Regression Analysis

Fall 2021

Department of Mathematics
Hong Kong Baptist University

Regression

The history of regression analysis started from a eugenics study by Francis Galton (1822-1911), a cousin of Charles Darwin. From Wikipedia on *regression toward the mean*, we know that:

- The concept of regression comes from genetics and was popularized by Sir Francis Galton during the late 19th century with the publication of *Regression towards mediocrity in hereditary stature*.
- Galton observed that extreme characteristics (e.g., height) in parents are not passed on completely to their offspring. Rather, the characteristics in the offspring *regress* towards a *mediocre* point (a point which has since been identified as the mean).
- By measuring the heights of hundreds of people, he was able to quantify regression to the mean, and estimate the size of the effect. Galton wrote that, "the average regression of the offspring is a constant fraction of their respective mid-parental deviations".
- For height, Galton estimated this correlation coefficient to be about $\frac{2}{3}$: the height of an individual will measure around a mid-point that is two thirds of the parents deviation from the population average.

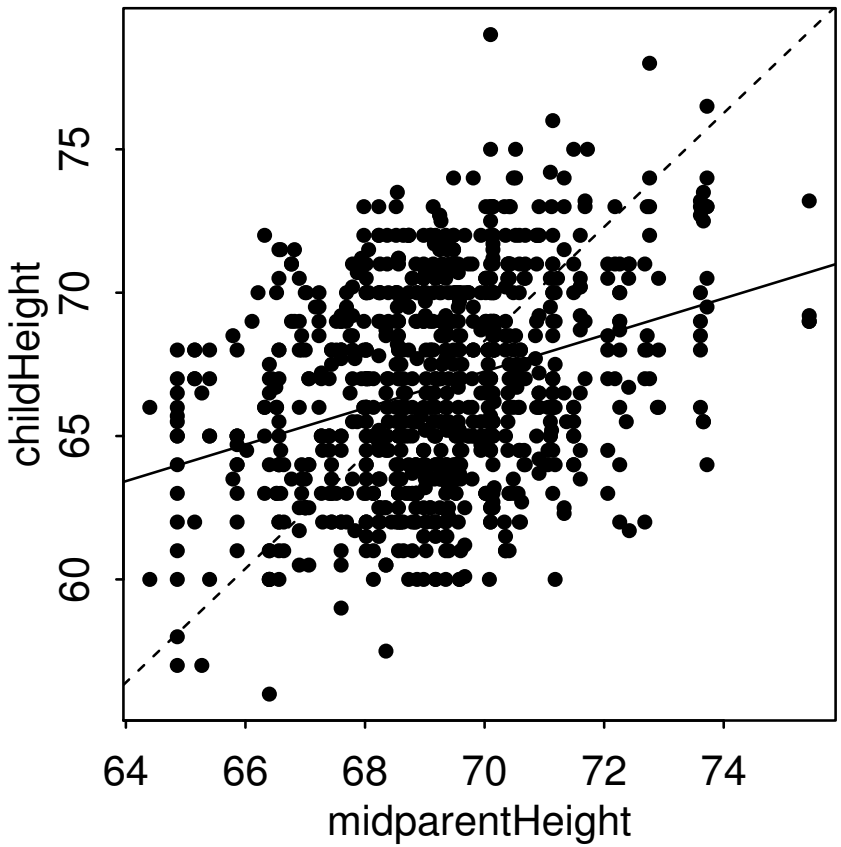

Figure: The height of child is plotted against a combined parental height defined as (father's height + 1.08 $\times$ mother's height)/2.

