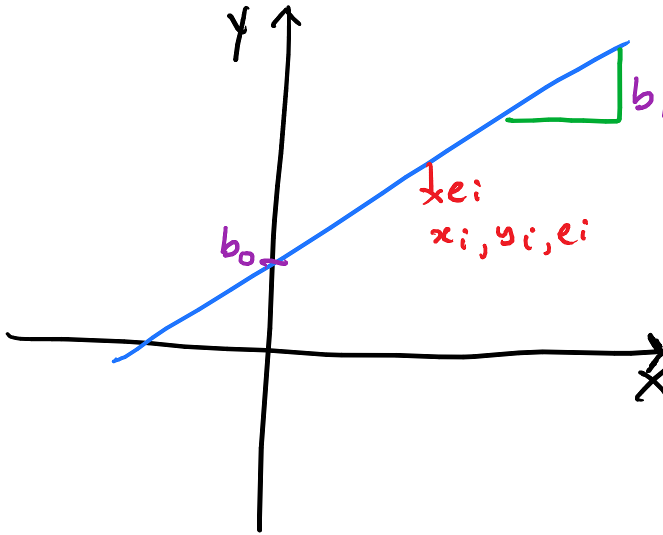


Notation

- Random varibles X, Y, E
- observations / sample x, y, e $\quad i = 1 \dots n$
- Sample vector $\underline{x}, \underline{y}, \underline{e}$

2. Simple linear regression

2.1 Regression model

Covariance and correlation give some idea of direction and strength of linear dependencies. **Simple linear regression** tries to give an explicit linear function that explains one random variable Y by another random variable X

Random veribles

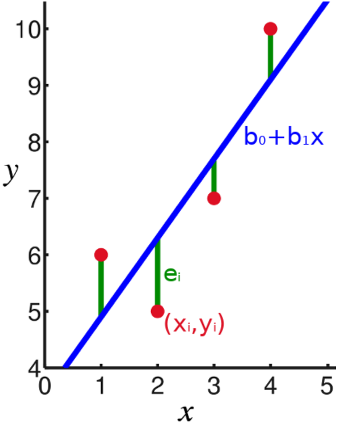
$$Y = b_0 + b_1 X + E.$$

Here E are random variations in Y that cannot be explained by X.

Notation

- Random variables are denoted by capital letters, e.g. X, Y, E.
- Individual observations are denoted by lower-case letters often with an index, e.g. x_i, y_i, e_i.
- Samples are given by vectors, e.g. $\mathbf{x}, \mathbf{y}, \mathbf{e}$.

Regression model



For **random variables** X, Y, E

$$Y = b_0 + b_1 X + E$$

For **samples** x_i, y_i, e_i

$$y_i = b_0 + b_1 x_i + e_i, \quad i = 1, 2, \ldots, n$$

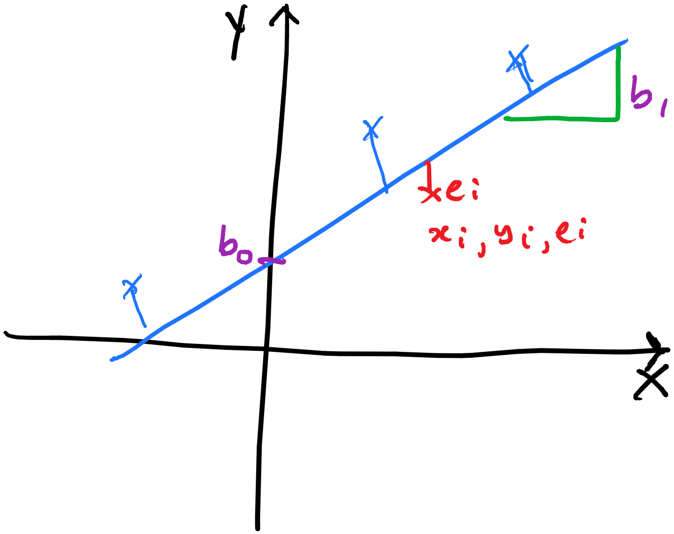
Applications

- b_0 and b_1 can contain interesting information about the data generating process, e.g. constant in the ideal gas law.
- Predict new values of Y given potential values of X, e.g. predict electricity consumption for the next day.

Regression model

Notation

- X is called **independent variable**
  or **regressor**
  or **explanatory variable**
  or **predictor**.
- Y is called **dependent variable**
  or **regressand**
  or **explained variable**
  or **responds variable**.
- E is called **error term** or **unobservable variable**.
- b_0 and b_1 are called **regression coefficients**.
  b_0 is the **intercept** and b_1 the **slope**.
- The regression is called **simple** if only one independent variable X is involved. Otherwise it is called **multiple regression** (next chapter).



Observations: x_i, y_i

unknown: b_0, b_1, e_i

$$\underline{e} = \underline{y} - X \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$$

$$e_i = y_i - b_0 - b_1 x_i \qquad i = 1 \ldots n$$

Idea: minimise $\|\underline{e}\|^2$

$$\min_{b_0, b_1} \|\underline{e}\|^2 = \|\underline{y} - b_0 \underline{1} - b_1 \underline{x}\|^2$$

$$= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i\|^2$$

Regression model for samples

We observe samples of X and Y but not for b_0, b_1, or E.
Given two data vectors $\mathbf{x}$, $\mathbf{y}$ we have the regression equation for the samples

$$y_i = b_0 + b_1 x_i + e_i, \qquad i = 1, 2, \ldots, n$$

$$\text{or } \mathbf{y} = b_0 \mathbf{1} + b_1 \mathbf{x} + \mathbf{e}$$

with $\mathbf{e} = (e_1, e_2, \ldots, e_n)^\top$.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} b_0 \\ \vdots \\ b_0 \end{bmatrix} + \begin{bmatrix} b_1 x_1 \\ b_1 x_2 \\ \vdots \\ b_1 x_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Equivalently,

$$\mathbf{y} = \mathbf{X} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} + \mathbf{e}$$

matrix X

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = b_0 \underline{1} + b_1 \underline{x}$$

vectors of ones

with the $n \times 2$ matrix

$$\mathbf{X} = [\mathbf{1}\ \mathbf{x}] = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

$$\Rightarrow \underline{y} = \underline{X} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} + \underline{e}$$

2.2 Estimation

Notation

Estimated quantities get a hat, e.g.

$$\widehat{b}_0, \quad \widehat{b}_1, \quad \widehat{\mathbf{e}}$$

to distinguish them from the true values b_0, b_1, and $\mathbf{e}$.

Idea

- The values $\mathbf{e} = (e_1, e_2, \ldots, e_n)^\top$ describe by how much $\mathbf{y}$ differs from the linear relation $b_0 + b_1\mathbf{x}$.
- Estimate $\widehat{\mathbf{e}}$, $\widehat{b}_0$, and $\widehat{b}_1$ in a way that keeps $\widehat{\mathbf{e}}$ as small as possible. We will minimise

$$\|\widehat{\mathbf{e}}\|^2 = \sum_{i=1}^{n} \widehat{e}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{b}_0 - \widehat{b}_1 x_i)^2.$$

- This will make $\overline{\widehat{\mathbf{e}}} = 0$ and minimise $var(\widehat{\mathbf{e}})$.

