Statistics 120
Fitting a Straight Line

The Problem

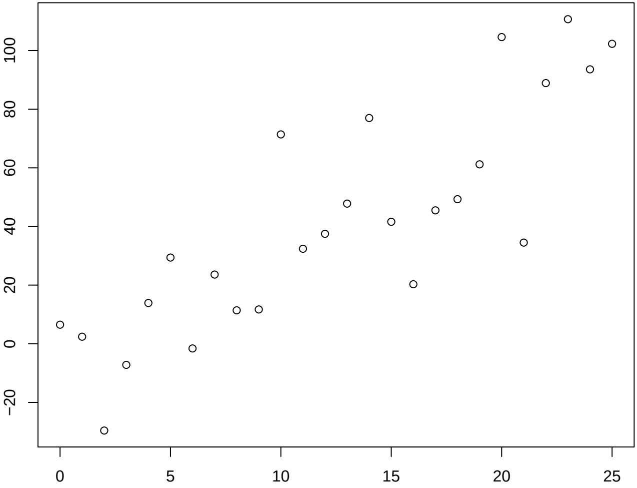
Given a set of points

$$(x_1, y_1), \ldots, (x_n, y_n),$$

how do we find a straight line

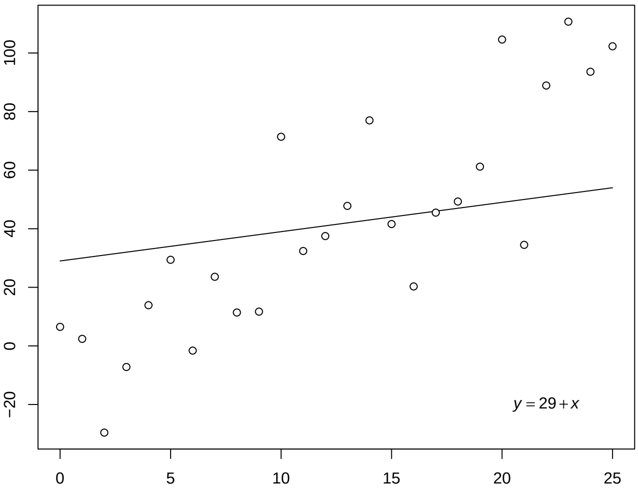
$$y = a + bx$$

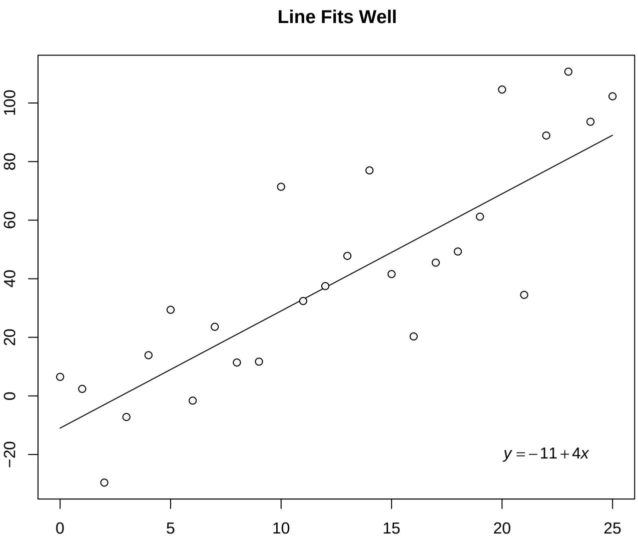
which provides a good description of the general trend underlying the points?



100
80
60
40
20
0
−20
0
5
10
15
20
25

Line Fits Badly





Line Fits Well
100
80
60
40
20
0
−20
0
5
10
15
20
25
$y = -11 + 4x$

Fitting Criteria

- Any assessment of how well a line fits a set of points must be based on how far the line deviates from the points.

$$d_i = y_i - (a + bx_i), \quad i = 1, \ldots, n$$

- It makes sense to use the absolute deviations $|d_i|$ rather than the raw deviations d_i.

- There are many different measures of how well a line fits a set of points.

