Lecture 1 | 17.02.2026

The method of least squares

(a simple linear regression model)

Least squares

- ❑ Introduced independently by A.M. Legendre (1805) and C.F. Gauss (1809)
- ❑ Originally developed for astronomical orbit determination and later theoretically justified via the normal distribution by Gauss
- ❑ Different mathematical perspectives of **least squares**:
    - ❑ geometric projection
    - ❑ algebraic system of equations
    - ❑ probabilistic likelihood (under normality)
    - ❑ numerical (convex) optimization
    - ❑ statistically the best estimate for unknown truth
- ❑ **However, not exclusively related to the regression fremework only...**

Regression (overview)

- Historically, an accidental word used by Francis Galton (1822 – 1911) because the heights of sons, while following the tendency of their parents (tall parents had tall sons, small parents had small sons), tend to return – "regress" – towards the mediocrity/median/average (population stability).
- Nowadays, "regression" is understood as a technique for fitting functional relationships (not necessarily linear, nor parametric ones) to some data (regardless of whether the "slope" or the direction is positive, or negative).
- Mathematically, the regression provides an explicit analytical expression for a (stochastic) relationship between one or more 'input' variable(s) – typically denoted as $\boldsymbol{X} \in \mathbb{R}^p$ and an 'output' variable $Y \in \mathbb{R}$.
- Generally, this relationship (i.e., regression model) can be expressed as

$$Y = f(\boldsymbol{X}) + error$$

for some well-specified (but unknown) function f (model) and some unobserved random noise (errors, fluctuations, or disturbances).

Simple regression model fundaments

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 univariate random variable which represents the independent variable (explanatory variable/model covariate)
- ❑ $\varepsilon \in \mathbb{R}$ represents a latent random variable – an unobserved random error

Simple regression model fundaments

General/generic model formulation (ordinary linear regression)

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

- ❑ $Y \in \mathbb{R}$ is a random variable, the covariate of interest (dependent variable)
- ❑ $X \in \mathbb{R}$ is a univariate random variable which represents the independent variable (explanatory variable/model covariate)
- ❑ $\varepsilon \in \mathbb{R}$ represents a latent random variable – an unobserved random error

- ❑ the simplest parametric model can take the form $f(x) = a + bx$ which is the **systematic part** (a straight line fitted through the data)
- ❑ the **non-systematic part** is an irreducible (unobserved) error – meaning that even if we observe specific realizations of X and Y and we know $a, b \in \mathbb{R}$ there is still some uncertainty not captured by the model

Simple regression model fundaments

General/generic model formulation (ordinary linear regression)

$$E\left[Y|X=x\right]=a+bx$$

❏ $Y \in \mathbb{R}$ is a random variable, the covariate of interest (dependent variable)

❏ $X \in \mathbb{R}$ is a univariate random variable which represents the independent variable (explanatory variable/model covariate)

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

❏ the simplest parametric model can take the form $f(x)=a+bx$ which is the **systematic part** (a straight line fitted through the data)

❏ the **non-systematic part** is an irreducible (unobserved) error – meaning that even if we observe specific realizations of X and Y and we know $a, b \in \mathbb{R}$ there is still some uncertainty not captured by the model

❏ instead of using the model $f(x)=a+bx$ for just one value "$X=x$" we would like to estimate some more useful characteristic of the whole distribution of Y when (conditioning on) "$X=x$" – **conditional mean**

Simple (ordinary) linear regression model

- ❑ Theoretical (population model)

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

- ❑ Random sample from the population (i.e., a joint distribution $F_{(Y,X)}$):

$$\mathcal{S} = \{(Y_i, X_i);\ i = 1, \dots, n\}$$

- ❑ Empirical (data) model counterpart (sometimes also with y_i and x_i)

$$Y_i = a + bX_i + \varepsilon_i \qquad i = 1, \dots, n \in \mathbb{N}$$

Principal goals:

- ❑ Estimation of the unknown parameters $a, b \in \mathbb{R}$
- ❑ Estimation of distributional characteristics of $Y|X$ – e.g., $E[Y|X = x]$
- ❑ Prediction of a future outcome of Y_0, for a given $X_0 = x_0$ (known)
- ❑ Forecasting outcomes of Y_0 given $X_0 = x_0$ (uncertainty statement)

