Correlation

MATH 2441, BCIT

Technical Mathematics for Food Technology

April 19, 2018

Correlation

A correlation exists between two variables when the values of one variable are somehow associated with the values of the other variable.

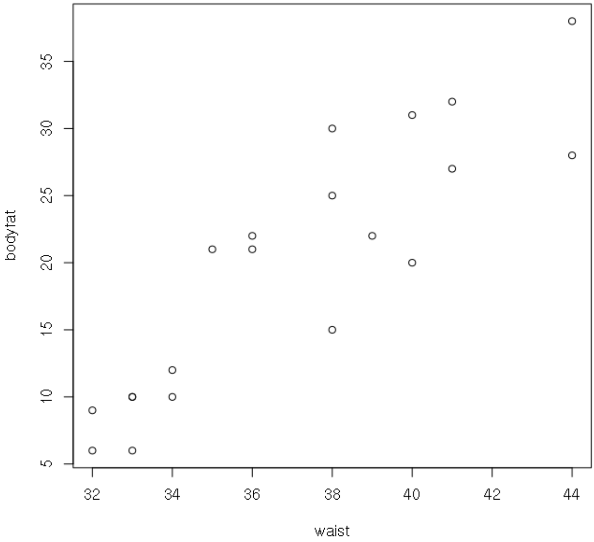
A linear correlation exists between two variables when there is a correlation and the plotted points of paired data result in a pattern that can be approximated by a straight line.

Correlation

Here is the data for waist (in inches), weight (in pounds), and body fat (in percent) for 20 test subjects.

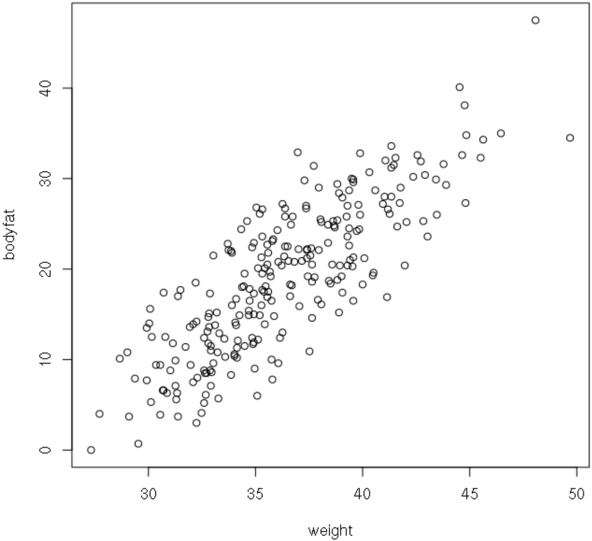
33	188	10	33	160	10	40	192	31	32	175	6
40	240	20	41	215	27	41	205	32	36	181	21
36	175	22	34	159	12	35	173	21	38	200	15
32	168	9	34	146	10	38	187	25	33	159	6
44	246	38	44	219	28	38	188	30	39	196	22

Correlation

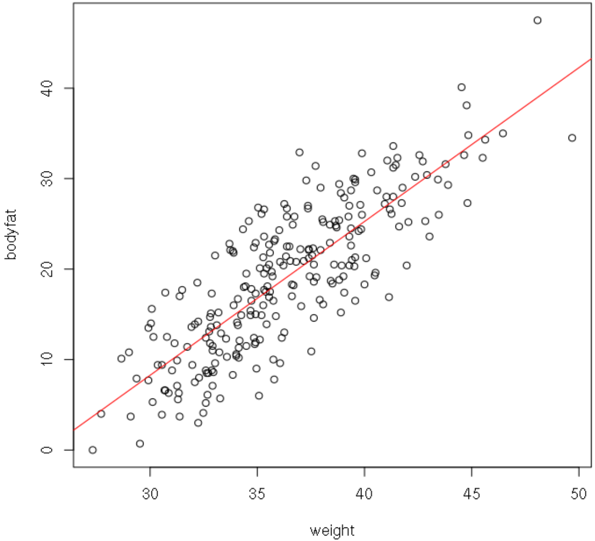


In the previous slide, you can see the data from 20 test subjects. In the following slide, you can see the data from 250 test subjects. It appears that there is a relationship between waist and body fat.

