The theory of linear models

ECON306 – Slides 3
Studenmund Ch. 4–5

Bruno Salcedo

The Pennsylvania State University

Summer 2014

[0]

Classical assumptions

1	Correct specification	$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$	$*$
2	Unbiased errors	$\mathbb{E}[\,\varepsilon_i\,] = 0$	$-$
3	Orthogonality	$\mathbb{E}[\,x_i \varepsilon_i\,] = 0$	$***$
4	No serial correlation	$\mathbb{E}[\,\varepsilon_i \varepsilon_j\,] = 0$	$***$
5	Homoskedasticity	$\mathbb{V}[\,\varepsilon_i\,] = \mathbb{V}[\,\varepsilon_j\,]$	$**$
6	No multicollinearity	$\mathbb{E}\left[\,x_i^2\,\right] \neq 0$	$***$
7	Normality	$\varepsilon_i \sim N(0, \sigma_i^2)$	$*$

Example: all the assumptions hold

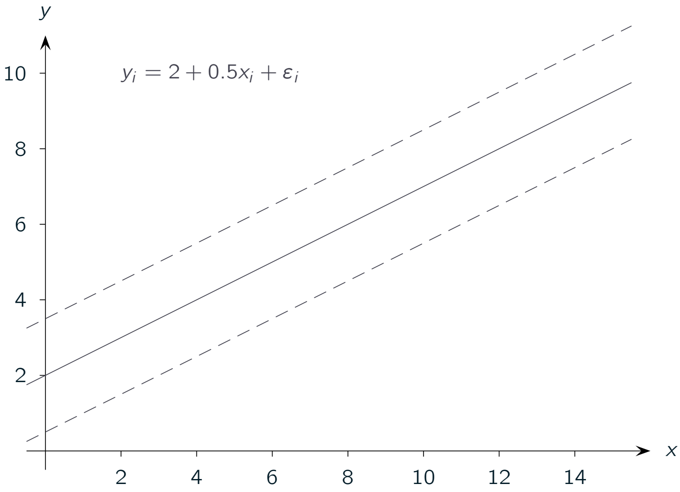
Data generating process

- $\{x_i, \varepsilon_i\}$ are i.i.d.
- x_i is distributed uniformly on $(0, 15)$
- ε_i is distributed $N(0, 0.75)$
- x_i and ε_i are independent
- y_i is given by:

$$y_i = 2 + 0.5x_i + \varepsilon_i$$

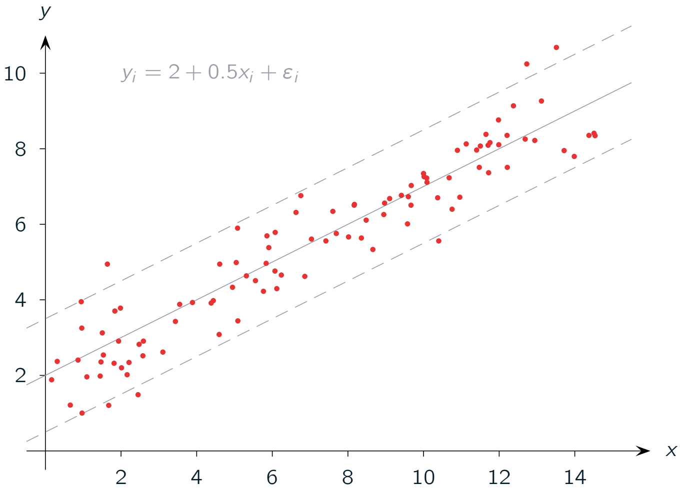
Example: all the assumptions hold

Data generating process



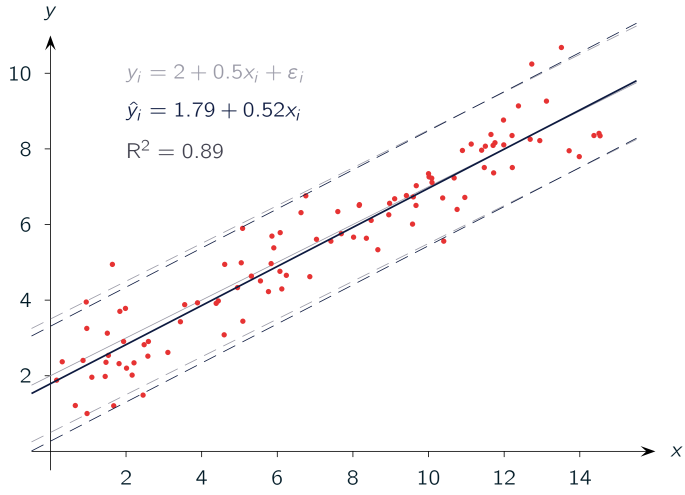
Example: all the assumptions hold

Realized sample with $n = 100$



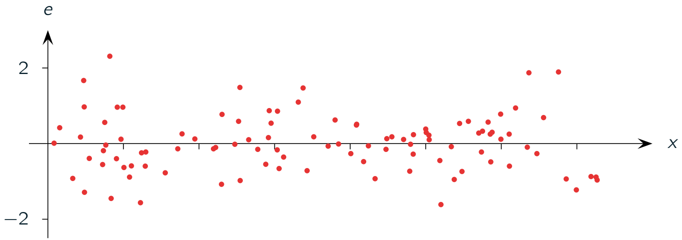
Example: all the assumptions hold

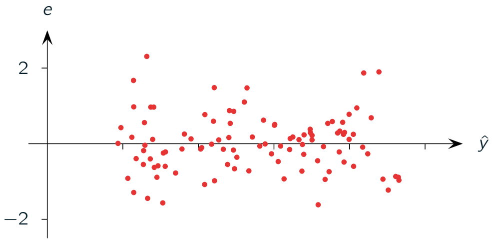
Estimated model ($n = 100$)



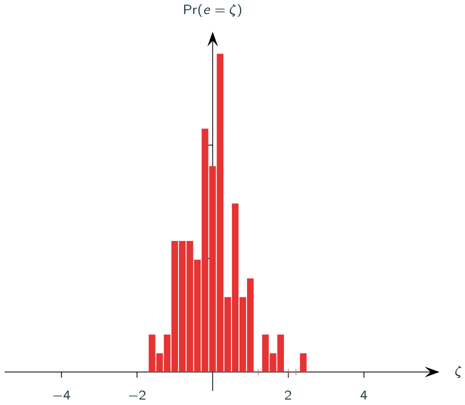
Example: all the assumptions hold

Residuals vs. predictions ($n = 100$)



