Section 2.2 – Least–Squares Regression

Basic Concept of Regression

Two variables are sometimes related in a ***deterministic*** way, meaning that given a value for one variable, the value of the other variable is exactly determined from a given equation.
The regression equation expresses a relationship between x (called the ***explanatory*** variable, ***predictor*** variable or ***independent*** variable), and y (called the ***response*** variable or ***dependent*** variable).

The typical equation of a straight line
$y = mx + b$ is expressed in the form
$\hat{y} = b_0 + b_1 x$, where b_0 is the y-intercept and b_1 is the slope.

Definitions

- **Regression Equation**
  Given a collection of paired data, the regression equation algebraically describes the relationship between the two variables.

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

- **Regression Line**
  The graph of the regression equation is called the regression line (or line of best fit, or least squares line).

$$b_1 = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} \qquad b_0 = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

Notation for Regression Equation

	Population Parameter	***Sample Statistic***
Slope of regression equation	β_1	$b_1 = r \cdot \frac{s_y}{s_x}$
y-intercept of regression equation	β_0	$b_0 = \bar{y} - b_1 x$
Equation of the regression line	$y = \beta_0 + \beta_1 x$	$\hat{y} = b_0 + b_1 x$

Requirements

1. The sample of paired (x, y) data is a random sample of quantitative data.
2. Visual examination of the scatterplot shows that the points approximate a straight-line pattern.
3. Any outliers must be removed if they are known to be errors. Consider the effects of any outliers that are not known errors.

Formulas for b_0 and b_1

Slope: $b_1 = r\dfrac{s_y}{s_x}$

y-intercept: $b_0 = \bar{y} - b_1\bar{x}$

Where r is the linear correlation coefficient,

s_y is the standard deviation of the y values, and

s_x is the standard deviation of the x values

***Special Property*:** The regression line fits the sample points best.

Rounding the *y*-intercept b_0 and the Slope b_1

Example

Using the following sample data

x	0	2	3	5	6	6
y	5.8	5.7	5.2	2.8	1.9	2.2

a) Find a linear equation that relates x and y by selecting 2 points and finding the equation of the line containing the points.

b) Graph the equation on the scatter diagram

c) Use the equation to predict y if $x = 3$

Solution

a) Using the 2 points (2, 5.7) and (6, 1.9)

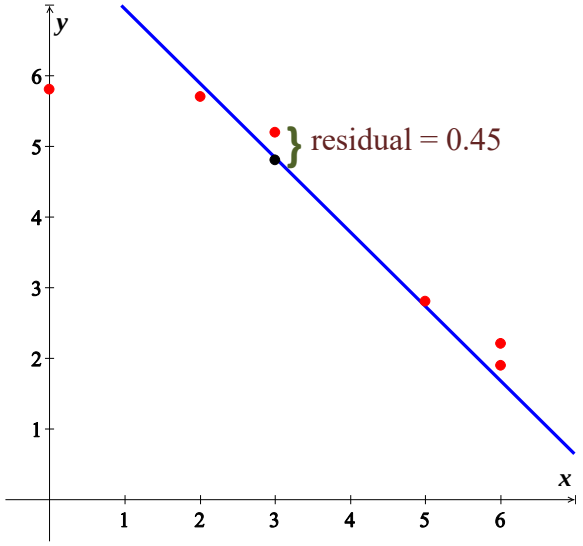
$$m = \frac{5.7 - 1.9}{2 - 6} = -0.95$$

$$y = m(x - x_1) + y_1$$

$$\underline{y = -0.95(x - 2) + 5.7 \ = -0.95x + 7.6}$$

b) Graph →

c) $\underline{y = -0.95(3) + 7.6 \ = 4.75}$



The difference between the observed value of y and the predicted value of y is the error, or ***residual***.