Simple (ordinary) linear regression model

- ❑ Theoretical (population model)

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

- ❑ Random sample from the population (i.e., a joint distribution $F_{(Y,X)}$):

$$\mathcal{S} = \{(Y_i, X_i);\ i = 1, \dots, n\}$$

- ❑ Empirical (data) model counterpart (sometimes also with y_i and x_i)

$$Y_i = a + bX_i + \varepsilon_i \qquad i = 1, \dots, n \in \mathbb{N}$$

Principal goals:

- ❑ Estimation of the unknown parameters $a, b \in \mathbb{R}$
- ❑ Estimation of distributional characteristics of $Y|X$ – e.g., $E[Y|X = x]$
- ❑ Prediction of a future outcome of Y_0, for a given $X_0 = x_0$ (known)
- ❑ Forecasting outcomes of Y_0 given $X_0 = x_0$ (uncertainty statement)

↪ both, the estimation and the prediction can be given in terms of some specific point (e.g., point estimate, point prediction) but the forecasting is typically given in terms of some region (interval estimate, interval prediction respectively) with given credibility guarantees

From the mean to a simple regression

- each random variable $Y \sim (\mu, \sigma^2)$ can be decomposed into two parts: a **systematic** (deterministic) and a **non-systematic** (stochastic) part

$$Y = \mu + \varepsilon$$

where $\mu \in \mathbb{R}$ captures the location and $\varepsilon \sim (0, \sigma^2)$ is a zero-mean error term that captures the underlying volatility (uncertainty around the mean) *(i.e., multiple observations taken under the same conditions – the same mean/variance)*

From the mean to a simple regression

- ❑ each random variable $Y \sim (\mu, \sigma^2)$ can be decomposed into two parts: a **systematic** (deterministic) and a **non-systematic** (stochastic) part

$$Y = \mu + \varepsilon$$

where $\mu \in \mathbb{R}$ captures the location and $\varepsilon \sim (0, \sigma^2)$ is a zero-mean error term that captures the underlying volatility (uncertainty around the mean) *(i.e., multiple observations taken under the same conditions – the same mean/variance)*

- ❑ very analogous principle also holds for observations taken under different conditions where such conditions are controled for by the value of $X = x$

$$Y_x = \mu_x + \varepsilon$$

where $\mu_x \in \mathbb{R}$ captures the location of Y under the conditions $X = x$ and, again, $\varepsilon \sim (0, \sigma_x^2)$ models the volatility (under the situation that $X = x$) *(for simplicity, an explicit analytic form, e.g. $\mu_x = a + bx$, is assumed together with $\sigma_x^2 \equiv \sigma^2$)*

Least squares for the mean

- A well known (empirical) estimate for the mean $\mu \in \mathbb{R}$ is the average...
- Taking observations (e.g. a random sample) $Y_1, \dots, Y_n$ under the same conditions where $Y = \mu + \varepsilon$ (in other words, $Y_i \sim (\mu, \sigma^2)$), a typical estimate for $\mu \in \mathbb{R}$ is the sample mean (average)

$$\widehat{\mu} \equiv \overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

- What is the purpose of such estimate? Why is it defined in this way?

Least squares for the mean

- A well known (empirical) estimate for the mean $\mu \in \mathbb{R}$ is the average...
- Taking observations (e.g. a random sample) $Y_1, \dots, Y_n$ under the same conditions where $Y = \mu + \varepsilon$ (in other words, $Y_i \sim (\mu, \sigma^2)$), a typical estimate for $\mu \in \mathbb{R}$ is the sample mean (average)

$$\widehat{\mu} \equiv \overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

- What is the purpose of such estimate? Why is it defined in this way? It solves the minimization problem

$$\widehat{\mu} = \operatorname*{Arg\,min}_{m \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - m)^2$$

which is the empirical version of its theoretical counterpart

$$\mu = \operatorname*{Arg\,min}_{m \in \mathbb{R}} E(Y - m)^2$$

(recall, that $VarY = E(Y - EY)^2$ *thus, the mean has some specific relation wrt. uncertainty)*

