

18.650 – Fundamentals of Statistics

6. Linear regression

Goals

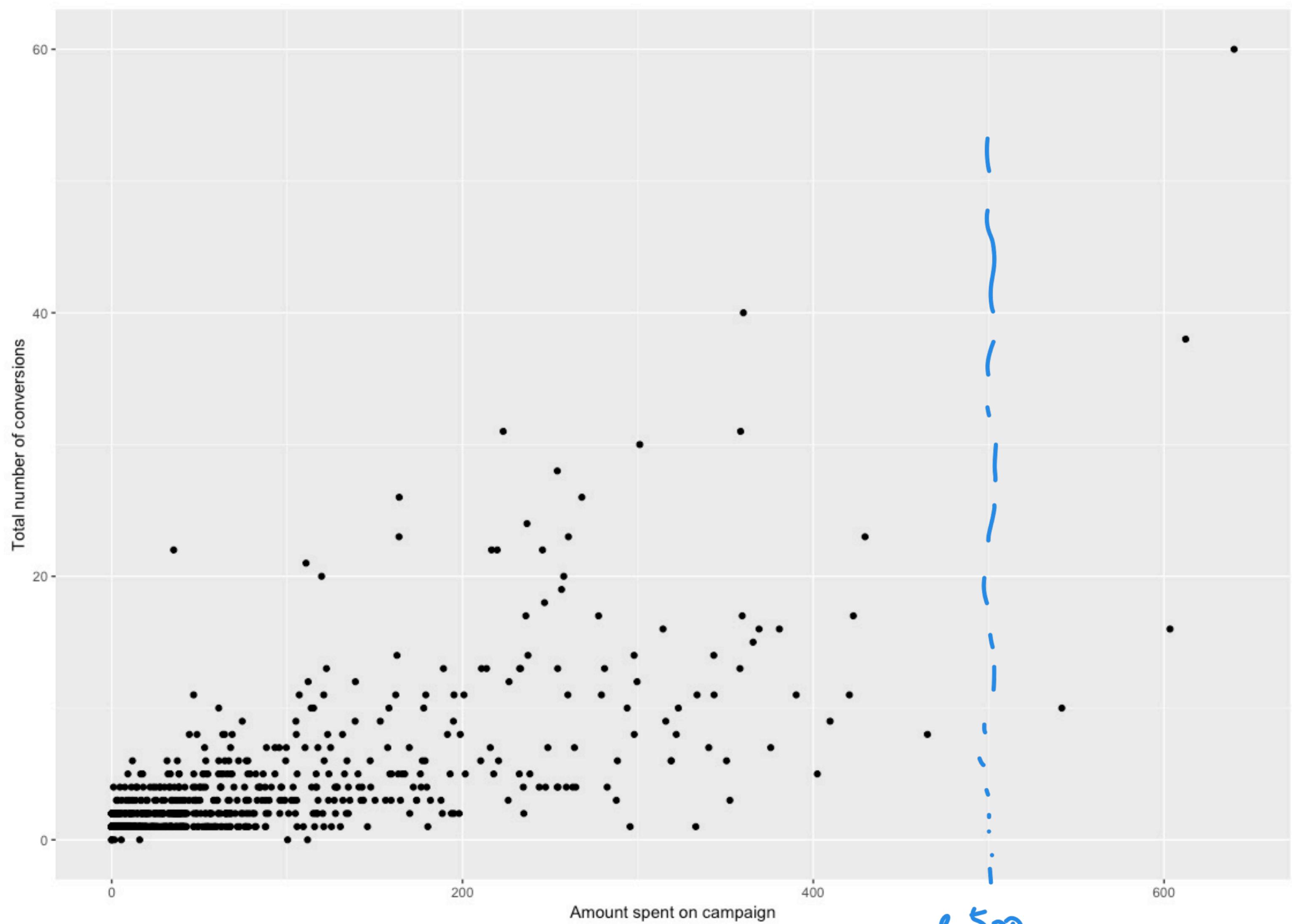
Consider two random variables X and Y . For example,

1. X is the amount of \$ spent on Facebook ads and Y is the total conversion rate
2. X is the age of the person and Y is the number of clicks

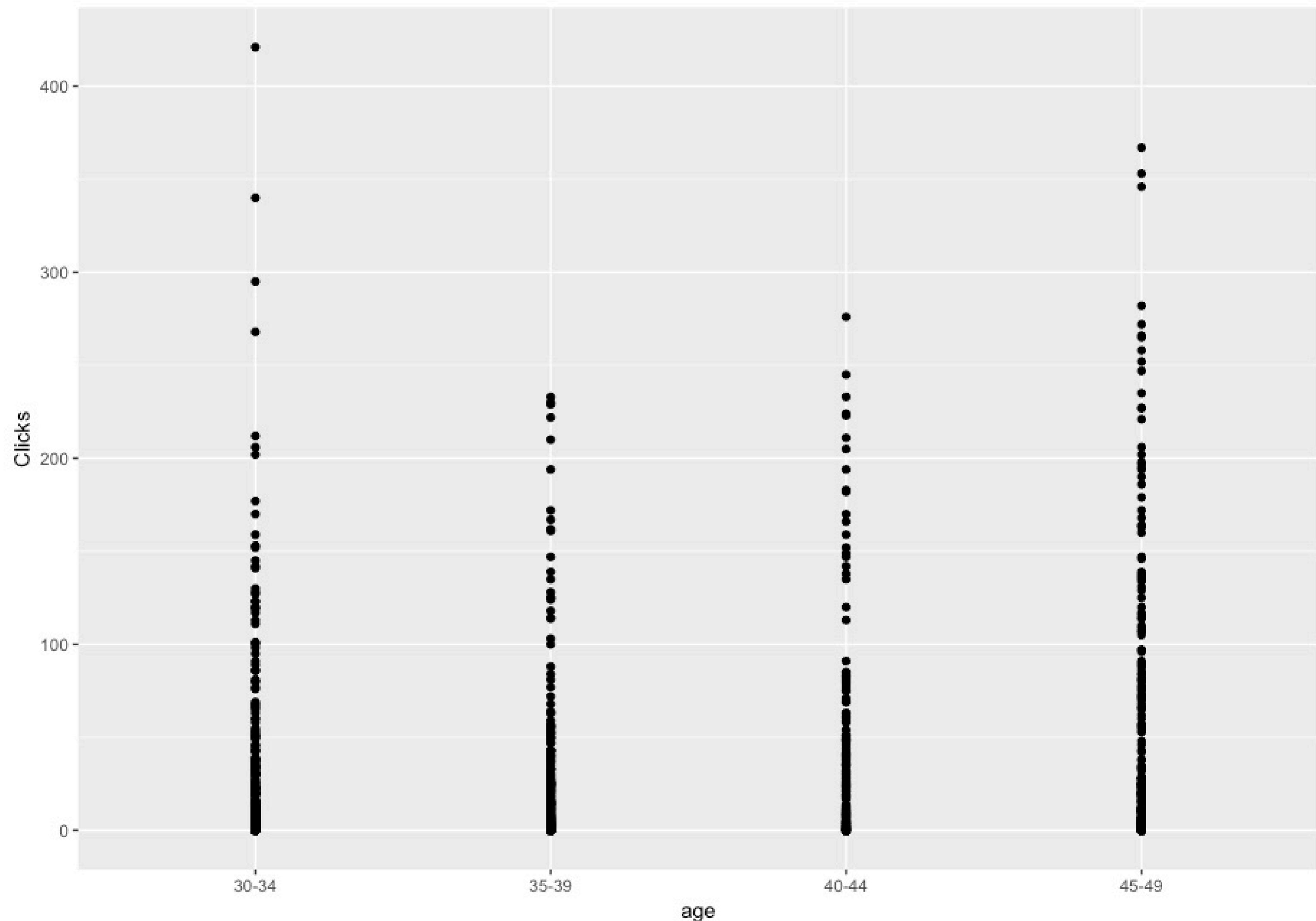
Given two random variables (X, Y) , we can ask the following questions:

- ▶ How to predict Y from X ?
- ▶ Error bars around this prediction?
- ▶ How much more conversions Y for an additional dollar?
- ▶ Does the number of clicks even depend on age?
- ▶ What if X is a random vector? For example, $X = (X_1, X_2)$ where X_1 is the amount of \$ spent on Facebook ads and X_2 is the duration in days of the campaign.

