Simple Linear Regression

Simple Linear Regression: concept

Simple linear regression seeks to summarize the relationship between the predictand (y) and the predictor (x) by a single straight line.

$$\hat{y} = a + bx$$

The regression procedure chooses that line producing the least error for predictions of y given observations of x (as indicated by "e" in the figure)

$$e_i = y_i - \hat{y}(x_i).$$

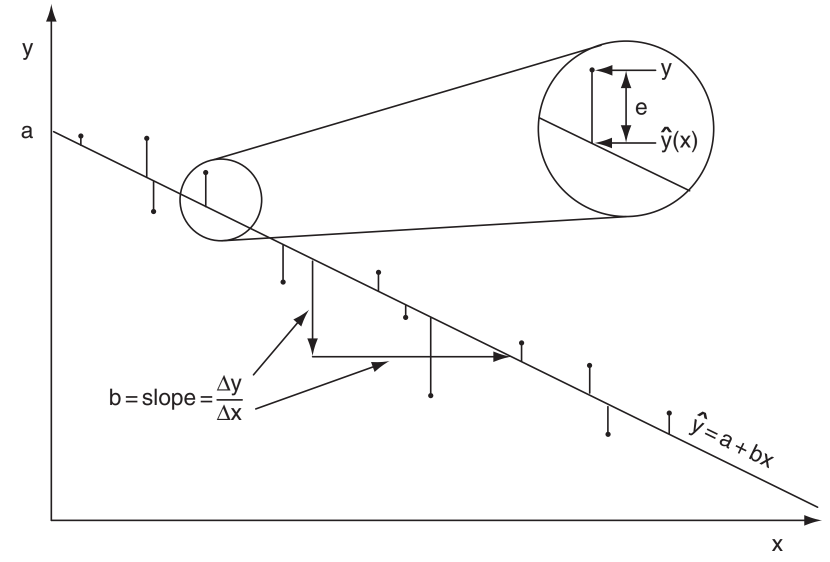


FIGURE 6.1 Schematic illustration of simple linear regression. The regression line, $\hat{y} = a + bx$, is chosen as the one minimizing some measure of the vertical differences (the residuals) between the points and the line. In least-squares regression that measure is the sum of the squared vertical distances. Inset shows the residual, *e*, as the difference between the data point and the regression line.

Commonly used error measures:

- Minimizing the sum of the squared errors (least squares)
- least absolute deviation (LAD): minimize the sum of the absolute errors

Analytic Expressions for the linear least-squares regression

The sum of squared residuals (errors) can be written as

$$\hat{y} = a + bx$$

$$\sum_{i=1}^{n}(e_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - [a + bx_i])^2,$$

To minimize the sum of squared residuals, we set the derivatives with respect to the parameters a and b to zero:

$$\frac{\partial \sum_{i=1}^{n}(e_i)^2}{\partial a} = \frac{\partial \sum_{i=1}^{n}(y_i - a - bx_i)^2}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0$$

$$\frac{\partial \sum_{i=1}^{n}(e_i)^2}{\partial b} = \frac{\partial \sum_{i=1}^{n}(y_i - a - bx_i)^2}{\partial b} = -2\sum_{i=1}^{n}[x_i(y_i - a - bx_i)] = 0.$$

Analytic Expressions for the linear least-squares regression

Solving the previous two equations, we have

$$b = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n}(x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