Data: $(x_1, Y_1), \ldots, (x_n, Y_n)$

Model: $Y = \mathbb{E}(Y|x) + \varepsilon$, $\mathbb{E}(Y|x) = f(x)$, i.e.

$$Y = f(x) + \varepsilon. \tag{1}$$

Y: Dependent variable, or response variable

x: Independent variable, explanatory variable, or covariate

$f(x)$: Regression function

ε: Unexplainable, or random, error

Data $(x_1, Y_1), \ldots, (x_n, Y_n)$ are independent observations on (x, Y), which follows model (1), and we have

$$Y_i = f(x_i) + \varepsilon_i,\ i = 1, \ldots, n.$$

More generally, we have data $(x_{1i}, x_{2i}, \ldots, x_{ki}, Y_i)$, $i = 1, \ldots, n$, observed from the model

$$Y = f(x_1, x_2, \ldots, x_k) + \varepsilon\,.$$

Simplest case:

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

called simple linear regression, in which there is only one x and the regression function is linear in the parameters β_0 and β_1.

Simple linear regression

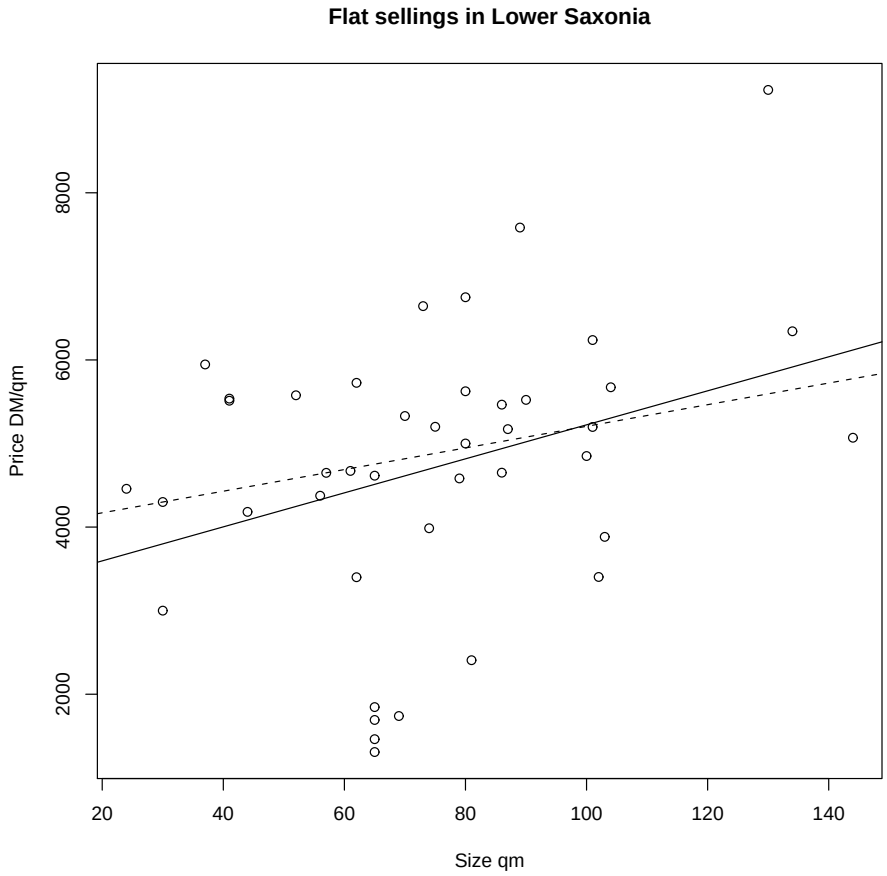

Figure: Lower Saxonian flat prices data with least squares (solid) and least absolute deviations (dashed) regression line.

Least Absolute Deviations Estimation (LAD)

Model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, \ldots, n.$

Choose the values of β_0 and β_1 to minimise $\mathcal{A}(\beta_0, \beta_1) = \sum_{i=1}^{n} |r_i|$, where the residuals r_i are given by $r_i = Y_i - \beta_0 - \beta_1 x_i$ for any given values of β_0 and β_1.

Hard to analyse mathematically.

