# Regression Models
## Fitting Techniques

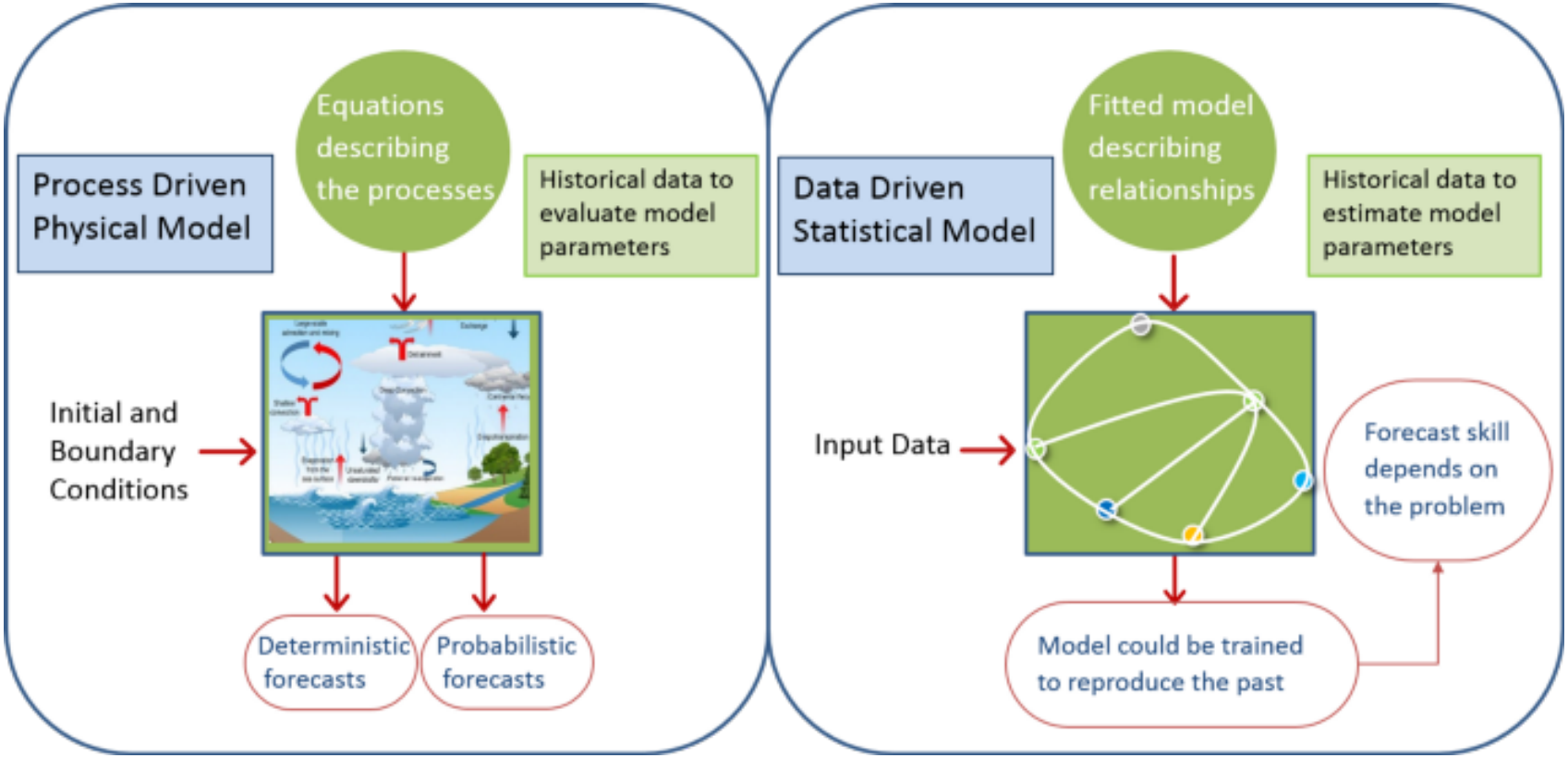
Alexander Trofimov
PhD, professor, NRNU MEPhI

lab@neuroinfo.ru
http://datalearning.ru

Course "Machine Learning"

November 2022

# Theoretical and Statistical Models




## Types of mathematical models:

- Theoretical models (physical models, economical models, etc.)
- Statistical models (data-driven)

# What is Statistical Model?

## Definition

Statistical models are mathematical models used to describe patterns of variability that random variables or data may display

In fact, statistical model is represented as a collection of probability distributions $\mathscr{P} = \{P\}$ of random variable or vector $X$

Data that are observations of random variable or vector $X$, are used to select a specific distribution $\hat{P} \in \mathscr{P}$

The statistical model is called parametrized if

$$\mathscr{P} = \{P_\theta, \theta \in \Theta\}$$

where $\Theta$ is the parameter space

The statistical assumptions about the process under modelling are usually needed to construct its statistical model

# Example 1: Binomial Statistical Model

**Experiment:**

The manufactured items are inspected for defects. $N$ is a number of observed items, $n$ is the number of defected items among them

What statistical model can be proposed for this experiment?

# Example 1: Binomial Statistical Model

**Experiment:**
The manufactured items are inspected for defects. $N$ is a number of observed items, $n$ is the number of defected items among them

What statistical model can be proposed for this experiment?

**Description:**
Let $X$ be the number of defected items among $N$ items, and $p$ is the probability of defect for every single item (that is unknown)

Then, the probabilities for $X$:

$$P(k,p) = P[X = k] = C_N^k p^k (1-p)^{N-k}, \quad k \in \{0, 1, ..., N\}$$

and the random variable $X$ has a binomial distribution $B(N,p)$, and $n$ is an observation of the random variable $X$

# Example 1: Binomial Statistical Model

**Statistical model:**

$$X \sim B(N, p)$$

$$\mathscr{P} = \{P(k, p), p \in [0, 1]\}$$

**Fitting to the data:**

The specific distribution $P$ from $\mathscr{P}$ is obtained by fitting the parameter $p$ to the given data:

$$\hat{p} = \frac{n}{N}$$

The $\hat{p}$ is an observed proportion of defected items that is an estimation of probability $p$ of defect for every single item

**The probabilities for $X$ after fitting:**

$$\hat{P}(k) = C_N^k \hat{p}^k (1 - \hat{p})^{N-k}, \quad k \in \{0, 1, ..., N\}$$

# Example 2: Multinomial Statistical Model

**Experiment:**

Three persons play cards $N = 10$ times. The 1-st player won in 2 games, the 2-nd player won in 3 games, and the 3-rd player won in 5 games

What statistical model can be proposed for this experiment?

# Example 2: Multinomial Statistical Model

**Experiment:**

Three persons play cards $N = 10$ times. The 1-st player won in 2 games, the 2-nd player won in 3 games, and the 3-rd player won in 5 games

What statistical model can be proposed for this experiment?

**Description:**

Let $X = (X_1, ..., X_M)^T$ be the numbers of games won by players $1, ..., M$ respectively, $p_j$ is the probability of winning for $j$-th player (that is unknown), $j = \overline{1, M}$, and $M$ is the number of players

Then, the probabilities for $X$:

$$P(x) = P[X = x] = \frac{N!}{\prod_{j=1}^{M} x_j!} \prod_{j=1}^{M} p_j^{x_j}$$

where $x = (x_1, ..., x_M)^T$, $x_j \in \{0, ..., M\}$, $j = \overline{1, M}$, $\sum_{j=1}^{M} x_j = N$

# Example 2: Multinomial Statistical Model

**Statistical model:**

$$X \sim Mult(N, p_1, ..., p_M)$$

$$\mathscr{P} = \left\{ P_X(x, p),\ p \in [0,1]^M,\ \sum_{j=1}^{M} p_j = 1 \right\}$$

where $p = (p_1, ..., p_M)$ is a vector of parameters

**Fitting to the data:**

The estimation $\hat{p}_j$ of probability $p_j$ is a frequency of wins for $j$-th player:

$$\hat{p}_j = \frac{n_j}{N}, \quad j = \overline{1, M}$$

**The probabilities for $X$ after fitting:**

$$\hat{P}(x) = \frac{N!}{\prod_{j=1}^{M} x_j!} \prod_{j=1}^{M} \hat{p}_j^{x_j}$$

# Example 3: One-Sample Normal Statistical Model

**Experiment:**
The sample $x_1, ..., x_n$ are independent measurements of a physical constant $\mu$ in a scientific experiment

What statistical model can be proposed for this experiment?

# Example 3: One-Sample Normal Statistical Model

**Experiment:**
The sample $x_1, ..., x_n$ are independent measurements of a physical constant $\mu$ in a scientific experiment

What statistical model can be proposed for this experiment?

**Description:**
Let $X$ be the result of measurement of a physical constant $\mu$.
Suppose, that measurement errors are additive Gaussian white noise with variance $\sigma^2$:

$$X = \mu + \varepsilon$$

where $\varepsilon$ is a random variable, $\varepsilon \sim N(0, \sigma^2)$

# Example 3: One-Sample Normal Statistical Model

**Statistical model:**

$$X \sim N(\mu, \sigma^2)$$

$$\mathscr{P} = \left\{ P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left( -\frac{(x-\mu)^2}{2\sigma^2} \right),\ \mu \in \mathbb{R},\ \sigma > 0 \right\}$$

where $\mu$ and $\sigma$ are unknown constants

**Fitting to the data:**

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

**The probability distribution of $X$ after fitting:**

$$\hat{P}(x) = \frac{1}{s\sqrt{2\pi}} \exp\left( -\frac{(x-\bar{x})^2}{2s^2} \right)$$

# Example 4: Two-Sample Normal Statistical Model

**Experiment:**

The samples $x_1, ..., x_{n_1}$ and $y_1, ..., y_{n_2}$ are independent measurements of the same device characteristic before and after its tuning

What statistical model can be proposed for this experiment?

# Example 4: Two-Sample Normal Statistical Model

**Experiment:**

The samples $x_1, ..., x_{n_1}$ and $y_1, ..., y_{n_2}$ are independent measurements of the same device characteristic before and after its tuning

What statistical model can be proposed for this experiment?

**Description:**

Let random variables $X$ and $Y$ be the results of measurements before and after tuning respectively. Suppose, that the device characteristic before tuning was equal to $\mu_1$ and after tuning became equal to $\mu_2$. Also suppose the measurement errors are additive Gaussian white noise with variance $\sigma^2$:

$$X = \mu_1 + \varepsilon, \quad Y = \mu_2 + \varepsilon$$

where $\varepsilon$ is a random variable, $\varepsilon \sim N(0, \sigma^2)$

# Example 4: Two-Sample Normal Statistical Model

**Statistical model:**

$$X \sim N(\mu_1, \sigma^2), \quad Y \sim N(\mu_2, \sigma^2)$$

where $\mu_1$, $\mu_2$ and $\sigma$ are unknown constants

**Fitting to the data:**

$$\hat{\mu}_1 = \bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i, \quad \hat{\mu}_2 = \bar{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i$$

$$\hat{\sigma}^2 = s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

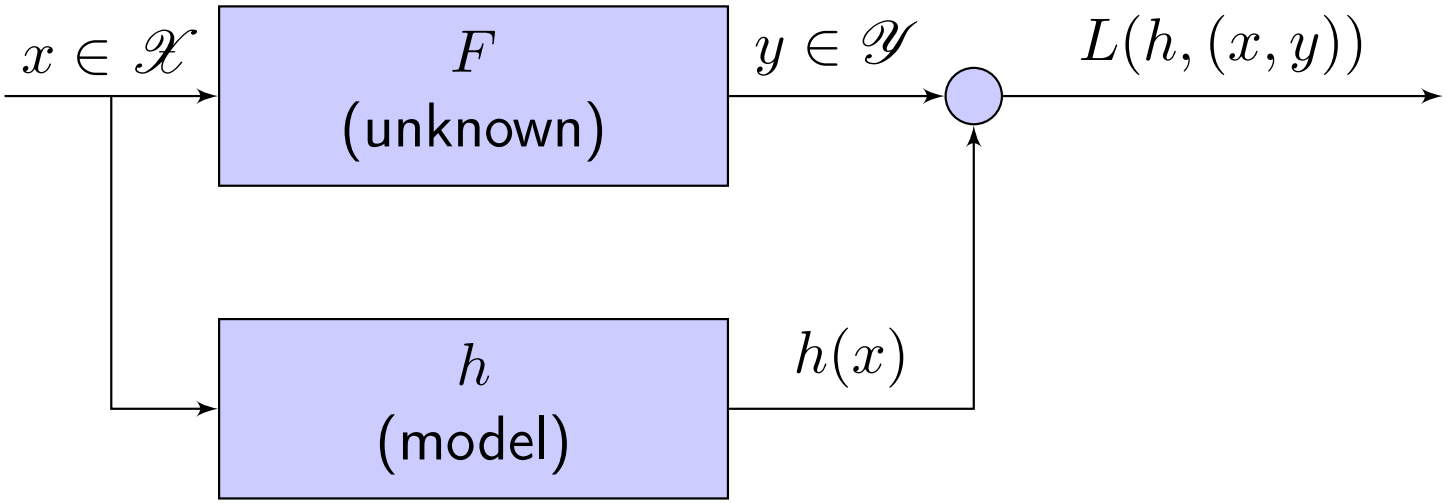
where $s_1^2$ and $s_1^2$ are estimations of $\sigma_1^2$ and $\sigma_2^2$. The estimation $s^2$ is called as pooled estimation of variance $\sigma^2$

**The probability distribution of $X$ and $Y$ after fitting:**

$$X \sim N(\bar{x}, s^2), \quad Y \sim N(\bar{y}, s^2)$$

# Regression Models

Particular case of statistical models are regression models




$x \in \mathscr{X}$ is a vector of independent variables called as regressors, or predictors, or explanatory variables

$y \in \mathscr{Y}$ is a dependent variable called as outcome, or response. The response domain $\mathscr{Y}$ is a set of real numbers, $\mathscr{Y} = \mathbb{R}$

$L(h,(x,y))$ is a loss function associated with model $h$

# Regression Models. Problem Statement

**Given:**

$\mathscr{D}_T = \{(x_1, y_1), ..., (x_n, y_n)\}$ is a training data sample

$\mathscr{H}$ is a class of hypotheses (e.g., linear functions)

$L(h, (x, y)) = (h(x) - y)^2$ is a quadratic loss function

**Objective:**

Find hypothesis $h \in \mathscr{H}$ that minimizes empirical risk $R^*$ over training sample $\mathscr{D}_T$:

$$R^*(h) = \frac{1}{n} \sum_{i=1}^{n} (h(x_i) - y_i)^2 \to \min_{h \in \mathscr{H}}$$

If $\mathscr{H}$ is parametrized, $\mathscr{H} = \{h_\beta, \beta \in \mathbb{R}^k\}$, then

$$R^*(\beta) = \frac{1}{n} \sum_{i=1}^{n} (h(x_i, \beta) - y_i)^2 \to \min_{\beta \in \mathbb{R}^k}$$

## Risk for Quadratic Loss

The target mapping $F$ is considered to be stochastic, and model output $h(x)$ is deterministic for any $x \in \mathscr{X}$

Risk for quadratic loss function $L(h,(x,Y))$ at given $x \in \mathscr{X}$:

$$\begin{aligned} R(h,x) &= \mathrm{M}[L(h,(x,Y))|x] = \mathrm{M}\left[(h(x)-Y)^2|x\right] \\ &= h^2(x) - 2h(x)\mathrm{M}[Y|x] + \mathrm{M}[Y^2|x] \\ &= h^2(x) - 2h(x)\mathrm{M}[Y|x] + \mathrm{D}[Y|x] + (\mathrm{M}[Y|x])^2 \\ &= (h(x)-\mathrm{M}[Y|x])^2 + \sigma_x^2 \end{aligned}$$

$(h(x)-\mathrm{M}[Y|x])^2$ is a error of model $h$ at given $x \in \mathscr{X}$

$\sigma_x^2 = \mathrm{D}[Y|x]$ is a noise, doesn't depend on $\mathscr{D}$ or $h$

**The risk $R(h,x)$ is minimal if**

$h(x) = \mathrm{M}[Y|x], \forall x \in \mathscr{X} \Leftrightarrow h(x)$ is a regression function of $Y$ on $x$

# Regression Function

Let $(X, Y)$ be a random vector, $(X, Y) \sim F_{XY}(x, y)$

## Definition

Regression function $\varphi(x)$ of $Y$ on $x$ is a conditional expectation of random variable $Y$ as a function of $x$:

$$\varphi(x) = \mathrm{M}[Y|x], \quad x \in \mathscr{X}$$

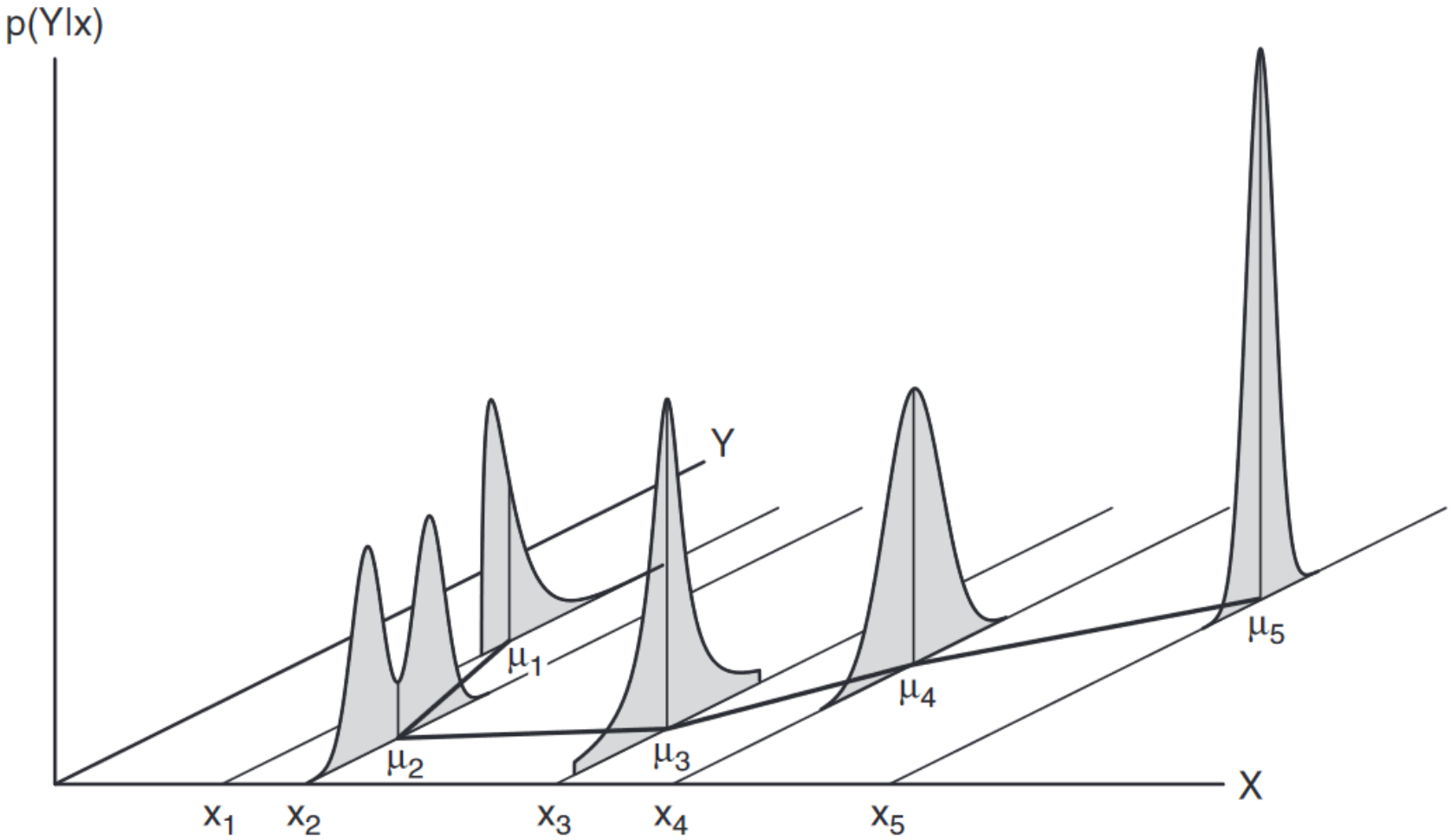
**Probability theory background:**

$$\mathrm{M}[Y|x] = \int_{-\infty}^{\infty} y f_Y(y|x) dy$$

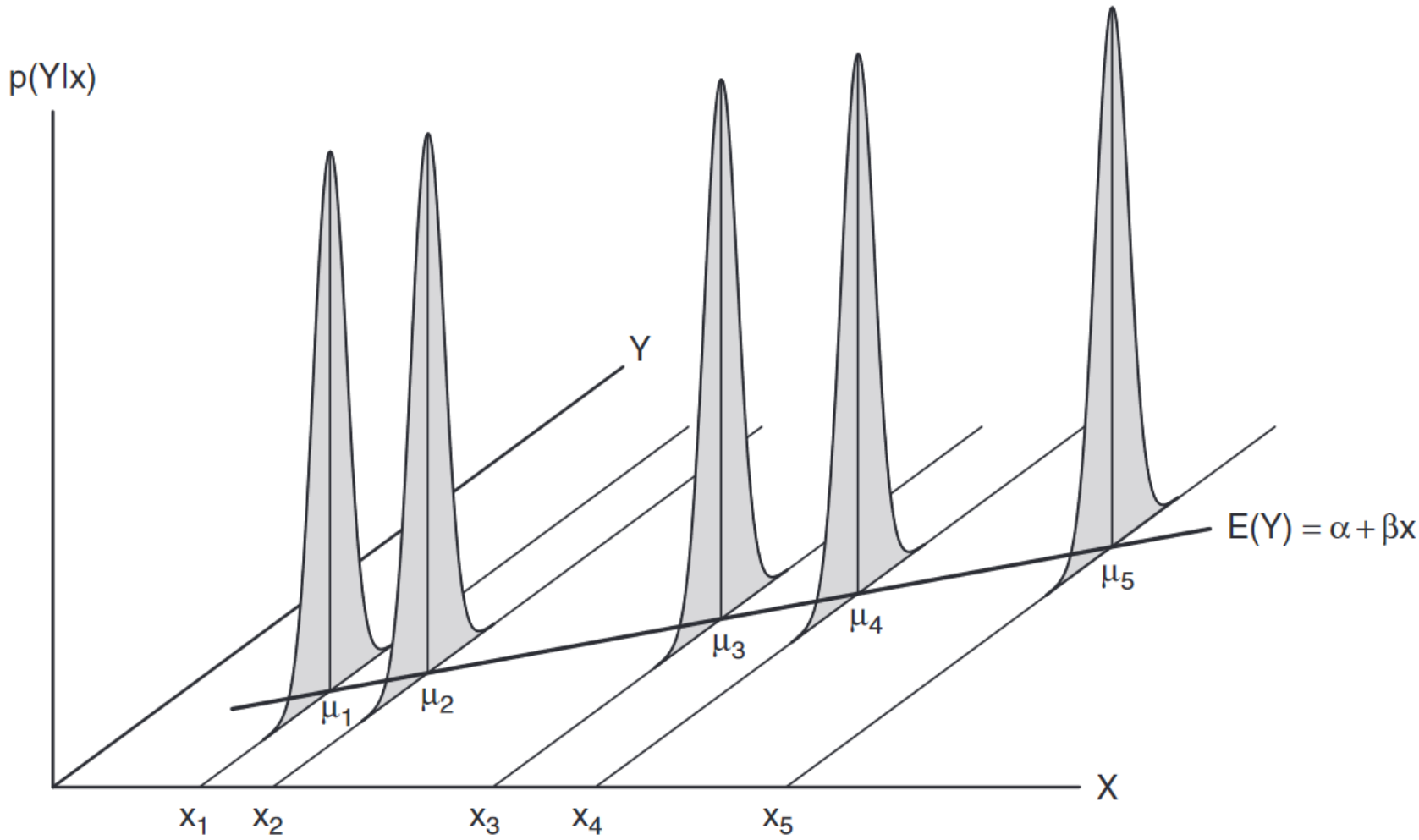
$$f_Y(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy, \quad f_{XY}(x, y) = \frac{\partial F_{XY}(x, y)}{\partial x \partial y}$$

