Intro to Correlation and Linear Regression

12.1 Bivariate Data and Correlation

When conducting research, we might only be interested in only one variable and therefore only take one measurement from each of our study participants (or from each observation). However, we can certainly take more than one measurement from each subject (or observation). When two measurements are taken from each subject, the data are called bivariate data. Bivariate data involves two variables measured on the same subjects or observations. The prefix bi- means "two," and variate refers to variables.

For example, we might ask a random set of students how long they studied before their Statistics final exam and then record the grade they earn on the exam. This set of measurements might look like this:

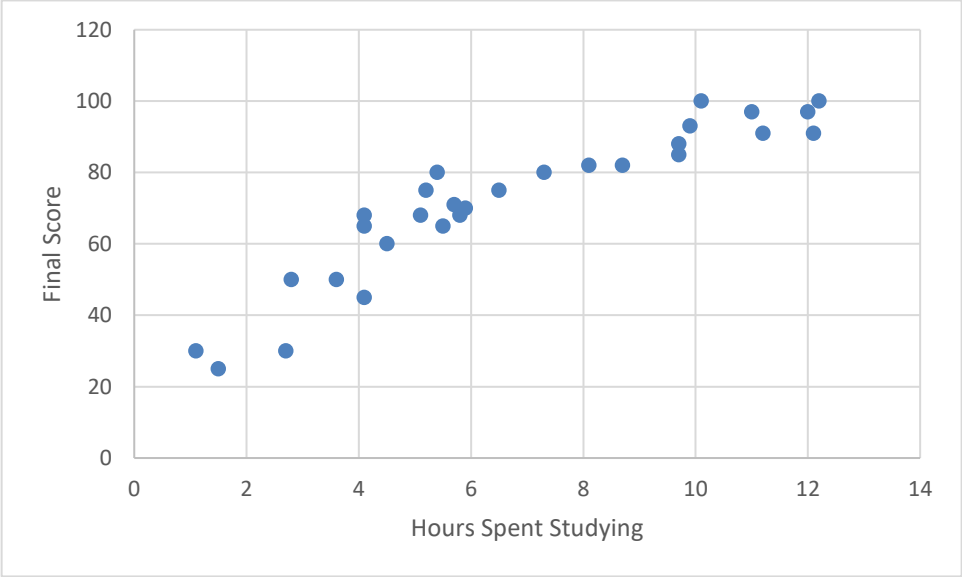
Student	Hours (x)	Score (y)
2548947	12	97
5428763	5.1	68
5426987	5.8	68
4412536	5.7	71
9947532	12.1	91
0218665	4.1	65
0478951	4.5	60
1379405	12.2	100
1986405	5.5	65
2576430	5.4	80
1494847	6.5	75
1024708	8.7	82
3601245	4.1	68
9460572	9.7	88
3046257	11.2	91
1052487	5.2	75
1094410	9.9	93
2403587	1.5	25
6541278	1.1	30
3046200	11	97
2104578	2.8	50
4019875	5.9	70
3026547	10.1	100
9047852	3.6	50
2413056	4.1	45
7549021	2.7	30
1904367	9.7	85
6625780	8.1	82
8014257	7.3	80

In our example study of the relationship between hours studied and exam scores, each student provides a pair of data points—making the data bivariate. Using this data, we can determine if there is an association between numbers of hours spent studying for an exam and the grade earned on the exam. This association will be referred to as a correlation between the two variables. Correlation expresses how changes in one variable are associated with changes in the other. Two variables can have a linear relationship, a non-linear relationship, or no relationship at all. We will focus our attention on the linear relationship (i.e., linear correlation). We will use an analytical measure developed by Karl Pearson, denoted by the variable **r**, to describe the linear relationship between x and y.

Scatter plots

Before introducing the method to calculate Pearson's correlation coefficient, r, to describe the correlation between x (hours of study) and y (final score), we should plot these (x,y) pairs to see if there is some discernable pattern. We plot the points on a standard xy-plane.

Example 12.1 Creating a Scatter Plot Using Excel: Create a scatterplot for the Hours Spent Studying/Final Exam Grade data and discuss the apparent correlation.