Conversions vs. amount spent



Clicks vs. age



Modeling assumptions

$(X_i, Y_i), i = 1, \dots, n$ are i.i.d from some **unknown joint distribution** \mathbb{P} .

\mathbb{P} can be described **entirely** by (assuming all exist)

- ▶ Either a joint PDF $h(x, y)$
- ▶ The marginal density of X $h(x) = \int h(x, y) dy$ and the **conditional density**

$$h(y|x) = \frac{h(x, y)}{h(x)}$$

$h(y|x)$ answers all our questions. It contains all the information about Y given X .

Partial modeling

We can also describe the distribution only *partially*, e.g., using

- ▶ The expectation of Y : $E[Y]$
- ▶ The conditional expectation of Y given $X = x$: $E[Y|X=x]$
The function

$$x \mapsto f(x) := \mathbb{E}[Y|X = x] = \int y h(y|x) dy$$

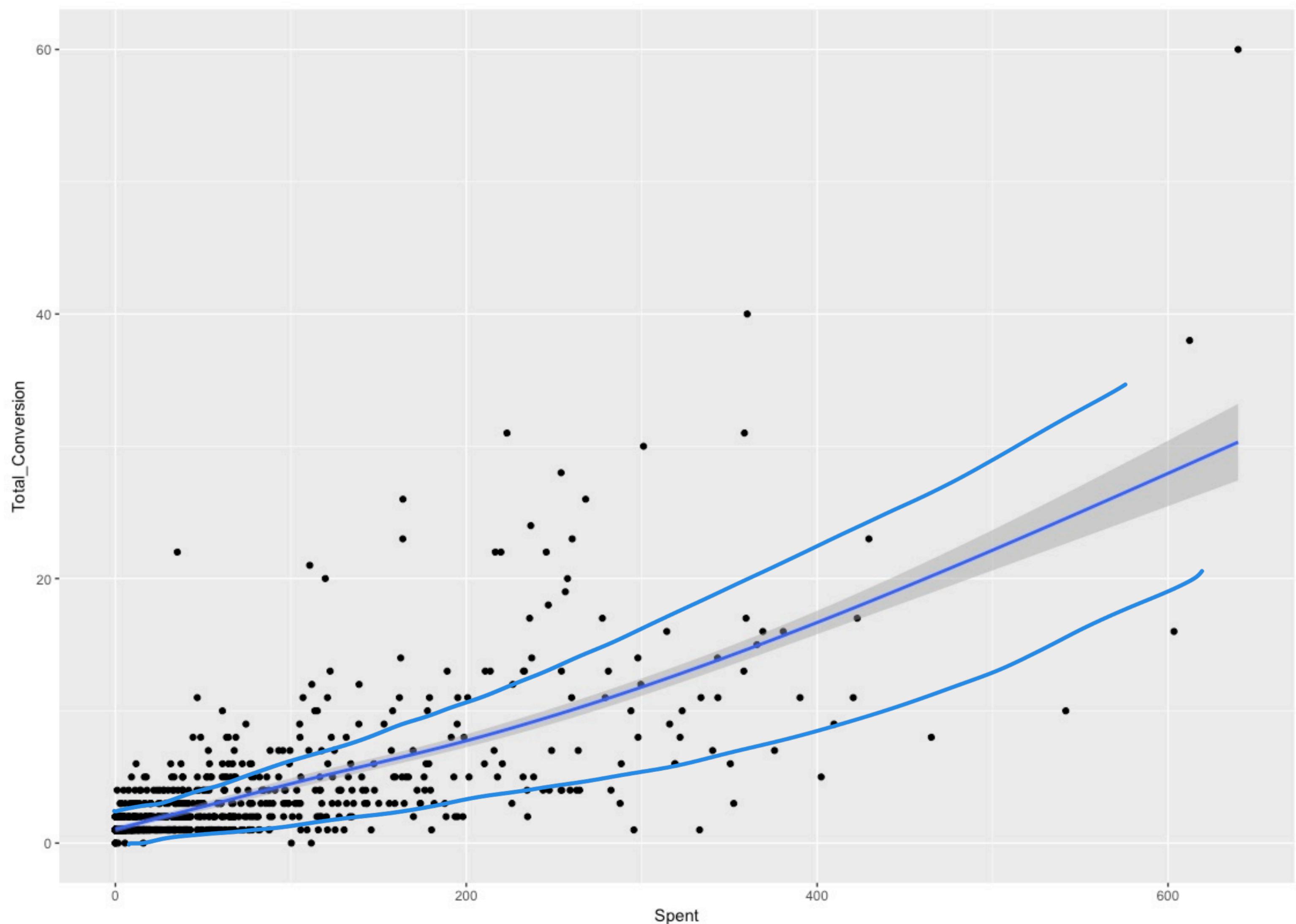
is called *regression function*.

- ▶ Other possibilities:
 - ▶ The conditional median: $m(x)$ such that

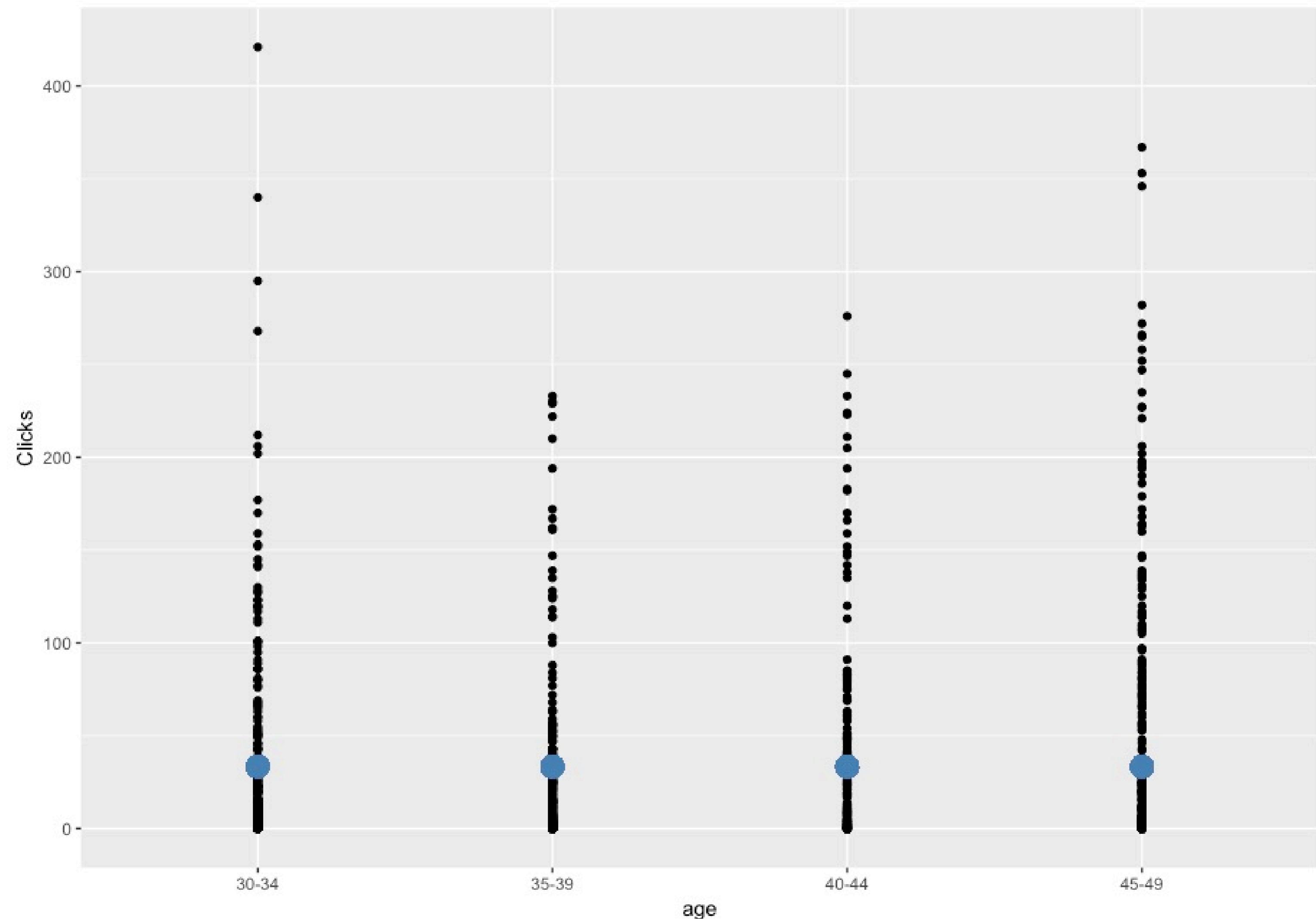
$$\int_{-\infty}^{m(x)} h(y|x) dy = \frac{1}{2}$$