Measures of Fit Quality

- Sum of Absolute Deviations

$$P(a,b) = \sum_{i=1}^{n} |\, y_i - (a + bx_i)\, |$$

- Sum of Squared Deviations

$$Q(a,b) = \sum_{i=1}^{n} |\, y_i - (a + bx_i)\, |^2$$

- Maximum Deviation

$$R(a,b) = \max |\, y_i - (a + bx_i)\, |$$

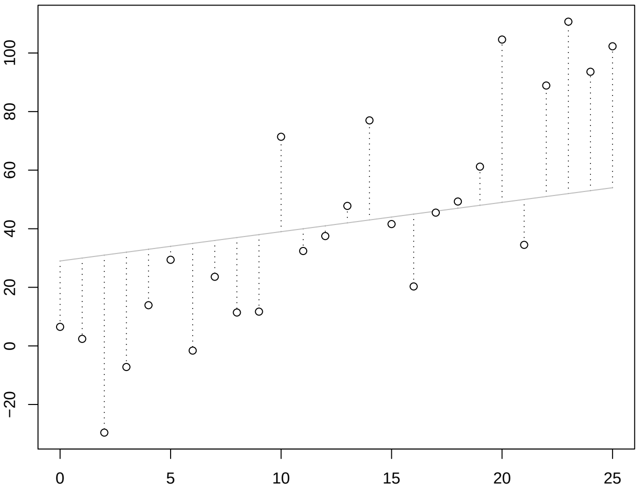
Least Squares

- The most commonly used fitting criterion is that of *least squares*.
- This means that we find the best fitting line by choosing a and b to minimise

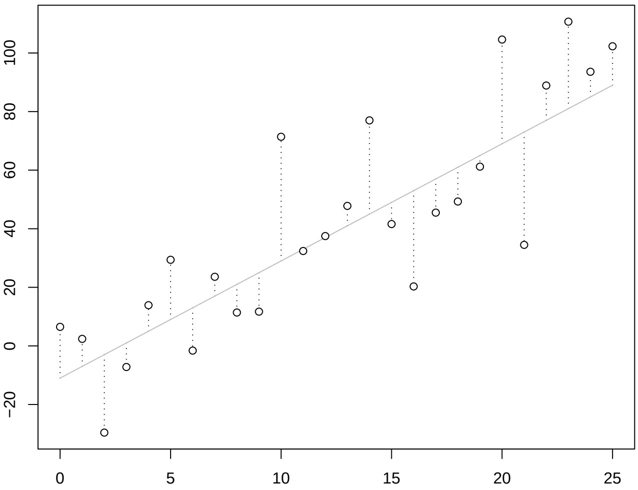
$$Q(a,b) = \sum_{i=1}^{n} |\, y_i - (a + bx_i) \,|^2$$

- The justification for using this choice is that it produces the simplest statistical theory.