Using the line from the last example, and the predicted value at $x = 3$:

residual = observed y – predicted y

= 5.2 – 4.75

= 0.45

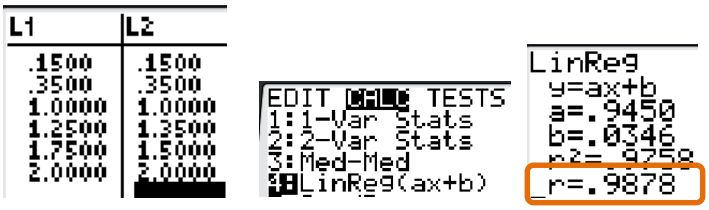
Example

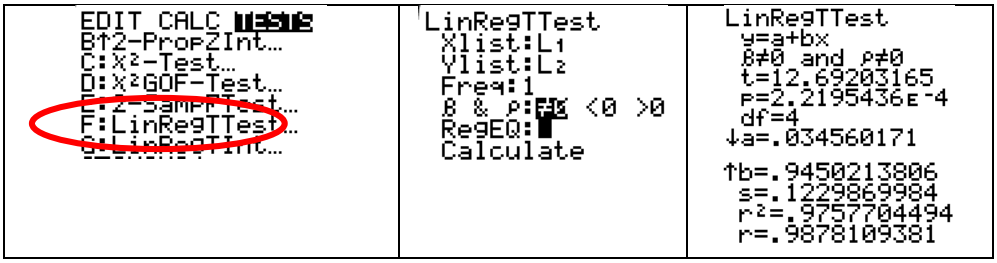
Using the pizza subway fare costs in Table below, Use technology to find the equation of the regression line in which the explanatory variable (or *x* variable) is the cost of a slice of pizza and the response variable (or *y* variable) is the corresponding cost of a subway fare. What proportion of the variation in the subway fare can be explained by the variation in the costs of a slice of pizza?

Table *– Cost of a Slice of Pizza, subway Fare, and the CPI*						
Year	**1960**	**1973**	**1986**	**1995**	**2002**	**2003**
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00
Subway Fare	0.15	0.35	1.00	1.35	1.50	2.00
CPI	30.2	48.3	112.3	162.2	191.9	197.8

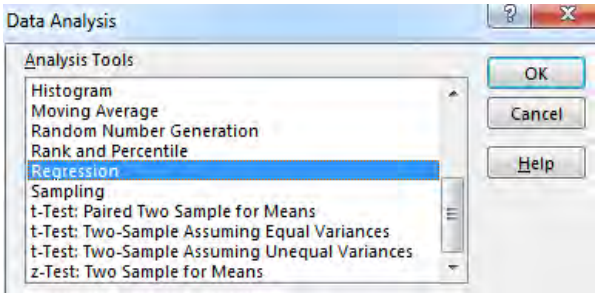
Solution

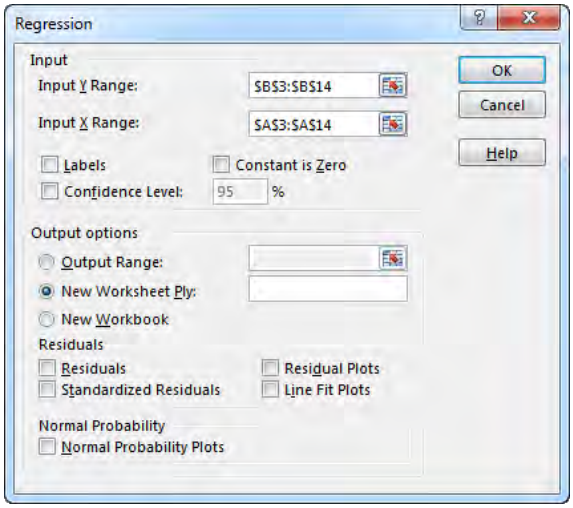
Requirements are satisfied: simple random sample; scatterplot approximates a straight line; no outliers
Here are results from four different technologies





Excel





SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9878109
R Square	0.9757704
Adjusted R	0.9697131
Standard E	0.122987
Observatic	6

