Linear Regression

Al Nosedal
University of Toronto

Summer 2019

My momma always said: "Life was like a box of chocolates. You never know what you're gonna get."

Forrest Gump.

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.

Review of Straight Lines

Suppose that y is a response variable (plotted on the vertical axis) and x is an explanatory variable (plotted on the horizontal axis). A straight line relating y to x has an equation of the form

$$y = a + bx$$

In this equation, b is the *slope*, the amount by which y changes when x increases by one unit. The number a is the *intercept*, the value of y when $x = 0$.

City mileage, highway mileage

We expect a car's highway gas mileage (mpg) to be related to its city gas mileage. Data for all 1040 vehicles in the government's 2010 Fuel Economy Guide give the regression line
highway mpg $= 6.554 + (1.016 \times$ city mpg)
for predicting highway mileage from city mileage.
a) What is the slope of this line? Say in words what the numerical value of the slope tells you.
b) What is the intercept? Explain why the value of the intercept is not statistically meaningful.
c) Find the predicted highway mileage for a car that gets 16 miles per gallon in the city. Do the same for a car with a city mileage of 28 mpg.

Solutions

a) The slope is 1.016. On average, highway mileage increases by 1.016 mpg for each additional 1 mpg change in city mileage.
b) The intercept is 6.554 mpg. This is the highway mileage for a nonexistent car that gets 0 mpg in the city. Although this interpretation is valid, such prediction would be invalid because it involves considerable extrapolation.
c) For a car that gets 16 mpg in the city, we predict highway mileage to be:

$$6.554 + (1.016)(16) = 22.81 \text{ mpg}.$$

For a car that gets 28 mpg in the city, we predict highway mileage to be:

$$6.554 + (1.016)(28) = 35.002 \text{ mpg}.$$

What's the line?

You use the same bar of soap to shower each morning. The bar weighs 80 grams when it is new. Its weight goes down by 5 grams per day on the average. What is the equation of the regression line for predicting weight from days of use?

Solution

The equation is:

$$\text{weight} = 80 - 5 \times \text{days}$$

The intercept is 80 grams (the initial weight), and the slope is -5 grams/day.

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.

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*

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

Coral reefs

We have previously discussed 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) Use your calculator to find the mean and standard deviation of both sea surface temperature x and growth y and the correlation r between x and y. Use these basic measures to find the equation of the least-squares line for predicting y from x.
b) Enter the data into your software or calculator and use the regression function to find the least-squares line. The result should agree with your work in a) up to roundoff error.

Solutions

a) $\bar{x} = 30.28$ $S_x = 0.4296$
$\bar{y} = 2.4557$ $S_y = 0.1578$
$r = -0.8914$.
Hence,

$$b = r\frac{S_y}{S_x} = (-0.8914)\left(\frac{0.1578}{0.4296}\right) = -0.3274$$

$$a = \bar{y} - b\bar{x} = 2.4557 - (-0.3274)(30.28) = 12.3693$$

b) Slope = - 0.3276 and intercept = 12.3758.

Reading our data

```
# Step 1. Entering data;

# url of coral growth data;

coral_url=
"https://mcs.utm.utoronto.ca/~nosedal/data/coral.txt"

# importing data into R;

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

Least-squares Regression Line

```
response=data$Coral_growth;

explanatory=data$Avg_summer;

coral.reg=lm(response~explanatory);
```

Means

```
# Finding means;

mean(response);

## [1] 2.515714

mean(explanatory);

## [1] 30.28
```

Standard deviations and r

```
# Finding standard deviations and r;

sd(response);

## [1] 0.15076

sd(explanatory);

## [1] 0.4296122

cor(explanatory, response);

## [1] -0.8635908
```

R code

```
names(coral.reg);

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

a and b

```
coral.reg$coef;

## (Intercept) explanatory
##   11.6921347   -0.3030522
```

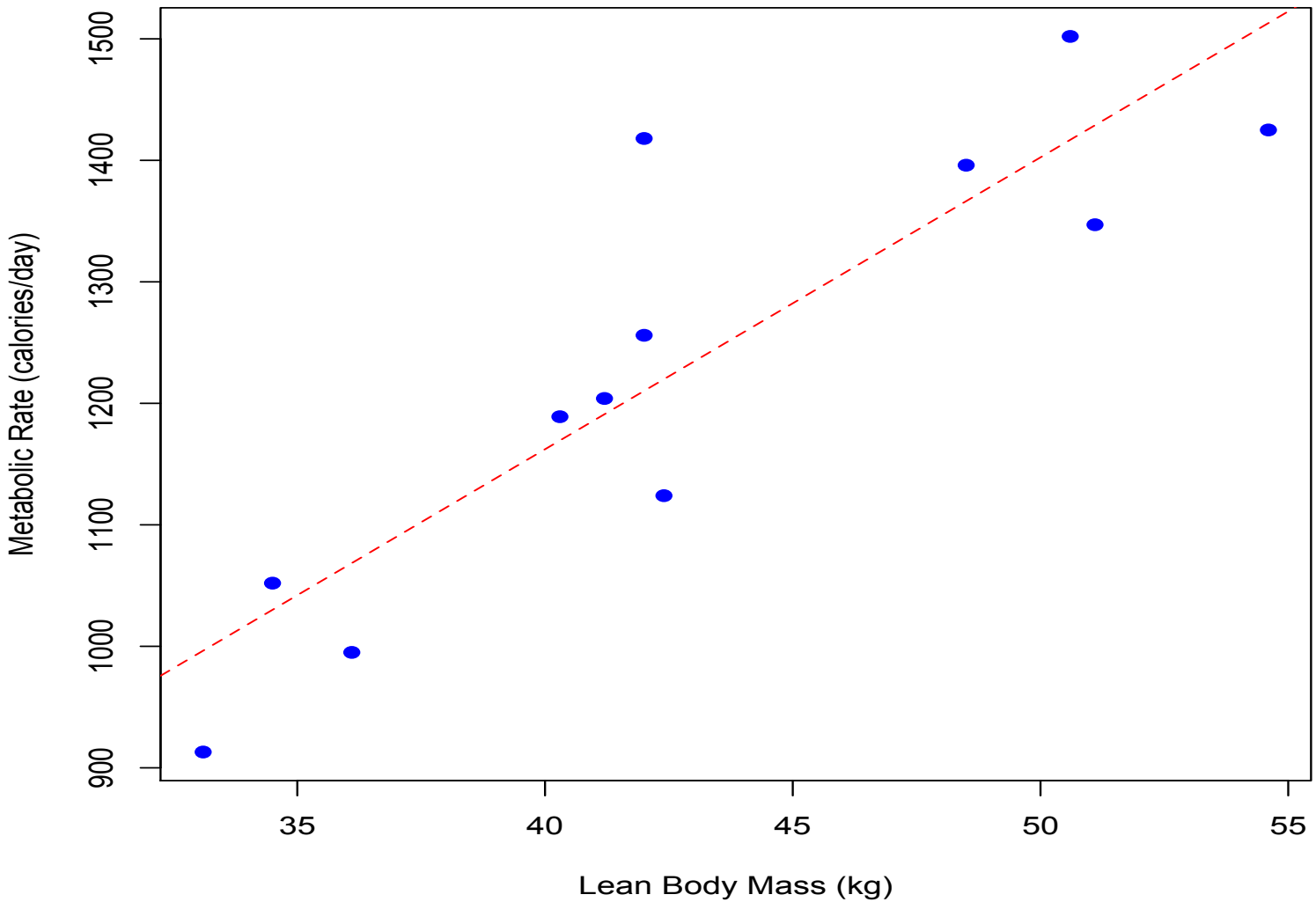
Do heavier people burn more energy?

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?

Scatterplot



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.

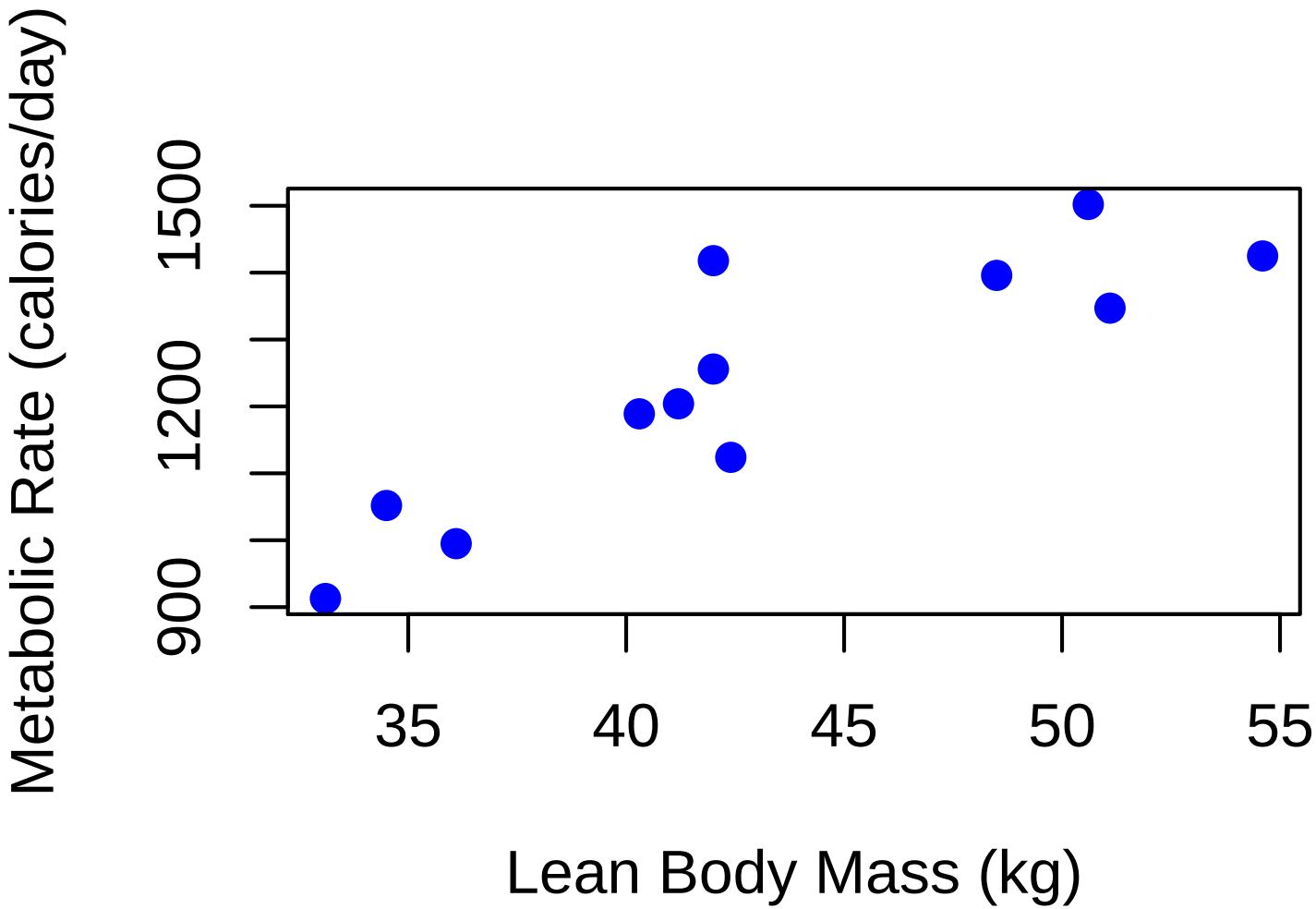
R code

```
# Step 1. Entering data;

mass=c(36.1, 54.6, 48.5, 42.0, 50.6, 42.0,
40.3, 33.1, 42.4,  34.5, 51.1, 41.2);

rate=c(995, 1425, 1396, 1418, 1502, 1256,
1189, 913, 1124, 1052, 1347, 1204);
```