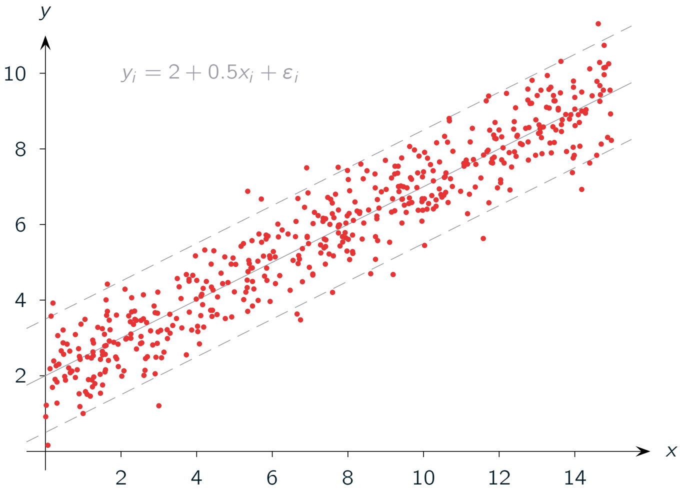


Residual histogram ($n = 100$)



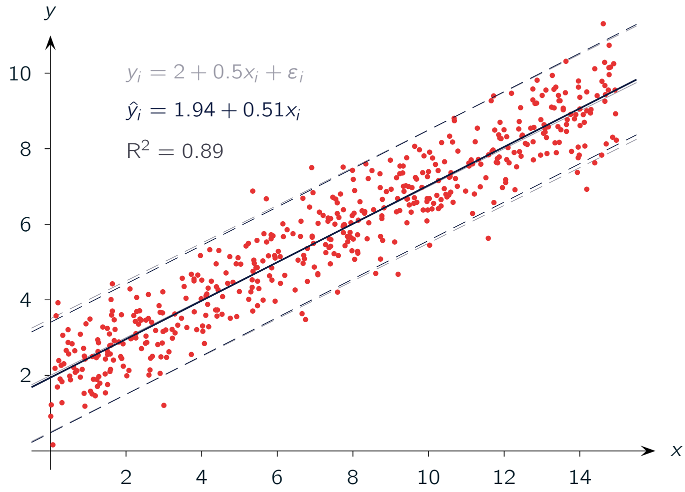
Example: all the assumptions hold

Realized sample with $n = 500$



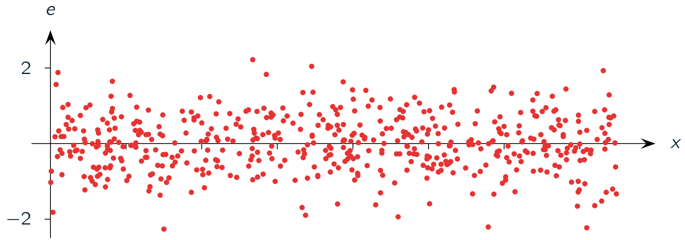
Example: all the assumptions hold

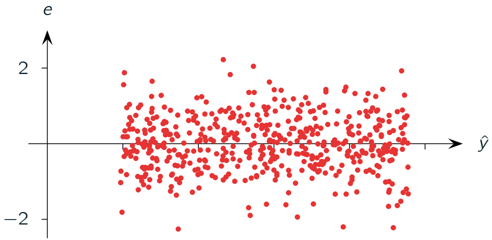
Estimated model ($n = 500$)



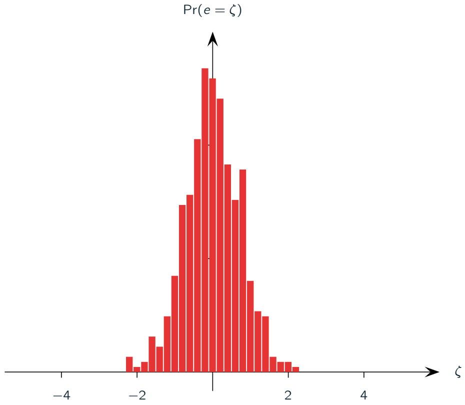
Example: all the assumptions hold

Residuals vs. predictions ($n = 500$)





Residual histogram ($n = 100$)



Correct specification

Correct specification

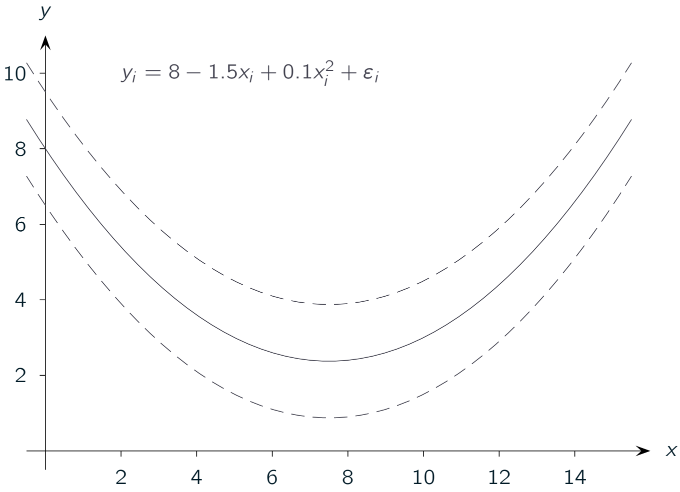
We assume that y_i has a linear relationship with x_i:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- If this is not true we can still run OLS and interpret the coefficients
- However, the interpretation is less appealing
- We can often adjust by making variable transformations

