Linear Regression – Matrix Form

Consider the model

$$Y = X\beta + \epsilon$$

where $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$ $X = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix}$ $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$

Then using matrix calculus we find that the least squares estimate for β is given by

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Hence, the least squares regression line is $\hat{Y} = X\hat{\beta}$.

Influence Matrix

Linear Regression – Matrix Form - Issues

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Hence, the least squares regression line is $\hat{Y} = X\hat{\beta}$.

- N samples, p-dimensional (what if p > N ?)
- Complexity of matrix inversion (what if N very large ?)
- Collinearity

Gradient Descent

$$\hat{y} = \beta_0 + \beta_1 \mathbf{x}$$

$$\mathcal{L}(\mathbf{w}) = \mathcal{L}(\beta_0, \beta_1)$$

$$= \frac{1}{n} \sum_{i=1}^{n} (\hat{y} - y)^2$$

1. Initialize the parameters w to some random values.
2. Update the parameters using gradient descent rule

   $$\beta(t+1) = \beta(t) - \eta \nabla_{\beta} L\big(\beta(t)\big)$$

3. Repeat 2 until $|\nabla_{\beta} L\big(\beta(t)\big)|$ is close to 0

Gradient Descent

1. Initialize the parameters w to some random values.
2. Update the parameters using gradient descent rule

   $$\beta(t+1) = \beta(t) - \eta \nabla_\beta L\big(\beta(t)\big)$$

3. Repeat 2 until $|\nabla_\beta L\big(\beta(t)\big)|$ is close to 0

$$\hat{y} = \beta_0 + \beta_1 \mathbf{x}$$

$$\begin{aligned}
\mathcal{L}(\beta_0, \beta_1) & \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1\mathbf{x} - y)^2 \\
\implies \frac{\partial \mathcal{L}}{\partial \beta_0} &= \frac{1}{n}\sum_{i=1}^{n} 2(\beta_0 + \beta_1\mathbf{x} - y) \\
\implies \frac{\partial \mathcal{L}}{\partial \beta_1} &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}(\beta_0 + \beta_1\mathbf{x} - y)
\end{aligned}$$