R code

```
# 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);
```

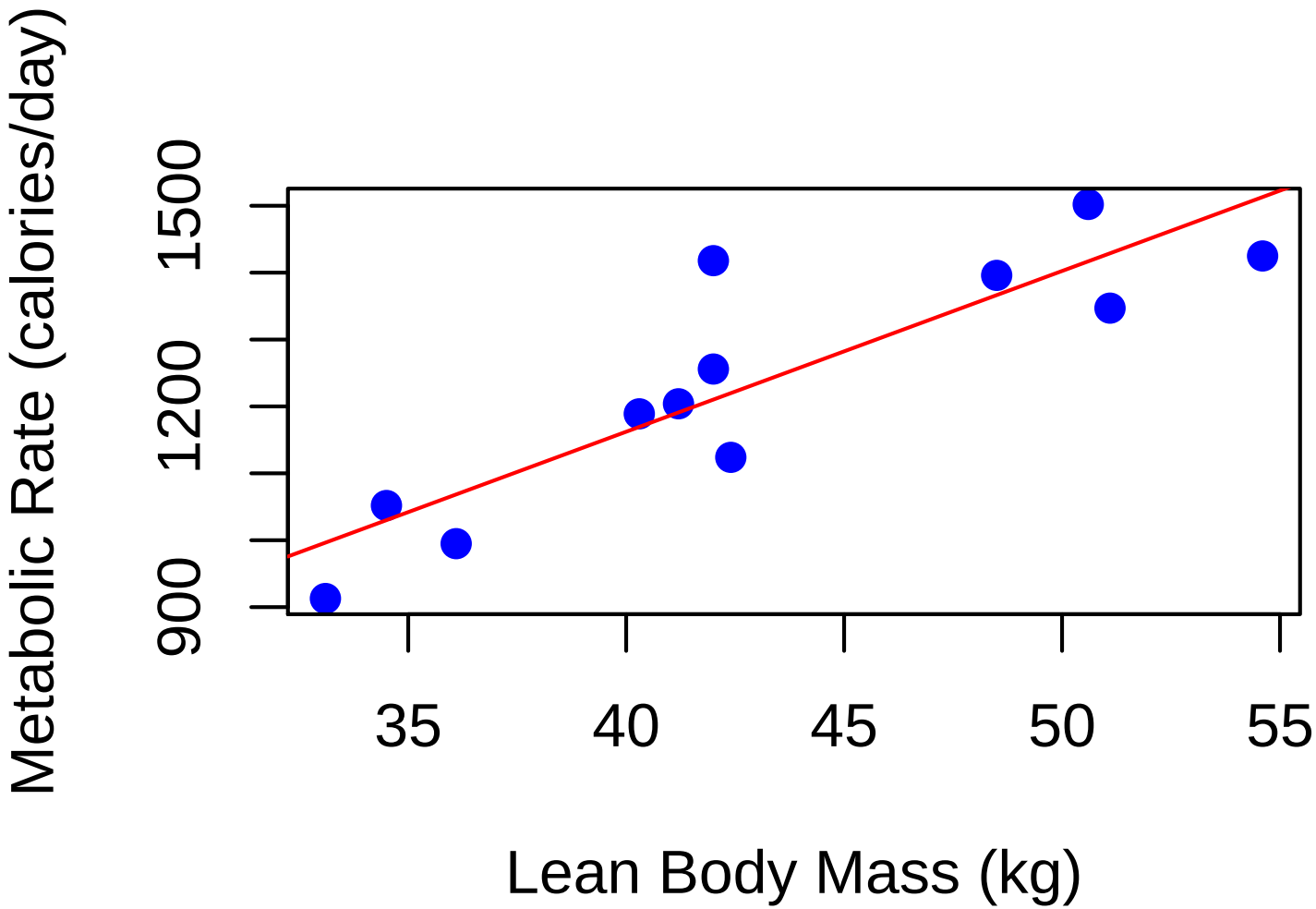
a and b

```
metabolic.reg$coef;

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

Scatterplot with least-squares line

```
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
```

Facts about Least-Squares Regression

1. *The distinction between explanatory and response variables is essential in regression.*
2. *The least-squares regression line always passes through the point* $(\bar{x}, \bar{y})$ *on the graph of* y *against* x.
3. *The square of the correlation,* r^2, *is the fraction of the variation in the values of* y *that is explained by the least-squares regression of* y *on* x.

What's my grade?

In Professor Krugman's economics course the correlation between the student's total scores prior to the final examination and their final-examination scores is $r = 0.5$. The pre-exam totals for all students in the course have mean 280 and standard deviation 40. The final-exam scores have mean 75 and standard deviation 8. Professor Krugman has lost Julie's final exam but knows that her total before the exam was 300. He decides to predict her final-exam score from her pre-exam total.

a) What is the slope of the least-squares regression line of final-exam scores on pre-exam total scores in this course? What is the intercept?

b) Use the regression line to predict Julie's final-exam score.

c) Julie doesn't think this method accurately predicts how well she did on the final exam. Use r^2 to argue that her actual score could have been much higher (or much lower) than the predicted value.

Solutions

a) $b = r\frac{S_y}{S_y} = (0.5)\frac{8}{40} = 0.1$
$a = \bar{y} - b\bar{x} = 75 - (0.1)(280) = 47.$
Hence, the regression equation is $\hat{y} = 47 + 0.1x$.
b) Julie's pre-final exam total was 300, so we would predict a final exam score of

$$\hat{y} = 47 + (0.1)(300) = 77.$$

c) Julie is right ... with a correlation of $r = 0.5$, $r^2 = 0.25$, so the regression line accounts for only 25% of the variability in student final exam scores. That is, the regression line doesn't predict final exam scores very well. Julie's score could, indeed, be much higher or lower than the predicted 77.

Residuals

A *residual* is the difference between an observed value of the response variable and the value predicted by the regression line. That is, a residual is the prediction error that remains after we have chosen the regression line:
residual = observed y - predicted y
residual $= y - \hat{y}$.

Residual Plots

A *residual plot* is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess how well a regression line fits the data.

Residuals by hand

You have already found the equation of the least-squares line for predicting coral growth y from mean sea surface temperature x.
a) Use the equation to obtain the 7 residuals step-by-step. That is, find the prediction $\hat{y}$ for each observation and then find the residual $y - \hat{y}$.
b) Check that (up to roundoff error) the residuals add to 0.
c) The residuals are the part of the response y left over after the straight-line tie between y and x is removed. Show that the correlation between the residuals and x is 0 (up to roundoff error). That this correlation is always 0 is another special property of least-squares regression.

```
coral.coeff=coral.reg$coeff;

coral.coeff;

## (Intercept) explanatory
##  11.6921347  -0.3030522

coral.residuals=coral.reg$residuals;

coral.residuals[1];

##           1
## -0.0675456

coral.residuals[2];

##           2
## -0.05996569
```

Solutions (residuals by hand)

a) The residuals are computed in the table below using $\hat{y} = -0.3030522x + 11.6921347$.

x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
29.68	2.63	2.6975456	-0.0675456
29.87	2.58	2.6399657	-0.0599657
30.16	2.60	2.5520805	0.1279195
30.22	2.48	2.5338974	0.0661026
30.48	2.26	2.4551038	0.0248962
30.65	2.38	2.403585	-0.023585
30.90	2.26	2.3278219	-0.0678219

b) $\sum(y_i - \hat{y}_i) = 5.5511151 \times 10^{-17}$ (they sum to zero, except for rounding error.

c) From software, the correlation between x_i and $y_i - \hat{y}_i$ is -0.0000854, which is zero except for rounding.

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.

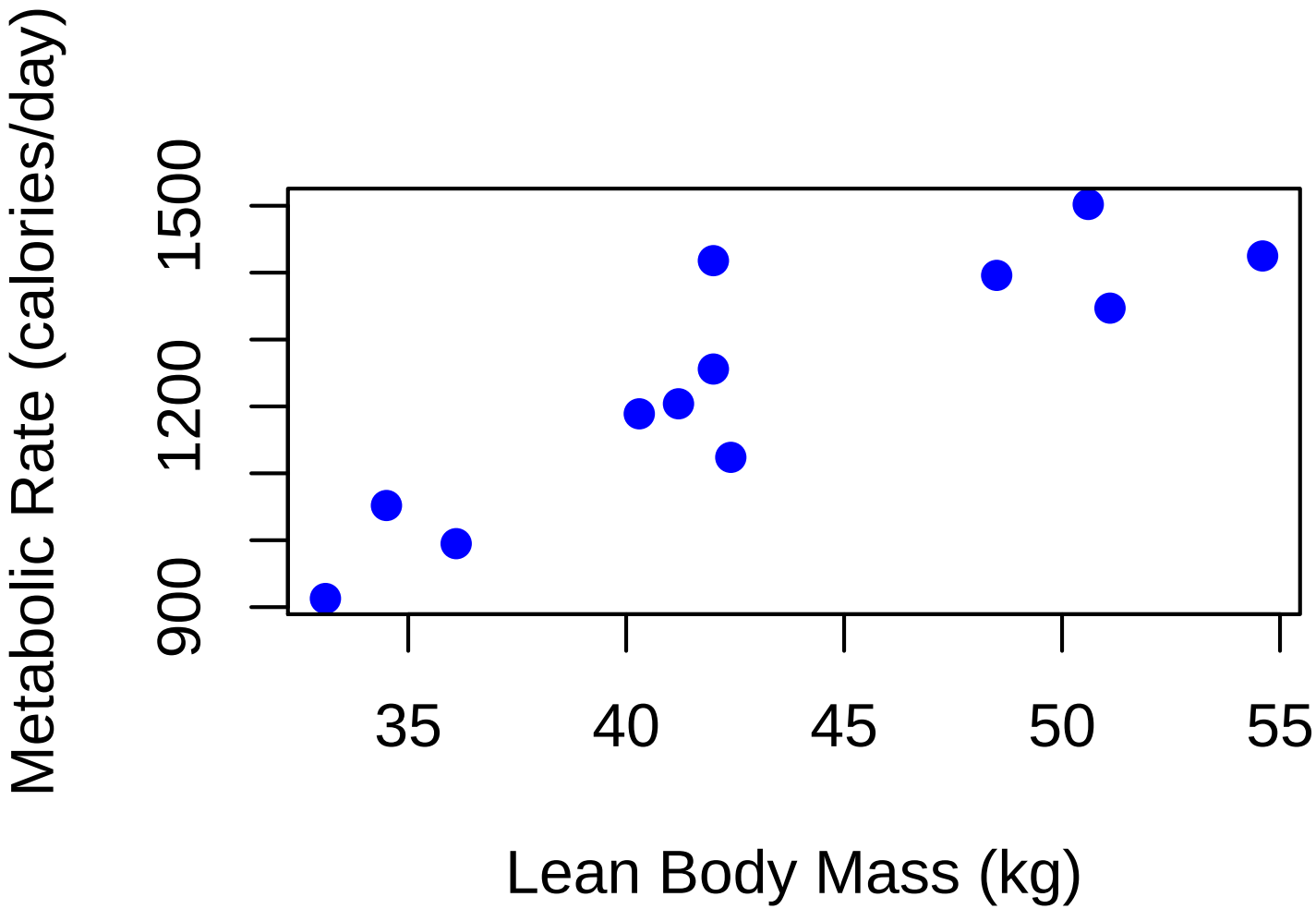
Reading our data

