# Chapter 5. Regression

## Regression Lines

**Definition.** 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 $x$.

**Note.** The text gives a review of the algebra and geometry of lines on pages 117 and 118. We give a quick example.

**Example.** Graph the line with slope $m = -1/2$ and passing through the point $(x_0, y_0) = (2, 8)$.

## The Least-Squares Regression Line

**Definition.** 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.

**Note.** Notice that the roles of $x$ and $y$ are very distinct! The idea here is that the $x$ values (the explanatory variable) is measured precisely, but there is variation in the $y$ values (the response variable). If you have had some exposure to calculus, then you might recognize the "as small as possible" comment as part of a minimization problem. By performing the minimization procedure (involving partial derivatives and critical points), we are lead to the following formulae.

**Definition.** We have data on an explanatory variable $x$ and a response variable $y$ for $n$ individuals. From the data, calculate the means $\overline{x}$ and $\overline{y}$, 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 the **slope** $b = r\dfrac{s_y}{s_x}$ and **intercept** $a = \overline{y} - b\overline{x}$. (We use $\hat{y}$ in the equation to represent the fact that it is the *predicted* response $\hat{y}$ for given $x$.

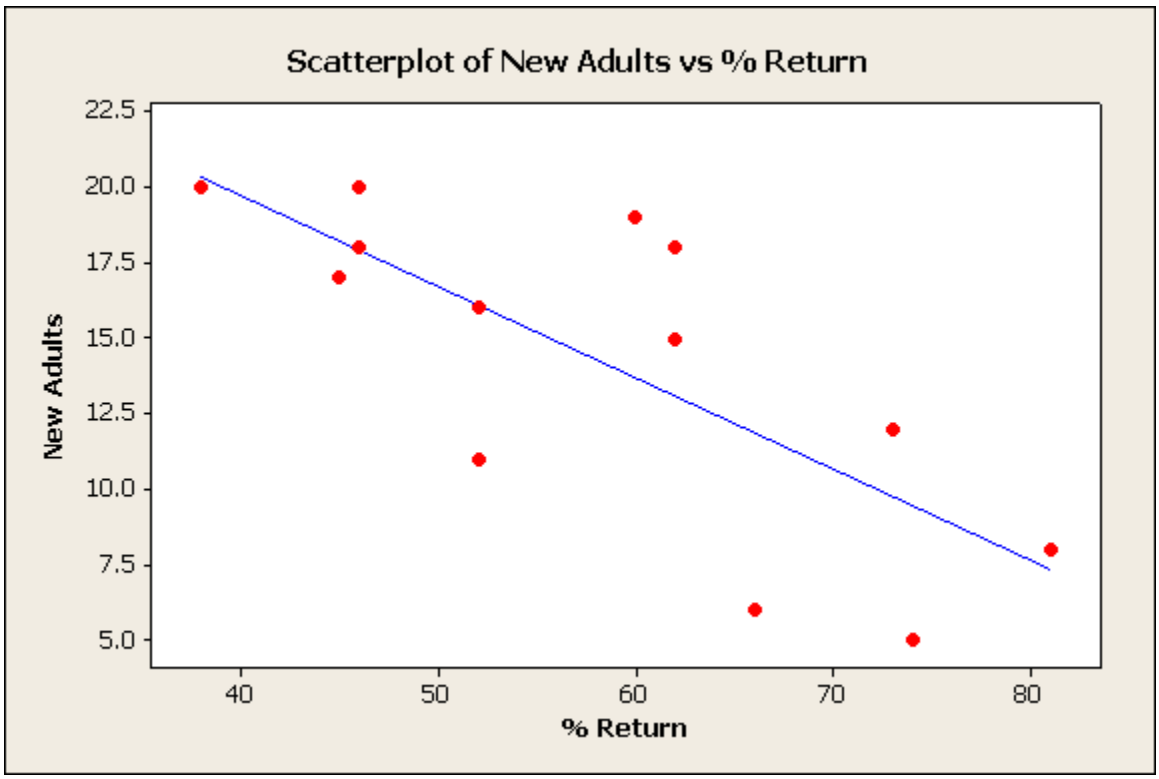
**Example.** Exercise 5.4 page 122.