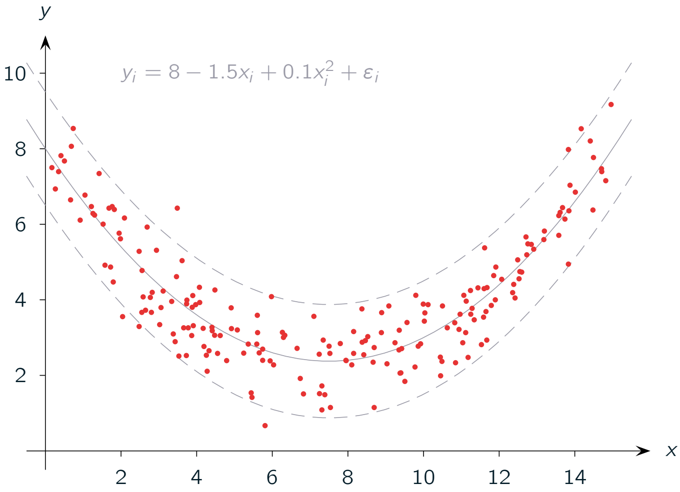
Example: incorrect specification

Data generating process



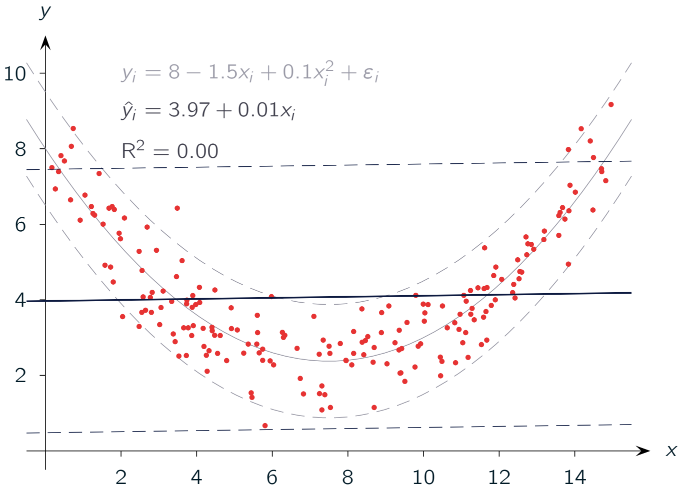
Example: incorrect specification

Realized sample



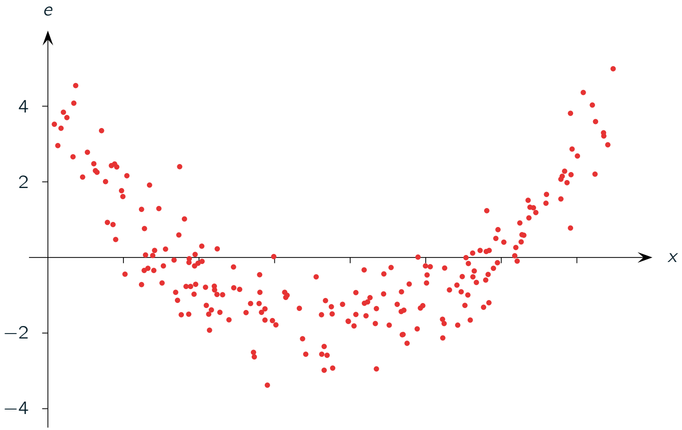
Example: incorrect specification

Estimated model



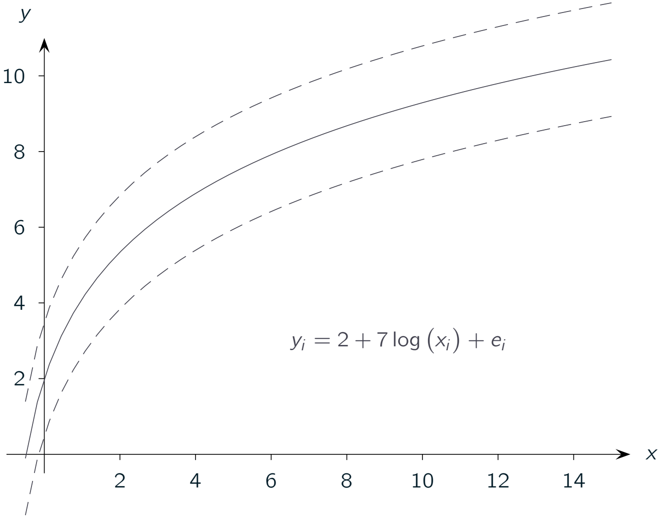
Example: incorrect specification

Residuals vs. regressors



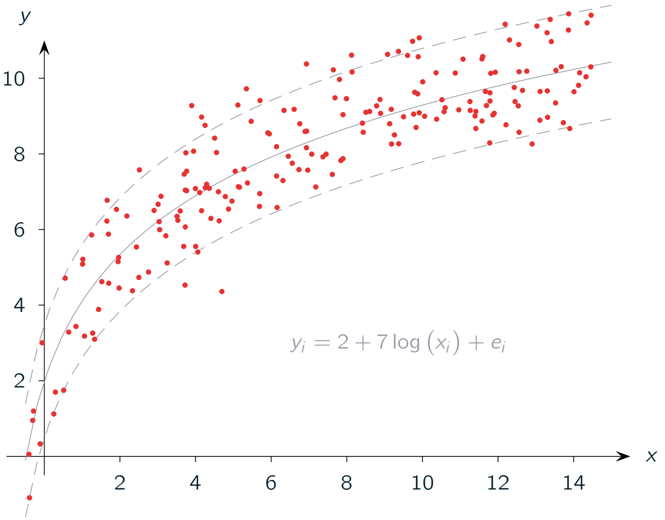
Example: incorrect specification 2

Data generating process



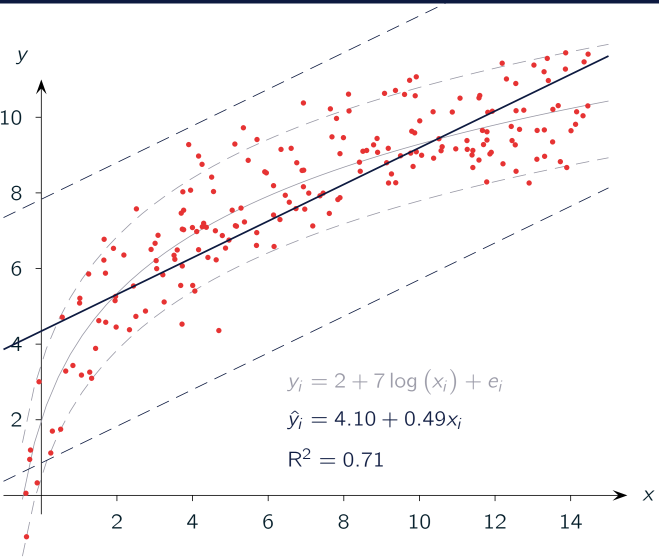
Example: incorrect specification 2

Realized sample