```
# Step 1. Entering data;

# url of metabolic rate data;

meta_url=
"https://mcs.utm.utoronto.ca/~nosedal/data/metabolic.txt"

# importing data into 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



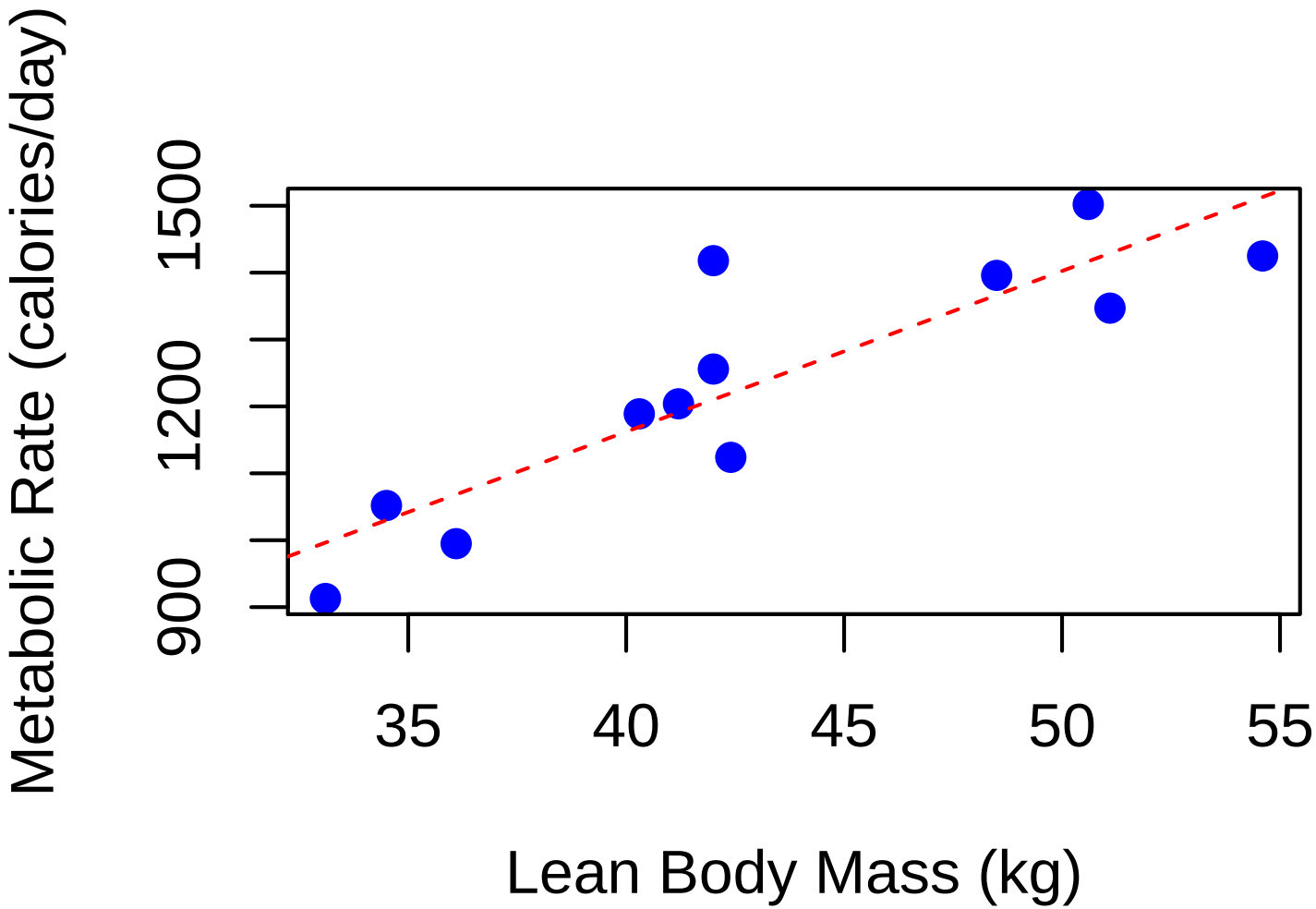
Least-Squares Regression Line

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

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

Scatterplot + L-S Regression Line

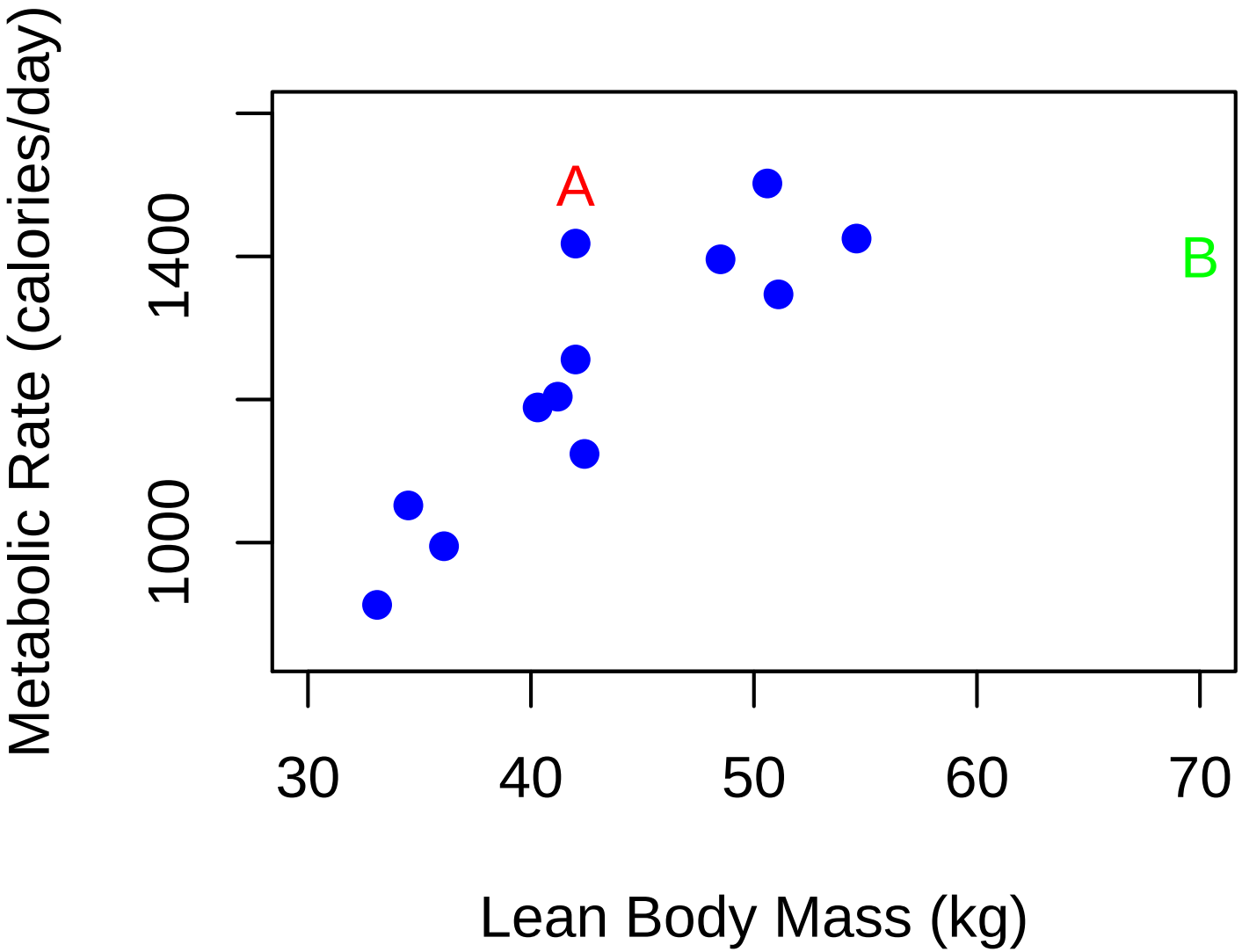
```
plot(data,pch=19,col="blue",
xlab="Lean Body Mass (kg)",
ylab="Metabolic Rate (calories/day)");

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