Remember: $\tilde{x} \cdot \perp = 0$ (orthogonal)

$$\text{span}\{\tilde{x}, \perp\} = \text{span}\{\tilde{x} + \bar{x}\perp, \perp\}$$
$$= \text{span}\{x, \perp\}$$

$\frac{\tilde{x}}{\|\tilde{x}\|}, \frac{\perp}{\|\perp\|}$ orthonormal basis

Estimation

Question

Why do we minimise the sum of squares $\sum_{i=1}^{n} \widehat{e}_i^2$ and not $\sum_{i=1}^{n} |\widehat{e}_i|$ or something similar?

Because minimising $\sum_{i=1}^{n} \widehat{e}_i^2$ leads ot a least squares problem and this is "easy" to compute.

Least squares

$$\underline{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

orthonormal basis of span $\{\underline{1}, \underline{x}\}$

Since

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \qquad \text{with} \qquad \mathbf{X} = [\mathbf{1}\ \mathbf{x}]\,,$$

we are looking for the least squares solution to

$$\mathbf{y} = \mathbf{X}\begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$$

This can be found by projecting $\mathbf{y}$ on an orthonormal basis for the span of the columns of $\mathbf{X}$.

last lecture

- Regression model

$$Y = b_0 + b_1 X + E$$

Y: regrassor / dependant varible

X: regressor / independant varible

E: error term / unobserved varible

- Observations: $y_i, x_i \quad i = 1 \dots n$

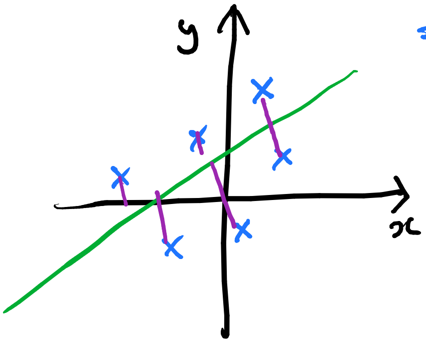
unobserved: $b_0, b_1 e_i \quad i = 1 \dots n$

$y_i = b_0 + b_1 x_i + e_i \qquad i = 1 \dots n$

$\underline{y} = b_0 + b_1 \underline{x} + \underline{e}$

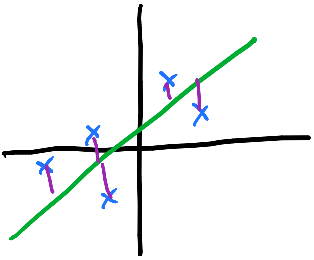
$\underline{y} = \underline{X} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} + \underline{e} \qquad \underline{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$

- Estimation: Find $\hat{b}_0, \hat{b}_1, \hat{e}$



which one to choose

$\|\hat{\underline{e}}\|^2$ is small

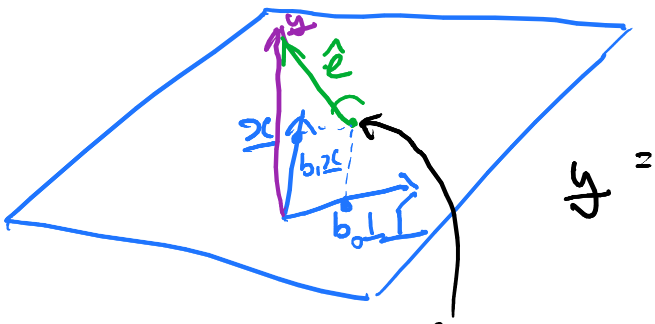


scatter plots

Vector diagram

$\text{span}\{\mathbb{1}, x\}$

"all linear functions in x"



$y = b_0 + b_1 x + e$

$\hat{y}$ orthogonal projection

$\hat{b}_0, \hat{b}_1$ is the solution to the least squares problem

$y = X\begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$ because $e = y - X\begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$

$$\min \|\hat{e}\|^2 = \min \left\| y - X\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \right\|^2$$

- option 1 = for least squares

$$X^T X \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = X^T y \quad \text{normal equation}$$

- option 2 = Orthogonal projection of y on a orthonormal basis of the columns of X

$$\text{span}\left\{\frac{\mathbb{1}}{\|\mathbb{1}\|}, \frac{\tilde{x}}{\|\tilde{x}\|}\right\} = \text{span}\{\mathbb{1}, x\}$$

orthonormal $\qquad$ columns X

$\hat{y}$ orthogonal projection of y on $\frac{\mathbb{1}}{\|\mathbb{1}\|}, \frac{\tilde{x}}{\|\tilde{x}\|}$

$$\hat{y} = \frac{y \cdot \tilde{x}}{\|\tilde{x}\|} \cdot \frac{\tilde{x}}{\|\tilde{x}\|} + \frac{y \cdot \mathbb{1}}{\|\mathbb{1}\|} \cdot \frac{\mathbb{1}}{\|\mathbb{1}\|}$$

$$= \frac{\tilde{x}}{\|\tilde{x}\|^2}(x - \bar{x}\mathbb{1}) + \underbrace{\frac{y \cdot \mathbb{1}}{\|\mathbb{1}\|^2}}_{\frac{1}{n}\sum_{i=1}^n y_i = \bar{y}} \cdot \mathbb{1}$$

$$\|\mathbb{1}\| = \left\| \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right\| = \sqrt{\sum_{i=1}^n 1} = \sqrt{n}$$

$$y \cdot \mathbb{1} = \sum_{i=1}^n 1 \cdot y_i$$

$$= \frac{y \cdot \tilde{x}}{\|\tilde{x}\|^2} \cdot x + \left(\bar{y} - \bar{x}\frac{y \cdot \tilde{x}}{\|\tilde{x}\|^2}\right) \cdot \mathbb{1}$$

$$= \hat{b}_0 \mathbb{1} + \hat{b}_1 x$$

$$\hat{b}_1 = \frac{y \cdot \tilde{x}}{\|\tilde{x}\|^2} \qquad \hat{b}_0 = (\bar{y} - \bar{x}\hat{b})$$

Estimation

Least squares

best estimation

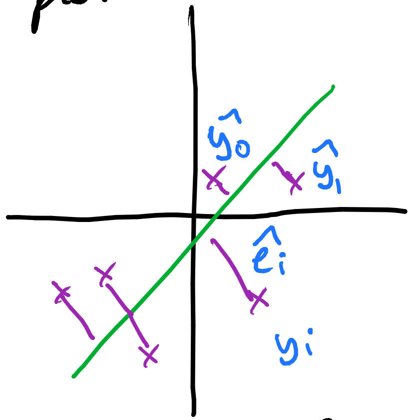
The two vectors

$$\frac{\mathbf{1}}{\|\mathbf{1}\|}, \qquad \frac{\widetilde{\mathbf{x}}}{\|\widetilde{\mathbf{x}}\|} = \frac{\mathbf{x} - \overline{\mathbf{x}}\mathbf{1}}{\|\mathbf{x} - \overline{\mathbf{x}}\mathbf{1}\|}$$

are orthonormal and have the same span as the columns of $\mathbf{X} = [\mathbf{1}\ \mathbf{x}]$. Let $\widehat{\mathbf{y}}$ be the orthogonal projection onto this basis.