Least squares for a simple regression line

- ❑ The same idea is also applied when a specific model, e.g. $f(x) = a + bx$, is used to model the mean μ_x of Y_i's taken under different conditions *(thus, the conditional expectation (mean) of Y given a value of X – expressed as $E[Y|X]$)*

Least squares for a simple regression line

- The same idea is also applied when a specific model, e.g. $f(x) = a + bx$, is used to model the mean μ_x of Y_i's taken under different conditions *(thus, the conditional expectation (mean) of Y given a value of X – expressed as $E[Y|X]$)*
- **Least squares** can be seen as a meassure of the quality of the fit in terms of some "goodness-of-fit" criterion – specifically:
  - **Mean Squared Error:** $f = \operatorname{Arg\,min}_{g \in \mathcal{C}} E[Y - g(X)]^2$ (theoretical functional)
  - **Least Squares:** $\hat{f}_n = \operatorname{Arg\,min}_{g \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} [Y_i - g(X_i)]^2$ (empirical functional)

  for a pre-specified class of functions $\mathcal{C} = \{f(x);\ f(x) = a + bx; a, b \in \mathbb{R}\}$
  - linear line with the intercept parameter a and the slope parameter b
  - for $b = 0$ everything reduces to a simple mean (sample average)

Least squares for a simple regression line

- ❑ The same idea is also applied when a specific model, e.g. $f(x) = a + bx$, is used to model the mean μ_x of Y_i's taken under different conditions *(thus, the conditional expectation (mean) of Y given a value of X – expressed as $E[Y|X]$)*
- ❑ **Least squares** can be seen as a meassure of the quality of the fit in terms of some "goodness-of-fit" criterion – specifically:
  - ❑ **Mean Squared Error:** $f = \operatorname{Arg\,min}_{g \in \mathcal{C}} E[Y - g(X)]^2$ (theoretical functional)
  - ❑ **Least Squares:** $\hat{f}_n = \operatorname{Arg\,min}_{g \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} [Y_i - g(X_i)]^2$ (empirical functional)

  for a pre-specified class of functions $\mathcal{C} = \{f(x);\ f(x) = a + bx; a, b \in \mathbb{R}\}$
  - ❑ linear line with the intercept parameter a and the slope parameter b
  - ❑ for $b = 0$ everything reduces to a simple mean (sample average)
- ❑ **Aim:** Find $\hat{f}_n \in \mathcal{C}$ while using the available data $\{(Y_i, X_i);\ i = 1, \dots, n\}$
  - ❑ restricting on $\mathcal{C}$ we are looking for $\widehat{a}, \widehat{b} \in \mathbb{R}$, such that $\hat{f}_n(x) = \widehat{a} + \widehat{b}x$
  - ❑ the problem reduces to solving a convex minimization problem

$$\min_{g \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} [Y_i - g(X_i)]^2 = \min_{a,b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} [Y_i - (a + bX_i)]^2 \equiv \min_{a,b \in \mathbb{R}} \mathcal{L}(a, b, \mathcal{S})$$

Least squares for a simple regression line

- ❑ The same idea is also applied when a specific model, e.g. $f(x) = a + bx$, is used to model the mean μ_x of Y_i's taken under different conditions *(thus, the conditional expectation (mean) of Y given a value of X – expressed as $E[Y|X]$)*
- ❑ **Least squares** can be seen as a meassure of the quality of the fit in terms of some "goodness-of-fit" criterion – specifically:
  - ❑ **Mean Squared Error:** $f = \operatorname{Arg\,min}_{g \in \mathcal{C}} E[Y - g(X)]^2$ (theoretical functional)
  - ❑ **Least Squares:** $\hat{f}_n = \operatorname{Arg\,min}_{g \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} [Y_i - g(X_i)]^2$ (empirical functional)

  for a pre-specified class of functions $\mathcal{C} = \{f(x);\ f(x) = a + bx; a, b \in \mathbb{R}\}$
  - ❑ linear line with the intercept parameter a and the slope parameter b
  - ❑ for $b = 0$ everything reduces to a simple mean (sample average)
- ❑ **Aim:** Find $\hat{f}_n \in \mathcal{C}$ while using the available data $\{(Y_i, X_i);\ i = 1, \dots, n\}$
  - ❑ restricting on $\mathcal{C}$ we are looking for $\widehat{a}, \widehat{b} \in \mathbb{R}$, such that $\hat{f}_n(x) = \widehat{a} + \widehat{b}x$
  - ❑ the problem reduces to solving a convex minimization problem

$$\min_{g \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} [Y_i - g(X_i)]^2 = \min_{a,b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} [Y_i - (a + bX_i)]^2 \equiv \min_{a,b \in \mathbb{R}} \mathcal{L}(a, b, \mathcal{S})$$