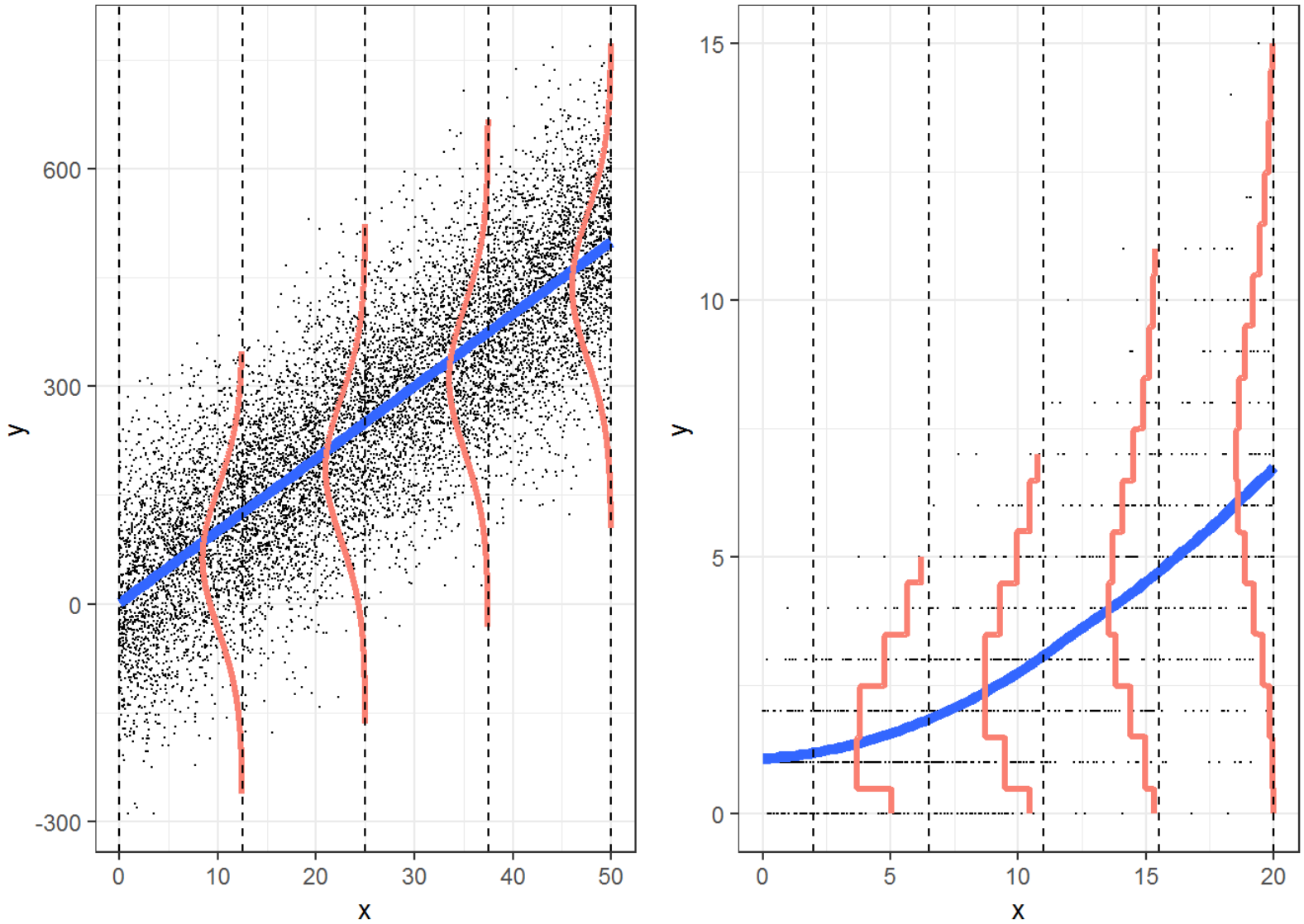
# Regression Function. Illustration 1



# Regression Function. Illustration 2



# Regression Function. Illustration 3



# Optimality of Regression Models

**Optimal case:**

$$h(x) = \varphi(x) = \mathrm{M}[Y|x]$$

**The regression model of response $Y$ for a given $x \in \mathscr{X}$:**

$$Y|x = \varphi(x) + \varepsilon(x) = \mathrm{M}[Y|x] + \varepsilon(x)$$

where $\varepsilon(x)$ is a random variable (noise)

**The regression error:**

$$\varepsilon(x) = Y|x - \mathrm{M}[Y|x]$$

$$\mathrm{M}[\varepsilon(x)] = \mathrm{M}[Y|x] - \mathrm{M}[Y|x] = 0$$

$$\mathrm{D}[\varepsilon(x)] = \mathrm{M}[\varepsilon(x)^2] = \mathrm{M}[(Y|x - \mathrm{M}[Y|x])^2] = \mathrm{D}[Y|x] = \sigma_x^2$$

The regression model has minimum variance of model errors $\mathrm{D}[\varepsilon(x)]$ among all statistical models (prove it!)

# Regression Analysis

The objective of regression analysis is to estimate the relationship between a quantitative response variable $Y$ and one or more explanatory variables $X_1, ..., X_k$

**Purposes of regression analysis:**

- **Description or explanation**
  Finding the explanatory variables that matter, estimating their effect on the outcome, concluding the mechanism, law
- **Prediction**
  Estimating missing data within the range of $X$ values or predict data outside the range (extrapolation)
- **Auxiliary purposes**
  Data reduction, filter linear effects, etc.

## How to Find the Regression Function?

For a given joint distribution $F_{XY}(x, y)$ the regression function $\varphi(x)$ can be derived analytically

In practice we usually don't know $F_{XY}(x, y)$, and the regression function $\varphi(x)$ should be estimated, or fitted to the data

**Fitted regression model:**

$$Y|x = \hat{\varphi}(x) + \hat{\varepsilon}(x)$$

where $\hat{\varphi}(x)$ is an estimate of regression function $\varphi(x)$ based on the given training data sample

The problem of function estimation is ill-posed, some distributional assumptions about $F_Y(y|x)$ are needed

# Types of Regression Analysis

- Parametric
  The family of distributions $\{F_Y(y|x)\}$ is fully specified up to unknown parameters $\beta$, that leads to a parametric class of hypothesis $\mathscr{H} = \{h(x, \beta), \beta \in \mathbb{R}^k\}$, e.g.:

$$Y|x \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

- Semi-parametric
  Some aspects of the distribution $F_Y(y|x)$ are described by parameters $\beta$, but others are left unspecified, e.g.:

$$\mathrm{M}[Y|x] = \beta_0 + \beta_1 x, \quad \mathrm{D}[Y|x] = \sigma^2$$

- Non-parametric
  No distributional assumptions about $F_Y(y|x)$

# Regression Analysis Pipeline

- **Step 1.** Specify outcome $Y$ and explanatory variables $x_1, ..., x_k$
- **Step 2.** Propose distributional assumptions about $F_Y(y|x)$ and specify the class of regression models $\mathscr{H}$
- **Step 3.** Estimate the regression function $\varphi(x)$ or its parameters $\beta$ using the data
- **Step 4.** Validate the regression model (regression diagnostic)
- **Step 5.** Interpret the regression results
- **Step 6.** If necessary modify model and/or distributional assumptions
- **Step 7.** Use the regression model

## Simple Linear Regression

In simple linear regression the relationship between response $Y$ and explanatory variable $x$ is modelled as

$$Y|x = \beta_0 + \beta_1 x + \varepsilon(x)$$

where $\beta_0, \beta_1$ are parameters to be estimated (or fitted) using the data, $\beta_0$ is intercept, $\beta_1$ is slope (or rate of change), $\varepsilon(x)$ is a random error

The regression function $\varphi(x)$ is assumed to be linear:

$$\varphi(x) = \mathrm{M}[Y|x] = \beta_0 + \beta_1 x$$

Differences between actual and modelled responses are called regression errors:

$$\varepsilon_i = y_i - \varphi(x_i) = y_i - \beta_0 - \beta_1 x_i, \quad i = 1, ..., n$$

# Ordinary Least Squares

The ordinary least squares (OLS) method finds the optimal parameter values by minimizing the sum (or mean) of squared residuals:

$$E(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \varphi(x_i))^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

The minimum is found by setting the gradient to zero:

$$\frac{\partial E(\beta_0, \beta_1)}{\partial \beta_0} = -\frac{1}{n}\sum(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial E(\beta_0, \beta_1)}{\partial \beta_1} = -\frac{1}{n}\sum x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

**The system of linear equations (so called normal equations):**

$$\begin{cases} \beta_0 n + \beta_1 \sum x_i = \sum y_i, \\ \beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i \end{cases}$$

## Parameters Estimation

**The system of normal equations in matrix form:**

$$X^T X\beta = X^T y$$

where $\beta = (\beta_0, \beta_1)^T$ is vector of parameters, $y = (y_1, ..., y_n)^T$ is vector of responses, $X$ is design matrix:

$$X = \begin{pmatrix} 1 & x_1 \\ ... & ... \\ 1 & x_n \end{pmatrix}$$

**The closed-form solution:**

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

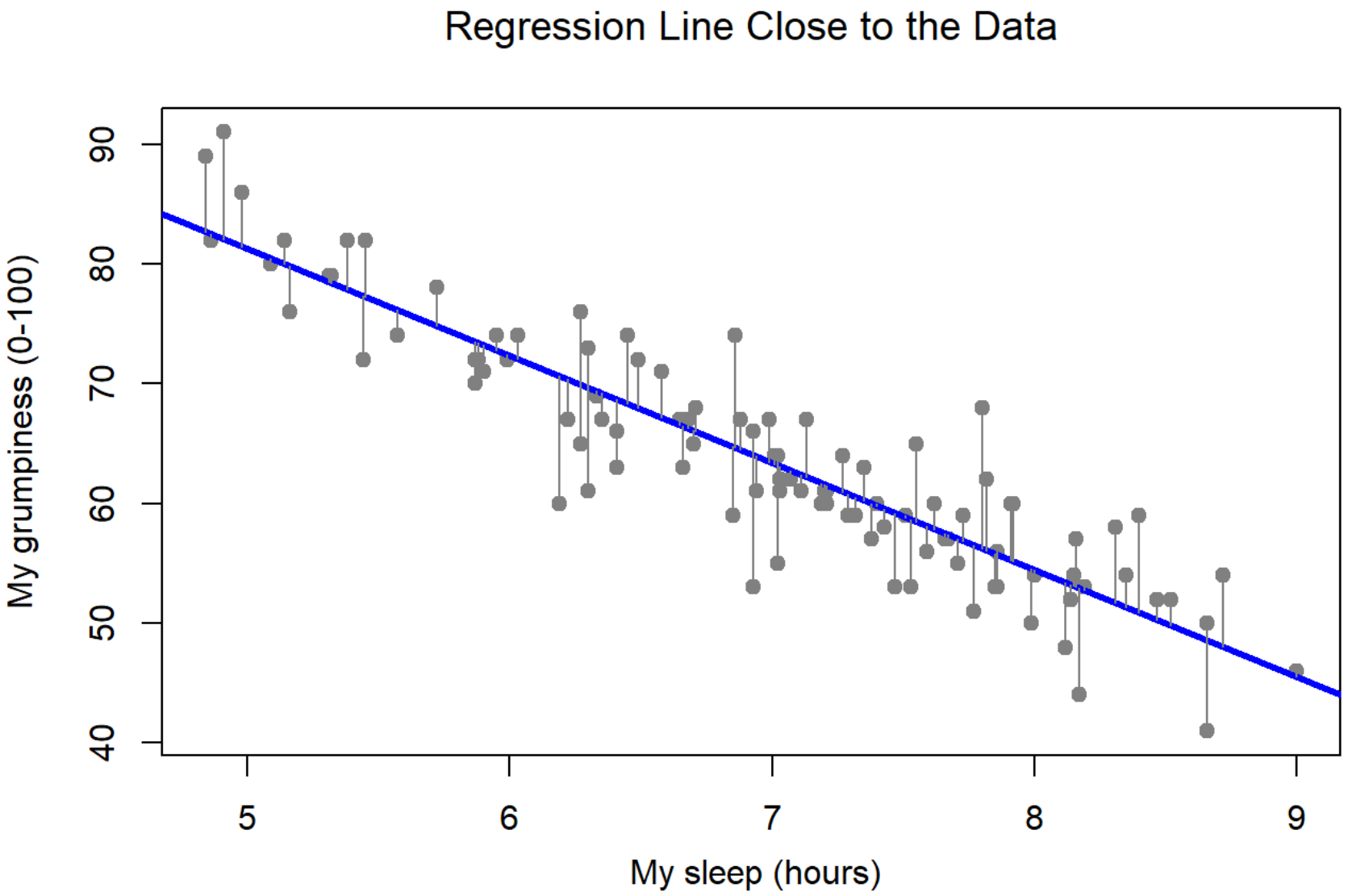
**In scalar form:**

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

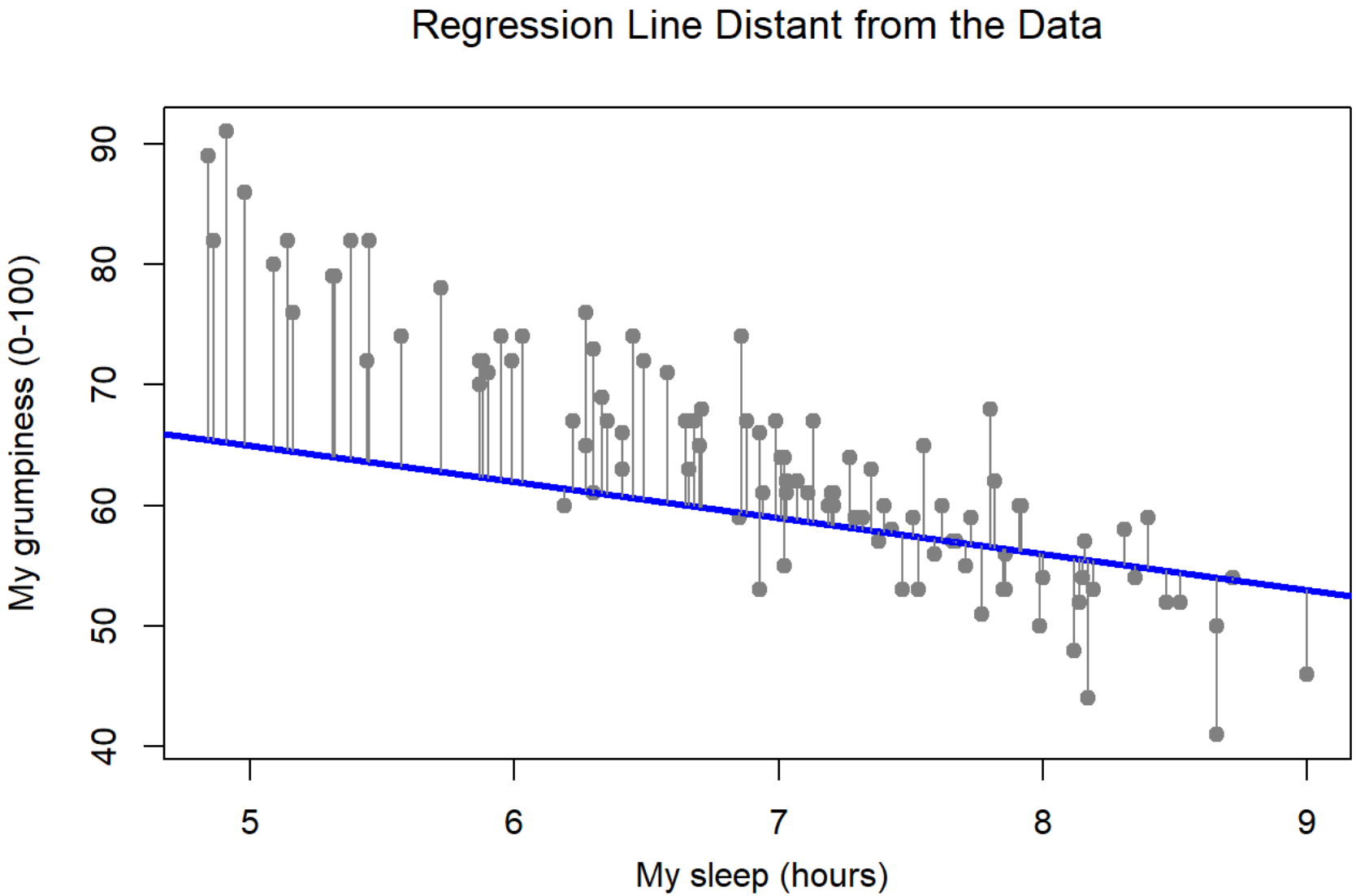
$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \rho_{XY} \frac{s_Y}{s_X}$$

where $\rho_{XY}$ is sample correlation, $s_X$ and $s_Y$ are standard deviations

# Simple Linear Regression. Illustration 1

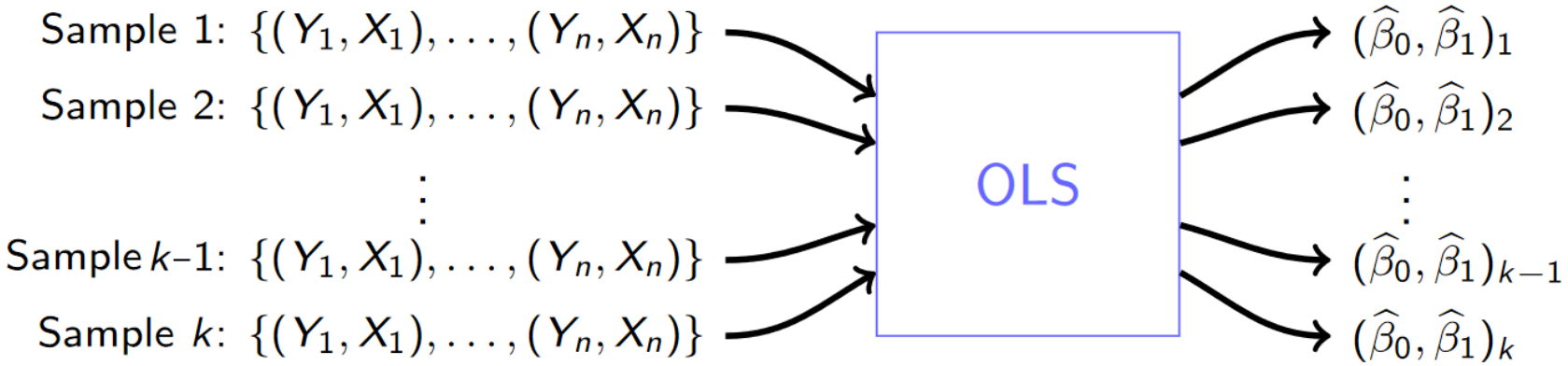


# Simple Linear Regression. Illustration 2



# OLS as an Estimator

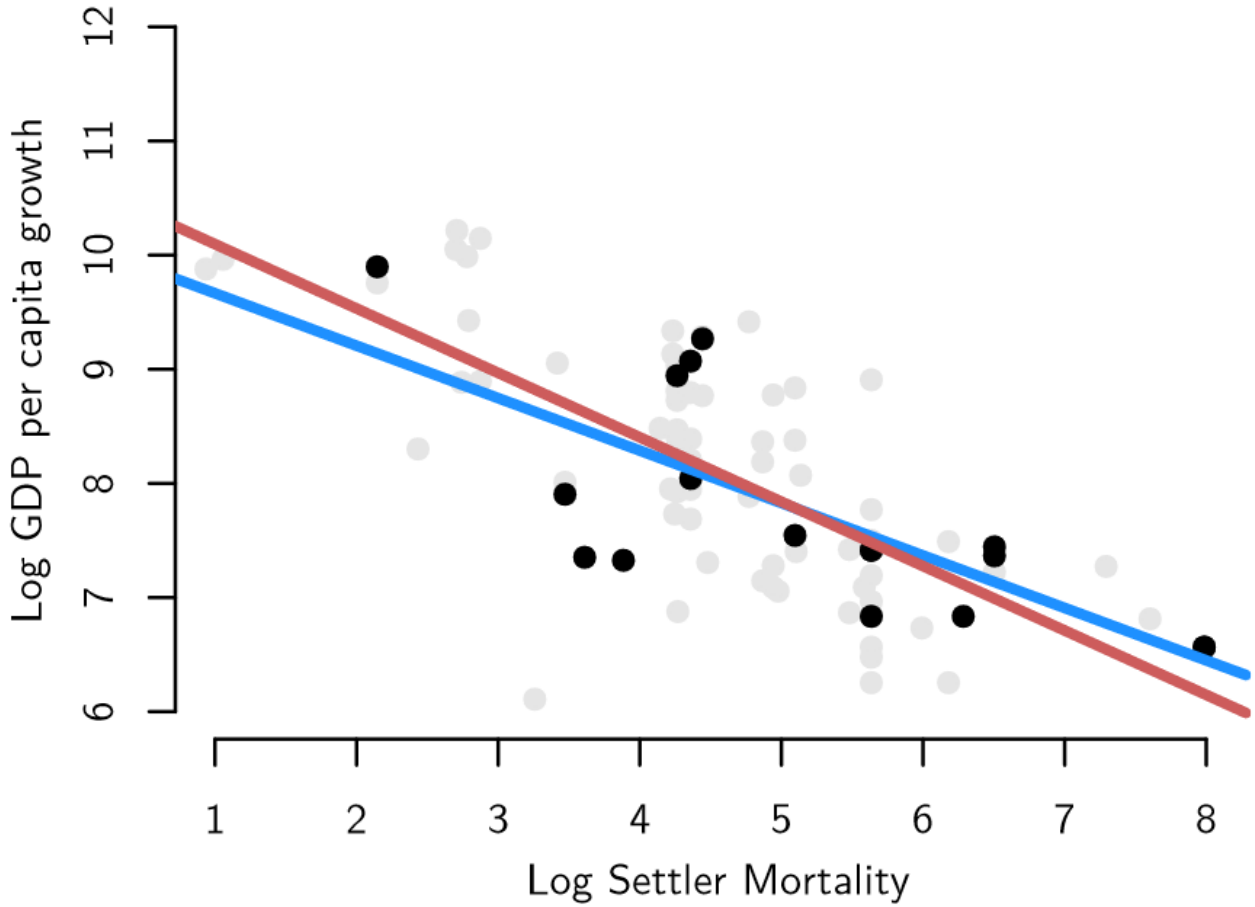
OLS is a method for parameters estimation, it can be viewed as a machine that we plug data into and we get out estimates