Q(29, 1) = 25003.35



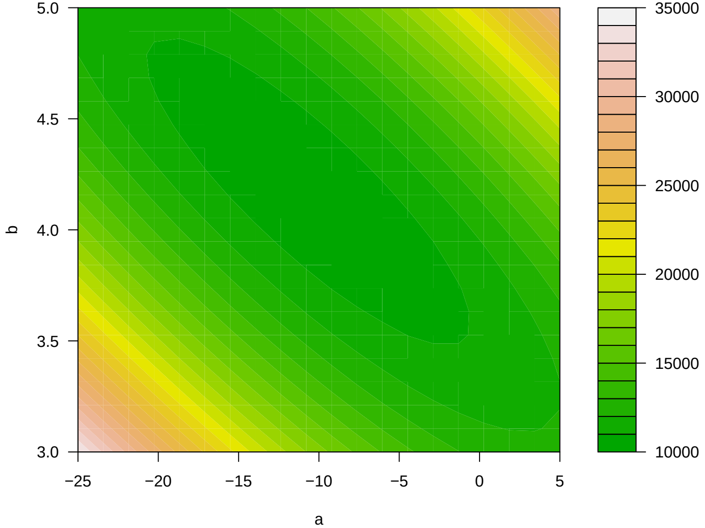
Q(−11, 4) = 10490.95



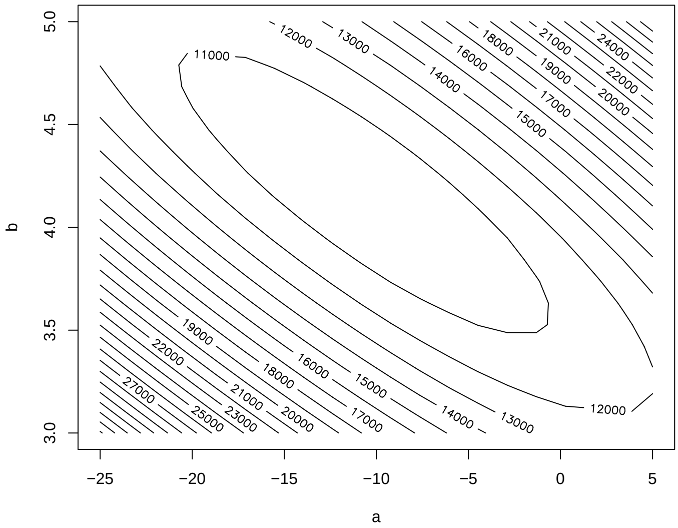
Finding the Best Slope and Intercept

- There are a number of ways of finding the best fitting slope and intercept.
- The simplest method is *exhaustive search*.
- To carry out this method, we compute the value of $Q(a,b)$ over a finely spaced grid.
- The results can be displayed with a *contour plot*.

A Contour Plot of Q(a,b)



A Contour Plot of Q(a,b)



Precise Determination of a and b

- The contour plots show that the best values of a and b are in the region of -10 and 4.2, but it is hard to be more precise.
- It is possible to finer and finer grids to zero-in on the best values, but it is possible to derive an exact formula for the best values.
- This will require a small diversion into mathematics.

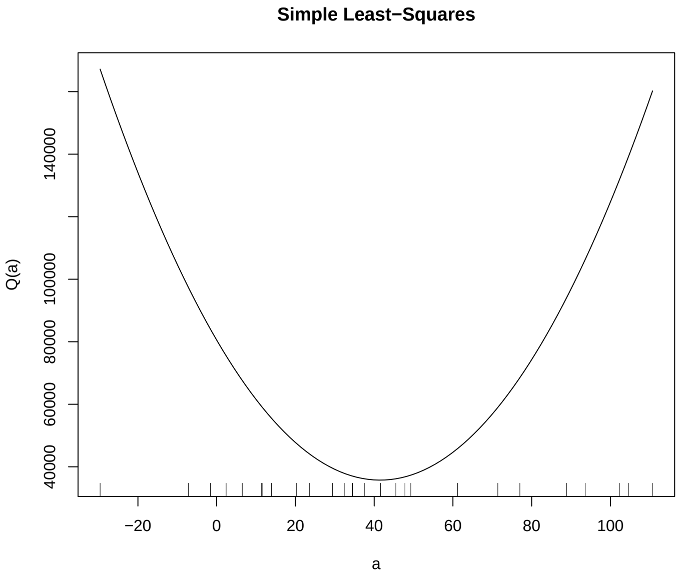
A Simplified Problem

- Suppose we have data values $y_i, \ldots, y_n$ and we want to locate the point which minimises

$$Q(a) = \sum_{i=1}^{n} (y_i - a)^2$$

- One way to proceed is to simply plot $Q(a)$ as a function of a.

- In practise we compute $Q(a)$ at a grid of points and we join up the dots.


Simple Least−Squares
Q(a)
40000
60000
80000
100000
140000
−20
0
20
40
60
80
100
a

Formal Minimisation

- $Q(a)$ is a smooth function of a, so it can be minimised using calculus.

Formal Minimisation

- $Q(a)$ is a smooth function of a, so it can be minimised using calculus.
- We want the point where $Q'(a) = 0$.

$$Q'(a) = \frac{d}{da}\sum_{i=1}^{n}(y_i - a)^2 = 2\sum_{i=1}^{n}(y_i - a)$$

Formal Minimisation

- $Q(a)$ is a smooth function of a, so it can be minimised using calculus.

- We want the point where $Q'(a) = 0$.

$$Q'(a) \; = \; \frac{d}{da}\sum_{i=1}^{n}(y_i - a)^2 \; = \; 2\sum_{i=1}^{n}(y_i - a)$$

- The equation $Q'(a) = 0$ can be solved for a.

$$\sum_{i=1}^{n}(y_i - a) = 0, \qquad \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} a, \qquad \sum_{i=1}^{n} y_i = na,$$

The Sample Mean

- We have just shown that $\bar{y}$ is the value of a which minimises the function:

$$Q(a) = \sum_{i=1}^{n} (y_i - a)^2$$

- The sample mean is the solution of a least-squares minimisation problem.

