Lecture 2 | 03.03.2026

Simple linear regression model

(continuous/binary/categorical covariate $X \in \mathbb{R}$)

Simple (ordinary) linear regression model

General/generic model formulation

$$Y = a + bX + \varepsilon$$

- ❑ $Y \in \mathbb{R}$ is a random variable, the covariate of interest (dependent variable)
- ❑ $X \in \mathbb{R}$ is a random variable, the model covariate (explanatory variable)
  (regression models can be also used with non-random $x \in \mathbb{R}$ however, within a slightly more complex framework)
- ❑ $\varepsilon \in \mathbb{R}$ represents a latent random variable – an unobserved random error

Simple (ordinary) linear regression model

General/generic model formulation

$$Y = a + bX + \varepsilon$$

- ❑ $Y \in \mathbb{R}$ is a random variable, the covariate of interest (dependent variable)
- ❑ $X \in \mathbb{R}$ is a random variable, the model covariate (explanatory variable)
  (regression models can be also used with non-random $x \in \mathbb{R}$ however, within a slightly more complex framework)

- ❑ $\varepsilon \in \mathbb{R}$ represents a latent random variable – an unobserved random error

- ❑ Using a random sample $\mathcal{D} = \{(Y_i, X_i);\ i = 1, \dots, n\}$ (i.e., the "**data**"), the estimates for $a, b \in \mathbb{R}$ can be obtained by least squares

$$(\widehat{a}, \widehat{b})^\top = \underset{(a,b)\in\mathbb{R}^2}{\mathrm{Arg\,min}} \sum_{i=1}^{n} [Y_i - (a + bX_i)]^2$$

with **explicit solution** (i.e., expressions) for $\widehat{a}$ and $\widehat{b}$

Simple (ordinary) linear regression model

General/generic model formulation

$$Y = a + bX + \varepsilon$$

- ❑ $Y \in \mathbb{R}$ is a random variable, the covariate of interest (dependent variable)
- ❑ $X \in \mathbb{R}$ is a random variable, the model covariate (explanatory variable)
  (regression models can be also used with non-random $x \in \mathbb{R}$ however, within a slightly more complex framework)
  $\hookrightarrow$ **and X can also take only finitelly many values** (quantitative vs. qualitative information)
- ❑ $\varepsilon \in \mathbb{R}$ represents a latent random variable – an unobserved random error

- ❑ Using a random sample $\mathcal{D} = \{(Y_i, X_i);\ i = 1, \dots, n\}$ (i.e., the "**data**"), the estimates for $a, b \in \mathbb{R}$ can be obtained by least squares

$$(\widehat{a}, \widehat{b})^\top = \operatorname*{Arg\,min}_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n} [Y_i - (a + bX_i)]^2$$

with **explicit solution** (i.e., expressions) for $\widehat{a}$ and $\widehat{b}$

Simple (ordinary) linear regression model

General/generic model formulation

$$Y = f(X) + \varepsilon$$

- ❑ $Y \in \mathbb{R}$ is a random variable, the covariate of interest (dependent variable)
- ❑ $X \in \mathbb{R}$ is a random variable, the model covariate (explanatory variable)
  (regression models can be also used with non-random $x \in \mathbb{R}$ however, within a slightly more complex framework)
  ↪ **and X can also take only finitelly many values** (quantitative vs. qualitative information)
- ❑ $\varepsilon \in \mathbb{R}$ represents a latent random variable – an unobserved random error

- ❑ Using a random sample $\mathcal{D} = \{(Y_i, X_i);\ i = 1, \dots, n\}$ (i.e., the "**data**"), the estimate for a parametric function f can be obtained by least squares

$$\hat{f} = \operatorname*{Arg\,min}_{f \in \mathcal{C}} \sum_{i=1}^{n} [Y_i - f(X_i)]^2$$

with **explicit solution** for $\widehat{f}$ in terms of some estimated parameters

Maximum likelihood estimation

- **The underlying model:** $Y = a + bX + \varepsilon$ (i.e., straight line)
- **Normality assumption:** $\varepsilon \sim N(0, \sigma^2)$ (i.e., distributional property)

Maximum likelihood estimation

- ❑ **The underlying model:** $Y = a + bX + \varepsilon$ (i.e., straight line)
- ❑ **Normality assumption:** $\varepsilon \sim N(0, \sigma^2)$ (i.e., distributional property)
- ❑ **Conditional normality of Y given X:** $Y|X \sim N(a + bX, \sigma^2)$

Maximum likelihood estimation

- ❑ **The underlying model:** $Y = a + bX + \varepsilon$ (i.e., straight line)
- ❑ **Normality assumption:** $\varepsilon \sim N(0, \sigma^2)$ (i.e., distributional property)
- ❑ **Conditional normality of Y given X:** $Y|X \sim N(a + bX, \sigma^2)$

Using the random sample in $\mathcal{D}$, drawn from the joint distribution $F_{(Y,X)}$ that can be factorised as $F_{(Y,X)} \equiv F_{Y/X} \cdot F_X$ where $F_{Y|X} \equiv N(a + bX, \sigma^2)$ and F_X does not depend on $a, b \in \mathbb{R}$:

- ❑ Intercept and slope estimates are

$$\widehat{a} = \overline{Y}_n - \widehat{b}\overline{X}_n \qquad \text{and} \qquad \widehat{b} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y}_n)(X_i - \overline{X}_n)}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}$$

- ❑ Variance parameter estimate is

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - (\widehat{a} + \widehat{b}X_i))^2$$

Maximum likelihood estimation

- ❑ **The underlying model:** $Y = a + bX + \varepsilon$ (i.e., straight line)
- ❑ **Normality assumption:** $\varepsilon \sim N(0, \sigma^2)$ (i.e., distributional property)
- ❑ **Conditional normality of Y given X:** $Y|X \sim N(a + bX, \sigma^2)$