$$\begin{aligned}
\widehat{\mathbf{y}} &= \frac{\mathbf{y} \cdot (\mathbf{x} - \overline{\mathbf{x}}\mathbf{1})}{\|\mathbf{x} - \overline{\mathbf{x}}\mathbf{1}\|} \frac{\mathbf{x} - \overline{\mathbf{x}}\mathbf{1}}{\|\mathbf{x} - \overline{\mathbf{x}}\mathbf{1}\|} + \frac{\mathbf{y} \cdot \mathbf{1}}{\|\mathbf{1}\|} \frac{\mathbf{1}}{\|\mathbf{1}\|} \\
&= \frac{\mathbf{y} \cdot (\mathbf{x} - \overline{\mathbf{x}}\mathbf{1})}{\|\mathbf{x} - \overline{\mathbf{x}}\mathbf{1}\|^2} \mathbf{x} + \left(\frac{\mathbf{y} \cdot \mathbf{1}}{\|\mathbf{1}\|^2} - \overline{\mathbf{x}} \frac{\mathbf{y} \cdot (\mathbf{x} - \overline{\mathbf{x}}\mathbf{1})}{\|\mathbf{x} - \overline{\mathbf{x}}\mathbf{1}\|^2} \right) \mathbf{1} \\
&= \underbrace{\frac{\mathbf{y} \cdot \widetilde{\mathbf{x}}}{\|\widetilde{\mathbf{x}}\|^2}}_{\widehat{b}_1} \mathbf{x} + \underbrace{\left(\overline{\mathbf{y}} - \overline{\mathbf{x}} \frac{\mathbf{y} \cdot \widetilde{\mathbf{x}}}{\|\widetilde{\mathbf{x}}\|^2} \right)}_{\overline{\mathbf{y}} - \widehat{b}_1 \overline{\mathbf{x}} = \widehat{b}_0} \mathbf{1}
\end{aligned}$$

scatter plot



$\hat{b}_0 + \hat{b}_1 x$

why is $\underline{y} \cdot \tilde{x} = \tilde{y} \cdot \tilde{x}$ True Because

$$\tilde{y} \cdot \tilde{x} = (\underline{y} - \bar{y}\underline{1}) \cdot \tilde{x}$$

$$= \underline{y} \cdot \tilde{x} - \bar{y}\underbrace{\underline{1} \cdot \tilde{x}}_{=0}$$

$$\rightarrow \hat{b}_1 = \frac{\underline{y} \cdot \tilde{x}}{\|\tilde{x}\|^2} = \frac{\tilde{y} \cdot \tilde{x}}{\|\tilde{x}\|^2} = \mathrm{corr}(\ldots)$$

$$= \mathrm{corr}(\underline{x}, \underline{y}) \cdot \frac{\mathrm{sd}(\underline{y})}{\mathrm{sd}(\underline{x})}$$

$$= \frac{\mathrm{cov}(\underline{x}, \underline{y})}{\mathrm{var}(\underline{x})}$$

Estimation

Result of least squares

$$\widehat{\mathbf{y}} = \widehat{b}_0\mathbf{1} + \widehat{b}_1\mathbf{x} \qquad \text{or}$$
$$\widehat{y}_i = \widehat{b}_0 + \widehat{b}_1 x_i \qquad i = 1, 2, \ldots, n$$

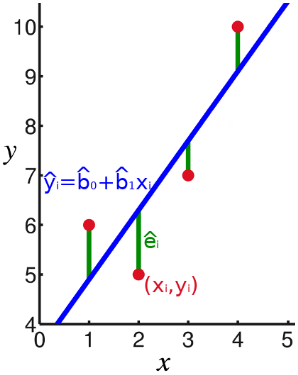
$$\mathbf{y} = \widehat{b}_0\mathbf{1} + \widehat{b}_1\mathbf{x} + \widehat{\mathbf{e}} \qquad \text{or}$$
$$y_i = \widehat{b}_0 + \widehat{b}_1 x_i + \widehat{e}_i \qquad i = 1, 2, \ldots, n$$

with

- $\widehat{b}_1 = \dfrac{\mathbf{y}\cdot\widetilde{\mathbf{x}}}{\|\widetilde{\mathbf{x}}\|^2} = \dfrac{\widetilde{\mathbf{y}}\cdot\widetilde{\mathbf{x}}}{\|\widetilde{\mathbf{x}}\|^2} = corr(\mathbf{x},\mathbf{y})\dfrac{sd(\mathbf{y})}{sd(\mathbf{x})}$
- $\widehat{b}_0 = \overline{\mathbf{y}} - \widehat{b}_1\overline{\mathbf{x}}$
- $\widehat{e}_i = y_i - \widehat{y}_i$
- $\widehat{\mathbf{e}} = \mathbf{y} - \widehat{\mathbf{y}}$

Estimation

E = error term



Notation We call $\hat{e}_i$ = residual

- $\widehat{b}_0$, $\widehat{b}_1$ - ordinary least squares (OLS) estimators
- $\widehat{y}_i$ - fitted values
- $\widehat{\mathbf{y}}$ - vector of fitted values
- $\widehat{e}_i$ - residuals
- $\widehat{\mathbf{e}}$ - vector of residuals
- $\widehat{y} = \widehat{b}_0 + \widehat{b}_1 x$, with $x \in \mathbb{R}$ - regression line.

Interpretation

The regression coefficients have a straight forward interpretation.

- For every increase of x by one unit a change of y by $\widehat{b}_1$ units can be expected on average.
- If $x = 0$, then on average $y = \widehat{b}_0$.

last lecture

- OLS estimations

$$\hat{b}_1 = \frac{\underline{y} \cdot \underline{\tilde{x}}}{\|\underline{\tilde{x}}\|^2}$$

$$\hat{b}_0 = \bar{y} - \hat{b}\,\bar{x}$$

- regression line

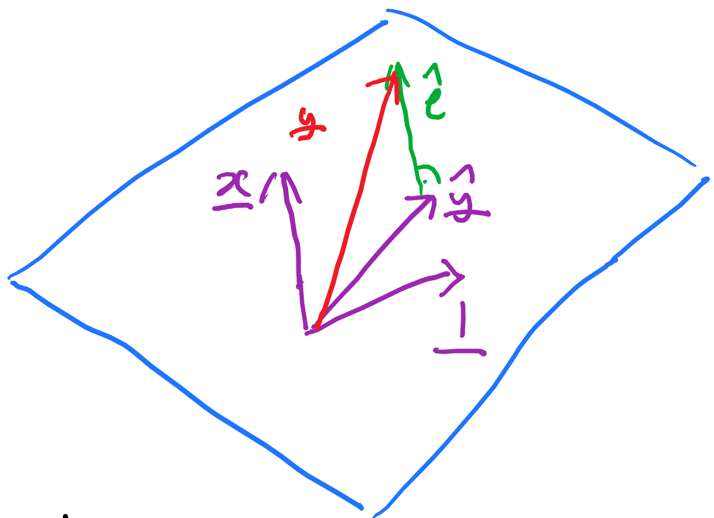
$$\hat{b}_0 + \hat{b}_1 x \qquad x \in \mathbb{R}$$

- Fitted values

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$$

- residuals $\hat{e}_i = y_i - \hat{y}_i$

Hence, $y_i = \hat{b}_0 + \hat{b}_1 x_i + \hat{e}_i$



By construction

- $\hat{\underline{e}}$ & $\underline{1}$
- $\hat{\underline{e}}$ & $\underline{x}$
- $\hat{\underline{e}}$ & $\underline{\tilde{x}}$ are orthogonal
- $\hat{\underline{e}}$ & $\hat{\underline{y}}$