The Least-Squares Intercept and Slope

- Using methods from calculus it is possible to derive explicit estimates of slope and intercept.

$$\begin{aligned}\widehat{\alpha} &= \bar{y} - \widehat{\beta}\bar{x} \\ \widehat{\beta} &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}.\end{aligned}$$

Least Squares in R

- The function `lm` computes the least-squares estimates of slope and intercept.

- Given variables `x` and `y` the least squares estimates can be computed with the statements.

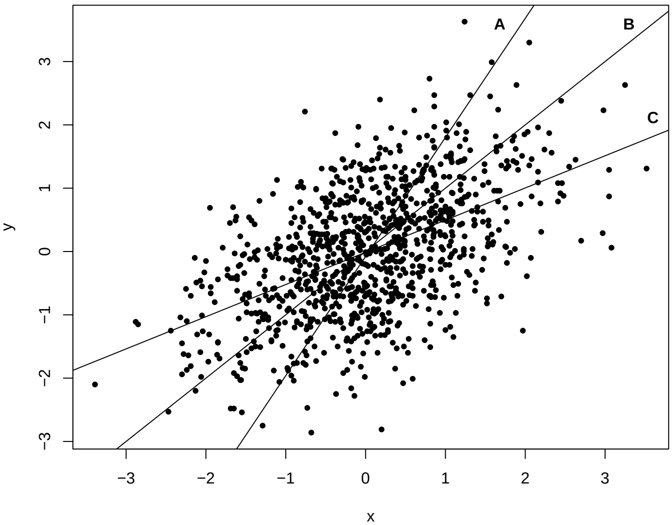
```
> res = lm(y ~ x)
> coef(res)
```

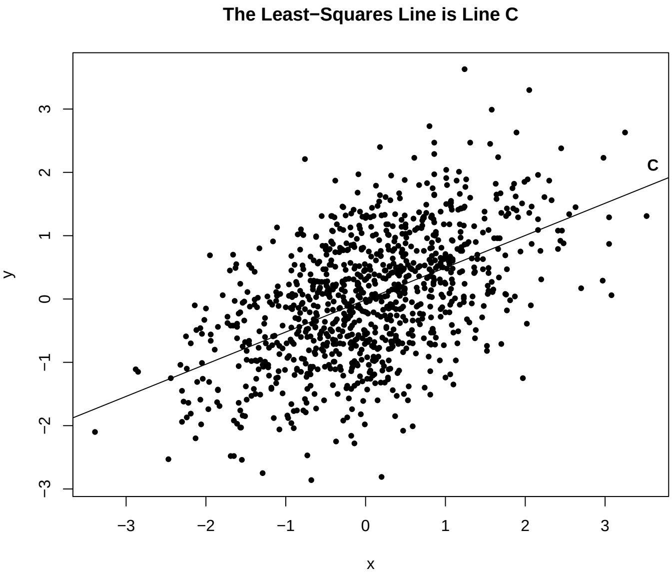
- `coef(res)` returns a vector with the intercept and slope as its first two elements.

The Position of the Least-Squares Line

- It is useful to gain some intuition about the location of the least-squares line in a plot of the points it is fitted to.
- We will do this by creating a set of random numbers and seeing where the least-squares line passes through them.

Three Lines Through a Set of Points



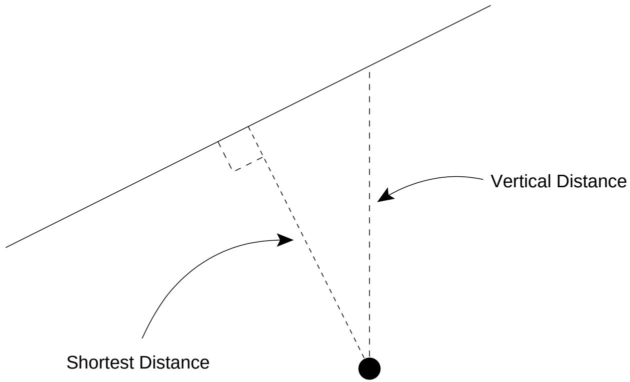


The Least−Squares Line is Line C
C
y
x
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

The Position of the Regression Line

- It is a common misconception that the least-squares line runs down the axis of symmetry of the cloud of points it is fitted to.
- Even quite experienced statisticians make this mistake.
- The slope of the least-squares line is less steep than the line down the axis of symmetry.
- The reason that the least-squares line is not the axis of symmetry is that it is based on vertical distances from the points to the line, rather than the shortest distances.