The estimates $\hat{\beta}_0, \hat{\beta}_1$ are random variables just like the sample mean or the sample variance or any other statistics

As a result, the estimate $\hat{\varphi}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ of conditional expectation $\varphi(x) = \mathrm{M}[Y|x]$ will also be a random variable

# Randomness of Estimates. Illustration




Red line is true (population) regression, blue line is an estimate for random sample

## Regression Errors

Regression errors $\varepsilon_1, ..., \varepsilon_n$ are the deviations of the observations $y_1, ..., y_n$ of dependent variable $Y$ from values $\varphi(x_1), ..., \varphi(x_n)$ of the regression function $\varphi(x)$:

$$\varepsilon_i = y_i - \varphi(x_i), \quad i = \overline{1, n}$$

As soon as the regression function $\varphi(x)$ is unknown, the errors $\varepsilon_1, ..., \varepsilon_n$ are unknown

Random regression errors $\varepsilon_1, ..., \varepsilon_n$ are the deviations of random variables $Y|x_1, ..., Y|x_n$ from values $\varphi(x_1), ..., \varphi(x_n)$ of the regression function $\varphi(x)$:

$$\varepsilon_i = Y|x_i - \varphi(x_i) = Y_i - \varphi(x_i), \quad i = \overline{1, n}$$

As soon as $Y|x_1, ..., Y|x_n$ are random variables, the errors $\varepsilon_1, ..., \varepsilon_n$ are random variables too

## Regression Residuals

Regression residuals $\hat{\varepsilon}_1, ..., \hat{\varepsilon}_n$ are the differences between the observations $y_1, ..., y_n$ and estimated values $\hat{\varphi}(x_1), ..., \hat{\varphi}(x_n)$ of the regression function $\varphi(x)$:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\varphi}(x_i), \quad i = \overline{1, n}$$

Regression residuals are observed, they form a sample of values

Random regression residuals $\hat{\varepsilon}_1, ..., \hat{\varepsilon}_n$ are the differences between the random variables $Y|x_1, ..., Y|x_n$ and estimates $\hat{\varphi}(x_1), ..., \hat{\varphi}(x_n)$ of the regression function $\varphi(x)$:

$$\hat{\varepsilon}_i = Y|x_i - \hat{\varphi}(x_i) = Y_i - \hat{\varphi}(x_i), \quad i = \overline{1, n}$$

The estimates $\hat{\varphi}(x_1), ..., \hat{\varphi}(x_n)$ are also random as soon as $\hat{\varphi}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and the estimates $\hat{\beta}_0, \hat{\beta}_1$ are random variables

## Assumptions in Regression Analysis

OLS procedure only provides estimates of the coefficients $\beta$, it makes no assumptions about the random errors $\varepsilon_1, ..., \varepsilon_n$
The statistical assumptions are needed to make statistical inferences for regression

- **Linearity**
  The regression function: $\varphi(x) = \mathrm{M}[Y|x] = \beta_0 + \beta_1 x$

$$Y|x = \varphi(x) + \varepsilon(x)$$

$$\mathrm{M}[Y|x] = \mathrm{M}[\varphi(x) + \varepsilon(x)] = \varphi(x) + \mathrm{M}[\varepsilon(x)]$$

  It's equivalent to the centerness of random errors:

$$\mathrm{M}[\varepsilon(x)] = 0 \quad \forall x \in \mathscr{X}$$

- **Exogeneity of regressors**
  The explanatory variable $X$ and the error $\varepsilon(x)$ are independent $\forall x \in \mathscr{X}$. It means that the sampled values of explanatory variable are independent on model errors

## Assumptions in Regression Analysis

- **Homoscedasticity**
  The variance of the errors is the same regardless of the value of explanatory variable:

$$\mathrm{D}[\varepsilon(x)] = \sigma^2 \quad \forall x \in \mathscr{X}$$

  It means that the response variable $Y$ is also homoscedastic:

$$\mathrm{D}[Y|x] = \mathrm{D}[\beta_0 + \beta_1 x + \varepsilon(x)] = \mathrm{D}[\varepsilon(x)] = \sigma^2 \quad \forall x \in \mathscr{X}$$

- **Independence of errors**
  Any pair of errors $\varepsilon(x_i)$ and $\varepsilon(x_j)$ are independent. It's the same as any pair of responses $Y|x_i$ and $Y|x_j$ are independent. Weaker assumption is uncorrelatedness:

$$cov[\varepsilon(x_i), \varepsilon(x_j)] = cov[Y|x_i, Y|x_j] = 0 \quad \forall x_i, x_j \in \mathscr{X},\ i \neq j$$

## Assumptions in Regression Analysis

- **Normality**
  The errors are normally distributed for every $x \in \mathscr{X}$:

$$\varepsilon(x) \sim N(0, \sigma^2) \quad \forall x \in \mathscr{X}$$

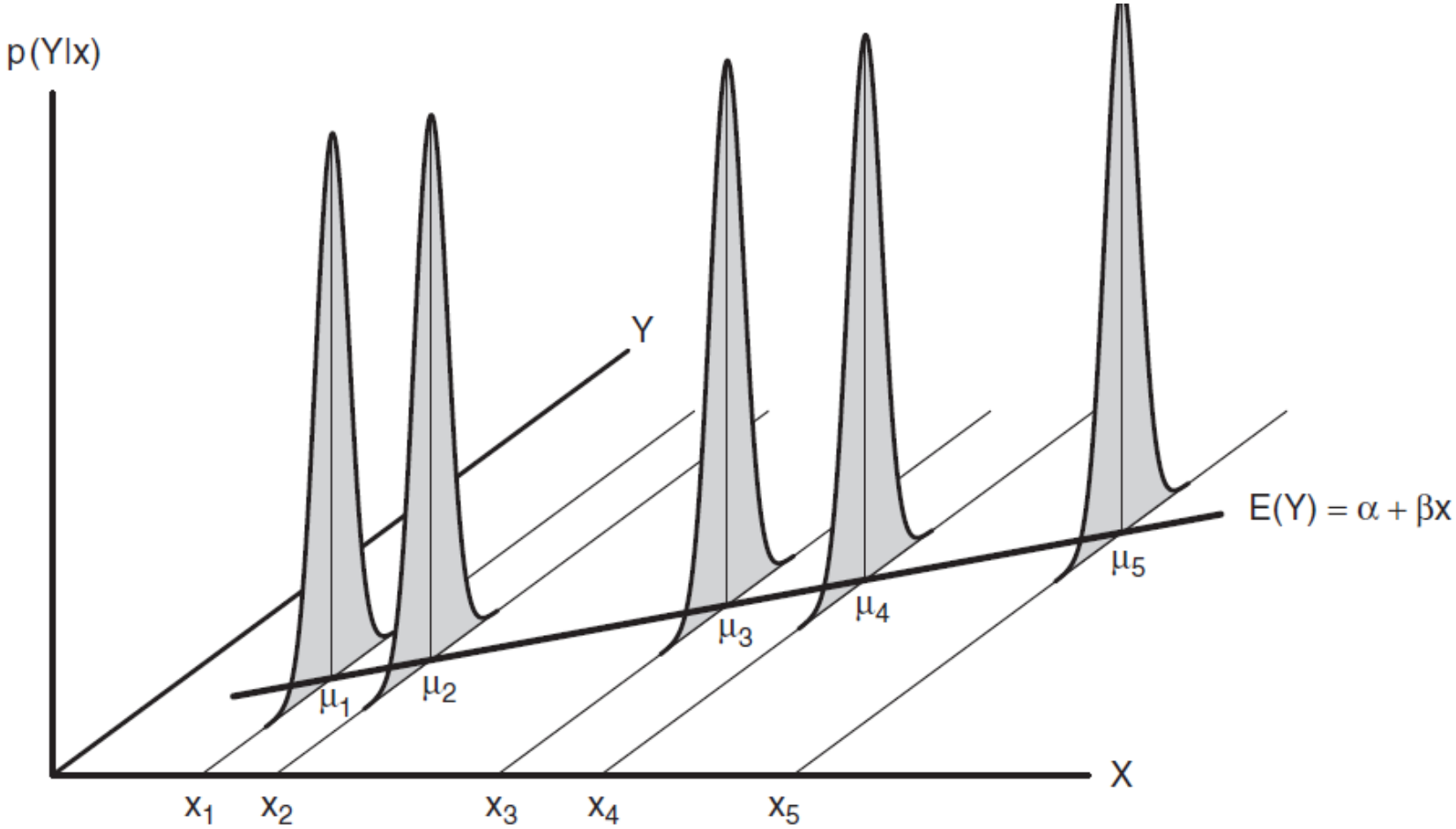
  Equivalently:

$$Y|x \sim N(\beta_0 + \beta_1 x, \sigma^2) \quad \forall x \in \mathscr{X}$$

  Under assumption of normality the independence and uncorrelatedness of errors are equivalent

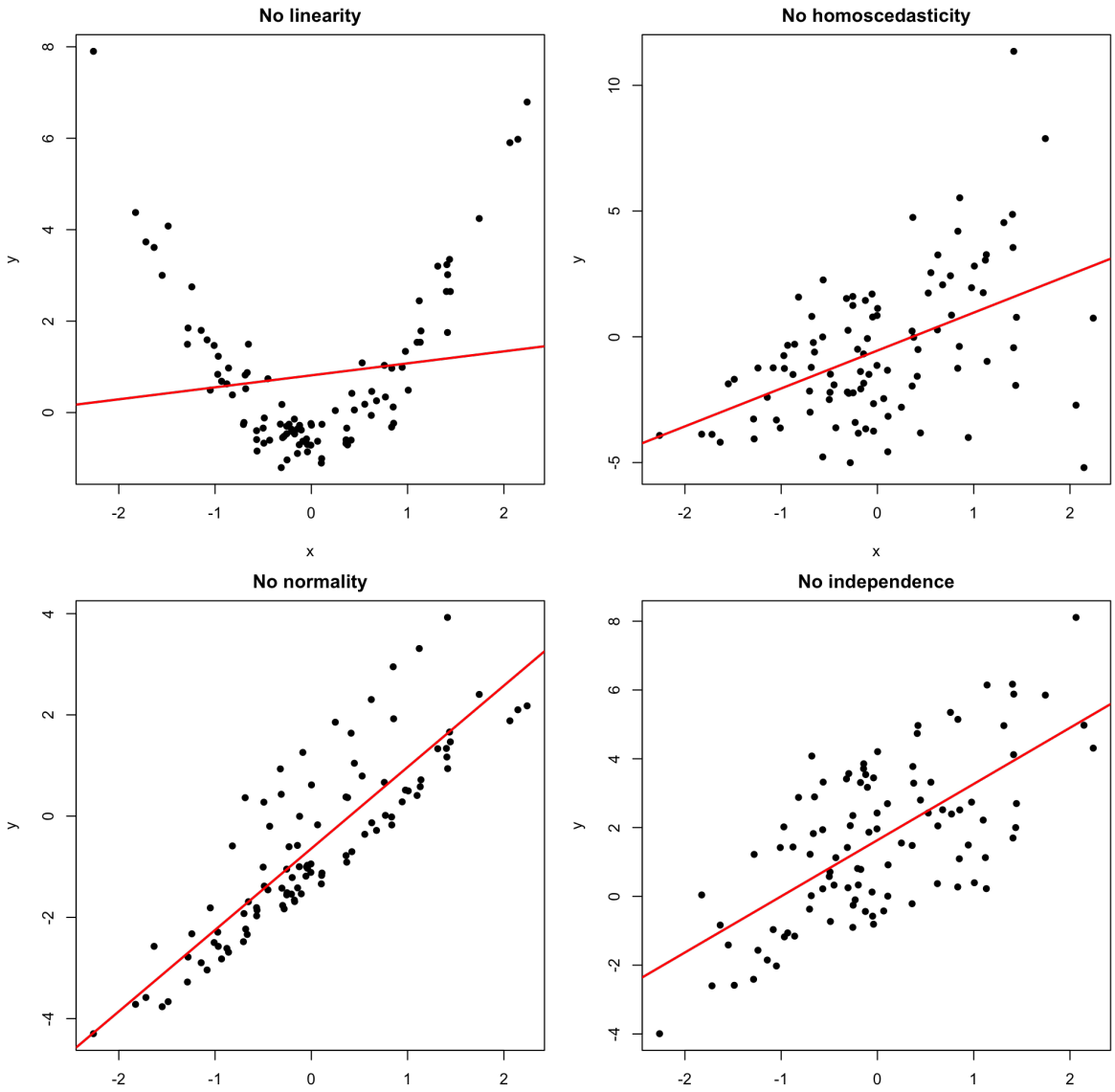
- **Variability of regressors**
  The explanatory variables must have non-zero variance. The regression has no sense if $X$ is constant

# Assumptions in Regression Analysis. Illustration




The assumptions of linearity, constant variance, and normality in simple regression are fulfilled

# Violations of Assumptions. Illustrations



# Properties of the OLS Estimator: Linearity

**Assumption:** linearity

The coefficients $\hat{\beta}_0, \hat{\beta}_1$ are linear estimators (i.e. linear functions of the observations $y_1, ..., y_n$):

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

where $B = (X^T X)^{-1} X^T$ is $2 \times n$ matrix

$$\hat{\beta}_0 = (1\ 0)\hat{\beta} = (1\ 0)By = \sum_{i=1}^{n} b_{0i} y_i$$

$$\hat{\beta}_1 = (0\ 1)\hat{\beta} = (0\ 1)By = \sum_{i=1}^{n} b_{1i} y_i = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} y_i$$

Derive the expression for $b_{0i}$!

# Properties of the OLS Estimator: Consistency

**Assumptions:** linearity, exogeneity of regressors

The coefficients $\hat{\beta}_0, \hat{\beta}_1$ are consistent estimators:

$$\hat{\beta}_0 \xrightarrow{P} \beta_0, \quad \hat{\beta}_1 \xrightarrow{P} \beta_1$$

$$P[|\hat{\beta}_j - \beta_j| < \delta] \to 1 \quad \forall \delta > 0 \quad \text{as} \quad n \to \infty, \quad j \in \{0, 1\}$$

**Proof:**

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum(x_i - \bar{x})Y_i}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i + \varepsilon_i)}{\sum(x_i - \bar{x})^2} \\
&= \frac{\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\beta_0 + \frac{\sum(x_i - \bar{x})x_i}{\sum(x_i - \bar{x})^2}\beta_1 + \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2} \\
&= 0 + \beta_1 + \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2} \to \beta_1 + \frac{cov[X, \varepsilon]}{\mathrm{D}[X]} = \beta_1 + 0 = \beta_1
\end{aligned}$$

Prove the consistency of $\hat{\beta}_0$!

# Properties of the OLS Estimator: Unbiasedness

**Assumption:** linearity

The coefficients $\hat{\beta}_0, \hat{\beta}_1$ are unbiased estimators:

$$\mathrm{M}[\hat{\beta}_0] = \beta_0, \quad \mathrm{M}[\hat{\beta}_1] = \beta_1$$

**Proof:**

$$\begin{aligned}
\mathrm{M}[\hat{\beta}_1] &= \mathrm{M}\left[\frac{\sum(x_i - \bar{x})Y_i}{\sum(x_i - \bar{x})^2}\right] = \frac{\sum(x_i - \bar{x})\mathrm{M}[Y_i]}{\sum(x_i - \bar{x})^2} \\
&= \frac{\sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i + \mathrm{M}[\varepsilon_i])}{\sum(x_i - \bar{x})^2} \\
&= \frac{\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\beta_0 + \frac{\sum(x_i - \bar{x})x_i}{\sum(x_i - \bar{x})^2}\beta_1 = 0 + \beta_1 = \beta_1
\end{aligned}$$

Prove the unbiasedness of $\hat{\beta}_0$!

# Properties of the OLS Estimator: Efficiency

**Assumptions:** linearity, exogeneity of regressors, homoscedasticity, independence of errors

**The sampling variances of estimators:**

$$\mathrm{D}[\hat{\beta}_0] = \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2} = \frac{\sigma^2 \alpha_{2X}^*}{n D_X^*}$$

$$\mathrm{D}[\hat{\beta}_1] = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} = \frac{\sigma^2}{n D_X^*}$$

Prove it!

## Gauss-Markov theorem

Of all the linear unbiased estimators, the OLS estimators are the most efficient, that is, they have the smallest sampling variance. Under assumption of normality, moreover, they are the most efficient among all unbiased estimators

## Properties of the OLS Estimator: Normality

**Assumptions:** linearity, homoscedasticity, independence of errors, normality of errors

The coefficients $\hat{\beta}_0, \hat{\beta}_1$ are normally distributed random variables:

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \alpha^*_{2X}}{nD^*_X}\right), \quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{nD^*_X}\right)$$

Even if the errors $\varepsilon(x)$ are not normally distributed, the distributions of $\hat{\beta}_0, \hat{\beta}_1$ are approximately normal

The variance $\sigma^2$ of the random error $\varepsilon(x)$ is usually unknown. It can be shown that the unbiased estimate of $\sigma^2$ is residual variance:

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

and $\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2(n-2)$, and $S_e^2$ is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$

## OLS Estimator and MLE

**Assumptions:** linearity, homoscedasticity, independence of errors, normality of errors

The estimates $\hat{\beta}_0, \hat{\beta}_1$ are maximum likelihood estimates (MLE) of coefficients $\beta_0, \beta_1$

**Statistical model:**

$$\varepsilon(x) \sim N(0, \sigma^2), \quad \varepsilon(x) = Y|x - \beta_0 - \beta_1 x$$

**Sample:** $\varepsilon_1, ..., \varepsilon_n, \quad \varepsilon_i = y_i - \beta_0 - \beta_1 x_i, \quad i = \overline{1, n}$

Likelihood function:

$$\mathscr{L}(\varepsilon_1, ..., \varepsilon_n, \beta_0, \beta_1) = \prod_{i=1}^{n} f_\varepsilon\left(\varepsilon_i | \beta_0, \beta_1\right)$$

MLE of $\beta_0, \beta_1$ is a solution of optimization problem:

$$\mathscr{L}(\varepsilon_1, ..., \varepsilon_n, \beta_0, \beta_1) \to \max_{\beta_0, \beta_1}$$

# MLE of Simple Regression Parameters

Log-likelihood function:

$$\ln \mathscr{L}(\varepsilon_1, ..., \varepsilon_n, \beta_0, \beta_1) = \sum_{i=1}^{n} \ln f_\varepsilon(\varepsilon_i | \beta_0, \beta_1) \to \max_{\beta_0, \beta_1}$$

$$\ln \mathscr{L}(\varepsilon_1, ..., \varepsilon_n, \beta_0, \beta_1) = \sum_{i=1}^{n} \ln \left[ \frac{1}{\sigma\sqrt{2\pi}} \exp\left( -\frac{\varepsilon_i^2}{2\sigma^2} \right) \right]$$

$$= n \ln \left( \frac{1}{\sigma\sqrt{2\pi}} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \varepsilon_i^2 = c - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\begin{cases} \frac{\partial \ln \mathscr{L}(\varepsilon_1, ..., \varepsilon_n, \beta_0, \beta_1)}{\partial \beta_0} = \frac{1}{\sigma^2} \sum (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial \ln \mathscr{L}(\varepsilon_1, ..., \varepsilon_n, \beta_0, \beta_1)}{\partial \beta_1} = \frac{1}{\sigma^2} \sum x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases}$$

It's the same system of linear equations like it was in OLS

Derive the MLE-estimation of $\sigma^2$!

