Section 5.3
The Least Squares Regression Line

Ind. var (x)

A **regression line** is a line that describes the relationship between the explanatory variable and the response variable.

Dep. var (y)

The most common mathematical method for fitting the best fit line is the **least squares method**. The **least squares regression line (LSRL)** is a regression line that makes the vertical distances of the points in a scatter plot from the line as small as possible.

The least squares line is of the form: $\hat{y} = a + bx$, where the slope $b = r\frac{s_y}{s_x}$ and the y-intercept $a = \bar{y} - b\bar{x}$

sd of y
sd of x
correlation coeff
mean y
mean x

Example 1: Use the following statistics to find the equation of the LSRL.

$$\bar{x} = 3, \quad s_x = 3, \quad \bar{y} = 4, \quad s_y = 5.29, \quad r = 0.189$$

Recall: $b = r\frac{s_y}{s_x}$ $\qquad$ $a = \bar{y} - b\bar{x}$

$b = 0.189\left(\frac{5.29}{3}\right)$ $\qquad$ $a = 4 - .3333(3)$

$a = 3.0002$

$b = .3333$

LSRL: $\hat{y} = 3.0002 + .3333x$

Example2: Find the correlation coefficient given: $b = 0.123, \quad s_x = 5.01, \quad s_y = 1.02$

r ?

Recall: $b = r\frac{s_y}{s_x}$

$\left(\frac{5.01}{1.02}\right) 0.123 = r \left(\frac{1.02}{5.01}\right)\left(\frac{5.01}{1.02}\right)$

$r = 0.6041$

We can create the LSRL using R. The command is: lm(*y*~*x*)

Example 3: Recall the following example from Sections 5.1 and 5.2:
The bivariate data given below relate the high temperature reached on a given day and the number of water bottles sold from a particular vending machine. Find the LSRL.

Temperature (in degrees)	Bottled Water (16 oz)
90	30
91	32
88	29
93	33
92	31
89	29
90	30
91	31
92	32
94	34

In R, we created two lists:
temp=c(90,91,88,93,92,89,91,92,94)
bottles=c(30,32,29,33,31,29,30,31,32,34)

Command: lm(bottles ~ temp)

Result:

Coefficients:	
(Intercept)	*temp*
-47.7667	0.8667

To build up $\hat{y} = a + bx$, the value under "(Intercept)" is *a* and the value under the name of the *x* variable is the slope *b*.

So, LSRL is: $\hat{y} = -47.7667 + 0.8667x$

Let's interpret the meaning of the value of the slope of the LSRL.

On average, for each degree increase in temp. for a given day there is an increase of .8667 bottles of H_2O sold

The LSRL can also be used to predict future values. Given an *x*, you can predict a *y*. However, if given an *x* far larger or smaller than the other *x* values, the predication for *y* will not be a very good one. This is called **extrapolation**.

Let's predict the number of bottled water sold on a day with a temperature of 87 degrees.

$\hat{y}(87) = -47.7667 + 0.8667(87)$
$= 27.6362$ 28 bottles

The correlation r, when squared, is called the **coefficient of determination**, r^2 where $0 \le r^2 \le 1$. It's the fraction of variation in the y values that is explained by the regression line and the explanatory variable. It allows us to determine how certain one can be in making predications with the line of best fit.

When r^2 is close to 1, it implies that the model may be very useful. When r^2 is close to 0, it implies that the model may not be very useful.

cor (temp, bottles)

For the problem in Example 3, we got that $r = 0.9513$, squaring this we get $r^2 = 0.9050$. This means that 90.50% of the variation in the number of bottled water sold is explained by the least squares regression line.

close to 1

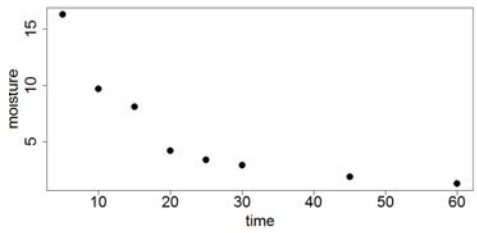
Example 4: The frying time of food x, in minutes, and the amount of moisture retained when done frying y are recorded as follows:

Frying time, x	5	10	15	20	25	30	45	60
Moisture, y	16.3	9.7	8.1	4.2	3.4	2.9	1.9	1.3

a. Make a scatter plot of the data.
Commands:

```
time=c(5,10,15,20,25,30,45,60)
moisture=c(16.3,9.7,8.1,4.2,3.4,2.9,1.9,1.3)
plot(time,moisture,pch=16,cex=2,cex.lab=2,cex.axis=2)
```

plot (time, moisture)



b. Compute the LSRL.
Command:

lm (moisture ~ time) $\hat{y} = 11.8599 - 0.2242x$

c. Provide an interpretation of the slope of this line.

On average, for each increase of 1 minute, moisture decreases by .02242 of a unit

r

d. Find the correlation coefficient for the relationship. Interpret this number.
Command: Answer:

cor(time, moisture) -0.8104 => strong negative linear association

r^2

e. Find the coefficient of determination for the relationship. Interpret this number.
Command: Answer:

cor(time, moisture)^2 0.6568

65.68% of the variation in the amount of moisture in food is explained by LSRL

In Mosaic:

lm(y ~ x, data = "package name")

cor(x, y, data = "package name")