- ▶ Conditional *quantiles*
- ▶ Conditional variance (not informative about location)

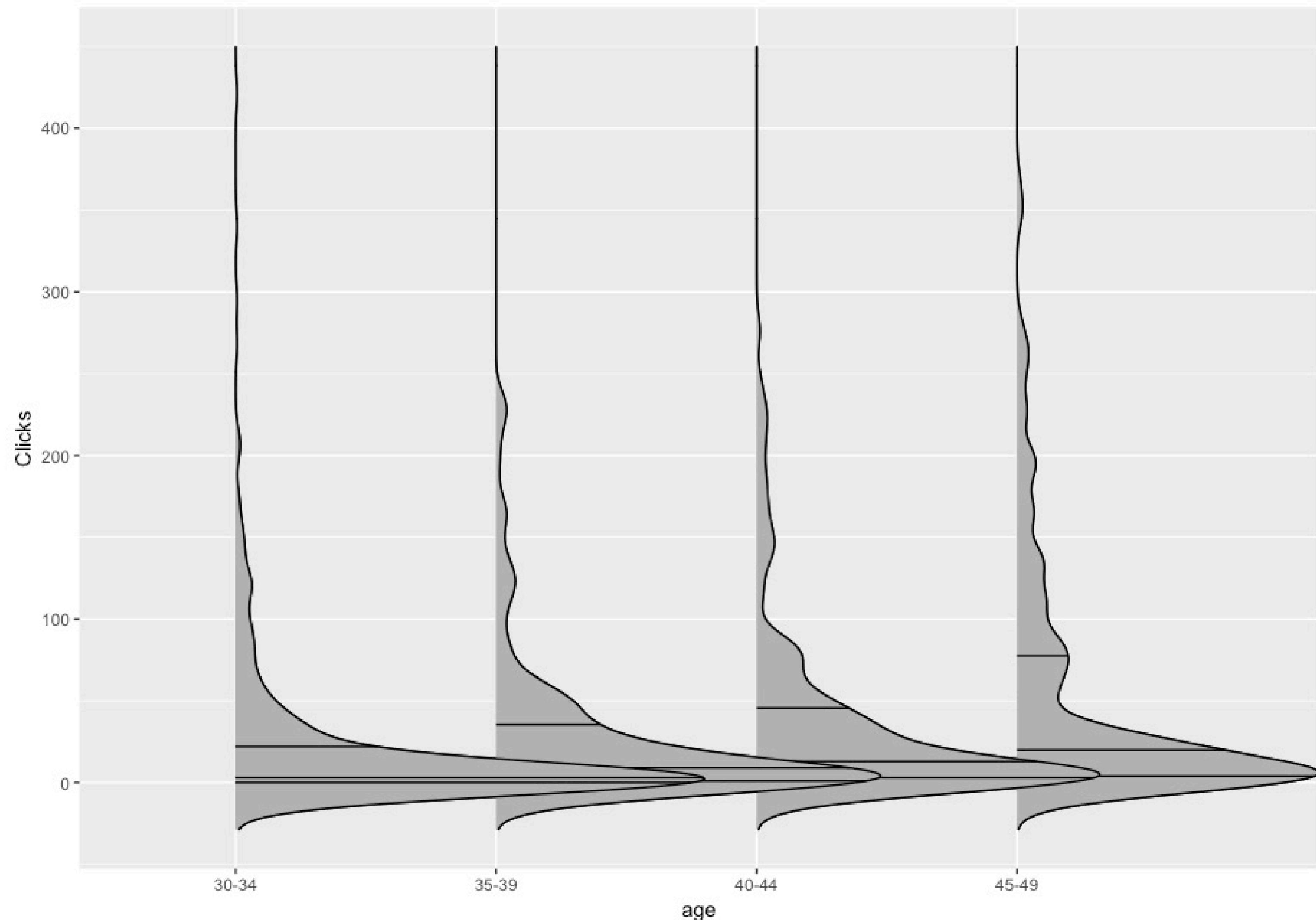
Conditional expectation and standard deviation



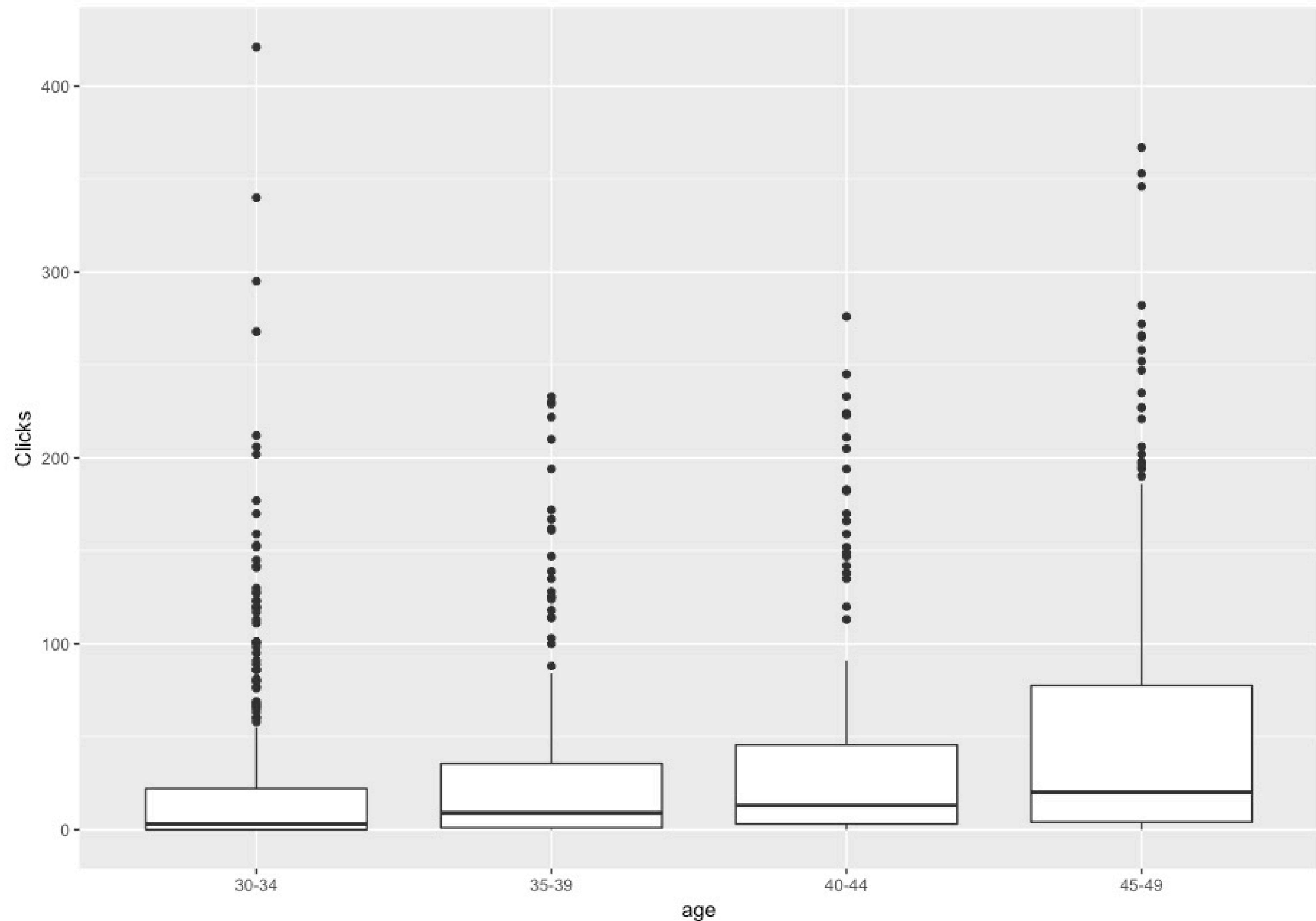
Conditional expectation



Conditional density and conditional quantiles



Conditional distribution: boxplots



Linear regression

We first focus on modeling the regression function

$$f(x) = E[Y | X=x]$$

- ▶ Too many possible regression functions f (nonparametric)
- ▶ Useful to restrict to **simple** functions that are described by a few parameters
- ▶ Simplest:

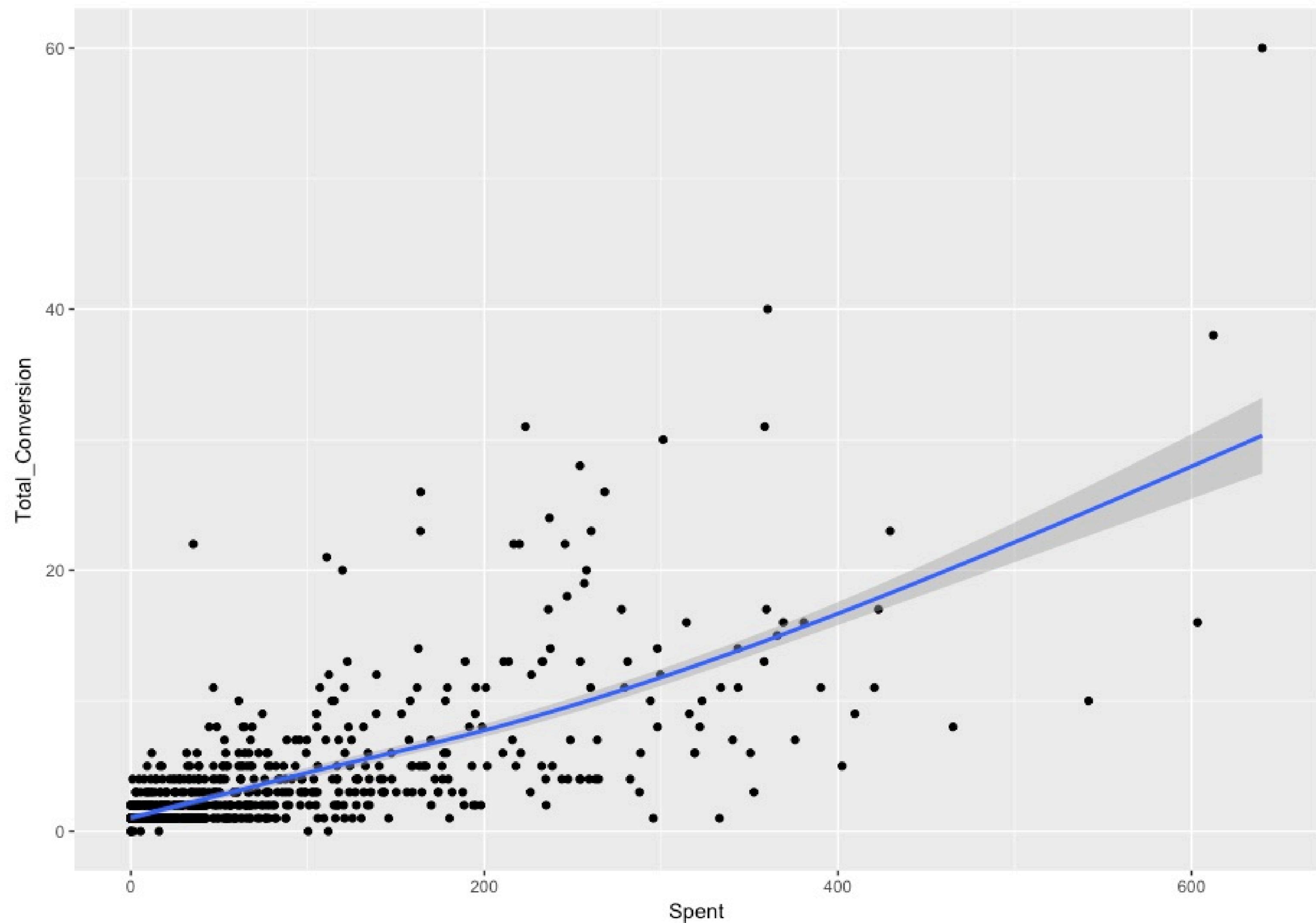
$$f(x) = a + bx$$

linear (or affine) function

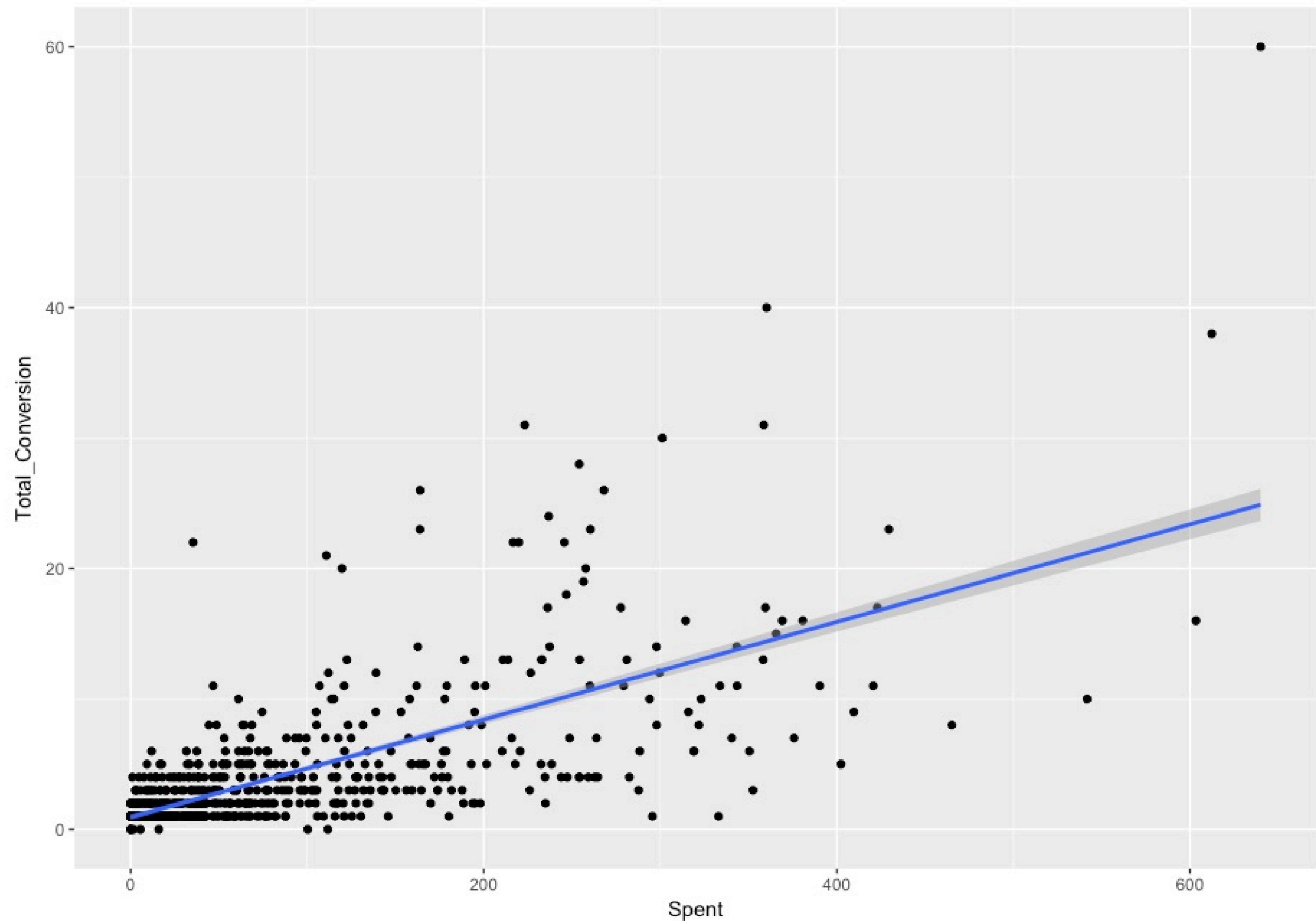
Under this assumption, we talk about

linear regression

Nonparametric regression



Linear regression



Probabilistic analysis

- ▶ Let X and Y be two real r.v. (not necessarily independent) with two moments and such that $\text{var}(X) > 0$.