**Partial Solution.** Using Minitab commands `Stat`, `Regression`, `Regression`, we get the output:

```
The regression equation is    New Adults = 31.9 - 0.304 % Return
```

In other words, with $x$ as 'Percent Return' and $y$ as 'New adults', the least-squares regression line is $y = -0.304x + 31.9$. The correlation is $r = -0.748$.

Using Minitab commands `Graph`, `Scatterplot`, `With Regression` we get:



## Facts about Least-Squares Regression

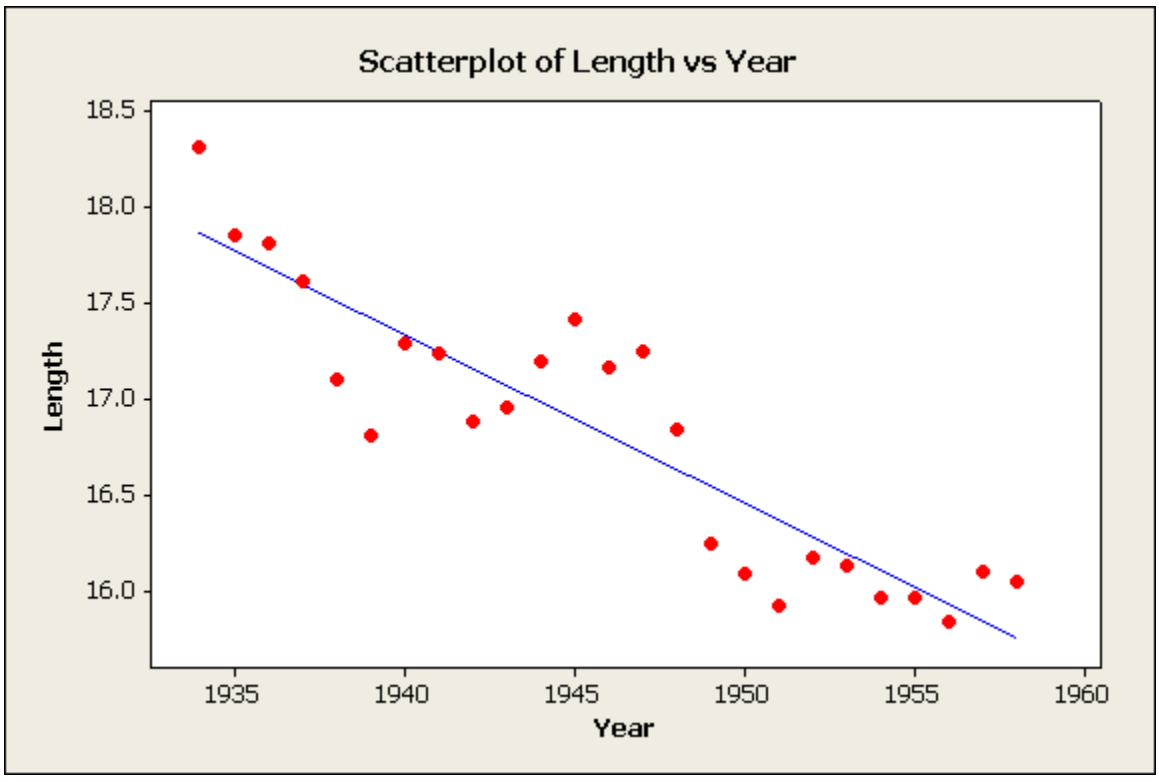
**Note.** The text mentions the following facts:

1. The distinction between explanatory and response variables is essential in regression.
2. The least-squares regression line always passes through the point $(\overline{x}, \overline{y})$.
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 $x$.