# Confidence Intervals for Regression Parameters

**The sampling variances of estimators:**

$$S^2[\hat{\beta}_0] = \frac{S_e^2 \alpha_{2X}^*}{nD_X^*}, \quad S^2[\hat{\beta}_1] = \frac{S_e^2}{nD_X^*}$$

**Central statistic**:

$$T_j = \frac{\hat{\beta}_j - \beta_j}{S[\hat{\beta}_j]} \sim T(n-2), \quad j \in \{0,1\}$$

**Confidence intervals for $\beta_j$, $j \in \{0,1\}$:**

$$P\left[\hat{\beta}_j - t_{1-\alpha/2,n-2} S[\hat{\beta}_j] < \beta_j < \hat{\beta}_j + t_{1-\alpha/2,n-2} S[\hat{\beta}_j]\right] = 1-\alpha$$

where $1-\alpha$ is a confidence level, $t_{1-\alpha/2,n-2}$ is $(1-\alpha/2)$-th quantile of Student's distribution with $n-2$ degrees of freedom

# Confidence and Prediction Intervals for Regression Function

**$(1-\alpha)$-confidence interval for conditional expectation $\varphi(x)$:**

$$\varphi(x) \in \left[\hat{\beta}_0 + \hat{\beta}_1 x \mp t_{1-\alpha/2,n-2} S_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{nD_X^*}}\right]$$

**$(1-\alpha)$-prediction interval for $Y|x$:**

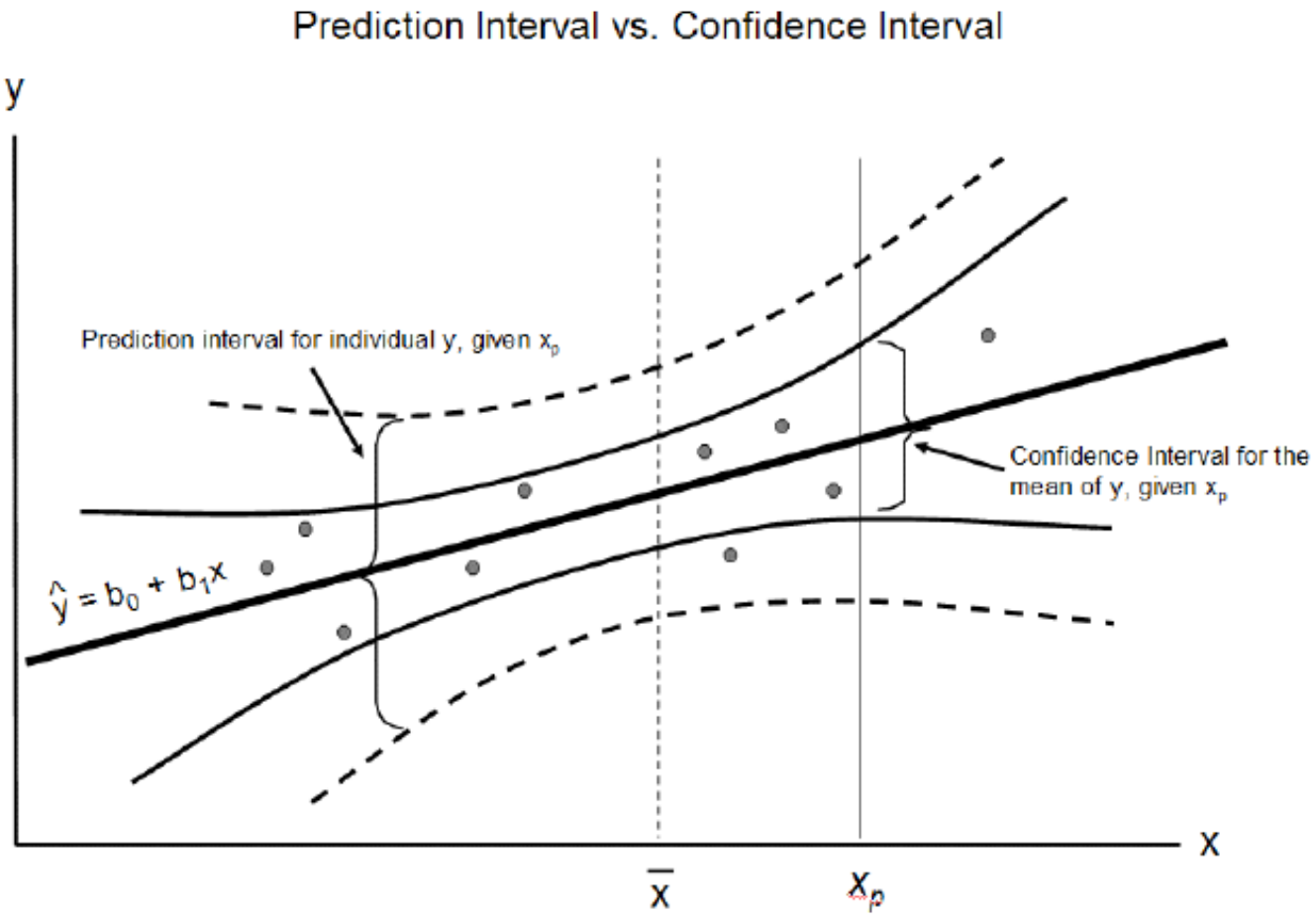
$$Y|x \in \left[\hat{\beta}_0 + \hat{\beta}_1 x \mp t_{1-\alpha/2,n-2} S_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{nD_X^*}}\right]$$

Confidence interval gives a range of values for an unknown conditional expectation $\mathrm{M}[Y|x]$

Prediction interval is an estimate of an interval in which a future observation $Y|x$ will fall

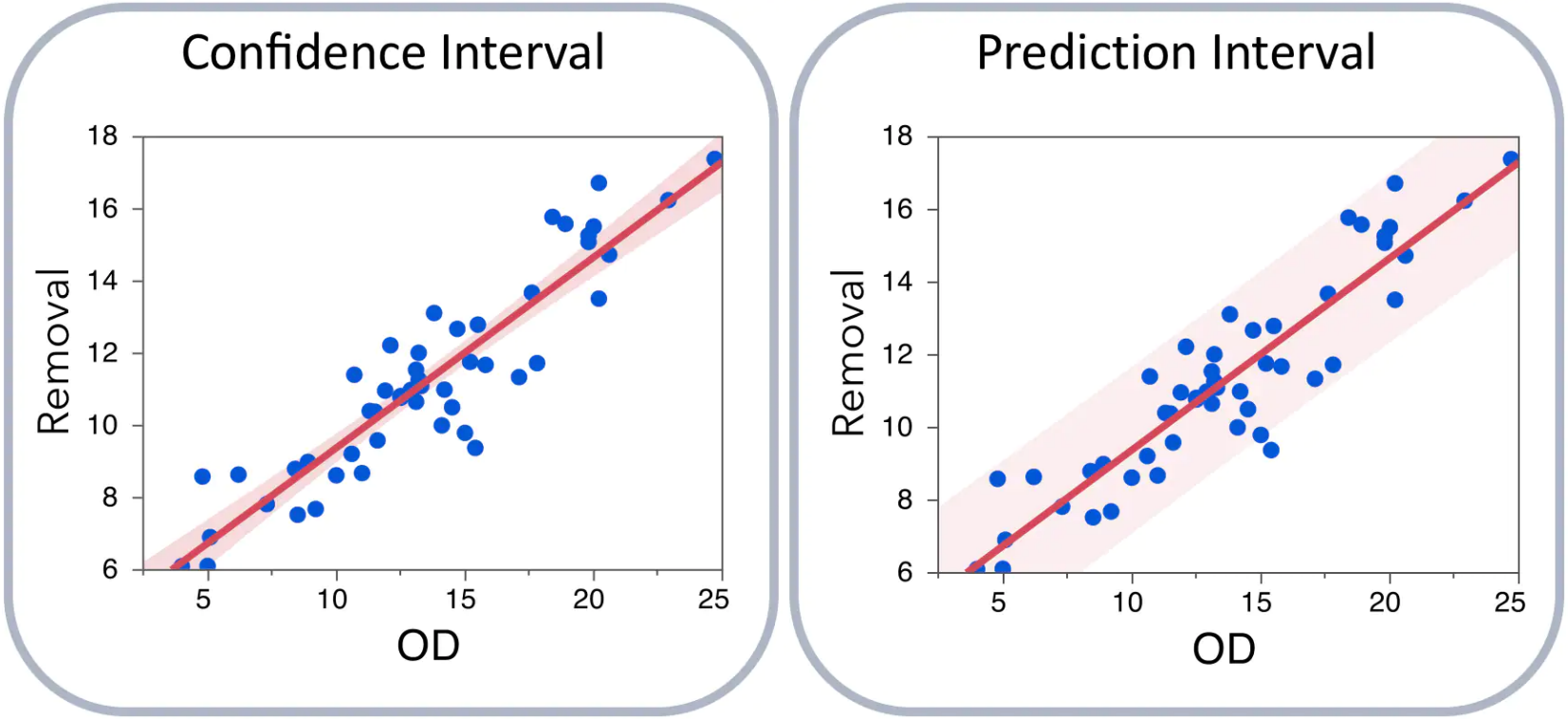
Confidence interval is always narrower than the corresponding prediction interval

# Confidence and Prediction Intervals. Illustration 1




Confidence interval is for conditional expectation $\varphi(x) = \mathrm{M}[Y|x]$, prediction interval is for a single value of $Y|x$

# Confidence and Prediction Intervals. Illustration 2




The further $x$ is from $\bar{x}$, the wider the intervals will be

If any of the statistical assumptions is violated, then the confidence intervals and prediction intervals may be invalid as well. This is why it's important to check them by examining the residuals, etc.

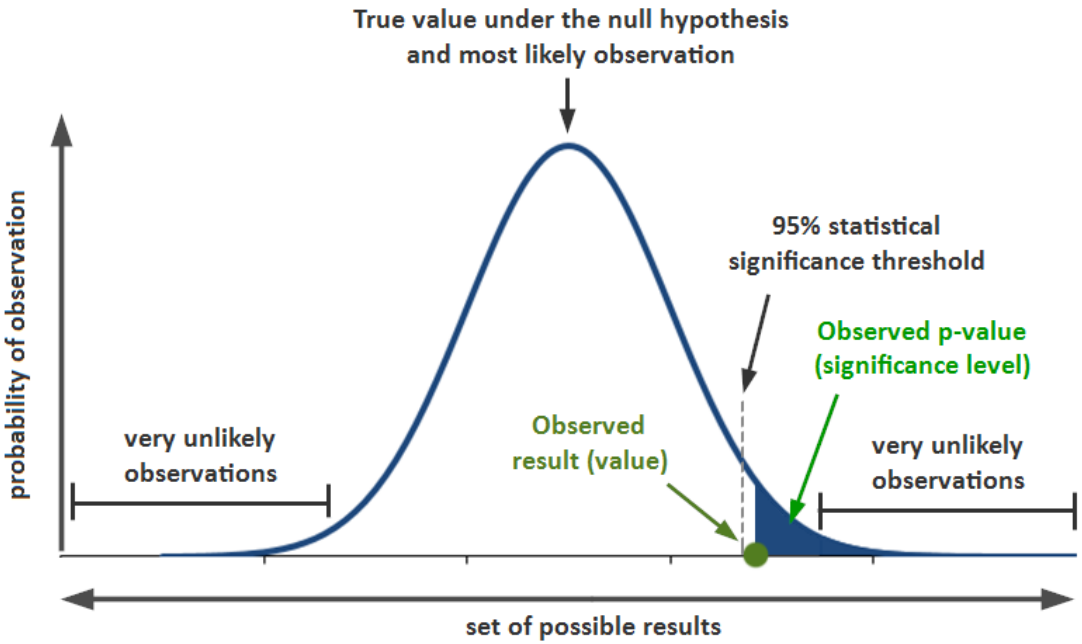
# Single Parameter Tests

## Statistical hypothesis:

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H' : \beta_j \neq 0, \quad j \in \{0, 1\}$$

## Test statistic:

$$Z_j = \frac{\hat{\beta}_j}{S[\hat{\beta}_j]}, \quad Z_j | H_0 \sim T(n-2), \quad j \in \{0, 1\}$$



## Multiple Linear Regression

In multiple linear regression the relationship between response $Y$ and explanatory variables $X_1, ..., X_k$ is modelled as

$$Y|x = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon(x) = x\beta + \varepsilon(x)$$

where $x = (1, x_1, ..., x_k)$ is a vector of regressors, $\beta = (\beta_0, ..., \beta_k)^T$ is a vector of parameters to be estimated using the data

The regression function $\varphi(x)$ is assumed to be linear:

$$\varphi(x) = \mathrm{M}[Y|x] = x\beta$$

The regression errors:

$$\varepsilon_i = y_i - \varphi(x_i) = y_i - x_i\beta, \quad i = 1, ..., n$$

where $x_i = (1, x_{1i}, ..., x_{ki})$ is a vector of regressors for $i$-th observation

# OLS for Multiple Linear Regression

**Least-squares criterion:**

$$E(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \varphi(x_i))^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\beta)^2 \to \min_{\beta}$$

The minimum is found by setting the gradient to zero:

$$\frac{\partial E(\beta_0, ..., \beta_j)}{\partial \beta_j} = 0, \quad j = \overline{0, k}$$

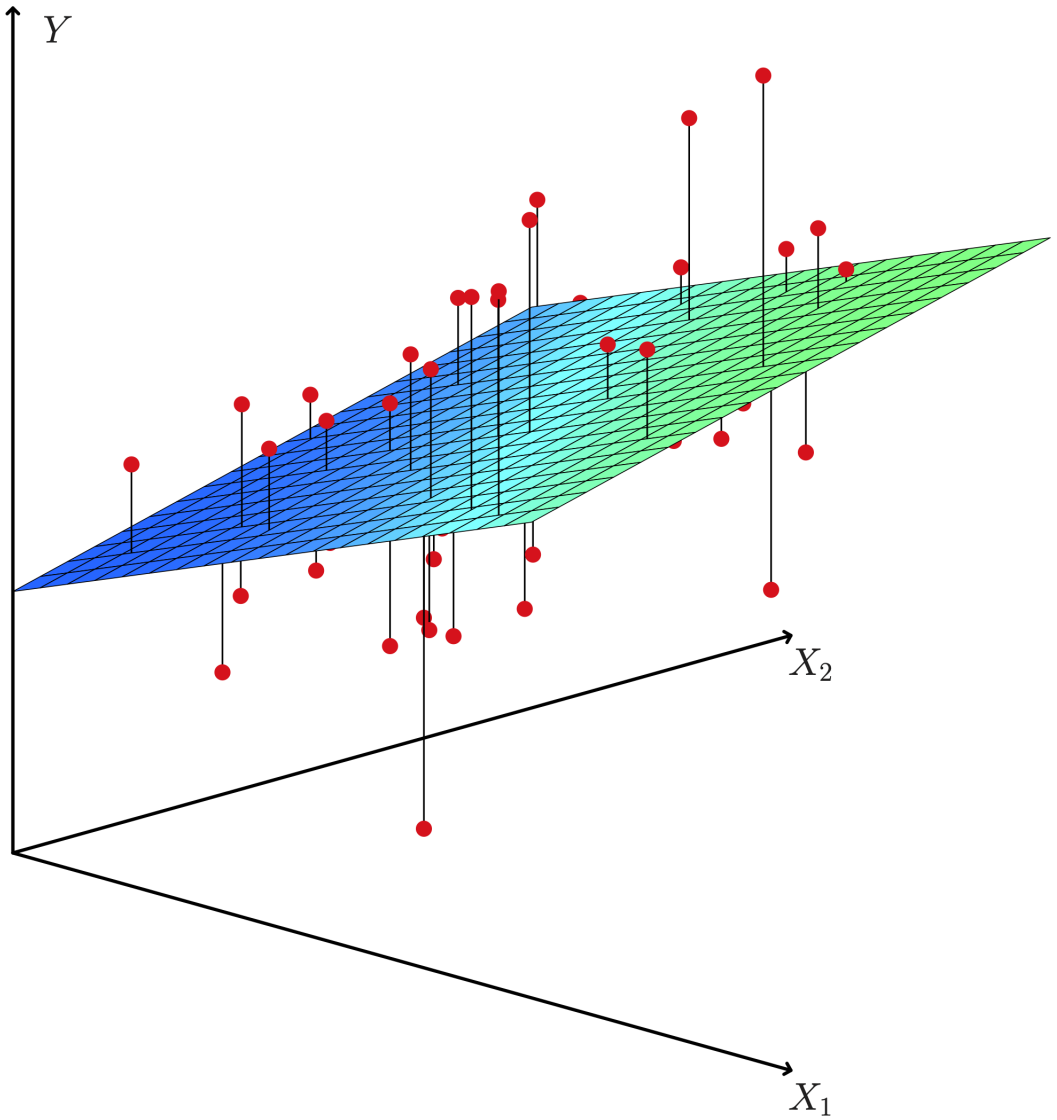
**The system of normal equations:**

$$X^T X\beta = X^T y$$

where $y = (y_1, ..., y_n)^T$ is vector of responses, $X$ is design matrix:

$$X = \begin{pmatrix} 1 & x_{11} & ... & x_{k1} \\ ... & ... & ... & ... \\ 1 & x_{1n} & ... & x_{kn} \end{pmatrix}$$

# Multiple Linear Regression. Illustration



## Parameters Estimation

**The closed-form solution:**

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

**The predicted responses:**

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

where matrix $H = X(X^T X)^{-1} X^T$ is called as projection matrix, or hat matrix

The predicted responses $\hat{y}$ are linear functions of observed responses $y$

**The vector of residuals:**

$$\hat{\varepsilon} = y - \hat{y} = y - Hy = (I - H)y$$

where $\hat{\varepsilon} = (\hat{\varepsilon}_1, ..., \hat{\varepsilon}_n)^T, \quad \hat{\varepsilon}_i = \hat{y}_i - y_i, \quad i = \overline{1, n}$

# Covariance Matrix of Regression Parameter Estimates

It can be shown that the covariance matrix of estimates $\hat{\beta}$:

$$C_{\hat{\beta}} = cov[\hat{\beta}] = \sigma^2 (X^T X)^{-1}$$

where $\sigma^2$ is the variance of the random error, $\sigma^2 = \mathrm{D}[\varepsilon(x)]$

It can be shown that the unbiased estimate of $\sigma^2$ is the variance of residuals (residual variance):

$$s_e^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} (y_i - x_i \hat{\beta})^2$$

and $\frac{(n-k-1)S_e^2}{\sigma^2} \sim \chi^2(n-k-1)$, and $S_e^2$ is independent of $\hat{\beta}_0, ..., \hat{\beta}_k$

**Estimate of covariance matrix $C_{\hat{\beta}}$:**

$$\hat{C}_{\hat{\beta}} = s_e^2 (X^T X)^{-1}$$

# Confidence Intervals for Regression Parameters

**The sampling variances of estimators:**

$$S^2[\hat{\beta}_j] = \hat{c}_{jj}, \quad j = \overline{0,k}$$

where $\hat{c}_{jj}$ is the $j$-th diagonal element of matrix $\hat{C}_{\hat{\beta}}$

**Central statistic**:

$$T_j = \frac{\hat{\beta}_j - \beta_j}{S[\hat{\beta}_j]} \sim T(n-k-1), \quad j = \overline{0,k}$$

**Confidence intervals for $\beta_j$, $j = \overline{0,k}$:**

$$P\left[\hat{\beta}_j - t_{1-\alpha/2,n-k-1} S[\hat{\beta}_j] < \beta_j < \hat{\beta}_j + t_{1-\alpha/2,n-k-1} S[\hat{\beta}_j]\right] = 1-\alpha$$

where $1-\alpha$ is a confidence level, $t_{1-\alpha/2,n-k-1}$ is $(1-\alpha/2)$-th quantile of $T(n-k-1)$-distribution

# Confidence and Prediction Intervals for Regression Function

**$(1-\alpha)$-confidence interval for conditional expectation $\varphi(x)$:**

$$\varphi(x) \in \left[ x\hat{\beta} \mp t_{\alpha/2,n-k-1} S_e \sqrt{x(X^T X)^{-1} x^T} \right]$$

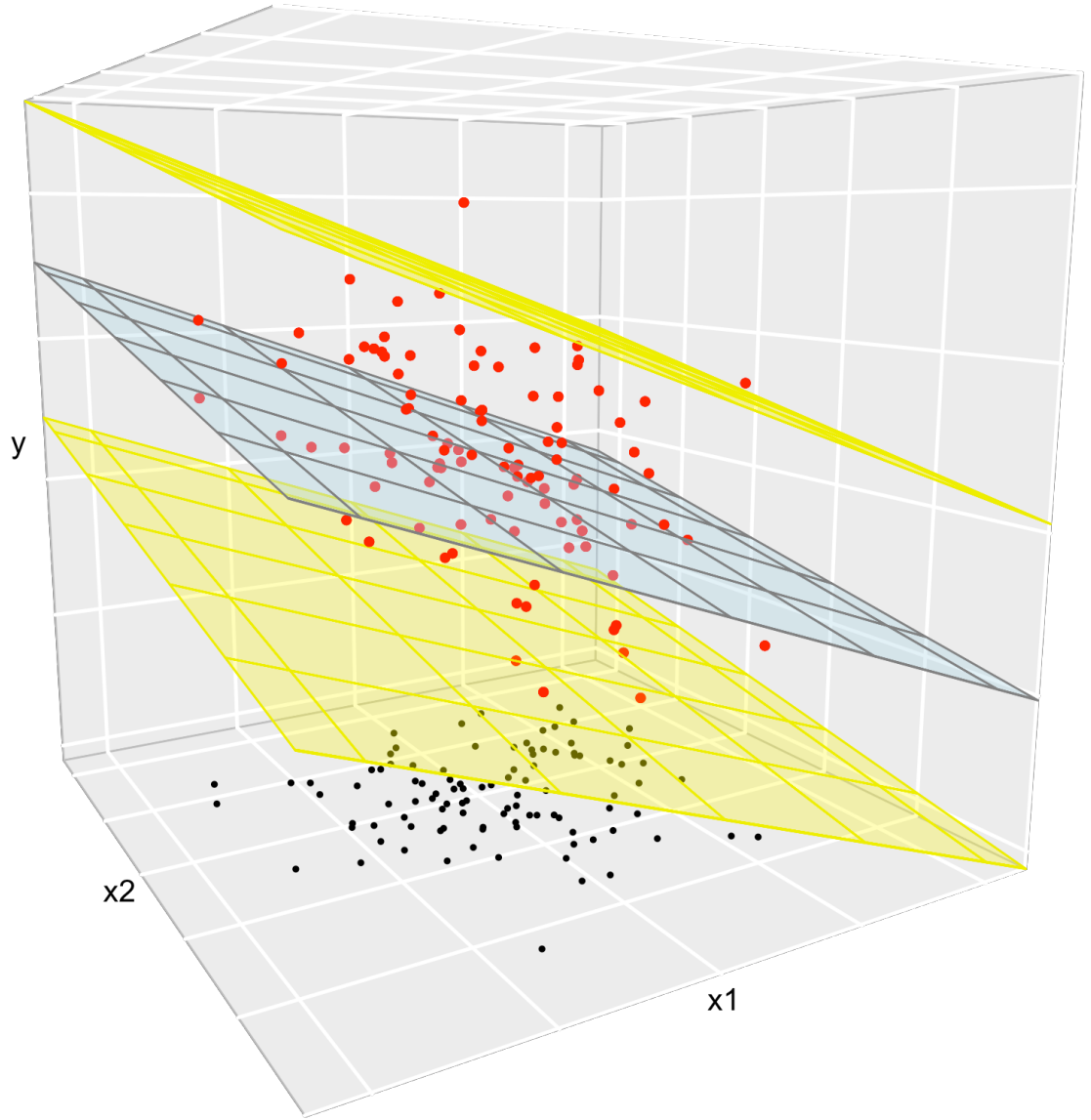
**$(1-\alpha)$-prediction interval for $Y|x$:**

$$Y|x \in \left[ x\hat{\beta} \mp t_{\alpha/2,n-k-1} S_e \sqrt{1 + x(X^T X)^{-1} x^T} \right]$$

Confidence interval gives a range of values for an unknown conditional expectation $\mathrm{M}[Y|x]$

Prediction interval is an estimate of an interval in which a future observation $Y|x$ will fall

# Prediction Intervals for Multiple Regression. Illustration



# Using Features in Multiple Linear Regression

The relationship between response $Y$ and explanatory variables $X_1, ..., X_k$ can be modelled as

$$Y|x = \beta_0 + \beta_1 f_1(x) + ... + \beta_m f_m(x) + \varepsilon(x) = f(x)\beta + \varepsilon(x)$$

where $f(x) = (1, f_1(x), ..., f_m(x))$ is a vector of features

**The system of normal equations in OLS:**

$$X^T X\beta = X^T y$$

where $y = (y_1, ..., y_n)^T$ is vector of responses, $\beta = (\beta_0, ..., \beta_m)^T$ is a vector of parameters, $X$ is design matrix:

$$X = \begin{pmatrix} 1 & f_1(x_1) & ... & f_m(x_1) \\ ... & ... & ... & ... \\ 1 & f_1(x_n) & ... & f_m(x_n) \end{pmatrix}$$

The solution and statistical inference are the same as for multiple linear regression

# Multiple Linear Regression. Examples

**Example 1.** Polynomial regression

$$\varphi(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$f(x) = (1, x, x^2)$$

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \dots & \dots & \dots \\ 1 & x_n & x_n^2 \end{pmatrix}$$