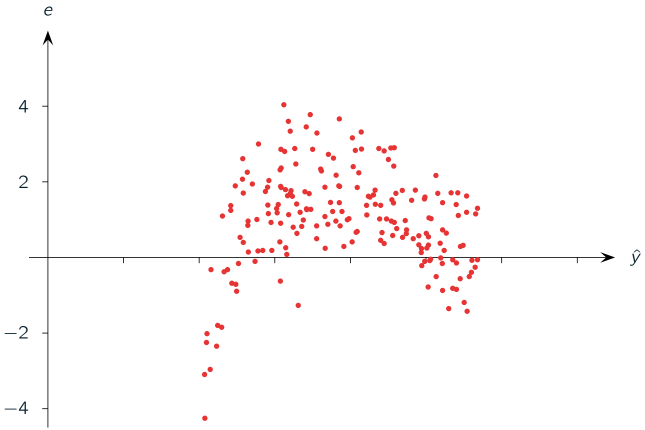
Example: incorrect specification 2

Estimated model



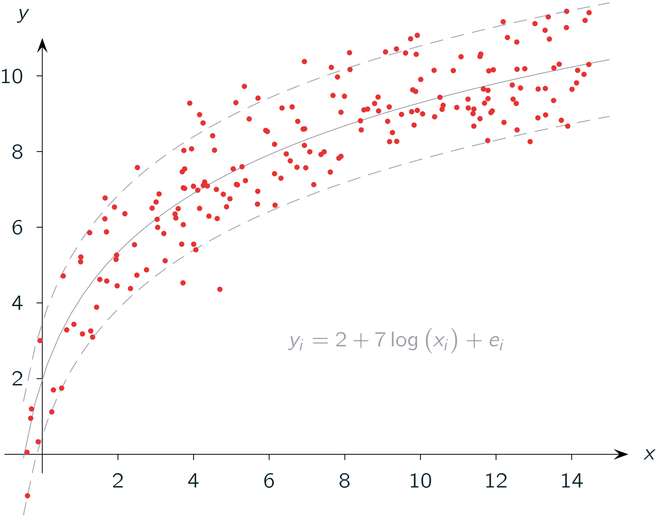
Example: incorrect specification 2

Residuals vs. predictions



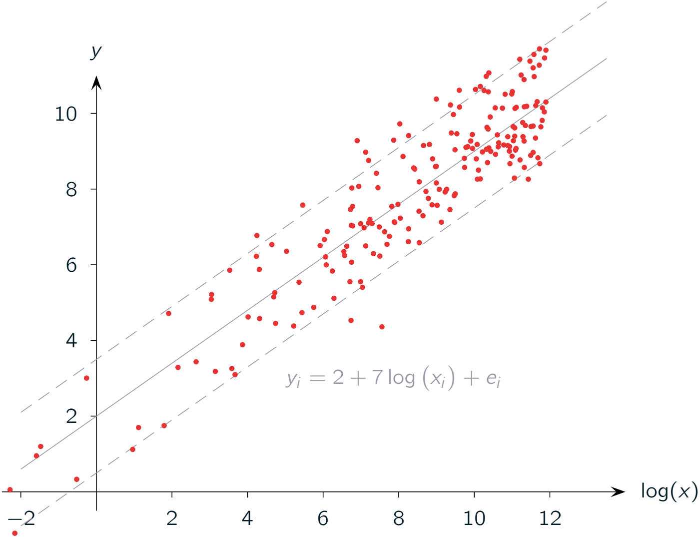
Example: incorrect specification 2

Realized sample



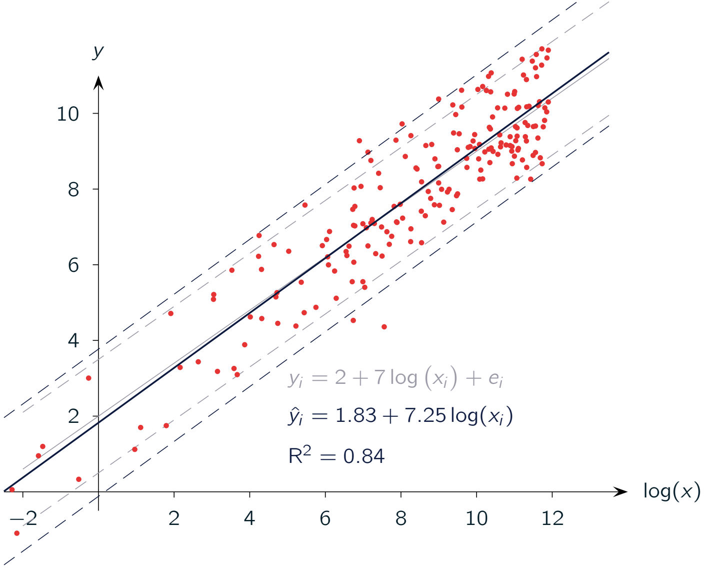
Example: incorrect specification 2

Change of variable



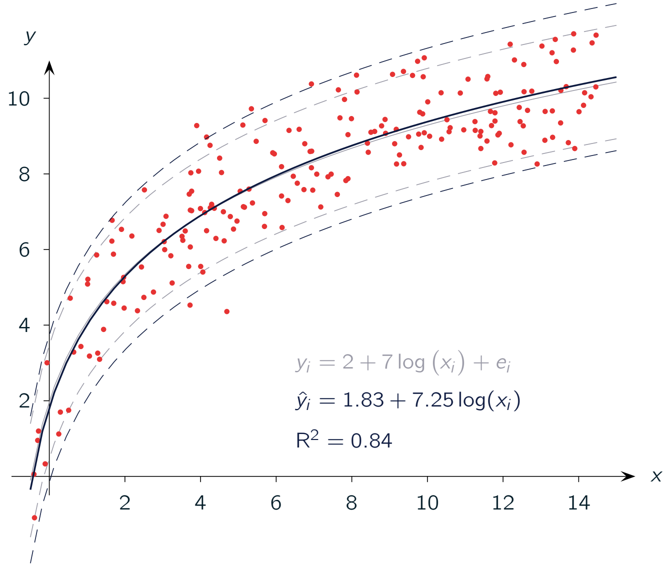
Example: incorrect specification 2

Estimated model



Example: incorrect specification 2

Estimated model



Unbiased errors

Unbiased errors

We assume that the error term has zero mean:

$$\mathbb{E}[\varepsilon_i] = 0$$

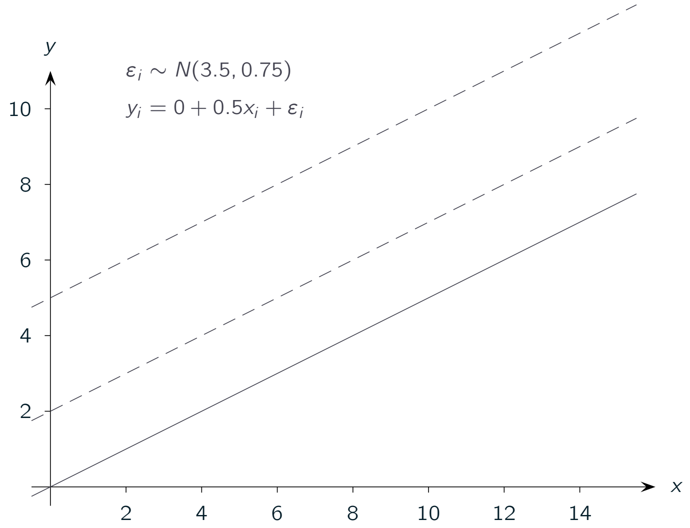
- This is a nominal assumption if we do not care about β_0
- We can still estimate β_1 as long as we include an intercept in our regression
- Simply relabel $\beta_0' = \beta_0 + \mu_\varepsilon$ and $\varepsilon_i' = \varepsilon_i - \mu_\varepsilon$

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \\ &= \left(\beta_0 + \mu_\varepsilon\right) + \beta_1 x_i + \left(\varepsilon_i - \mu_\varepsilon\right) \\ &= \beta_0' + \beta_1 x_i + \varepsilon_i' \end{aligned}$$

$$\mathbb{E}\left[\varepsilon_i'\right] = \mathbb{E}[\varepsilon_i - \mu_\varepsilon] = \mu_\varepsilon - \mu_\varepsilon = 0$$

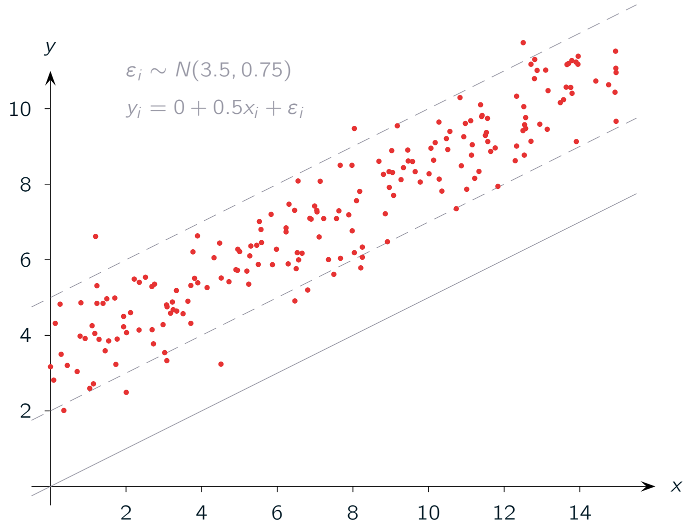
Example: based errors

Data generating process



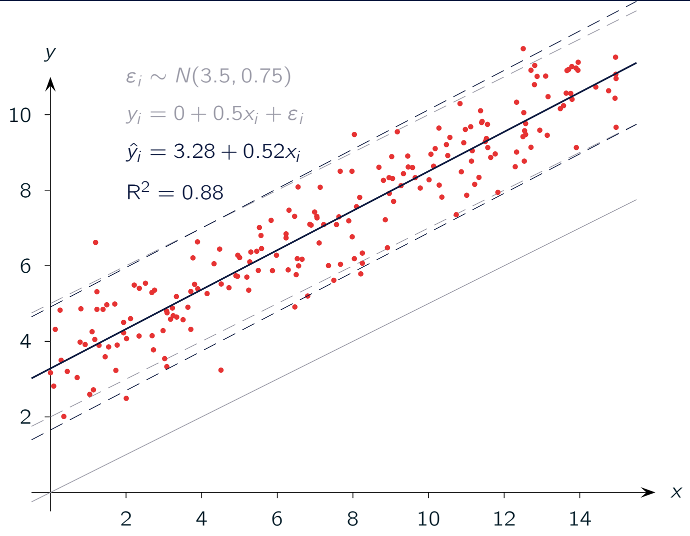
Example: based errors

Realized sample



Example: biased errors

Estimated model



Orthogonality

Orthogonality

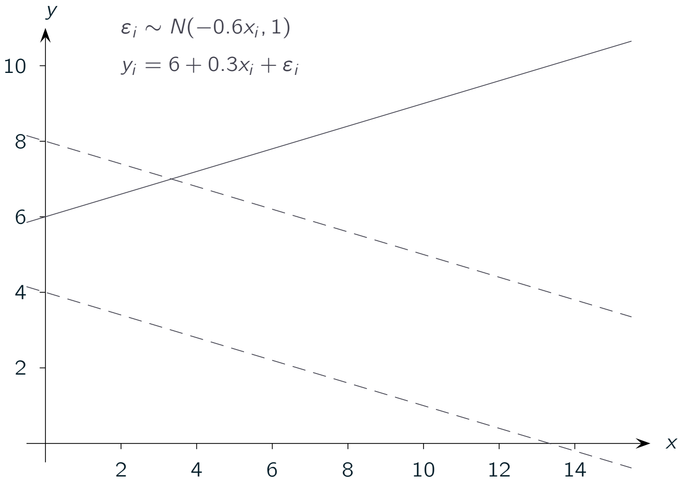
We assume that the regressors are uncorrelated with the error term

$$\mathbb{E}[x_i \varepsilon_i] = 0$$

- x_i is exogenous if this holds, and otherwise endogenous
- Endogeneity is commonly caused by omission of important variables
- When a regressor is endogenous, OLS may attribute to x variation that is actually due to ε
- This may result in bad estimates both for β_0 and for β_1