- ▶ The **theoretical linear regression** of Y on X is the line $x \mapsto \underbrace{a^*}_{\text{ }} + \underbrace{b^*}_{} x$ where

$$(a^*, b^*) = \underset{(a,b) \in \mathbb{R}^2}{\operatorname{argmin}} \mathbb{E} \left[(Y - \underbrace{a}_{\text{ }} - \underbrace{bX}_{\text{ }})^2 \right]$$

- ▶ Setting partial derivatives to zero gives

$$\triangleright b^* = \frac{\text{cov}(X, Y)}{\text{var}(X)},$$

$$\triangleright a^* = \mathbb{E}[Y] - b^* \mathbb{E}[X] = \mathbb{E}[Y] - \frac{\text{cov}(X, Y)}{\text{var}(X)} \mathbb{E}[X].$$

Noise

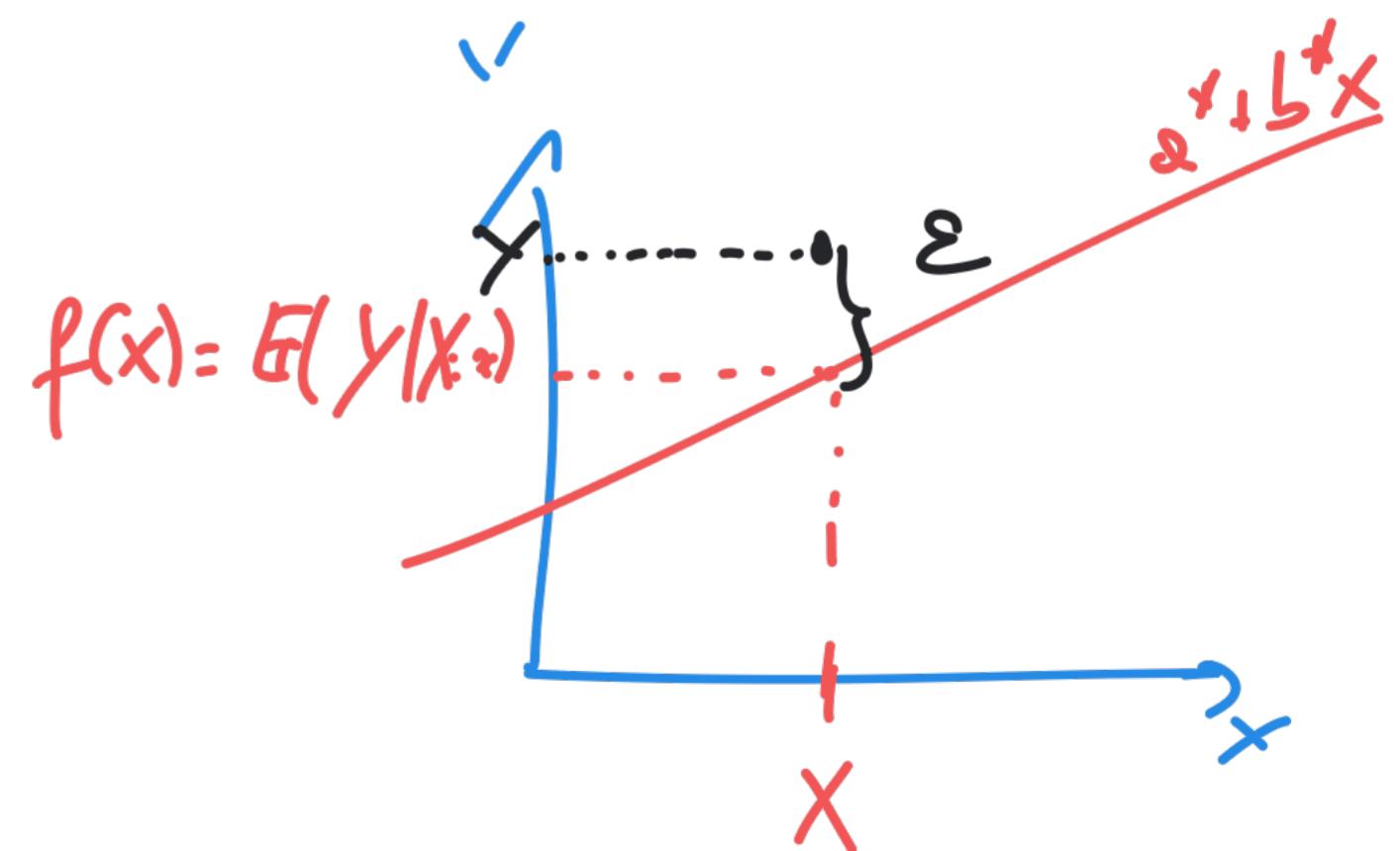
Clearly the points are not exactly on the line $x \mapsto a^* + b^*x$ if $\text{Var}(Y|X = x) > 0$. The random variable $\varepsilon = Y - (a^* + b^*X)$ is called *noise* and satisfies

$$Y = a^* + b^*X + \varepsilon,$$

with

- $\mathbb{E}[\varepsilon] = 0$ and
- $\text{cov}(X, \varepsilon) = 0$. (exercise)

$$\mathbb{E}[\varepsilon] = \mathbb{E}[Y] - a^* - b^* \mathbb{E}[X] = 0$$



Statistical problem

In practice a^*, b^* need to be estimated from data.

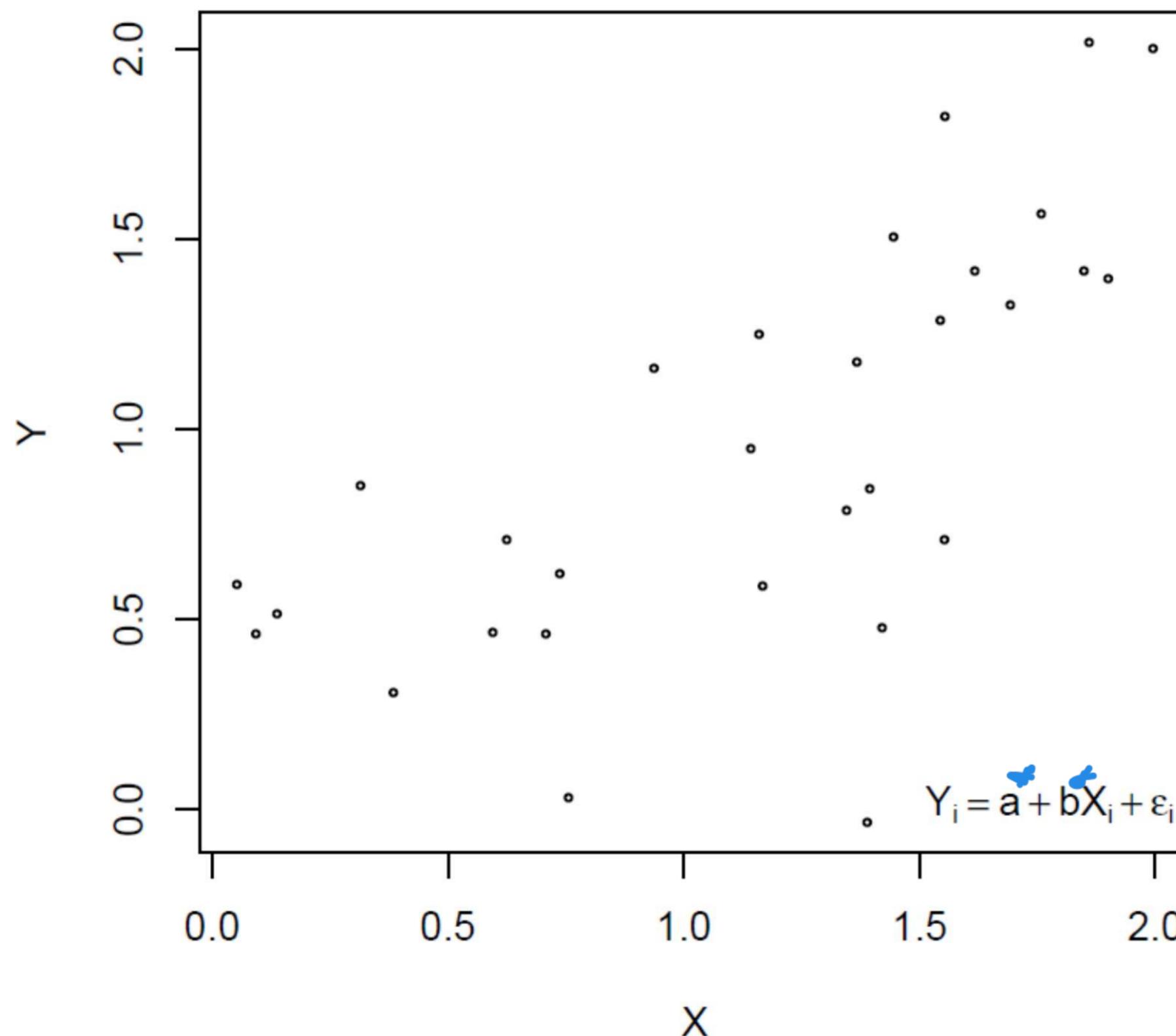
- ▶ Assume that we observe n i.i.d. random pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ with same distribution as (X, Y) :