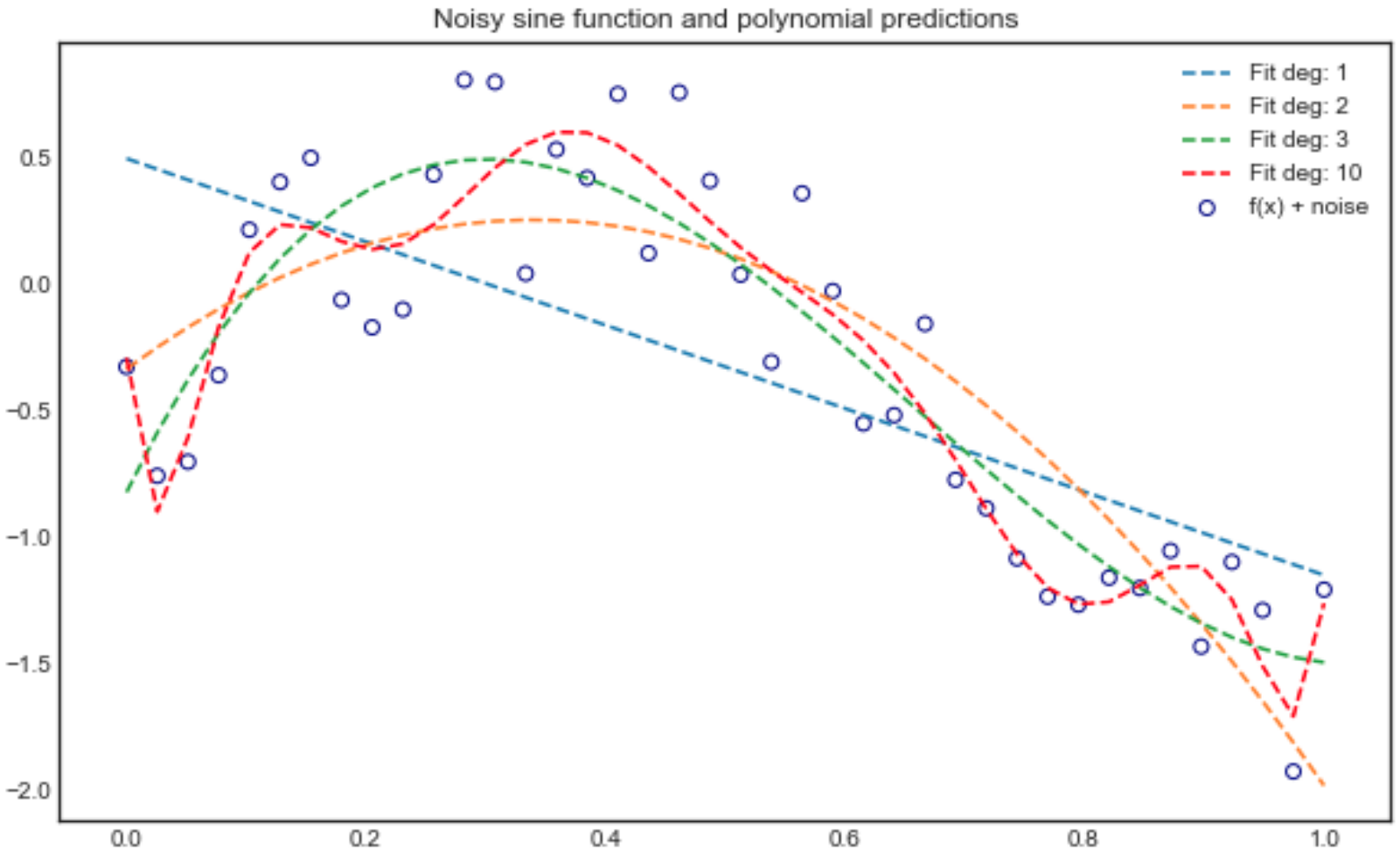
**Example 2.** Linear regression with mixed term

$$\varphi(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

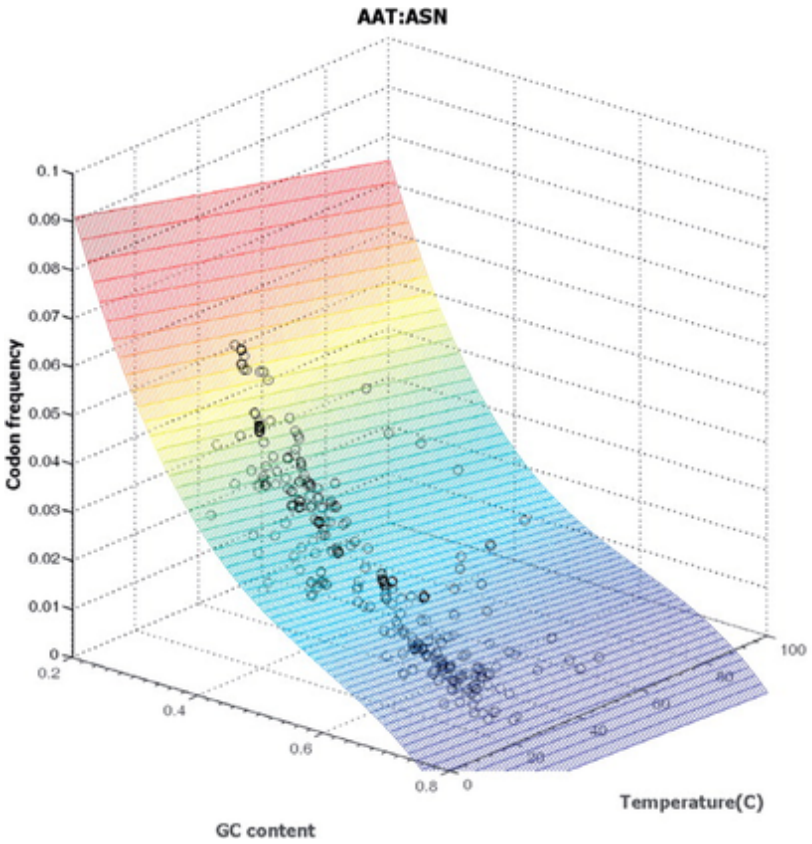
$$f(x) = (1, x_1, x_2, x_1 x_2)$$

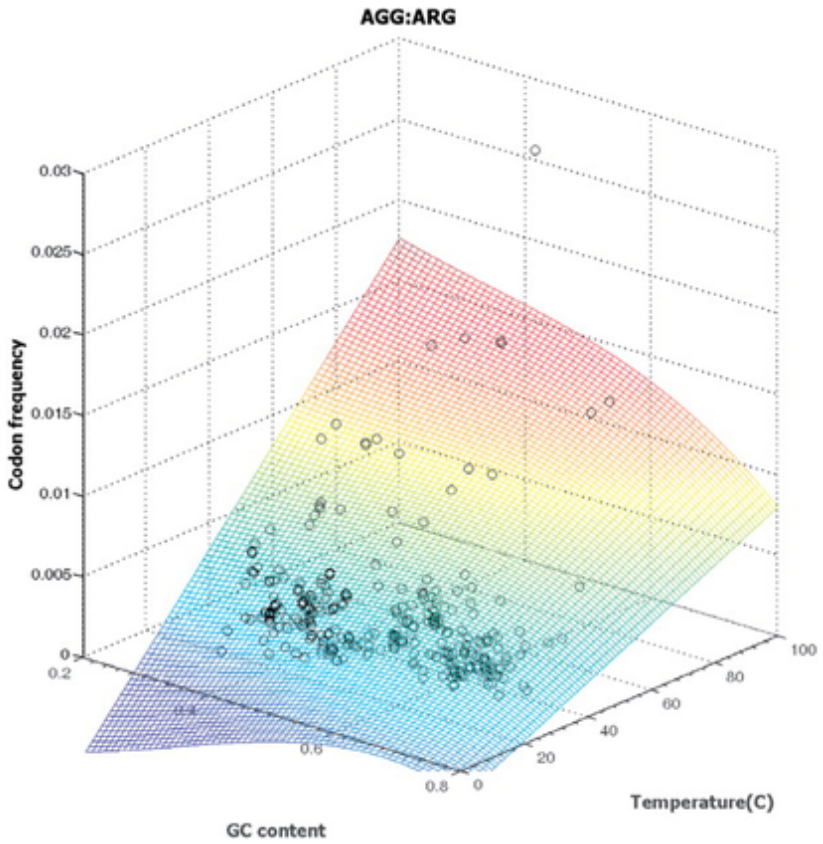
$$X = \begin{pmatrix} 1 & x_{11} & x_{21} & x_{11}x_{21} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & x_{1n}x_{2n} \end{pmatrix}$$

# Polynomial Regression. Illustration 1



# Polynomial Regression. Illustration 2



# Statistical Tests in Multiple Linear Regression

## I. Single parameter tests:

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H' : \beta_j \neq 0, \quad j \in \{0, k\}$$

Test statistic:

$$Z_j = \frac{\hat{\beta}_j}{S[\hat{\beta}_j]}, \quad Z_j | H_0 \sim T(n-2), \quad j \in \{0, k\}$$

## II. Test for significance of regression:

$$H_0 : \beta_1 = ... = \beta_k = 0 \quad \text{vs} \quad H' : \beta_1^2 + ... + \beta_k^2 > 0$$

Test statistic:

$$Z = \frac{R^2/k}{(1-R^2)/(n-k-1)}, \quad Z | H_0 \sim F(k, n-k-1)$$

where $R^2$ is the coefficient of determination:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{(n-k-1)s_e^2}{nD_Y^*}$$

# Importance of Assumptions in Regression Analysis

**Linear regression model:**

$$Y|x = x\beta + \varepsilon(x)$$

**Assumptions in linear regression analysis:**

- Linearity
- Exogeneity of regressors
- Homoscedasticity
- Independence of errors
- Normality
- Variability of regressors

If these assumptions are violated, then the statistical inference may be invalid

The assumptions should be tested in exploratory data analysis or after fitting the regression model to the data

# The Problem of Heteroscedasticity

Heteroscedasticity means that conditional variance of the outcome is not constant:

$$\mathrm{D}[Y|x] \neq const \Leftrightarrow \mathrm{D}[\varepsilon(x)] \neq const$$

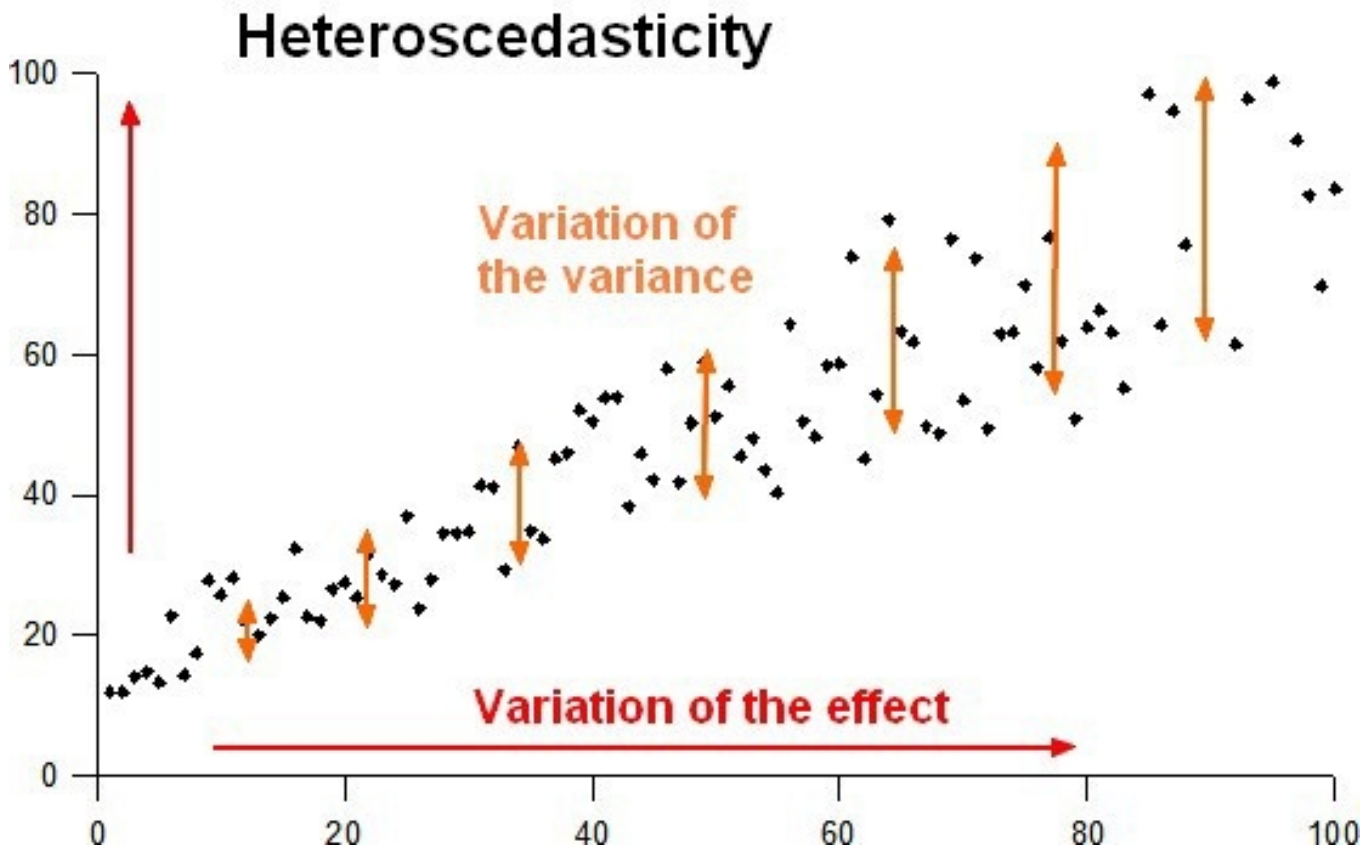
**Why heretoscedasticity is not desirable while building the regression model?**

- The OLS model is no longer efficient, i.e. it is not guaranteed to be the best unbiased linear estimator for your data
- The standard errors of the model's parameters become incorrect, hence, the confidence intervals, prediction intervals and test statistics become wrong

**Questions:**

- How to identify heteroscedasticity?
- How to fix heteroscedasticity?

# Heteroscedasticity. Illustration




Identification of heteroscedasticity is one of the tasks of the regression diagnostics

# Statistical Tests for Heteroscedasticity

Heteroscedasticity can be identified visually on scatter plots or by using statistical tests

**Statistical tests for heteroscedasticity:**

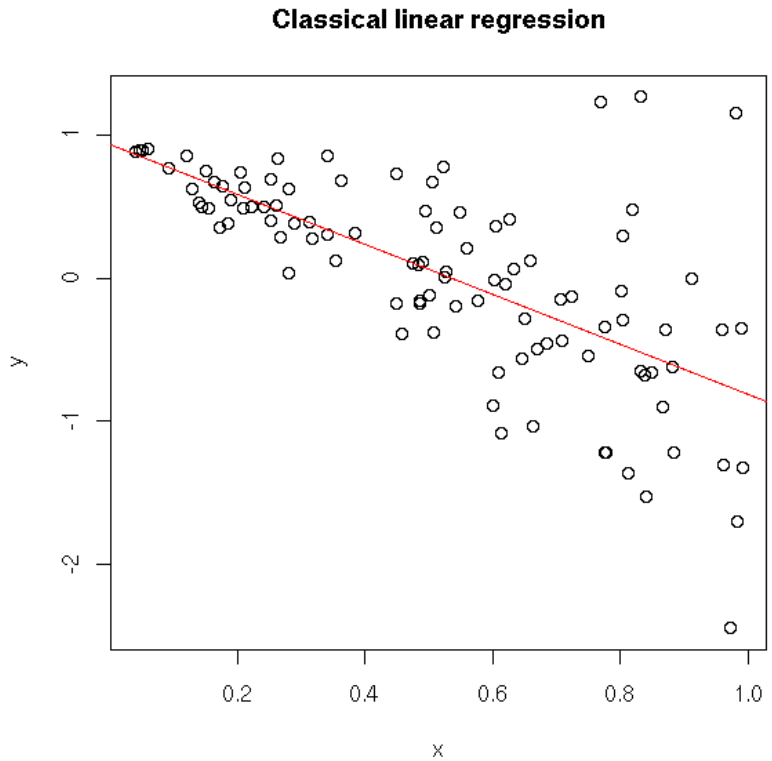
- White test
- Breush-Pagan test
- Park test
- Glejser test
- Goldfeld-Quandt test
- ...

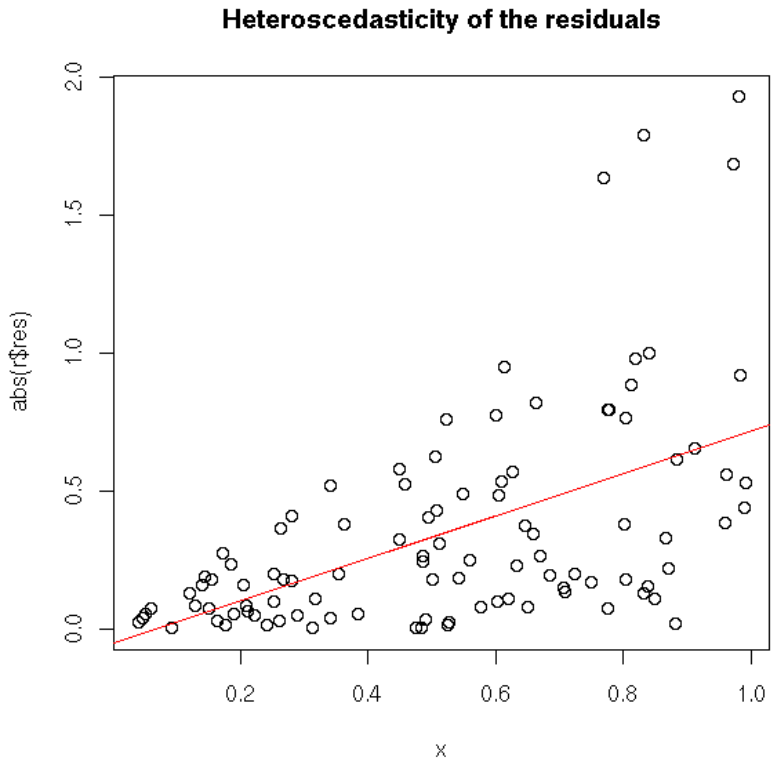
The statistical tests are usually applied to the regression errors $\varepsilon_1, ..., \varepsilon_n$, the null hypothesis:

$$H_0 : \mathrm{D}[\varepsilon_1] = ... = \mathrm{D}[\varepsilon_n] = \sigma^2$$

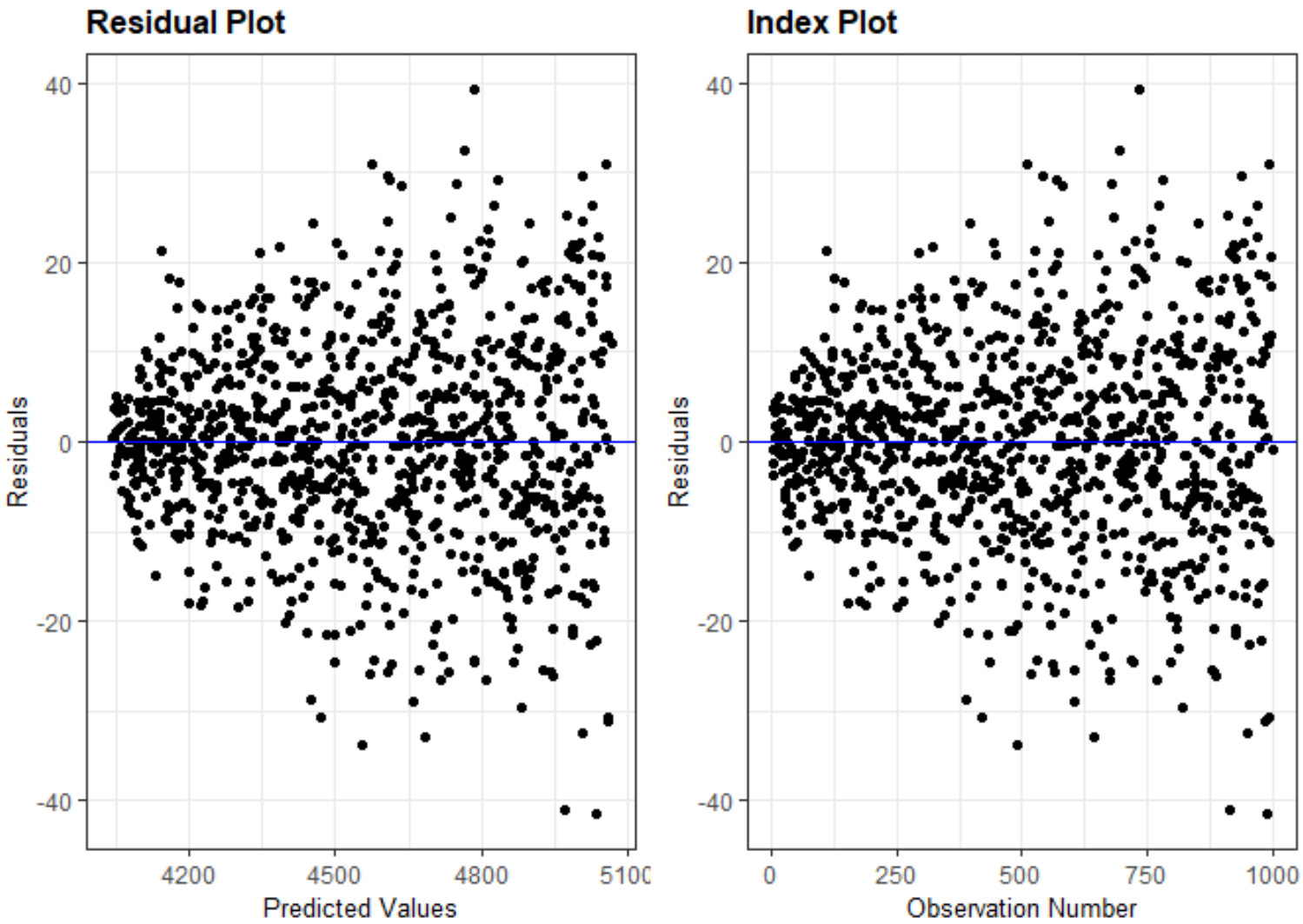
# Heteroscedasticity of Residuals. Illustration 1

**Idea:** If the sample $\varepsilon_1, ..., \varepsilon_n$ is heteroscedastic, then the squared sample $\varepsilon_1^2, ..., \varepsilon_n^2$ (or $|\varepsilon_1|, ..., |\varepsilon_n|$) has significant linear regression on predicted response $\hat{y}$ and/or explanatory variables $x_1, ..., x_k$



# Heteroscedasticity of Residuals. Illustration 2



# Approaches to Fix Heteroscedasticity

## How to fix heteroscedasticity?

- **Log-transformation of the response** $y$
  It will dampen down some of the heteroscedasticity, then build OLS regression of $\log y$

  Other transformations can also be applied
- **Weighted least squares approach to fit regression model**
  Include extra non-negative constants (weights), associated with each data point, into the fitting criterion
- **Heteroscedasticity-consistent standard error estimators**
  Provides better estimates of the covariance matrix of regression parameters

Hayes A.F., Cai L. Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation // Behavior research methods. 2007, 39(4), 709-722.