The last fact tells us that $r^2$, not $r$, is the best description of how strong a linear relationship between $x$ and $y$ is. In Exercise 5.4, we have $r = -0.748$ and so $r^2 = 0.560$. This means that 56% of the variation in 'New adults' as explained by the linear relationship with 'Percent return'.

**Example S.5.1. Linear Stooges.**
Find the least-squares regression line between the explanatory variable 'year' and the response variable 'length' for the data of Example S.4.1. What percentage of the variation in length can be explained by the linear relationship with year?

**Solution.** Minitab yields the least squares line $y = -0.0879x + 188$ with year $x$ and length $y$. The correlation coefficient is $r = -0.904$. Hence $r^2 = 0.817$ and $0.817 \times 100\% = 81.7\%$ of the variation in length can be explained by the linear relationship with year. The graph is:



# Residuals

**Definition.** A **residual** is the difference between an observed value of the rsponse 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:

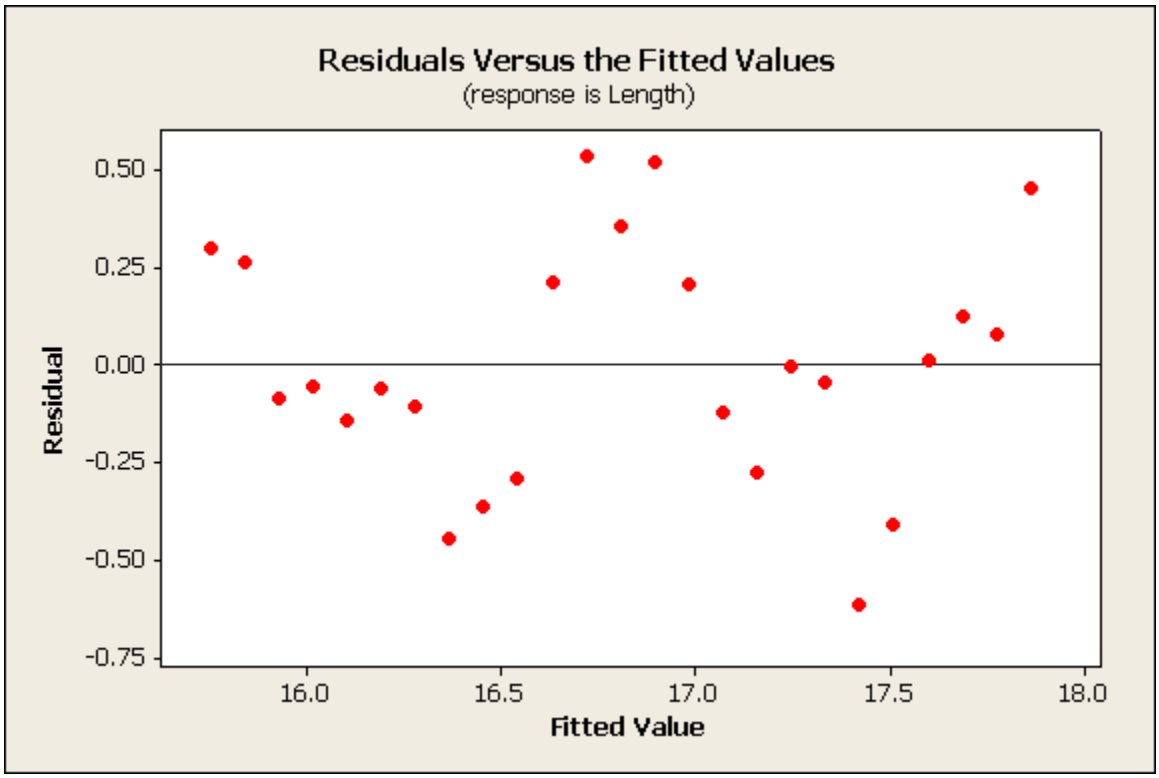
$$\text{residual} \;=\; \text{observed } y - \;\text{predicted } y = y - \hat{y}.$$

**Example S.5.2. Stooge Residual.**
Use Minitab to find the residuals for the year and length regression line of Example S.5.1.