ANOVA

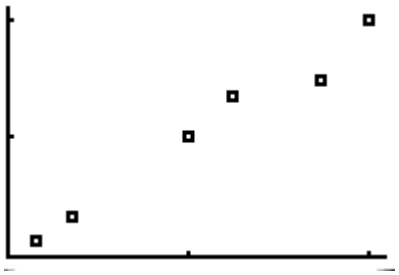
	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	2.436580126	2.4365801	161.08767	0.000222
Residual	4	0.060503207	0.0151258		
Total	5	2.497083333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Jpper 95.0%*
Intercept	0.0345602	0.095012806	0.3637422	0.7344608	-0.229238	0.298358	-0.229238	0.298358
X Variable	0.9450214	0.074457849	12.692032	0.000222	0.7382932	1.1517495	0.7382932	1.1517495

Example

Graph the regression equation $\hat{y} = 0.0346 + 0.945x$ (from the preceding Example) on the scatterplot of the pizza/subway fare data and examine the graph to subjectively determine how well the regression line fits the data.

Solution



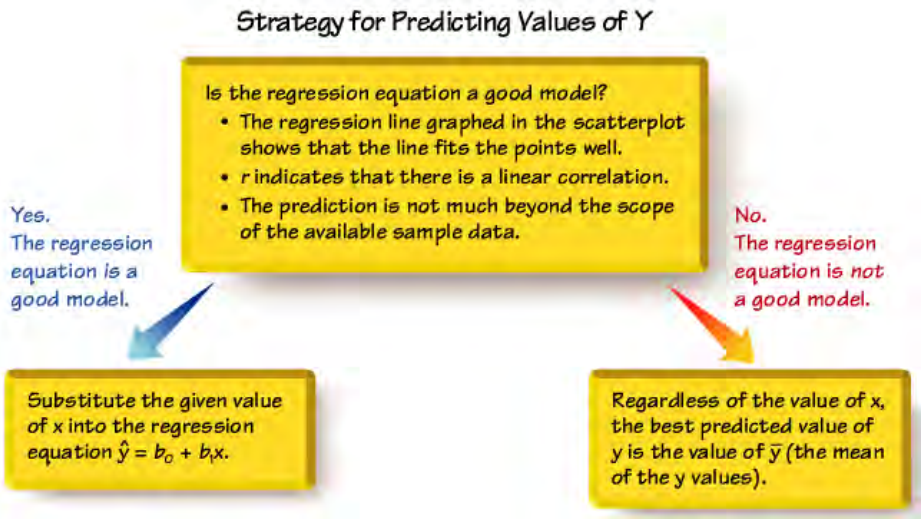
Using the Regression Equation for Predictions

1. Use the regression equation for predictions only if the graph of the regression line on the scatterplot confirms that the regression line fits the points reasonably well.
2. Use the regression equation for predictions only if the linear correlation coefficient *r* indicates that there is a linear correlation between the two variables.
3. Use the regression line for predictions only if the data do not go much beyond the scope of the available sample data. (Predicting too far beyond the scope of the available sample data is called *extrapolation*, and it could result in bad predictions.)
4. If the regression equation does not appear to be useful for making predictions, the best predicted value of a variable is its point estimate, which is its sample mean.

If the regression equation is not a good model, the best predicted value of y is simply y, the mean of the y values.

Remember, this strategy applies to linear patterns of points in a scatterplot.

If the scatterplot shows a pattern that is not a straight-line pattern, other methods apply.



Definitions

In working with two variables related by a regression equation, the marginal change in a variable is the amount that it changes when the other variable changes by exactly one unit. The slope b_1 in the regression equation represents the marginal change in y that occurs when x changes by one unit.

In a scatterplot, an outlier is a point lying far away from the other data points.

Paired sample data may include one or more influential points, which are points that strongly affect the graph of the regression line.