```
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;
```

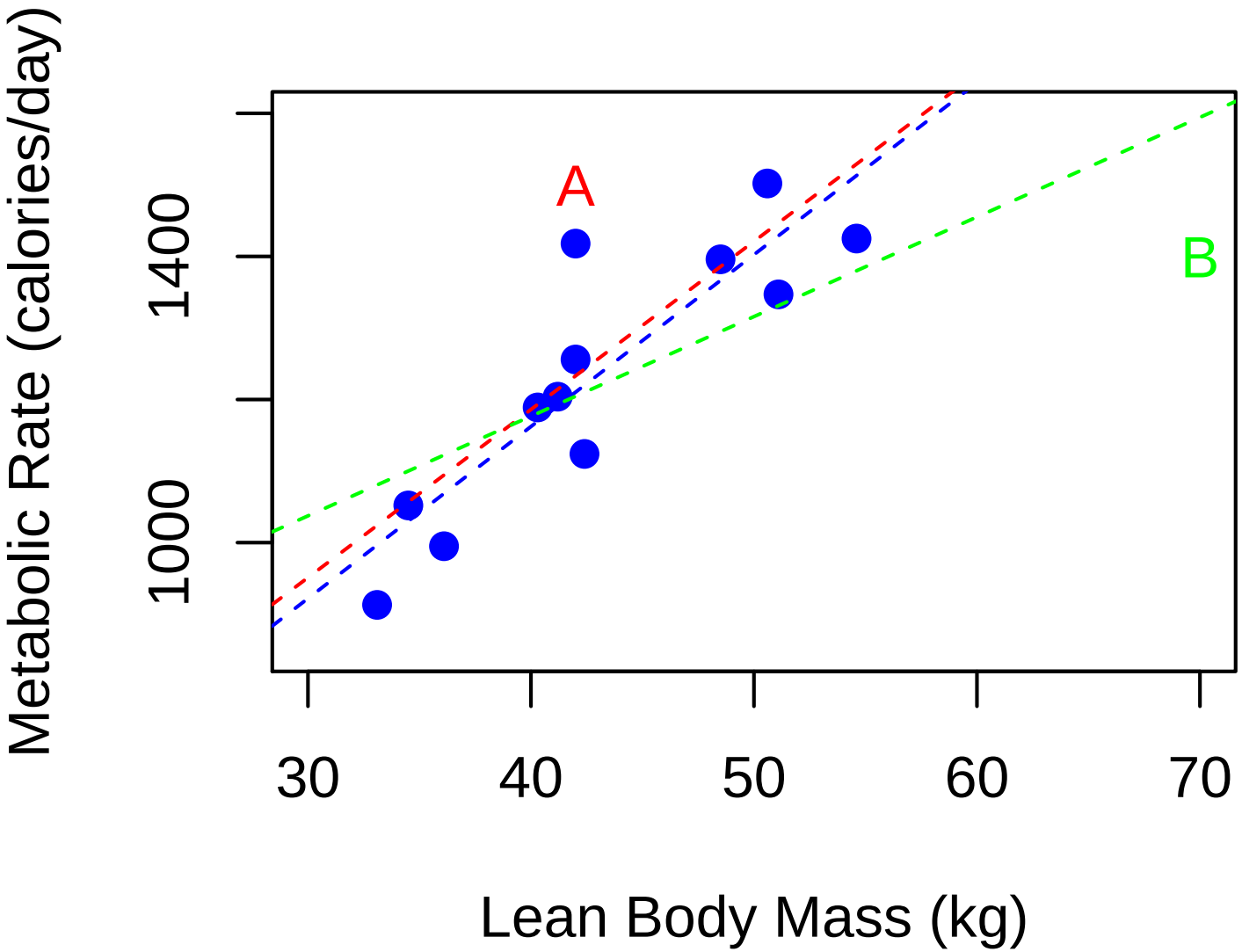

Metabolic Rate (calories/day)
1000
1400
A
B
30
40
50
60
70
Lean Body Mass (kg)

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;
```

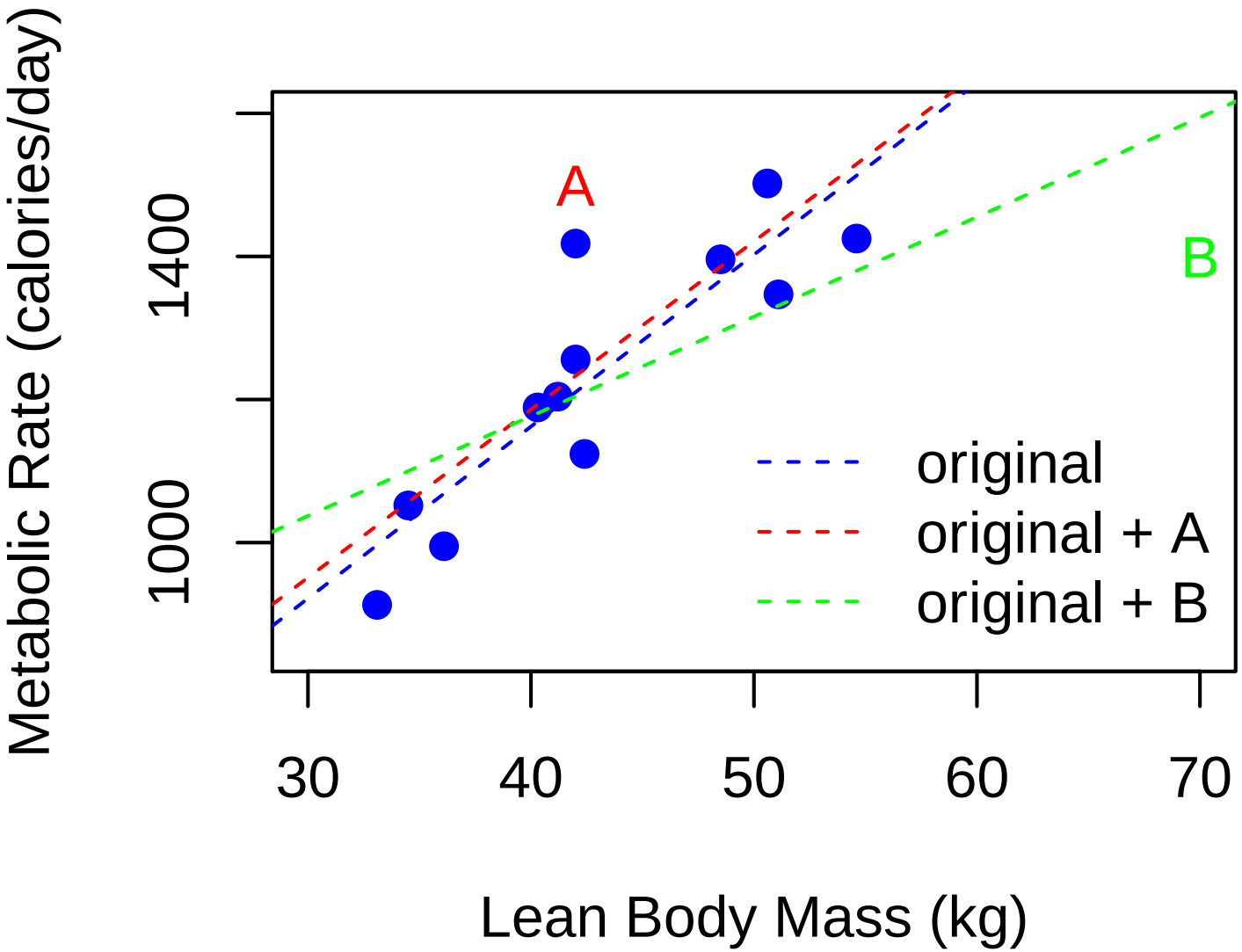
Scatterplot + A +B + L-S Regression Lines

```
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);
```


Metabolic Rate (calories/day)
1000
1400
A
B
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");
```


Metabolic Rate (calories/day)
1000
1400
A
B
original
original + A
original + B
30
40
50
60
70
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.

Example

The number of people living on American farms declined steadily during last century. Here are data on the farm population (millions of persons) from 1935 to 1980:

Year	Population
1935	32.11
1940	30.5
1945	24.4
1950	23.0
1955	19.1
1960	15.6
1965	12.4
1970	9.7
1975	8.9
1980	7.2

Example

a) Make a scatterplot of these data and find the least-squares regression line of farm population on year.
b) According to the regression line, how much did the farm population decline each year on the average during this period? What percent of the observed variation in farm population is accounted for by linear change over time?
c) Use the regression equation (trendline) to predict the number of people living on farms in 1990. Is this result reasonable? Why?

R Code

```
# Step 1. Entering Data;

year=seq(1935,1980,by=5);

population=c(32.11,30.5,24.4,23.0,19.1,
15.6,12.4,9.7,8.9,7.2);

# seq creates a sequence of numbers;
# which starts at 1935 and ends at 1980;
# we want a distance of 5 between numbers;
```

R Code, L-S Line

```
least.squares=lm(population~year);

least.squares

##
## Call:
## lm(formula = population ~ year)
##
## Coefficients:
## (Intercept)         year
##   1167.1418      -0.5869
```

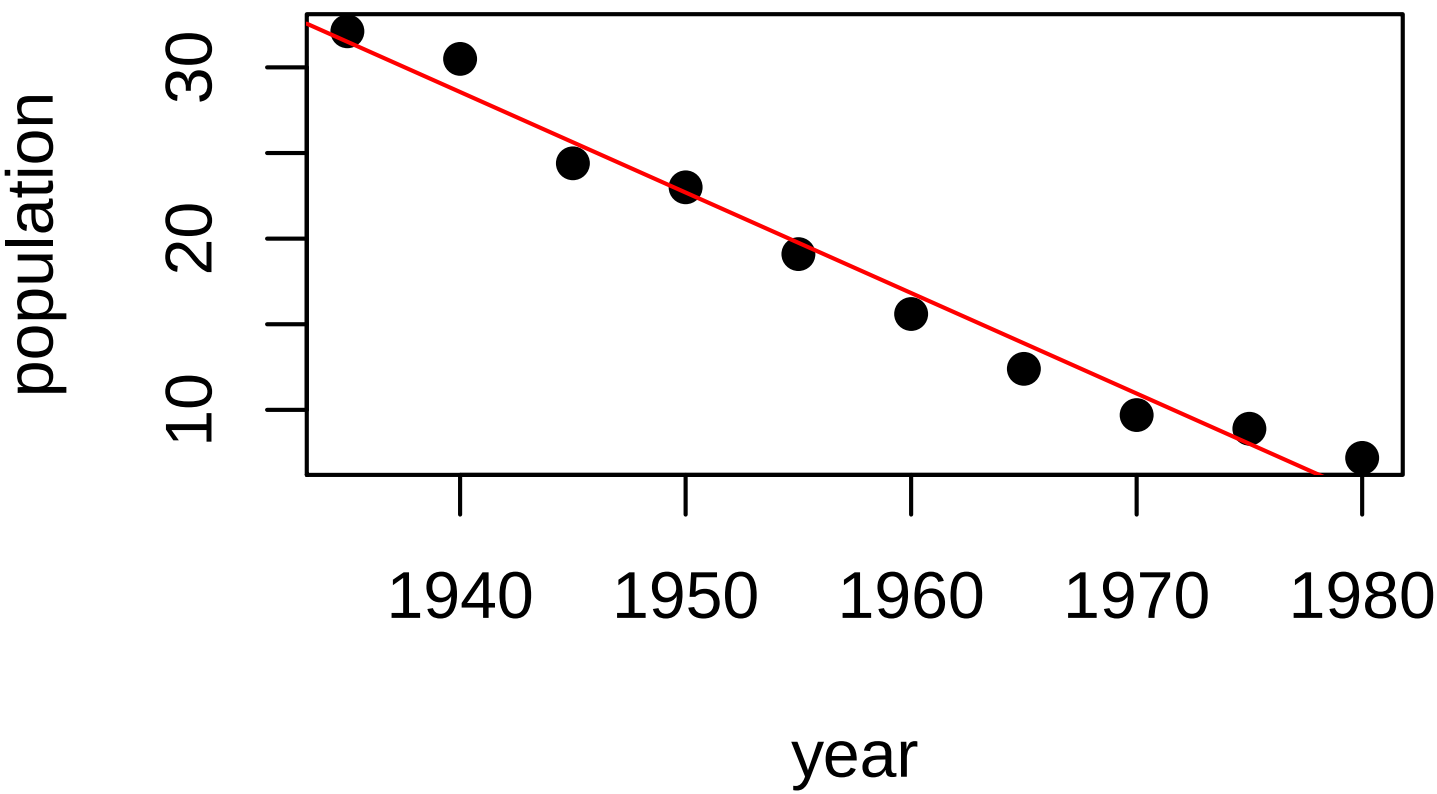
R Code, L-S Line