Correlation



Correlation



The red line is called the regression line. We will learn how to calculate it later. Here is its equation:

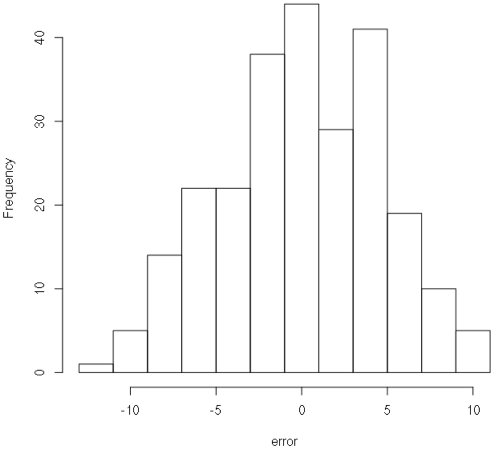
$$b = 1.7w - 42.73 \tag{1}$$

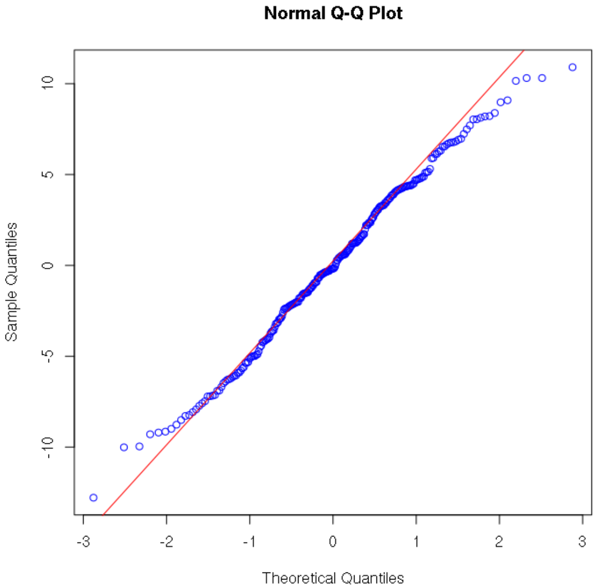
The regression line minimizes the mean square distance of the data points to the line (all other lines have a greater mean square distance from the data points). Have a look at three of the 250 test subjects.

	waist	bf (act)	bf (pred)	error
test subject 1	33.54	12.3	14.29	1.9936
test subject 2	32.68	6.1	12.82	6.7212
test subject 3	34.61	25.3	16.10	-9.1993

Correlation

It appears that the error is normally distributed (perhaps not quite on the margins).



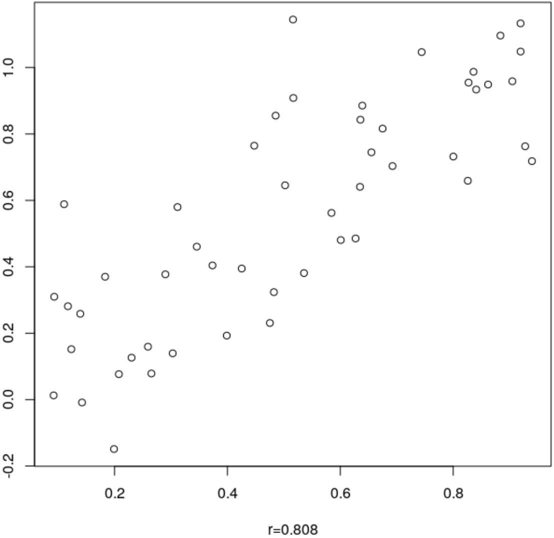

Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
10
5
0
-5
-10
-3
-2
-1
0
1
2
3

Correlation

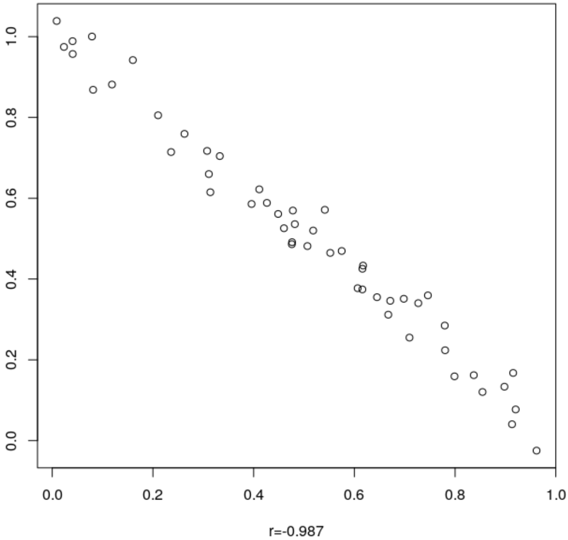
Let's have a look at a few scatterplots. Is there a correlation or not? Is there a linear correlation? The linear correlation coefficient r measures the strength of the linear correlation. It is a sample statistic. The linear correlation coefficient for the population is called ϱ ("rho" in the Greek alphabet).