- $\bar{\hat{e}} = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i = \frac{1}{n} \underbrace{\underline{\hat{e}} \cdot \underline{1}}_{=0} = 0$

- $\bar{\hat{y}} = \bar{y}$? Tutorial

- $\mathrm{cov}(\underline{x}, \underline{\hat{e}}) = \frac{1}{n-1} \underbrace{\underline{x} \cdot \underline{\hat{e}}}_{=0} = 0$

$$= \mathrm{corr}(\underline{x}, \underline{\hat{e}})$$

- $\mathrm{cov}(\underline{\hat{y}}, \underline{\hat{e}}) = \frac{1}{n-1} \underline{\hat{y}} \cdot \underline{\hat{e}} = 0$

$$= \mathrm{corr}(\underline{\hat{y}}, \underline{\hat{e}})$$

$$\tilde{\hat{e}} = \underline{\hat{e}}$$

$$\min \|\hat{e}\|^2 = \min \|\tilde{\hat{e}}\|^2 = \min\ n \times \mathrm{Var}(\hat{e})$$

Estimation

Matlab Function 'regress'

Properties

- $\overline{\widehat{\mathbf{e}}} = 0$
- $\overline{\widehat{\mathbf{y}}} = \overline{\mathbf{y}}$
- $cov(\mathbf{x}, \widehat{\mathbf{e}}) = 0$, i.e. $\widetilde{\mathbf{x}} \perp \widehat{\mathbf{e}}$
- $cov(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = 0$, i.e. $\widetilde{\widehat{\mathbf{y}}} \perp \widehat{\mathbf{e}}$

Hence, the residuals of the OLS estimator ...

- contain no information that could be explained by a linear function of $\mathbf{x}$.
- are on average 0.
- have the smallest variance possible.

• Example (Temperature, pressure)

$$\underline{x} = [0, 10, 20, 30, 40, 50]^T$$

$$\bar{x} = 25$$

$$\tilde{x} = x - 25 = [-25, -15, -5, 5, 10, 25]^T$$

$$\|\tilde{x}\|^2 = 25^2 + 15^2 + 5^2 + 5^2 + 15^2 + 25^2$$

$$= 1750$$

$$\underline{y} = [91 \ 95 \ 100 \ 101 \ 107 \ 112]^T$$

$$\bar{y} = 101$$

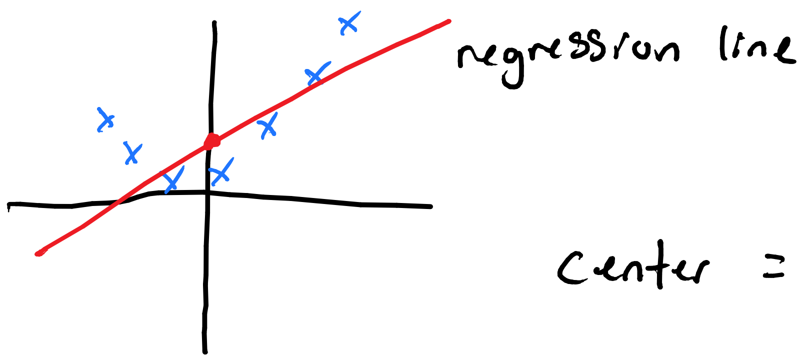
$$\hat{b}_1 = \underline{y} \cdot \tilde{x} = \frac{91(-25) + 95(-15) + 100(-5) + 101(5) + 107(15) + 112(25)}{1750}$$

$$\approx 0.406$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x} \approx 101 - 0.406 \times 25$$

$$\approx 90.9$$

$$\hat{y}_i = 90.9 + 0.406 \times x_i$$



center = temp - mean (temp)

Estimation

Example Temperature and pressure

$$
\begin{aligned}
\widetilde{\mathbf{x}} &= (-25, -15, -5, 5, 15, 25)^\top \\
\mathbf{y} &= (91, 95, 100, 101, 107, 112)^\top \\
\overline{\mathbf{y}} &= 101 \qquad \text{and} \qquad \overline{\mathbf{x}} = 25 \\
\mathbf{y} \cdot \widetilde{\mathbf{x}} = \widetilde{\mathbf{y}} \cdot \widetilde{\mathbf{x}} &= \Big((-25)91 + (-15)95 + (-5)100 \\
&\qquad\qquad + (5)101 + (15)107 + (25)112\Big) \\
&= 710 \\
\|\widetilde{\mathbf{x}}\|^2 &= 1750 \\
\widehat{b}_1 &= \frac{\mathbf{y} \cdot \widetilde{\mathbf{x}}}{\|\widetilde{\mathbf{x}}\|^2} = \frac{710}{1750} \approx 0.406 \\
\widehat{b}_0 &= \overline{\mathbf{y}} - \widehat{b}_1 \overline{\mathbf{x}} = 101 - \frac{710}{1750} 25 \approx 90.9.
\end{aligned}
$$

Hence, $\widehat{y} = 90.9 + 0.406x$.

Estimation

Example

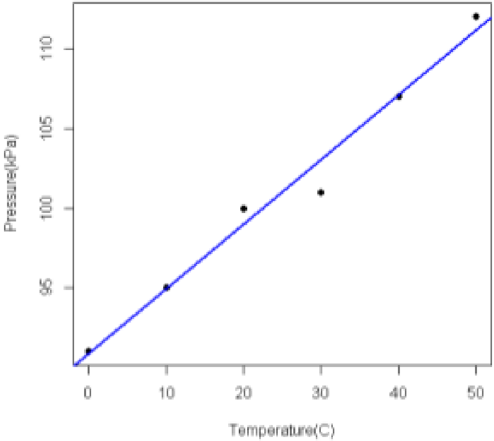


Figure 3: Running example: pressure versus temperature in a boiler.

Estimation

Example

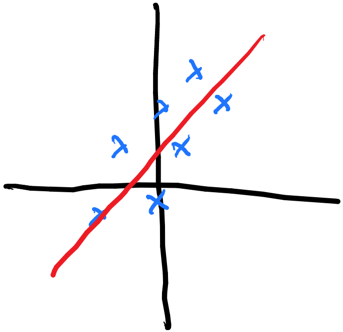
Fitted values

Temp (°C) x_i	Pressure (kPa) y_i	$\widehat{b}_0 + \widehat{b}_1 x_i$
0	91	90.9
10	95	94.9
20	100	99.0
30	101	103
40	107	107
50	112	111

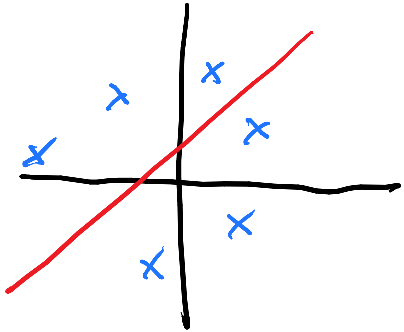
Table 1: Temperature, pressure, and predicted pressure

Interpretation

- The regression line is $\widehat{y} = 90.9 + 0.406x$.
- For every increase (decrease) of temperature by one degree Celsius an increase (decrease) of pressure by $b_1 = 0.406$ kPa can be expected.
- The pressure that we can expect at $0°$C is $b_0 = 90.9$ kPa.



good fit



bad fit

$$\mathrm{var}(y) = \frac{1}{n-1}\|\tilde{y}\|^2$$

$$TSS = (n-1)\,\mathrm{var}(y)$$

2.3 Goodness-of-fit

Notation

- **Total sum of squares** (TSS)

$$TSS = \|\widetilde{\mathbf{y}}\|^2 = \sum_{i=1}^{n}(y_i - \bar{\mathbf{y}})^2$$

This quantifies all variation in the sample **y**.