**Solution.** Go through `Stat`, `Regression`, and `Regression`. Click on the `Results` button and select `In addition, the full table of fits and residuals.` The output includes:

| Obs | Year | Length | Fit | Residual | Obs | Year | Length | Fit | Residual |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1934 | 18.3100 | 17.8613 | 0.4487 | 14 | 1947 | 17.2500 | 16.7189 | 0.5311 |
| 2 | 1935 | 17.8500 | 17.7734 | 0.0766 | 15 | 1948 | 16.8400 | 16.6310 | 0.2090 |
| 3 | 1936 | 17.8100 | 17.6856 | 0.1244 | 16 | 1949 | 16.2500 | 16.5432 | -0.2932 |
| 4 | 1937 | 17.6100 | 17.5977 | 0.0123 | 17 | 1950 | 16.0900 | 16.4553 | -0.3653 |
| 5 | 1938 | 17.1000 | 17.5098 | -0.4098 | 18 | 1951 | 15.9200 | 16.3674 | -0.4474 |
| 6 | 1939 | 16.8100 | 17.4219 | -0.6119 | 19 | 1952 | 16.1700 | 16.2795 | -0.1095 |
| 7 | 1940 | 17.2900 | 17.3341 | -0.0441 | 20 | 1953 | 16.1300 | 16.1917 | -0.0617 |
| 8 | 1941 | 17.2400 | 17.2462 | -0.0062 | 21 | 1954 | 15.9600 | 16.1038 | -0.1438 |
| 9 | 1942 | 16.8800 | 17.1583 | -0.2783 | 22 | 1955 | 15.9600 | 16.0159 | -0.0559 |
| 10 | 1943 | 16.9500 | 17.0704 | -0.1204 | 23 | 1956 | 15.8400 | 15.9280 | -0.0880 |
| 11 | 1944 | 17.1900 | 16.9826 | 0.2074 | 24 | 1957 | 16.1000 | 15.8402 | 0.2598 |
| 12 | 1945 | 17.4100 | 16.8947 | 0.5153 | 25 | 1958 | 16.0500 | 15.7523 | 0.2977 |
| 13 | 1946 | 17.1600 | 16.8068 | 0.3532 | | | | | |

Click on the `Graphs` button and select `Residuals versus fits` to get the following:




**Definition.** 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. The mean of the residuals is always 0.

## Influential Observations

**Definition.** An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in either the $x$ or $y$ 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.

## Cautions About Correlation and Regression

**Definition. 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.

**Definition.** 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.

### Example S.5.3. Lurking Stooges

In Example S.5.1, we found a fairly strong correlation between the length of a Three Stooges film and the year of filming. What could be a lurking variable in this case?

## Association Does Not Imply Causation

**Note.** An association (or correlation) 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 causes changes in $y$. The best way to get good evidence that $x$ causes $y$ is to do an experiment in which we change $x$ and keep lurking variables under control.

**Example S.5.4. Causal Stooges**
Discuss the possibility of a causal relationship between the year $x$ and the average length of a Stooges film $y$ in Example S.5.1.

**Solution.** There is no causal relationship between year and film length. It is not the year which is *causing* the films to be on average shorter, but probably some lurking variable. On page 88 of Fleming's book, it is commented that the budget for Stooges shorts declined over time. This is a more likely explanation for the decrease in the length of the films and a very good candidate for a lurking variable. Fleming's book also mentions that stock footage from earlier shorts was repeated in the later shorts, particularly around the mid 1950s. This repetition of footage will have an effect in our future dis-

cussion of independence in the setting off hypothesis testing (in Part III).

**Note.** The book gives a nice discussion of when an association is causal in Example 5.10: Does smoking cause lung cancer?

*rbg-12-21-2008*