The correlation coefficient for the sample of 250 test subjects measuring body fat and waist is $r = 0.8236847$.

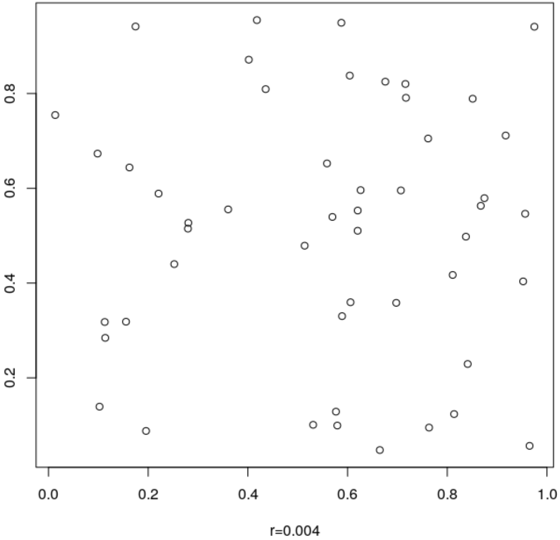
Scatterplot Examples



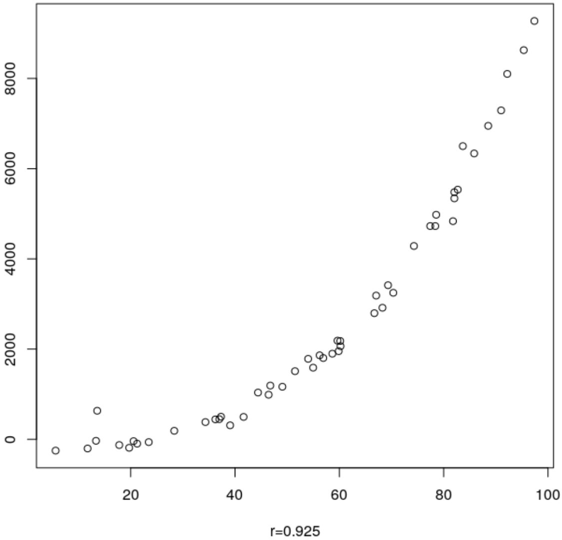
Scatterplot Examples



Scatterplot Examples



Scatterplot Examples



Notation

To determine whether there is a linear correlation between two variables, first take note of the following notation.

n	number of pairs of sample data
$\sum$	denotes addition of items indicated
$\sum x$	sum of all x-values
$\sum x^2$	sum of all x^2-values
$(\sum x)^2$	sum of all x-values squared
$\sum xy$	sum of all $x \cdot y$-values
r	linear correlation coefficient for sample data
ϱ	linear correlation coefficient for population of paired data

Requirements

Here are the requirements for the procedure that follows.

1. The sample of paired (x, y) data is a simple random sample of quantitative data.
2. Visual examination of the scatterplot confirms that the points approximate a straight-line pattern.
3. Outliers must be removed if they are known to be errors. The procedure is not robust with respect to erroneous outliers.

Requirements 2 and 3 are an intuitive summary of a more stringent requirement: the pairs of (x, y) data must have (or approximate) a bivariate normal distribution, which means that for a fixed value x, the corresponding y-values have a normal distribution, and vice versa. Think of the deviation of the actual y-value from a perfectly linear corresponding y-value as a normally distributed error.

Calculating r

Here is a simple formula for r that is difficult to calculate. Let z_x be the z-score of an individual x-value and z_y be the z-score of an individual y-value. Then

$$r = \frac{\sum (z_x z_y)}{n-1} \tag{2}$$

Here is a more difficult formula that makes calculation much easier.

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}} \tag{3}$$

Calculating r Example

Here are the data for females, shoe prints, and heights.