↪ the notation $\mathcal{L}(a, b, \mathcal{S})$ is used to denote a general (arbitrary) loss function $\mathcal{L}(\cdot)$, the set of unknown parameters $a, b \in \mathbb{R}$ and, also, the available dataset $\mathcal{S} = \{(Y_i, X_i);\ i = 1, \dots, n\}$. **The loss function $\mathcal{L}(\cdot)$ can be, however, defined differently.**

Least squares solution

❑ Convex minimization problem

- ❑ minimization of a convex function
- ❑ minimization with respect to a convex set

❑ Normal equations (score equations)

- ❑ partial derivative of $\mathcal{L}(a, b, \mathcal{S})$ with respect to the argument $a \in \mathbb{R}$
- ❑ partial derivative of $\mathcal{L}(a, b, \mathcal{S})$ with respect to the argument $b \in \mathbb{R}$
- ❑ both partial derivatives are set to be equal to zero and solved for $a, b \in \mathbb{R}$

❑ Solutions of the normal equations (under minimal assumptions)

- ❑ Intercept parameter estimate:

$$\widehat{a} = \overline{Y}_n - \widehat{b}\overline{X}_n$$

- ❑ Slope parameter estimate:

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

Least squares solution

Convex minimization problem

- minimization of a convex function
- minimization with respect to a convex set

Normal equations (score equations)

- partial derivative of $\mathcal{L}(a, b, \mathcal{S})$ with respect to the argument $a \in \mathbb{R}$
- partial derivative of $\mathcal{L}(a, b, \mathcal{S})$ with respect to the argument $b \in \mathbb{R}$
- both partial derivatives are set to be equal to zero and solved for $a, b \in \mathbb{R}$

Solutions of the normal equations (under minimal assumptions)

- Intercept parameter estimate:

$$\widehat{a} = \overline{Y}_n - \widehat{b}\overline{X}_n$$

- Slope parameter estimate:

$$\widehat{b} = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y}_n)(X_i - \overline{X}_n)}{\sum_{i=1}^{n} (X_i - \overline{X}_n)^2} = \frac{\widehat{Cov}(Y, X)}{\widehat{Var} X}$$

Least squares solution

❑ Convex minimization problem

- ❑ minimization of a convex function
- ❑ minimization with respect to a convex set

❑ Normal equations (score equations)

- ❑ partial derivative of $\mathcal{L}(a, b, \mathcal{S})$ with respect to the argument $a \in \mathbb{R}$
- ❑ partial derivative of $\mathcal{L}(a, b, \mathcal{S})$ with respect to the argument $b \in \mathbb{R}$
- ❑ both partial derivatives are set to be equal to zero and solved for $a, b \in \mathbb{R}$

❑ Solutions of the normal equations (under minimal assumptions)

- ❑ Intercept parameter estimate:

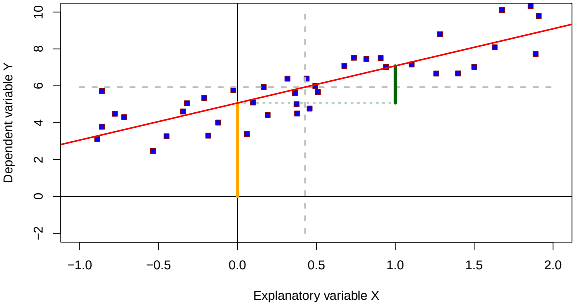
$$\widehat{a} = \overline{Y}_n - \widehat{b}\overline{X}_n$$

- ❑ Slope parameter estimate:

$$\widehat{b} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y}_n)(X_i - \overline{X}_n)}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2} = \frac{\widehat{Cov}(Y, X)}{\widehat{Var} X}$$

