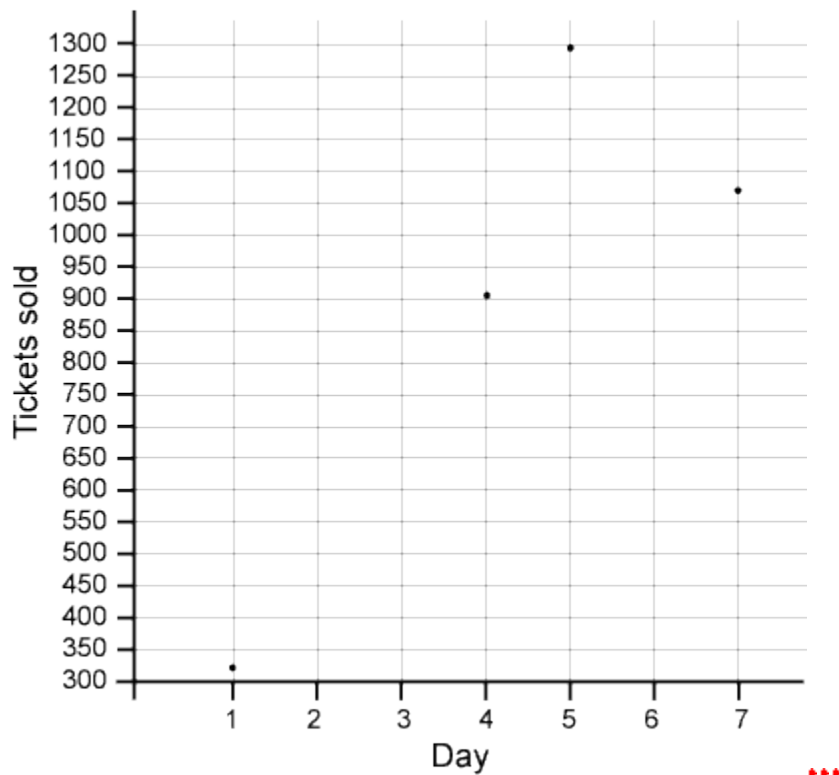
[1 point] 2 of 11 ID: MST.DCP.GRR.03.0040

A cinema manager wishes to investigate the pattern of ticket sales over the different days of the week. Starting on a Monday, she records the number of tickets sold each day for one week. The results are recorded below

Day	Tickets sold
1	321
2	545
3	412
4	901

5	1290
6	1215
7	1063

In the time-series plot below, only four of the values have been marked. Drag the three red markers onto the plot to complete it and represent the data for all seven days.



[1 point] 3 of 11 ID: MST.SLR.AL.R.02.0020

A study was conducted into the relationship between the age in years of a person and their average weekly after tax income.

You have been supplied with the following information regarding a regression model that had been developed as part of the study.

Regression analysis	
sample intercept	513.64
sample slope	2.89

A particular observed value from the data is:

$$x = 37, y = 608.18$$

Calculate the residual of this observed value of the response variable. Give your answer to 2 decimal places.

residual =

[1 point] 4 of 11 ID: MST.SLR.REE.06.0020

The following table shows the average petrol price and the number of online shopping orders over a given month:

Petrol Price and Online Shopping	
Average Petrol Price per gallon over a month (\$)	Number of Online Shopping Orders
4.89	4,098
2.2375	2,533
1.3725	1,488
1.9975	2,070
3.055	2,478
4.6125	3,819
4.635	3,935
2.9875	2,490
3.975	3,372
3.57	2,936

The relationship between the average petrol price and the number of online shopping orders in a given month is proposed to follow the simple regression equation below: [show variables](#)

