Linear regression

Part 2: The method of least squares

By: Noureddin Sadawi, PhD

University of London

Sample statistics

- Sample statistics, denoted by lower case letters *a* and *b*, are computed as estimates of the population parameters *A* and *B* respectively.
- Substituting the values *a* and *b* for the parameters *A* and *B* respectively, in the regression equation, we obtain the *estimated (simple linear) regression equation*.

Estimated regression equation

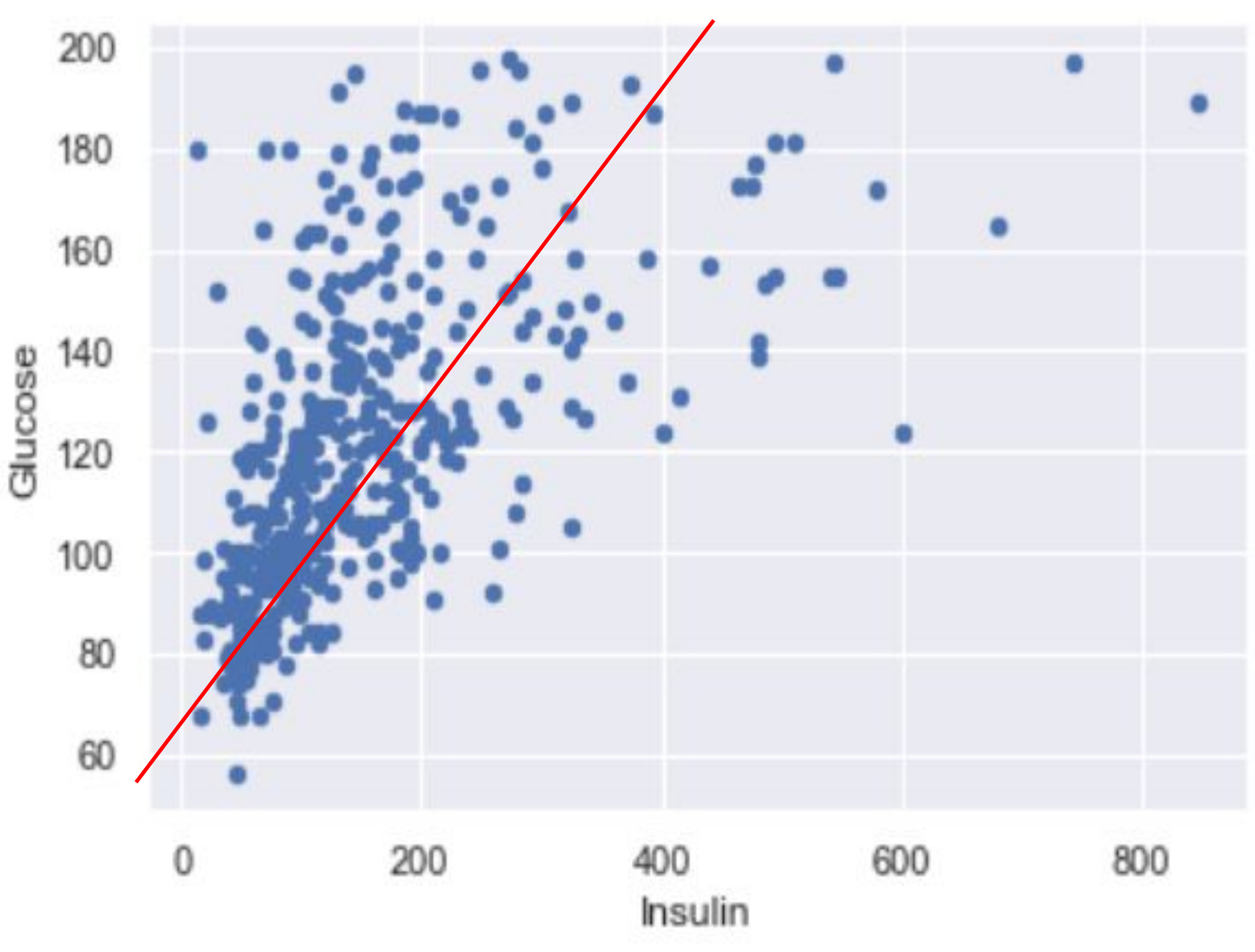
- The estimated regression equation is given by:

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

- The graph of the estimated regression equation is called the estimated regression line.
- *a* is the y intercept and *b* is the slope or gradient.
- In general, $\hat{y}$ is the point estimate of *E(Y)*, which is the mean value of *Y* for a given value of *X*.

Correlation coefficient is not enough

- Scatter diagram for regression analysis is constructed with the dependent variable *Y* on the vertical axis and the independent variable *X* on the horizontal axis.
- The scatter diagram allows us to observe the data graphically and to draw preliminary conclusions about the possible relationship between the variables.



Simple and multiple linear regression

- Simple linear regression involves one dependent variable and one independent variable.
- Multiple linear regression is one involving one dependent variable and two or more independent variables.
- Regression can be used for prediction, estimation, modelling causal relationships and hypothesis testing.

The method of least squares

- The method of least squares is a procedure which involves using sample data to find the estimated regression equation.

- It uses the sample data to provide the values of a and b that minimize the sum of squares of the deviations between the observed values of the dependent variable and the estimated values of the dependent variable.

The method of least squares

- This is expressed mathematically as:

$$Min \sum (y - \hat{y})^2$$

- Where *y* represents the observed value of the dependent variable and $\hat{y}$ represents the estimated value of the dependent variable.
- This expression is known as **the least squares criterion**.