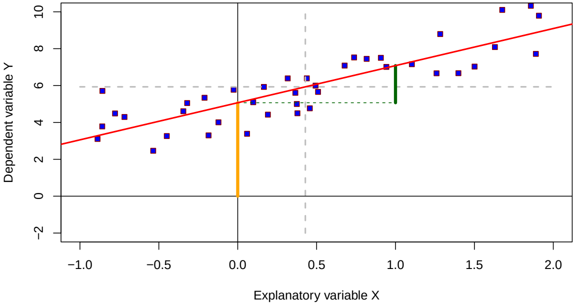
- ❑ the convexity of the optimization problem guarantees a unique solution

Least squares solution – visualization



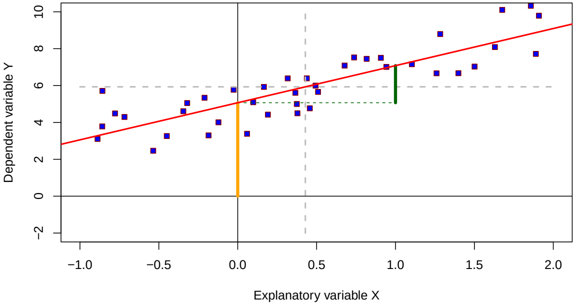
❑ random sample from $F_{(Y,X)}$ — the observed data $\{(Y_i, X_i);\ i = 1, \dots, n\}$

Least squares solution – visualization



- ❑ random sample from $F_{(Y,X)}$ — the observed data $\{(Y_i, X_i);\ i = 1, \dots, n\}$
- ❑ estimated regression model $\hat{f}(x) = \widehat{a} + \widehat{b}x$ $(y = 5.0 + 2.0x)$
- ❑ estimated intercept parameter $\widehat{a} \in \mathbb{R}$ $(\widehat{a} = 5.048)$
- ❑ estimated slope parameter $\widehat{b} \in \mathbb{R}$ $(\widehat{b} = 2.012)$

Least squares solution – visualization

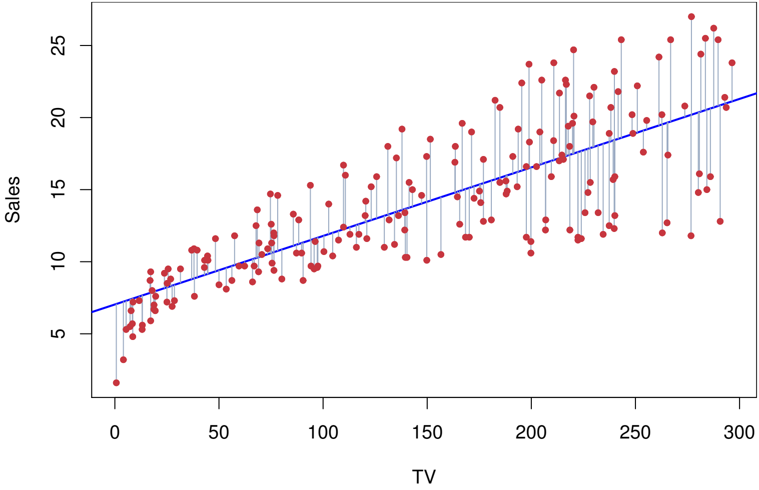


- ❑ random sample from $F_{(Y,X)}$ — the observed data $\{(Y_i, X_i);\ i = 1, \ldots, n\}$
- ❑ estimated regression model $\hat{f}(x) = \widehat{a} + \widehat{b}x$ $(y = 5.0 + 2.0x)$
- ❑ estimated intercept parameter $\widehat{a} \in \mathbb{R}$ $(\widehat{a} = 5.048)$
- ❑ estimated slope parameter $\widehat{b} \in \mathbb{R}$ $(\widehat{b} = 2.012)$
- ❑ the "unknown" true regression model is $f(x) = 5 + 2x$