Using the random sample in $\mathcal{D}$, drawn from the joint distribution $F_{(Y,X)}$ that can be factorised as $F_{(Y,X)} \equiv F_{Y/X} \cdot F_X$ where $F_{Y|X} \equiv N(a + bX, \sigma^2)$ and F_X does not depend on $a, b \in \mathbb{R}$:

- ❑ Intercept and slope estimates are

$$\widehat{a} = \overline{Y}_n - \widehat{b}\overline{X}_n \qquad \text{and} \qquad \widehat{b} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y}_n)(X_i - \overline{X}_n)}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}$$

- ❑ Variance parameter estimate is

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - (\widehat{a} + \widehat{b}X_i))^2$$

- ❑ And, moreover, it also holds that

$$\widehat{a} \sim N\left(a, \sigma^2\left[\frac{1}{n} + \frac{\overline{X}_n^2}{\sum_i (X_i - \overline{X}_n)^2}\right]\right) \qquad \text{and} \qquad \widehat{b} \sim N\left(b, \frac{\sigma^2}{\sum_i (X_i - \overline{X}_n)^2}\right)$$

Likelihood and log-likelihood

- Density of a normal $N(\mu, \sigma^2)$ distribution (for any $x \in \mathbb{R}$)

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Likelihood and log-likelihood

- Density of a normal $N(\mu, \sigma^2)$ distribution (for any $x \in \mathbb{R}$)

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- Likelihood $L(\mu, \sigma^2, \mathcal{S})$ for the dataset $\mathcal{S} = \{(Y_i, X_i);\ i = 1, \ldots, n\}$

$$L(\mu, \sigma^2, \mathcal{S}) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{(Y_i - (a + bX_i))^2}{2\sigma^2}\right\}\right]$$

Likelihood and log-likelihood

❑ Density of a normal $N(\mu, \sigma^2)$ distribution (for any $x \in \mathbb{R}$)

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

❑ Likelihood $L(\mu, \sigma^2, \mathcal{S})$ for the dataset $\mathcal{S} = \{(Y_i, X_i);\ i = 1, \ldots, n\}$

$$L(\mu, \sigma^2, \mathcal{S}) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{(Y_i - (a + bX_i))^2}{2\sigma^2}\right\}\right]$$

❑ The corresponding log-likelihood function $\ell(\mu, \sigma^2, \mathcal{S})$

$$\ell(\mu, \sigma^2, \mathcal{S}) = (-n/2)\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(Y_i - (a + bX_i))^2}{2\sigma^2}$$

Likelihood and log-likelihood

- ❑ Density of a normal $N(\mu, \sigma^2)$ distribution (for any $x \in \mathbb{R}$)

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- ❑ Likelihood $L(\mu, \sigma^2, \mathcal{S})$ for the dataset $\mathcal{S} = \{(Y_i, X_i);\ i = 1, \dots, n\}$

$$L(\mu, \sigma^2, \mathcal{S}) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{(Y_i - (a + bX_i))^2}{2\sigma^2}\right\}\right]$$

- ❑ The corresponding log-likelihood function $\ell(\mu, \sigma^2, \mathcal{S})$

$$\ell(\mu, \sigma^2, \mathcal{S}) = (-n/2)\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(Y_i - (a + bX_i))^2}{2\sigma^2}$$

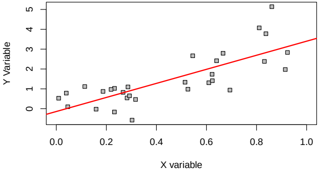
$\hookrightarrow$ note the notation difference between the likelihood $L(\cdot)$, log-likelihood $\ell(\cdot)$, and a general loss function $\mathcal{L}(\cdot)$

Quantitative vs. qualitative covariate $X \in \mathbb{R}$

- Simple regression model $Y_i = a + bX_i + \varepsilon_i$ for different values of X_i ...

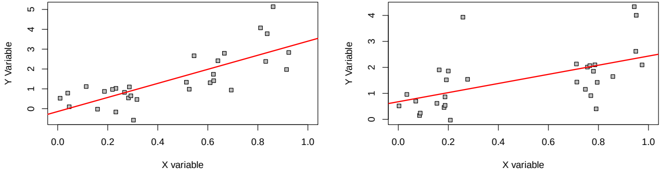
Quantitative vs. qualitative covariate $X \in \mathbb{R}$

- ❏ Simple regression model $Y_i = a + bX_i + \varepsilon_i$ for different values of X_i ...



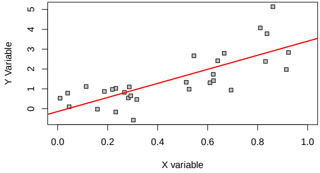
Quantitative vs. qualitative covariate $X \in \mathbb{R}$

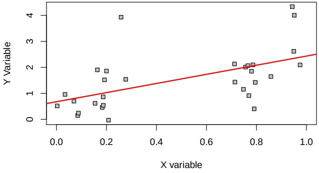
- ❑ Simple regression model $Y_i = a + bX_i + \varepsilon_i$ for different values of X_i …

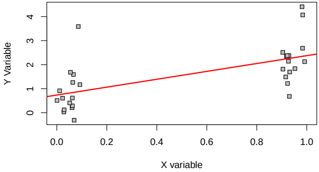


Quantitative vs. qualitative covariate $X \in \mathbb{R}$

- ❑ Simple regression model $Y_i = a + bX_i + \varepsilon_i$ for different values of X_i ...