Gradient Descent

1. Initialize the parameters w to some random values.
2. Update the parameters using gradient descent rule

$$w(t+1) = w(t) - \eta \nabla_w L(w(t))$$

3. Repeat 2 until $|\nabla_w L(w(t))|$ is close to 0

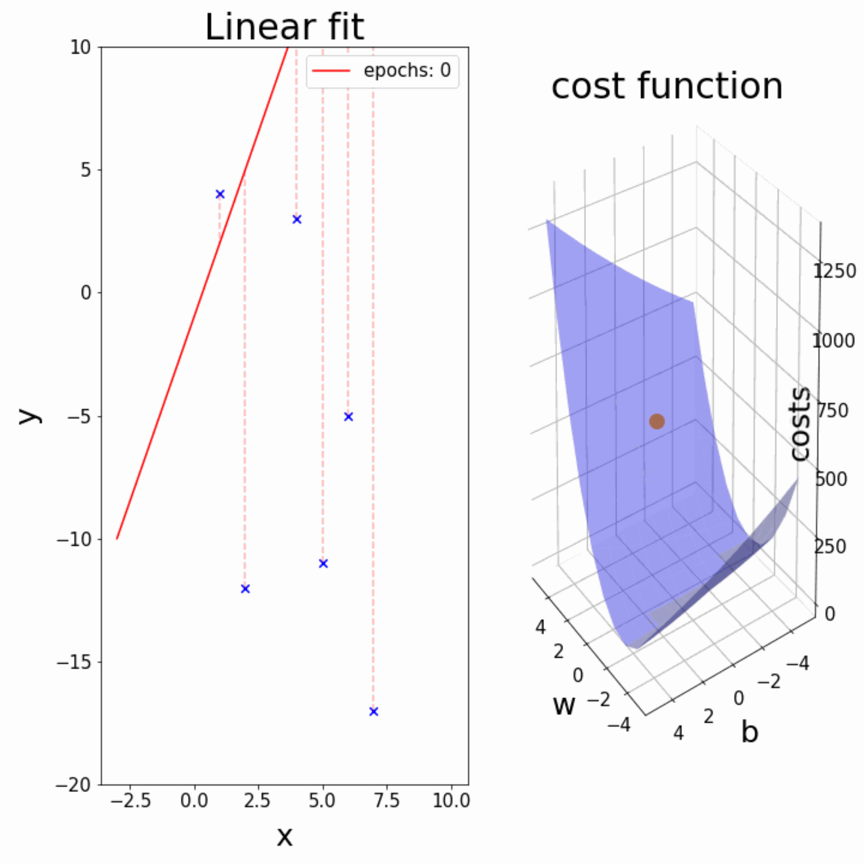
$$\mathbf{w} = [\beta_0 \ \ \beta_1]$$

$$\hat{y} = \beta_0 + \beta_1 \mathbf{x}$$

$$\begin{aligned}\mathcal{L}(\mathbf{w}) &= \mathcal{L}(\beta_0, \beta_1) \\ &= \frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1\mathbf{x} - y)^2\end{aligned}$$

$$\Longrightarrow \frac{\partial \mathcal{L}}{\partial \beta_0} = \frac{1}{n}\sum_{i=1}^{n} 2(\beta_0 + \beta_1\mathbf{x} - y)$$

$$\Longrightarrow \frac{\partial \mathcal{L}}{\partial \beta_1} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}(\beta_0 + \beta_1\mathbf{x} - y)$$



Gradient Descent

1. Initialize the parameters w to some random values.
2. Update the parameters using gradient descent rule

$$w(t+1) = w(t) - \eta \, \nabla_w L(w(t))$$

3. Repeat 2 until $|\nabla_w L(w(t))|$ is close to 0

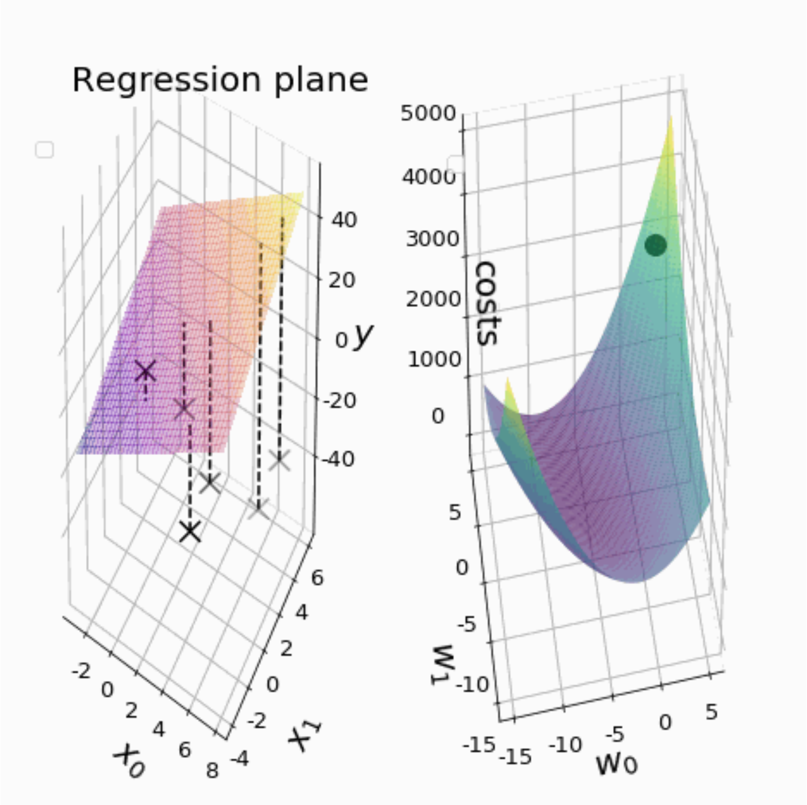
$$\mathbf{w} = [\beta_0 \;\; \beta_1]$$

$$\hat{y} = \beta_0 + \beta_1 \mathbf{x}$$

$$\begin{aligned}
\mathcal{L}(\mathbf{w}) &= \mathcal{L}(\beta_0, \beta_1) \\
&= \frac{1}{n} \sum_{i=1}^{n} (\hat{y} - y)^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 \mathbf{x} - y)^2
\end{aligned}$$

$$\implies \frac{\partial \mathcal{L}}{\partial \beta_0} = \frac{1}{n} \sum_{i=1}^{n} 2(\beta_0 + \beta_1 \mathbf{x} - y)$$

$$\implies \frac{\partial \mathcal{L}}{\partial \beta_1} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}(\beta_0 + \beta_1 \mathbf{x} - y)$$