# Weighted Least Squares

**Linear regression model:**

$$Y|x = x\beta + \varepsilon(x)$$

Weighted Least Squares (WLS) criterion:

$$E(\beta) = \frac{1}{n}\sum_{i=1}^{n} w_i(y_i - x_i\beta)^2 \to \min_{\beta}$$

where $w_i$ is a scale factor (weight) of $i$-th observation, $i = \overline{1,n}$

**The system of normal equations in matrix form:**

$$X^T W X\beta = X^T W y$$

**The solution:**

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

where $W = diag(w_1, ..., w_n)$ is a diagonal weight matrix

**How to choose weights $w_1, ..., w_n$?**

## Transformation to Homoscedasticity

**Idea:** choose weights that should transform the response variances to a constant value

$$Y_i = x_i\beta + \varepsilon_i \quad \Rightarrow \quad \mathrm{D}[Y_i] = \mathrm{D}[x_i\beta + \varepsilon_i] = \mathrm{D}[\varepsilon_i] = \sigma_i^2, \quad i = \overline{1,n}$$

$$\frac{Y_i}{\sigma_i} = \frac{x_i}{\sigma_i}\beta + \frac{\varepsilon_i}{\sigma_i}$$

$$Y_i' = x_i'\beta + \varepsilon_i'$$

where

$$Y_i' = \frac{Y_i}{\sigma_i}, \quad x_i' = \frac{x_i}{\sigma_i}, \quad \varepsilon_i' = \frac{\varepsilon_i}{\sigma_i}$$

The transformed model is homoscedastic now:

$$Y_i' = x_i'\beta + \varepsilon_i' \quad \Rightarrow \quad \mathrm{D}[Y_i'] = \mathrm{D}[\varepsilon_i'] = 1, \quad i = \overline{1,n}$$

# Weights in WLS

Let's apply OLS to the transformed model:

$$\hat{\beta} = (X'^T X')^{-1} X'^T y'$$

where

$$X' = \Sigma^{-1} X, \quad y' = \Sigma^{-1} y$$

and $\Sigma = diag(\sigma_1, ..., \sigma_n)$

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

Hence, the weight matrix should be

$$W = \Sigma^{-1} \Sigma^{-1} = (\Sigma^{-1})^2$$

and the individual weights of observations should be equal to the reciprocals of residual variances:

$$w_i = \frac{1}{\sigma_i^2}, \quad i = \overline{1, n}$$

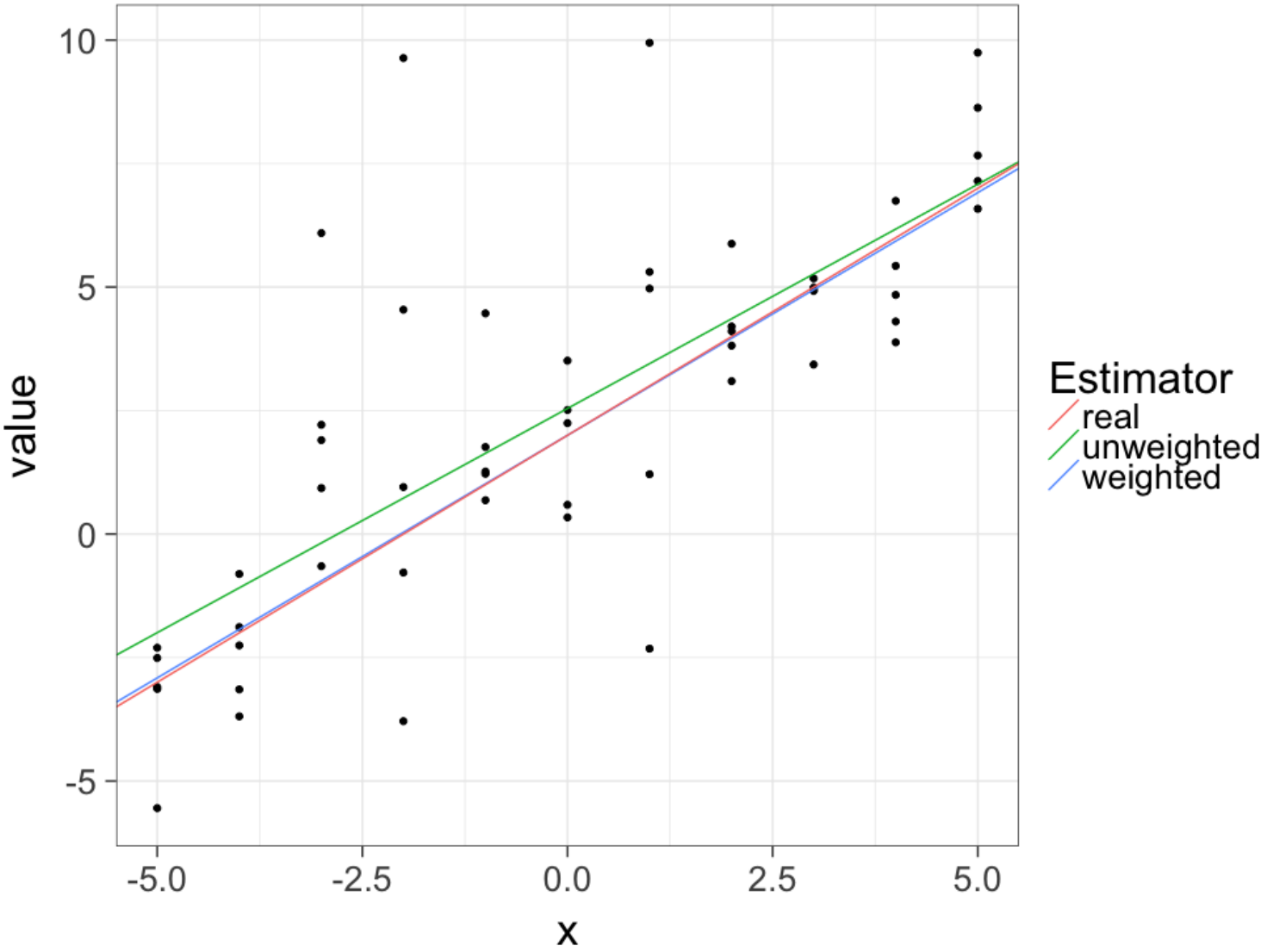
## Approaches to choose weights

In practice, the response variances $\sigma_1^2, ...\sigma_n^2$ are unknown. In this case:

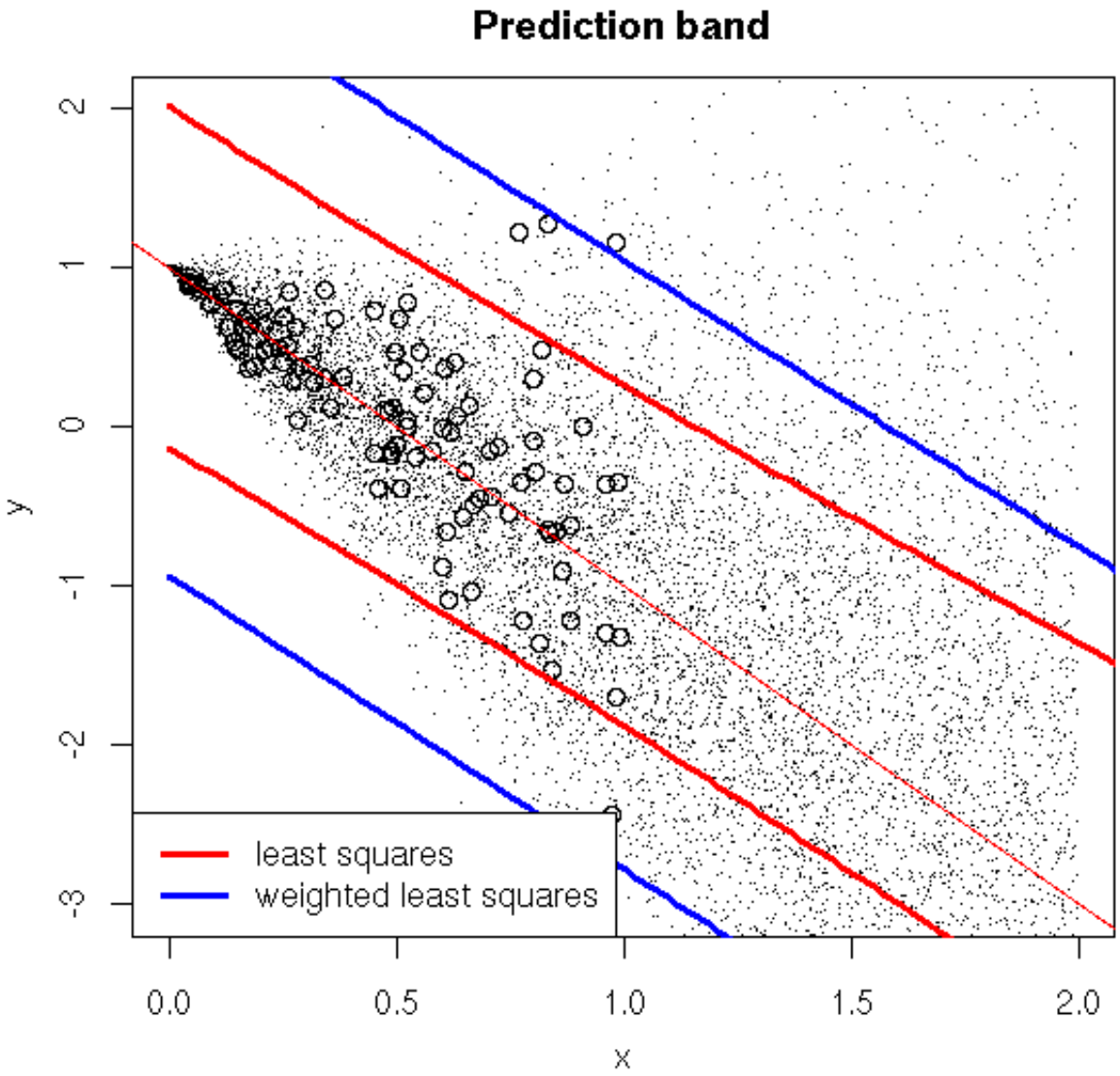
- Assume some model for weights based on visual analysis or using some extra knowledge, e.g. $w_i = \frac{1}{x_i}$, $i = \overline{1, n}$
- If your data occur only at discrete levels of $X$, estimate the variance directly at each level
- For continuous predictors, it's not recommended to use estimates of response variances due to sensitivity to outliers and randomness, especially for few data
- Build OLS regression to get the residuals $\hat{\varepsilon}_1, ..., \hat{\varepsilon}_n$, and then regress $\hat{\varepsilon}^2$ or $|\hat{\varepsilon}|$ on $x$ or $\hat{y}$. The predicted values of this regression model can be used instead of estimates of response variances

It's good practice to try various approaches and choose the best one based on, for instance, the distribution of residuals

# WLS. Illustration 1



# WLS. Illustration 2



# Weighted Least Squares. Notes

- In WLS, it's assumed that covariance matrix of errors is diagonal $cov[\varepsilon] = diag(\sigma_1^2, ..., \sigma_n^2)$
- To apply WLS, we need to know the weights $w_1, ..., w_n$
- If the weights are the reciprocals of residual variances, then WLS overcomes the issue of non-constant error variances $\sigma_1^2, ..., \sigma_n^2$
- Points with low variance will be given higher weights and points with higher variance are given lower weights
- WLS solution is the same as the OLS solution for the transformed model
- WLS gives us an easy way to remove some observations from a model by setting their weights equal to 0
- We can also downweight outliers or influential points to reduce their impact on the overall model

## Generalized Least Squares

In generalized least squares (GLS) it's assumed that residual covariance matrix $cov[\varepsilon]$ may be non-diagonal

**Linear regression model:**

$$Y|x = x\beta + \varepsilon(x)$$

**GLS criterion:**

$$E(\beta) = (y - X\beta)^T W (y - X\beta) \to \min_{\beta}$$

where $W$ is non-diagonal weight matrix. If $W = cov[\varepsilon]^{-1}$, then GLS solution is the same as the OLS solution of the transformed model:

$$Y_i' = x_i'\beta + \varepsilon_i', \quad i = \overline{1, n}$$

where

$$X' = \Sigma^{-1}X, \quad y' = \Sigma^{-1}y, \quad \varepsilon' = \Sigma^{-1}\varepsilon$$

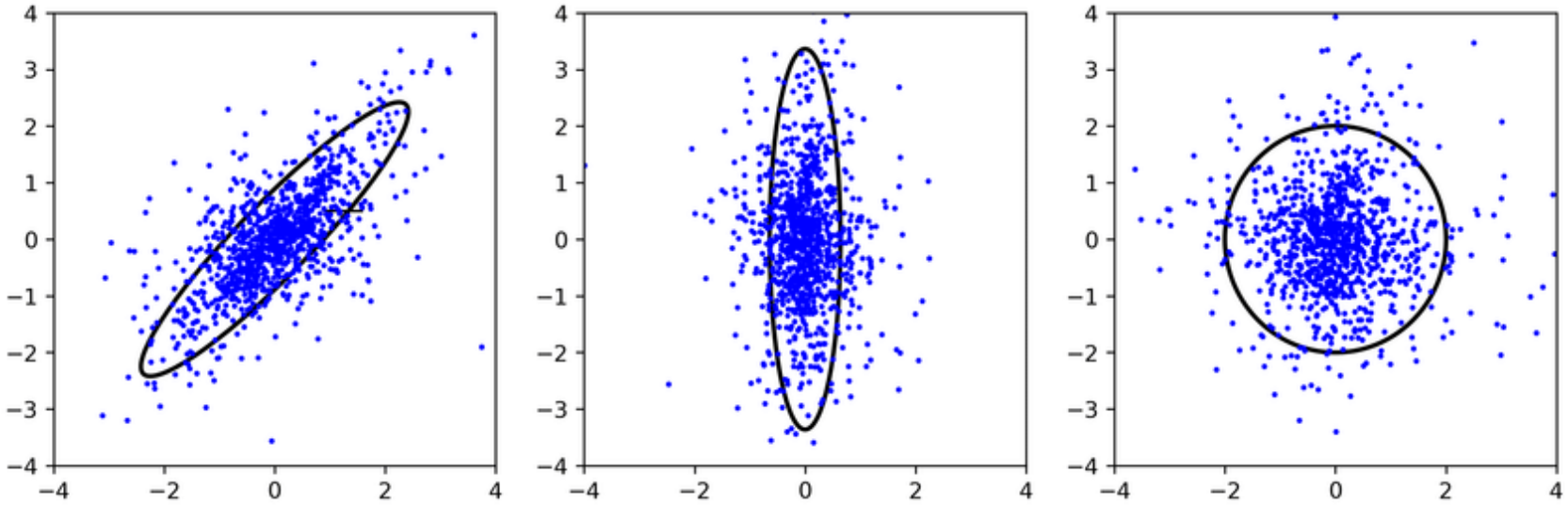
and $\Sigma$ is a square root of covariance matrix: $cov[\varepsilon] = \Sigma\Sigma^T$

# Whitening Transformation. Illustration

If the random vector $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$, $\mathrm{M}[\varepsilon] = 0$, has a covariance matrix $cov[\varepsilon] = \Sigma\Sigma^T$, then the transformation

$$\varepsilon' = \Sigma^{-1}\varepsilon$$

is called as whitening transformation. The transformed variables $\varepsilon'_1, ..., \varepsilon'_n$ are uncorrelated unit-variance random variables



GLS regression = error whitening + OLS regression

# Generalized Least Squares. Notes

- WLS is a particular case of GLS where the error covariance matrix $cov[\varepsilon]$ is assumed to be diagonal
- To apply GLS, we need to know all covariances $cov[\varepsilon_i, \varepsilon_j]$, $i, j = \overline{1, n}$
- In practice, the error covariance matrix $cov[\varepsilon]$ is unknown
- It's not recommended to use estimates of error covariance matrix due to sensitivity to outliers and randomness, the better choice is to assume some model for covariances with low number of parameters
- GLS solution is the same as the OLS solution of the transformed model (with whitened errors)
- The GLS estimator is unbiased, consistent, efficient and asymptotically normal since OLS is applied to data with homoscedastic uncorrelated errors and the Gauss–Markov theorem applies

# Robust Regression Techniques

In OLS, WLS and GLS it's assumed that the response errors follow a normal distribution, and that extreme values are rare. But in real data, extreme values called outliers do occur

Outliers have a large influence on the LS fit because squaring the residuals magnifies the effects of these extreme data points. To minimize their influence, robust regression techniques can be used:

- Least absolute residuals (LAR) method
- Iteratively reweighted least-squares (IRLS) algorithm
- RANSAC regression
  Tries to separate data into outliers and inliers and fits the model on the inliers
- Theil-Sen Regression
  Involves fitting multiple regression models on subsets of the training data and combining the coefficients
- ...

# Least Absolute Residuals Regression

**Idea:** use $L_1$-norm in loss function

**Linear regression model:**

$$Y|x = x\beta + \varepsilon(x)$$

Least Absolute Residuals (LAR) criterion:

$$E(\beta) = \frac{1}{n}\sum_{i=1}^{n} |y_i - x_i\beta| \to \min_{\beta}$$

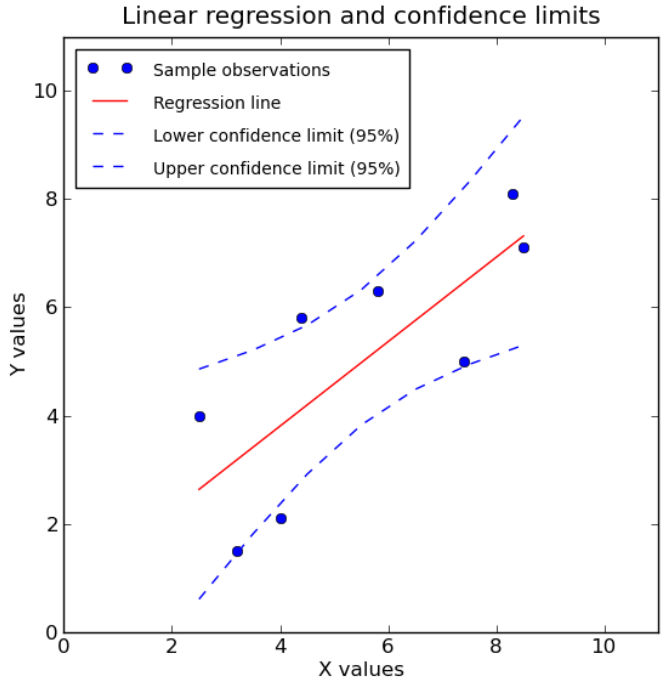
LAR regression does not have a closed-form solution, iterative training approach is required (based on simplex method, descend methods, etc.)

There are possibly multiple solutions of LAR regression, they depend on the initial point and search algorithm in the optimization procedure

# WLS for Robust Regression

**Idea:** minimize a weighted sum of squares, where the weight given to each data point depends on how far it is from the fitted line

**How to measure the distance of an observation to the fitted line?**




The euclidean distance is not the best metric, since the residual variance is not constant across the values of regressors

## Variances of Estimated Residuals

**The estimated residuals:**

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - x_i\hat{\beta}, \quad i = \overline{1,n}$$

where $\hat{\beta} = (X^T X)^{-1} X^T y$ is OLS-estimate of $\beta$

Under the assumptions of homoscedasticity and uncorrelatedness of errors:

$$\mathrm{D}[\varepsilon_i] = \mathrm{D}[Y|x_i] = \sigma^2, \quad cov[\varepsilon_i, \varepsilon_j] = 0 \quad \forall i,j = \overline{1,n}, \quad i \neq j$$

the variances of estimated residuals $\hat{\varepsilon} = (\hat{\varepsilon}_1, ..., \hat{\varepsilon}_n)^T$:

$$\mathrm{D}[\hat{\varepsilon}] = \mathrm{D}[Y - X(X^T X)^{-1} X^T Y] = (I - H)\mathrm{D}[Y] = (I - H)\sigma^2$$

$$\mathrm{D}[\hat{\varepsilon}_i] = \sigma^2(1 - h_i), \quad i = \overline{1,n}$$

where $H = X(X^T X)^{-1} X^T$ is hat matrix and $h_i$ is its $i$-th diagonal element called as leverage of observation $x_i$, $i = \overline{1,n}$

## Standardized Residuals

Under the assumption of normality

$$\hat{\varepsilon}_i \sim N\left(0, \sigma^2(1-h_i)\right), \quad i = \overline{1,n}$$

As soon as $\sigma$ is unknown, we use its unbiased estimate $S_e$:

$$S_e^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2$$

**The standardized residuals:**

$$t_i = \frac{\hat{\varepsilon}_i}{S[\hat{\varepsilon}_i]} = \frac{\hat{\varepsilon}_i}{S_e\sqrt{1-h_i}}, \quad t_i \sim T(n-k-1), \quad i = \overline{1,n}$$

The value of $t_i$ (also called as studentized residual) is a better metric of how far the $i$-th observation is from the fitted line

# Iteratively Reweighted Least-Squares (IRLS) Algorithm

- **Step 1.** Set weight matrix to be identical, $W = I$
- **Step 2.** Fit the model by WLS
- **Step 3.** Compute standartized residuals $t_1, ..., t_n$ or robust standartized residuals $t'_1, ..., t'_n$:

$$t'_i = \frac{\hat{\varepsilon}_i}{S'_e\sqrt{1-h_i}}, \quad i = \overline{1,n}$$

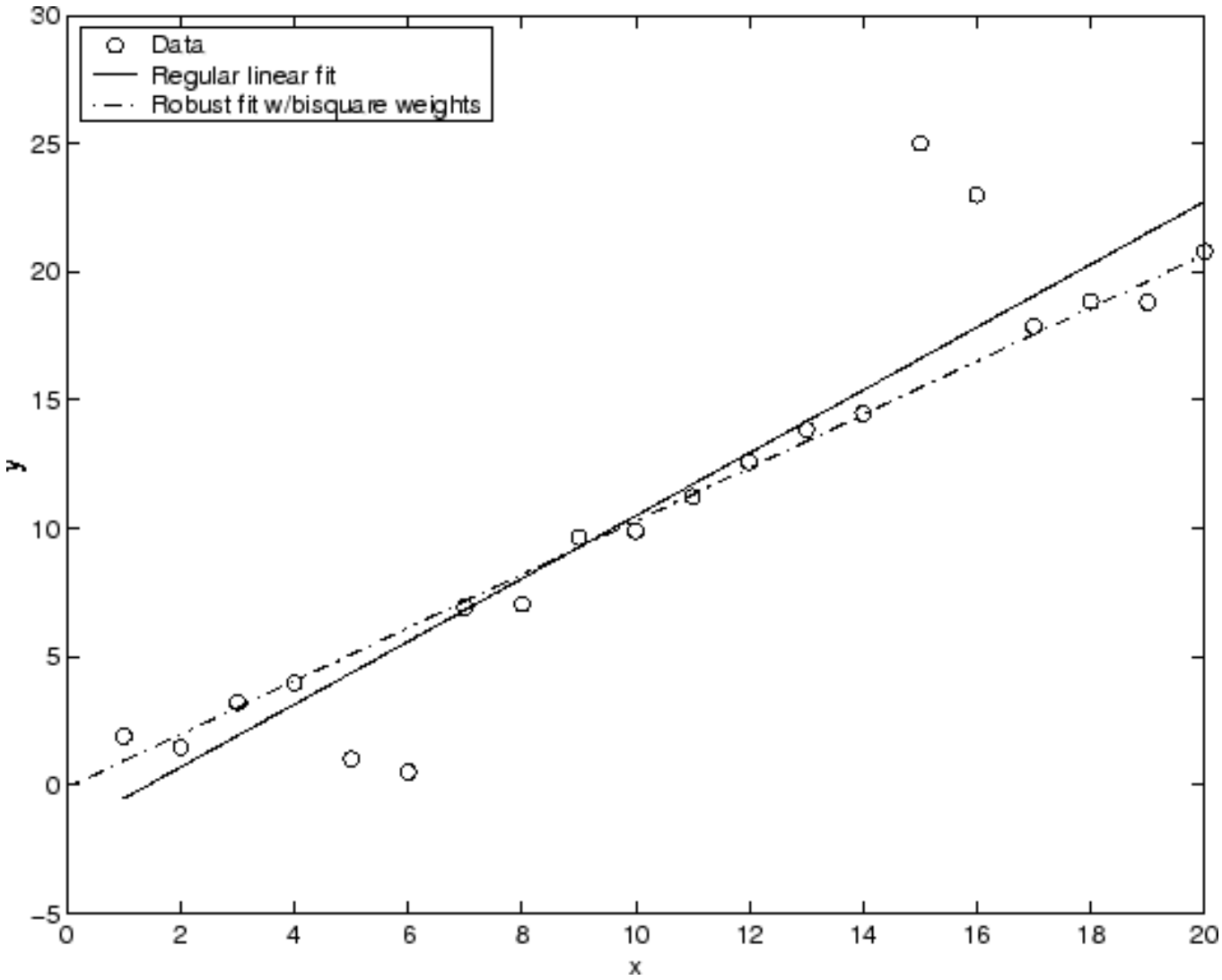
where $S'_e = \frac{med(|\hat{\varepsilon}_1|,...,|\hat{\varepsilon}_n|)}{0.675}$ is the robust standard deviation