24.8	165.1	28.1	179.1
28.6	166.4	27.6	175.9
25.4	177.8	26.5	166.4
26.7	167.6	26.5	167.6
26.7	168.3	28.4	162.6
27.9	165.7	26.5	167.6
27.9	165.1	26.0	165.1
28.9	165.1	27.0	172.7
27.9	165.1	25.1	157.5
25.9	152.4	27.9	167.6
25.4	162.6		

Calculating r Example

Now calculate the following . . .

$\sum x$	565.7	$\sum y$	3503.3
$\sum x^2$	15268.45	$\sum y^2$	585173.7
$(\sum x)^2$	320016.5	$(\sum y)^2$	12273111
$\sum xy$	94404.95		

. . . and fill in the formula

$$r = \frac{21 \cdot 94404.95 - 565.7 \cdot 3503.3}{\sqrt{21 \cdot 15268.45 - 320016.5}\sqrt{21 \cdot 585173.7 - 12273111}} \tag{4}$$

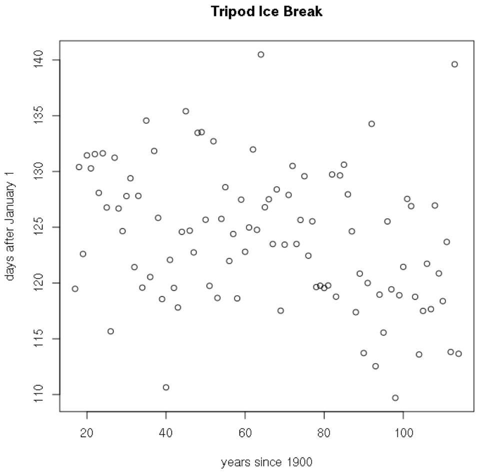
There are many opportunities here to make an error. It is better to use statistics software. In R Statistics, for example, the relevant command is `cor(x,y)`. The result, in either case, is $r = 0.22122$.

Hypothesis Testing for Correlation

Use the critical values for the Pearson Correlation Coefficient r to determine whether there is a correlation between the two variables or not. In the case of female shoe prints and heights, $n = 20$, and therefore, at a significance level $\alpha = 0.05$, the critical value is $r = 0.444$. The null hypothesis is $\varrho = 0$.

Since our test statistic is only $r^* = 0.221$, we fail to reject the null hypothesis that there is no correlation. There is not enough evidence to show that there is a linear correlation (remember to check the requirements first).

Example 1: Nenana Tripod Ice Break. Have a look at `http://www.nenanaakiceclassic.com/`. Is there a correlation? (Might it support the theory that the Earth is warming?)



Hypothesis Testing in R

- Step 1 The null hypothesis is $\varrho = 0$. The alternative hypothesis is $\varrho < 0$ (the Earth is warming, therefore the tripod will break up the ice earlier in the year the more recently we measure). We will test the null hypothesis at a significance level of $\alpha = 0.01$.
- Step 2 The test statistic is $r = -0.3130899$ (calculated using R Statistics).
- Step 3 The critical value of the Pearson Correlation Coefficient r is approximately 0.256 at $\alpha = 0.01$ and $n = 98$ (consult the table).

Decision: reject the null hypothesis. The data supports the hypothesis that there is a linear correlation between years after 1900 and the days after January 1 when the tripod breaks the ice (assuming that the data have a bivariate normal distribution and a linear rather than some other correlation).

Hypothesis Testing in R

Here is how you can do the hypothesis testing in R Statistics. Let y be the years after 1900 and d be the days after January 1 when the tripod breaks the ice. Try the command $\mathrm{summary(lm(d\ y))}$. The output is on the next slide. Notice the p-value. It clearly suggests that we should reject H_0.

Hypothesis Testing in R

```
Residuals:
     Min       1Q   Median       3Q      Max
-15.2035  -3.6805  -0.2056   4.0684  18.7533

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 128.58108    1.50954   85.18   <2e-16 ***
y            -0.06834    0.02116   -3.23   0.0017 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.925 on 96 degrees of freedom
Multiple R-squared:  0.09803,Adjusted R-squared:  0.08863
F-statistic: 10.43 on 1 and 96 DF,  p-value: 0.001695
```

Hypothesis Testing in R

Let's have a closer look at the summary on the last slide. The residuals are the errors of our predictions using the regression line. For each x-value (independent variable), there is a y-value (dependent variable) and a $\hat{y}$-value (prediction using the regression line. The residual is $\hat{y} - y$. The residual standard deviation s_e is a measure how much the data scatters along the regression line:

$$s_e = \sqrt{\frac{\sum(\hat{y} - y)^2}{n-2}} \tag{5}$$

The residual standard deviation for the Nenana data is large, 5.925 days, because even if you know the regression line it's hard to predict the date when the ice will be broken in a particular year.

