STA258H5

Al Nosedal
and Alison Weir

Winter 2017

SIMPLE LINEAR REGRESSION

Scatterplot

A *scatterplot* shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.

Always plot the explanatory variable, if there is one, on the horizontal axis (the x axis) of a scatterplot. As a reminder, we usually call the explanatory variable x and the response variable y. If there is no explanatory-response distinction, either variable can go on the horizontal axis.

Examining Scatterplot

In any graph of data, look for the **overall pattern** and for striking **deviations** from that pattern.
You can describe the overall pattern of a scatterplot by the **form**, **direction**, and **strength** of the relationship.
An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern of the relationship.

Positive association, negative association

Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other, and below-average values also tend to occur together.
Two variables are **negatively associated** when above-average values of one tend to accompany below-average values of the other, and vice versa.

Do heavier people burn more energy?

Metabolic rate, the rate at which the body consumes energy, is important in studies of weight gain, dieting, and exercise. We have data on the lean body mass and resting metabolic rate for 12 women who are subjects in a study of dieting. Lean body mass, given in kilograms, is a person's weight leaving out all fat. Metabolic rate is measured in calories burned per 24 hours.

The researchers believe that lean body mass is an important influence on metabolic rate. Make a scatterplot to examine this belief.

Do heavier people burn more energy?

The study of dieting described earlier collected data on the lean body mass (in kilograms) and metabolic rate (in calories) for both female and male subjects.

Mass	Rate	Sex	Mass	Rate	Sex
36.1	995	F	40.3	1189	F
54.6	1425	F	33.1	913	F
48.5	1396	F	42.4	1124	F
42.0	1418	F	34.5	1052	F
50.6	1502	F	51.1	1347	F
42.0	1256	F	41.2	1204	F

More data

Mass	Rate	Sex	Mass	Rate	Sex
51.9	1867	M	47.4	1322	M
46.9	1439	M	48.7	1614	M
62.0	1792	M	51.9	1460	M
62.9	1666	M			

a) Make a scatterplot of metabolic rate versus lean body mass for all 19 subjects. Use separate symbols to distinguish women and men. (This is a common method to compare two groups of individuals in a scatterplot)
b) Does the same overall pattern hold for both women and men? What is the most important difference between women and men?

Reading our data

```r
# Step 1. Entering data;

# url of metabolic rate data;

meta_url = "http://www.math.unm.edu/~alvaro/metabolic2.txt"

# import data in R;

data = read.table(meta_url, header = TRUE);
```

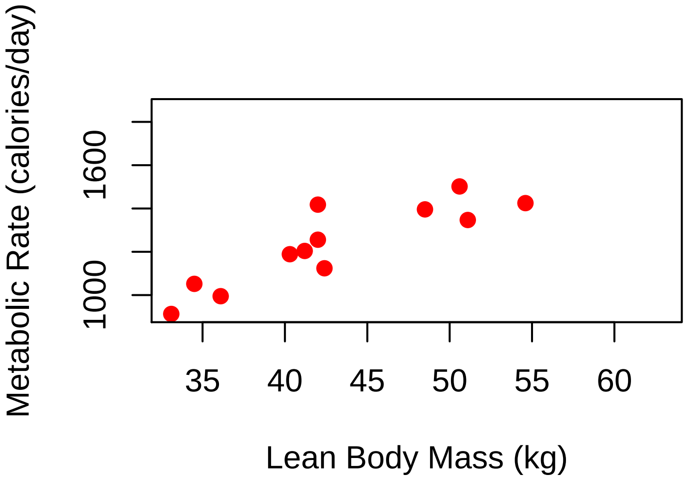
Reading our data

```
# Step 2. Formating data;

x.min=min(data$Mass);
x.max=max(data$Mass);

y.min=min(data$Rate);
y.max=max(data$Rate);

female=data[1:12, ];
male=data[13:19, ];
```

Scatterplot

```r
# Step 3. Making scatterplot;

plot(female$Mass,female$Rate,pch=19,col="red",
xlab="Lean Body Mass (kg)",
ylab="Metabolic Rate (calories/day)",
xlim=c(x.min,x.max),ylim=c(y.min,y.max));
```

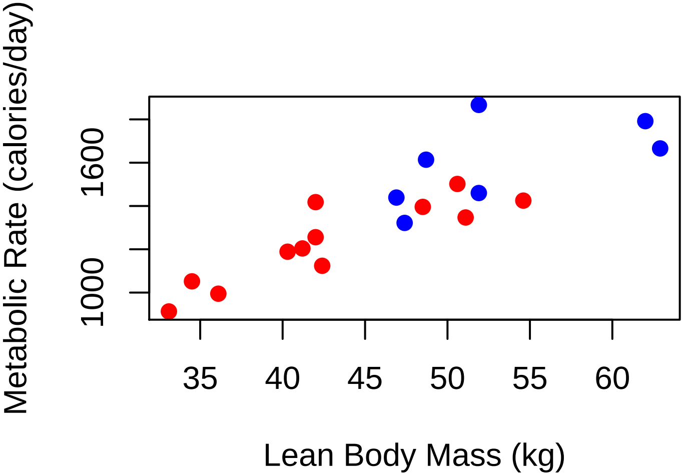
Scatterplot



Scatterplot

```
# Step 3. Making scatterplot;

plot(female$Mass,female$Rate,pch=19,col="red",
xlab="Lean Body Mass (kg)",
ylab="Metabolic Rate (calories/day)",
xlim=c(x.min,x.max),ylim=c(y.min,y.max));

points(male$Mass,male$Rate,pch=19,col='blue');

# pch=19 tells R that you want solid circles;
```