Linear Regression

- Linear Regression → Linear in coefficients and NOT variables

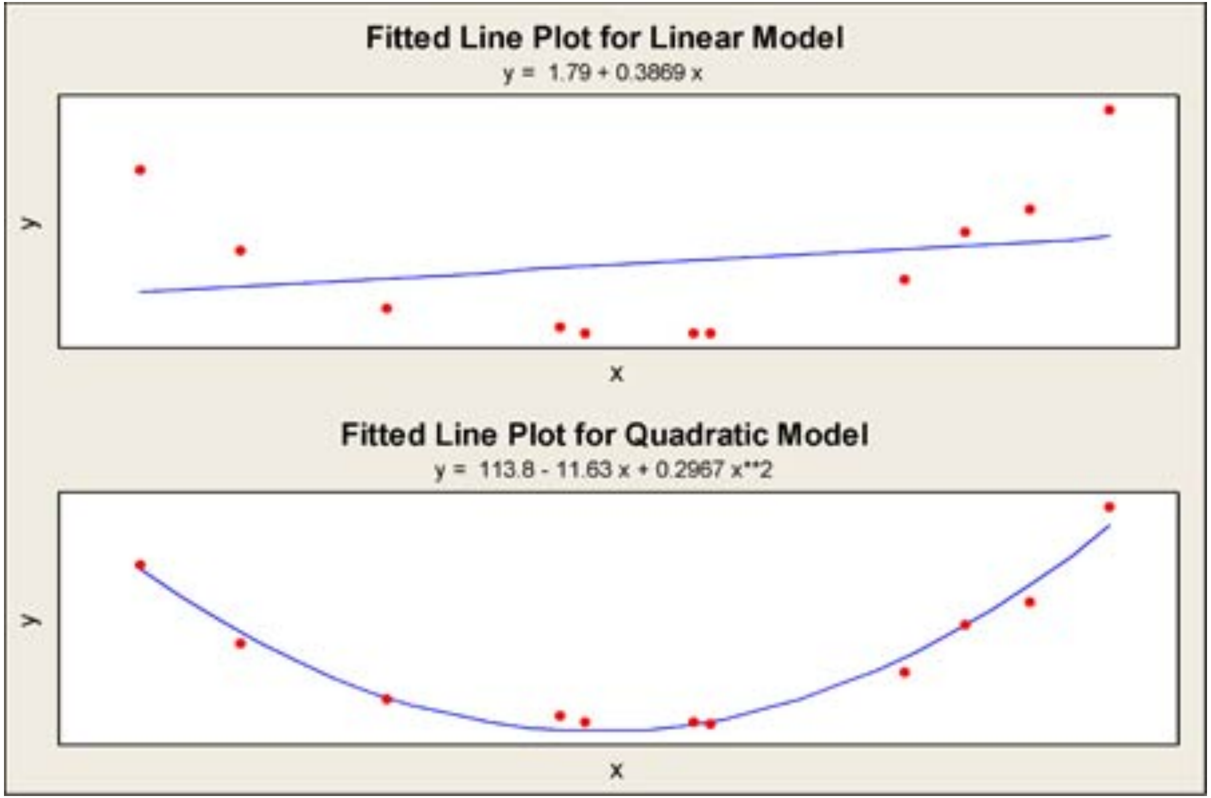
- A second-order model (quadratic model):

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

- β_1: Linear effect parameter.
- β_2: Quadratic effect parameter.

kth order polynomial model in one variable

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon$$



Fitted Line Plot for Linear Model
y = 1.79 + 0.3869 x
y
x
Fitted Line Plot for Quadratic Model
y = 113.8 - 11.63 x + 0.2967 x**2
y
x

