XV. Linear Regression

Scenario. This section is a continuation of the previous two sections on correlation. The scenario is the same: you will have taken a pair of observations on each member of the sample. Each observation generally measures a different but related response.

Data. For linear regression you will need to either compute or be given the following data:

	1st observation	2nd observation
means	$\bar{x}$	$\bar{y}$
standard deviations	s_x	s_y

In addition you will need to know the correlation coefficient r between the two variables.

Research Objective. To find the equation of the regression line

$$y = mx + b$$

which best fits the data. In addition, you will usually also want to use this equation to predict a value for y from a given value for x. If you are trying to do a prediction, you will be given only the x observation on a specific individual and will want to predict the value of y.

Solution Template

Step 1. Make a dictionary which assigns values to the variables. Before making your table, check to see if you are trying to do a prediction. If so, the quantity you are trying to predict must be y and the quantity used to do the prediction must be x.

average for x	$\bar{x}$
st. dev. for x	s_x
average for y	$\bar{y}$
st. dev. for y	s_y
correlation coeef.	r
indiv. x observation	x_0

The individual x_0 observation is the observation for a specific individual which will be used to predict y.

Step 2. Find the parameter m:

$$m = \frac{r s_y}{s_x}$$

Step 3. Find the parameter b:

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

Step 4. Write down the regression equation:

$$y = mx + b$$

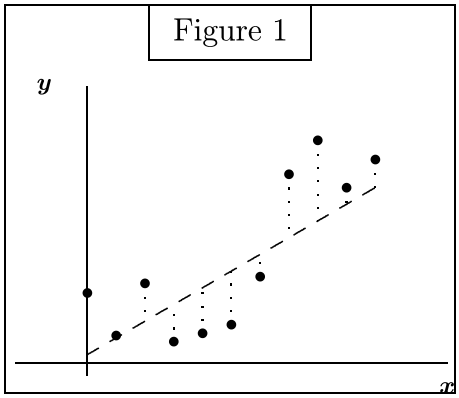
Step 5. Perform the prediction by substituting x_0 into the regression equation:

$$y = mx_0 + b$$

End of Solution Template

Interpretation. The regression line provides the line which "best fits" the data.

For example, given a scatter plot and regression line which look like:



The dashed line represents the regression line. The dotted vertical lines represent the error between the predicted value of y (on the line) and the actually observed value of y (the dot). Some dots fall above the line (positive error) and some fall below the line (negative error). If we just add up all of the error terms, the result will be zero (just as it was for standard deviations). To measure the absolute magnitude of the error, we would need to square the errors. The regression line is the straight line which minimizes the sum of the squares of the errors. For this reason, the regression line is sometimes called the *least squares line.*

You can also assign error bounds (similar to confidence intervals) on the estimate for y obtained from x. This involves first finding the *Standard Error of the Estimate.* A convenient formula for the standard error of the estimate is

$$s_{xy} = s_y\sqrt{(1-r^2)\frac{n}{n-2}}.$$

(This is a measure of the variability in the y's which is *not* due to the dependence on x.) This can be used to construct a "confidence interval" about the regression line. For example, $y \pm 1.64 s_{xy}$ would give a 90% confidence interval about the regression line, the 1.64 corresponding to a 90% confidence limit from the normal tables. Thus we would be 90% confident that the true y values fall between the upper and lower limits.

Example. (This is a continuation of the example in the section dealing with correlations.) A counselor is treating individuals with an eating disorder by administering a drug which enhances appetite. The counselor tries eleven different dosages; the

average daily dose was 0.5 mg with a standard deviation of 0.316 mg . The average weight loss was 0.500 lbs with a standard deviation of 3.223 pounds. The variables "dosage" and "weight change" are correlated with $r = 0.922$.

You administer a daily dosage of 0.85 mg to a patient. Predict the change in the patient's weight.

Solution.