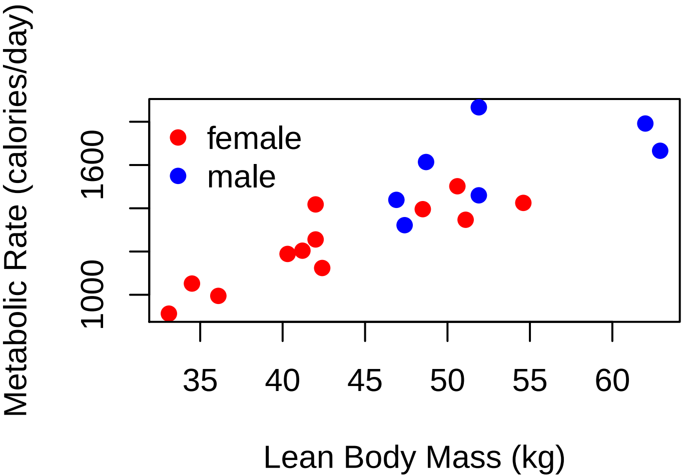
Scatterplot



Scatterplot

```r
# Step 3. Making scatterplot;

plot(female$Mass,female$Rate,pch=19,col="red",
xlab="Lean Body Mass (kg)",
ylab="Metabolic Rate (calories/day)",
xlim=c(x.min,x.max),ylim=c(y.min,y.max));

points(male$Mass,male$Rate,pch=19,col='blue');

legend("topleft",c("female","male"),pch=c(19,19),
col=c('red','blue'),bty="n");

# legend tells R that you want to add a legend to
# your graph;
# topleft, where you want to position legend;
# bty="n" NO box around legend;
```

Scatterplot



b) Solution

For both men and women, the association is linear and positive. The women's points show a stronger association. As a group, males typically have larger values for both variables (they tend to have more mass, and tend to burn more calories per day).

Correlation

The *correlation* measures the direction and strength of the linear relationship between two quantitative variables. Correlation is usually written as r.

Suppose that we have data on variables x and y for n individuals. The values for the first individual are x_1 and y_1, the values for the second individual are x_2 and y_2, and so on.

The means and standard deviations of the two variables are $\bar{x}$ and S_x for the x-values, and $\bar{y}$ and S_y for the y-values. The correlation r between x and y is

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_x} \right) \left(\frac{y_i - \bar{y}}{S_y} \right)$$

Facts about correlation

The correlation r measures the strength and direction of the linear association between two quantitative variables x and y. Although you calculate a correlation for any scatterplot, **r measures only straight-line relationships**.

Correlation indicates the direction of a linear relationship by its sign: $r > 0$ for a positive association and $r < 0$ for a negative association. Correlation always satisfies $-1 \leq r \leq 1$ and indicates the strength of a relationship by how close it is to -1 or 1. Perfect correlation, $r = \pm 1$, occurs only when the points on a scatterplot lie exactly on a straight line.

Coral reefs

This example is about a study in which scientists examined data on mean sea surface temperatures (in degrees Celsius) and mean coral growth (in millimeters per year) over a several-year period at locations in the Red Sea. Here are the data:

Sea Surface Temperature	Growth
29.68	2.63
29.87	2.58
30.16	2.60
30.22	2.48
30.48	2.26
30.65	2.38
30.90	2.26

a) Make a scatterplot. Which is the explanatory variable?
b) Find the correlation r step-by-step. Explain how your value for r matches your graph in a).
c) Enter these data into your calculator and use the correlation function to find r (or use R to find r).

Reading our data

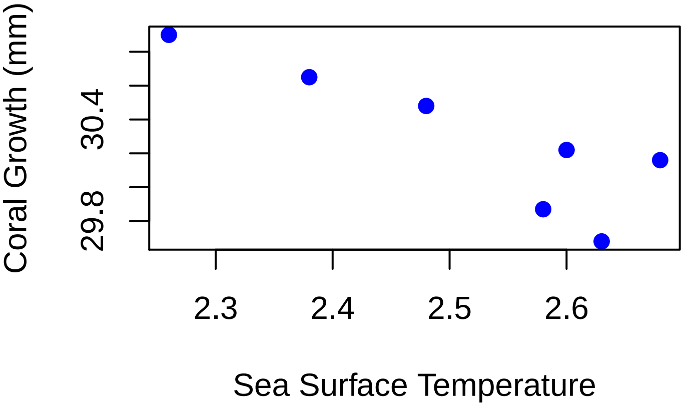
```
# Step 1. Entering data;

# url of coral rate data;

coral_url = "http://www.math.unm.edu/~alvaro/coral.txt"

# import data in R;

data = read.table(coral_url, header = TRUE);
```

Scatterplot

Temperature is the explanatory variable.

```r
# Step 2. Making scatterplot;

plot(data,xlab="Sea Surface Temperature",
ylab="Coral Growth (mm)",pch=19,col="blue");
```

Scatterplot



$\bar{x}$ and $\bar{y}$

First, let's find $\bar{x}$ and $\bar{y}$

$$\bar{x} = \frac{29.68 + 29.87 + 30.16 + 30.22 + 30.48 + 30.65 + 30.90}{7} = 30.28$$

$$\bar{y} = \frac{2.63 + 2.58 + 2.60 + 2.48 + 2.26 + 2.38 + 2.26}{7} = 2.4557$$

S_x

Now, let's find S_x and S_y

$$S_x^2 = \frac{(29.68 - 30.28)^2 + ... + (30.65 - 30.28)^2 + (30.90 - 30.28)^2}{6}$$

$$S_x^2 = 0.1845$$

$$S_x = 0.4296$$

S_y

$$S_y^2 = \frac{(2.63 - 2.4557)^2 + ... + (2.38 - 2.4557)^2 + (2.26 - 2.4557)^2}{6}$$

$$S_y^2 = 0.0249$$

$$S_y = 0.1578$$

Sample Covariance

Finally, we find the sample covariance and r

$$\begin{aligned}\sum_{i=1}^{7} x_i y_i &= (29.68)(2.63) + ... + (30.65)(2.38) + (30.90)(2.26) \\ &= 520.1504\end{aligned}$$

$$\begin{aligned}\textit{Sample covariance} &= \frac{\sum_{i=1}^{n} x_i y_i}{n-1} - \frac{n\bar{x}\bar{y}}{n-1} \\ &= \frac{520.1504}{6} - \frac{(7)(30.28)(2.4557)}{6} \\ &= 86.6917 - 86.7516 = -0.0599\end{aligned}$$

Correlation

$$r = \frac{\text{Sample Covariance}}{S_x S_y} = \frac{-0.0599}{(0.4296)(0.1578)} = -0.8835$$

This is consistent with the strong, negative association depicted in the scatterplot.

c) R will give a value of $r = -0.8635908$.