Least Squares Estimation (LSE)

Model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \ldots, n$.
Choose the values of β_0 and β_1 to minimise $\mathcal{S}(\beta_0, \beta_1) = \sum_{i=1}^{n} r_i^2$, where the residuals r_i are given by $r_i = Y_i - \beta_0 - \beta_1 x_i$ for any given values of β_0 and β_1. Taking partial derivatives of $\mathcal{S}(\beta_0, \beta_1)$ w.r.t. β_0 and β_1 and setting to zero:

$$0 = \frac{\partial \mathcal{S}}{\partial \beta_0}\Big|_{\beta_0=\hat{\beta}_0, \beta_1=\hat{\beta}_1} = -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i),$$

$$0 = \frac{\partial \mathcal{S}}{\partial \beta_1}\Big|_{\beta_0=\hat{\beta}_0, \beta_1=\hat{\beta}_1} = -2\sum_{i=1}^{n} x_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i).$$

Then we obtain the normal equations:

$$\sum_{i=1}^{n} Y_i = \hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} x_i\,, \qquad \sum_{i=1}^{n} x_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2.$$

Solving the normal equations yields

$$\hat{\beta}_1 = \frac{C_{xY}}{C_{xx}}, \ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}, \tag{2}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, C_{xY} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}), C_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Least squares regression line:

$$y = \hat{\beta}_0 + \hat{\beta}_1 \, x.$$

Estimator for Y_i:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \, x_i \, , i = 1, \ldots, n.$$

Remark. The method of least squares was discovered independently by Garl Friedrich Gauss (1777-1855) and Adrien Marie Legendre (1752-1833). See Section 1.8 of Draper and Smith (1998).

Let $C_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$. Then the least squares regression line can be written as

$$y = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{Y} + \frac{C_{xY}}{\sqrt{C_{xx}}\sqrt{C_{YY}}} \sqrt{C_{YY}} \frac{x - \bar{x}}{\sqrt{C_{xx}}}$$

which can be rewritten as

$$\frac{y - \bar{Y}}{\sqrt{S_Y^2}} = \frac{S_{XY}}{\sqrt{S_Y^2 S_x^2}} \frac{x - \bar{x}}{\sqrt{S_x^2}},$$

where S_{xY} is the sample covariance of x_i and Y_i, and S_x^2 and S_Y^2 are the sample variances of x_i and Y_i respectively. Estimator for response Y given x, or prediction equation:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x \,.$$

The interpretation of the slope estimate $\hat{\beta}_1$ is as follows: There is a $\hat{\beta}_1$-unit increase in the mean of Y for every 1-unit increase in x.

Note. Francis Galton's finding on parent–child heights mentioned earlier (taken from Wikipedia), in terms of the notation used above, is

$$\frac{\hat{Y} - \bar{Y}}{\text{sd}(Y)} = \frac{2}{3}\frac{(x - \bar{x})}{\text{sd}(x)},$$

where $x =$ weighted average of mother's and father's heights, $Y =$ child's height, and $\text{sd}(x)$ and $\text{sd}(Y)$ are the sample standard deviations of x_i's and Y_i's, respectively.

Properties of LSE

The residuals obtained from the least squares line are

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\,, \;\; i = 1, \ldots, n.$$

Properties:

(i) $\sum_{i=1}^n e_i = 0$,
(ii) $\sum_{i=1}^n x_i e_i = 0$,
(iii) $\sum_{i=1}^n \hat{Y}_i e_i = 0$.

Properties (i) and (ii) follow from the normal equations:

$$0 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^{n} e_i\,,$$

$$0 = \sum_{i=1}^{n} x_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^{n} x_i e_i\,.$$

Property (iii) follows from properties (i) and (ii):
$\sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i) e_i = \hat{\beta}_0 \sum_{i=1}^n e_i + \hat{\beta}_1 \sum_{i=1}^n x_i e_i = 0.$