Interpreting a Scatter Plot

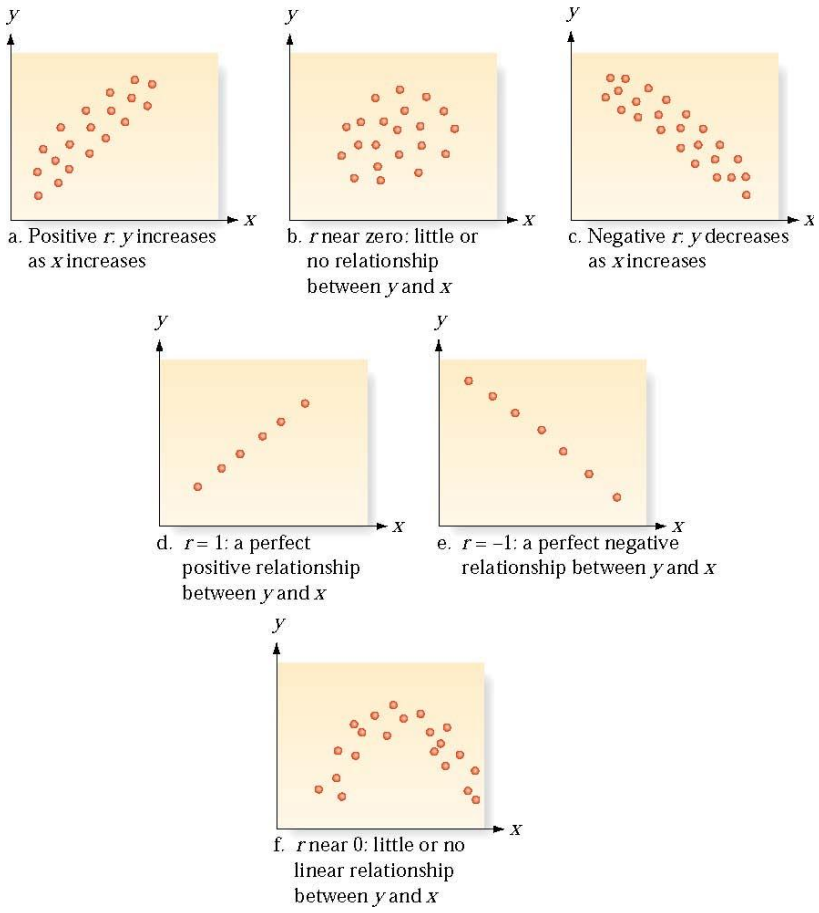
A scatter plot can show four main types of relationships between two variables:

1. Positive Linear Relationship – As one variable increases, the other also increases, and the points roughly follow an upward-sloping line.

2. Negative Linear Relationship – As one variable increases, the other decreases, and the points roughly follow a downward-sloping line.

3. No Relationship – The points are scattered randomly, showing no clear pattern or association between the variables.

4. Non-Linear Relationship – The variables are related, but the pattern follows a curve (e.g., U-shaped or exponential) rather than a straight line.



12.2 Pearson's Correlation Coefficient

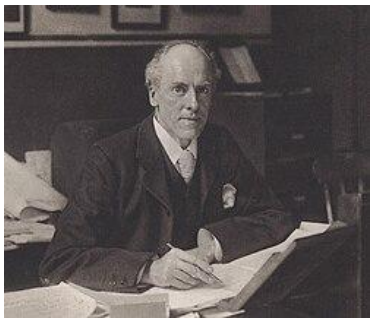


Photo of Karl Pearson in 1910

Now that we have seen that the scatterplot of our example data appears to have a positive linear relationship, we can use ***Pearson's correlation coefficient*** to measure the strength of the linear relationship. The correlation coefficient (r) can range from -1 to 1. When r is negative, it implies that as one variable increases, the other decreases. That is called a negative linear relationship. When r is positive, it implies that as one variable increases, the other also increases. When x and y move up or down together, they are said to have a positive linear relationship.

The Coefficient of Correlation

The coefficient of correlation, $r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$ is a measure of the strength of the linear relationship between two variables x and y.

Where the Sum of Squares (SS) are defined as:

$$SS_{yy} = \sum (y_i - \overline{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

$$SS_{xx} = \sum (x_i - \overline{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$SS_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

Example 12.2: Calculate r for the hours of study and final score data using Excel.

Student	Hours (x)	Score (y)	x^2	xy	y^2
2548947	12	97	144	1164	9409
5428763	5.1	68	26.01	346.8	4624
5426987	5.8	68	33.64	394.4	4624
4412536	5.7	71	32.49	404.7	5041
9947532	12.1	91	146.41	1101.1	8281
218665	4.1	65	16.81	266.5	4225
478951	4.5	60	20.25	270	3600
1379405	12.2	100	148.84	1220	10000
1986405	5.5	65	30.25	357.5	4225
2576430	5.4	80	29.16	432	6400
1494847	6.5	75	42.25	487.5	5625
1024708	8.7	82	75.69	713.4	6724
3601245	4.1	68	16.81	278.8	4624
9460572	9.7	88	94.09	853.6	7744
3046257	11.2	91	125.44	1019.2	8281
1052487	5.2	75	27.04	390	5625
1094410	9.9	93	98.01	920.7	8649
2403587	1.5	25	2.25	37.5	625
6541278	1.1	30	1.21	33	900
3046200	11	97	121	1067	9409
2104578	2.8	50	7.84	140	2500
4019875	5.9	70	34.81	413	4900
3026547	10.1	100	102.01	1010	10000
9047852	3.6	50	12.96	180	2500
2413056	4.1	45	16.81	184.5	2025
7549021	2.7	30	7.29	81	900
1904367	9.7	85	94.09	824.5	7225
6625780	8.1	82	65.61	664.2	6724
8014257	7.3	80	53.29	584	6400
Sum:	195.6	2081	1626.36	15838.9	161809