- **Explained sum of squares** (ESS)

$$ESS = \|\widetilde{\widehat{\mathbf{y}}}\|^2 = \sum_{i=1}^{n}(\widehat{y}_i - \bar{\mathbf{y}})^2$$

The variation in **y** that is explained by the regression line.

- **Residual sum of squares** (RSS)

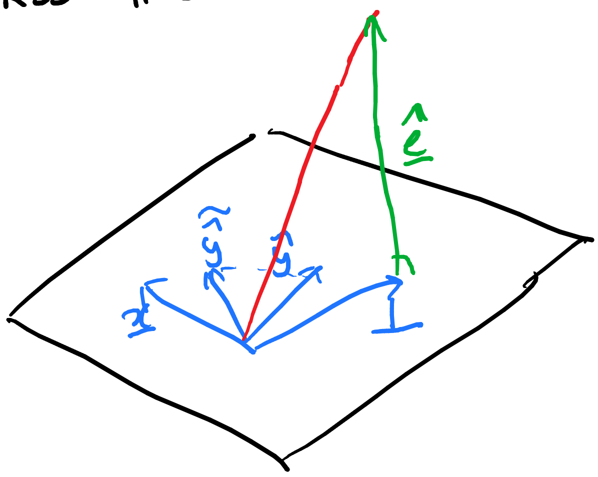
$$RSS = \|\widehat{\mathbf{e}}\|^2 = \sum \widehat{e}_i^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

The variation in **y** that is not explained by the regression line.

last lecture

- properties of OLS $\hat{b}_0, \hat{b}_1, \hat{e}$
- examples
- $TSS = \|\tilde{y}\|^2$, $ESS = \|\hat{\tilde{y}}\|^2$

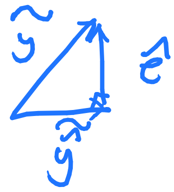
  $RSS = \|\hat{e}\|^2$



(i) $\hat{e}$ and $\hat{\tilde{y}}$ are orthogonal

(ii) $\tilde{y} = \underbrace{\hat{y} + \hat{e}}$

$\tilde{y} = \hat{y} + \hat{e} = \hat{y} + \tilde{\hat{e}} = \hat{\tilde{y}} + \hat{e}$



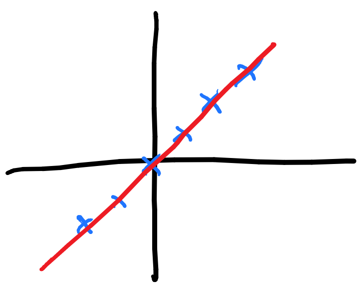
$$\|\tilde{y}\|^2 = \|\hat{\tilde{y}}\|^2 + \|\hat{e}\|^2$$

$$TSS = ESS + RSS$$

$$\Rightarrow R^2 = \underbrace{\frac{ESS}{TSS}}_{\geq 0} = \underbrace{1 - \underbrace{\frac{RSS}{TSS}}_{\geq 0}}_{\leq 1}$$

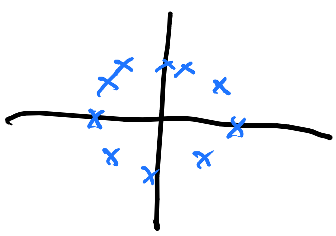
$0 \leq R^2 \leq 1$

$R^2 = 1$



perfect fit

$R^2 = 0$



regression is useless

- Calculating $\boxed{R^2}$

$$R^2 = \frac{\|\tilde{\hat{y}}\|^2}{\|\tilde{y}\|^2} = \frac{\mathrm{var}(\underline{\hat{y}})}{\mathrm{var}(\underline{y})}$$

$$\hat{y} = X\begin{pmatrix}\hat{b}_0 \\ \hat{b}_1\end{pmatrix}$$

Goodness-of-fit

We want to measure how good the regression line fits the data by taking the ratio of explained and total sum of squares.
A large value indicates good fit, a small value indicates a less good fit.

Definition The **R-squared** of the regression is

$$R^2 = \frac{ESS}{TSS}$$

Remember $\quad \widetilde{\widehat{\mathbf{y}}} \perp \widehat{\mathbf{e}}$

Pythagorean theorem

$$\|\mathbf{y} - \overline{y}\mathbf{1}\|^2 = \|\overline{y}\mathbf{1} - \widehat{\mathbf{y}}\|^2 + \|\widehat{\mathbf{y}} - \mathbf{y}\|^2, \quad \text{i.e.} \quad \|\widetilde{\mathbf{y}}\|^2 = \|\widetilde{\widehat{\mathbf{y}}}\|^2 + \|\widehat{\mathbf{e}}\|^2$$

Hence, $TSS = ESS + RSS$ and

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}.$$

Goodness-of-fit

Properties of the R-squared

- $0 \leq R^2 \leq 1$
- If $R^2 = 1$, the fit is perfect. All data points are on the regression line and all residuals equal 0.
- If $R^2 = 0$, the regression line does not fit the data and is useless.
- If R^2 is close to 1, it is very likely that the regression found an actual relation in the data and will probably work well for prediction.
- If R^2 is small, it is unclear whether we found an actual relation or not. Prediction will very likely not work well.

2.4 Confidence intervals

Remember

We assume the regression equation with the true b_0 and b_1 for the random variables X and Y

$$Y = b_0 + b_1 X + E.$$

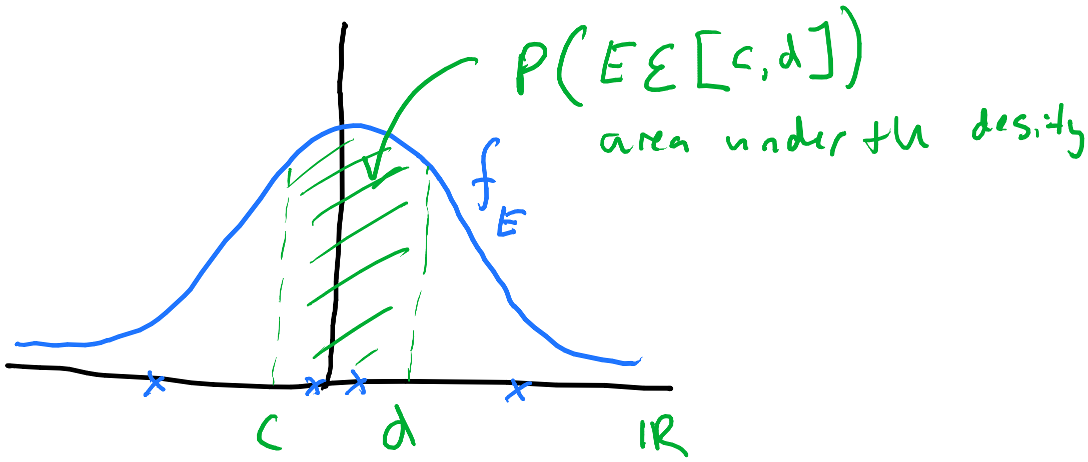
With samples $\mathbf{x}$ and $\mathbf{y}$ we estimate $\widehat{b}_0$ and $\widehat{b}_1$ such that