R Code

```
growth=data[ ,1];

# data[ ,1] gives you the first column of data;

temp=data[ ,2];

# data[ ,2] gives you the 2nd column of data;

cov(growth, temp);

## [1] -0.05593333

cor(growth, temp);

## [1] -0.8635908
```

Regression Line

A *regression line* is a straight line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the value of y for a given value of x.

Equation of the Least-Squares Regression Line

We have data on an explanatory variable x and a response variable y for n individuals. From the data, calculate the means $\bar{x}$ and $\bar{y}$ and the standard deviations S_x and S_y of the two variables, and their correlation r. The least-squares regression line is the line

$$\hat{y} = a + bx$$

with *slope*

$$b = r\frac{S_y}{S_x}$$

and *intercept*

$$a = \bar{y} - b\bar{x}$$

Least-Squares Regression Line

The **least-squares regression line** of y on x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

Do heavier people burn more energy?

We have data on the lean body mass and resting metabolic rate for 12 women who are subjects in a study of dieting. Lean body mass, given in kilograms, is a person's weight leaving out all fat. Metabolic rate, in calories burned per 24 hours, is the rate at which the body consumes energy.

Mass	Rate	Mass	Rate
36.1	995	40.3	1189
54.6	1425	33.1	913
48.5	1396	42.4	1124
42.0	1418	34.5	1052
50.6	1502	51.1	1347
42.0	1256	41.2	1204

a) Make a scatterplot that shows how metabolic rate depends on body mass. There is a quite strong linear relationship, with correlation $r = 0.876$.

b) Find the least-squares regression line for predicting metabolic rate from body mass. Add this line to your scatterplot.

c) Explain in words what the slope of the regression line tells us.

d) Another woman has a lean body mass of 45 kilograms. What is her predicted metabolic rate?

Solutions

b) the regression equation is

$$\hat{y} = 201.2 + 24.026x$$

where $y=$ metabolic rate and $x=$ body mass.
c) The slope tells us that on the average, metabolic rate increases by about 24 calories per day for each additional kilogram of body mass.
d) For $x = 45$ kg, the predicted metabolic rate is
$\hat{y} = 1282.4$ calories per day.

Reading our data

```
# Step 1. Entering data;

# url of metabolic rate data;

meta_url = "http://www.math.unm.edu/~alvaro/metabolic2.txt"

# import data in R;

data = read.table(meta_url, header = TRUE);
```

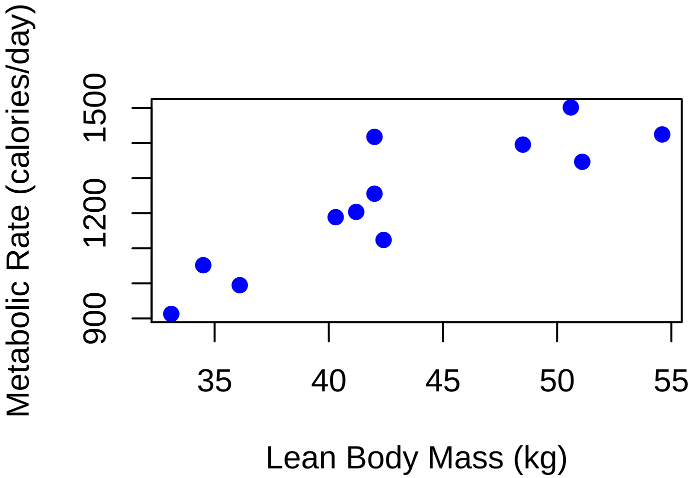
Reading our data

```
# Step 2. Formating data;

female = data[1:12, ];

mass = female$Mass;

rate = female$Rate;
```

R code

```r
# Step 2. Making scatterplot;

plot(mass, rate ,pch=19,col="blue",
xlab="Lean Body Mass (kg)",
ylab="Metabolic Rate (calories/day)");
```

Scatterplot



Regression Equation (R Code)

```
# Step 3. Finding Regression Equation;

metabolic.reg=lm(rate~mass);
```

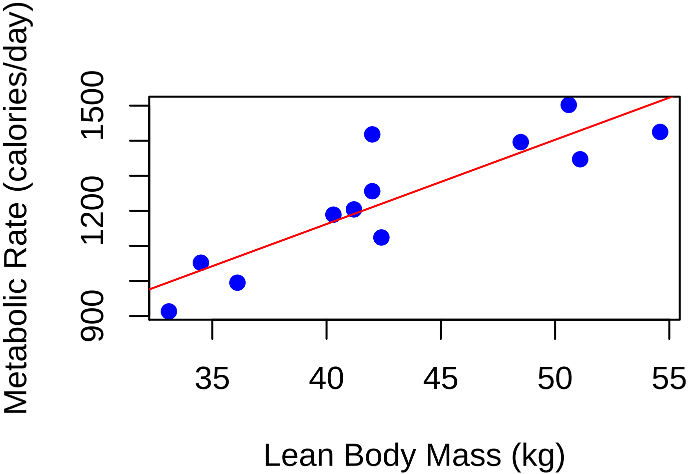
a and b

```
metabolic.reg$coef;

## (Intercept)        mass
##   201.16160    24.02607
```

Scatterplot with least-squares line

```r
plot(mass,rate,
pch=19,col="blue", xlab="Lean Body Mass (kg)",
ylab="Metabolic Rate (calories/day)");

abline(metabolic.reg$coef, col="red");
```

Scatterplot with least-squares line



Prediction

```
new<-data.frame(mass=45);

predict(metabolic.reg,newdata=new);

##        1
## 1282.335
```

Do heavier people burn more energy?