Beyond the Basics of Regression

Definitions

In working with two variables related by a regression equation, the marginal change in a variable is the amount that it changes when the other variable changes by exactly one unit. The slope b_1 in the regression equation represents the marginal change in y that occurs when x changes by one unit.

In a scatterplot, an ***outlier*** is a point lying far away from the other data points. Paired sample data include one or more ***influential points***, which are points that strongly affect the graph of the regression line.

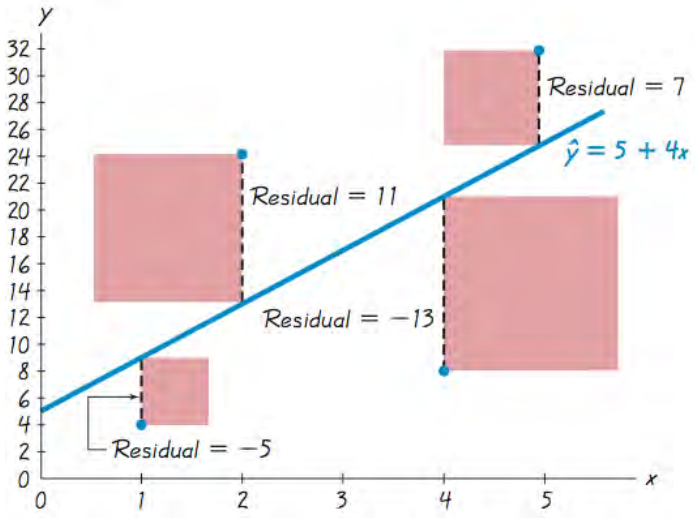
Residuals and the Least-Squares Property

Definition

For a pair of sample x and y values, the residual is the difference between the *observed* sample value of y and the y-value that is *predicted* by using the regression equation. That is,

$$residual = observed\ y - predicted\ y = y - \hat{y}$$

Residuals



Definitions

A straight line satisfies the least-squares property if the sum of the squares of the residuals is the smallest sum possible.