Hypothesis Testing in R

s_e is one measure of the relationship between x-values and y-values. The correlation coefficient r is another one. In the R summary it is called "Multiple R-squared" and equals r^2. The reason why it is squared is because one could say that the correlation accounts for r^2 of the variation in the y-values. In the Nenana example, which year it is accounts for 9.8% of the variation in the number of days it takes for the ice to break.

A theorem in statistics tells us that

$$\frac{b_1 - \beta_1}{\frac{s_e}{s_x\sqrt{n-1}}} \tag{6}$$

is distributed according to a t-distribution with degree of freedom $df = n - 2$. b_1 is the slope of the regression line calculated from the sample; β_1 is the slope of the regression line hypothesized for the population. The R summary tells us that the slope of the regression line for the sample is $b_1 = -0.06834$ and the p-value for the hypothesis that $\beta_1 = 0$ is 0.0017 (two-tailed). We reject the hypothesis that $\beta_1 = 0$, which is similar to rejecting the hypothesis that $r = 0$. It is usually not interesting to investigate the hypothesis that the y-intercept is zero.

Hypothesis Testing in R

Here is yet another hypothesis test whether there is a linear correlation or not. A relatively complicated formula gives us the F-statistic of the regression analysis. The F-distribution (named after Ronald Fisher) looks similar to the chi-squared distribution. It has two degrees of freedom. In R Statistics, `qf(0.95,5,2)` gives you the F-statistic for which 95% of the area under the curve is to the left of the F-statistic, with degrees of freedom 5 and 2. `pf(19.29641,5,2)` is the reverse procedure which gives you the area under the curve to the left of the F-statistic 19.29641. We will meet this distribution again when we cover ANOVA.

Exercise 1: The table below lists measured amounts of redshift and the distances (billions of light-years) to randomly selected clusters of galaxies. Is there sufficient evidence to conclude that there is a linear correlation between amounts of redshift and distances to clusters of galaxies?

Redshift	Distance
0.0233	0.32
0.0539	0.75
0.0718	1.00
0.0395	0.55
0.0438	0.61
0.0103	0.14

The Regression Line

To find the regression line $\hat{y} = b_0 + b_1 x$, use the following formula for the slope b_1 and the y-intercept b_0:

$$b_1 = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} \tag{7}$$

$$b_0 = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} \tag{8}$$

You may wonder where these equations come from. They identify the line equation which best fits the data using the least squares method. The least squares method identifies the line that best fits the data by measuring the distance that each data point is away from the line, squaring it, and then adding all of those numbers. The line that scores lowest on this fitness test is the regression line.

Regression Line Example

Example 2: Galaxy Distances. It is clear in the hypothesis test that there is a linear correlation between redshift and galaxy distances, even with a small sample size. What is the regression line? How could we predict the distance of a galaxy, knowing its redshift?

$$\begin{aligned} b_0 &= -0.004396 \\ b_1 &= 13.999899 \end{aligned} \tag{9}$$

Each thousandth unit of redshift adds fourteen million light-years to the distance.

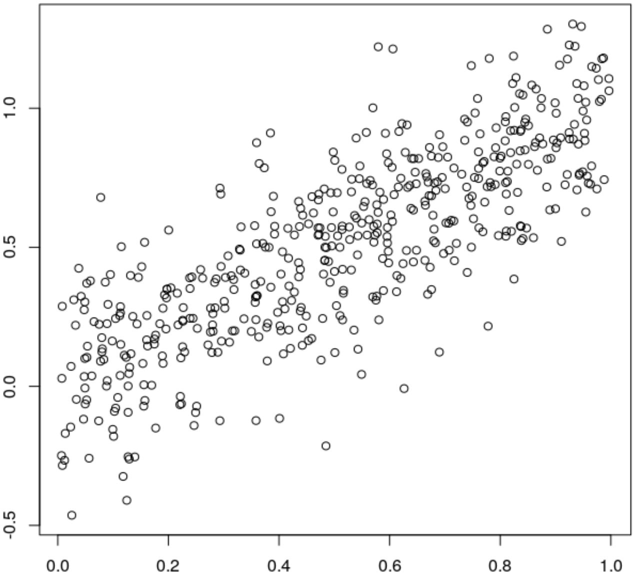
In R Statistics, you can create a scatterplot of data sets `x` and `y` using the command `plot(x,y)`. Adding the command `abline(lm(y∼x),col="red")` will add the regression line to the plot (in red colour). `lm(y∼x)` will give you both the *y*-intercept and the slope of the regression line.