Using the sums obtained above from the 29 ordered pairs ($\Sigma x = 195.6$, $\Sigma y = 2081$, $\Sigma x^2 = 1626.36$, $\Sigma xy = 15838.9$, and $\Sigma y^2=161809$), we can fill in the sum of square formulas. By hand this would look like this:

$SSxx = \Sigma x^2 - (\Sigma x * \Sigma x)/n = 1626.36 – 195.6*195.6/29 = 307.0717241$

$SSyy = \Sigma y^2 - (\Sigma y * \Sigma y)/n = 161809 – 2081*2081/29 = 12479.31034$

$SSxy = \Sigma xy - (\Sigma x * \Sigma y)/n = 15838.9 – 195.6*2081/29 = 1802.913793$

Finally,

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = r = \frac{1802.913793}{\sqrt{307.0717241 * 12479.31034}} = \mathbf{0.921}$$

This value for r indicates that the numbers of hours spent studying and the final exam score variables have a strong positive linear relationship. The positive relationship implies that more hours spent studying tend to appear with higher final exam scores. You are probably tempted to say that this result indicates that studying more causes students to get higher scores on their final exam. That conclusion is plausible; however, this correlation coefficient cannot confirm that inference. There are several possible reasons for two variables to show a strong linear relationship like this.

Pearson's correlation coefficient (r) and typical interpretations:

- r ≈ +1.0 → Perfect positive linear relationship (points lie exactly on an upward-sloping line).
- r between +0.7 and +0.9 → Strong positive linear relationship.
- r between +0.4 and +0.6 → Moderate positive linear relationship.
- r between +0.1 and +0.3 → Weak positive linear relationship.
- r near 0 → No linear relationship.
- r between –0.1 and –0.3 → Weak negative linear relationship.
- r between –0.4 and –0.6 → Moderate negative linear relationship.
- r between –0.7 and –0.9 → Strong negative linear relationship.
- r ≈ –1.0 → Perfect negative linear relationship (points lie exactly on a downward-sloping line).

Possible reasons why r indicates a strong linear relationship:

1. The x variable might have a causal relationship with y. This means x is a cause for y (x causes y).
2. The y variable might have a causal relationship with x. This means y is a cause for x (y causes x).
3. There is a third variable that causes both x and y. This variable is called a lurking variable.
4. The relationship is purely a coincidence. This is called a spurious relationship. Sometimes weird patterns turn up in data that are do purely to random chance.

Examples of each:

Hours spent studying (x) are correlated with grades (y) because studying causes us to learn more and to perform better on exams.

Hours spent exercising (x) might be negatively correlated with BMI (body mass index) (y) not because exercise causes one to lose weight, but because when we are heavier, we find it too hard to exercise. In other words, it might be harder to sustain exercise habits due to the extra weight and the corresponding stress on our joints.

Ice cream sales (x) and drowning deaths (y) are correlated in many places around the world not because ice cream causes drowning or because people eat ice cream after someone drowns as a coping mechanism. These variables are correlated because both variables increase during warm months and likely decrease during cold months.

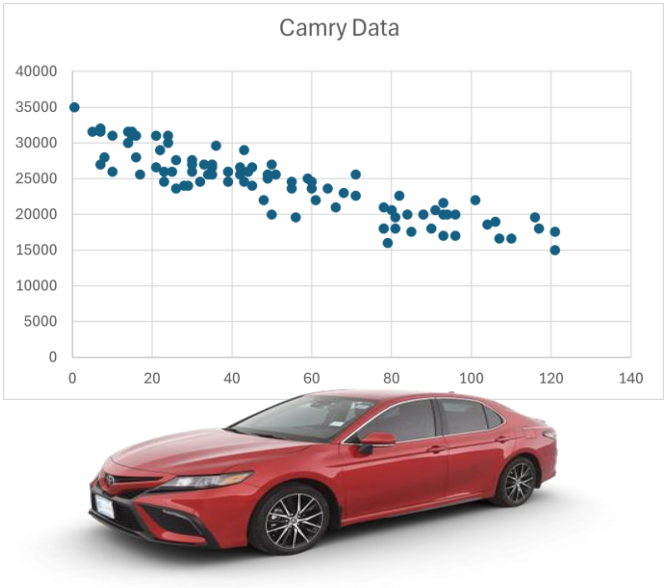
Consumption of margarine in a state and the divorce rate for the state are correlated not because of a relationship between the consumption of margarine and marital dissatisfaction but most likely due to a mere coincidence.