Some useful jargon

- ❑ **Fitted values**: $\widehat{Y}_i = \widehat{a} + \widehat{b}X_i$
  ($\widehat{Y}_i$ are "estimates" for Y_i values, projected Y_i values onto a line $\widehat{a} + \widehat{b}x$)
- ❑ **Residuals**: $u_i = Y_i - \widehat{Y}_i$
  (u_i are "estimates" for ε_i, projections of Y_i into orthogonal complement)
- ❑ **Residual sum of squares (RSS)**: $\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$
  (the sum of squared residuals – minimization criterion – least squares)
- ❑ **Residual variance**: $\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$ (RSS divided by degrees of freedom)
  (the empirical estimate of the unknown variance of the error term)
- ❑ **Residual standard error (RSE)**: $\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}$
  (estimate for the standard error – a square root of theresidual variance)
- ❑ **Total sum of squares (SST)**: $\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2$
  (the overall data variability with respect to Y when "scaled" by $n-1$)
- ❑ **Multiple R^2 value**: $R^2 = 1 - RSE/SST = (SST - RSE)/SST$
  (relative proportion of the variability explained by the model – the value ($SST - RSE$) represents the overall variability explained by the model and it is given relatively wrt the total variability in the denomitator – SST)

Regression example



Projection from 3D onto 2D – illustration

- For three data points only, $(Y_1, X_1), (Y_2, X_2)$ and (Y_3, X_3), the whole dataset can be represented in terms of two points in the three dimensional (3D) real space, $\boldsymbol{y} = (Y_1, Y_2, Y_3)^\top \in \mathbb{R}^3$ and $\boldsymbol{x} = (X_1, X_2, X_3)^\top \in \mathbb{R}^3$

Projection from 3D onto 2D – illustration

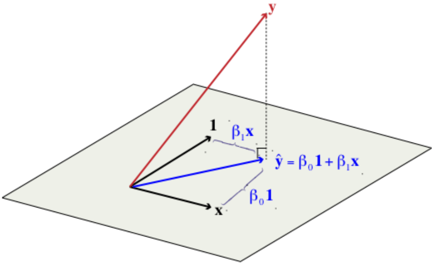
- For three data points only, $(Y_1, X_1), (Y_2, X_2)$ and (Y_3, X_3), the whole dataset can be represented in terms of two points in the three dimensional (3D) real space, $\boldsymbol{y} = (Y_1, Y_2, Y_3)^\top \in \mathbb{R}^3$ and $\boldsymbol{x} = (X_1, X_2, X_3)^\top \in \mathbb{R}^3$
- The underlying model is (still) the simple regression line $f(x) = a + bx$

Projection from 3D onto 2D – illustration

- For three data points only, $(Y_1, X_1), (Y_2, X_2)$ and (Y_3, X_3), the whole dataset can be represented in terms of two points in the three dimensional (3D) real space, $\boldsymbol{y} = (Y_1, Y_2, Y_3)^\top \in \mathbb{R}^3$ and $\boldsymbol{x} = (X_1, X_2, X_3)^\top \in \mathbb{R}^3$
- The underlying model is (still) the simple regression line $f(x) = a + bx$
- Geometric interpretation of the regression is a projection from $\mathbb{R}^3$ into $\mathbb{R}^2$

Projection from 3D onto 2D – illustration

- For three data points only, $(Y_1, X_1), (Y_2, X_2)$ and (Y_3, X_3), the whole dataset can be represented in terms of two points in the three dimensional (3D) real space, $\boldsymbol{y} = (Y_1, Y_2, Y_3)^\top \in \mathbb{R}^3$ and $\boldsymbol{x} = (X_1, X_2, X_3)^\top \in \mathbb{R}^3$
- The underlying model is (still) the simple regression line $f(x) = a + bx$
- Geometric interpretation of the regression is a projection from $\mathbb{R}^3$ into $\mathbb{R}^2$



y
1
$\beta_1 \mathbf{x}$
$\hat{\mathbf{y}} = \beta_0 \mathbf{1} + \beta_1 \mathbf{x}$
$\beta_0 \mathbf{1}$
x

Statistical properties of the estimates $\widehat{a}$ and $\widehat{b}$

- ❑ **The underlying model:** $Y = a + bX + \varepsilon$ (i.e., straight line)
- ❑ **Assumptions:** $E\varepsilon = 0$ and $Var\varepsilon = \sigma^2 < \infty$ (random error properties)

Obtaining now the random sample (Y_i, X_i) with at least two unique values of X_i for $i = 1, \dots, n$ (because the straight line is determined by two unique points) it holds, under the assumptions above, that