$$Y_i = a^* + b^* X_i + \varepsilon_i$$

- ▶ We want to estimate a^* and b^* .

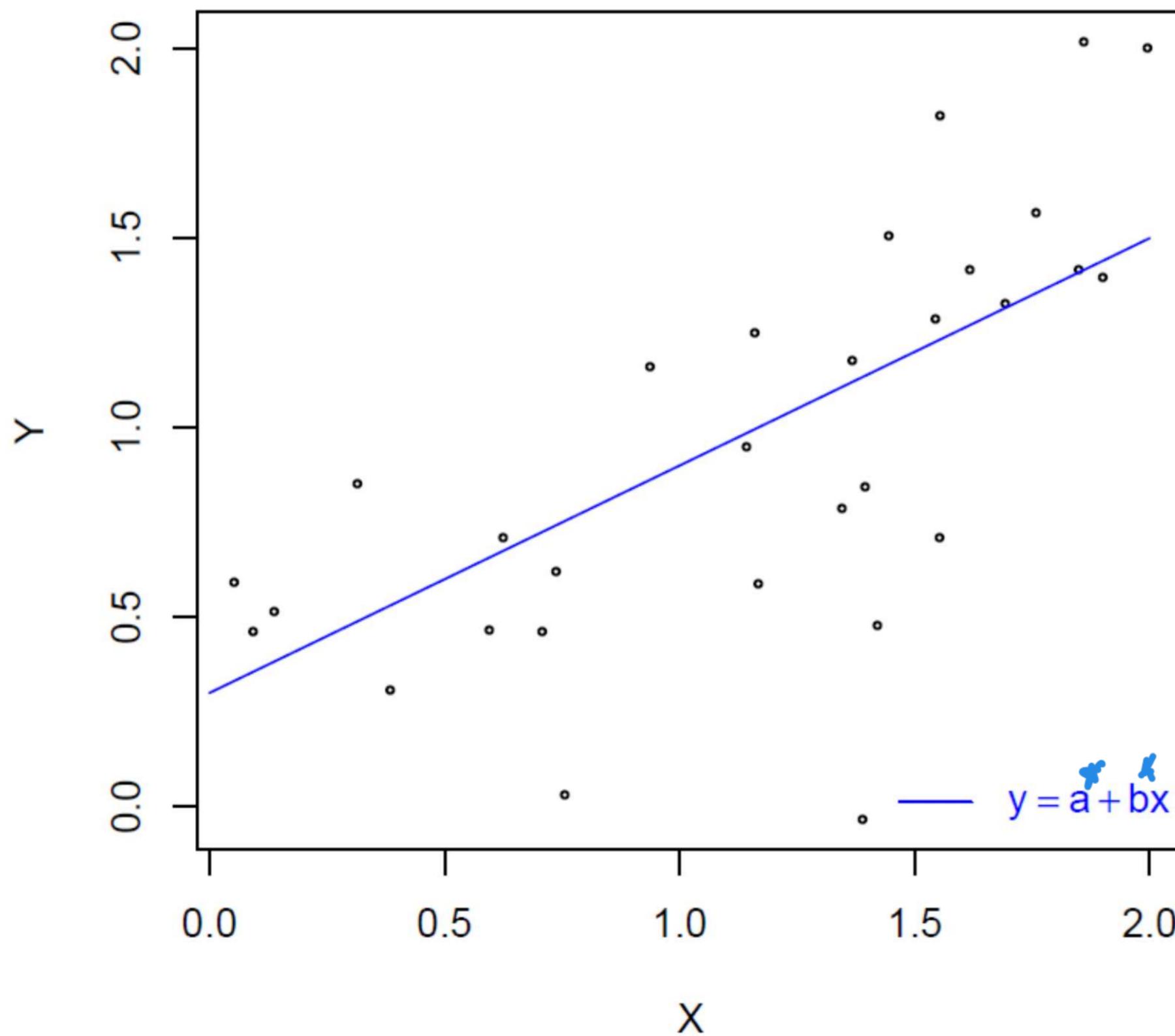
Statistical problem

$$Y_i = a^* + b^* X_i + \varepsilon_i$$



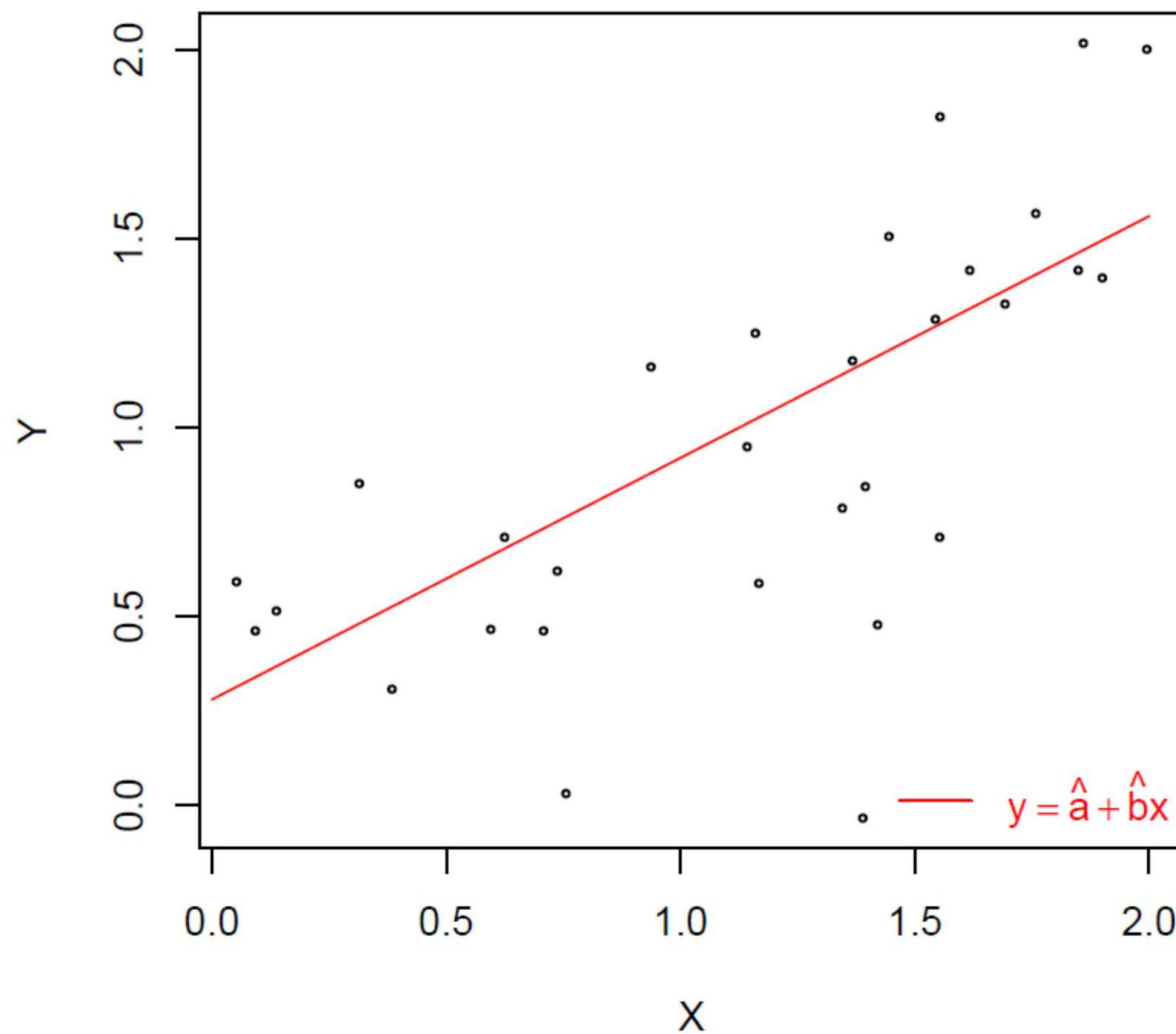
Statistical problem

$$Y_i = a^* + b^* X_i + \varepsilon_i$$



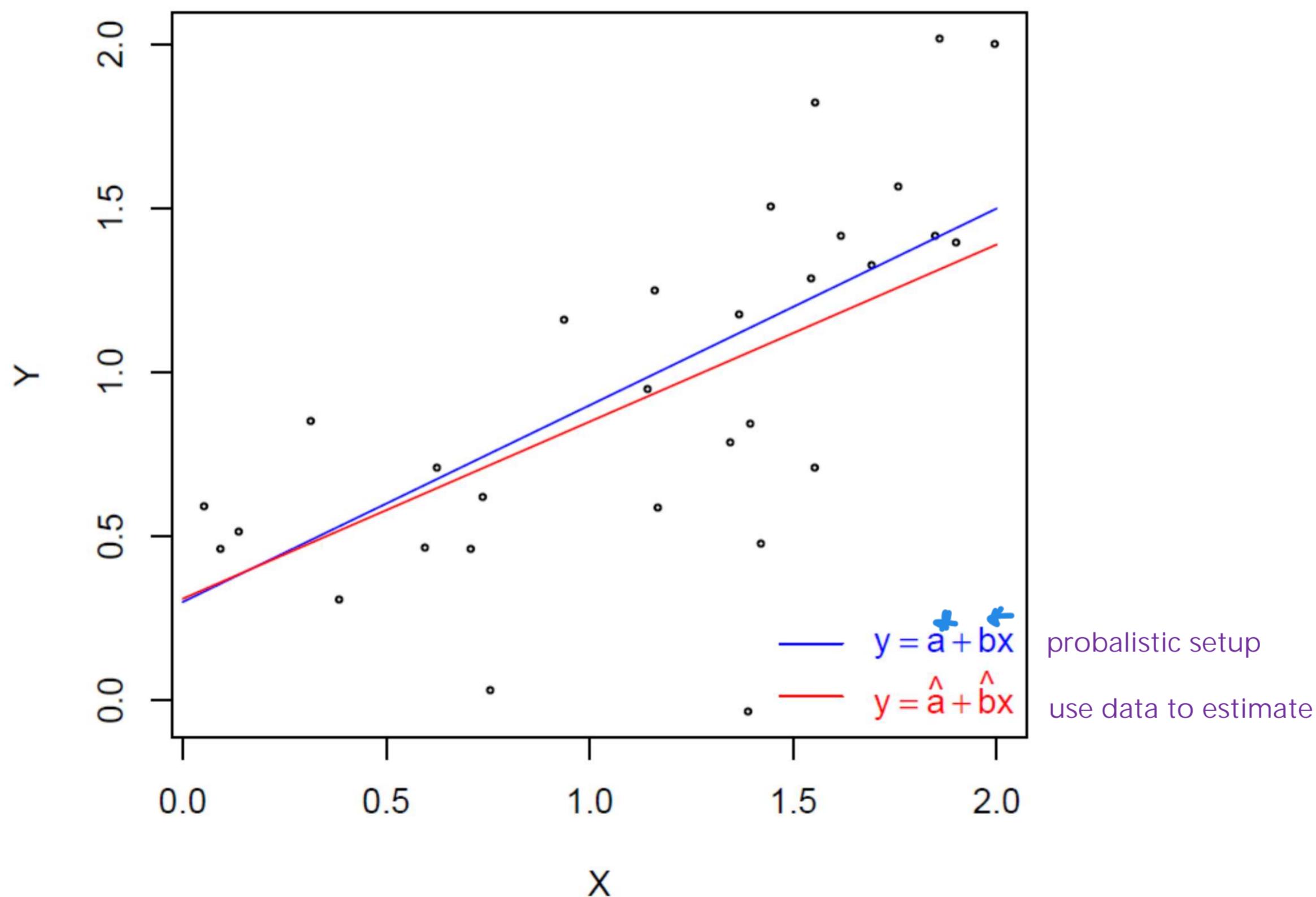
Statistical problem

$$Y_i = a^* + b^* X_i + \varepsilon_i$$