Example 12.3 Calculating r and classify the association:

The following set of data was derived from a used car website. The bivariate data set includes the mileage numbers (in thousands) and the listed sale price in 2025 dollars for the Toyota Camry SE. Use the following sum of squares to calculate r and discuss the type of correlation. $SSxx = 99657.6224$

SSxy = -12326703.13, and SSyy = 1931256250

Miles	Price	yr
5	31590	25
15	30990	25
7	31990	25
16	30990	25
10	30990	25
7	31590	25
24	30990	25
0.5	34990	25
21	30990	25
15	31590	25
23	25990	24
39	25990	24
14	29990	24
39	25990	24
30	26990	24
35	26990	24
14	31590	24



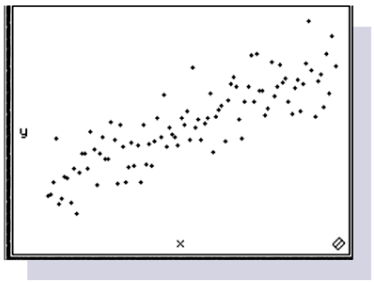
33	26990	24
26	27590	24
44	25990	24
42	26590	24
21	26590	24
25	25990	23
24	29990	23
49	24990	23
34	25590	23
8	27990	23
36	29590	23
43	24590	23
44	25990	23
14	29990	23
51	25590	23
49	25590	23
30	27590	23
22	28990	23
35	25590	23
35	26590	23
59	24990	23
60	24590	22
43	28990	22
16	27990	22
50	26990	22
42	25590	22
55	24590	22
71	25590	22
55	23590	22
101	21990	22
64	23590	22
32	24590	22
39	24590	21
82	22590	21
71	22590	21
93	21590	21
60	23590	21
106	18990	21
30	25990	21
68	22990	21
116	19590	21
104	18590	20
45	26590	20

45	23990	20
7	26990	20
93	19990	20
91	20590	20
23	24590	19
48	21990	19
10	25990	19
17	25590	19
78	20990	19
88	19990	19
29	23990	19
88	19990	19
61	21990	19
94	19990	19
80	20590	19
26	23590	18
28	23990	18
96	19990	18
84	19990	18
121	17590	18
117	17990	18
81	19590	18
66	20990	18
107	16590	17
85	17590	17
93	16990	17
96	16990	17
81	17990	17
56	19590	16
78	17990	16
78	17990	16
50	19990	16
90	17990	16
110	16590	16
79	15990	14
121	14990	14

$$r = \frac{SSxy}{\sqrt{SSxxSSyy}} = \frac{-12326703.13}{\sqrt{99657.6224 * 1931256250}} = -.889$$

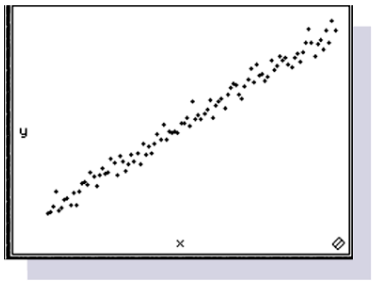
Below are some more examples of r values and scatter plots:

ActivStats



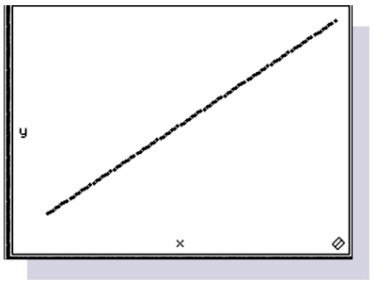
(a) Positive correlation: $r = 0.851$

ActivStats



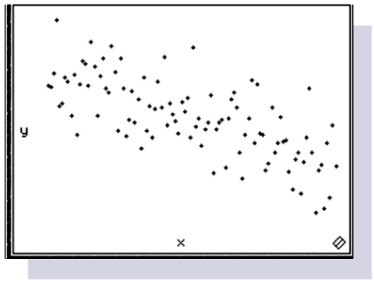
(b) Positive correlation: $r = 0.991$

ActivStats



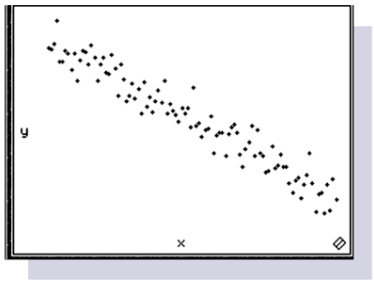
(c) Perfect positive correlation: $r = 1$