```
cor(year,population);

## [1] -0.9884489
```

Scatterplot

```
plot(year,population,pch=19);

abline(least.squares$coeff,col="red");

# pch=19 tells R to draw solid circles;

# abline tells R to add trendline;
```

Scatterplot



Solution

a) The scatterplot shows a strong negative association with a straight-line pattern. The regression line (trendline) is $\hat{y} = 1167.14 - 0.587x$.
b) This is the slope - about 0.587 million $(587,000)$ per year during this period. Because $r \approx -0.9884$, the regression line explains $r^2 \approx 97.7\%$ of the variation in population.
c) Substituting, $x = 1990$ gives $\hat{y} = 1167.14 - 0.587(1990) = -0.99$, an impossible result because a population must be greater than or equal to 0. The rate of decrease in the farm population dropped in the 1980s. Beware of extrapolation.

The endangered manatee

The table shown below gives 33 years of data on boats registered in Florida and manatees killed by boats. If we made a scatterplot for this data set, it would show a strong positive linear relationship. The correlation is $r = 0.951$.

a) Find the equation of the least-squares line for predicting manatees killed from thousands of boats registered. Because the linear pattern is so strong, we expect predictions from this line to be quite accurate - but only if conditions in Florida remain similar to those of the past 33 years.

b) In 2009, experts predicted that the number of boats registered in Florida would be 975,000 in 2010. How many manatees do you predict would be killed by boats if there are 975,000 boats registered? Explain why we can trust this prediction.

c) Predict manatee deaths if there were no boats registered in Florida. Explain why the predicted count of deaths is impossible.

Table

Year	Boats	Manatees	Year	Boats	Manatees
1977	447	13	1988	675	43
1978	460	21	1989	711	50
1979	481	24	1990	719	47
1980	498	16	1991	681	53
1981	513	24	1992	679	38
1982	512	20	1993	678	35
1983	526	15	1994	696	49
1984	559	34	1995	713	42
1985	585	33	1996	732	60
1986	614	33	1997	755	54
1987	645	39	1998	809	66

Table (cont.)

Year	Boats	Manatees
1999	830	82
2000	880	78
2001	944	81
2002	962	95
2003	978	73
2004	983	69
2005	1010	79
2006	1024	92
2007	1027	73
2008	1010	90
2009	982	97

Solutions

a) The regression line is $\hat{y} = -43.172 + 0.129x$.
b) If $975,000$ boats are registered, then by our scale, $x = 975$, and $\hat{y} = -43.172 + (0.129)(975) = 82.6$ manatees killed. The prediction seems reasonable, as long as conditions remain the same, because "975" is within the space of observed values of x on which the regression line was based. That is, this is not extrapolation.
c) If $x = 0$ (corresponding to no registered boats), then we would "predict" -43.172 manatees to be killed by boats. This is absurd, because it is clearly impossible for fewer than 0 manatees to be killed. Note that $x = 0$ is well outside the range of observed values of x on which the regression line was based.

Extrapolation

Extrapolation is the use of a regression line for prediction far outside the range of values of the explanatory variable x that you used to obtain the line. Such predictions are often not accurate.

Association does not imply causation

An association between an explanatory variable x and a response variable y, even if it is very strong, is not by itself good evidence that changes in x actually cause changes in y.

Example

Measure the number of television sets per person x and the average life expectancy y for the world's nations. There is a high positive correlation: nations with many TV sets have higher life expectancies.
The basic meaning of causation is that by changing x we can bring about a change in y. Could we lengthen the lives of people in Rwanda by shipping them TV sets? No. Rich nations have more TV sets than poor nations. Rich nations also have longer life expectancies because they offer better nutrition, clean water, and better health care. There is no cause-and-effect tie between TV sets and length of life.

Is math the key to success in college?

A College Board study of 15,941 high school graduates found a strong correlation between how much math minority students took in high school and their later success in college. New articles quoted the head of the College Board as saying that "Math is the gatekeeper for success in college." Maybe so, but we should also think about lurking variables. What might lead minority students to take more or fewer high school math courses? Would these same factors influence success in college?

Solution

A student's intelligence may be a lurking variable: stronger students (who are more likely to succeed when they get to college) are more likely to choose to take these math courses, while weaker students may avoid them. Other possible answers may be variations on this idea; for example, if we believe that success in college depends on a student's self-confidence, and perhaps confident students are more likely to choose math courses.

Lurking Variable

A *lurking variable* is a variable that is not among the explanatory or response variables in a study and yet may influence the interpretation of relationships among those variables.

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

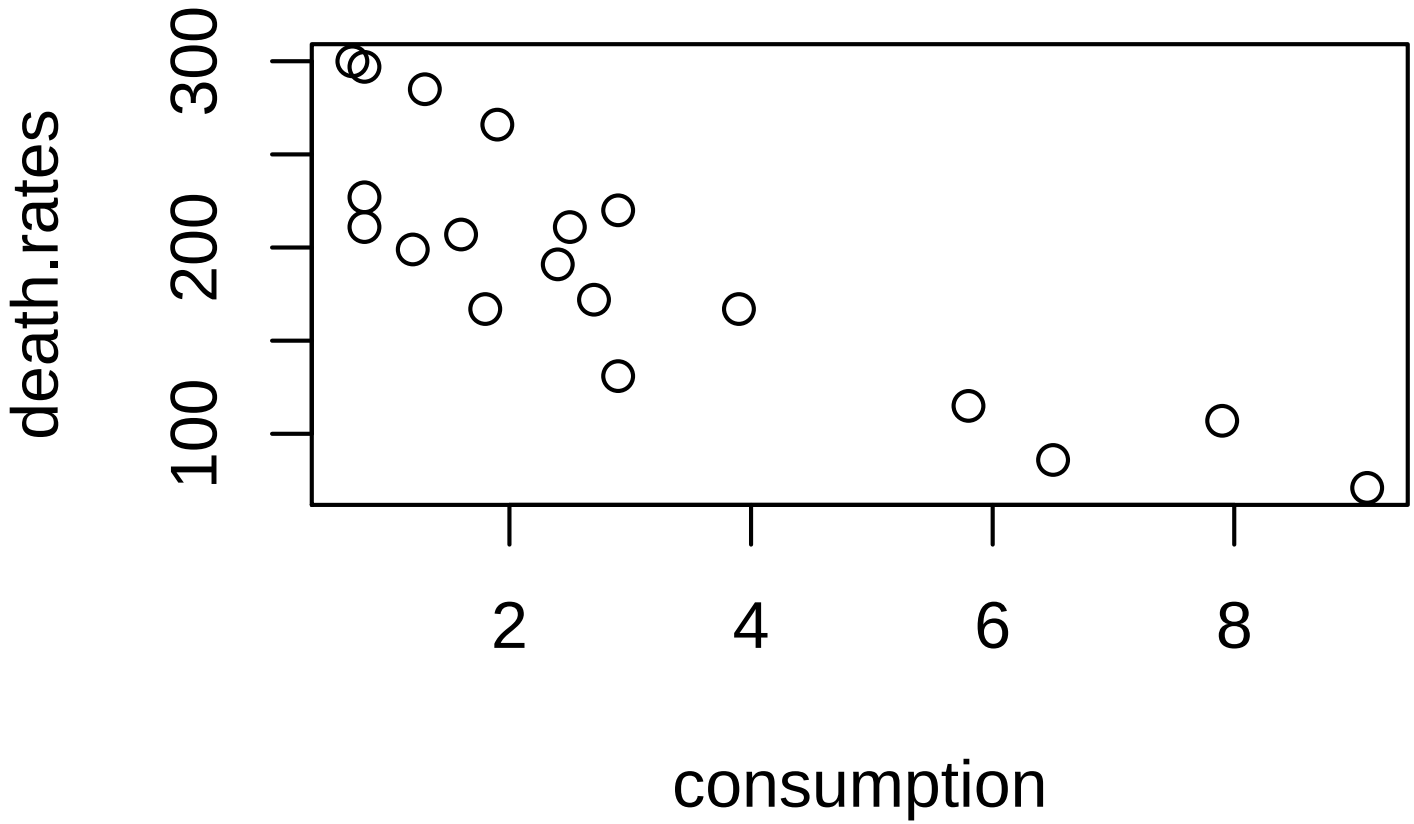
Solution (Bar chart)