$\hat{y}$ = predicted response variable

x = explanatory variable

b_0 = sample intercept

b_1 = sample slope

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

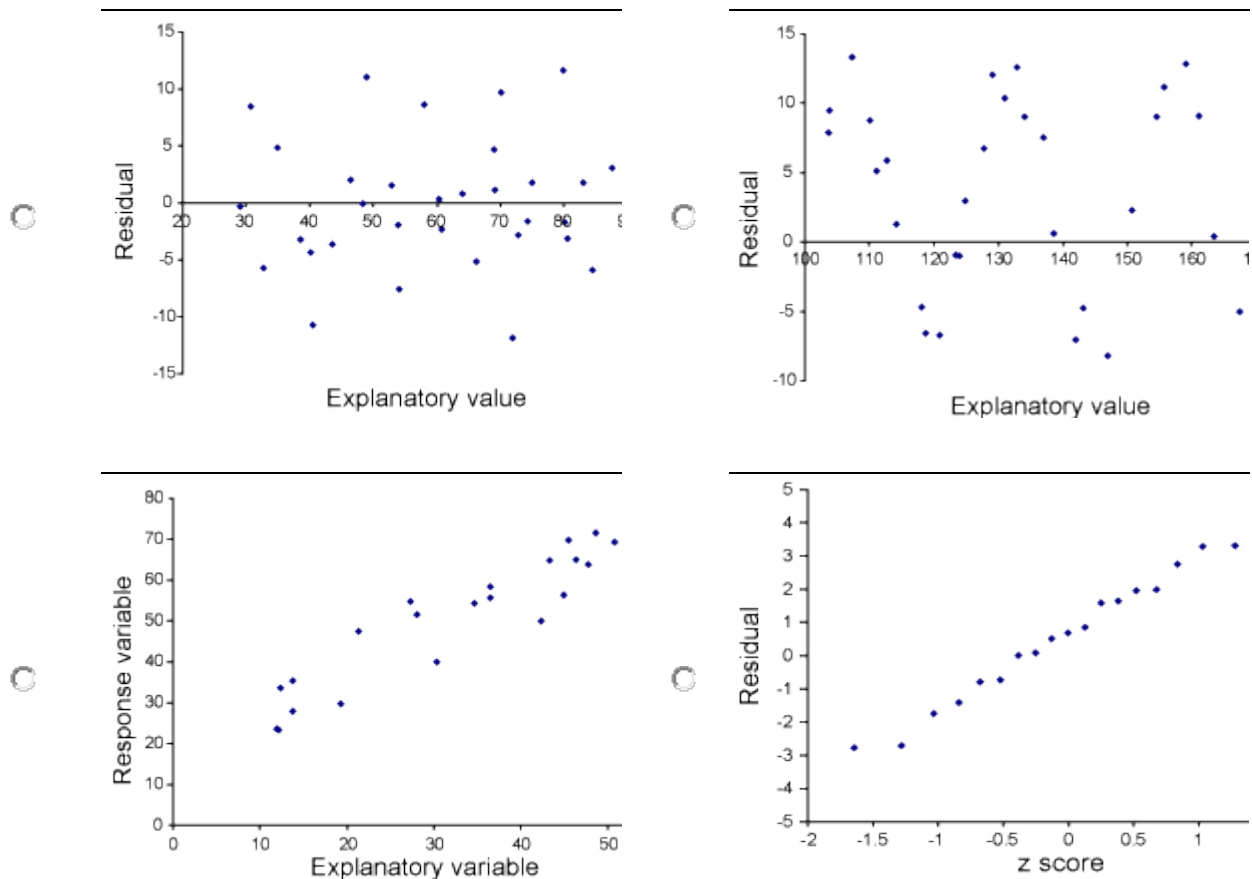
Calculate the percentage of variability in the number of online shopping orders that is **not** explained by the average petrol price. Give your answer as a percentage to 1 decimal place.

Percentage = %

[1 point] 5 of 11 ID: MST.SLR.ALR.01.0030c.STAT301

You have constructed a simple linear regression model and are testing whether the assumption of normality of the residuals is reasonably satisfied.

Select the scatter plot that indicates normality of the residuals:



[2 points] 6 of 11 ID: MST.SLR.REE.04.0030

Two variables, x and y , are investigated for an association and are found to be strongly correlated with positive correlation. A simple linear regression model is constructed. The regression model can explain 91.3% of variation in the values of y based upon the values of x .

Select all the statements that can be reasonably drawn from this analysis:

- ☐ x and y are causally related. Furthermore, due to the positive correlation, larger values in x cause larger values of y to occur
- ☐ x and y have a strong association. Furthermore, larger values of x tend to occur with larger values of y .
- ☐ small values of x tend to occur with large values of y
- ☐ small values of x tend to occur with small values of y
- ☐ the value of y tends to be greater than the value of x
- ☐ if x is small then it is impossible for y to be very large
- ☐ the value of x tends to be greater than the value of y

[1 point] 7 of 11 ID: MST.NDM.CC.08.0010



An investigation has been conducted to determine whether there is a relationship between the salary paid to a Chief Executive Officer (CEO) and the productivity of that CEO as measured by the change in profits from the time the CEO was employed.