ActivStats



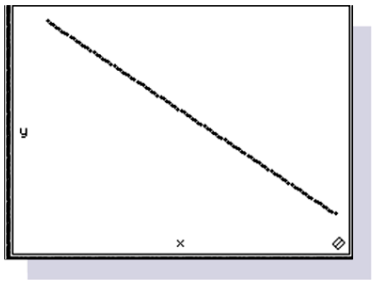
(d) Negative correlation: $r = -0.702$

ActivStats



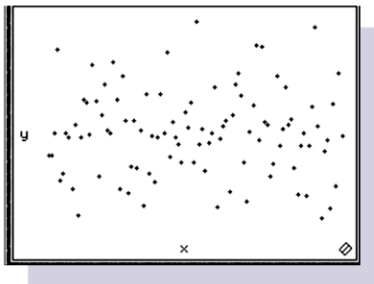
(e) Negative correlation: $r = -0.965$

ActivStats



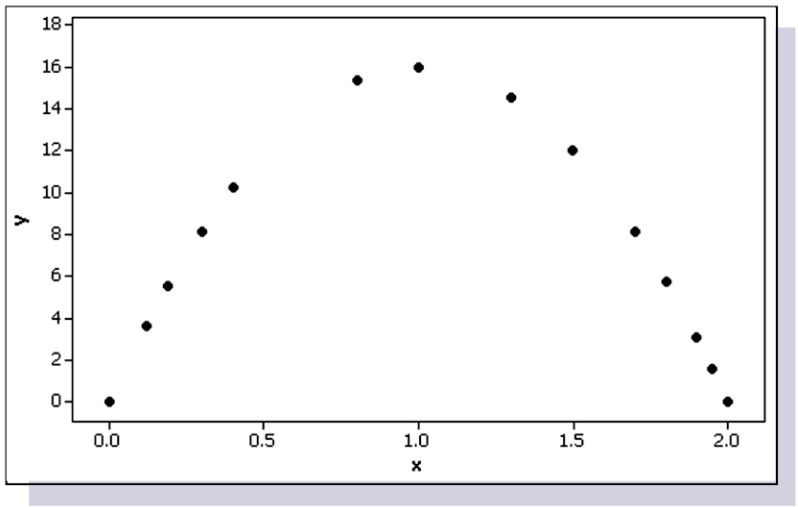
(f) Perfect negative correlation: $r = -1$

ActivStats



(g) No correlation: $r = 0$

Minitab



(h) Nonlinear relationship: $r = -0.087$

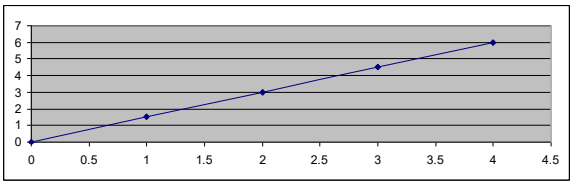
12.3 Linear Regression

Probabilistic Models

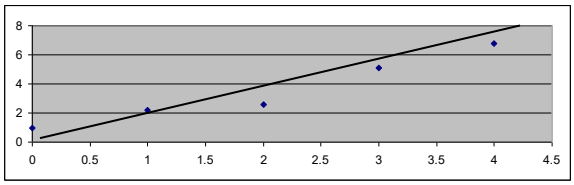
In this section, we will try to model the relationship between two variables. In algebra, you worked with many models that were ***deterministic*** in nature. For example, the model: $y = 1.07x$ is a deterministic model that will give the after-tax price for an item purchased in Palm Beach, Florida. X here represents the pre-tax price of an item. Y is the final price of the item post-tax. This model is deterministic because given a pre-tax price we can exactly (there is no error in this prediction) determine the value of the item after tax. Recall that the y variable is called the dependent variable because it depends upon the independent variable x.

Deterministic models are great when we can get them, but often we do not know all the factors affecting the dependent variable (even if we did know them all, often it is not possible to include them all in a model). In those cases, we will not be able to predict y without error. This means we will need to create a ***probabilistic*** model:

$$y = \text{deterministic model} + Random\ error$$

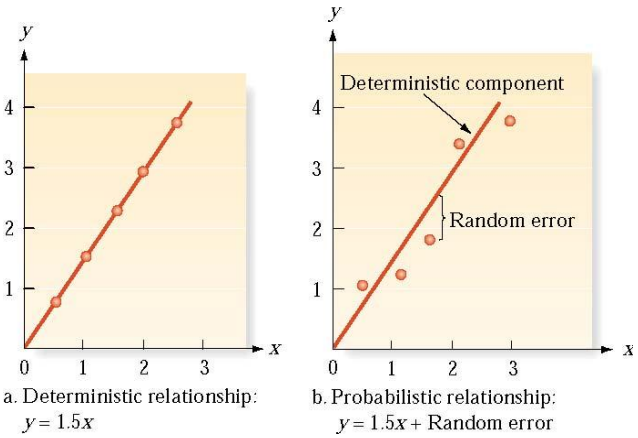


Deterministic model



Probabilistic Model

In the probabilistic model graph, the difference between the height of the line and the height of our individual points is due to the random error.



a. Deterministic relationship: $y = 1.5x$

b. Probabilistic relationship: $y = 1.5x +$ Random error

In this section, we will look at the simplest form of a probabilistic model:

A First-Order (Straight-Line) Probabilistic Model

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

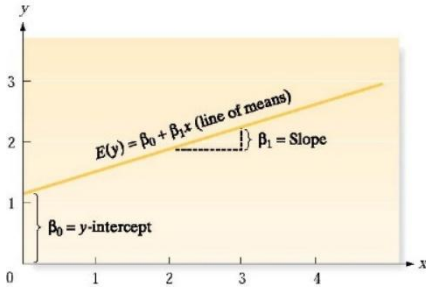
y = Dependent variable, x = independent variable, β_1 = slope, β_0 = y-intercept, and ε = random error component.