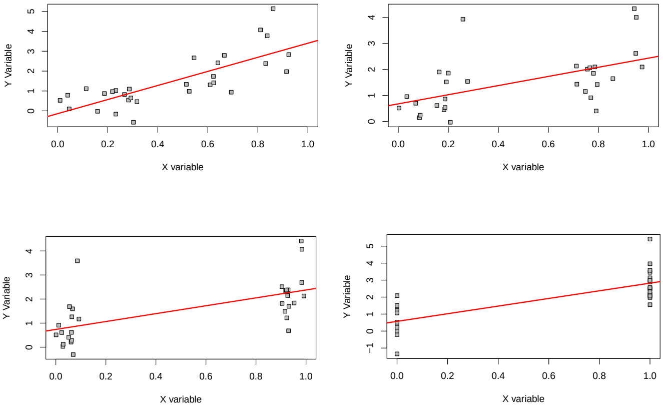






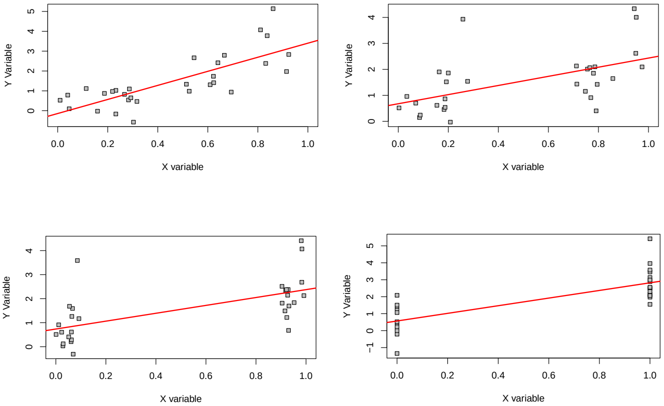
Quantitative vs. qualitative covariate $X \in \mathbb{R}$

- Simple regression model $Y_i = a + bX_i + \varepsilon_i$ for different values of X_i …



Quantitative vs. qualitative covariate $X \in \mathbb{R}$

❑ Simple regression model $Y_i = a + bX_i + \varepsilon_i$ for different values of X_i …



The regression line $f(x) = a + bx$ provides and estimate of the conditional mean $E[Y|X = x]$ for any $x \in [0, 1]$ (or $x \in \mathbb{R}$) … **But what if all values beside $x \in \{0, 1\}$ do not make any sense… ?**

Regression model for a categorical covariate

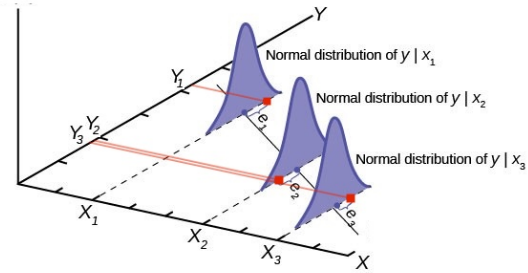
- **Two sample problem**
  if X only takes two values (e.g., $X \in \{0,1\}$ or $X = \pm 1$), random sample $\mathcal{D}$ – observations (Y_i, X_i) for $i = 1, \dots, n$ – can be split into two parts: The values of Y_i for which $X_i = 0$ and the values of Y_i for which $X = 1$ and a simple average is calculated in both groups **(two-sample problem)**

- **Multiple samples**
  if X takes finitely many different values (i.e., X is a categorical variable with $K \in \mathbb{N}$ different levels/labels, $X \in \{l_1, \dots, l_K\}$), the random sample $\{(Y_i, X_i)\}_{i=1}^{n}$ can be split into K disjoint groups and, again, simple averages can be calculated for each of K groups **(analysis of variance – ANOVA)**

- **Continuous explanatory variable**
  if X is a continuous variable (taking infinitely/uncountable many values), the sample can not be split into all possible groups – for almost all "$X = x$" there will be no observations of Y available in $\mathcal{D}$
  $\Longrightarrow$ **borrowing power from the neighbors** **(regression problem)**

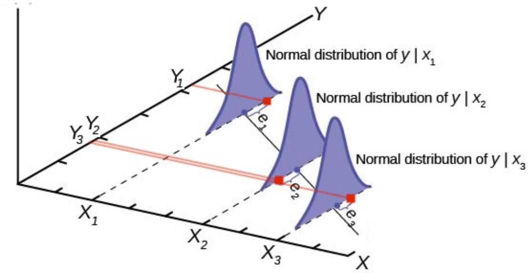
Conditional distribution of Y given $X = x$

❑ To say **what is the mean of Y under the condition that $X = x$** (which may never be observed in the data) we need to borrow power from other ocurrences of Y (i.e, observations $Y_1, \dots, Y_n$) taken under the situations in which X takes values observed in $X_1 \dots, X_n$
(all being potentially unique and all being different from $x \in \mathbb{R}$)


Y
Normal distribution of $y \mid x_1$
Normal distribution of $y \mid x_2$
Normal distribution of $y \mid x_3$
Y_1
Y_2
Y_3
e_1
e_2
e_3
X_1
X_2
X_3
X

Conditional distribution of Y given $X = x$

- To say **what is the mean of Y under the condition that $X = x$** (which may never be observed in the data) we need to borrow power from other ocurrences of Y (i.e, observations $Y_1, \dots, Y_n$) taken under the situations in which X takes values observed in $X_1 \dots, X_n$
  (all being potentially unique and all being different from $x \in \mathbb{R}$)


Y
Normal distribution of y | x₁
Normal distribution of y | x₂
Normal distribution of y | x₃
Y₁
Y₃
Y₂
e₁
e₂
e₃
X₁
X₂
X₃
X

- Borrowing power can be also expressed as a form of a learning process...
  (e.g., statistical learning, machine learning, reinforcement learning, ...)

Ordinary rgression as a two sample problem

- If the explanatory variable $X \in \mathbb{R}$ only takes two distinct values – e.g., one for TRUE and zero for FALSE – the regression model $f(x) = a + bx$ can be still considered in a straightforward way...