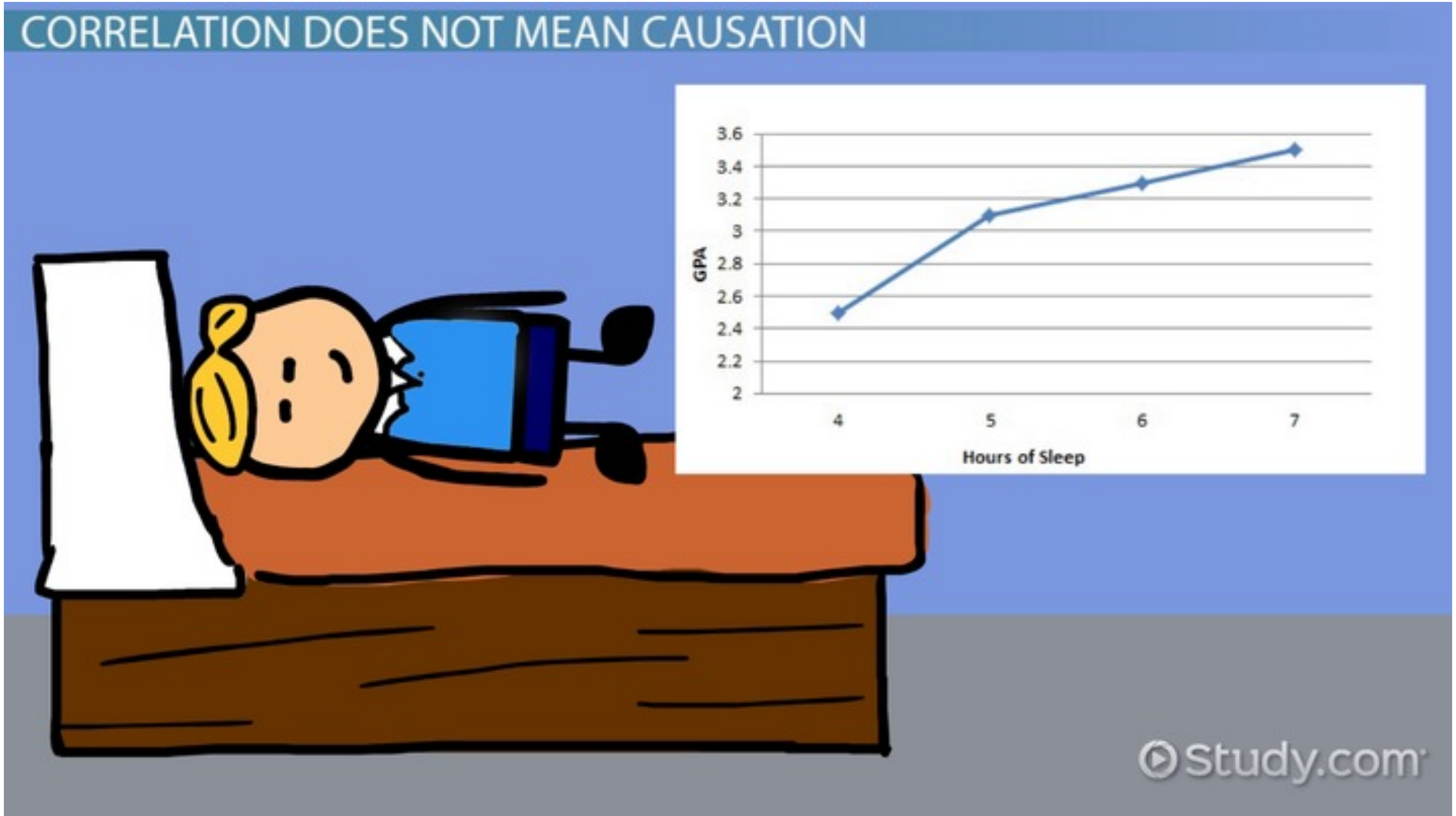
A quadratic polynomial regression function

$$Y_i = \beta_0 + \beta_1 X_i + \beta_{11} X_i^2 + \varepsilon_i$$

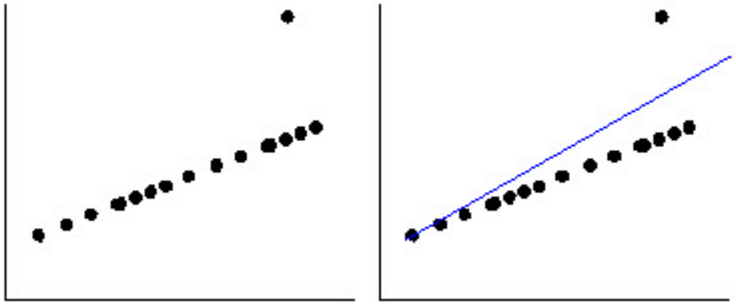
where:

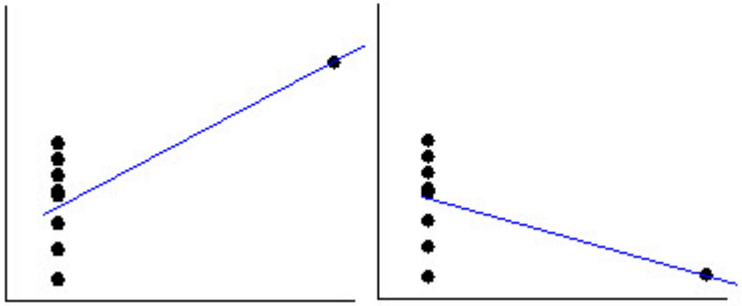
- Y_i = amount of immunoglobin in blood (mg)
- X_i = maximal oxygen uptake (ml/kg)
- typical assumptions about error terms ("INE")

Careful: X may not be causing y !

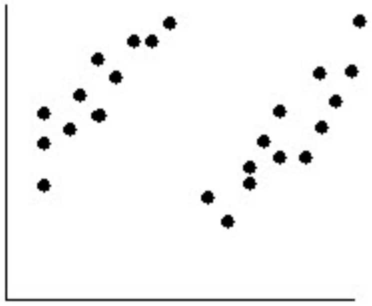


