Linear Regression: Preamble

The "best" guess about Y, if we know only the distribution of Y, is $E[Y]$.

If "best" is Mean Squared Error.

More precisely, the value of a that minimizes $E[(Y-a)^2]$ is $a = E[Y]$.

Proof:

Let $\hat{Y} := Y - E[Y]$.
Then, $E[\hat{Y}] = E[Y - E[Y]] = E[Y] - E[Y] = 0$.
So, $E[\hat{Y}c] = 0, \forall c$. Now,

$$\begin{aligned}
E[(Y-a)^2] &= E[(Y - E[Y] + E[Y] - a)^2] \\
&= E[(\hat{Y} + c)^2] \text{ with } c = E[Y] - a \\
&= E[\hat{Y}^2 + 2\hat{Y}c + c^2] = E[\hat{Y}^2] + 2E[\hat{Y}c] + c^2 \\
&= E[\hat{Y}^2] + 0 + c^2 \geq E[\hat{Y}^2].
\end{aligned}$$

Hence, $E[(Y-a)^2] \geq E[(Y - E[Y])^2], \forall a$. $\square$

Linear Regression: Preamble

Thus, if we want to guess the value of Y, we choose $E[Y]$.

Now assume we make some observation X related to Y.

How do we use that observation to improve our guess about Y?

The idea is to use a function $g(X)$ of the observation to estimate Y.

The simplest function $g(X)$ is a constant that does not depend of X.

The next simplest function is linear: $g(X) = a + bX$.

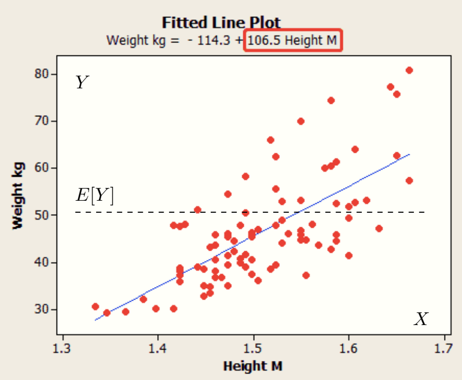
What is the best linear function? That is our next topic.

A bit later, we will consider a general function $g(X)$.

Linear Regression: Motivation

Example 1: 100 people.

Let (X_n, Y_n) = (height, weight) of person n, for $n = 1, \ldots, 100$:



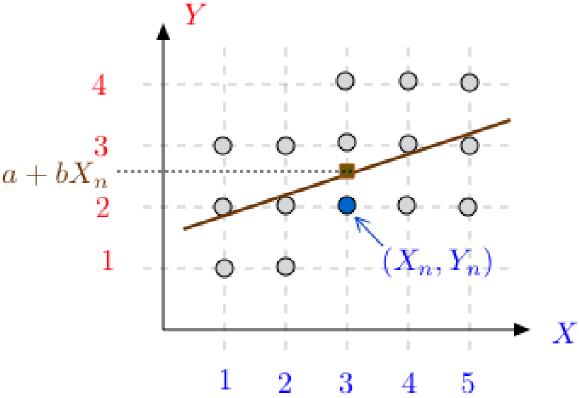
The blue line is $Y = -114.3 + 106.5X$. (X in meters, Y in kg.)

Best linear fit: Linear Regression.

Motivation

Example 2: 15 people.

We look at two attributes: (X_n, Y_n) of person n, for $n = 1, \ldots, 15$:



The line $Y = a + bX$ is the linear regression.

LLSE

$LLSE[Y|X]$ - best guess for Y given X.

Theorem

Consider two RVs X, Y with a given distribution $Pr[X = x, Y = y]$. Then,

$$L[Y|X] = \hat{Y} = E[Y] + \frac{cov(X,Y)}{var(X)}(X - E[X]).$$

Proof 1:

$Y - \hat{Y} = (Y - E[Y]) - \frac{cov(X,Y)}{var[X]}(X - E[X])$. $E[Y - \hat{Y}] = 0$ by linearity.

Also, $E[(Y - \hat{Y})X] = 0$, after a bit of algebra. (next slide)

Combine brown inequalities: $E[(Y - \hat{Y})(c + dX)] = 0$ for any c, d.
Since: $\hat{Y} = \alpha + \beta X$ for some α, β, so $\exists c, d$ s.t. $\hat{Y} - a - bX = c + dX$.
Then, $E[(Y - \hat{Y})(\hat{Y} - a - bX)] = 0, \forall a, b$. Now,

$$\begin{aligned} E[(Y - a - bX)^2] &= E[(Y - \hat{Y} + \hat{Y} - a - bX)^2] \\ &= E[(Y - \hat{Y})^2] + E[(\hat{Y} - a - bX)^2] + 0 \geq E[(Y - \hat{Y})^2]. \end{aligned}$$

This shows that $E[(Y - \hat{Y})^2] \leq E[(Y - a - bX)^2]$, for all (a, b).
Thus $\hat{Y}$ is the LLSE. $\square$

A Bit of Algebra

$$Y - \hat{Y} = (Y - E[Y]) - \frac{cov(X,Y)}{var[X]}(X - E[X]).$$

Hence, $E[Y - \hat{Y}] = 0$. We want to show that $E[(Y - \hat{Y})X] = 0$.

Note that

$$E[(Y - \hat{Y})X] = E[(Y - \hat{Y})(X - E[X])],$$

because $E[(Y - \hat{Y})E[X]] = 0$.

Now,

$$\begin{aligned} & E[(Y - \hat{Y})(X - E[X])] \\ & \quad = E[(Y - E[Y])(X - E[X])] - \frac{cov(X,Y)}{var[X]} E[(X - E[X])(X - E[X])] \\ & \quad =^{(*)} cov(X,Y) - \frac{cov(X,Y)}{var[X]} var[X] = 0. \quad \square \end{aligned}$$

$^{(*)}$ Recall that $cov(X,Y) = E[(X - E[X])(Y - E[Y])]$ and $var[X] = E[(X - E[X])^2]$.