- Let $X = 1$ for TRUE (group 1) and $X = 0$ for FALSE (group 2)
  - For $X = 0$, the model reduces to $E[Y|X = 0] = f(0) = a + b \cdot 0 = a$
    (*i.e., $a \in \mathbb{R}$ stands for the mean of the sub-population for which the covariate equals to FALSE*)
  - For $X = 1$, the model reduces to $E[Y|X = 1] = f(1) = a + b \cdot 1 = a + b$
    (*i.e., $a + b \in \mathbb{R}$ stands for the the mean of the sub-population where the covariate is TRUE*)

Ordinary rgression as a two sample problem

- If the explanatory variable $X \in \mathbb{R}$ only takes two distinct values – e.g., one for TRUE and zero for FALSE – the regression model $f(x) = a + bx$ can be still considered in a straightforward way...

- Let $X = 1$ for TRUE (group 1) and $X = 0$ for FALSE (group 2)
  - For $X = 0$, the model reduces to $E[Y|X = 0] = f(0) = a + b \cdot 0 = a$

    (*i.e., $a \in \mathbb{R}$ stands for the mean of the sub-population for which the covariate equals to FALSE*)
  - For $X = 1$, the model reduces to $E[Y|X = 1] = f(1) = a + b \cdot 1 = a + b$

    (*i.e., $a + b \in \mathbb{R}$ stands for the the mean of the sub-population where the covariate is TRUE*)

- There are infinitely many different parametrizations that can be used to encode the binary variable X — for instance, it can take two values ± 1

  (*thus, $a - b$ stands for the mean of the first and $a + b$ for the mean of the second sub-population*)

  What would be the interpretation of the $a \in \mathbb{R}$ parameter itself?

Ordinary rgression as a two sample problem

- If the explanatory variable $X \in \mathbb{R}$ only takes two distinct values – e.g., one for TRUE and zero for FALSE – the regression model $f(x) = a + bx$ can be still considered in a straightforward way...

- Let $X = 1$ for TRUE (group 1) and $X = 0$ for FALSE (group 2)
  - For $X = 0$, the model reduces to $E[Y|X = 0] = f(0) = a + b \cdot 0 = a$
    (*i.e., $a \in \mathbb{R}$ stands for the mean of the sub-population for which the covariate equals to FALSE*)
  - For $X = 1$, the model reduces to $E[Y|X = 1] = f(1) = a + b \cdot 1 = a + b$
    (*i.e., $a + b \in \mathbb{R}$ stands for the the mean of the sub-population where the covariate is TRUE*)

- There are infinitely many different parametrizations that can be used to encode the binary variable X — for instance, it can take two values ± 1
  (*thus, $a - b$ stands for the mean of the first and $a + b$ for the mean of the second sub-population*)
  What would be the interpretation of the $a \in \mathbb{R}$ parameter itself?

- In general, the binary explanatory variable X reduces the ordinary linear regression model into a classical two sample problem of the form

$$Y = a + \beta_1 \mathbb{I}_{\{X = TRUE\}} + \beta_2 \mathbb{I}_{\{X = FALSE\}} + \varepsilon = f(X) + \varepsilon$$

Parametrizations of the binary variable

- There are infinitely many different ways/parametrizations how to identify parameters a, β_1, β_2 (to assign some interpretation to these parameters)
- In other words, the binary explanatory variable X reduces the ordinary linear regression model into a classical two sample problem of the form

$$Y = \mu_1 \mathbb{I}_{\{\texttt{group 1}\}} + \mu_2 \mathbb{I}_{\{\texttt{group 2}\}} + \varepsilon$$

  where $\mu_1 \in \mathbb{R}$ represents the unknonw (true) mean of the first random sample, and $\mu_2 \in \mathbb{R}$ stands for the unknown mean of the second sample

Parametrizations of the binary variable

- There are infinitely many different ways/parametrizations how to identify parameters a, β_1, β_2 (to assign some interpretation to these parameters)
- In other words, the binary explanatory variable X reduces the ordinary linear regression model into a classical two sample problem of the form

$$Y = \mu_1 \mathbb{I}_{\{\texttt{group 1}\}} + \mu_2 \mathbb{I}_{\{\texttt{group 2}\}} + \varepsilon$$

  where $\mu_1 \in \mathbb{R}$ represents the unknonw (true) mean of the first random sample, and $\mu_2 \in \mathbb{R}$ stands for the unknown mean of the second sample

- **How to express μ_1 and μ_2 in terms of the model $f(x) = a + bx$?**
  - Parametrization #1: let $\texttt{TRUE} = 0$ and $\texttt{FALSE} = 1$

    $\Longrightarrow \mu_1 = E[Y|X = \texttt{TRUE}] = a$ and $\mu_2 = E[Y|X = \texttt{FALSE}] = a + b$

Parametrizations of the binary variable

- There are infinitely many different ways/parametrizations how to identify parameters a, β_1, β_2 (to assign some interpretation to these parameters)
- In other words, the binary explanatory variable X reduces the ordinary linear regression model into a classical two sample problem of the form

$$Y = \mu_1 \mathbb{I}_{\{\texttt{group 1}\}} + \mu_2 \mathbb{I}_{\{\texttt{group 2}\}} + \varepsilon$$

where $\mu_1 \in \mathbb{R}$ represents the unknonw (true) mean of the first random sample, and $\mu_2 \in \mathbb{R}$ stands for the unknown mean of the second sample

- **How to express μ_1 and μ_2 in terms of the model $f(x) = a + bx$?**
  - Parametrization #1: let $\texttt{TRUE} = 0$ and $\texttt{FALSE} = 1$
    $\Longrightarrow \mu_1 = E[Y|X = \texttt{TRUE}] = a$ and $\mu_2 = E[Y|X = \texttt{FALSE}] = a + b$
  - Parametrization #2: let $\texttt{TRUE} = 1$ and $\texttt{FALSE} = 0$
    $\Longrightarrow \mu_1 = E[Y|X = \texttt{TRUE}] = a + b$ and $\mu_2 = E[Y|X = \texttt{FALSE}] = a$