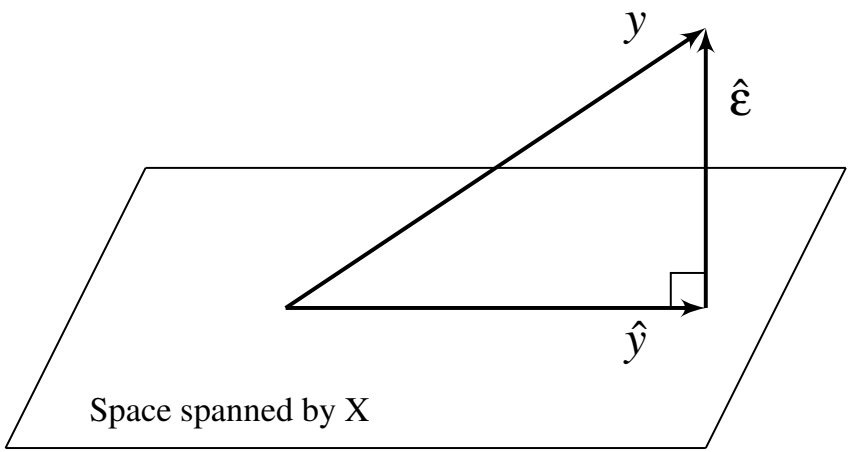

Figure: Geometrical representation of the least squares estimation. The data vector $\boldsymbol{Y} = (\hat{Y}_1, \ldots, \hat{Y}_n)'$ is projected orthogonally onto the model space spanned by $\boldsymbol{X}$, whose columns are $(1, ..., 1)'$ and $(x_1, \ldots, x_n)'$. The fit is represented by projection $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)'$ with the difference between the fit and the data represented by the residual vector $\boldsymbol{e} = (e_1, \ldots, e_n)'$.

Properties (i) and (ii) mean that the vector $(e_1, \ldots, e_n)$ is orthogonal to the vectors $(1, ..., 1)$ and $(x_1, \ldots, x_n)$. This implies that nothing in $(e_1, \ldots, e_n)$ can be affected by $(1, ..., 1)$ and $(x_1, \ldots, x_n)$. If the orthogonality does not hold, then there are still some variation in $(Y_1, \ldots, Y_n)$ that can be explained by $(1, ..., 1)$ and $(x_1, \ldots, x_n)$ meaning that we had not yet squeezed out all information on $(Y_1, \ldots, Y_n)$ provided by $(1, ..., 1)$ and $(x_1, \ldots, x_n)$.

Property (iii) means the vectors $(\hat{Y}_1, \ldots, \hat{Y}_n)$ and $(e_1, \ldots, e_n)$ are orthogonal. This, along with the fact that $(\hat{Y}_1, \ldots, \hat{Y}_n)$ lies in the space generated by $(1, ..., 1)$ and $(x_1, \ldots, x_n)$, says $(\hat{Y}_1, \ldots, \hat{Y}_n)$ is the orthogonal projection of $(Y_1, \ldots, Y_n)$ onto the space generated by $(1, ..., 1)$ and $(x_1, \ldots, x_n)$, and so among all feasible vectors in the space generated by $(1, ..., 1)$ and $(x_1, \ldots, x_n)$, $(\hat{Y}_1, \ldots, \hat{Y}_n)$ is the closest (measured in the Euclidean distance) to the observed vector $(Y_1, \ldots, Y_n)$.

LSE in Matrix Form

The simple linear regression model can be rewritten as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \text{ and } \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Denote the transpose of a matrix $\boldsymbol{A}$ by $\boldsymbol{A}'$. Write $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)'$, $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \ldots \hat{Y}_n)'$, $\boldsymbol{e} = (e_1, \ldots, e_n)'$. Then we have

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \\ \hat{\boldsymbol{Y}} &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \equiv \boldsymbol{H}\boldsymbol{Y}, \\ \boldsymbol{e} &= \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}, \end{aligned}$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix, and the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is a projection matrix onto the linear space spanned by the columns of the design matrix $\boldsymbol{X}$ (i.e. $\boldsymbol{H}$ is symmetric, $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$, $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$, $\boldsymbol{H}(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{0}$, where $\boldsymbol{0}$ denotes the $n \times n$ zero matrix).

Note that $\boldsymbol{e}'\boldsymbol{X} = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{Y}'(\boldsymbol{X} - \boldsymbol{X}) = \boldsymbol{0}$, where $\boldsymbol{0}$ denotes the 2×1 zero matrix.