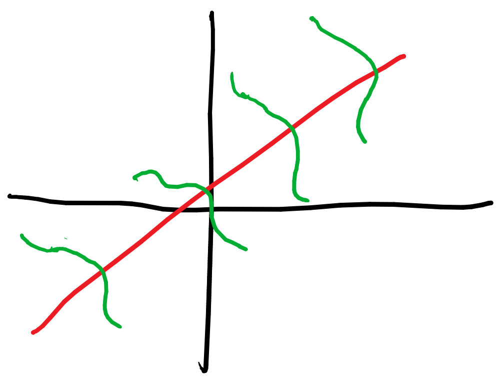
$$\mathbf{y} = \widehat{b}_0 \mathbf{1} + \widehat{b}_1 \mathbf{x} + \widehat{\mathbf{e}}.$$

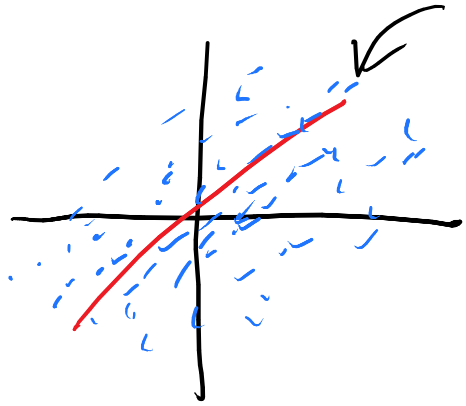
Estimation errors

- Every time we collect data for the random variables X and Y, we will get different data vectors $\mathbf{x}$ and $\mathbf{y}$ and different estimates $\widehat{b}_0$ and $\widehat{b}_1$.
- The estimators $\widehat{b}_0$ and $\widehat{b}_1$ can be understood as random variables.
- The estimation errors $\widehat{b}_0 - b_0$ and $\widehat{b}_1 - b_1$ will be random.
- How much can we trust $\widehat{b}_0$ and $\widehat{b}_1$?

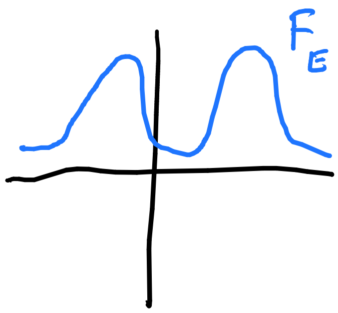
Density is a non negative function

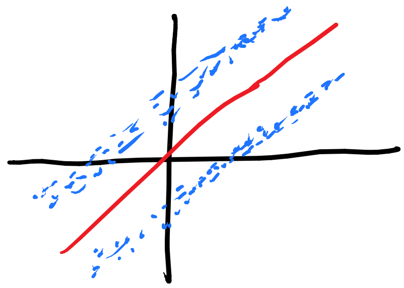




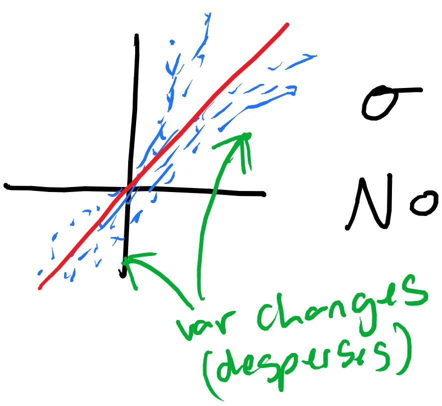


regression line looks okay





Not okay



σ depends on X

Not okay

The assumption holds in many applications, e.g

measurement errors.

Confidence intervals

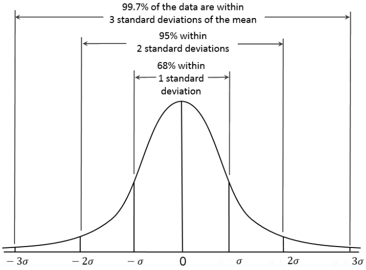
We will answer this question only for b_1 and only under the following assumption.

Assumption

The relation $Y = b_0 + b_1 X + E$ holds with $E \sim \mathcal{N}(0, \sigma^2)$, i.e. normally distributed with mean 0 and variance σ^2.

Interpretation

- When Y is on average equal to $b_0 + b_1 X$, E will have mean 0.
- E has the probability density $f_E(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$



68% chance that $E \in [-\sigma, \sigma]$

95% chance that $E \in [-1.96\sigma, 1.96\sigma]$

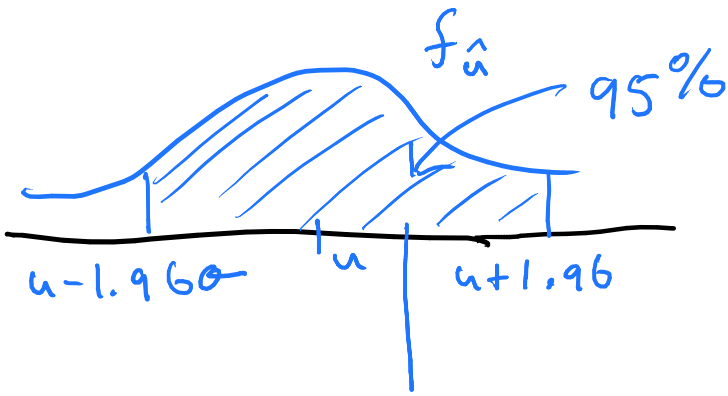
99.7% chance that $E \in [3\sigma, 3\sigma]$

$\alpha = 0.05 \rightarrow 95\%$

$\alpha = 0.1 \rightarrow 90\%$

$\alpha = 0.01 \rightarrow 99\%$

depends on the engineering field



with probability 0.95

$\hat{u} \in [u - 1.96\sigma, u + 1.96\sigma]$

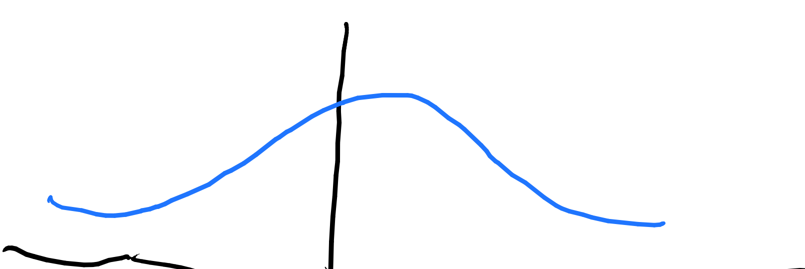
$\Rightarrow |-u - \hat{u}| \leq 1.96\sigma$

$\Rightarrow u \in [\hat{u} - 1.96\sigma, \hat{u} + 1.96\sigma]$

95% confidence interval for u.