Statistical properties of the estimates $\hat{a}$ and $\widehat{b}$

- ❏ **The underlying model:** $Y = a + bX + \varepsilon$ (i.e., straight line)
- ❏ **Assumptions:** $E\varepsilon = 0$ and $Var\varepsilon = \sigma^2 < \infty$ (random error properties)

Obtaining now the random sample (Y_i, X_i) with at least two unique values of X_i for $i = 1, \dots, n$ (because the straight line is determined by two unique points) it holds, under the assumptions above, that

1. **Unbiased estimates:** $E\widehat{a} = a$ and $E\widehat{b} = b$ for all $a, b \in \mathbb{R}$
2. **Linear estimates:** $\widehat{a}$ and $\widehat{b}$ can be expressed as linear functions of Y_i
3. **Best estimates:** $\widehat{a}$ and $\widehat{b}$ are the best linear estimates (they have the smallest variance from all linear unbiased estimates)

- ❏ The result is also known as the Gauss–Markov Theorem – the estimates are known as **BLUE** – Best Linear Unbiased Estimates (**BLUE – nejlepší nestranný lineárný odhad**)

(a formal proof will be given for a multiple linear regression model with multiple predictor variables)

Statistical inference in a regression model

The ultimate task of a regression model is to perform a statistical inference that is based on the underlying (regression) model... This mainly includes confidence (prediction) intervals/regions and hypotheses tests (i.e., managing uncertainty)...

Statistical inference in a regression model

The ultimate task of a regression model is to perform a statistical inference that is based on the underlying (regression) model... This mainly includes confidence (prediction) intervals/regions and hypotheses tests (i.e., managing uncertainty)...

❑ Confidence intervals

(random interval which covers an unknown non-random quantity with a pre-defined probability)

- ❑ typically for the unknown (but fixed) parameters $a, b \in \mathbb{R}$
- ❑ also for the conditional mean parameter $\mu_x = E[Y|X = x]$
- ❑ or some reasonable linear combination, e.g. $c_1 a + c_2 bx$, for $c_1, c_2 \in \mathbb{R}$

❑ Hypothesis tests

(null vs. alternative hypothesis about the unknown but non-random parameters)

- ❑ typically in the form $H_0 : c_1 a + c_2 bx = d$ against a general (both-sided) alternative $H_A : c_1 a + c_2 bx \neq d$
- ❑ performed in terms of a test statistic which is sensitive (large) under the violation of the null hypothesis H_0

Model utilization for prediction

- ## Point prediction
  (one realization of the random variable to somehow characterize another random quantity)
  - what can be the expected outcome/realization of Y if we restrict to a sub-population given by $X = x_0$
  - typically, Y_0 (an outcome of Y when $X = x_0$) is predicted as the estimated conditional mean of Y given $X = x_0$ (i.e., $\widehat{Y}_0 = \widehat{a} + \widehat{b}x_0$)
  - other characteristics can be used of course

- ## Interval prediction
  (random interval which covers unknown but random quantity with a pre-defined probability)

Summary

- the dependent variable Y and the explanatory variable X are assumed to follow (jointly) some (known/unknown) distribution $F_{(Y,X)}(y,x)$
- simple linear regression model $Y = a + bX + \varepsilon$ (population version) (for a continuous response $Y \in \mathbb{R}$ and continuous or binary $X \in \mathbb{R}$)
- random sample (Y_i, X_i), $i = 1, \ldots, n \Longrightarrow Y_i = a + bX_i + \varepsilon_i$ (data model) (realizations $Y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}$ drawn from a joint distribution of (Y, X))
- estimates for the unknown parameters $a, b \in \mathbb{R}$ via convex minimization (minimization based on the mean squared error/least squares respectively)
- under the normal model the estimation based on the maximum likelihood (distribution properties of the estimates $\widehat{a}$ and $\widehat{b}$ given straightforwardly)
- typical inference regarding the parameters $a, b \in \mathbb{R}$ or $E[Y|X = x]$ (performed in terms of confidence intervals or statistical tests respectively)
- utilization of the regression model for estimation/prediction/forecasting (the application is relatively straightforward due to intuitive parameters)