

2.3 Least-Square regressions

Example 2.10 How do children grow? The pattern of growth varies from child to child, so we can best understanding the general pattern by following the average height of a number of children. Here is the mean heights of a group of children in Kalama, an Egyptian village that was the site of a study of nutrition in developing countries. The data were obtained by measuring the heights of 161 children from the village each month from 18 to 29 months of age.

Data		Scatterplot
Age x in months	Height y in centimeters	
18	76.1	
19	77.0	
20	78.1	
21	78.2	
22	78.8	
23	79.7	
24	79.9	
25	81.1	
26	81.2	
27	81.8	
28	82.8	
29	83.5	

Correlation $r = 0.9944$

Regression line: $\hat{y} = 64.93 + 0.635x$

lecture 5

1

1. Least-square Regression

Suppose we have n observations on two variables x and y : $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Assume there is a strong linear association between x and y . Then we may fit a straight line

$\hat{Y} = a + bX$ to the data. Where **a**---intercept, **b**---slope

We use least square method to calculate a and b : i.e. minimize

$$\sum (y_i - a - bx_i)^2$$

We have

$$b = r \frac{s_y}{s_x} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

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

The equation $\hat{y} = a + bx$ is called the least-squares (LS) regression line.

2. Prediction with LS regression line

We can use a regression line to predict the response y for a specific value of the explanatory variable x .

But be aware of **extrapolation** which 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 reliable.

lecture 5

2

3. Interpreting the regression line

For LS regression line $\hat{y} = a + bx$,

b----- the amount by which y changes when x increases by one unit.

a----- the value of y when x=0.

Note: the LS regression line always passes through the point (\bar{x}, \bar{y}) on the graph of y against x. so the regression line may be written in an alternative way:

$$y - \bar{y} = r \frac{S_y}{S_x} (x - \bar{x})$$

4. Correlation and regression

r^2 -----coefficient of determinations $r^2 = \frac{S_{\hat{y}}^2}{S_y^2}$

It is the fraction of the variation in the values of y that is explained by the LS regression of y on x.

Variation of y has two sources:

- Values of y vary as x changes.
- Values of y are scattered above and below \hat{y} .

2.4 Cautions about Regression and Correlation

1. Residuals

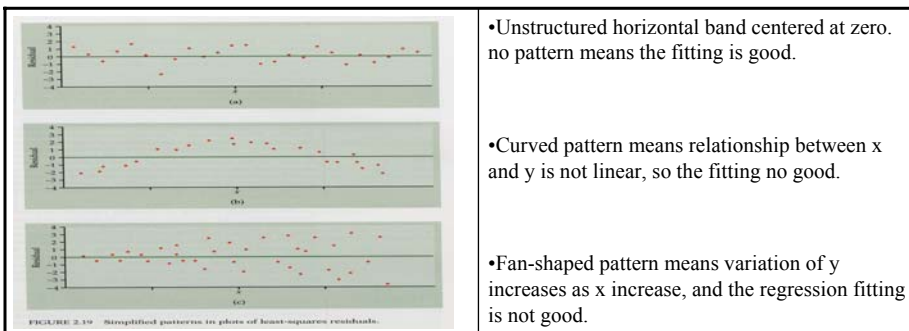
A residual is the difference between an observed value of the response variable and the value predicted by the regression line. That is,

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

2. Residual plot

A residual plot is a scatter plot of the regression residuals against the explanatory variable.

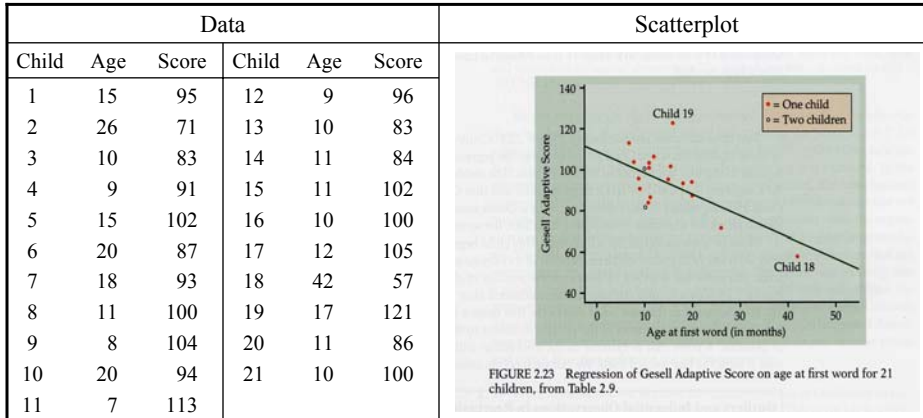
The typical pattern of residual plots:



3. Outlier and influential observation

Example:

Does the age at which a child begins to talk predict a later score on a test of mental ability? A study of cognitive development in young children recorded the age in months at which each of 21 children spoke their first word and their Gesell Adaptive Score, the result of an ability test taken much later.

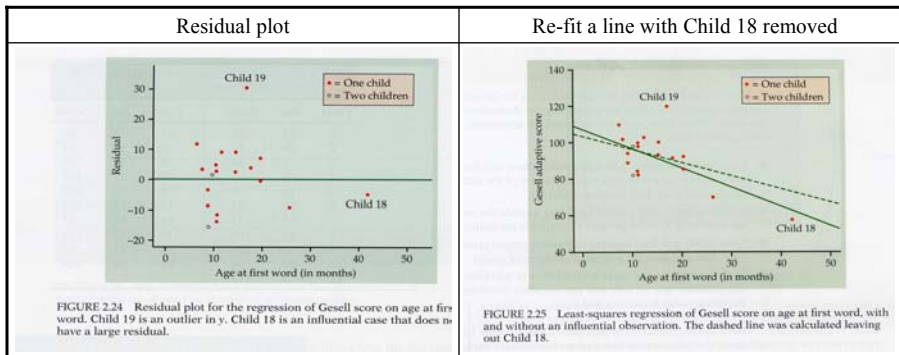


The fitted regression line is: $\hat{y} = 109.87 - 1.127x$, $r^2=0.41$

lecture 5

5

Example (continue)



- An **outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the **y direction** of a scatter plot have large regression residuals, but other outliers need not have large residuals.

- An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in the **x direction** of a scatterplot are often influential for the least squares regression line.

lecture 5

6

4. Lurking variables

It 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.

Example:

The Math. Department of a large state university must plan the number of sections and instructors required for its elementary courses. The department hopes that number of students in those courses can be predicted from number of first-year students. Here is the data for several years.

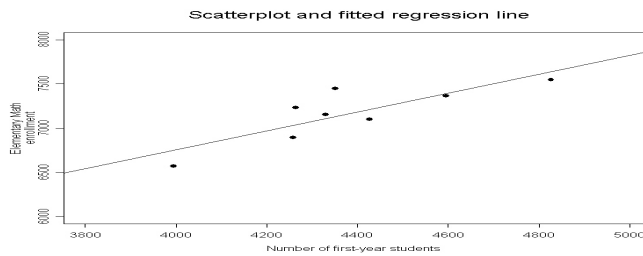
X----no. of first-year students, Y----no. of students who enroll in elementary Math. courses.

Year	1993	1994	1995	1996	1997	1998	1999	2000
X	4595	4827	4427	4258	3995	4330	4265	4351
Y	7364	7547	7099	6894	6572	7156	7232	7450

Regression line:

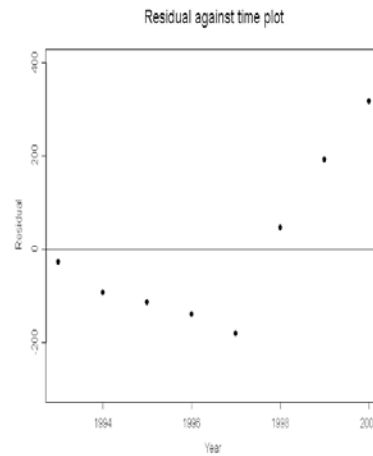
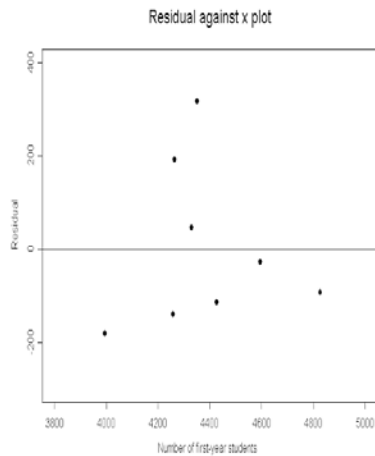
$$\hat{Y} = 2492.69 + 1.0663x$$

$$r^2 = 0.694$$



lecture 5

7



Comments: The 2nd residual plot suggests that a change took place between 1997 and 1998 that caused a higher proportion of students to take math courses beginning in 1998.