Parametrizations of the binary variable

- There are infinitely many different ways/parametrizations how to identify parameters a, β_1, β_2 (to assign some interpretation to these parameters)
- In other words, the binary explanatory variable X reduces the ordinary linear regression model into a classical two sample problem of the form

$$Y = \mu_1 \mathbb{I}_{\{\texttt{group 1}\}} + \mu_2 \mathbb{I}_{\{\texttt{group 2}\}} + \varepsilon$$

where $\mu_1 \in \mathbb{R}$ represents the unknonw (true) mean of the first random sample, and $\mu_2 \in \mathbb{R}$ stands for the unknown mean of the second sample

- **How to express μ_1 and μ_2 in terms of the model $f(x) = a + bx$?**
  - Parametrization #1: let $\texttt{TRUE} = 0$ and $\texttt{FALSE} = 1$
    $\Longrightarrow \mu_1 = E[Y|X = \texttt{TRUE}] = a$ and $\mu_2 = E[Y|X = \texttt{FALSE}] = a + b$
  - Parametrization #2: let $\texttt{TRUE} = 1$ and $\texttt{FALSE} = 0$
    $\Longrightarrow \mu_1 = E[Y|X = \texttt{TRUE}] = a + b$ and $\mu_2 = E[Y|X = \texttt{FALSE}] = a$
  - Parametrization #3: let $\texttt{TRUE} = -1$ and $\texttt{FALSE} = 1$
    $\Longrightarrow \mu_1 = E[Y|X = \texttt{TRUE}] = a - b$ and $\mu_2 = E[Y|X = \texttt{FALSE}] = a + b$

Parametrizations of the binary variable

- There are infinitely many different ways/parametrizations how to identify parameters a, β_1, β_2 (to assign some interpretation to these parameters)

- In other words, the binary explanatory variable X reduces the ordinary linear regression model into a classical two sample problem of the form

$$Y = \mu_1 \mathbb{I}_{\{\texttt{group 1}\}} + \mu_2 \mathbb{I}_{\{\texttt{group 2}\}} + \varepsilon$$

  where $\mu_1 \in \mathbb{R}$ represents the unknonw (true) mean of the first random sample, and $\mu_2 \in \mathbb{R}$ stands for the unknown mean of the second sample

- **How to express μ_1 and μ_2 in terms of the model $f(x) = a + bx$?**
  - Parametrization #1: let $\texttt{TRUE} = 0$ and $\texttt{FALSE} = 1$
    $\Longrightarrow \mu_1 = E[Y|X = \texttt{TRUE}] = a$ and $\mu_2 = E[Y|X = \texttt{FALSE}] = a + b$
  - Parametrization #2: let $\texttt{TRUE} = 1$ and $\texttt{FALSE} = 0$
    $\Longrightarrow \mu_1 = E[Y|X = \texttt{TRUE}] = a + b$ and $\mu_2 = E[Y|X = \texttt{FALSE}] = a$
  - Parametrization #3: let $\texttt{TRUE} = -1$ and $\texttt{FALSE} = 1$
    $\Longrightarrow \mu_1 = E[Y|X = \texttt{TRUE}] = a - b$ and $\mu_2 = E[Y|X = \texttt{FALSE}] = a + b$
  - Parametrization #4: let $\texttt{TRUE} = v_1$ and $\texttt{FALSE} = v_2$
    $\Longrightarrow \mu_1 = E[Y|X = \texttt{TRUE}] = a + bv_1$ and $\mu_2 = E[Y|X = \texttt{FALSE}] = a + bv_2$

Parametrizations of the binary variable

- There are infinitely many different ways/parametrizations how to identify parameters a, β_1, β_2 (to assign some interpretation to these parameters)

- In other words, the binary explanatory variable X reduces the ordinary linear regression model into a classical two sample problem of the form

$$Y = \mu_1 \mathbb{I}_{\{\texttt{group 1}\}} + \mu_2 \mathbb{I}_{\{\texttt{group 2}\}} + \varepsilon$$

  where $\mu_1 \in \mathbb{R}$ represents the unknonw (true) mean of the first random sample, and $\mu_2 \in \mathbb{R}$ stands for the unknown mean of the second sample

- **How to express μ_1 and μ_2 in terms of the model $f(x) = a + bx$?**
  - Parametrization #1: let $\texttt{TRUE} = 0$ and $\texttt{FALSE} = 1$
    $\Longrightarrow \mu_1 = E[Y|X = \texttt{TRUE}] = a$ and $\mu_2 = E[Y|X = \texttt{FALSE}] = a + b$
  - Parametrization #2: let $\texttt{TRUE} = 1$ and $\texttt{FALSE} = 0$
    $\Longrightarrow \mu_1 = E[Y|X = \texttt{TRUE}] = a + b$ and $\mu_2 = E[Y|X = \texttt{FALSE}] = a$
  - Parametrization #3: let $\texttt{TRUE} = -1$ and $\texttt{FALSE} = 1$
    $\Longrightarrow \mu_1 = E[Y|X = \texttt{TRUE}] = a - b$ and $\mu_2 = E[Y|X = \texttt{FALSE}] = a + b$
  - Parametrization #4: let $\texttt{TRUE} = v_1$ and $\texttt{FALSE} = v_2$
    $\Longrightarrow \mu_1 = E[Y|X = \texttt{TRUE}] = a + bv_1$ and $\mu_2 = E[Y|X = \texttt{FALSE}] = a + bv_2$
  - Parametrization #5: let $\texttt{TRUE} = \dots$ and $\texttt{FALSE} = \dots$
    (infinitely many different parametrizations can be used... So, which one to chose?)