Step 1. Since you are trying to predict change in weight, this must be the y variable. You are using the dosage to predict the weight loss, so the dosage must be the x variable. Our dictionary thus looks like:

average for x	$\bar{x}$	0.5
st. dev. for x	s_x	0.316
average for y	$\bar{y}$	-0.5
st. dev. for y	s_y	3.223
correlation coeef.	r	0.922
indiv. x observation	x_0	0.85

Note that the average change in weight is -0.5 pounds since the average weight *loss* is +0.5 pounds.

Step 2. The parameter m is

$$\begin{aligned} m &= \frac{r s_y}{s_x} \\ &= \frac{0.922 \times 3.223}{0.316} \\ &= 9.404. \end{aligned}$$

Step 3. The parameter b is

$$\begin{aligned} b &= \bar{y} - m\bar{x} \\ &= -0.5 + 9.404 \times 0.5 \\ &= 4.22. \end{aligned}$$

Step 4. This means that the regression equation is

$$y = 9.404x + 4.22.$$

Step 5. The predicted change in weight for a dosage of 0.85 is thus

$$\begin{aligned} y &= 9.404 \times 0.85 + 4.22 \\ &= 12.213 \end{aligned}$$

and so we expect our client to gain 12.213 pounds. ∎

Remark. The standard error of estimate in this problem is

$$\begin{aligned} s_{xy} &= s_y\sqrt{(1-r^2)\frac{n}{n-2}} \\ &= 3.223\sqrt{(1-0.922)^2\frac{11}{9}} \\ &= 3.223\sqrt{(0.00608)(1.222)} \\ &= 3.223 \times .0862 \\ &= 0.278. \end{aligned}$$

Thus, 95% error bounds for our prediction in the previous example are

$$12.213 \pm 1.960 \times 0.278$$

or

$$11.668 \quad \text{to} \quad 12.758.$$

Problems

1. Suppose that the following data are gathered by taking paired measurements x and y on subjects:

$\bar{x}$	567
s_x	111
$\bar{y}$	3.21
s_y	0.97

Suppose in addition that x and y are shown to be correlated at $r = 0.41$. Given an x observation of 483, predict the corresponding y value. (These are GRE/gradepoint data for the physical science and engineering at OU.)

2. Suppose that the following data are gathered by taking paired measurements x and y on subjects:

$\bar{x}$	68
s_x	12
$\bar{y}$	75
s_y	15

Suppose in addition that x and y are shown to be correlated at $r = -0.65$. Given an x observation of 56, predict the corresponding y value.

3. Suppose that the following data are gathered by taking paired measurements x and y on subjects:

$\bar{x}$	31.3
s_x	1.2
$\bar{y}$	14.6
s_y	4.8

Suppose in addition that x and y are shown to be correlated at $r = 0.45$. Given an x observation of 23, predict the corresponding y value.

4. The University of Oklahoma uses ACT scores to help place students in mathematics classes. The average grade in Calculus I is 2.31 (on a 4.0 scale) with a standard deviation of 0.89. Among all students in Calculus I, the average ACT math score is 26.2 with a standard deviation of 4.32. ACT scores and Calculus I grades are known to be correlated at $r = 0.43$.

You are confronted with a Native American student whose ACT score in mathematics is 22. Predict this student's Calculus I grade. This student is planning to enroll in Electrical Engineering (which requires five semesters of mathematics starting with Calculus I). The student is also an AFROTC cadet. Would you advise the student to enroll in Calculus I? Would you inform the student about the prediction which you just made?

5. In a group of 63 children selected at random from the ninth grade in the Norman Public Schools, the average IQ was 108 with a standard deviation of 16.9. For these same children, school records for their mothers show an average IQ of 103 with a standard deviation of 15.8. In addition, the IQs of the children and their mothers are correlated at $r = 0.87$. You are confronted with a ninth grader from Norman Public Schools whose mother's IQ is 108; predict the child's IQ.

Video Assignment.

View the following program(s) from the series *Against All Odds*:

Program	*Title*
8	*Describing Relationships*
9	*Correlation*
11	*The Question of Causation*