Statistical problem

$$Y_i = a^* + b^* X_i + \varepsilon_i$$



Least squares

Definition

The ~~least squared error~~ (LSE) estimator of (a, b) is the minimizer of the sum of squared errors:

$$\sum_{i=1}^n (Y_i - a - bX_i)^2.$$

(\hat{a}, \hat{b}) is given by

$$\hat{b} = \frac{\bar{XY} - \bar{X}\bar{Y}}{\bar{X^2} - \bar{X}^2}$$

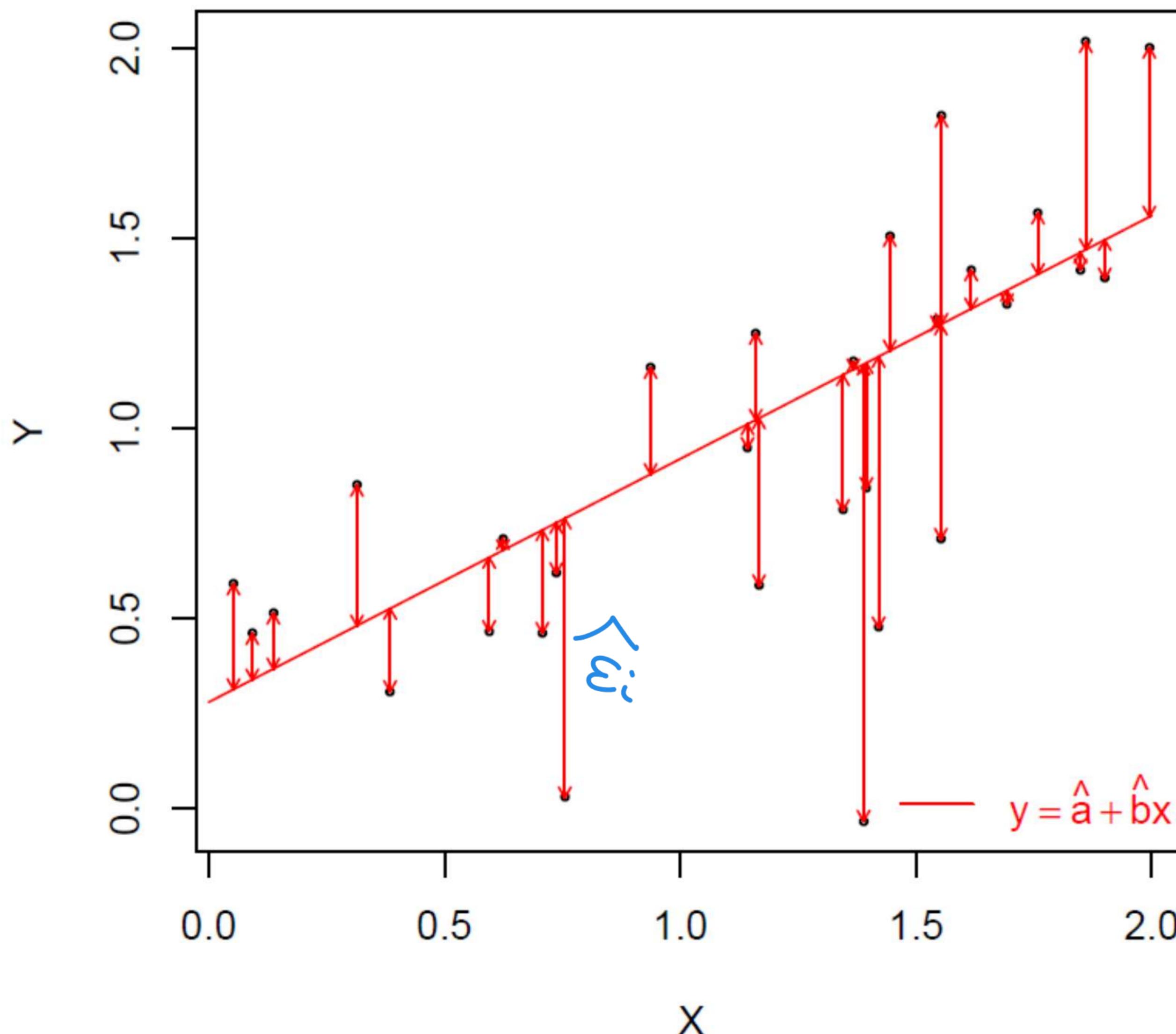
(exercise)