Regression Line Example

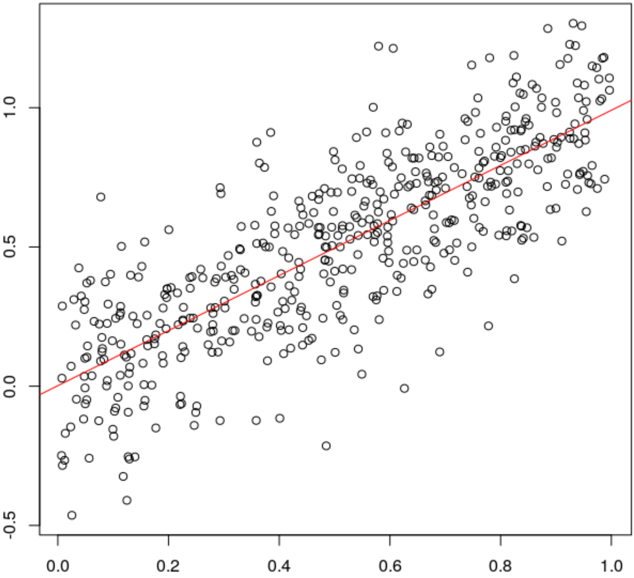
Here is some R Statistics code:

```
a<-runif(500,0,1)

b<-rnorm(500,a,0.2)

plot(a,b)

abline(lm(b~a),col="red")
```

Regression Line Example



Regression Line Example



Exercise 2: Consider the data on the next slide. These are the results for two successive term tests (the names are randomly made up by a computer program, but the grades are real). Answer the following questions:

1. Is there a linear correlation between the first and the second term test? Answer the question for a significance level of $\alpha = 0.05$. If you were doing this problem with a significance level $\alpha = 0.01$, what would be the decision and what type of error (type I or type II) would it make less likely compared to using the higher significance level?
2. What is the equation of the regression line?
3. If William Jones (again, fake name but real grade) had a score of 85 on the first term test, what score is the point estimate for the second term test given the linear correlation? His true score for the second term test was 77.

Regression Line Exercise

Nancy Rogge	83	60		Arnold Murray	73	82
Elizabeth Rushing	98	95		Ann Coburn	60	52
Katy Nunez	68	62		Kim Lazzari	63	62
Michael Preuss	68	90		Valentina Martinez	68	88
Edna Phipps	68	90		Eric Mumford	78	91
George Thompson	45	85		Alyssa Warner	98	90
James Newman	42	20		Kevin Ellis	65	72
Nathan Stowman	108	90		Susan Ervin	90	80
Kimberly Gaitor	83	90		Albert Gutierrez	55	52
Leland Garner	60	65		Robin Calderon	95	100
Bryan Veilleux	53	75		Jennifer Blackburn	60	57
Mary Watts	73	86		Doris Larkin	83	85
Jerry Brown	58	60		James Miller	63	42
Jacob Ludwick	47	52		Gregory Myklebust	70	87
Wayne Vega	100	85		Rita Swinton	90	90
Kathryn Wilson	55	80		Barbara Richardson	63	82
Tony Bateman	20	11		Ora Tidmore	108	87

Regression Line Exercise

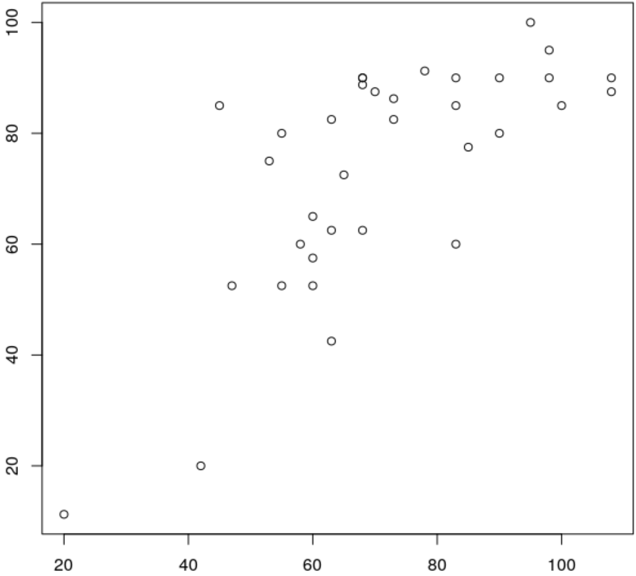
$\sum x$	2496	$\sum y$	2580
$\sum x^2$	191344	$\sum y^2$	204700
$(\sum x)^2$	6230016	$(\sum y)^2$	6656400
$\sum xy$	193904	n	34