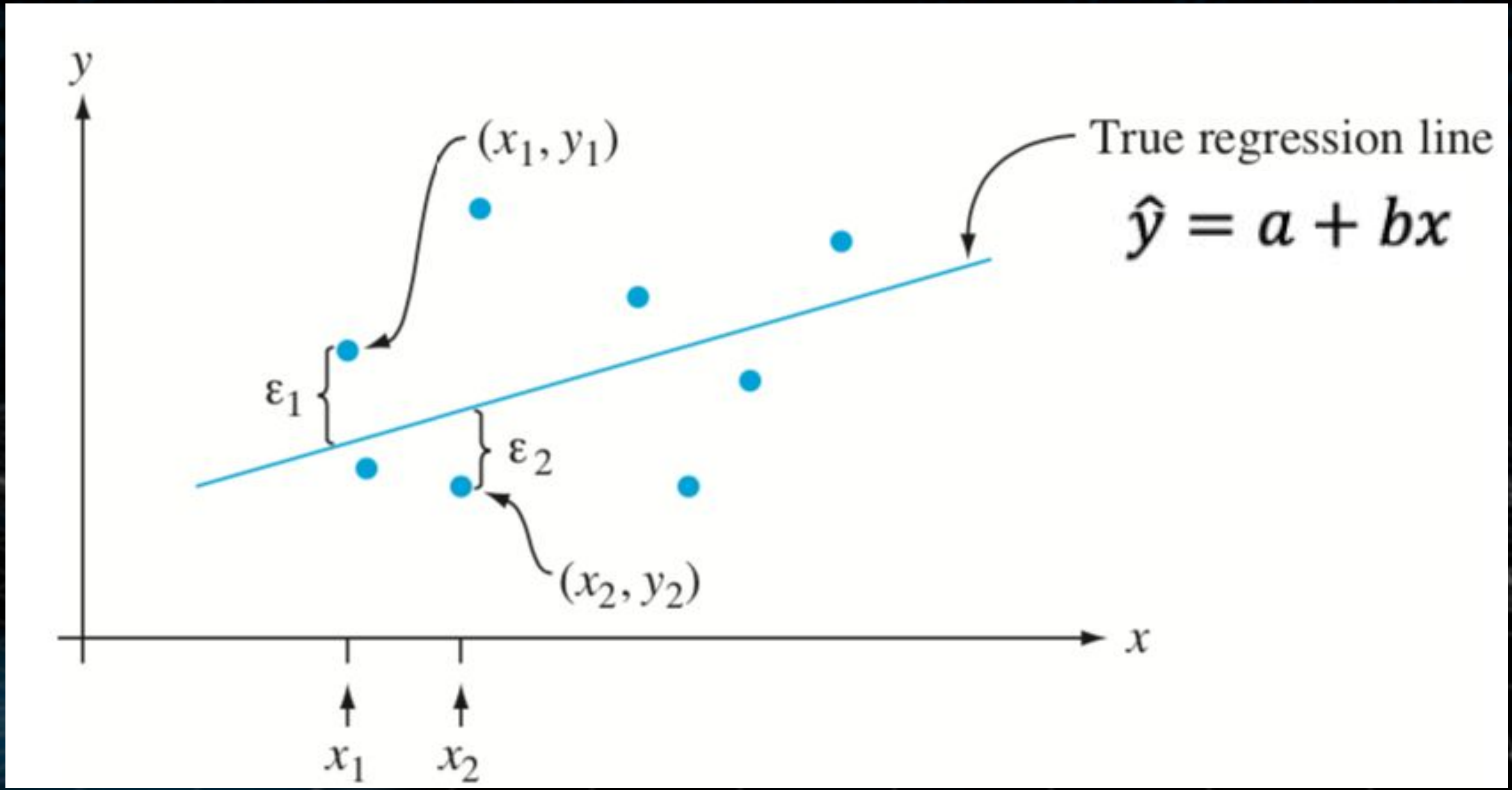
The method of least squares

- The values of a and b that minimise the least squares criterion are given by the equations:

$$b = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

- Where *x* represents the observed values of the independent variable.
- *y* represents the observed values of the dependent variable.
- $\bar{x}$ is the mean for variable *x* and $\bar{y}$ is the mean for variable *y*.

The simple linear regression model



The null hypothesis in linear regression

- The H_0 in linear regression is: the slope is zero.
- In other words, there is no significant linear relationship between the independent variable(s) and the dependent variable.
- It is possible to compute the standard error (s.e.) of the slope, and therefore compute a C.I. to test the null hypothesis (we spoke about the standard error in Topic 3).
- It is also possible to compute a p-value for the coefficient(s) of the regression equation.

Example

- Suppose we are studying the relationship between the minimum body temperature (in ^{o}C) and the uninterrupted sleep duration in hours for a number of very young children.
- We collect the minimum body temperature and the sleep duration.
- Let us assume that a = 218.73 and b = - 6.05

```
unint. sleep duration = 218.73 - 6.05 * min body temp
```

- This means that uninterrupted sleep decreases by 6.05 hours for every degree increase in the minimum body temperature.
- The s.e. of the slope = 1.41.
- A 95% C.I. for the slope is given by: - 6.05 +/- (1.96*1.41) = (-8.81, -3.29).

Example

- If we test against the null hypothesis (i.e. zero slope or b = 0):

$$(-\ 6.05 - 0)\ /\ 1.41 = -4.29,\ \text{p-value} < 0.001$$

- The slope of the line is significantly different from zero.
- This means we have enough evidence to believe that sleep duration decreases on average as the child's minimum body temperature increases.
- The average drop in sleep duration for a 1 ^{o}C increase in minimum temperature is between 3.29 to and 8.81 hours (feel free to convert it to minutes).
- Similar full study: https://adc.bmj.com/content/66/4/521