Return to the example about lean body mass and metabolic rate. We will use these data to illustrate influence.
a) Make a scatterplot of the data that is suitable for predicting metabolic rate from body mass, with two new points added. Point A: mass 42 kilograms, metabolic rate 1500 calories. Point B: mass 70 kilograms, metabolic rate 1400 calories. In which direction is each of these points an outlier?
b) Add three least-squares regression lines to your plot: for the original 12 women, for the original women plus Point A, and for the original women plus Point B. Which new point is more influential for the regression line? Explain in simple language why each new point moves the line in the way your graph shows.

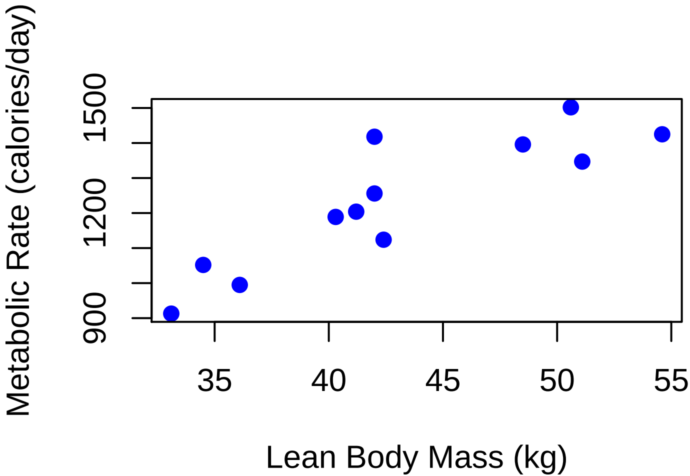
Reading our data

```r
# Step 1. Entering data;

# url of metabolic rate data;

meta_url = "http://www.math.unm.edu/~alvaro/metabolic.txt"

# import data in R;

data = read.table(meta_url, header = TRUE);
```

Scatterplot

```
plot(data,pch=19,col="blue",
xlab="Lean Body Mass (kg)",
ylab="Metabolic Rate (calories/day)");
```

Scatterplot



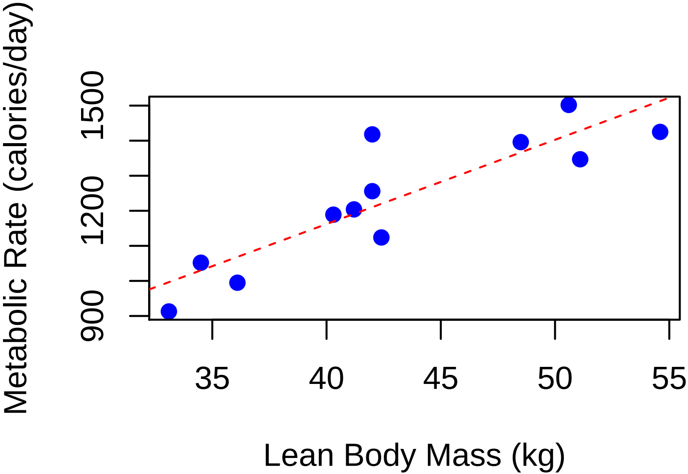
Least-Squares Regression Line

```
# Step 3. Finding L-S Regression Line;

mod=lm(data$Rate~data$Mass);
```

Scatterplot + L-S Regression Line

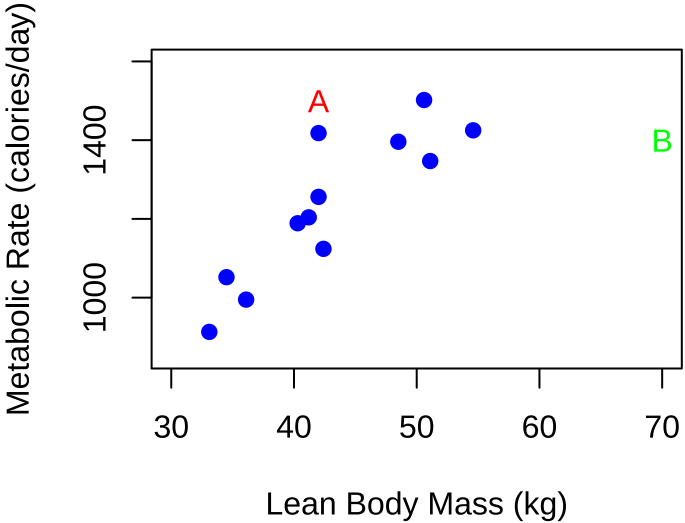
```
plot(data,pch=19,col="blue",
xlab="Lean Body Mass (kg)",
ylab="Metabolic Rate (calories/day)");

abline(mod$coeff,col="red",lty=2);

# abline tells R to add a line to your
# scatterplot;
# lty= 2 is used to draw a dashed-line;
```

Scatterplot + L-S Regression Line



Scatterplot + A +B

```r
plot(data,pch=19,col="blue",
xlab="Lean Body Mass (kg)",
ylab="Metabolic Rate (calories/day)",
xlim=c(30,70),ylim=c(850,1600 ));

points(42,1500,pch="A",col="red");
#point A;

points(70,1400,pch="B",col="green");
#point B;
```

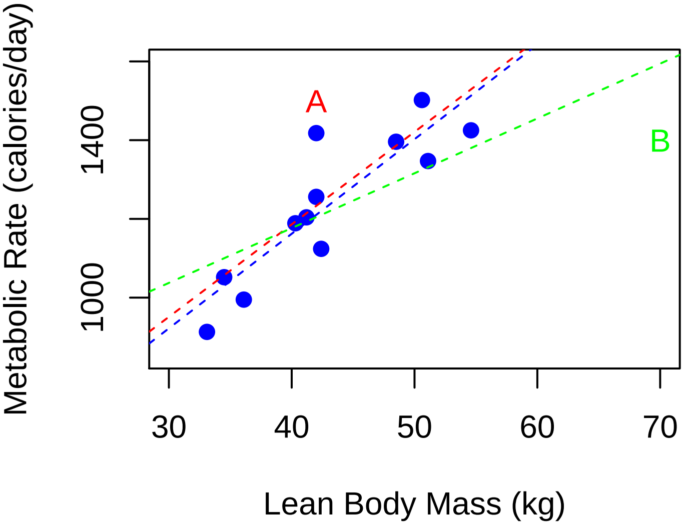

A
B
Metabolic Rate (calories/day)
1000
1400
30
40
50
60
70
Lean Body Mass (kg)