Using data to come up with estimates of the parameters β_0 and β_1 to form an equation is called **regression analysis**. The goal of **regression analysis** is to find the straight line that comes closest to all the points in a scatter plot simultaneously.

Fitting the Model: The Least Squares Approach

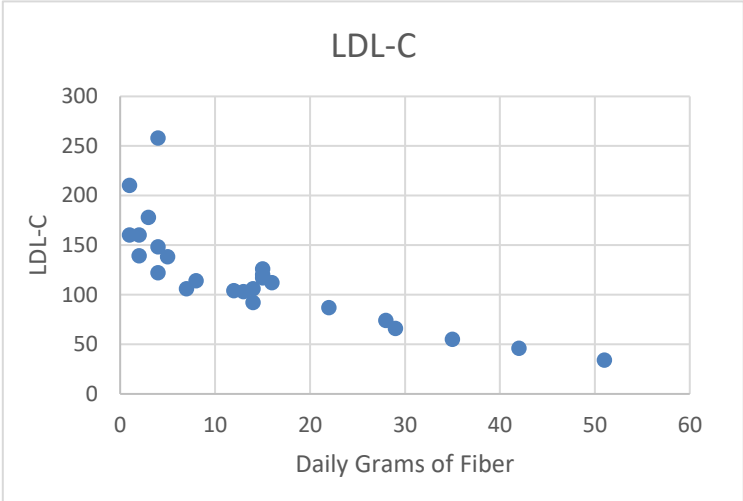
$$y = \beta_0 + \beta_1 x + \varepsilon$$

To fit the straight-line model, we need to find a way to estimate the unknown parameters: β_1 & β_0.

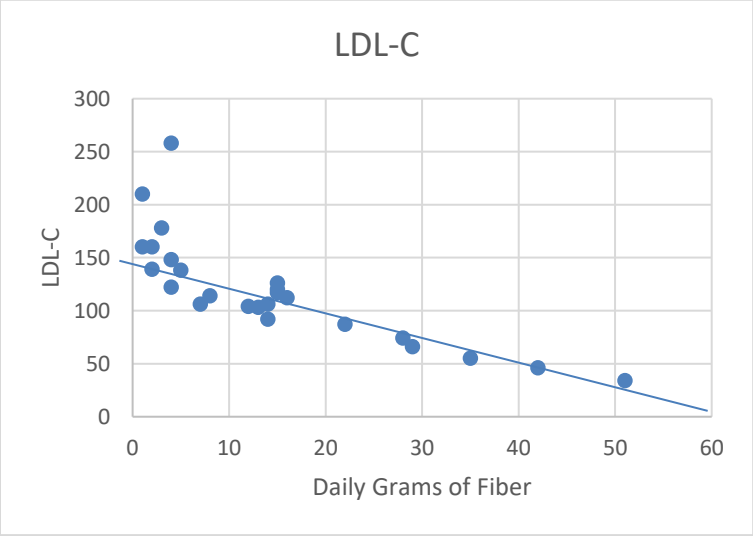


We will demonstrate the method using data from a study looking at the relationship between average daily fiber intake (estimated from food journals) and LDL cholesterol levels in 25 adults between the ages of 49 -51.

Subject	Fiber Intake	LDL-C
1	1	160
2	22	87
3	28	74
4	42	46
5	15	126
6	13	103
7	12	104
8	14	106
9	2	139
10	15	120
11	16	112
12	29	66
13	15	117
14	14	92
15	2	160
16	4	258
17	3	178
18	51	34
19	5	138
20	4	122
21	1	210
22	7	106
23	4	148
24	8	114
25	35	55



We could try to fit some arbitrary line to the points above. For Example:



The line was created by running the line between the following two points in the data set: (2, 139) and (42,46). Using these points, we can derive the slope and eventually the equation of the line.

Slope $\frac{(y_2-y_1)}{(x_2-x_1)} = \frac{(139-46)}{(2-42)} \approx -2.325$. Using the point-slope formula from Algebra, we get the linear equation: $\tilde{y} = -2.325x + 143.65$. This model was obtained visually (i.e.-we guessed). We could have made several other guesses at the appropriate equation, so we should assess our guess.

Let's then compare the observed and predicted values for the visual model we found. In the table below, the X and Y are the actual values represented by the dots in our graph above. The $\tilde{y}$ (y-tilda) is the value that results when we plug the x value from the leftmost column of the table into our model. The first of the last two columns gives us the difference between the actual y-value from the point and the predicted value from our line. We call that difference the ***error*** of our prediction. For example, our line model says that when x is 1 we should have y at 141.325, but the y value at x = 1 was 160. This means our error is 18.675. The last column squares these differences (or errors).