Linear Regression – Outliers

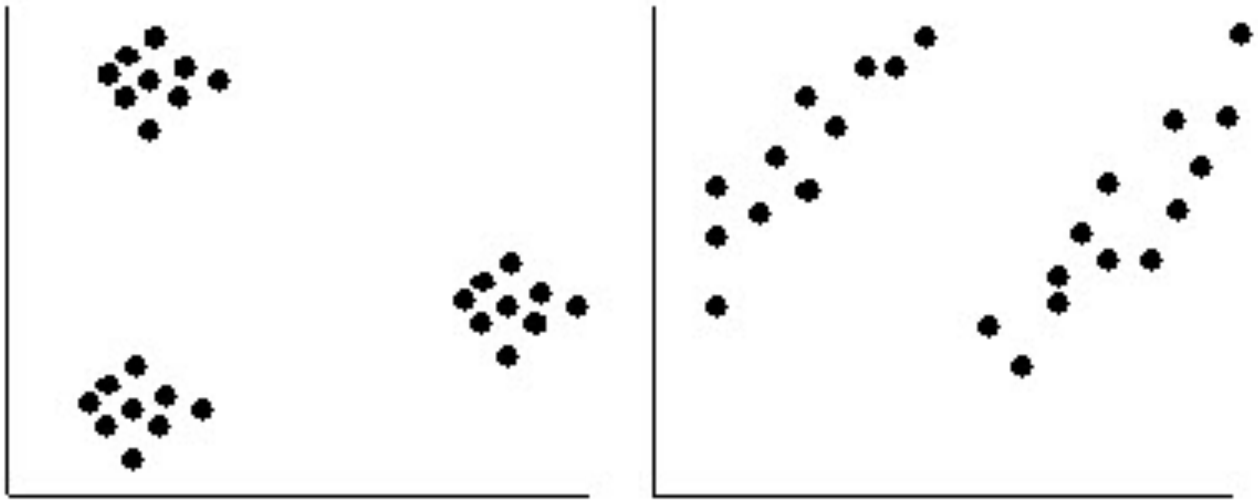




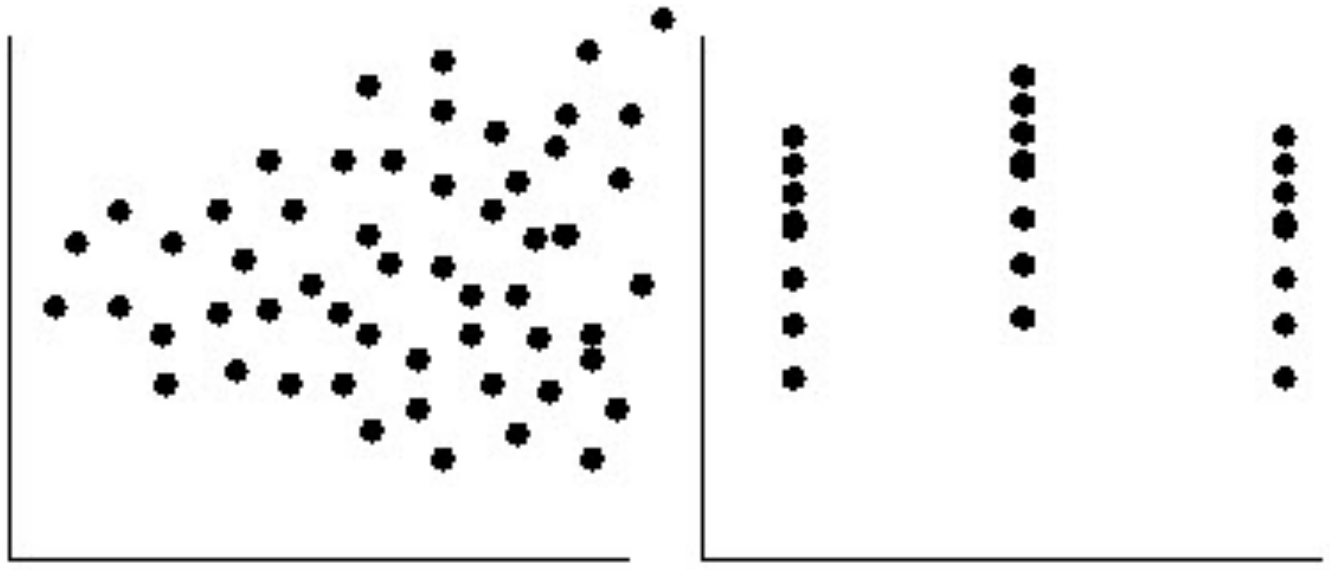
Linear Regression is problematic in many other cases



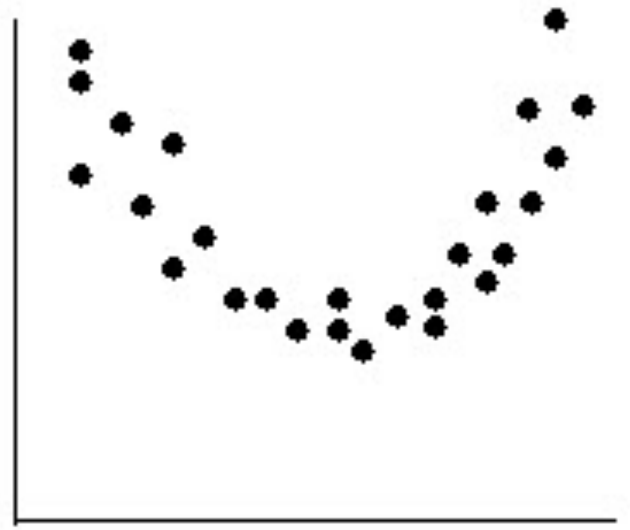
Linear Regression is problematic in many other cases