Least-Squares Regression Lines

```
# Step 3. Finding L-S Regression Line;

mod=lm(data$Rate~data$Mass);
# original;

modA=lm(c(data$Rate,1500)~c(data$Mass,42));
# point A;

modB=lm(c(data$Rate,1400)~c(data$Mass,70));
# point B;
```

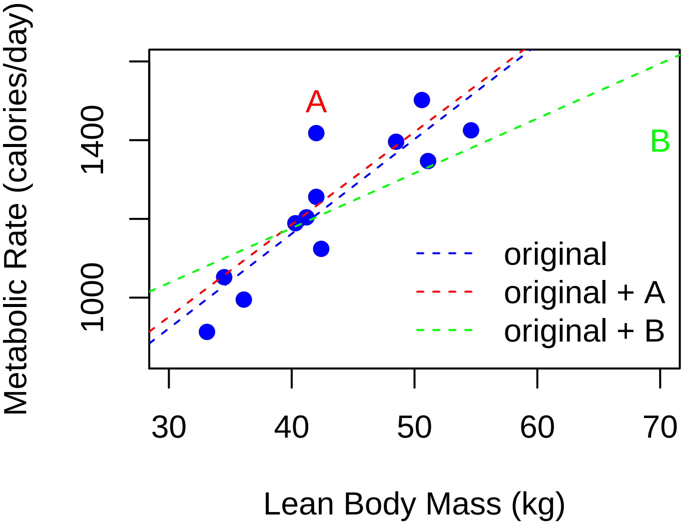
Scatterplot + A +B + L-S Regression Lines

```
plot(data,pch=19,col="blue",
xlab="Lean Body Mass (kg)",
ylab="Metabolic Rate (calories/day)",
xlim=c(30,70),ylim=c(850,1600 ));

points(42,1500,pch="A",col="red");
points(70,1400,pch="B",col="green");

abline(mod$coeff,col="blue",lty=2);
abline(modA$coeff,col="red",lty=2);
abline(modB$coeff,col="green",lty=2);
```


A
B
Metabolic Rate (calories/day)
1000
1400
30
40
50
60
70
Lean Body Mass (kg)

Adding a legend

```
legend("bottomright",
c("original","original + A","original + B"),
col=c("blue","red","green"),
lty=c(2,2,2),bty="n");
```


A
B
original
original + A
original + B
1000
1400
30
40
50
60
70
Metabolic Rate (calories/day)
Lean Body Mass (kg)

Solutions

a) Point A lies above the other points; that is, the metabolic rate is higher than we expect for the given body mass. Point B lies to the right of the other points; that is, it is an outlier in the x (mass) direction, and the metabolic rate is lower than we would expect.
b) In the plot, the dashed blue line is the regression line for the original data. The dashed red line slightly above that includes Point A; it has a very similar slope to the original line, but a slightly higher intercept, because Point A pulls the line up. The third line includes Point B, the more influential point; because Point B is an outlier in the x direction, it "pulls" the line down so that it is less steep.

Influential observations

An observation is *influential* for a statistical calculation if removing it would markedly change the result of the calculation.
The result of a statistical calculation may be of little practical use if it depends strongly on a few influential observations.
Points that are outliers in either the x or the y direction of a scatterplot are often influential for the correlation. Points that are outliers in the x direction are often influential for the least-squares regression line.

Another example

There is some evidence that drinking moderate amounts of wine helps prevent heart attacks. A table shown below gives data on yearly wine consumption (liters of alcohol from drinking wine, per person) and yearly deaths from heart disease (deaths per 100,000 people) in 19 developed nations*.

Another example

a) Make a scatterplot that shows how national wine consumption helps explain heart disease death rates.
b) Describe the form of the relationship. Is there a linear pattern? How strong is the relationship?
c) Is the direction of the association positive or negative? Explain in simple language what this says about wine and heart disease. Do you think these data give good evidence that drinking wine **causes** a reduction in heart disease deaths? Why?

Table

Country	Alcohol from wine	Heart disease deaths
Australia	2.5	211
Austria	3.9	167
Belgium	2.9	131
Canada	2.4	191
Denmark	2.9	220
Finland	0.8	297
France	9.1	71
Iceland	0.8	211
Ireland	0.7	300
Italy	7.9	107
Netherlands	1.8	167
New Zealand	1.8	266
Norway	0.8	227
Spain	6.5	86
Sweden	1.6	207
Switzerland	5.8	115
United Kingdom	1.3	285
United States	1.2	199
West Germany	2.7	172

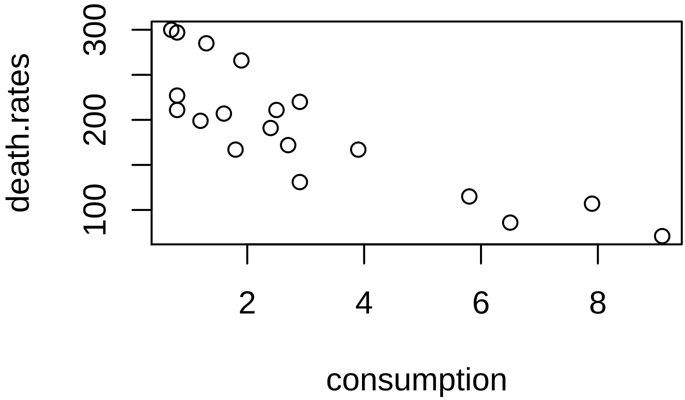
Solution (Bar chart)

```
# Step 1. Entering data;

consumption=c(2.5, 3.9, 2.9, 2.4, 2.9, 0.8, 9.1,
0.8, 0.7, 7.9, 1.8, 1.9, 0.8, 6.5, 1.6, 5.8, 1.3, 1.2, 2.7);

death.rates=c(211, 167, 131, 191, 220, 297, 71,
211, 300, 107,167, 266, 227, 86, 207, 115, 285, 199, 172);
```

Scatterplot (R code)

```r
plot(consumption,death.rates);
```

Scatterplot (R code)



Another example (cont.)