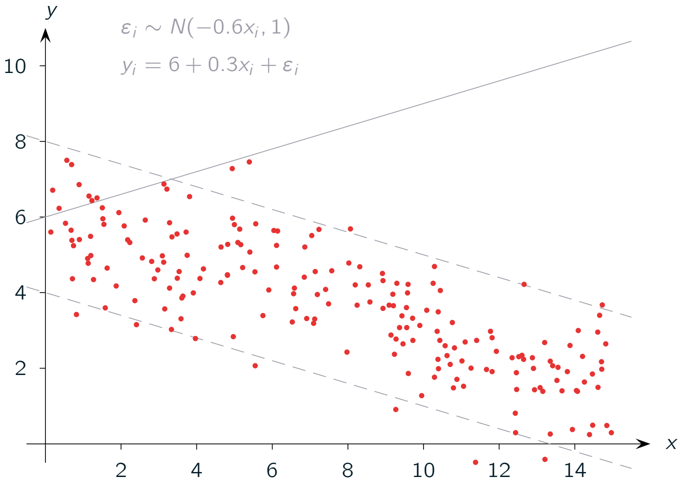
Example: correlation between x and ε

Data generating process



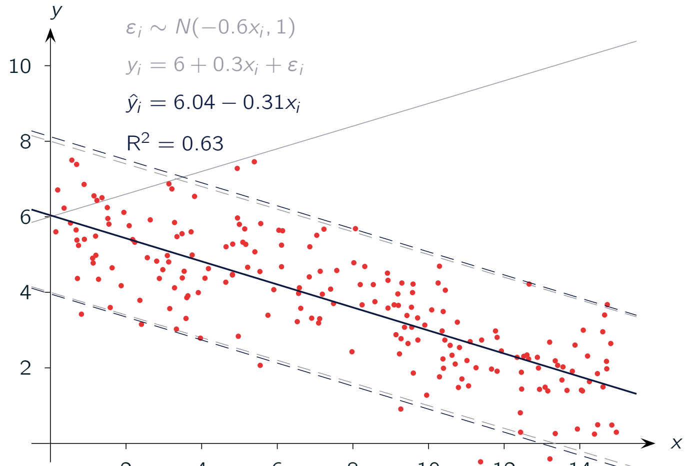
Example: correlation between x and ε

Realized sample



Example: correlation between x and ε

Estimated model



No serial correlation

No serial correlation

We assume that the data comes from a random sample, in particular:

$$\mathbb{E}[\,\varepsilon_i \varepsilon_j\,] = 0$$

- This may be a bad assumption for time series
- The realization of the error in one period may depend on the realization in the past period
- This makes the interpretation of OLS estimates problematic

Homoskedaticity

Homoskedaticity

We assume that the error term has constant variance:

$$\mathbb{V}[\varepsilon_i] = \mathbb{V}[\varepsilon_j] = \sigma_\varepsilon^2$$

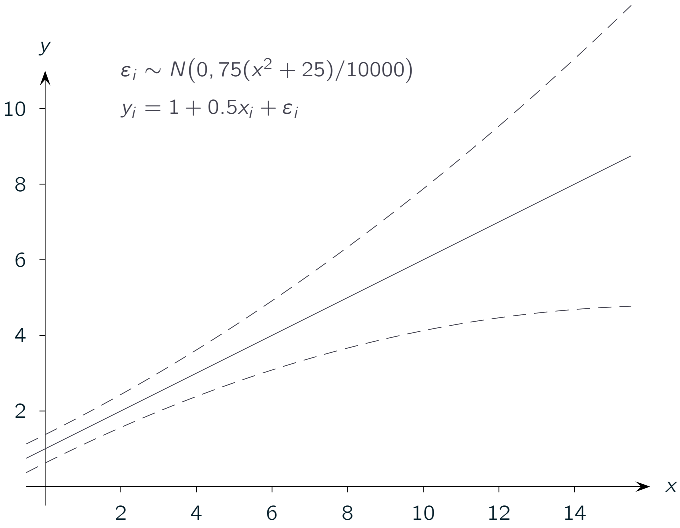
- (Homo = equal) + (skedasticity = variance)
- Otherwise we say that we have heteroskedasticity

- It is not important for estimation
- We don't use/need any assumptions to compute OLS or interpret the coefficients

- It is important for inference but is easily fixed using robust variance estimators

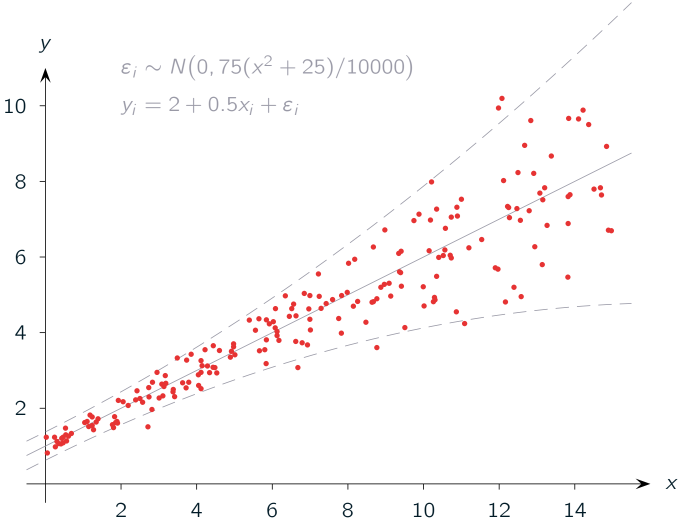
Example: Heteroskedasticity

Data generating process



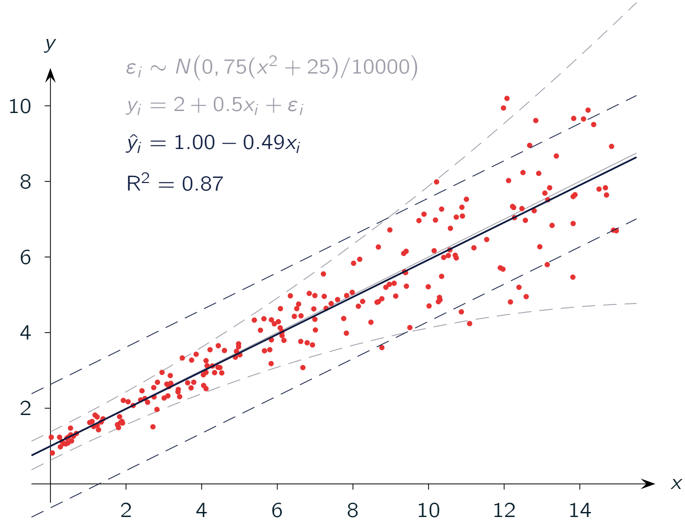
Example: Heteroskedasticity

Realized sample



Example: correlation between x and ε

Estimated model



Multicolinearity

No multicolinearity

We assume that the regressors have positive variance:

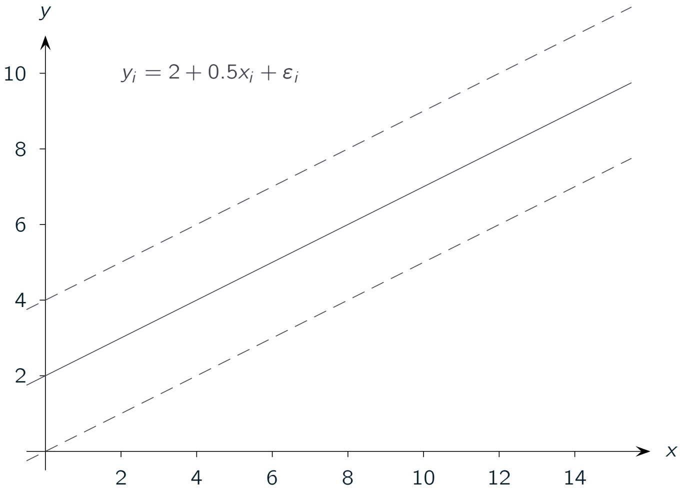
$$\mathbb{E}\left[x_i^2\right] > 0$$

- To measure the impact of changes in x on y, x has to change
- OLS divides by the variance of x, it can't be done if it is exactly 0

- Problems may arise with imperfect colinearity: when $\mathbb{V}\left[x\right]$ is small
- The estimation and numerical errors may generate inaccurate estimates!!

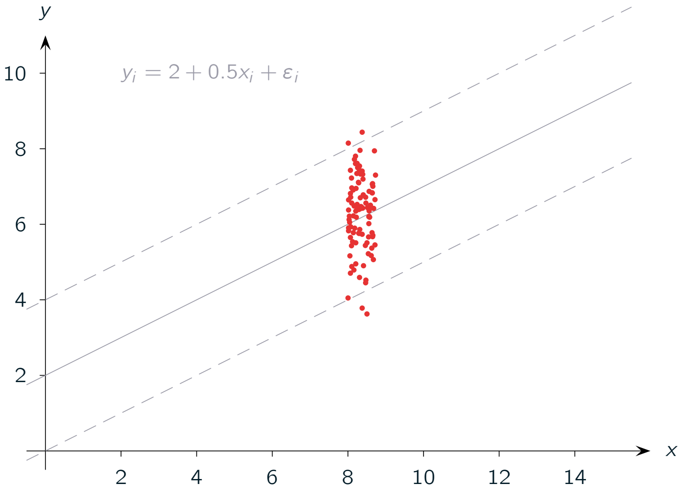
Example: multicolinearity

Data generating process



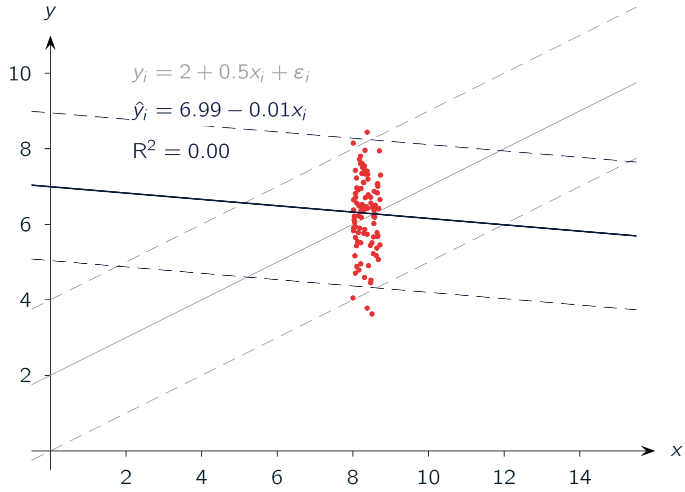
Example: all the assumptions hold

Realized sample



Example: all the assumptions hold

Estimated model



Normality

Normality

The error terms are normally distributed:

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

- This assumption allows to determine the (finite sample) distribution of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$
- It is important for inference but not for estimation
- It can be replaced with the assumption of having a large sample (asymptotic distribution)

[0]

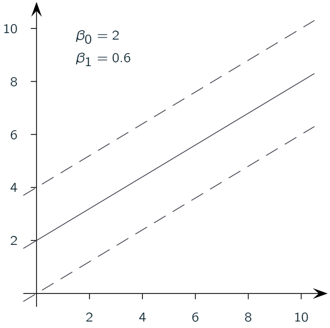
Sampling distribution

- Assume that the classical assumptions 1–7 hold
- What can we say about the OLS estimates?
    - Are they good estimates of the true data generating process?
        - Are they unbiased?
        - Are they efficient?
        - Are they consistent?
    - Can we use the OLS estimates to make inference?
- To answer these questions we need to understand their sampling distribution

Example: sampling distribution

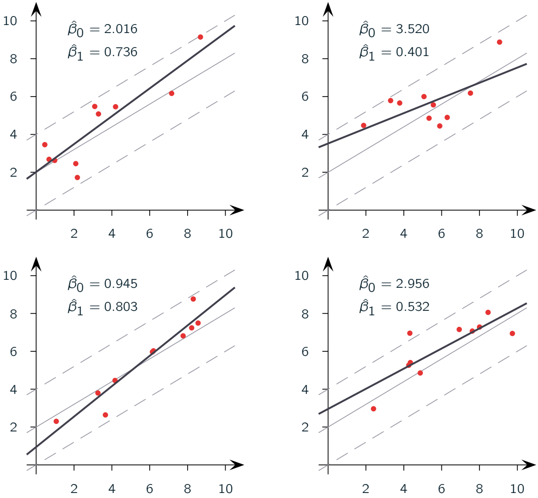
Data generating process

$$y_i = 2 + 0.6x_i + \varepsilon_i$$
$$x_i \sim U(0, 10)$$
$$\varepsilon_i \sim N(0, 1)$$