- **Step 4.** Compute the robust weights as a function of $t'$. The bisquare weights:

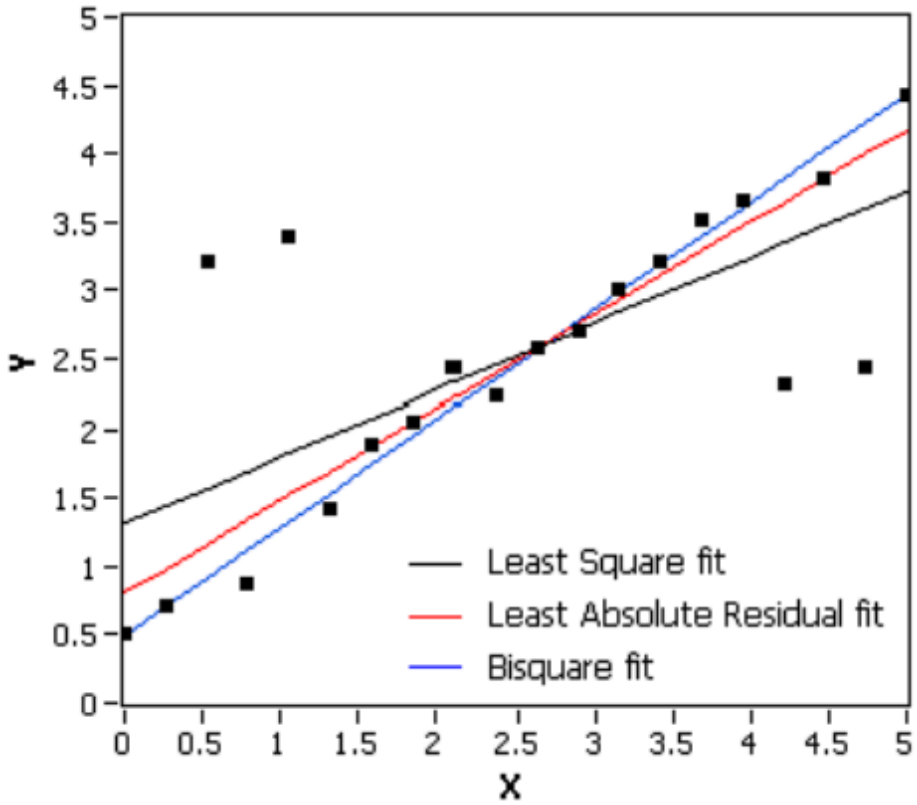
$$w_i = \begin{cases} \left(1-(t'_i/c_B)^2\right)^2, & |t'_i| < c_B, \\ 0, & otherwise \end{cases} \qquad (c_B = 4.685)$$

- **Step 5.** Go to step 2 until the fit converges

# Bisquare Weight Regression



# Robust Regression Techniques. Illustration




For most cases, the bisquare weight method is preferred over LAR because it simultaneously seeks to find a curve that fits the bulk of the data using the usual LS approach, and it minimizes the effect of outliers

# Non-linear Regression Model

**Non-linear regression model:**

$$Y|x = \varphi(x, \beta) + \varepsilon(x)$$

where $\beta = (\beta_0, ..., \beta_k)^T$ is a vector of parameters and $\varphi(x, \beta)$ is a non-linear function of $\beta_0, ..., \beta_k$

**Criterion:**

$$E(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \varphi(x_i, \beta))^2 \to \min_{\beta}$$

**Examples:**

$$\varphi(x, \beta) = \frac{\beta_0 x}{\beta_1 + x}$$

$$\varphi(x, \beta) = \beta_0 + \beta_1 x^{\beta_2}$$

$$\varphi(x, \beta) = \beta_0 \sin(\beta_1 + \beta_2 x)$$

## Non-linear Least Squares

Non-linear models are more difficult to fit than linear models because the coefficients $\beta_0, ..., \beta_k$ cannot be estimated using simple matrix techniques. Instead, iterative training techniques are used:

- **Step 1.** Assume initial estimate for each coefficient $\beta_0, ..., \beta_k$
- **Step 2.** Produce the fitted curve $\varphi(x, \hat{\beta})$ for the current set of coefficients $\hat{\beta}$
- **Step 3.** Adjust the coefficients $\hat{\beta}$ by an optimization algorithm and determine whether the fit improves
- **Step 4.** Iterate the process by returning to step 2 until the fit reaches the specified convergence criteria

The class of parametric non-linear regression functions must be specified for non-linear least squares

If you do not achieve a reasonable fit, you should experiment with different starting points, optimization algorithm and convergence criteria

## Transformation to Linearity

In particular cases the non-linear model can be transformed to linearity, for example:

$$\varphi(x, \beta_0, \beta_1) = \beta_0 e^{\beta_1 x}$$

$$\ln \varphi(x, \beta_0, \beta_1) = \ln \beta_0 + \beta_1 x$$

$$\varphi'(x, \beta_0', \beta_1') = \beta_0' + \beta_1' x$$

where $\varphi'(x, \beta_0, \beta_1) = \ln \varphi(x, \beta_0, \beta_1), \quad \beta_0' = \ln \beta_0, \quad \beta_1' = \beta_1$

It's a linear model and OLS can be used to estimate $\beta' = (\beta_0', \beta_1')^T$:

$$\beta' = (X^T X)^{-1} X^T y'$$

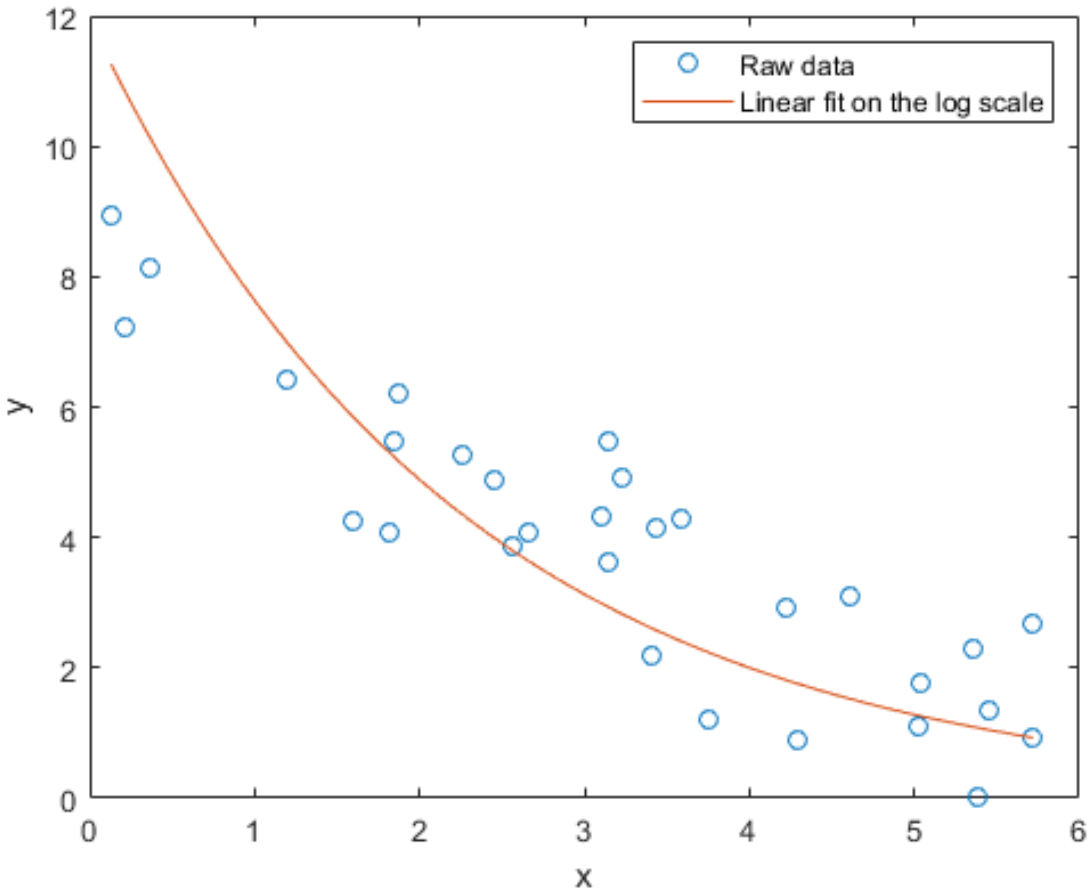
where $y' = (\ln y_1, ..., \ln y_n)^T$ and $X = \begin{pmatrix} 1 & x_1 \\ ... & ... \\ 1 & x_n \end{pmatrix}$

The original parameters $\beta_0 = e^{\beta_0'}, \quad \beta_1 = \beta_1'$

# Transformation to Linearity. Illustration 1

**Non-linear regression function:** $\varphi(x, \beta_0, \beta_1) = \beta_0 e^{\beta_1 x}$

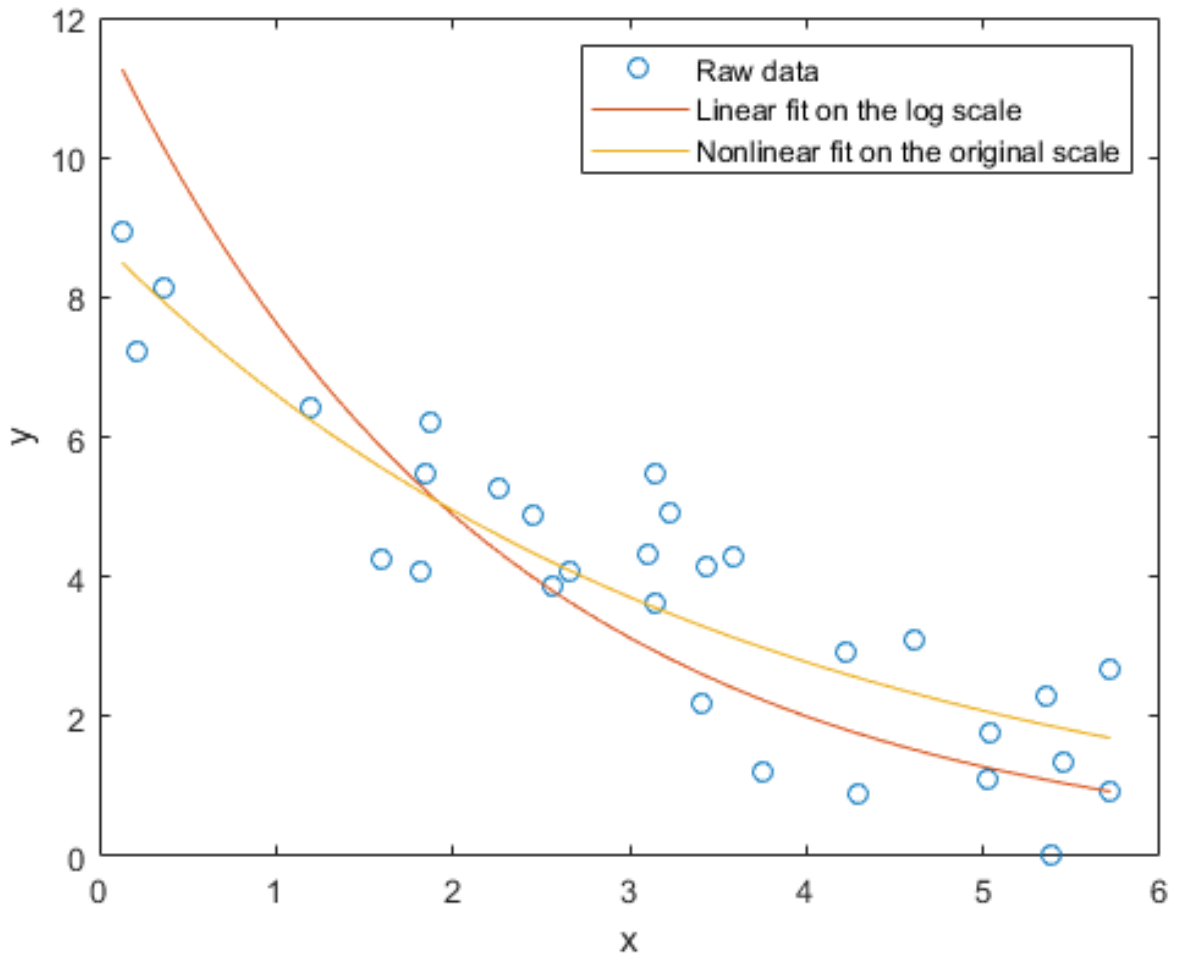
**Transformed to linearity:** $\ln \varphi(x, \beta_0, \beta_1) = \ln \beta_0 + \beta_1 x$




What is wrong with this regression?

# Transformation to Linearity. Illustration 2

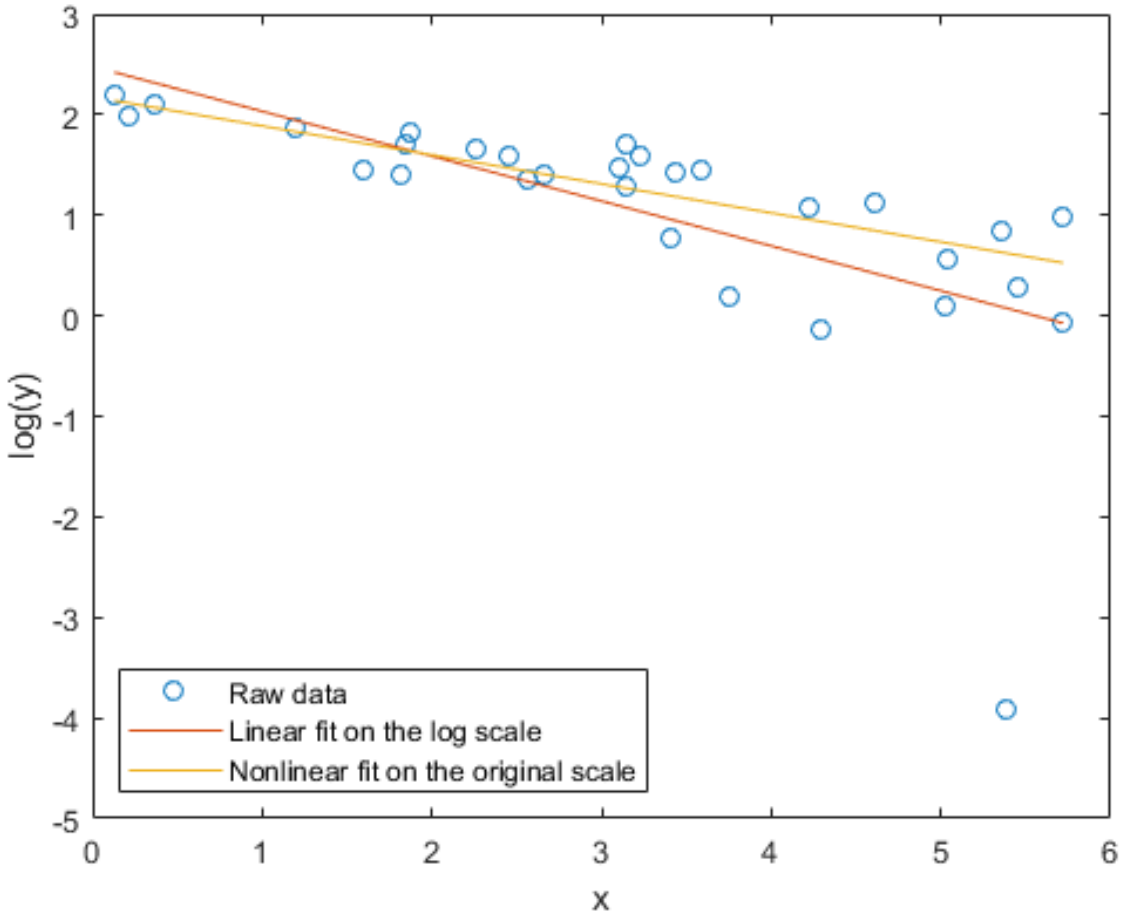
**Non-linear regression function:** $\varphi(x, \beta_0, \beta_1) = \beta_0 e^{\beta_1 x}$




Why the regressions are different?

# Transformation to Linearity. Illustration 3

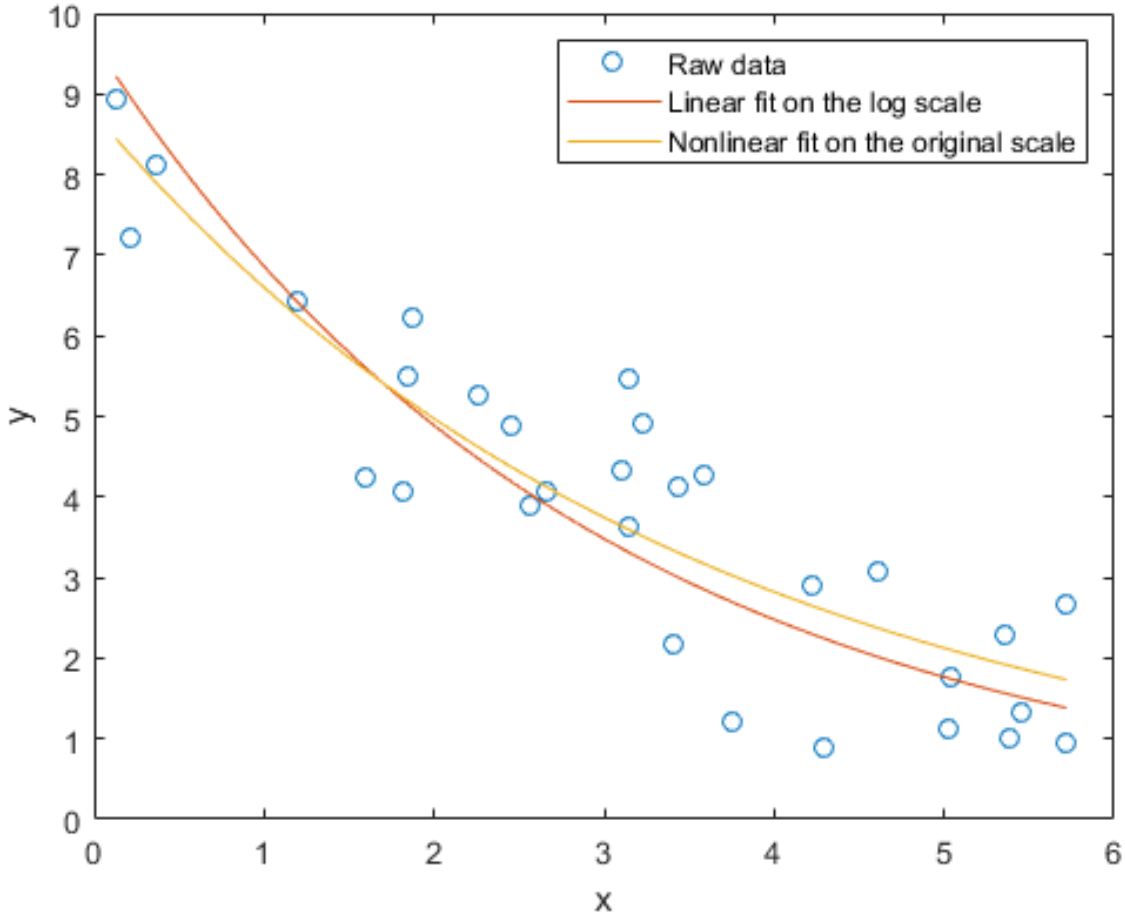
Symmetric measurement errors on the original scale have become asymmetric on the log scale




Linear fit on log scale is very much affected by an outlier

# Transformation to Linearity. Illustration 4

Let's remove the outlier




The outlier is removed, but the regressions are still different. Which one is "right"?