Fiber Intake	**LDL-C**	$\tilde{y}$	$y - \tilde{y}$	$(y - \tilde{y})^2$
1	160	141.325	18.675	348.7556
22	87	92.5	-5.5	30.25
28	74	78.55	-4.55	20.7025
42	46	46	0	0
15	126	108.775	17.225	296.7006
13	103	113.425	-10.425	108.6806
12	104	115.75	-11.75	138.0625
14	106	111.1	-5.1	26.01
2	139	139	0	0
15	120	108.775	11.225	126.0006
16	112	106.45	5.55	30.8025
29	66	76.225	-10.225	104.5506
15	117	108.775	8.225	67.65062
14	92	111.1	-19.1	364.81
2	160	139	21	441
4	258	134.35	123.65	15289.32
3	178	136.675	41.325	1707.756
51	34	25.075	8.925	79.65562
5	138	132.025	5.975	35.70062
4	122	134.35	-12.35	152.5225
1	210	141.325	68.675	4716.256
7	106	127.375	-21.375	456.8906
4	148	134.35	13.65	186.3225
8	114	125.05	-11.05	122.1025
35	55	62.275	-7.275	52.92563
Sum of Errors			**225.4**	**24903.43**

One way to determine quantitatively how well a straight line fits a set of points is to note the extent to which the data points deviate from the line. The quantity $\sum(y-\tilde{y})$ in the table above gives us the total deviation between our observed values and our predicted values. The $\sum(y-\tilde{y})$ should equal zero (the sum of errors should equal zero). In our visual model, that is not the case, which is something that would eliminate it as a candidate model for this set of data points. We will want all our prediction lines to have the property that **the sum of errors equals zero**. This will ensure on average our prediction error is zero. The quantity $\sum(y-\tilde{y})^2$ is called the ***sum of squares of the errors (SSE)***. It gives another measure of deviation which gives greater emphasis to larger deviations from the line.

Remember, we visually selected the model (line) above, so it is no wonder that the sum of errors was not zero. However, there are usually multiple models possible that have the property that $\sum(y-\tilde{y})$=0. Since we can find more than one model with the property $\sum(y-\tilde{y})$= 0, we need additional criteria to choose the best fitting line. It turns out that it can be shown that there is one line for which the sum of errors is zero and SSE is a *minimum*. This line is called the **least squares line**.

The **least squares line** has the following properties:

1. The sum of errors (SE) equals zero.
2. The sum of squared errors (SSE) is smaller than that for any other straight-line model.

The following formulas will give the Least Squares Estimates for the population β_1& β_0. We will use a "hat" symbol to denote the estimates, that is to say $\hat{\beta}_1$estimates β_1 and $\hat{\beta}_0$ estimates β_0.

Formulas for the Least Squares Estimates

Slope: $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$

y-intercept: $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$

where $SS_{xy} = \sum(x_i-\overline{x})(y_i-\overline{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$

and $SS_{xx} = \sum(x_i-\overline{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$

Example 12.4 Calculate the least squares line for the fiber – LDL data using Excel: (Data Tab Video)

Preliminary computations to help us do our work are provided to speed things up:

$\sum x_i y_i$= 30,045

$\sum x_i^2$= 9,540

$\sum x_i$= 362

$\sum y_i$= 2,975

$\bar{x}$= 14.48

$\bar{y}$= 119

Using the numbers from above we can get:

$SS_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$= 30,045 – (362)(2975)/25 = -13,033

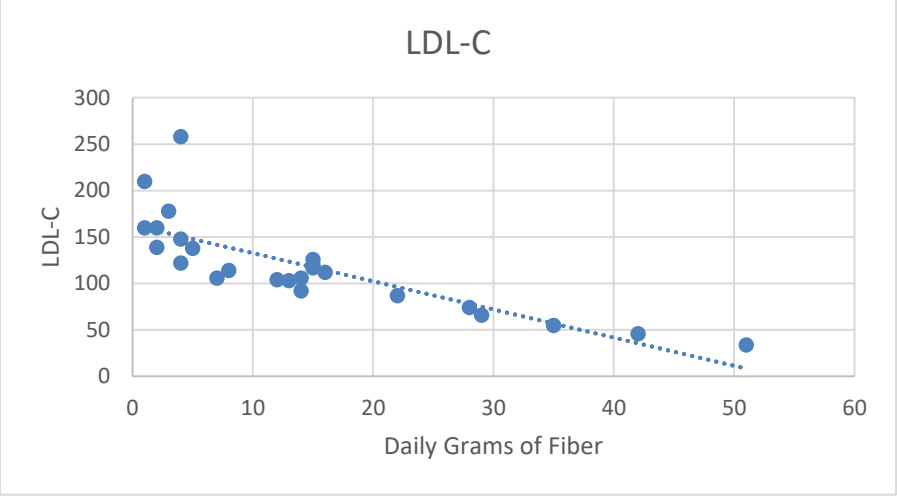
$SS_{xx} = \sum(x_i - \overline{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$= 9,540 –(362)(362)/25= 4,298.24

then $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$= -13033/4298.24 = -3.03217

and $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = \frac{(\sum y_i)}{n} - \hat{\beta}_1 \frac{(\sum x_i)}{n}$= $119 - (-3.03217)(14.48) = 162.9058$