```
# 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)

```
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.dat contains the total discharge from these rivers each year from 1936 to 1999. Discharge is measured in cubic kilometers of water.

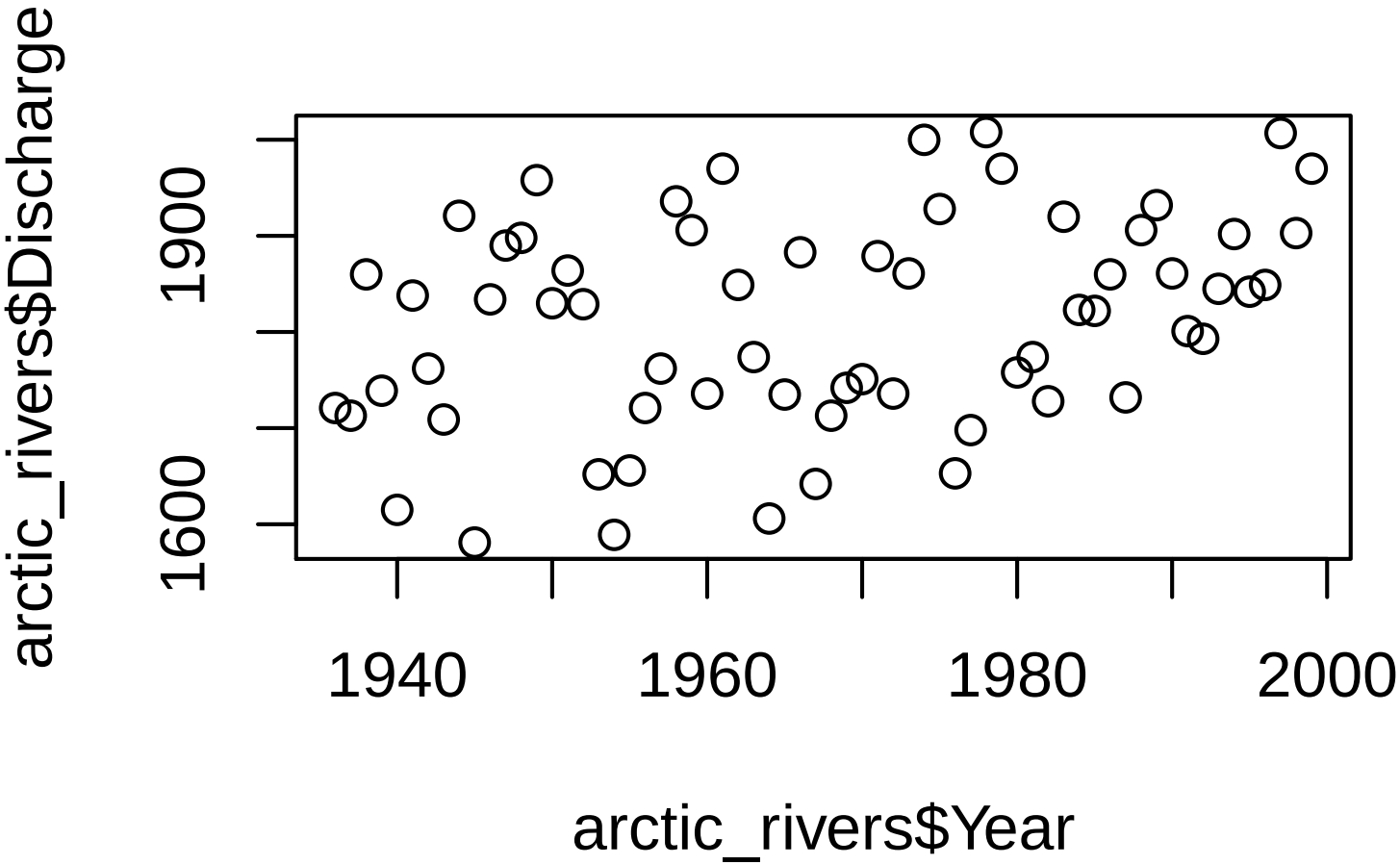
Reading our data

```
# url of arctic rivers data;

riv_url=
"https://mcs.utm.utoronto.ca/~nosedal/data/arctic-rivers.txt"


# importing data into R;

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

Scatterplot (R code)

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

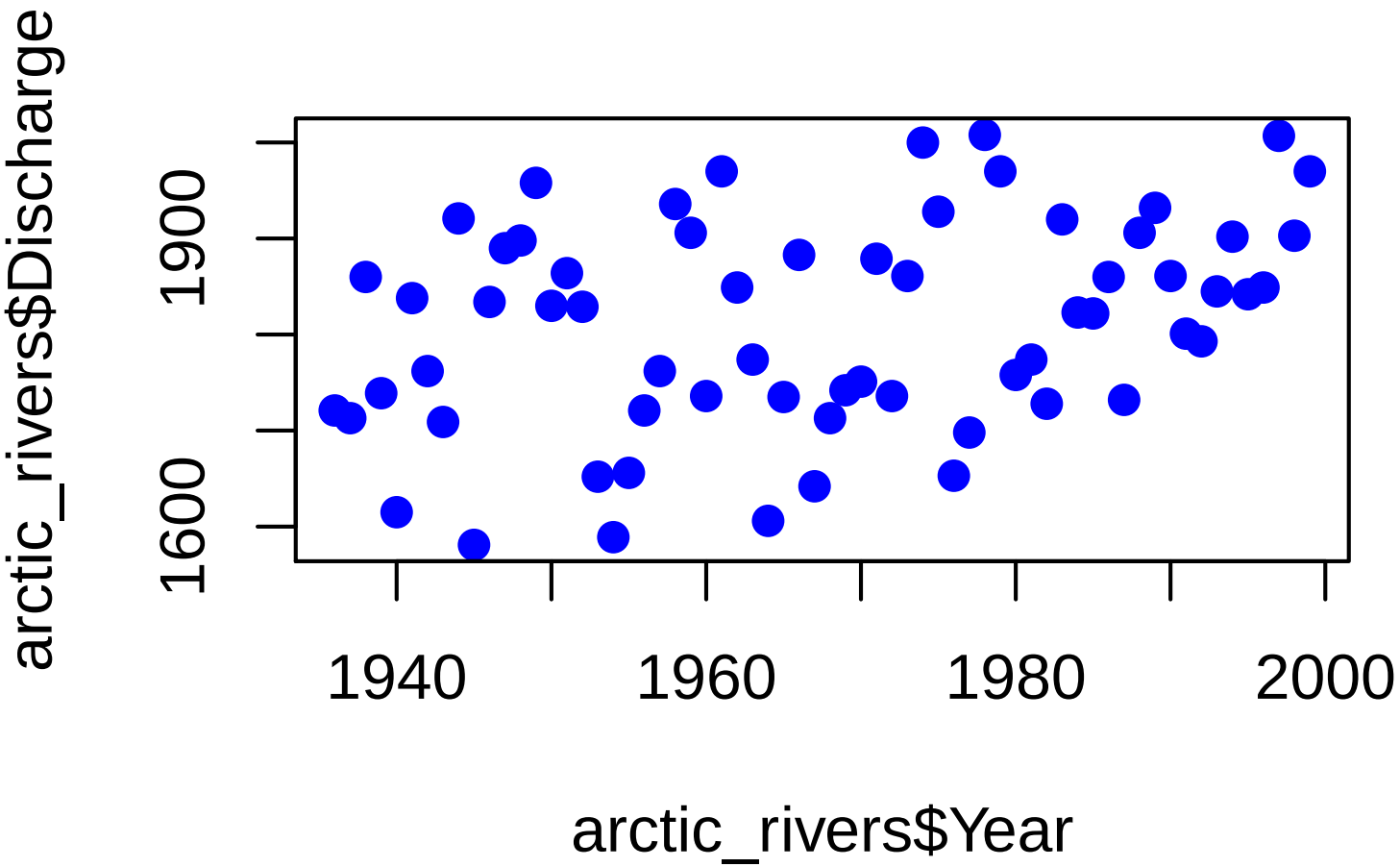
Scatterplot (R code)



Scatterplot (R code)

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

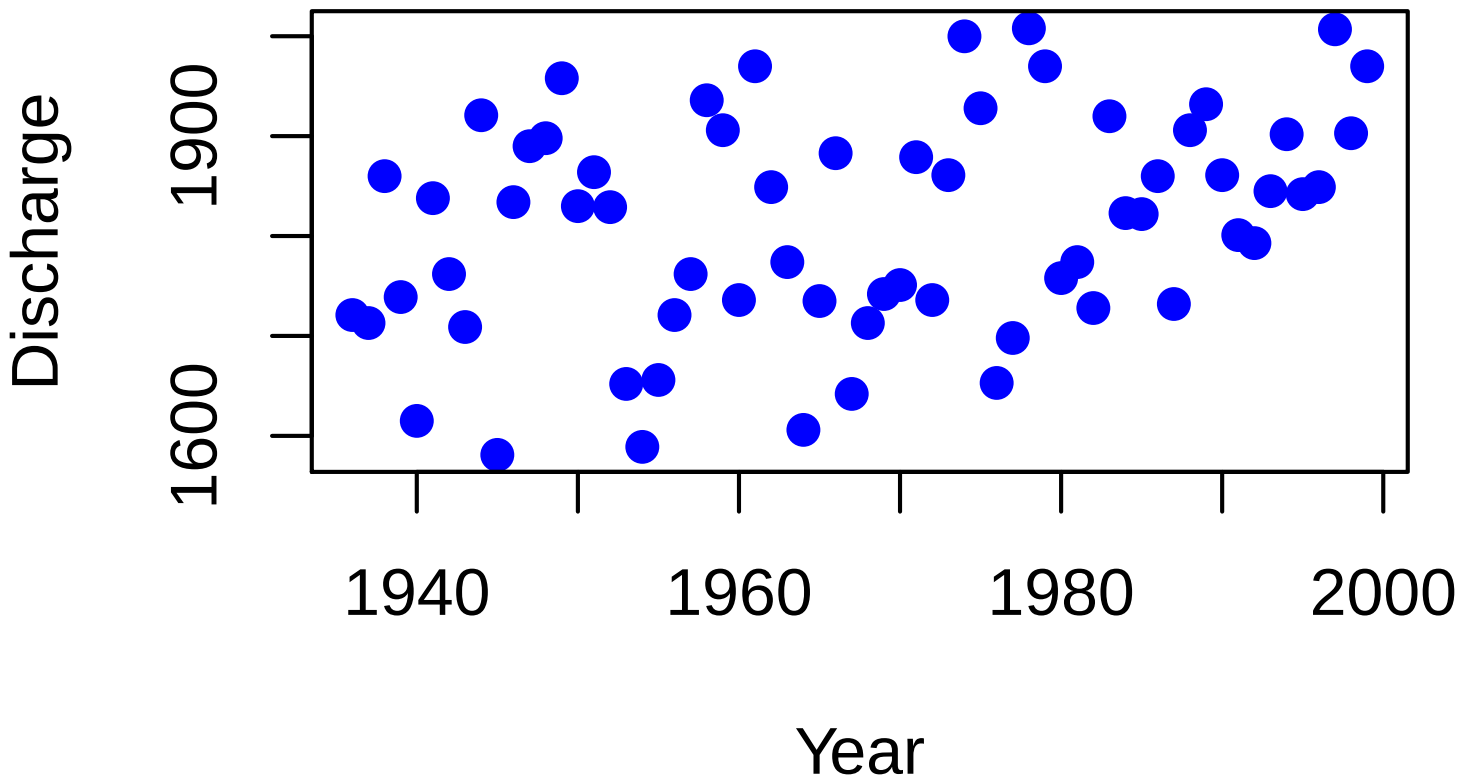
Scatterplot (R code)