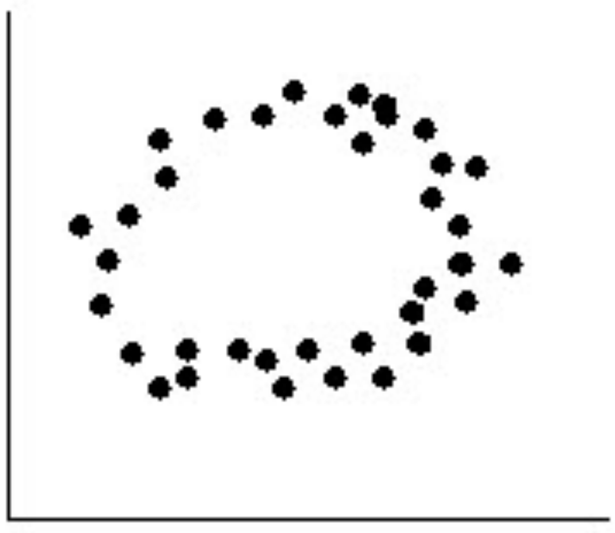


Linear Regression is problematic in many other cases

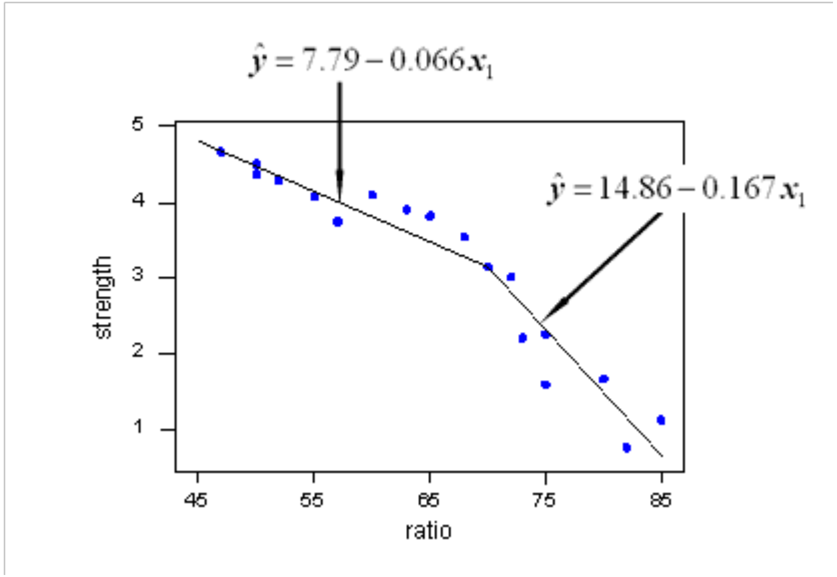


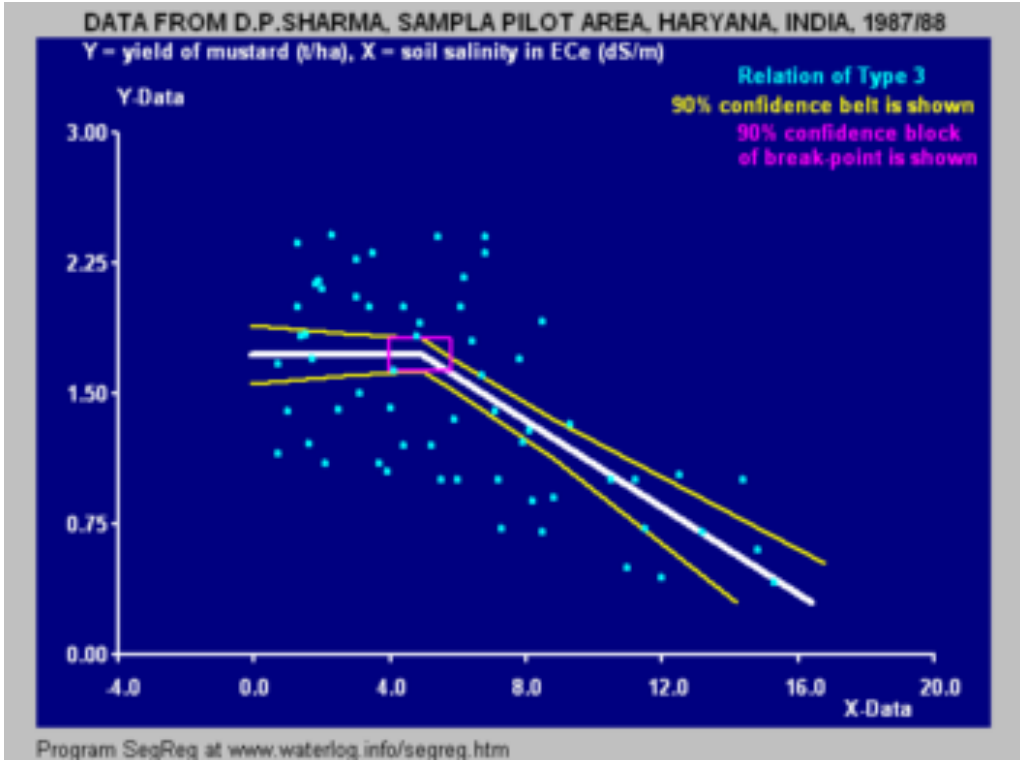
Linear Regression is problematic in many other cases





Piecewise Linear Regression





Also ref: Multivariate Adaptive Regression Splines (MARS)

Bivariate and multivariate models

Bivariate or simple regression model

(Education) $x \longrightarrow y$ (Income)

Multivariate or multiple regression model

(Education) x_1

(Gender) x_2

(Experience) x_3

(Age) x_4

$\longrightarrow$ y (Income)

Model with simultaneous relationship

Price of wheat $\longleftrightarrow$ Quantity of wheat produced

Types of Regression Models