Over-parametrization problem

- Typically it is assumed that the regression model contains the intercept parameter $a \in \mathbb{R}$, thus the model is of the form

$$Y = a + \beta_1 \mathbb{I}_{\{X=\mathtt{TRUE}\}} + \beta_2 \mathbb{I}_{\{X=\mathtt{FALSE}\}} + \varepsilon$$

for three unknown parameters $a, \beta_1, \beta_2 \in \mathbb{R}$

Over-parametrization problem

- Typically it is assumed that the regression model contains the intercept parameter $a \in \mathbb{R}$, thus the model is of the form

$$Y = a + \beta_1 \mathbb{I}_{\{X=\texttt{TRUE}\}} + \beta_2 \mathbb{I}_{\{X=\texttt{FALSE}\}} + \varepsilon$$

  for three unknown parameters $a, \beta_1, \beta_2 \in \mathbb{R}$

- Having a qualitative (binary) explanatory variable $X \in \{\texttt{TRUE}, \texttt{FALSE}\}$, the population of $Y \in \mathbb{R}$ can be only split into two sub-populations (using X)
  - sub-population $E[Y|X = \texttt{TRUE}]$ and sub-population $E[Y|X = \texttt{FALSE}]$
  - there are only 2 subgroups (population means – aka equations) to be used
  - but there are three unknown parameters that should be estimated
  - thus, three parameters can not be uniquely estimated from 2 groups

Over-parametrization problem

- Typically it is assumed that the regression model contains the intercept parameter $a \in \mathbb{R}$, thus the model is of the form

$$Y = a + \beta_1 \mathbb{I}_{\{X = \texttt{TRUE}\}} + \beta_2 \mathbb{I}_{\{X = \texttt{FALSE}\}} + \varepsilon$$

  for three unknown parameters $a, \beta_1, \beta_2 \in \mathbb{R}$

- Having a qualitative (binary) explanatory variable $X \in \{\texttt{TRUE}, \texttt{FALSE}\}$, the population of $Y \in \mathbb{R}$ can be only split into two sub-populations (using X)
  - sub-population $E[Y|X = \texttt{TRUE}]$ and sub-population $E[Y|X = \texttt{FALSE}]$
  - there are only 2 subgroups (population means – aka equations) to be used
  - but there are three unknown parameters that should be estimated
  - thus, three parameters can not be uniquely estimated from 2 groups

- this is known as the **over-parametrization** problem and it is typically solved by introducing some additional equation
  (having 3 unknown parameters to estimate and 2 + 1 equations to use)

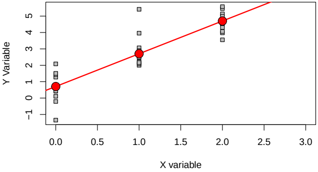
  ↪ for instance, $\beta_1 = 0$ for Parametrization #1; $\beta_2 = 0$ for Parametrization #2; or $\beta_1 + \beta_2 = 0$ for Parametrization #3

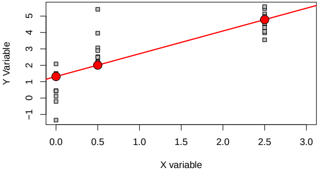
Categorical explanatory variable

❑ Three groups of data (three random samples):

- $X_1, \ldots, X_n \sim (\mu_1, \sigma^2)$
- $Y_1, \ldots, Y_n \sim (\mu_2, \sigma^2)$
- $Z_1, \ldots, Z_n \sim (\mu_3, \sigma^2)$

❑ Different inference about the mean parameters using continuous covariate:

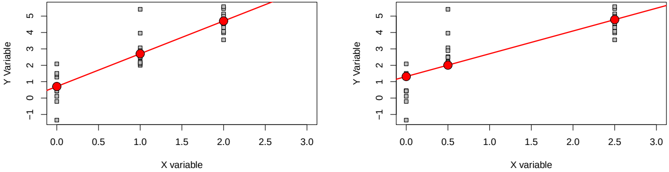




Categorical explanatory variable

- Three groups of data (three random samples):
  - $X_1, \ldots, X_n \sim (\mu_1, \sigma^2)$
  - $Y_1, \ldots, Y_n \sim (\mu_2, \sigma^2)$
  - $Z_1, \ldots, Z_n \sim (\mu_3, \sigma^2)$

- Different inference about the mean parameters using continuous covariate:



- Values $0, 1,$ and 2 (left) and values $0, 0.5,$ and 2.5 (right) only distinguish the groups of observations... they should have no effect on the unknown parameter estimation or the consequent inference...

More general model for $\boldsymbol{X} \in \mathbb{R}$

- For a multi-categorical explanatory variable $X \in \{l_1, \dots l_K\}$ for $K \in \mathbb{N}$ the linear model can be expressed in a straightforward way as

$$Y = a + \beta_1 \mathbb{I}_{\{X = l_1\}} + \beta_2 \mathbb{I}_{\{X = l_2\}} + \dots + \beta_K \mathbb{I}_{\{X = l_K\}} + \varepsilon$$

  for $K \in \mathbb{N}$ different groups (random samples) but using $K + 1$ unknown parameters (i.e., parameters $a, \beta_1, \dots, \beta_k \in \mathbb{R}$)

- Analogous over-parametrization problem occurs again... the parameters are not uniquely defined... one additional equation is needed again

- Similarly, as in the binary case, there are infinitelly different options but only some of them have a usefull interpretation

- For a notation simplicity, the model can be also expressed as

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

  $\hookrightarrow$ where $\boldsymbol{X} = (1, \mathbb{I}_{\{X = l_1\}}, \dots, \mathbb{I}_{\{X = l_K\}})^\top \in \mathbb{R}^{K+1}$ and $\boldsymbol{\beta} = (a, \beta_1, \dots, \beta_K)^\top \in \mathbb{R}^{K+1}$

Model choice – suitability/accuracy

What if a straight line is not enough? How to improve the model/fit?