The scatter plot plots the salary paid (x) against the change in profits (y) of a sample of CEOs. Without doing any calculations and according only to this scatter diagram, a reasonable coefficient of correlation (r) between x and y would be:

- ☐ $r = 0.33$
- ☐ $r = -0.68$
- ☐ $r = 2.23$
- ☐ $r = -1.26$

[3 points] 8 of 11 ID: MST.SLR.REE.04.0020b

The regression equation:

$$\hat{y} = 5 + 1.2x$$

was calculated from a sample. It is part of a regression model that has been developed in order to predict the score in an end-of-year exam based on the score in a mid-year exam for a particular university course. In the sample, mid-year scores ranged from 50 to 80.

Select whether or not each of the following conclusions are correct from the regression analysis:

		Correct	Not correct
a)	For an increase by one in the mid-year score, the predicted increase in end-of-year score is 1.2.	☐	☐
b)	If a student achieves a score of 45 in the mid-year exam then we know that the student will achieve a score of 59 in the end-of-year exam.	☐	☐
c)	If a student achieves a score of 60 in the mid-year exam then we know that the student will achieve a score of 77 in the end-of-year exam.	☐	☐

[1 point] 9 of 11 ID: MST.SLR.TM.01.0030

It has been hypothesised that there is a linear relationship between the average annual interest rate and the volume of car sales, and that relationship can be represented as follows: [show variables](#)

$\hat{y}$ = volume of car sales predicted by regression model (millions)

x = average annual interest rate (%)

b_0 = sample intercept

b_1 = sample slope

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

The following table lists the average interest rates recorded in a particular year and the corresponding number of car sales:

Interest Rates and Car Sales	
Interest Rate (% per annum)	Car Sales Volume (millions)
4.8	3.258
5.9	3.564
3.4	4.304
8.2	2.022
7.8	1.898
8.1	2.366
6	3.01
8.5	2
6.2	3.152
4.6	3.956

Using the data provided, and at $\alpha = 0.05$, the null hypothesis that a significant linear relationship between the average annual interest rate and the volume of car sales does **not** exist is .

[6 points] 10 of 11 ID: MST.SLR.REE.03.0040

A biologist is studying the levels of heavy metal contaminants among a population of the South Nakaratuan Chubby Bat. The biologist is interested in constructing a simple linear regression model to investigate the relationship between weight of an animal and the level of heavy metal contamination. In the proposed regression model the level of contaminant is the response variable and weight is the explanatory variable. The contaminant level is measured in parts per billion (ppb) and weight in grams.

A random sample of 20 individuals is selected and measurements are taken.

Point, Contaminant level, Weight,
 1, 108, 128,
 2, 76, 121,
 3, 83, 104,
 4, 118, 144,
 5, 82, 111,
 6, 130, 140,
 7, 110, 136,
 8, 81, 103,
 9, 113, 138,

10,76,108,
11,103,130,
12,89,106,
13,70,109,
14,109,124,
15,66,104,
16,97,122,
17,113,138,
18,114,121,
19,107,121,
20,176,104

[Download the data](#)

Contaminant study in South Nakaratuan Chubby Bat	
Contaminant level (ppb)	Weight (g)
108	128
76	121
83	104
118	144
82	111
130	140
110	136
81	103
113	138
76	108
103	130
89	106
70	109
109	124
66	104
97	122
113	138
114	121
107	121
176	104

Plotting the data, the researcher notices an obvious outlier. They decide to do the regression analysis with and without the outlier and compare the results.

Calculate the slope (b_1) and intercept (b_0) of the simple regression equation using the data provided. Give your answers to 2 decimal places.

a) Slope = b_1 =

b) Intercept = b_0 =

Find the proportion of variation in the values of contaminant level that is explained by the regression model. Give your answer as a decimal to 2 decimal places.

c) R^2 =

Repeat this process omitting the outlier:

d) Slope = b_1 =

e) Intercept = b_0 =

Find the proportion of variation in the values of contaminant level that is explained by the regression model. Give your answer as a decimal to 2 decimal places.

f) R^2 =

[2 points] 11 of 11 ID: MST.SLR.REE.03.0030

A study is conducted to determine the simple linear regression relation (if any) between average temperature over a week and ice-cream sales in a particular city. The following table provides the regression data required to construct the model:

Average Temperature and Ice-cream Sales	
Average Temperature over a Week (°F)	Weekly Ice-cream Sales (\$'000s)
77	205.229
85	220.895
39	129.593
68	187.391
89	229.703
91	232.347
61	173.527
38	128.631
46	143.537
42	136.589
88	227.441

94	238.713
----	---------

Calculate the slope (b_1) and intercept (b_0) of the simple regression equation using the data provided. Give your answers to 2 decimal places.