Note also that $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y} = \boldsymbol{X}\hat{\beta}$ lies in the linear subspace spanned by the columns of $\boldsymbol{X}$, and $\hat{\boldsymbol{Y}}$ and $\boldsymbol{e}$ are orthogonal:

$$\hat{\boldsymbol{Y}} \cdot \boldsymbol{e} = \hat{\boldsymbol{Y}}'\boldsymbol{e} = 0\,,$$

where $\cdot$ denotes dot product.

Simple linear regression with normal errors

Model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \ , \ i = 1, \ldots, n,$$

where x_i are fixed, ε_i are i.i.d. with $\mathcal{N}(0, \sigma^2)$ distribution, and β_0, β_1 and σ^2 are unknown constant parameters.
Equivalently:

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), \ i = 1, \ldots, n, \ \text{independent.}$$

Y: Dependent variable, response variable
x: Independent variable, explanatory variable, covariate
ε: Unexplainable, or random, error
Deterministic: $\beta_0 + \beta_1 x_i$
Random: ε_i
Observed: $(x_1, Y_1), \cdots, (x_n, Y_n)$
Unobserved: $\varepsilon_1, \cdots, \varepsilon_n$

Maximum Likelihood Estimation

The likelihood function given $(x_1, Y_1), \ldots, (x_n, Y_n)$ is

$$\ell(\beta_0, \beta_1, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2}.$$

Log likelihood function is

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \log \ell(\beta_0, \beta_1, \sigma^2) \\ &= -n \log \sqrt{2\pi} - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2 . \end{aligned}$$

For any fixed σ^2, maximising $L(\beta_0, \beta_1, \sigma^2)$ is equivalent to minimising the sum of squares $\mathcal{S} = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2$. Hence, the MLE of β_0 and β_1 is the same as the LSE.

Maximum likelihood estimator for σ^2 is obtained by maximising $L(\beta_0, \beta_1, \sigma^2)$ w.r.t. σ^2, by keeping β_0 and β_1 fixed at $\hat{\beta}_0$ and $\hat{\beta}_1$. It is

$$\frac{SSE}{n},$$

where $SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$, which is biased.

An unbiased estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{SSE}{n-2},$$

in which the denominator is the value of the degrees of freedom; we lose two degrees of freedom because we estimate two parameters in order to get $\hat{Y}_i$.

Different parametrisations

One aspect of the linear model that can cause some difficulty at first is the fact that there is often more than one way to write the same model.
For example, an alternative form for simple linear regression is

$$Y_i = \gamma_0 + \gamma_1(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\bar{x} = \frac{1}{n}\sum x_i$ is the mean of the x_i. By equating the systematic parts of the two models, we have

$$\gamma_0 + \gamma_1(x_i - \bar{x}) = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n,$$

if and only if $\gamma_1 = \beta_1$ and $\gamma_0 - \gamma_1\bar{x} = \beta_0$ (i.e. $\gamma_1 = \beta_1$ and $\gamma_0 = \beta_0 + \beta_1\bar{x}$). The slope parameters are the same in the two models, but the intercepts differ. One, β_0, is the intercept at $x = 0$, the other, γ_0, is the intercept at $x = \bar{x}$. If you fit either model by least squares you will get the same straight line, but described by different parameters.

Another parametrisation (by centering both Y_i and x_i) is

$$Y_i - \bar{Y} = \eta_0 + \eta_1(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, \ldots, n.$$

Then, by equating the systematic parts, we have

$$\bar{Y} + \eta_0 + \eta_1(x_i - \bar{x}) = \beta_0 + \beta_1 x_i,$$

and thus

$$\eta_1 = \beta_1 \quad \text{and} \quad \bar{Y} + \eta_0 - \eta_1 \bar{x} = \beta_0,$$

i.e.

$$\eta_1 = \beta_1 \quad \text{and} \quad \eta_0 = \beta_0 + \beta_1 \bar{x} - \bar{Y}\,.$$