- So far, the model $Y = f(X) + \varepsilon$ with $f(x) = a + bx$ was considered
- To **improve** the model, we can use some more flexible function $f(\cdot)$ ...
  (as far as the model is linear in the unknown parameters, e.g., $f(x) = a + bx + cx^2$, or $f(x) = ax^- + bx^+$)

Model choice – suitability/accuracy

What if a straight line is not enough? How to improve the model/fit?

- ❏ So far, the model $Y = f(X) + \varepsilon$ with $f(x) = a + bx$ was considered
- ❏ To **improve** the model, we can use some more flexible function $f(\cdot)$ ...
  (as far as the model is linear in the unknown parameters, e.g., $f(x) = a + bx + cx^2$, or $f(x) = ax^- + bx^+$)

How to access the quality of $\hat{f}(\cdot)$ (the model accuracy) quantitatively?
(with a sufficiently high $p \in \mathbb{N}$, the function $f(x) = a_0 + a_1x + \cdots + a_px^p$ can will interpolate unique data)

- ❏ Using the **training (learning) data** $\{(Y_i, X_i);\ i = 1, \dots, n\}$
- ❏ Using a fresh **testing data** $\{(Y_i, X_i);\ i = n + 1, \dots, N\}$

Model choice – suitability/accuracy

What if a straight line is not enough? How to improve the model/fit?

- ❑ So far, the model $Y = f(X) + \varepsilon$ with $f(x) = a + bx$ was considered
- ❑ To **improve** the model, we can use some more flexible function $f(\cdot)$ ...
  (as far as the model is linear in the unknown parameters, e.g., $f(x) = a + bx + cx^2$, or $f(x) = ax^- + bx^+$)

How to access the quality of $\hat{f}(\cdot)$ (the model accuracy) quantitatively?

(with a sufficiently high $p \in \mathbb{N}$, the function $f(x) = a_0 + a_1x + \cdots + a_px^p$ can will interpolate unique data)

- ❑ Using the **training (learning) data** $\{(Y_i, X_i);\ i = 1, \ldots, n\}$
- ❑ Using a fresh **testing data** $\{(Y_i, X_i);\ i = n + 1, \ldots, N\}$

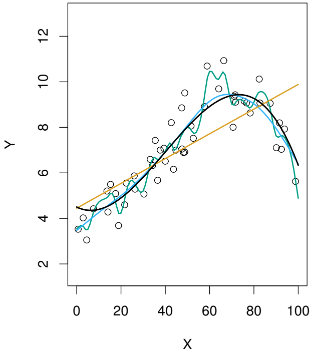
How to access the model quality (its accuracy) qualitatively?

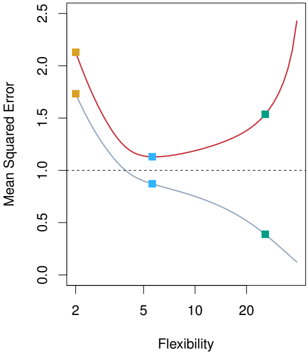
- ❑ Using mathematical/stochastic theory and various statistical tools
- ❑ Using expert knowledge, previous experience, common sense, etc.

Model prediction error – Example I

- □ unknown theoretical model $f(\cdot)$
- □ linear model estimate $\hat{f}_1(\cdot)$
- □ cubic model estimate $\hat{f}_2(\cdot)$
- □ polynomial model estimate $\hat{f}_3(\cdot)$

- □ least squares on training data
- □ least squares on **testing data**

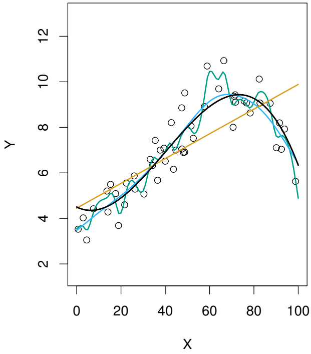


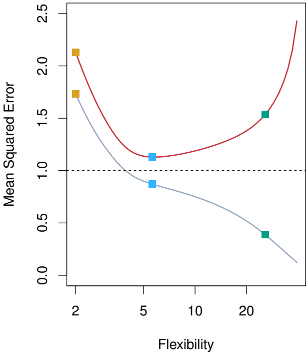


Model prediction error – Example I

- unknown theoretical model $f(\cdot)$
- linear model estimate $\hat{f}_1(\cdot)$
- cubic model estimate $\hat{f}_2(\cdot)$
- polynomial model estimate $\hat{f}_3(\cdot)$

- least squares on training data
- least squares on **testing data**



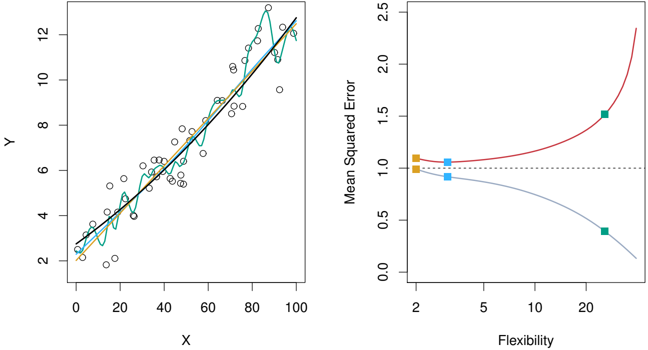


Note, that words linear, cubic, or polynomial refer to the (geometrical) type of the curve that is fitted through the data... but we are still speaking about **linear** regression models here!