(a) Slope = b_1 =

(b) Intercept = b_0 =

homework.

[2 points] 1 of 7 ID: MST.MR.TM.05.0030

A regression model is constructed with the goal of predicting the number of motor vehicle accidents in a city per year based upon the population of the city, the number of recorded traffic offenses per year, the number of vehicles per capita in the city and the average annual temperature in the town. A random sample of 50 cities were studied for this purpose.

Here is an analysis output on the regression model:

ANOVA					
	DF	SS	MS	F	Probability
Regression	4	174.555	43.63875	9.83869129...	< 0.001
Residual	45	199.594	4.43542222...		
Total	49	374.149			

Regression analysis	
R^2	0.46653873...
s	2.10604421...

Regression coefficients				
	Estimate	Standard Error	t	Probability
Intercept	16.90	4.899	3.44968361...	0.00123025...
Population of city	1.879	0.2272	8.27024648...	< 0.001
No. of vehicles per capita	0.880	0.2184	4.02930403...	< 0.001
No. of traffic offenses	0.935	0.2860	3.26923077...	0.0020711...
Average annual temp.	0.283	0.4948	0.57194826...	0.57020328...

a) At a level of significance of 0.05, the result of the F test for this model is that the null hypothesis

rejected.

b) Suppose you are going to construct a new model by removing the most insignificant variable. You would first remove:

- ☐ population of city
- ☐ no. of vehicles per capita
- ☐ no. of traffic offenses
- ☐ average annual temp.

[3 points] 2 of 7 ID: MST.MR.MRE.04.0010

IHI Insurance have conducted a multiple linear regression analysis to predict the value of the number of annual claims incurred by a motor insurance policyholder (y) based on the age of the policyholder in years (x_1) and the *value of the car insured in thousands of dollars* (x_2). The analysis was based on a random sample of 500 policyholders. The ages in the sample ranged from 16 to 70 and the values of the cars in the sample ranged from \$1,100 to \$52,000. The multiple linear regression equation corresponding to IHI's analysis is: $\hat{y}_i = 13 + (-0.5)x_{1i} + (0.1)x_{2i}$.

Select whether the following statements regarding the multiple linear regression equation are true or false:

		True	False
a)	Holding x_1 constant, a one thousand dollar increase in x_2 will result in a decrease of 0.5 in the predicted value of y .	☐	☐
b)	A 22 year old policyholder with a car worth \$20,000 is predicted to make 4 claims.	☐	☐
c)	Holding x_2 constant, a one year increase in x_1 will result in an increase of 0.1 in the predicted value of Y .	☐	☐

[1 point] 3 of 7 ID: MST.MR.MRE.01.0010b.STAT301

A multiple linear regression model is to be constructed to determine if there is a relationship between a dependent variable (y) and two independent variables (x_1 and x_2). A random sample of size n has been collected and the values of x_{1i} , x_{2i} and y_i for $i = 1, 2, \dots, n$ have been recorded. The residuals (e_i) in this analysis are defined as the difference between the observed values of y and the values of y predicted by the regression equation.

Select the condition that is one of the assumptions of a valid multiple linear regression model:

- ☐ the relationship between the dependent and independent variables is quadratic
- ☐ the independent variables are independent of the dependent variable
- ☐ the residuals are constant

☐ the residuals have a constant variance

[3 points] 4 of 7 ID: MST.MR.TM.04.0010a

A multiple regression model is developed in order to study the relationship between the number of power outages that occur in a rural town per year (y) and two independent variables: the distance of that town from a major metropolitan area (x_1) and the amount spent on utilities each year for that town (x_2).

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

A sample of 50 rural towns is collected. For each town, the distance (in hundreds of miles) of the town from a major metropolitan area and the amount spent (in hundreds of thousands of dollars) on utilities that year are recorded. Also, the number of reported power outages is recorded. The following regression equation was developed:

$$\hat{y} = 10 + 1.2x_1 - 2.7x_2$$

A t test is to be conducted in order to assess the significance of distance from a major metropolitan area in predicting number of power outages. Select whether the following statements about the test are true or false:

		True	False
a)	The test will assess whether β_1 is zero.	☐	☐
b)	Assuming the null hypothesis is true, the test statistic follows the t distribution with 48 degrees of freedom.	☐	☐
c)	The test is two-tailed.	☐	☐

[14 points] 5 of 7 ID: MST.MR.CM.01.0060

Joe Smith is considering opening a franchise of a very popular gourmet fruit drink store. He is interested in constructing a multiple regression model to predict the number of drinks sold in a day based upon the volume of foot traffic passing the store, the daily maximum temperature and the average drink price.