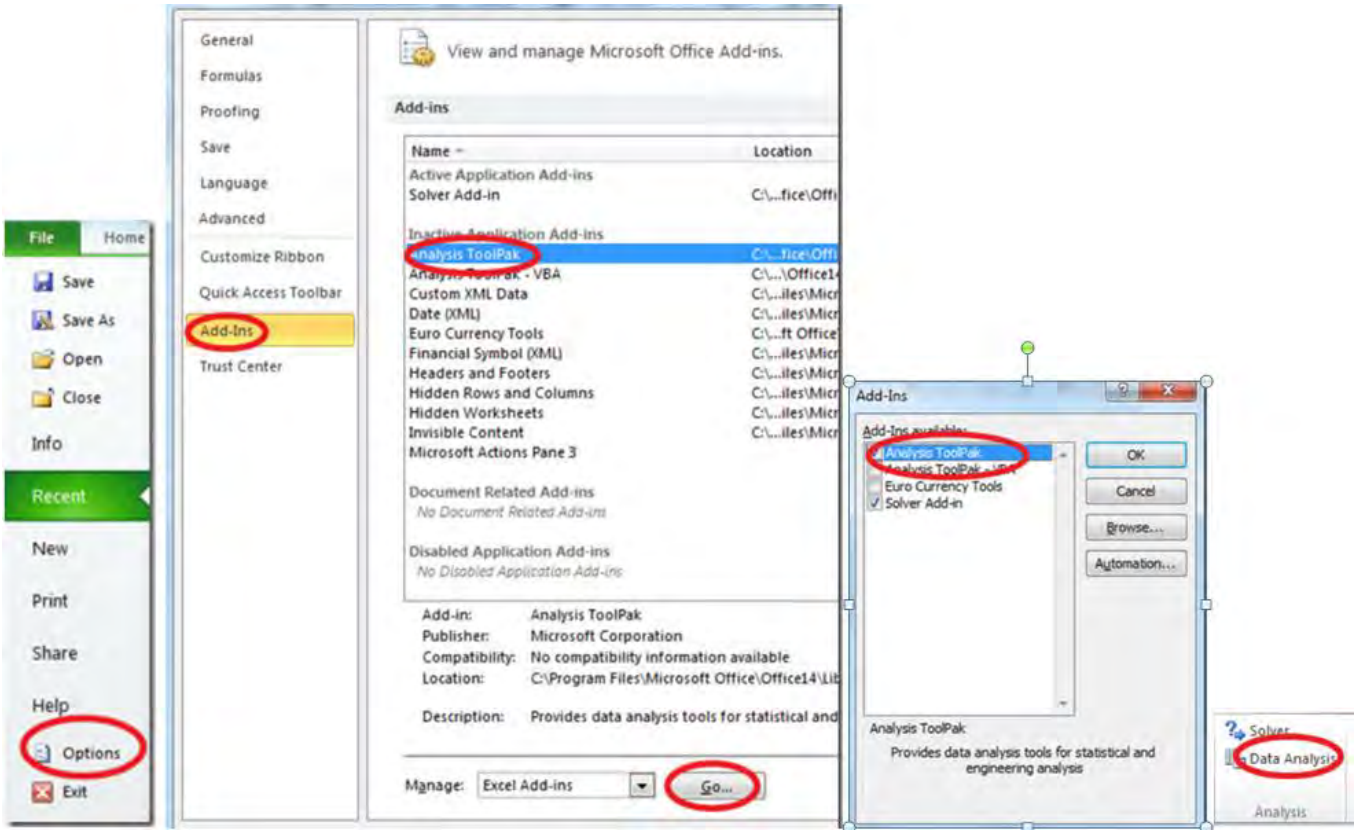
A ***residual plot*** is a scatterplot of the (x, y) values after each of the y-coordinate values has been replaced by the residual value $y - \hat{y}$ (where y denotes the predicted value of y). That is, a residual plot is a graph of the points $(x,\ y - \hat{y})$.

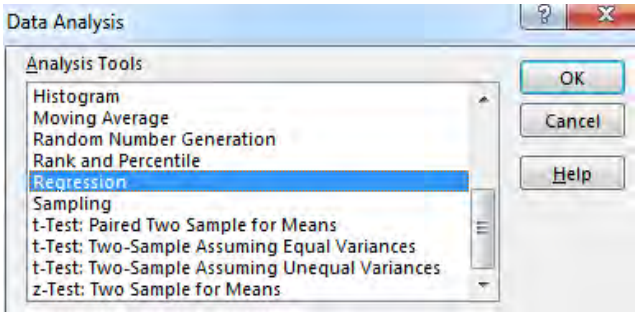
Residual Plot Analysis

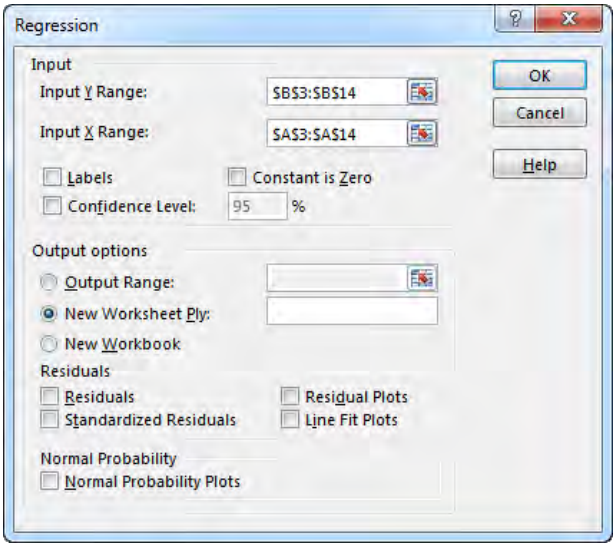
When analyzing a residual plot, look for a pattern in the way the points are configured, and use these criteria:

- The residual plot should not have an obvious pattern that is not a straight-line pattern.
- The residual plot should not become thicker (or thinner) when viewed from left to right.