## Non-linear Regression or Transformation to Linearity?

**Non-linear model 1:** $Y|x = \beta_0 e^{\beta_1 x} + \varepsilon(x)$

$$\ln Y|x = \ln\left(\beta_0 e^{\beta_1 x} + \varepsilon(x)\right)$$

If the original noise $\varepsilon(x)$ is additive, the log-transformation is inappropriate

**Non-linear model 2:** $Y|x = \beta_0 e^{\beta_1 x}\varepsilon(x)$

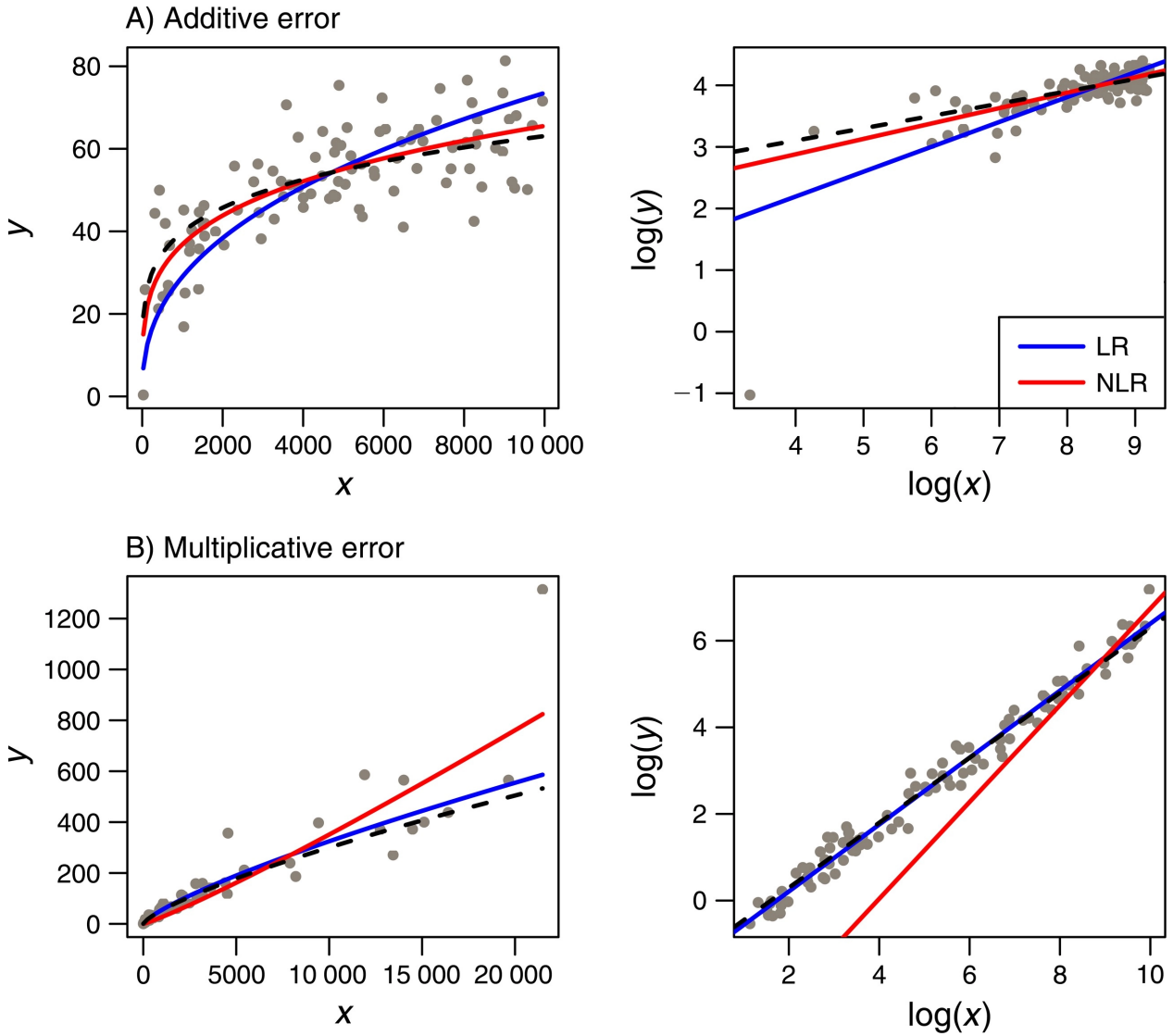
$$\ln Y|x = \beta_0' + \beta_1' x + \varepsilon'(x)$$

where $\beta_0' = \ln\beta_0, \quad \beta_1' = \beta_1, \quad \varepsilon'(x) = \ln\varepsilon(x)$

$$\varepsilon'(x) \sim N(0,\sigma^2) \Leftrightarrow \varepsilon(x) \sim Lognormal(0,\sigma^2)$$

If the original noise $\varepsilon(x)$ is multiplicative and log-normally distributed, then log-transformation results to a linear model with additive normal noise

# Non-linear Regression vs Transformation to Linearity. Illustration




(A) additive normal error, (B) multiplicative lognormal error

## Generalized Linear Models

Particular case of non-linear models is generalized linear models (GLM). The GLMs allow the linear model to be related to the response variable via a non-linear link function

**Linear models:**

- For each $x$, the response $Y|x$ has a normal distribution
- A coefficient vector $\beta$ defines a linear combination $x\beta$ of the predictors $x$
- The regression function is linear: $\mathrm{M}[Y|x] = x\beta$

**Generalized linear models:**

- For each $x$, the response $Y|x$ has a distribution that can be normal, binomial, Poisson, etc.
- A coefficient vector $\beta$ defines a linear combination $x\beta$ of the predictors $x$
- The transformed regression function is linear: $g(\mathrm{M}[Y|x]) = x\beta$, where $g(\mu)$ is the link function

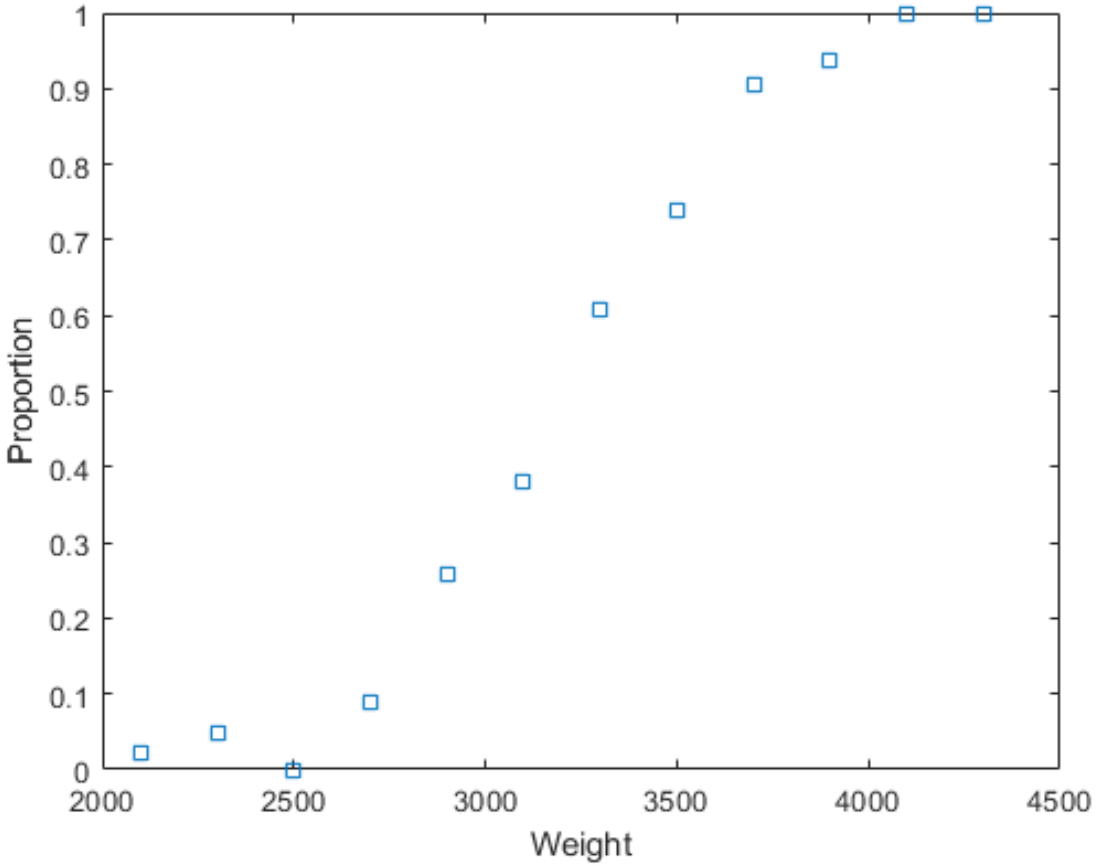
# GLM Fitting Pipeline

- **Step 1.** Prepare data
  Specify predictors $x$ and response variable $y$
- **Step 2.** Specify distribution of response variable $Y|x$
  Binomial, Poisson, gamma, etc.
- **Step 3.** Specify link function $f(\mu)$
  Logit, probit, log-log, etc.
- **Step 4.** Specify the linear model
  With or without intercept, select features, etc.
- **Step 5.** Choose fitting method
  OLS, WLS, etc.
- **Step 6.** Fit model to data
- **Step 7.** Examine quality of the fitted model
  Analysis of residuals, statistical tests, cross-validation

# GLM Example. Illustration 1

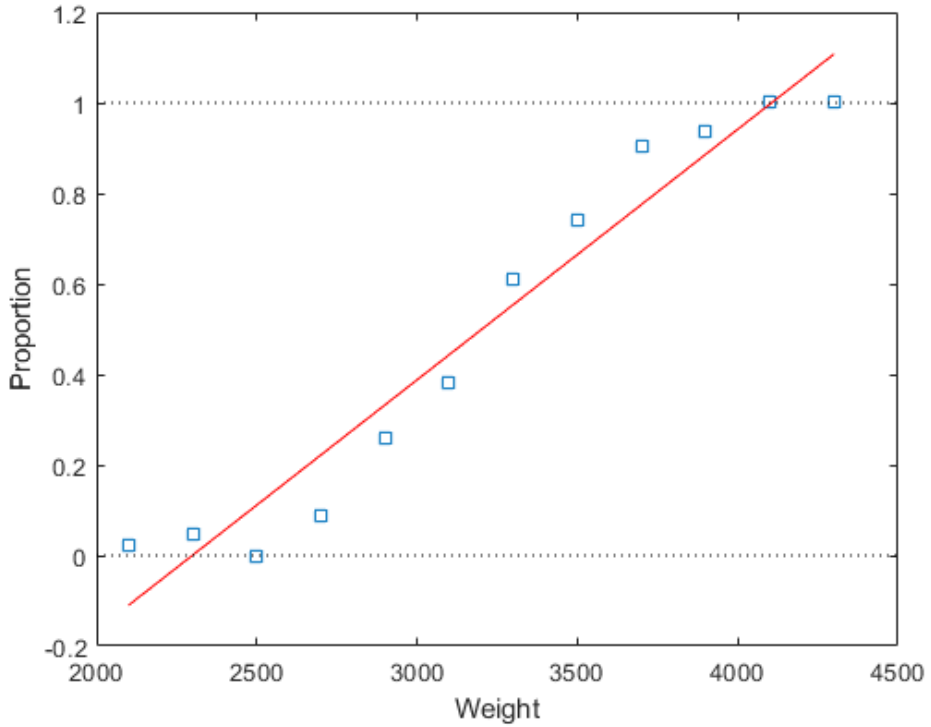
The predictor $x$ is car's weight, and the response variable $y$ is the proportion of cars of various weights that fail a mileage test

**The scatterplot $y$ vs $x$**



# GLM Example. Illustration 2

Let's build the linear model: $\varphi(x) = \beta_0 + \beta_1 x$

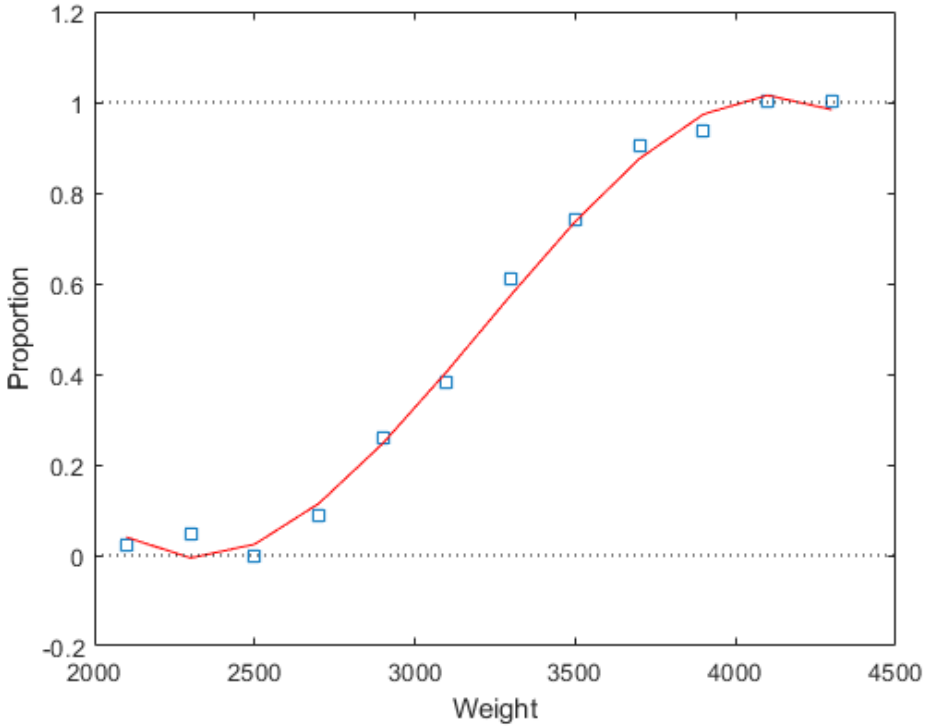



**Problems:**

- The line predicts proportions less than 0 and greater than 1
- The proportions are not normally distributed, since they are necessarily bounded. This violates one of the assumptions required for fitting a simple linear regression model

# GLM Example. Illustration 3

Let's build the linear model: $\varphi(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$




**Problems:**

- The fitted proportion starts to decrease as weight goes above 4000; in fact it will become negative for larger weight values
- The assumption of a normal distribution is still violated

# GLM Example. Choosing the Link Function

It's reasonable to assume that the failure counts came from a binomial distribution, with a probability parameter $p$ that increases with weight

So, the distribution $Y|x$ should be binomial (up to a multiplier)

Let's build GLM: $g(\varphi(x)) = \beta_0 + \beta_1 x$ with logit link function

$$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right), \quad 0 < \mu < 1$$

Logit function limits the predicted proportions to the range $(0, 1)$ and it's appropriate for the binomial distribution of the responses

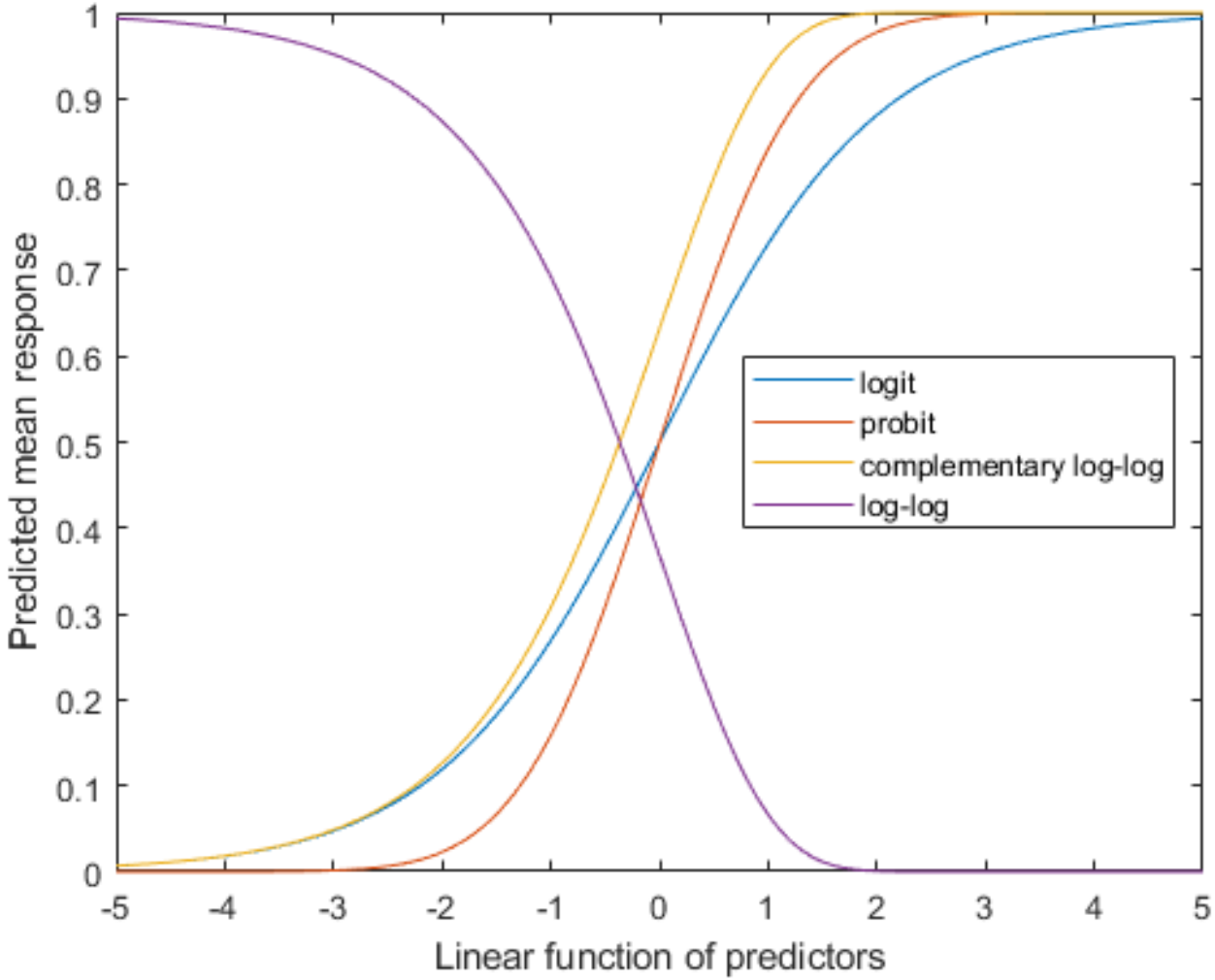
Inverse logit function is a logistic function: $g^{-1}(z) = \frac{1}{1+e^{-z}}$

**The regression function:**

$$\varphi(x) = g^{-1}(\beta_0 + \beta_1 x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x)}}$$

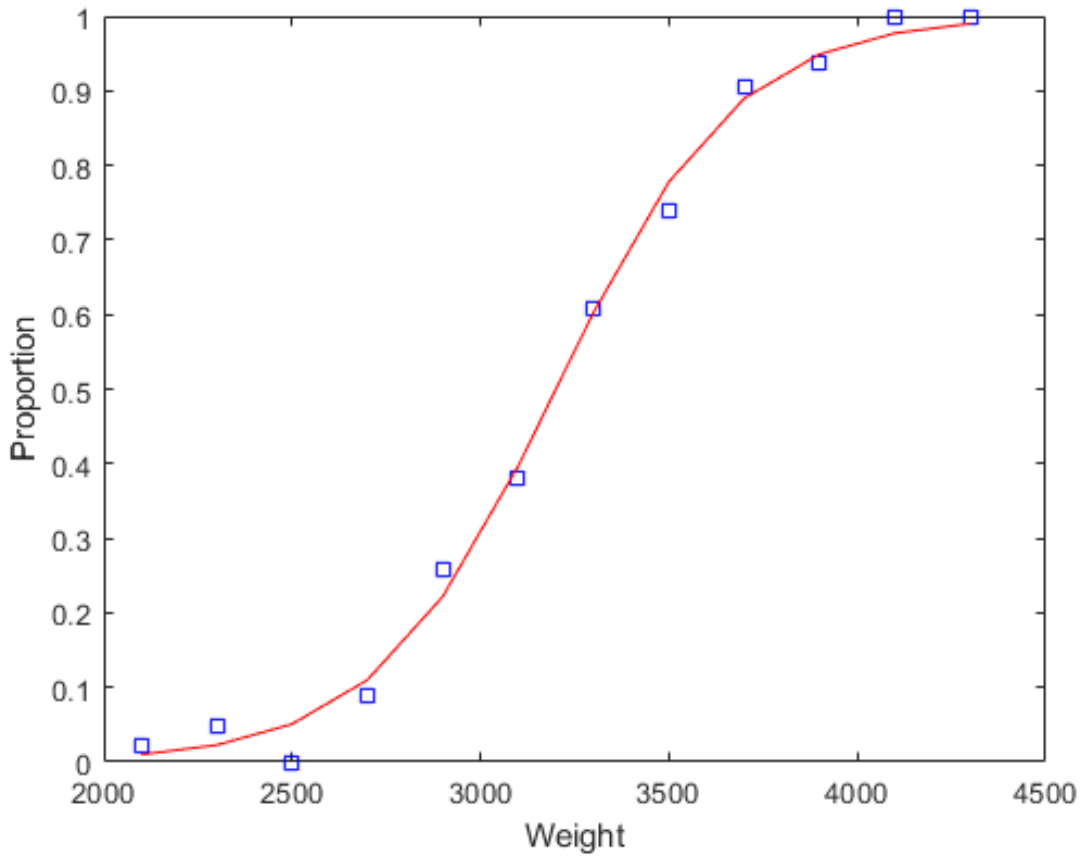
# GLM Example. Choosing the Link Function

Other link functions can also be choosen (e.g., probit)



# GLM Example. Illustration 4

GLM: $g(\varphi(x)) = \beta_0 + \beta_1 x$, where $g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$ and $Y|x$ has binomial distribution




The fitted proportions asymptote to zero and one as weight becomes small or large

## Neural Network Regression

One more particular case of non-linear models is neural network regression model:

$$Y|x = \varphi(x, \beta) + \varepsilon(x)$$

where $\beta = (\beta_0, ..., \beta_k)^T$ is a vector of parameters and $\varphi(x, \beta)$ is a neural network function, non-linear by $x$ and parameters $\beta_0, ..., \beta_k$

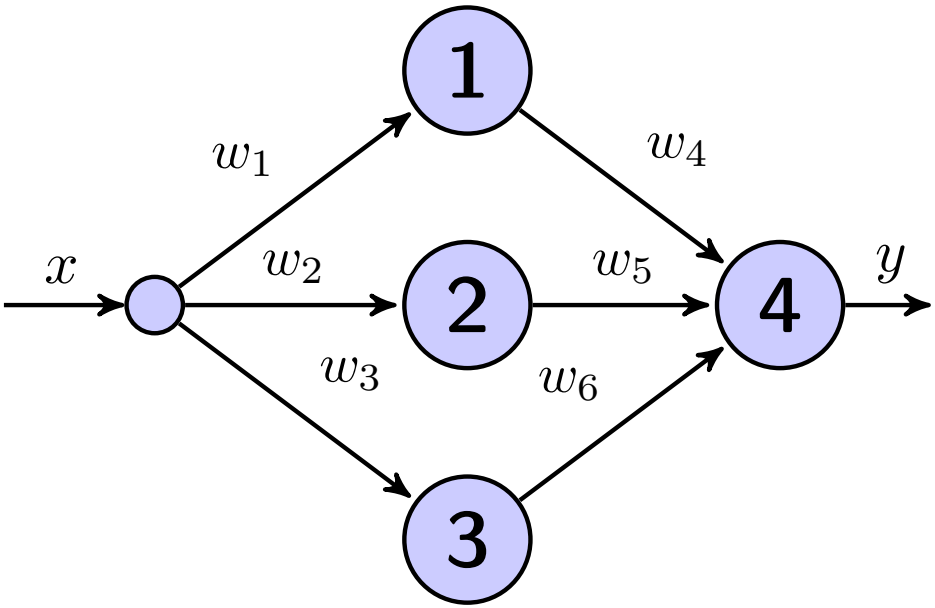
The function $\varphi(x, \beta)$ has a specific form, it's a multiple composition of simple non-linearities (like logistic functions, gaussians etc.)

The number $k$ of parameters to estimate can be very high, up to thousands or even millions

**OLS criterion:**

$$E(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \varphi(x_i, \beta))^2 \to \min_{\beta}$$

# Neural Network Regression. Example




**Model of the neuron:**

$$y = f\left(\sum_{j=1}^{M} w_j x_j\right)$$

$x_1, ..., x_M$ are neuron's inputs

$y$ is neuron's output

$f(\cdot)$ is neuron's transfer function

**The regression function:**

$$\varphi(x) = f_4\left(w_4 f_1(w_1 x) + w_5 f_2(w_2 x) + w_6 f_3(w_3 x)\right)$$

**The OLS training criterion:**

$$E(w_1, ..., w_6) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \varphi(x_i, w_1, ..., w_6))^2 \to \min_{w_1,...,w_6}$$

# Summary

**Types of regression models:**

- Simple linear regression model
- Multiple linear regression model
  Polynomial regression, exponential regression, etc.
- Non-linear regression models
  - Generalized linear models (GLM)
  - Neural network regression models
  - ...

**Types of fitting techniques:**

- Ordinary least squares (OLS)
- Weighted least squares (WLS)
- Generalized least squares (GLS)
- Robust fitting techniques (LAR, IRLS, etc.)
- Non-linear least squares
- ...