Our table gives data on wine consumption and heart disease death rates in 19 countries. A scatterplot shows a moderately strong relationship.
a) The correlation for these variables is $r = -0.843$. What does a negative correlation say about wine consumption and heart disease deaths?
b) The least-squares regression line for predicting heart disease death rate from wine consumption is

$$\hat{y} = 260.56 - 22.969x$$

Verify this using R. Then use this equation to predict the heart disease death rate in another country where adults average 4 liters of alcohol from wine each year.

a) Finding correlation

```
cor(consumption,death.rates);

## [1] -0.8428127
```

Least-squares Regression Line

```
explanatory<-consumption;

response<-death.rates;

wine.reg<-lm(response~explanatory);
```

R code

```
names(wine.reg);

##   [1] "coefficients"  "residuals"         "effects"           "rank'
##   [5] "fitted.values" "assign"            "qr"                "df.re
##   [9] "xlevels"       "call"              "terms"             "model
```

a and b

```
wine.reg$coef;

## (Intercept) explanatory
##   260.56338   -22.96877
```

Prediction

```
wine.reg$coef[1]+wine.reg$coef[2]*4;

## (Intercept)
##    168.6883
```

Prediction (again...)

```
new=data.frame(explanatory=4);

predict(wine.reg,newdata=new);

##        1
## 168.6883
```

c) The association is negative: Countries with high wine consumption have fewer heart disease deaths, while low wine consumption tends to go with more deaths from heart disease. This does not prove causation; there may be some other reason for the link.

Our main example

One effect of global warming is to increase the flow of water into the Arctic Ocean from rivers. Such an increase might have major effects on the world's climate. Six rivers (Yenisey, Lena, Ob, Pechora, Kolyma, and Severnaya Dvina) drain two-thirds of the Arctic in Europe and Asia. Several of these are among the largest rivers on earth. File arctic-rivers.txt contains the total discharge from these rivers each year from 1936 to 1999[2]. Discharge is measured in cubic kilometers of water.

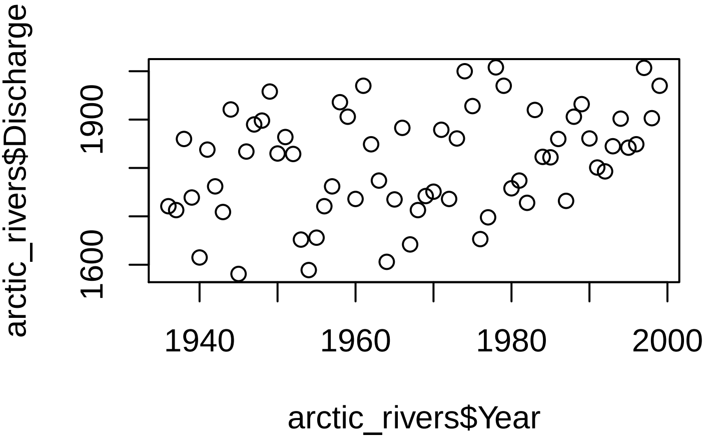
Reading our data

```r
# url of arctic rivers data;

riv_url = "http://www.math.unm.edu/~alvaro/arctic-rivers.txt"

# import data in R;

arctic_rivers = read.table(riv_url, header = TRUE);
```

Scatterplot (R code)

```r
plot(arctic_rivers$Year,arctic_rivers$Discharge);
```

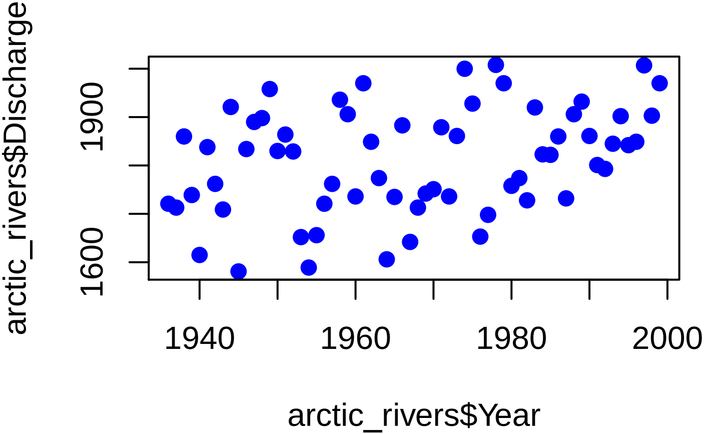
Scatterplot (R code)



Scatterplot (R code)

```r
plot(arctic_rivers$Year,arctic_rivers$Discharge,
pch=19,col="blue");
```

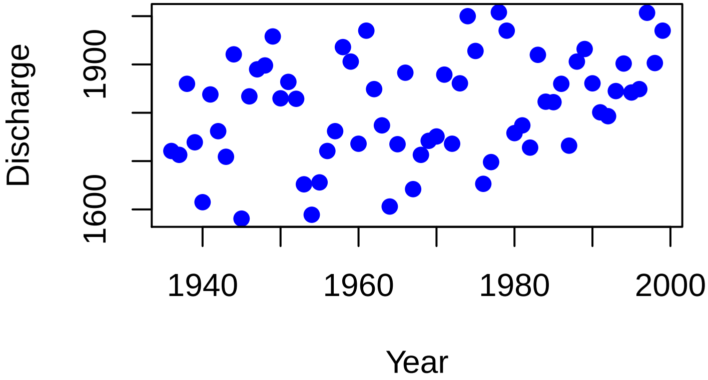
Scatterplot (R code)



Scatterplot (R code)

```r
plot(arctic_rivers$Year,arctic_rivers$Discharge,
pch=19,col="blue", xlab="Year",
ylab="Discharge");
```

Scatterplot (R code)



The scatterplot shows a weak positive, linear relationship.