Scatterplot (R code)

```
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
```

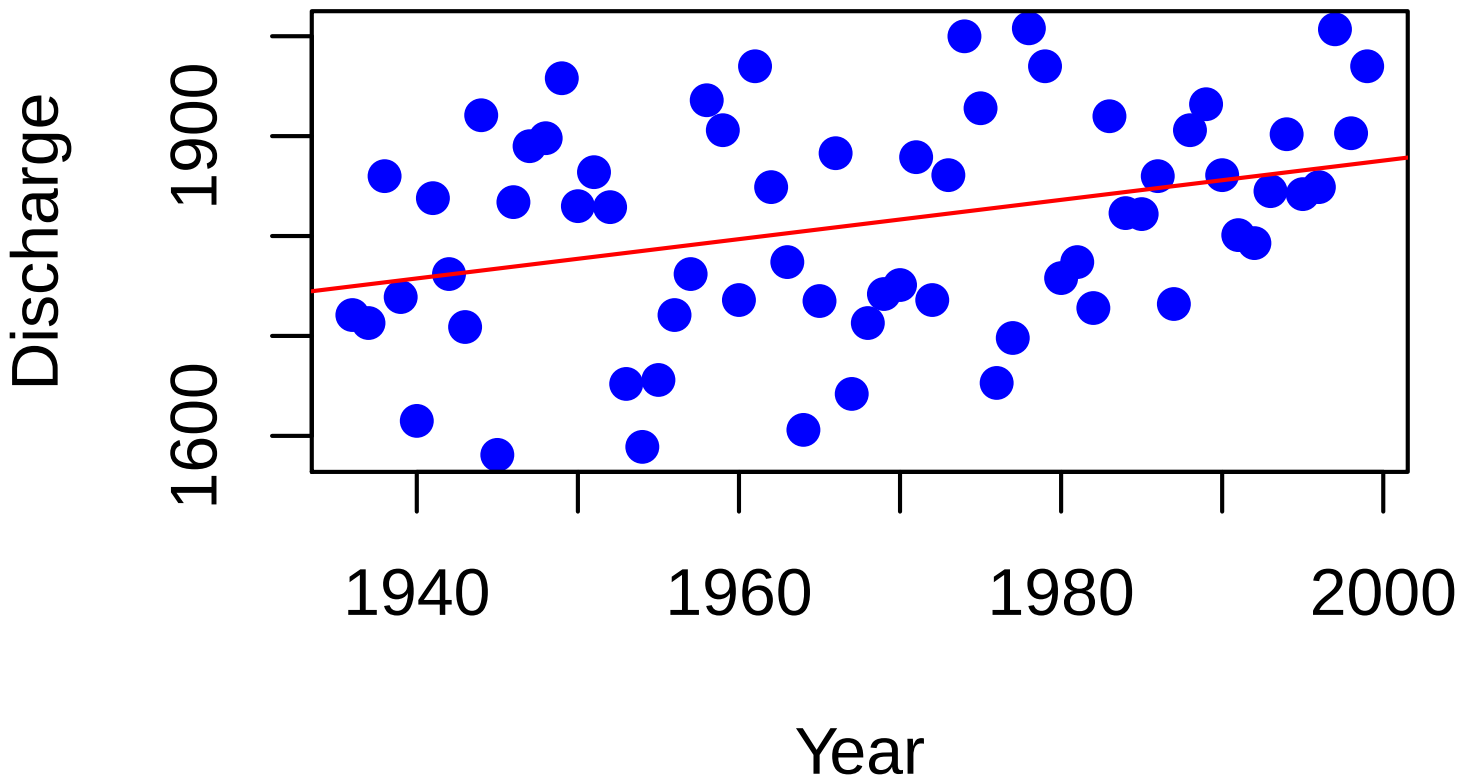
a and b

```
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

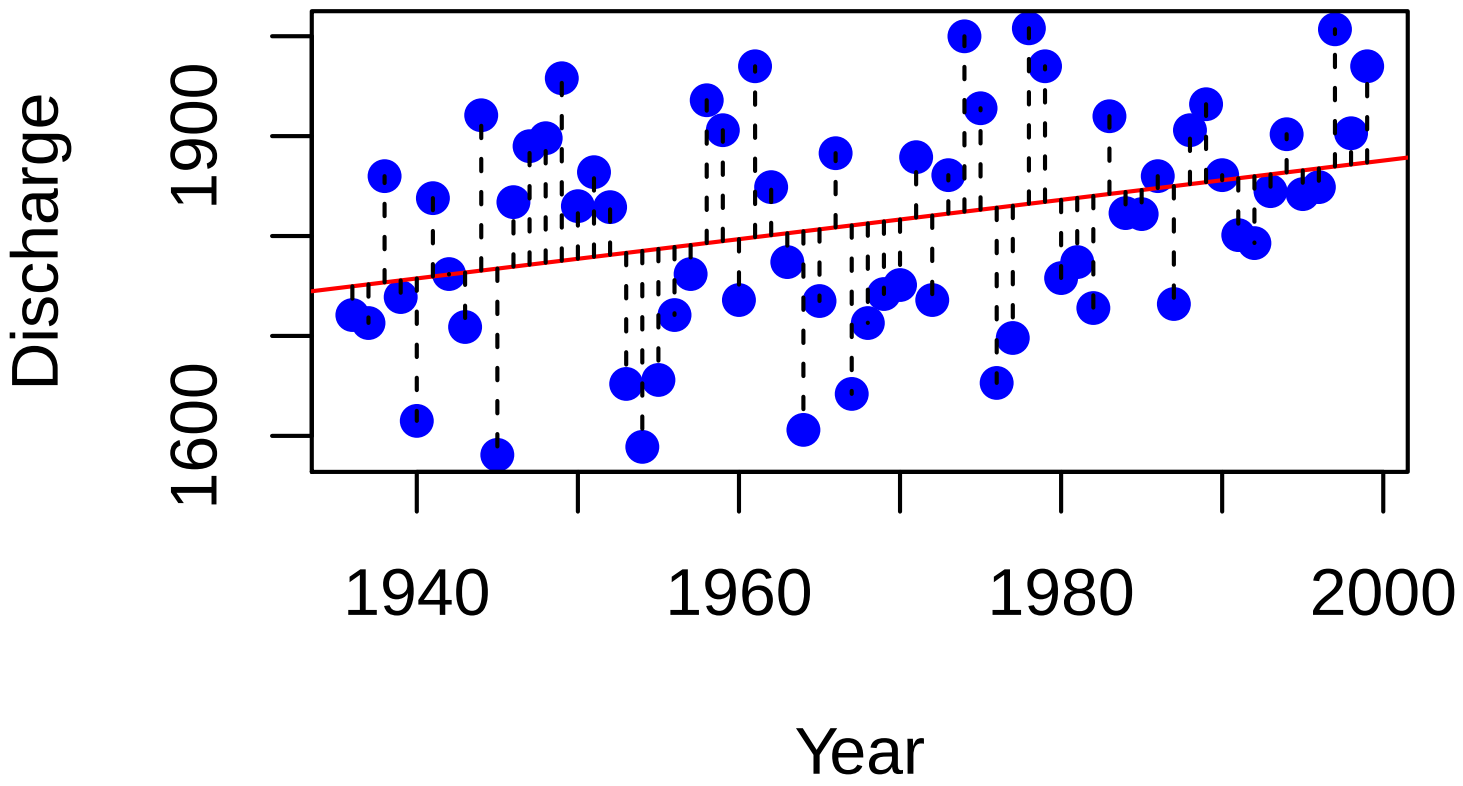
A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is,

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

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



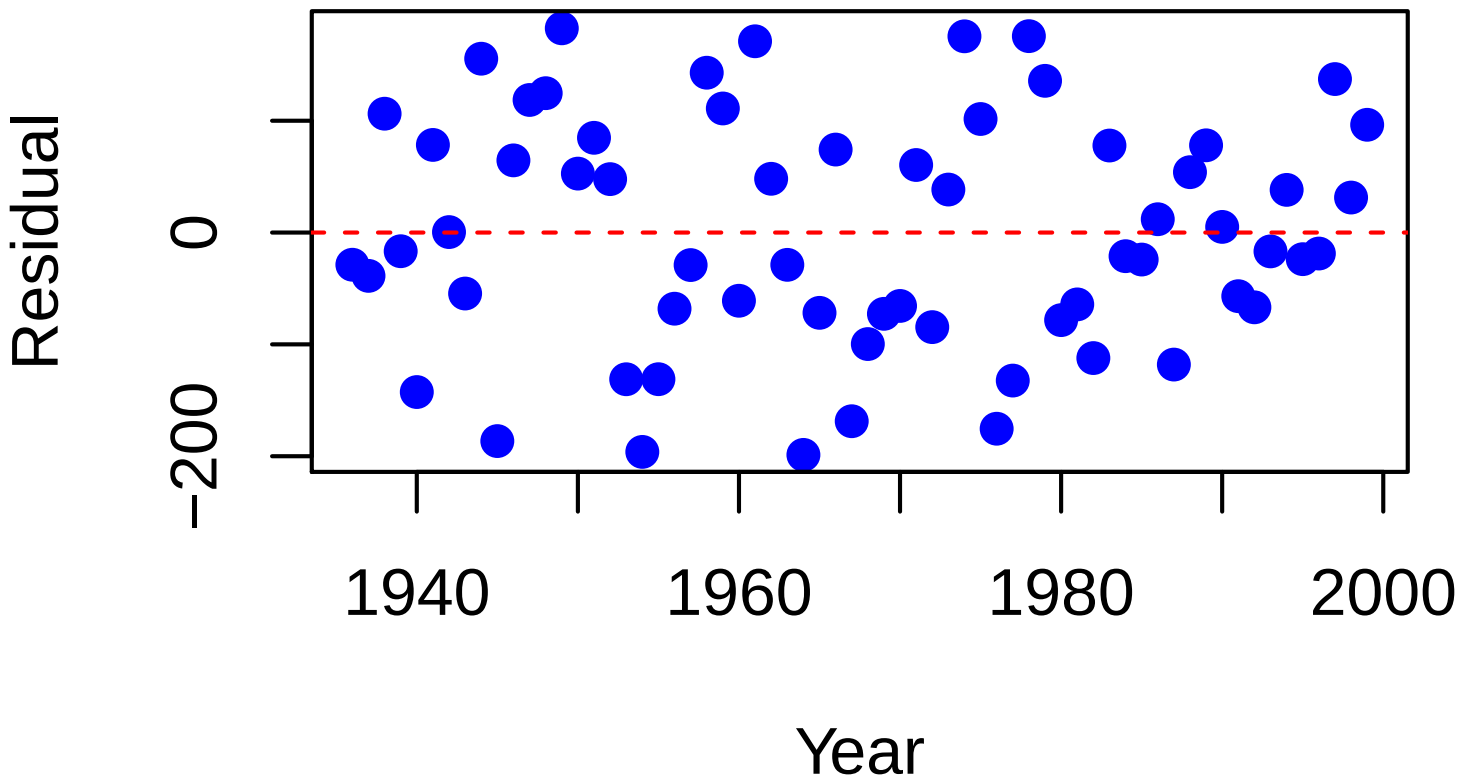
Residual Plots

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



Example: Counting carnivores

Ecologist look at data to learn about nature's patterns. One pattern they have found relates the size of a carnivore (body mass in kilograms) to how many of those carnivores there are in an area. The right measure of "how many" is to count carnivores per 10,000 kilograms of their prey in the area. Below we show a table that gives data for 25 carnivore species. To see the pattern, plot carnivore abundance against body mass. Biologist often find that patterns involving sizes and counts are simpler when we plot the **logarithms** of the data.

Table: Size and abundance of carnivores

Carnivore species	Body mass (kg)	Abundance
Least weasel	0.14	1656.49
Ermine	0.16	406.66
Small Indian mongoose	0.55	514.84
Pine marten	1.3	31.84
Kit fox	2.02	15.96
Channel Island fox	2.16	145.94
Arctic fox	3.19	21.63
Red fox	4.6	32.21
Bobcat	10	9.75
Canadian lynx	11.2	4.79
European badger	13	7.35
Coyote	13	11.65
Ethiopian wolf	14.5	2.7

Table: Size and abundance of carnivores

Carnivore species	Body mass (kg)	Abundance
Eurasian lynx	20	0.46
Wild dog	25	1.61
Dhole	25	0.81
Snow leopard	40	1.89
Wolf	46	0.62
Leopard	46.5	6.17
Cheetah	50	2.29
Puma	51.9	0.94
Bobcat	10	9.75
Spotted hyena	58.6	0.68
Lion	142	3.4
Tiger	181	0.33
Polar bear	310	0.6

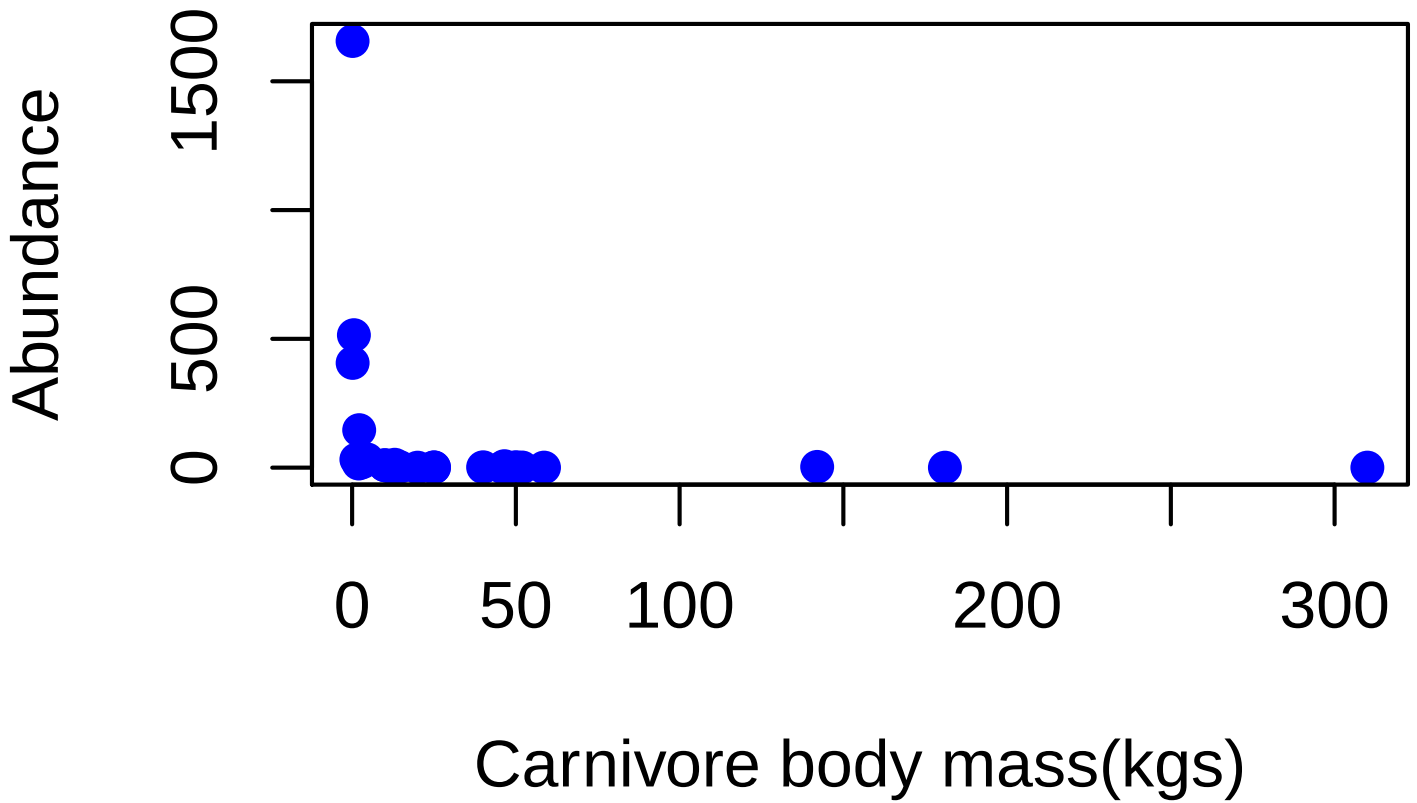
Reading our data

```
# Step 1. Entering data;

# url of carnivores;

carnivores_url=
"https://mcs.utm.utoronto.ca/~nosedal/data/carnivores.txt"


# importing data into R;

carnivores = read.table(carnivores_url, header = TRUE);
```

Abundance vs Body Mass

```
# Step 2. Making  scatterplot;

plot(carnivores$B.mass,carnivores$Abundance,pch=19,
col="blue",xlab="Carnivore body mass(kgs)",ylab="Abundance",
main=" ");


# main adds title to graph;
```

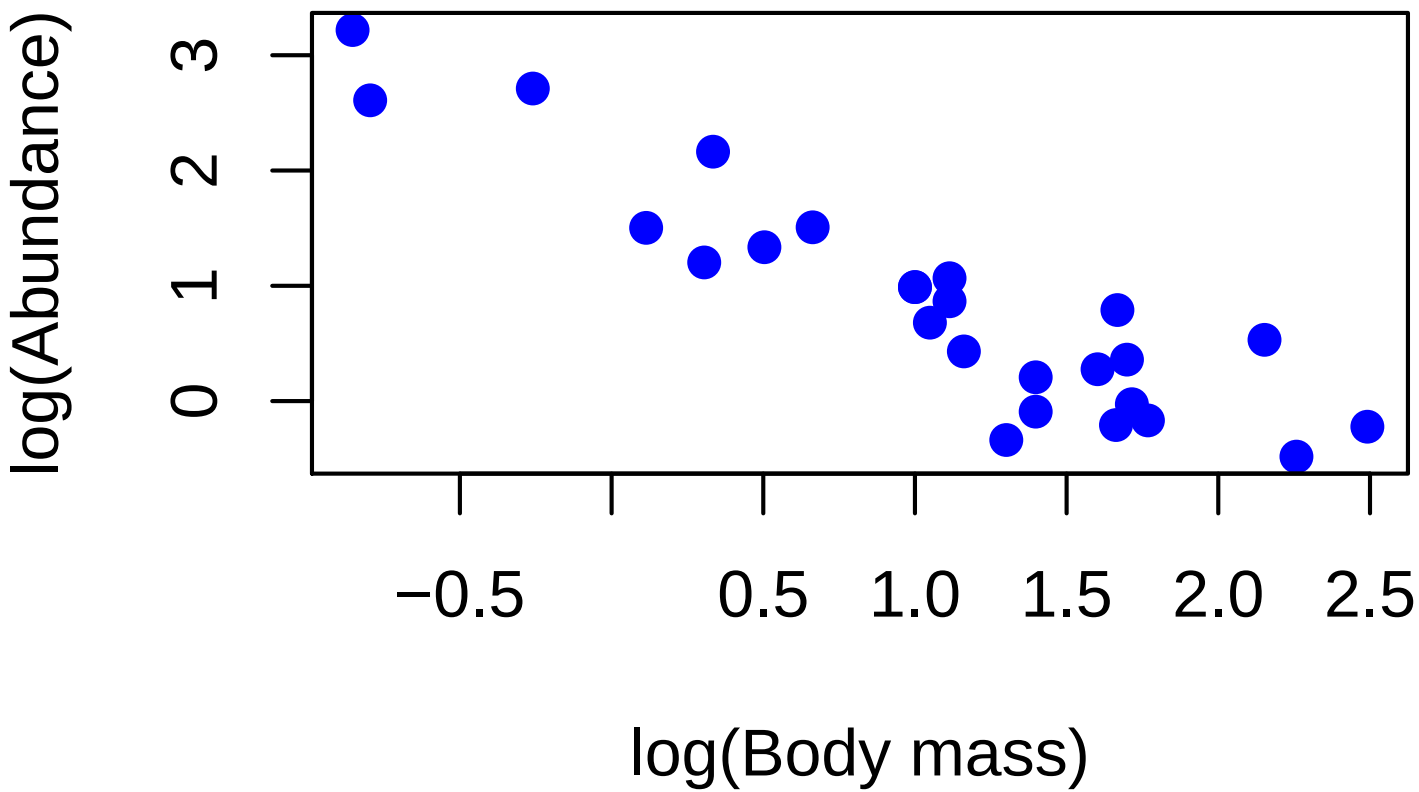
Abundance vs Body Mass



log(Abundance) vs log(Body Mass)

```
# Step 2. Making time plot;

plot(log10(carnivores$B.mass),log10(carnivores$Abundance),
pch=19, col="blue",xlab="log(Body mass)",
ylab="log(Abundance)", main=" ");


# main adds title to graph;
```

log(Abundance) vs log(Body Mass)



This scatterplot shows a **moderately strong negative association**. Bigger carnivores are less abundant. The form of the association is **linear**. It is striking that animals from many different parts of the world should fit so simple a pattern. We could use the straight-line pattern to predict the abundance of another carnivore species from its body mass (Homework?).