Regression Line Exercise

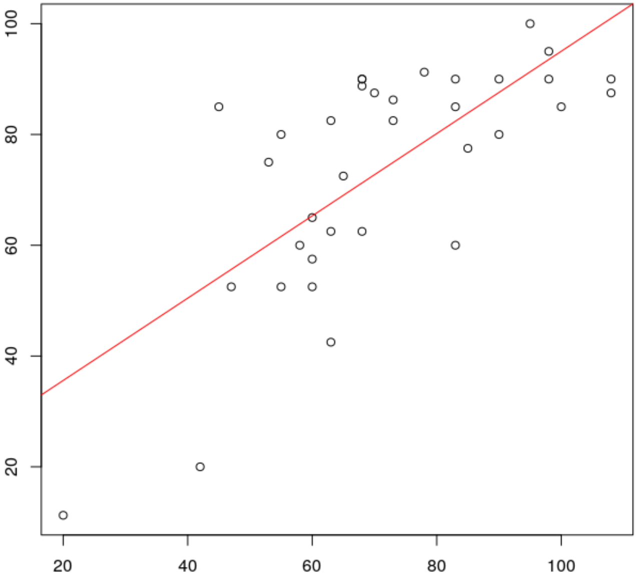
The solution for the linear correlation coefficient is $r = 0.7122366$.

The solution for the y-intercept and slope of the regression line is $b_0 = 20.7350$ and $b_1 = 0.7429$. The point estimate for William Jones' grade is 83.88.

Regression Line Exercise



Regression Line Exercise



Linear Correlation Exercise

Costs listed below are repair costs (in dollars) for cars crashed at 6 mi/h in full-front crash tests and the same cars crashed at 6 mi/h in full-rear crash tests (based on data from the Insurance Institute for Highway Safety). The cars are the Toyota Camry, Mazda 6, Volvo S40, Saturn Aura, Subaru Legacy, Hyundai Sonata, and Honda Accord. Is there sufficient evidence to conclude that there is a linear correlation between the repair costs from full-front crashes and full-rear crashes?

Front	936	978	2252	1032	3911	4312	3469
Rear	1480	1202	802	3191	1122	739	2767

Listed below are systolic blood pressure measurements (in mm Hg) obtained from the same woman (based on data from "Consistency of Blood Pressure Differences Between the Left and Right Arms," by Eguchi et al., Archives of Internal Medicine, Vol. 167). Is there sufficient evidence to conclude that there is a linear correlation between right and left arm systolic blood pressure measurements?

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

Linear Correlation Exercise

One classic application of correlation involves the association between the temperature and the number of times a cricket chirps in a minute. Listed below are the numbers of chirps in one minute and the corresponding temperatures in °F (based on data from "The Song of Insects" by George W. Pierce, Harvard University Press). Is there sufficient evidence to conclude that there is a linear correlation between the number of chirps in one minute and the temperature?

Chirps	882	1188	1104	864	1200	1032	960	900
°F	69.7	93.3	84.3	76.3	88 6	82.6	71.6	79.6

Lemons and Car Crashes. Find the best predicted crash fatality rate for a year in which there are 500 metric tons of lemon imports.

Lemon Imports	230	265	358	480	530
Crash Fatality Rate	15.9	15.7	15.4	15.3	14.9

Linear Correlation Exercise

Altitude and Temperature. At 6327 ft (or 6.327 thousand feet), Mario Triola, the author of many of these exercises, recorded the temperature. Find the best predicted temperature at that altitude. How does the result compare to the actual recorded value of 48°F?

Altitude	3	10	14	22	28	31	33
Temperature	57	37	24	-5	-30	-41	-54

End of Lesson

Next Lesson: Goodness of Fit