Our main example

```
r<-cor(arctic_rivers$Year,arctic_rivers$Discharge);

r;

## [1] 0.3343926
```

The scatterplot shows a weak positive, linear relationship, which is confirmed by r (0.3343926).

R code

```
explanatory<-arctic_rivers$Year;

response<-arctic_rivers$Discharge

rivers.reg<-lm(response~explanatory);
```

R code

```
names(rivers.reg);

##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.re
##  [9] "xlevels"       "call"          "terms"         "model
```

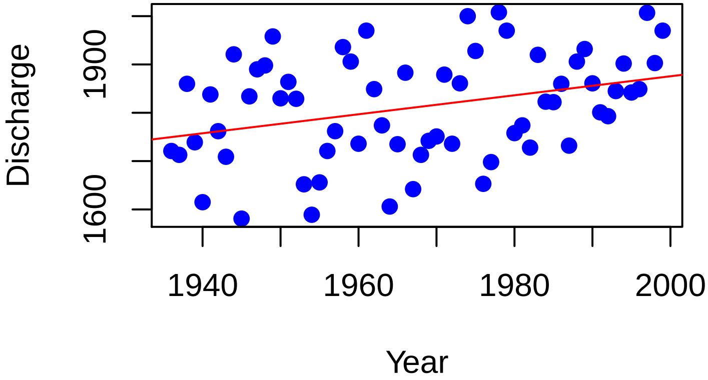
a and b

```
rivers.reg$coef;

##  (Intercept)  explanatory
## -2056.769460     1.966163
```

Scatterplot with least-squares line

```
plot(explanatory,response,
pch=19,col="blue", xlab="Year",
ylab="Discharge");

abline(rivers.reg$coef, col="red");
```

Scatterplot with least-squares line



Residuals

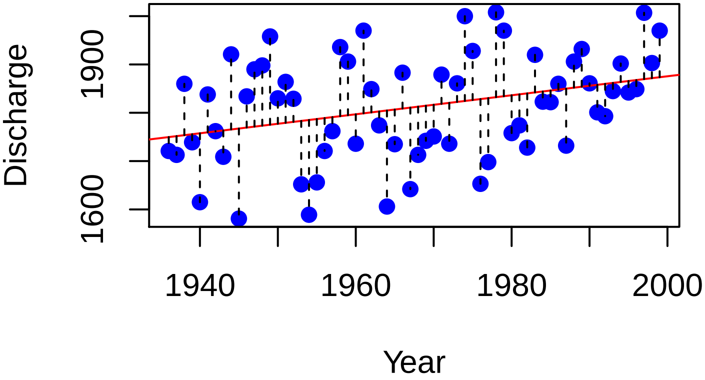
A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is,

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

Scatterplot with residual line segments

```
plot(explanatory,response,
pch=19,col="blue", xlab="Year",
ylab="Discharge");

abline(rivers.reg$coef, col="red");

segments(explanatory, fitted(rivers.reg),
explanatory,response, lty=2, col="black");
```

Scatterplot with residual line segments



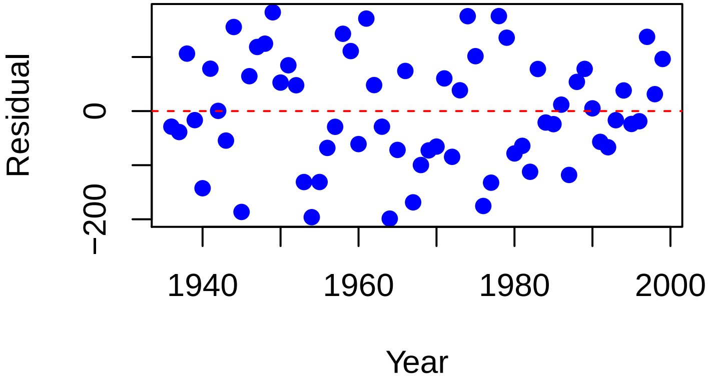
Residual Plots

A **residual plot** is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess the fit of a regression line.
A residual plot magnifies the deviations of the points from the line and makes it easier to see unusual observations and patterns.

Residual plot

```
plot(explanatory,resid(rivers.reg),
pch=19,col="blue", xlab="Year",
ylab="Residual");

abline(h=0, col="red",lty=2);
```

Residual plot



INFERENCE FOR REGRESSION.

The Regression Model

We have n observations on an explanatory variable x and a response variable y. Our goal is to study or predict the behavior of y for given values of x.

- For any fixed value of x, the response y varies according to a Normal distribution. Repeated measures y are independent of each other.
- The mean response μ_y has a straight-line relationship with x: $\mu_y = \alpha + \beta x$. The slope β and intercept α are **unknown** parameters.
- The standard deviation of y (call it σ) is the same for all values of x. The value of σ is **unknown**.
  The regression model has three parameters, α, β, and σ.

Regression Standard Error

The **regression standard error** is

$$s = \sqrt{\frac{1}{n-2} \sum \text{residual}^2} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y})^2}.$$

Use s to estimate the **unknown** σ in the regression model.

Regression Standard Error

```
residuals<-resid(rivers.reg);

n<-length(residuals);

s<-sqrt(sum(residuals^2)/(n-2));

s;

## [1] 104.0026
```

Regression Standard Error (another way)

```
summary(rivers.reg)$sigma

## [1] 104.0026
```

Confidence intervals for the regression slope

A level C confidence interval for the slope β of the true regression line is

$$b \pm t^* SE_b.$$

In this formula, the standard error of the least-squares slope b is

$$SE_b = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

and t^* is the critical value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

Example

We will use the data in arctic-rivers.txt to give a 90% confidence interval for the slope of the true regression of Arctic river discharge on year.