$\hat{b}_1 \sim t(n-2)$

t - distribution with $n-2$ d_f



Confidence intervals

Definition

A $100(1-\alpha)\%$ **confidence interval** for b_1 is an interval $[c, d]$ constructed from the data vectors $\mathbf{x}$, $\mathbf{y}$ such that the chance for b_1 to be in $[c, d]$ is $100(1-\alpha)\%$.

In other words

If we collect a lot of different samples and construct a confidence interval for each sample

$$\begin{aligned} \mathbf{x}_1, \mathbf{y}_1 &\rightsquigarrow [c_1, d_1] \\ \mathbf{x}_2, \mathbf{y}_2 &\rightsquigarrow [c_2, d_2] \\ &\vdots \end{aligned}$$

then $b_1 \in [a_j, b_j]$ can be expected in $100(1-\alpha)\%$ of the cases.

Remark

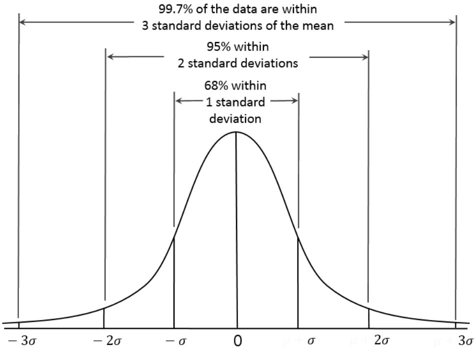
Typically, we chose $\alpha = 0.05$ or $\alpha = 0.1$ and construct 95% or a 90% confidence interval respectively.

Confidence intervals

Example

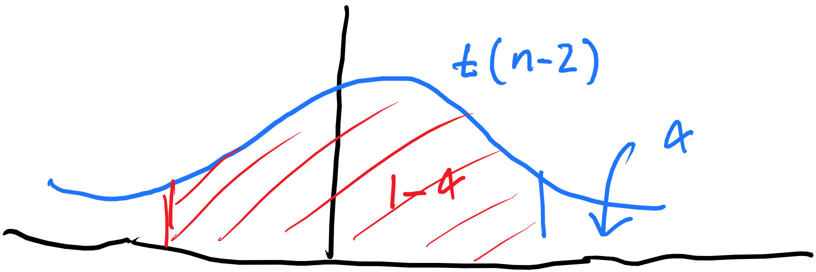
Let $\widehat{u}$ be an estimator for some value u with $\widehat{u} \sim \mathcal{N}(u, \sigma^2)$.

- Hence, with a 95% chance $|\widehat{u} - u| \leq 1.96\sigma$.
- Thus, with a 95% chance $u \in [\widehat{u} - 1.96\sigma, \widehat{u} + 1.96\sigma]$.
- $[\widehat{u} - 1.96\sigma, \widehat{u} + 1.96\sigma]$ is a 95% confidence interval.



$\hat{b}_1 \sim t(n-2)$

t-distribution with $n-2$ d_f



$t_{1-\alpha}(n-2)$

$\alpha/2$ $\alpha/2$

$-t_{1-\frac{\alpha}{2}}(n-1)$ $t_{1-\frac{\alpha}{2}}(n-2)$

Confidence intervals

t-distribution

Confidence intervals for b_1 work similarly, except that we cannot work with the normal distribution.

- Instead we have to work with the so called **t-distribution** with $n-2$ **degrees of freedom**.
- Replace 1.96 by the so called **critical value** $t_{1-\frac{\alpha}{2}}(n-2)$.
- The critical value can be looked up in a table or use the Matlab command `tinv(1-`α`/2,n-2)`.

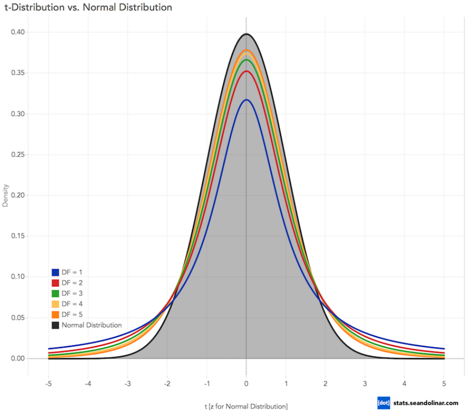
Theorem With probability $100(1-\alpha)\%$

$$b_1 \in \left[\widehat{b}_1 - t_{1-\frac{\alpha}{2}}(n-2)se(\widehat{b}_1)\ ,\ \widehat{b}_1 + t_{1-\frac{\alpha}{2}}(n-2)se(\widehat{b}_1)\right]$$

where $se(\widehat{b}_1) = \sqrt{\frac{1}{n-2}\frac{RSS}{\|\widetilde{\mathbf{x}}\|^2}}$ is the so called **standard error**.

Confidence intervals

t-distribution



- It is similar to the normal distribution but has heavier tails.
- Often called **Student's *t*-distribution**.
- First published 1908 by William Sealy Gosset under the pseudonym "Student" while working for the Guinness Brewery.

Confidence intervals

t-distribution

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Notice

Rows - degree of freedoms $(n-2)$

Columns - confidence level

Look for $t_{1-\frac{\alpha}{2}}$ in the first row.

Or look for α in the row "two-tails".

Or look for the confidence % in the last row.

Confidence intervals

Interpretation

- A small confidence interval indicates that the estimator $\widehat{b}_1$ is quite precise.
- A wide confidence interval indicates that the estimator $\widehat{b}_1$ is less reliable.
- If 0 is not in the confidence interval, we have strong evidence that $b_1 \neq 0$ and that X has an impact on Y.
  We say b_1 is **significant**.
- If we collect more data and increase the sample size n, the confidence interval becomes tighter.

Remember

$$b_1 \in \left[\widehat{b}_1 - t_{1-\frac{\alpha}{2}}(n-2)se(\widehat{b}_1)\ ,\ \widehat{b}_1 + t_{1-\frac{\alpha}{2}}(n-2)se(\widehat{b}_1)\right]$$

$$\text{with } se(\widehat{b}_1) = \sqrt{\frac{1}{n-2}\frac{RSS}{\|\widetilde{\mathbf{x}}\|^2}}$$

Confidence intervals

Example 95% confidence interval for temperature and pressure

$$\begin{aligned}
\widehat{b}_0 &= 90.9,\ \widehat{b}_1 = 0.406 \\
\widehat{y} &= 90.9 + 0.406x \\
\mathbf{y} &= (91, 95, 100, 101, 107, 112)^\top \\
\widehat{\mathbf{e}} &= (91 - 90.9, 95 - 95, 100 - 99, 101 - 103, 107 - 107, 112 - 111)^\top \\
&= (0.1, 0, 1, -2, 1)^\top \\
RSS &= 0.1^2 + 0^2 + 1^2 + -2^2 + 1^2 = 6.01
\end{aligned}$$

Hence, the standard error is

$$se(\widehat{b}_1) = \sqrt{\frac{1}{6-2}\frac{6.01}{1750}} \approx 0.029.$$

Confidence intervals

Example 95% confidence interval for temperature and pressure

We have $n-2=6-2=4$ degrees of freedom and $1-\frac{\alpha}{2}=1-\frac{0.05}{2}=0.975$.

$$t_{0.975}(4) \approx 2.78$$
$$t_{0.975}(4) se(\widehat{b_1}) \approx 2.78 \times 0.029 \approx 0.081$$

This gives the 95% confidence interval for b_1

$$[0.406-0.081,\ 0.406+0.081]=[0.325,\ 0.487].$$

We can see that b_1 is significant.