In fact, one of the schools in the university changed its program to require that entering students take another math course. This change is the lurking variable that explains the pattern we observed.

lecture 5

8

5. Causation, Common response, Confounding

▪Causation

- The association between two variables X and Y is explained by a direct cause-and-effect link. One variable X causes the other Y.

- Example: X = mother's body mass index
Y = daughter's body mass index

▪Common response

- The observed association between the variables X and Y is explained by a lurking variable Z. Both X and Y changes in response to changes in Z. We say that X and Y are common responses of Z. This common response creates an association even though there may be no direct causal link between X and Y.

- Example: X = a student's SAT score as a high school senior
Y = the student's first-year college grade point average.

▪Confounding

- Two variables are confounded when their effects on a response variable can't be distinguished from each other. The confounding variables may be either explanatory variables or lurking variables.

- Example: X = number of years of education a worker has
Y = the worker's income

lecture 5

9

Summary of Correlation and Regression:

Correlation and Regression are powerful tools for measuring the association between two variables and for expressing the dependence of one variable on the other. But they **must be used with an awareness of their limitations**. Below are some basic rules:

- Correlations measures **only linear association**, and fitting a straight line makes sense only when the overall pattern of the relationship is linear. Always plot your data before calculating.
- Extrapolation** (using a fitted model far outside the range of the data that we used to fit it) often produces unreliable predictions.
- Correlation and least-squares regression are **not resistant**. Always plot your data and look for potentially influential points.
- Lurking variables** can make a correlation or regression misleading. Plot the residual against time and against other variables that may influence the relationship between X and Y.
- When you calculate and interpret correlation and regression, be aware of possible **common response and confounding effects**. Even a very strong association between two variables is not by itself good evidence that there is a cause-and-effect link between the variables.
- How to establish causation?** The best method to establish a direct causal link between X and Y is to conduct a carefully designed experiment in which the effect of possible lurking variables are controlled. We will study this in next chapter.

lecture 5

10

Example: We have seen that investment theory uses the standard deviation of returns to describe the volatility or risk of an investment. To describe how the risk of a specific security is related to that of the market as a whole, we use least-squares regression. Figure 2.42 plots the monthly percent total return y on Philip Morris common stock against the monthly return x on the Standard & Poor's 500-stock index, which represents the market, for the period between August 1990 and June 1997. The one clear outlier turns out not to be very influential. Here are the basic descriptive measures:

$$\bar{x}=1.304 \quad s_x=3.392 \quad r=0.5251$$

$$\bar{y}=1.878 \quad s_y=7.554$$

- (a) Find the equation of the least-squares line from this information. What percent of the volatility in Philip Morris stock is explained by the linear relationship with the market as a whole?
- (b) Explain carefully what the slope of the line tells us about how Philip Morris stock responds to changes in the market. This slope is called 'beta' in investment theory.

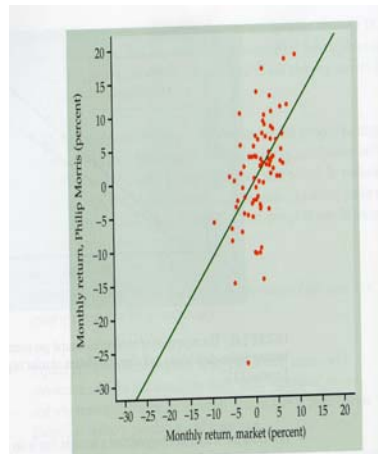


FIGURE 2.42 Monthly returns on an individual stock plotted against the returns for the stock market as a whole, with the least-squares line. See Exercise 2.118.