Installation analysis package to use regression







Exercises *Section* 2.2 – Least–Squares Regression

1. A physician measured the weights and cholesterol levels of a random sample of men. The regression equation is $\hat{y} = -116 + 2.44x$, where x represents weight (in pounds). What does the symbol $\hat{y}$ represent? What does the predictor variable represent? What does the response variable represent?

2. In what sense is the regression line the straight line that "best" fits the points in a scatterplot?

3. In a study, the total weight (in pounds) of garbage discarded in one week and the household size were recorded for 62 households. The linear correlation coefficient is $r = 0.759$ and the regression equation $\hat{y} = 0.445 + 0.119x$, where x represents the total weight of discarded garbage. The mean of the 62 garbage weights is 27.4 lb. and the 62 households have a mean size of 3.71 people. What is the best predicted number of people in a household that discards 50 lb. of garbage?

4. A sample of 8 mother/daughter pairs of subjects was obtained, and their heights (in inches) were measured. The linear correlation coefficient is 0.693 and the regression equation $\hat{y} = 69 - 0.0849x$, where x represents the height of the mother. The mean height of the mothers is 63.1 in. and the mean height of the daughters is 63.3 in. Find the best predicted height of a daughter given that the mother has a height of 60 in.

5. A sample of 40 women is obtained, and their heights (in inches) and pulse rates (in beats per minute) are measured. The linear correlation coefficient is 0.202 and the equation of the regression line is $\hat{y} = 18.2 + 0.920x$, where x represents height. The mean of the 40 heights is 63.2 in. and the mean of the 40 pulse rates is 76.3 beats per minute. Find the best predicted pulse rate of a woman who is 70 in. tall.

6. Heights (in inches) and weights (in pounds) are obtained from a random sample of 9 supermodels. The linear correlation coefficient is 0.360 and the equation of the regression line is $\hat{y} = 31.8 + 1.23x$, where x represents height. The mean of the 9 heights is 69.3 in. and the mean of the 9 weights is 117 lb. Find the best predicted weight of a supermodel with a height of 72 in.?

7. Find the equation of the regression line for the given data below

x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74

Examine the scatterplot and identify a characteristic of the data that is ignored by the regression line

8. Find the equation of the regression line for the given data below

x	10	8	13	9	11	14	6	4	12	7	5
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73

Examine the scatterplot and identify a characteristic of the data that is ignored by the regression line

9. Find the equation of the regression line for the given data below

CPI	30.2	48.3	112.3	162.2	191.9	197.8
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00

Let the first variable be the predictor (x) variable. Find the best indicated predicted cost of a slice of pizza when the Consumer Price Index (CPI) is 182.5 (in the year 2000).

10. Find the equation of the regression line for the given data below

CPI	30.2	48.3	112.3	162.2	191.9	197.8
Subway fare	0.15	0.35	1.00	1.35	1.5	2.00

Let the first variable be the predictor (x) variable. Find the best indicated predicted cost of a slice of pizza when the Consumer Price Index (CPI) is 182.5 (in the year 2000).

11. Listed below are systolic blood pressure measurements (in mm *HG*) obtained from the same woman.

Right Arm	102	101	94	79	79
Left Arm	175	169	182	146	144

Find the best predicted systolic blood pressure in the left arm given that the systolic blood pressure in the right arm is 100 mm Hg.

12. Find the best predicted height of runner-up Goldwater, given that the height of the winning presidential candidate is 75 in. Is the predicted height of Goldwater close to his actual height of 72 in.?

Winner	69.5	73	73	74	74.5	74.5	71	71
Runner-Up	72	69.5	70	68	74	74	73	76

13. Find the best predicted amount of revenue (in millions of dollars), given that the amount has a size 87 thousand ft^2. How does the result compare to the actual revenue of $65.1 million?