and

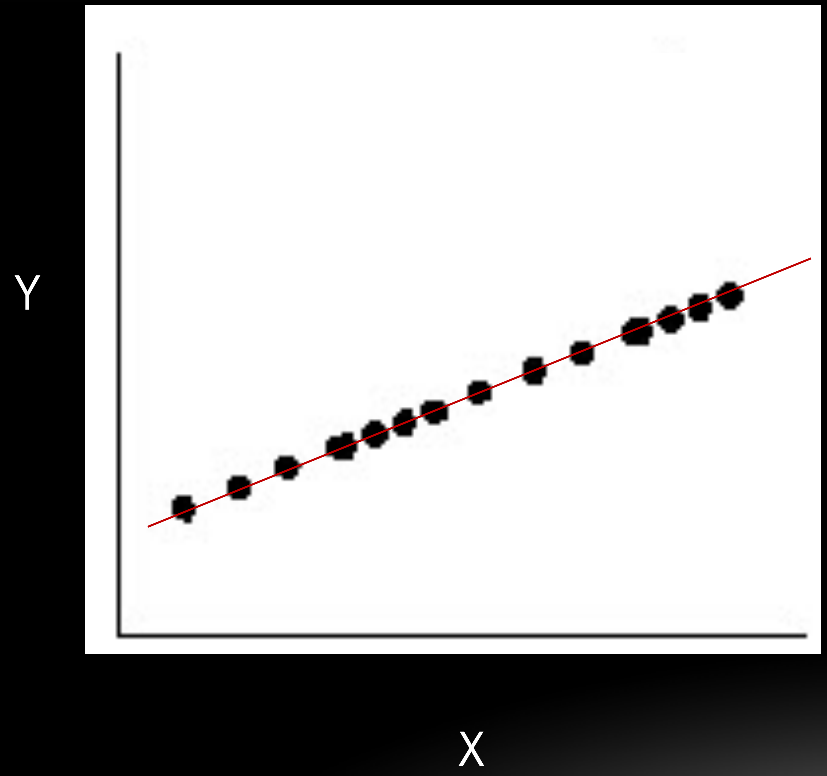
$$a = \bar{y} - b\bar{x}.$$

The existence of analytic solutions is an advantage of the linear least-squares regression.

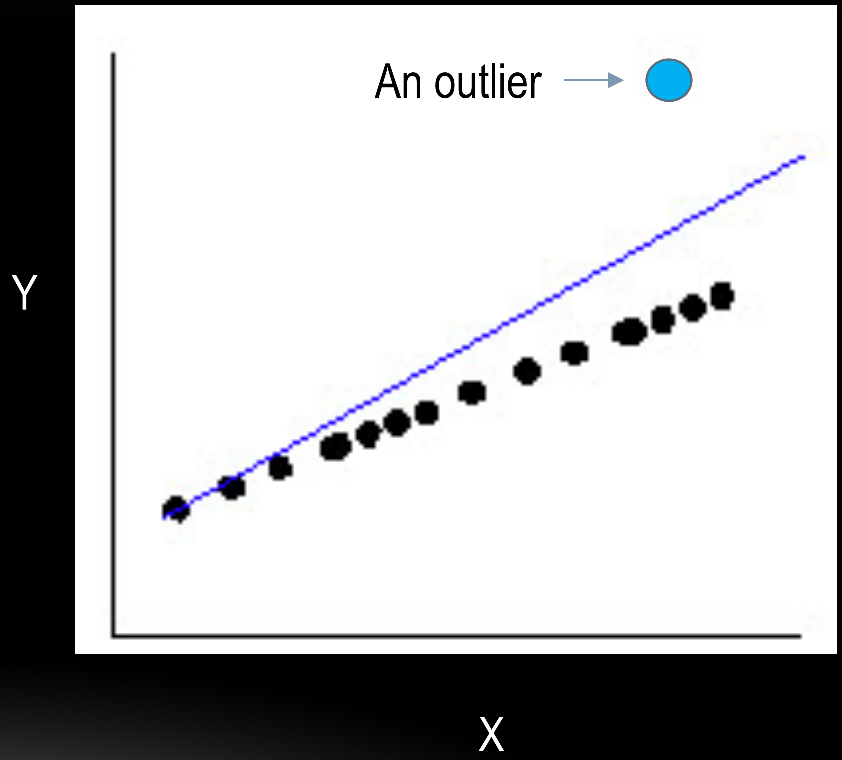
Least squares regression is not resistant to outliers!

because the errors are squared.