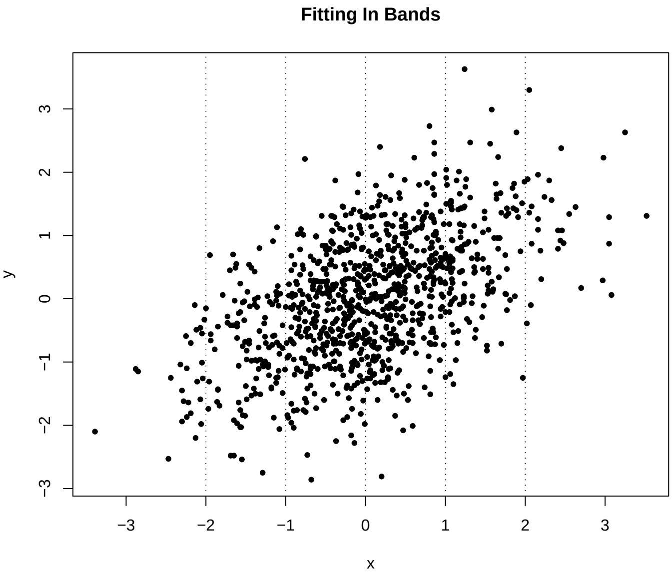
The Distances Considered In Least-Squares



The Position of the Least-Squares Line

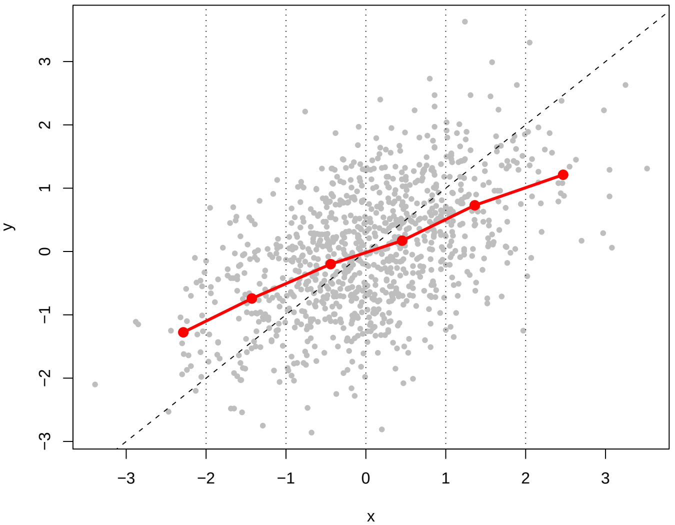
- We can show that the least-square line runs where it does by dividing the range of the x variable into small intervals and working separately within each interval.

- There is not much variability in y within each interval so we can estimate the position of the line by taking the the point defined by the means of the x and y values of the points in each interval.



Fitting In Bands
y
x
−3
−2
−1
0
1
2
3

Fiting In Bands



The Term "Regression"

- In early studies of population genetics, it was noticed that the (adult) daughters of the tallest women in a population were generally not quite as tall as their mothers, and the daughters of the shortest women were generally a little taller than their mothers.
- This phenomenon was observed for all kinds of population characteristics, and it was thought that there was some deep natural law at work.
- The phenomenon was called "regression toward the mean" and was studied using least-squares.
- Over time the two names became synonymous.

The Term “Regression”

- In fact there is nothing deep happening here.
- Provided that the relationship between mother’s height and daughter’s height is not a prefect straight line, the line which describes the average height of daughters as a function of mother’s height has a slope which is less than the line of equal mother-daughter height.
- This is a purely mathematical phenomenon and is completely unrelated to genetics.