Size	160	227	140	144	161	147	141
Revenue	189	157	140	127	123	106	101

14. Find the best predicted new mileage rating of a jeep given that old rating is 19 mi/gal. Is the predicted value close to the actual value of 17 mi/gal?

Old	16	27	17	33	28	24	18	22	20	29	21
New	15	24	15	29	25	22	16	20	18	26	19

15. Find the best predicted temperature for a recent year in which the concentration (in parts per million) of CO_2 is 370.9. Is the predicted temperature close to the actual temperature of 14.5° C??

CO_2	314	317	320	326	331	339	346	354	361	369
Temperature	13.9	14.0	13.9	14.1	14.0	14.3	14.1	14.5	14.5	14.4

16. Find the best predicted IQ score of someone with a brain size of 1275 cm^3

Brain Size	965	1029	1030	1285	1049	1077	1037	1068	1176	1105
IQ	90	85	86	102	103	97	124	125	102	114

17. Listed below are the word counts for men and women.

Male

27531	15684	5638	27997	25433	8077	21319	17572	26429	21966	11680	10818
12650	21683	19153	1411	20242	10117	20206	16874	16135	20734	7771	6792
26194	10671	13462	12474	13560	18876	13825	9274	20547	17190	10578	14821
15477	10483	19377	11767	13793	5908	18821	14069	16072	16414	19017	37649
17427	46978	25835	10302	15686	10072	6885	20848				

Female

20737	24625	5198	18712	12002	15702	11661	19624	13397	18776	15863	12549
17014	23511	6017	18338	23020	18602	16518	13770	29940	8419	17791	5596
11467	18372	13657	21420	21261	12964	33789	8709	10508	11909	29730	20981
16937	19049	20224	15872	18717	12685	17646	16255	28838	38154	25510	34869
24480	31553	18667	7059	25168	16143	14730	28117				

Find the best predicted word count of a woman given that her male partner speaks 6,000 words in a day.

18. According the least-squares property, the regression line minimizes the sum of the squares of the residuals. Listed below are the paired data consisting of the first 6 pulse and the first systolic blood pressures of males.

Pulse (***x***)	68	64	88	72	64	110
Systolic (***y***)	125	107	126	110	72	107

a) Find the equation of the regression line.
b) Identify the residuals, and find the sum of squares of the residuals.
c) Show that the equation $\hat{y} = 70 + 0.5x$ results in a larger sum of squares of residuals.

19. The scatter diagram for the data set below

x	0	2	3	5	5	5
y	7.3	5.1	6	4	5.3	3.6

Given that $\bar{x} = 3.333$, $s_x = 2.0655911$, $\bar{y} = 5.217$, $s_y = 1.3467244$, and $r = -0.8363944$, determine the least squares regression line.

20. The scatter diagram for the data set below

x	0	0	0	2	4	6
y	7.1	5.9	6.5	6.2	4.9	2.2

a) Determine the least squares regression line.
b) Graph the least-squares regression line on the scatter diagram

21. The scatter diagram for the data set below

x	4	5	6	7	9
y	5	8	9	11	14

a) Determine the least squares regression line.
b) Graph the least-squares regression line on the scatter diagram.
c) Compute the sum of the squared residuals for the least-squares regression line found in part (*a*).

22. A student at a junior college conducted a survey of 20 randomly selected full-time students to determine the relation between the number of hours of video game playing each week, *x*, and grade-point average, *y*. She found that a linear relation exists between the two variables. The least-squares regression line that describes this relation is $\hat{y} = -0.0531x + 2.9213$.

a) Predict the grade-point average of a student who plays video games 8 hours per week.
b) Interpret the slope
c) Interpret the appropriate *y*-intercept.
d) A student who plays video games 7 hours per week has a grade-point average of 2.67. Is the student grade-point average above or below average among all students who play video games 7 hours per week.