A perfect linear relationship between X and Y

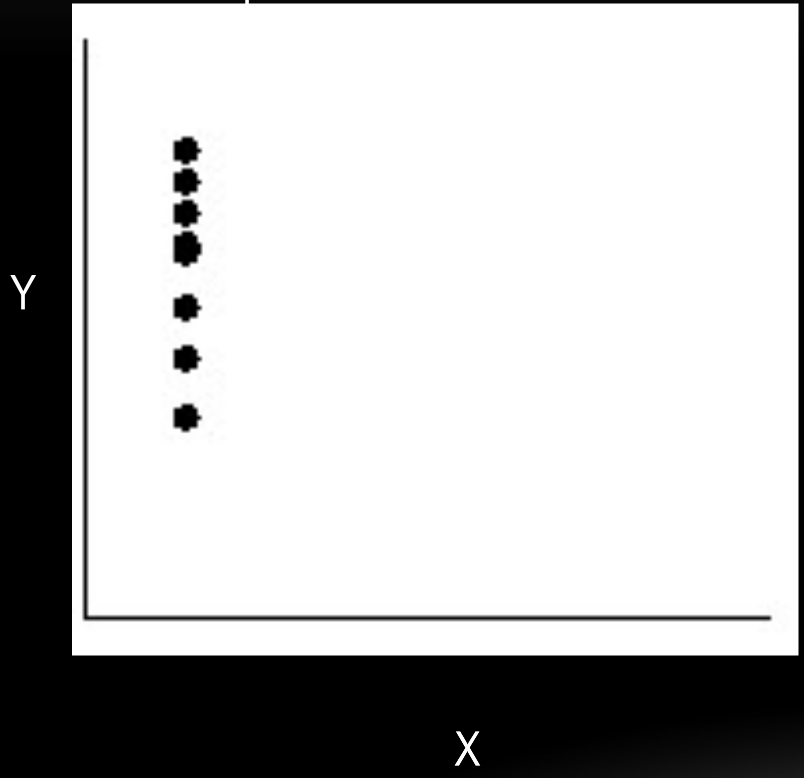


When we have an outlier

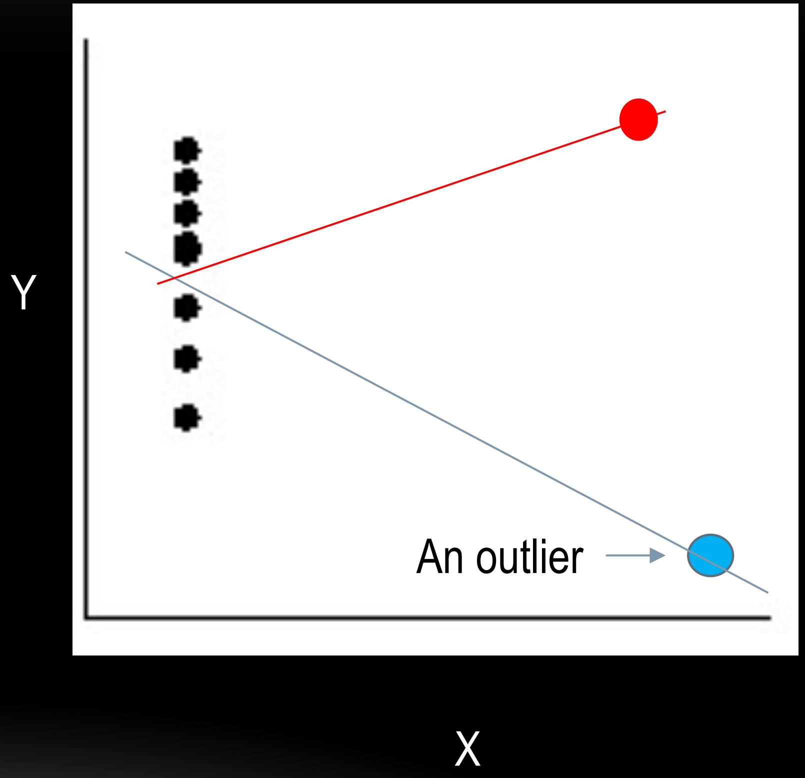


Another Example

Y is independent of X



When we have an outlier



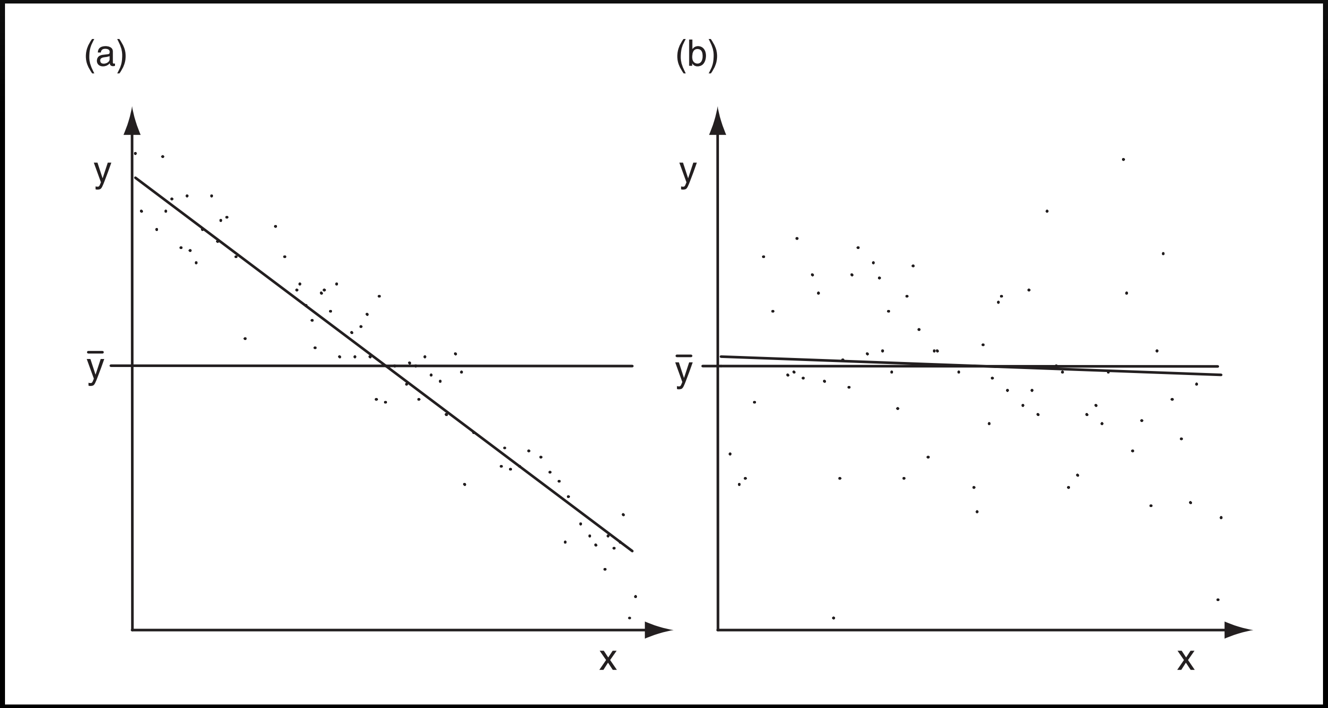
LAD regression is resistant to outliers, but the lack of analytical formulas for the regression function means that the estimation must be iterative.

How do we measure the fit of a regression?
Goodness-of-Fit Measures

- Mean squared error (MSE): the mean squared differences between the prediction and the observation, a common measure of forecast accuracy.

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}[y_i - \hat{y}(x_i)]^2.$$

- Coefficient of determination (R^2): the portion of variations of y described by the regression. For simple linear regression, it is equal to the square of the Pearson correlation between x and y.



Panel a: small MSE and large R^2
Panel b: large MSE and small R^2

References

- Wilks, 2011: “Statistical Methods in the Atmospheric Sciences”, Sections 6.2 and 6.4