2.5 *t*-test

Statisticians like to describe the statement that b_1 is significant in a different way. It is common to derive this statement from a statistical **hypothesis test**. Let us understand statistical testing step by step.

Hypothesis

- A statistical test is a tool that gives evidence **against** a hypothesis, which is called the **null hypothesis** H_0.
- If you want evidence that b_1 is significant, you need to start with the opposite statement.

$$H_0 : b_1 = 0$$

- Sometime the **alternative hypothesis** H_1 is introduced as well. Although, it is not necessary to state it explicitly because it is always the opposite of H_0.

$$H_1 : b_1 \neq 0$$

t-test

Test statistic

- We want to find evidence against H_0. However, we assume for a moment that H_0 is true.
- The next step is to find an estimator of which we know the distribution whenever H_0 is true. We know that

$$\frac{\widehat{b}_1}{se(\widehat{b}_1)} \sim t \quad \text{with } n-2 \text{ degrees of freedom}$$

  when $b_1 = 0$. This estimator is called **test statistic**.
- If $\widehat{b}_1 / se(\widehat{b}_1)$ takes a value that is very unlikely for the t-distribution, we have evidence that H_0 was probably not true in the first place.

t-test

Critical values

- Critical values have the property that there is only a $100\alpha\%$ chance that

  $$\widehat{b}_1/se(\widehat{b}_1) < -t_{1-\frac{\alpha}{2}}(n-2) \qquad \text{or} \qquad t_{1-\frac{\alpha}{2}}(n-2) < \widehat{b}_1/se(\widehat{b}_1)$$

  when $\frac{\widehat{b}_1}{se(\widehat{b}_1)} \sim t$ with $n-2$ degrees of freedom.
- Hence, when

  $$\left|\widehat{b}_1/se(\widehat{b}_1)\right| > t_{1-\frac{\alpha}{2}}(n-2)$$

  we **reject** the null-hypothesis on the $100(1-\alpha)\%$ **confidence level**.
- If

  $$\left|\widehat{b}_1/se(\widehat{b}_1)\right| \leq t_{1-\frac{\alpha}{2}}(n-2),$$

  we **fail to reject** the null-hypothesis. We have neither evidence for nor against the null-hypothesis!

t-test

Note

- The test described above is called **two sided *t*-test**.
- The test rejects H_0 when 0 is not in the $100(1-\alpha)\%$ confidence interval and fails to reject when 0 is in the interval.
- We cannot rule out a null-hypothesis with 100% certainty. We have to live with the fact that we can only reject on a certain confidence level, e.g. 90% or 95%.
- α is called the **significance level** of the test.

t-test

One sided t-test

Sometimes we want to confirm that b_1 is positive (or negative). In this case the **one sided t-test** is the right tool.

- If we want to show that $b_1 < 0$,

$$H_0 : b_1 \geq 0, \qquad H_1 : b_1 < 0.$$

- If H_0 is true, the "worst case" would be $b_1 = 0$. In this case

$$\widehat{b}_1 / se(\widehat{b}_1) \sim t \quad \text{with } n-2 \text{ degrees of freedom.}$$

- $\widehat{b}_1 / se(\widehat{b}_1) < -t_{1-\alpha}(n-2)$ with an α% chance.
- Note we need to use $1-\alpha$ for the one sided test!
- Reject H_0 if $\quad \widehat{b}_1 / se(\widehat{b}_1) < -t_{1-\alpha}(n-2)$.
- Fail to reject if $\quad -t_{1-\alpha}(n-2) \leq \widehat{b}_1 / se(\widehat{b}_1)$.

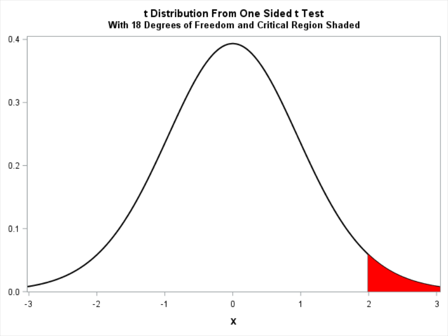
t-test

One sided t-test

- If we want to confirm that $b_1 > 0$,

$$H_0 : b_1 \leq 0.$$

- Reject H_0 if $\quad t_{1-\alpha}(n-2) < \widehat{b}_1 / se(\widehat{b}_1).$
- Fail to reject if $\quad \widehat{b}_1 / se(\widehat{b}_1) \leq t_{1-\alpha}(n-2).$



t-test

Example Temperature and pressure $\widehat{y} = 90.9 + 0.406x$
Can we confirm that b_1 is positive on the 99% confidence level?

$$H_0 : b_1 \leq 0$$

The critical value is $t_{0.99}(4) = 3.747$.

Compute the test statistic:

$$\begin{aligned} \widehat{b_1} &= 0.406 \\ se(\widehat{b_1}) &\approx 0.029 \\ \frac{\widehat{b_1}}{se(\widehat{b_1})} &\approx \frac{0.406}{0.029} \approx 14. \end{aligned}$$

Since $14 > 3.747$ the null hypothesis can be rejected. We have strong evidence that $b_1 > 0$.

2.6 Prediction intervals

In many applications $\widehat{b}_0$ and $\widehat{b}_1$ are estimated and used to predict Y for a new value x^* of X by the corresponding fitted value $\widehat{y}^*$

$$\widehat{y}^* = \widehat{b}_0 + \widehat{b}_1 x^*.$$

How precise is this prediction compared to the actual value y^*?

Theorem

A $100(1-\alpha)\%$ confidence interval for y^* is $[\widehat{y}^* - \tau, \widehat{y}^* + \tau]$ with

$$\tau = t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\frac{RSS}{n-2}}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{\mathbf{x}})^2}{\|\widetilde{\mathbf{x}}\|^2}}.$$

Notation

This interval is called **prediction interval**.

Prediction intervals

$$\widehat{y}^* \pm t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\frac{RSS}{n-2}}\sqrt{1+\frac{1}{n}+\frac{(x^*-\overline{\mathbf{x}})^2}{\|\widetilde{\mathbf{x}}\|^2}}$$

Properties

Prediction intervals are

- wide for small n and if x^* is far away from the mean $\overline{\mathbf{x}}$.
- small for large n and if x^* is close to $\overline{\mathbf{x}}$.

Rule of thumb

Interpolation works much better than extrapolation. Be careful if x^* is outside the range of the date $\mathbf{x}$.

Prediction intervals

$$\widehat{y}^* \pm t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\frac{RSS}{n-2}}\sqrt{1+\frac{1}{n}+\frac{(x^*-\overline{\mathbf{x}})^2}{\|\widetilde{\mathbf{x}}\|^2}}$$

Example Temperature and pressure $\widehat{y} = 90.9 + 0.406x$

Let $x^* = 24°\text{C}$. What is a 95% prediction interval?

The critical value is $t_{0.975}(4) = 2.78$. The prediction interval is given by

$$90.9 + 0.406 \times 24 \pm 2.78\sqrt{\frac{601}{4}}\sqrt{1+\frac{1}{6}+\frac{(24-25)^2}{1750}}$$
$$\approx 100.84 \pm 3.75$$

or equivalently $[97.79,\ 104.59]$.