$$\hat{a} = \bar{Y} - \hat{b}\bar{X}.$$

R code:

```
a <- lm(y ~ x)  
plot(a)  
summary(a)
```

Residuals



Multivariate regression

linear combination

$$Y_i = \mathbf{X}_i^\top \beta^* + \varepsilon_i, \quad i = 1, \dots, n.$$

$$\alpha^* + \beta^* X_i$$

$$\mathbf{X}_i = \begin{pmatrix} X_i^{(1)} \\ X_i^{(2)} \\ \vdots \\ X_i^{(p)} \end{pmatrix}$$

上标是自变量的数量
下标是数据的数量

$$\mathbf{X}_i^\top \beta^* = \sum_{j=1}^p X_i^{(j)} \beta_j^*$$

- Vector of **explanatory variables** or **covariates**: $\mathbf{X}_i \in \mathbb{R}^p$
(wlog, assume its first coordinate is 1).
not independent variable
not necessarily independent
- **Response / Dependent variable**: Y_i .
- $\beta^* = (a^*, \mathbf{b}^{*\top})^\top$; $\beta_1^* (= a^*)$ is called the **intercept**.
- $\{\varepsilon_i\}_{i=1,\dots,n}$: noise terms satisfying $\text{cov}(\mathbf{X}_i, \varepsilon_i) = 0$.

Definition

The **least squares estimator (LSE)** of β^* is the minimizer of the sum of square errors:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^n (Y_i - \mathbf{X}_i^\top \beta)^2$$

空间上Y的距离

LSE in matrix form

- ▶ Let $\mathbf{Y} = (Y_1, \dots, Y_n)^\top \in \mathbb{R}^n$.
- ▶ Let \mathbb{X} be the $n \times p$ matrix whose rows are $\mathbf{X}_1^\top, \dots, \mathbf{X}_n^\top$ (\mathbb{X} is called the **design matrix**). *\mathbb{X} has size $n \times p$*
- ▶ Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top \in \mathbb{R}^n$ (unobserved noise)

▶ $\mathbf{Y} = \mathbb{X}\beta^* + \boldsymbol{\varepsilon}$, β^* unknown.

▶ The LSE $\hat{\beta}$ satisfies:

$$\begin{array}{c|c|c|c} \mathbf{Y} & = & \mathbb{X}^\top & \mathbf{I} \\ & & \vdots & \\ & & \mathbb{X}^\top & \beta^* \\ & & \mathbb{X} & \\ & & & \boldsymbol{\varepsilon} \end{array}$$

矩阵X的每一行都是一次的观测值
第一列全是1
X和Y是一对一的，整个公式每一行是一组方程

$$\|\mathbf{w}\|_2^2 = \sum_{i=1}^n w_i^2$$

$w \in \mathbb{R}^n$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|\mathbf{Y} - \mathbb{X}\beta\|_2^2.$$

$$\sum_{i=1}^n (Y_i - X_i^\top \beta)^2$$

Closed form solution

- ▶ Assume that $\text{rank}(\mathbb{X}) = p$.
- ▶ Analytic computation of the LSE:

$$\hat{\beta} = \boxed{(\mathbb{X}^\top \mathbb{X})^{-1}} \mathbb{X}^\top \mathbf{Y}.$$

rank p, n > p

- ▶ Geometric interpretation of the LSE: $\mathbb{X}\hat{\beta}$ is the orthogonal projection of \mathbf{Y} onto the subspace spanned by the columns of \mathbb{X} :

$$\mathbb{X}\hat{\beta} = P\mathbf{Y},$$

where $P = \mathbb{X}(\mathbb{X}^\top \mathbb{X})^{-1}\mathbb{X}^\top$.

Statistical inference

$$y_i \not\propto \beta^* + \varepsilon$$

To make inference (confidence regions, tests) we need more assumptions.

Assumptions:

- ▶ The design matrix \mathbb{X} is deterministic and $\text{rank}(\mathbb{X}) = p$.
- ▶ The model is **homoscedastic**: $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d.
- ▶ The noise vector ε is Gaussian:

$$\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_n)$$

for some known or unknown $\sigma^2 > 0$.

$$\Rightarrow \mathbb{Y} \sim \mathcal{N}_n(\mathbb{X}\beta^*, \sigma^2 \mathbb{I})$$

Properties of LSE

- LSE = MLE
- Distribution of $\hat{\beta}$: $\hat{\beta} \sim N_p(\beta^*, \sigma^2 (\mathbb{X}^\top \mathbb{X})^{-1})$
- Quadratic risk of $\hat{\beta}$: $\mathbb{E} [\|\hat{\beta} - \beta\|_2^2] = \sigma^2 \text{tr} ((\mathbb{X}^\top \mathbb{X})^{-1})$.
- Prediction error: $\mathbb{E} [\|\mathbf{Y} - \mathbb{X}\hat{\beta}\|_2^2] = \sigma^2(n - p)$.
- Unbiased estimator of σ^2 : $\hat{\sigma}^2 = \frac{\|\mathbf{Y} - \mathbb{X}\hat{\beta}\|_2^2}{n - p} = \frac{1}{n - p} \sum_{i=1}^n \hat{\epsilon}_i^2$

Theorem

- $(n - p) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$.
- $\hat{\beta} \perp \hat{\sigma}^2$. (Cochran's theorem)

Significance tests

- Test whether the j -th explanatory variable is significant in the linear regression ($1 \leq j \leq p$).
- $H_0 : \beta_j = 0$ v.s. $H_1 : \beta_j \neq 0$.
- If γ_j is the j -th diagonal coefficient of $(\mathbb{X}^\top \mathbb{X})^{-1}$ ($\gamma_j > 0$):

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 \gamma_j}} \sim t_{n-p}$$

- Let $T_n^{(j)} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \gamma_j}}$.
- Test with non asymptotic level $\alpha \in (0, 1)$:

$$R_{j,\alpha} = \left\{ |T_n^{(j)}| > q_{\frac{\alpha}{2}}(t_{n-p}) \right\}$$

where $q_{\frac{\alpha}{2}}(t_{n-p})$ is the $(1 - \alpha/2)$ -quantile of t_{n-p} .

- We can also compute p-values.

Bonferroni's test

- ▶ Test whether a **group** of explanatory variables is significant in the linear regression.
- ▶ $H_0 : \beta_j = 0, \forall j \in S$ v.s. $H_1 : \exists j \in S, \beta_j \neq 0$, where $S \subseteq \{1, \dots, p\}$.
- ▶ *Bonferroni's test:* $R_{S,\alpha} = \bigcup_{j \in S} R_{j,\frac{\alpha}{k}}$, where $k = |S|$.
- ▶ This test has nonasymptotic level at most α .

Type I is $P_{H_0}[R_{S,\alpha}] \leq \sum_{j \in S} P_{H_0}[R_{j,\frac{\alpha}{k}}] = \alpha$

$\underbrace{= \alpha}_{= \alpha/k}$

Remarks

- ▶ Linear regression exhibits correlations, **NOT** causality
- ▶ Normality of the noise: One can use goodness of fit tests to test whether the residuals $\hat{\varepsilon}_i = Y_i - \mathbb{X}_i^\top \hat{\beta}$ are Gaussian.
- ▶ Deterministic design: If \mathbb{X} is not deterministic, all the above can be understood conditionally on \mathbb{X} , if the noise is assumed to be Gaussian, conditionally on X .