Information has been collected on 30 randomly selected days for several randomly chosen stores: [show data](#)

Point, No. drinks sold, Vol. traffic, Temp, Price,
1,690,413,65,2.3,
2,870,442,108,3.65,
3,499,240,104,3.15,
4,726,311,108,3.4,
5,773,437,108,2.85,
6,571,388,85,3.5,
7,519,276,74,1.8,
8,470,209,107,2.2,
9,507,251,109,2.65,
10,753,404,94,2.55,

11,737,363,107,3.15,
 12,373,182,63,2.25,
 13,733,333,110,3.45,
 14,718,407,68,2.45,
 15,692,315,95,3.05,
 16,479,201,80,2.95,
 17,701,443,65,2.15,
 18,931,499,94,2,
 19,684,367,70,2.05,
 20,845,490,95,2.95,
 21,807,453,61,1.65,
 22,887,441,103,3.6,
 23,893,480,105,3.9,
 24,710,453,66,2.3,
 25,523,207,94,2.55,
 26,730,349,108,3.35,
 27,678,324,93,4,
 28,482,303,86,3.75,
 29,541,187,113,2.45,
 30,872,498,110,2.1

Number of drinks sold	Volume of traffic (people per day)	Daily max temp °F	Average drink price
690	413	65	2.3
870	442	108	3.65
499	240	104	3.15
726	311	108	3.4
773	437	108	2.85
571	388	85	3.5
519	276	74	1.8
470	209	107	2.2
507	251	109	2.65
753	404	94	2.55
737	363	107	3.15
373	182	63	2.25
733	333	110	3.45
718	407	68	2.45
692	315	95	3.05
479	201	80	2.95
701	443	65	2.15
931	499	94	2
684	367	70	2.05

845	490	95	2.95
807	453	61	1.65
887	441	103	3.6
893	480	105	3.9
710	453	66	2.3
523	207	94	2.55
730	349	108	3.35
678	324	93	4
482	303	86	3.75
541	187	113	2.45
872	498	110	2.1

a) Find the multiple regression equation using all three explanatory variables. Assume that x_1 is volume of traffic, x_2 is maximum daily temperature and x_3 is average drink price. Give your answers to 3 decimal places.

$$\hat{y} = \boxed{} + \boxed{} \text{ vol. traffic} + \boxed{} \text{ max temp} + \boxed{} \text{ average drink price}$$

b) At a level of significance of 0.05, the result of the F test for this model is that the null hypothesis

rejected.

c) The value of R^2 for this model, to 3 decimal places, is equal to

d) The value of s for this model, to 3 decimal places, is equal to

e) The least significant explanatory variable in this model is:

- ☐ vol. traffic
☐ max temp
☐ average drink price

f) Construct a new multiple regression model by removing the variable *average drink price*. Give your answers to 3 decimal places.

The new regression model equation is:

$$\hat{y} = \boxed{} + \boxed{} \text{ vol. traffic} + \boxed{} \text{ max temp}$$

g) In the new model compared to the previous one, the value of R^2 (to 3 decimal places) is:

- ☐ increased
☐ decreased
☐ unchanged

h) In the new model compared to the previous one, the value of s (to 3 decimal places) is:

- ☐ increased
- ☐ decreased
- ☐ unchanged

i) The better model is the:

- ☐ original model
- ☐ reduced model

[1 point] 6 of 7 ID: MST.MR.TM.02.0030

Man Size has constructed a multiple regression model and has conducted an F test of the model at a level of significance of 0.05. The model has four independent variables. The result of the F test was that the null hypothesis was rejected. Select the appropriate conclusion that can be drawn:

- ☐ at least one of the regression coefficients is non-zero
- ☐ all of the regression coefficients are non-zero
- ☐ all of the regression coefficients are zero
- ☐ the dependent variable is not related to any of the independent variables
- ☐ exactly one of the regression coefficients is non-zero

[2 points] 7 of 7 ID: MST.MR.TM.03.0040.STAT301

Select all the scatter plots that indicate a violation of one or more of the assumptions of linear regression:

