Regression

A few fine points

Linear ***Relationship (Function)*** vs. Linear ***Regression***

In **linear algebra, a linear function** is a map f between two vector spaces that preserves vector addition and scalar multiplication:

$$f\ (x+y) = f\ (x) + f(y)$$
$$f\ (a * x) = a\ *\ f\ (x)$$

From Wikipedia

In **calculus, analytic geometry and related areas, a linear function** is a polynomial of degree one or less, including the zero polynomial (the latter not being considered to have degree zero):

$$y = a_0 + a_1 * x_1 + a_2 * x_2 + \cdots + a_n * x_n$$

From Wikipedia

A **linear equation** is an algebraic equation in which each term is either a constant or the product of a constant and (the first power of) a single variable:

$$y = m * x + b$$

From Wikipedia

To get parameters of linear equation, we use linear algebra …

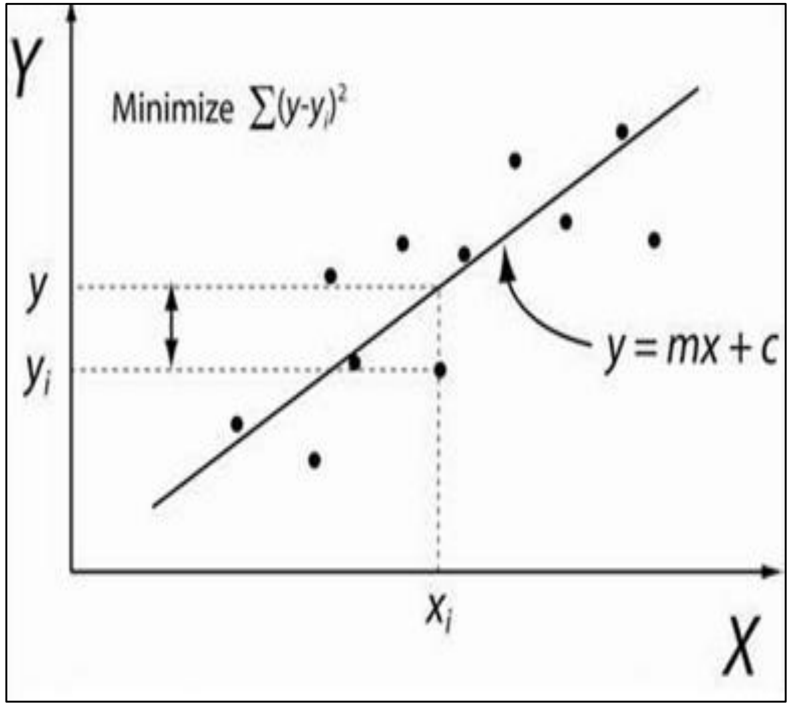
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

… which relies on a very different definition of linear function

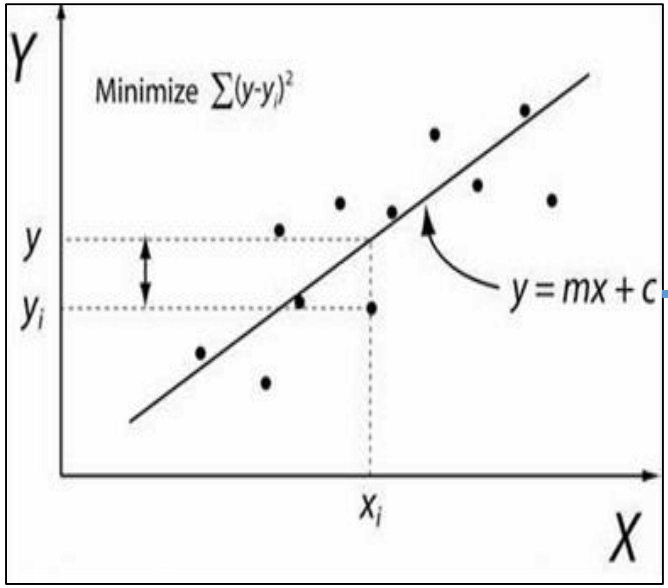
Why does it work ?



TRAIN:
$y(\boldsymbol{a} \mid x_{train}) = \boldsymbol{a_0} + \boldsymbol{a_1} * x_1 + \boldsymbol{a_2} * x_2 + \cdots + \boldsymbol{a_n} * x_n$

PREDICT:
$y(\boldsymbol{x_{test}} \mid a) = a_0 + a_1 * \boldsymbol{x_1} + a_2 * \boldsymbol{x_2} + \cdots + a_n * \boldsymbol{x_n}$

Main Idea



TRAIN:

$$y(\boldsymbol{a} \mid x_{train}) = \boldsymbol{a_0} + \boldsymbol{a_1} * x_1 + \boldsymbol{a_2} * x_2 + \cdots + \boldsymbol{a_n} * x_n$$

PREDICT:

$$Y_{predicted} = y(\boldsymbol{x_{test}} \mid a) = a_0 + a_1 * \boldsymbol{x_1} + a_2 * \boldsymbol{x_2} + \cdots + a_n * \boldsymbol{x_n}$$

$$R^2 \equiv 1 - \frac{Var(Y_{predicted} - Y_{observed})}{Var(Y_{observed})}$$

R^2 *is an* ***estimate***

- It has a stochastic component
- It is **not** normally distributed
- **=> cannot use T-test**

Compare Model Prediction with Test data labels

Is the model adequate ?

Fisher's Z-transformation:

$$z = \frac{1}{2} * \ln\left(\frac{1+r}{1-r}\right)$$

$$\sigma = \frac{1}{\sqrt{N-3}}$$

$r = \sqrt{R^2}$ - Pearson's correlation between model predictions and observations

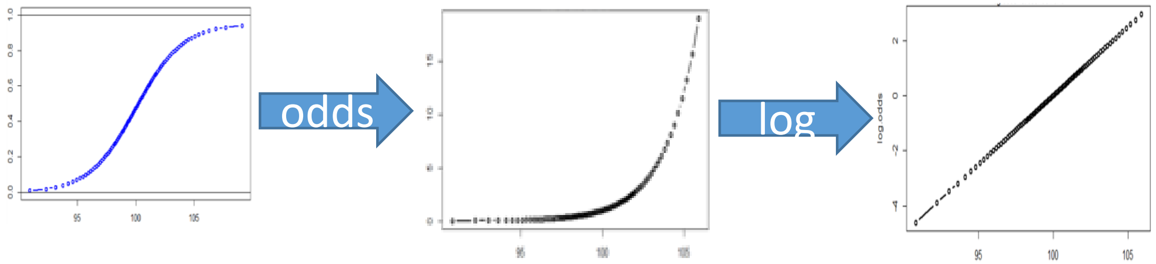
$$R_1^2 = 0.912$$
$$R_2^2 = 0.927$$

Is Model 2 better than Model 1?

Now we **can** use T-test

"Big 7" of Linearizable Equations

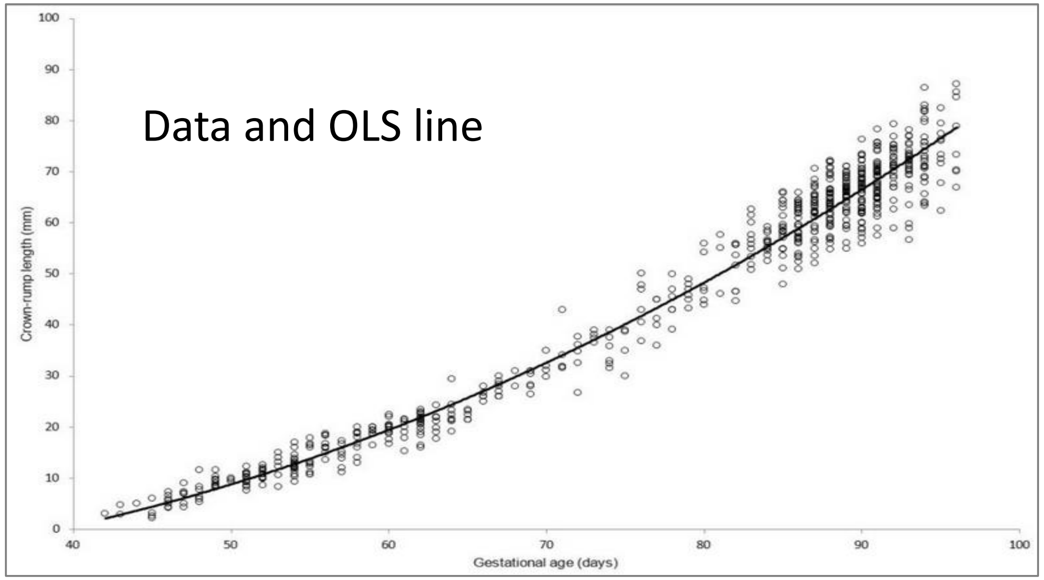
1. Straight-line relationship: $y = a_0 + a_1 * x$
2. Quadratic relationship: $y = a_0 + a_1 * x + a_2 * x^2$
3. Exponential relationship: $y = a_0 * e^{a_1 * x}$
4. Logarithmic relationship: $y = a_0 + a_1 * \ln(x)$
5. Power relationship: $y = a_0 * x^{a_1}$
6. Hyperbolic relationship: $y = \frac{a_0}{a_1 + a_2 * x}$
7. Sigmoid relationship: $P\{y\} = a_0 + a_1 * x$

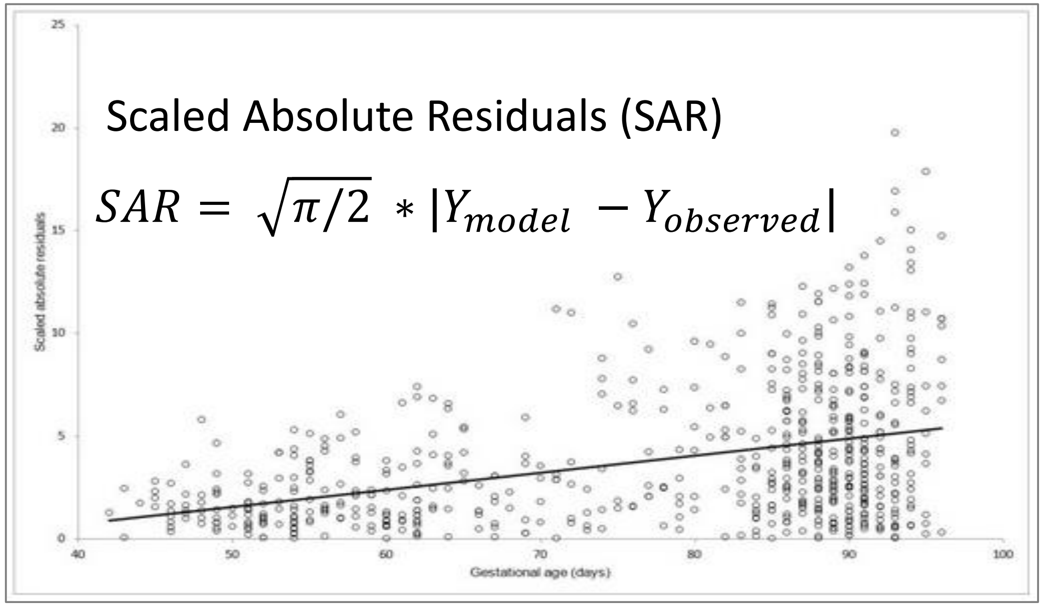


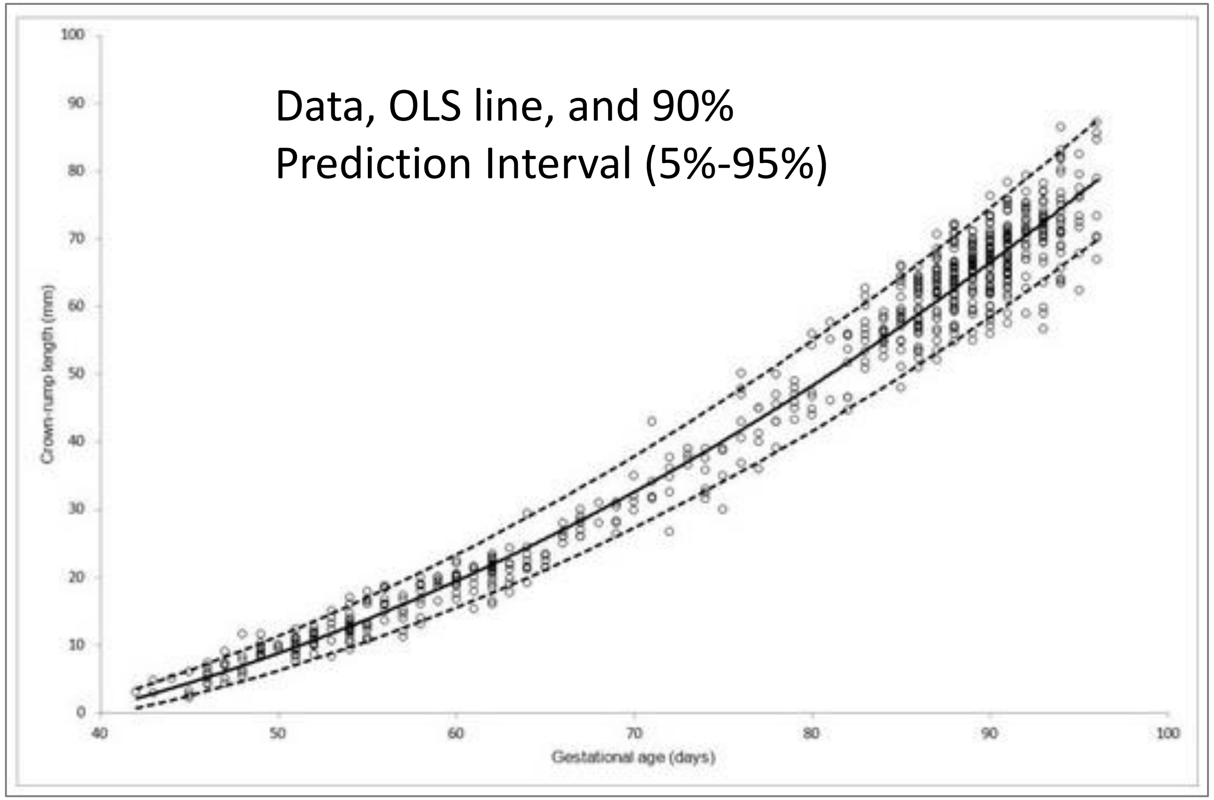
1. $\mathrm{y = a_0 + a_1 * x}$ - no action needed
2. $\mathrm{y = a_0 + a_1 * x + a_2 * x^2}$:
   - $\mathrm{x_1 = x; x_2 = x^2} \Rightarrow$
   - $\mathbf{y = a_0 + a_1 * x_1 + a_2 * x_2}$
3. Exponential relationship: $\mathrm{y = a_0 * e^{a_1 * x}}$:
   - $\mathrm{y' = \log(y)} \Rightarrow$
   - $\mathbf{y' = \log(a_0) + a_1 * x = a_0' + a_1 * x}$
4. Logarithmic relationship: $\mathrm{y = a_0 + a_1 * \ln(x)}$:
   - $\mathrm{y' = \exp(y)} \Rightarrow$
   - $\mathbf{y' = e^{a_0 + a_1 * \ln(x)} = a_0' + a_1' * x}$
5. Power relationship: $\mathrm{y = a_0 * x^{a_1}}$:
   - $\mathrm{y' = \log(y); x' = \log(x)} \Rightarrow$
   - $\mathbf{y_1 = \log(a_0) + a_1 * x' = a_0' + a_1 * x'}$
6. Hyperbolic relationship: $\mathrm{y = \frac{a_0}{a_1 + a_2 * x}}$:
   - $\mathrm{b_1 = \frac{a_1}{a_0}; b_2 = \frac{a_2}{a_0}; y' = 1/y} \Rightarrow$
   - $\mathbf{y' = b_1 + b_2 * x'}$
7. Sigmoid relationship: $\mathrm{P\{Evt\} = a_0 + a_1 * x}$:
   - $Odds\{Evt\} = \frac{P\{\mathrm{Evt}\}}{(1 - P\{\mathrm{Evt}\})}$;
   - $y' = \log(Odds) = logit\ (P\{Evt\})$
   - $\boldsymbol{y' = a_0 + a_1 * x}$

A case for Quantile Regression

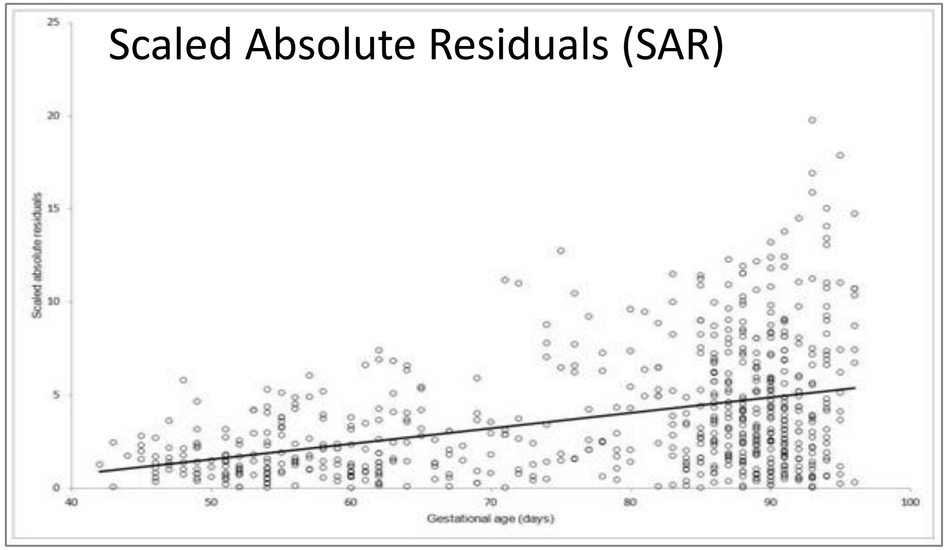
http://www.ejbi.org/en/ejbi/article/37-en-approaches-for-constructing-age-related-reference-intervals-and-centile-charts-for-fetal-size-.html







The main idea of Quantile Regression



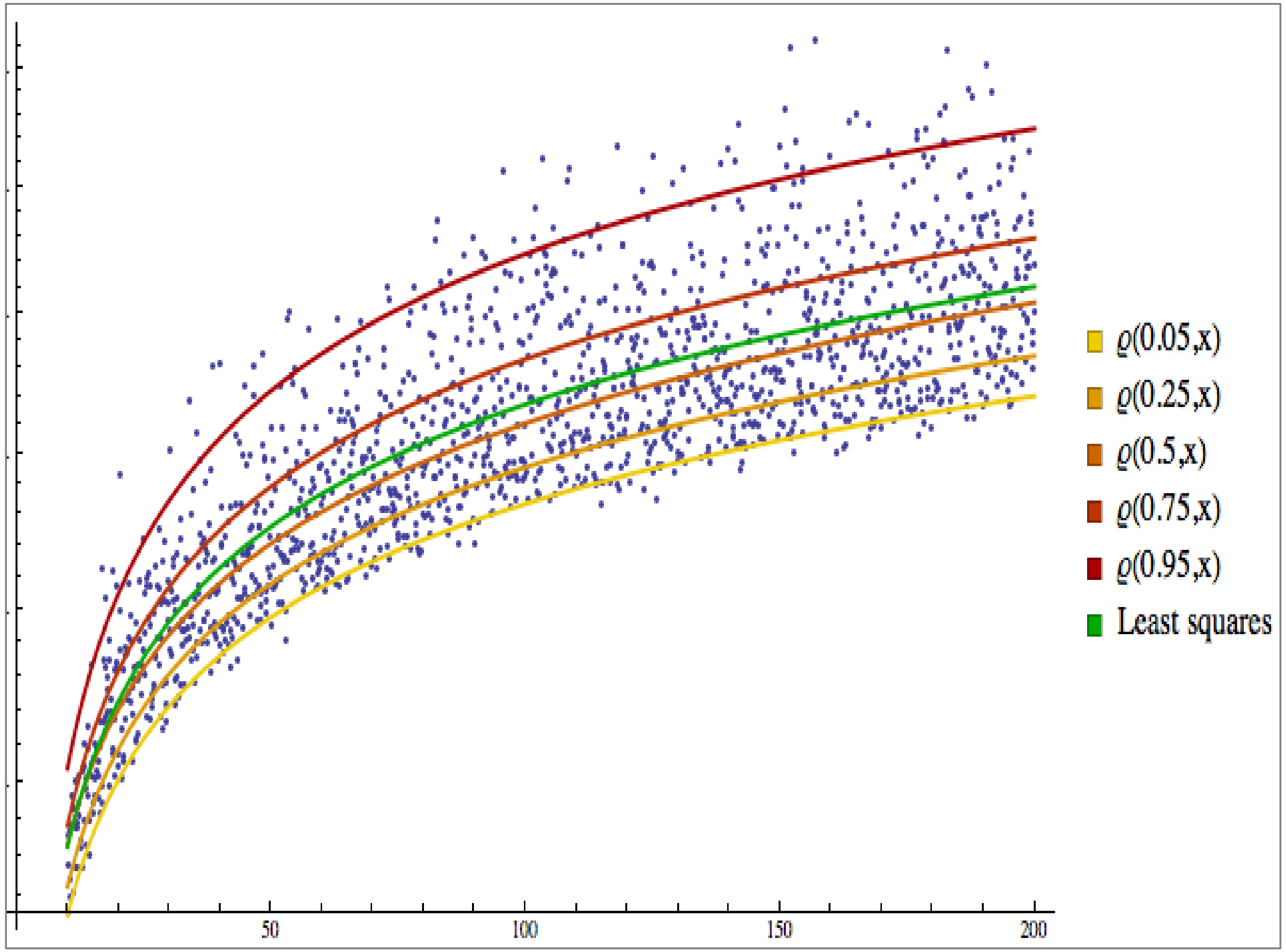
And a little bit more math

$$y(\boldsymbol{a}\,|\boldsymbol{p}, x_{train}) = \boldsymbol{a_0} + \boldsymbol{a_1} * x_1 + \boldsymbol{a_2} * x_2 + \cdots + \boldsymbol{a_n} * x_n$$

$$\varepsilon_p[i] = y_{p,model}[i] \ - y_{p,observed}[i]$$

$$a = argmin\left\{\boldsymbol{E}\left[|\varepsilon_p|\right]\right\}$$

$$\boldsymbol{E}\left[|\varepsilon_p|\right] = \sum_{i=1}^{N} p * \left|\varepsilon_p[i]\,|(\varepsilon_p[i] > 0)]\right| + \sum_{i=1}^{N} (1-p) * \left|[\varepsilon_p[i]|(\varepsilon_p[i] < 0)\right|]$$



Advantages:

Accurate representation of each quantile.

In-depth understanding of underlying behavior.

Disadvantages:

No closed-form solution.

To close the topic of Regression…