The **least squares line** is then given by:

$$\hat{y} = 162.91 - 3.03x$$



Let us find the SSE for this line to determine if it beats our visual model:

Fiber Intake	LDL-C	$\tilde{y}$	$y - \tilde{y}$	$(y - \tilde{y})^2$
1	160	159.87363	0.12637	0.015969
22	87	96.19806	-9.19806	84.60431
28	74	78.00504	-4.00504	16.04035
42	46	35.55466	10.44534	109.1051
15	126	117.42325	8.57675	73.56064
13	103	123.48759	-20.4876	419.7413
12	104	126.51976	-22.5198	507.1396
14	106	120.45542	-14.4554	208.9592
2	139	156.84146	-17.8415	318.3177
15	120	117.42325	2.57675	6.639641
16	112	114.39108	-2.39108	5.717264
29	66	74.97287	-8.97287	80.5124
15	117	117.42325	-0.42325	0.179141
14	92	120.45542	-28.4554	809.7109
2	160	156.84146	3.15854	9.976375
4	258	150.77712	107.2229	11496.75
3	178	153.80929	24.19071	585.1905
51	34	8.26513	25.73487	662.2835
5	138	147.74495	-9.74495	94.96405
4	122	150.77712	-28.7771	828.1226
1	210	159.87363	50.12637	2512.653
7	106	141.68061	-35.6806	1273.106
4	148	150.77712	-2.77712	7.712395
8	114	138.64844	-24.6484	607.5456
35	55	56.77985	-1.77985	3.167866
Sum of Errors			0.00054	20721.71

We can now confirm our Least Squares Model is better fitting than out visual model because our LSM has a Sum of Errors = 0, and even if both models had that trait, the least squares model has a lower SSE.

Finally, we created our model for the purpose of making predictions about the average y for a given x. In this example, we would want to do something like predict the average LDL level for people eating 15 grams of fiber daily. Unfortunately, we cannot jump into using this model until we confirm the model's variables have a significant linear association. Elsewhere in the course, we will be able to test this in a formal way, but for this brief introduction to simple linear regression, we will check if there is significant association by using n, r, and a table of critical r values.

Example 12.5 Checking for a significant r value:

First, we need to determine our correlation coefficient, r. Recall the formula is:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

We have already found SSxy and SSxx, so we only need to calculate SSyy:

SSyy = Σy^2 - (Σy*Σy)/n = 414,265 - 2,975*2,975/25 = 60,240

Using our three SS values:

SSyy = Σy^2 - (Σy*Σy)/n = 414,265 - 2,975*2,975/25 = 60,240

$SS_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$= 30,045 – (362)(2975)/25 = -13,033

$SS_{xx} = \sum(x_i - \overline{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$= 9,540 –(362)(362)/25= 4,298.24

We get:

$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$ = $r = \frac{-13033}{\sqrt{4298.24*60240}}$= -0.810

This looks like a strong, negative linear relationship, but to check if it is significant, we will take the n and r to the table below to confirm significance:

Critical Values of the Pearson Correlation Coefficient		
n	**α = 0.05**	**α = 0.01**
4	0.95	0.99
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875
8	0.707	0.834
9	0.666	0.798
10	0.632	0.765
11	0.602	0.735
12	0.576	0.708
13	0.553	0.684
14	0.532	0.661
15	0.514	0.641
16	0.497	0.623
17	0.482	0.606
18	0.468	0.59

19	0.456	0.575
20	0.444	0.561
25	0.396	0.505
30	0.361	0.463
35	0.335	0.43
40	0.312	0.402
45	0.294	0.378
50	0.279	0.361
60	0.254	0.33
70	0.236	0.305
80	0.22	0.286
90	0.207	0.269
100	0.196	0.256

If we look at the row for n = 25 and take the critical r value from the 0.05 column, we can determine if we have a significant linear correlation.

The table provides the value: 0.396. If the absolute value of our correlation coefficient (r) is greater than this value, we can conclude the correlation is significant. Since our r value was -0.810, its absolute value is larger than the critical value of 0.396.

Since $|-0.810| > 0.396$, we conclude there is a significant linear relationship between fiber intake and LDL cholesterol.

Now that we have confirmed the significant linear relationship exists between these two variables, we can use the model to make predictions.

Example 12.6 Using the model for prediction:

Using the model developed from the fiber and LDL cholesterol data set, determine the average LDL number for adults with a daily intake of fiber of 15 grams.

$$\hat{y} = 162.91 - 3.03x$$

Interpreting the Slope and y-Intercept

The standard interpretation of the slope is, "the slope represents the average change in y for a unit change in x." For example, for every additional gram of fiber eaten each day, the average change in LDL cholesterol is 3.03 mg/dL.

To interpret the y-intercept, we set x = 0. For example, the average LDL value for those eating 0 grams of fiber each day is 162.91 mg/dL. For some models, the y-intercept will not be interpretable. For example, if you create a model for predicting average shoe size (y) based on male height (x), it is nonsensical to set the height to zero.