Confidence interval for slope

```
summary(rivers.reg)$coef

##                  Estimate    Std. Error   t value   Pr(>|t|)
## (Intercept) -2056.769460 1384.6873683 -1.485367 0.14251371
## explanatory     1.966163    0.7037491  2.793841 0.00692068
```

Confidence interval for slope

```
b<-summary(rivers.reg)$coef[2,1];

SEb<-summary(rivers.reg)$coef[2,2];

# lower bound;
b-qt(0.95,df=n-2)*SEb;

## [1] 0.7910398

# upper bound;
b+qt(0.95,df=n-2)*SEb;

## [1] 3.141286
```

R output gives $b = 1.966163$ and $SE_b = 0.7037491$. There were $n = 64$ observations, so $df = 62$. Our 90% Confidence Interval for β is given by (0.7910398, 3.1412862). Because this interval does not contain 0, we have evidence that β (the rate at which discharge is increasing) is positive.

Testing the hypothesis of no linear relationship

We can also test hypotheses about the slope β. The most common hypothesis is

$$H_0 : \beta = 0.$$

A regression line with slope 0 is horizontal. That is, the mean of y does not change at all when x changes. So this H_0 says that there is no true linear relationship between x and y.

Significance test for regression slope

To test the hypothesis $H_0 : \beta = 0$, compute the t statistic

$$t = \frac{b}{SE_b}.$$

In terms of a random variable T having the $t(n-2)$ distribution, the P-value for a test of H_0 against $H_a : \beta \neq 0$ is $2P(T \geq |t|)$.

Our main example

The most important question we ask of the data in arctic-rivers.txt is this: Is the increasing trend visible in your plot statistically significant? If so, changes in the Arctic may already be affecting earth's climate. Use R to answer this question.

```
t.statistic<-b/SEb;

t.statistic;

## [1] 2.793841

p.value<-2*(1-pt(t.statistic,n-2));

p.value;

## [1] 0.00692068
```

```
summary(rivers.reg)$coef

##                 Estimate   Std. Error   t value   Pr(>|t|)
## (Intercept) -2056.769460 1384.6873683 -1.485367 0.14251371
## explanatory     1.966163    0.7037491  2.793841 0.00692068
```

The t statistic for testing $H_0 : \beta = 0$ is therefore $t = 2.7938409$. This has $df = 62$; R gives a P-value of 0.0069207. There is significant evidence (at $\alpha = 0.01$ significance level) that β is nonzero.

APPENDIX

Simple Linear Regression (SLR) Model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Y: Response or dependent variable

X: Predictor or independent variable, treat as fixed.

β_0 and β_1: Regression coefficients.

ϵ: Random error.

Goal: To be able to predict y for a given value of x.

Assumptions about ϵ

Error terms

- Are uncorrelated
- Have mean zero
- Have constant variance
- Don't require Normal distribution

$$E(\epsilon) = 0 \rightarrow E(Y) = E(\beta_0 + \beta_1 X + \epsilon) = \beta_0 + \beta_1 X \quad \text{"true line''}$$

The least-squares procedure for fitting a line through a set of *n* data points is similar to the method that we might use if we fit a line by eye; that is, we want the differences between the observed values and corresponding points on the fitted line to be "small" according to some criterion. A convenient way to accomplish this is to minimize the sum of squares of the vertical deviations from the fitted line.

Thus, if

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

is the predicted value of the ith y value, then the deviation of the observed value y_i from $\hat{y}_i$ is the difference $y_i - \hat{y}_i$ and the sum of squares of deviations to be minimized is

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2.$$

The quantity SSE is also called the **sum of squares for error**.

If SSE possesses a minimum, it will occur for values of β_0 and β_1 that satisfy the equations, $\frac{\partial SSE}{\partial \hat{\beta}_0} = 0$ and $\frac{\partial SSE}{\partial \hat{\beta}_1} = 0$. These equations are called the **least-squares equations** for estimating the parameters of a line.

You can verify that the solutions are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}.$$

(Further, it can be shown that the simultaneous solution for the two least-squares equations yields values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize SSE. I leave this for you to prove).

Least-Squares Estimators for Simple Linear Regression Model

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}},$$

where $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ and $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Fitted Values and Residuals

Fitted Value:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Residual:

$$\hat{\epsilon} = y_i - \hat{y}_i$$

Estimating $E(Y)$

- Point estimate $\hat{E(Y)} = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- Applies for any x, even ones we didn't observe
- Fitted values $\hat{y}_i$ estimate means $E(Y_i)$

Estimating σ^2

- Residuals $\hat{\epsilon}_i = e_i$ estimate errors ϵ_i
- Estimate of the common variance $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$
- The denominator is $n - 2$ to make it an un unbiased estimator
- Often called MSE, Mean Square Error
- Square root of MSE called residual standard error, or standard error of regression

Distribution of $\hat{\beta}_1$

$$\hat{\beta}_1 \text{ has a } N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

Sums of Squares

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$SSTO = SSR + SSE$$

Sum of Squares (Total) = Sum of Squares (Regression) + Sum of Squares (Error)

ANOVA Table

Analysis of Variance (ANOVA) Table.

Source	df	SS	MS	F
Regression	1	SSR	MSR	$F^* = MSR/MSE$
Error/Residual	n-2	SSE	MSE	
Total	n-1	SSTO		

ANOVA Table

Analysis of Variance (ANOVA) Table.

- $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$
- Test Statistic: $F^* = \text{MSR}/\text{MSE}$
- Reference distribution: $F_{1,n-2}$
- Note. $F^* = t_*^2$

Coefficient of Determination

The statistic r^2 is called the **coefficient of determination** and has an interesting and useful interpretation.
r^2 can be interpreted as the proportion of the total variation in the y_i's that is explained by the variable x in a simple linear regression model (for more details, see page 601).

```
summary(rivers.reg)$fstatistic;

##     value     numdf     dendf
##  7.805547  1.000000 62.000000

summary(rivers.reg)$r.squared;

## [1] 0.1118184
```

Another way would be typing in

```
summary(rivers.reg)
```

(that would show you everything but I don't have enough room here...)