Model prediction error – Example II

- ❑ unknown theoretical model $f(\cdot)$
- ❑ linear model estimate $\hat{f}_1(\cdot)$
- ❑ cubic model estimate $\hat{f}_2(\cdot)$
- ❑ polynomial model estimate $\hat{f}_3(\cdot)$

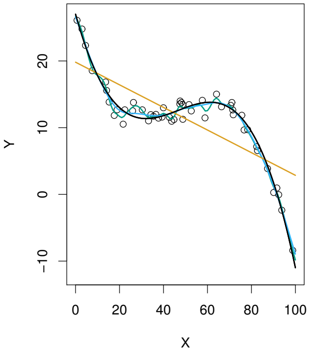
- ❑ least squares on training data
- ❑ least squares on **testing data**

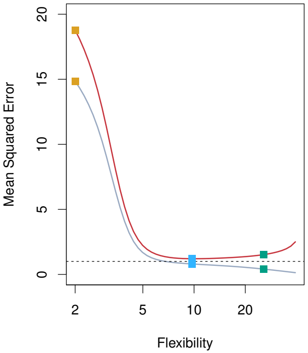


Model prediction error – Example III

- ❑ unknown theoretical model $f(\cdot)$
- ❑ linear model estimate $\hat{f}_1(\cdot)$
- ❑ cubic model estimate $\hat{f}_2(\cdot)$
- ❑ polynomial model estimate $\hat{f}_3(\cdot)$

- ❑ least squares on training data
- ❑ least squares on **testing data**

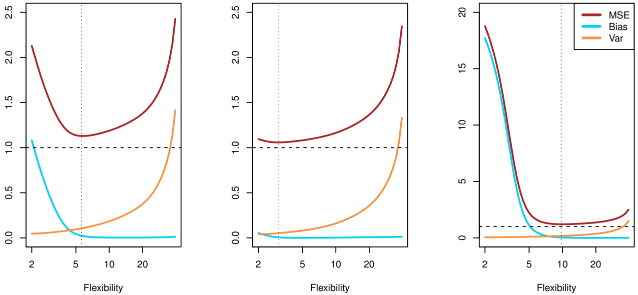




Bias-variance trade-off

Mean Squared Error (MSE):

$$
\begin{aligned}
E[Y - \hat{f}(X)]^2 &= E[(f(X) + \varepsilon - E\hat{f}(X)) - (\hat{f}(X) - f(X))]^2 \\
&= E[\hat{f}(X) - E\hat{f}(X)]^2 + \Big(E\hat{f}(X) - f(X)\Big)^2 + E\varepsilon^2 \\
&= Var\ \hat{f}(X) + \Big(Bias\ \hat{f}(X)\Big)^2 + Var\ \varepsilon
\end{aligned}
$$



Some trade-offs to keep in mind

- ❑ **Mathematics:** parsimonious models vs. “black-box” algorithms
  (transparent models are tractable by mathematical theory)
- ❑ **Probability:** bias vs. variability of the estimate
  (small bias means better accuracy, large variance means high uncertainty)
- ❑ **Utilization:** prediction purposes vs. explanation of the relationship
  (different models are build depending on the primary purpose)
- ❑ **Computation:** computational tractability and time efficiency
  (limitations in algorithmic computations do not allow for arbitrary models)
- ❑ **Interpretation:** simple models are easy to interpret but less accurate
  (complex models are challenging (or even impossible) to be well explained)

Some trade-offs to keep in mind

- ❑ **Mathematics:** parsimonious models vs. “black-box” algorithms
  (transparent models are tractable by mathematical theory)
- ❑ **Probability:** bias vs. variability of the estimate
  (small bias means better accuracy, large variance means high uncertainty)
- ❑ **Utilization:** prediction purposes vs. explanation of the relationship
  (different models are build depending on the primary purpose)
- ❑ **Computation:** computational tractability and time efficiency
  (limitations in algorithmic computations do not allow for arbitrary models)
- ❑ **Interpretation:** simple models are easy to interpret but less accurate
  (complex models are challenging (or even impossible) to be well explained)

“All models are wrong, but some are useful!”

George E. P. Box (1919 – 2013)

Vaguely: Regression vs. Classification

- ❑ What is the nature of the input variable $X \in \mathbb{R}$?
- ❑ What is the nature of the output variable $Y \in \mathbb{R}$?

Vaguely: Regression vs. Classification

❑ What is the nature of the input variable $X \in \mathbb{R}$?

❑ What is the nature of the output variable $Y \in \mathbb{R}$?

❑ ordinary linear regression model

❑ analysis of variance

❑ classification

❑ contingency table

Principal roles of the regression

Regression models and all kinds of data smoothing techniques (e.g., moving averages, weighted averages, splines, parametric smoothing, Whittaker-Henderson) are technically very similar but there is at least one principal and crucial difference – while the data smoothing techniques just smooth the empirical data the regression methods goes beyond as they try to learn important facts about the unknown population that is behind the data generating mechanism – the theoretical model behind the data.

Principal roles of the regression

Regression models and all kinds of data smoothing techniques (e.g., moving averages, weighted averages, splines, parametric smoothing, Whittaker-Henderson) are technically very similar but there is at least one principal and crucial difference – while the data smoothing techniques just smooth the empirical data the regression methods goes beyond as they try to learn important facts about the unknown population that is behind the data generating mechanism – the theoretical model behind the data.

- ❑ **Goal #1**
  with a good choice of the model (i.e., the regression function $f(\cdot)$) we can use the information contained in X (the explanatory variable) to say something relevant about Y (the dependent variable) **But why do we want to do so?**

- ❑ **Goal #2**
  if the set of potential explanatory variables is relatively rich, it can be useful to say which of them are relevant (which ones play a role) to say something about the conditional mean/distribution of Y **Why and how to select good ones?**

- ❑ **Goal #3**
  once we know which covariantes from $X_1, \dots, X_p \in \mathbb{R}$ have an important impact on Y it is often of interest to quantify the effect – i.e., to evaluate how a specific covariate affects the value of Y **Why is this useful in practice?**