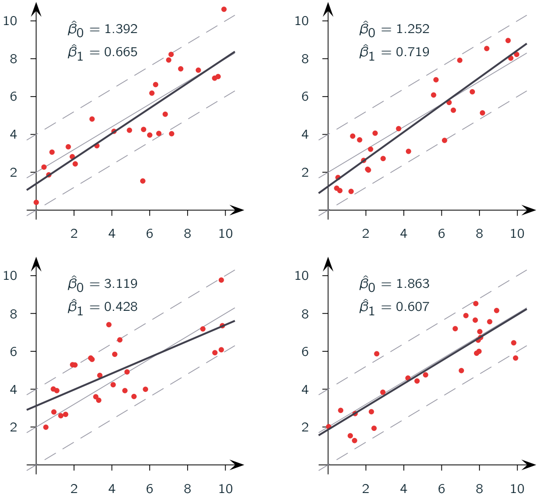
Example: sampling distribution

Four samples with $n = 10$



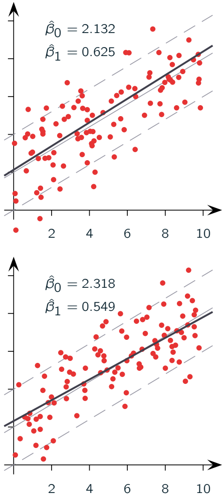
Example: sampling distribution

Four samples with $n = 25$



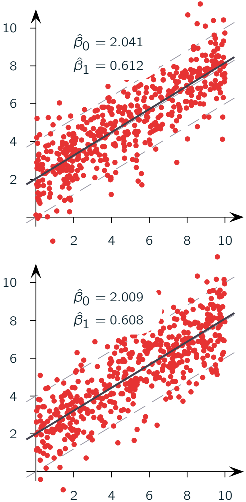
Example: sampling distribution

Four samples with $n = 100$



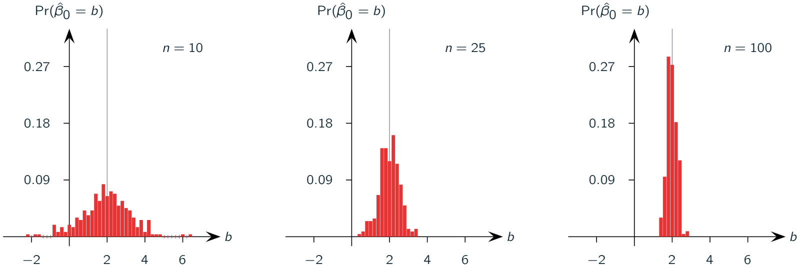
Example: sampling distribution

Four samples with $n = 500$



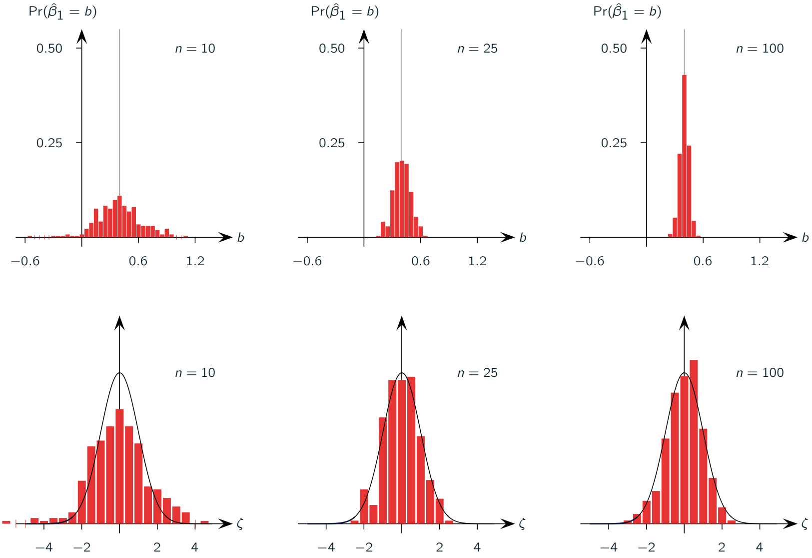
Example: A strange random variable

Sampling distribution of $\hat{\beta}_0$



Example: A strange random variable

Sampling distribution of $\hat{\beta}_1$



Unbiasedness

Theorem

The OLS estimates are unbiased:

$$\mathbb{E}\left[\hat{\beta}_0\right] = \beta_0 \qquad \mathbb{E}\left[\hat{\beta}_1\right] = \beta_1$$

- We can write:

$$\beta_0 = \mu_y - \beta_1 \mu_x \qquad \beta_1 = \frac{\sigma_{xy}}{\sigma_x^2}$$

- The OLS estimates are the corresponding sample analogues:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

- Sample averages are unbaiased (and consistent) estimators of means

Unbiasedness of $\hat{\beta}_1$

- Notice that:

$$
\begin{aligned}
\bar{y} &= \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon} \\
y_i - \bar{y} &= \beta_1 (x_i - \bar{x}) + \varepsilon_i - \bar{\varepsilon}
\end{aligned}
$$

- Substituting in the formula for $\hat{\beta}_1$:

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\
&= \frac{\sum (x_i - \bar{x}) \Big(\beta_1 (x_i - \bar{x}) + \varepsilon_i - \bar{\varepsilon} \Big)}{\sum (x_i - \bar{x})^2} \\
&= \beta_1 + \frac{\sum (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum (x_i - \bar{x})^2} = \beta_1 + \frac{\sum (x_i - \bar{x}) \varepsilon_i}{\sum (x_i - \bar{x})^2}
\end{aligned}
$$

- Taking expectation:

$$
\mathbb{E}\left[\hat{\beta}_1 \right] = \beta_1 + \mathbb{E}\left[\frac{\sum (x_i - \bar{x}) \varepsilon_i}{\sum (x_i - \bar{x})^2} \right] = \beta_1
$$

Variance of $\hat{\beta}_1$

- Under the classical assumptions, the variance of the OLS slope estimator is:

$$\mathbb{V}\left[\hat{\beta}_1\right] = \frac{1}{n} \cdot \frac{\mathbb{V}[\varepsilon_i]}{\mathbb{V}[x_i]}$$

- Notice two interesting things:
  - Increasing the variance of x increases efficiency
  - Increasing variance of ε (noise) decreases efficiency

Theorem

Under the classical assumptions, the OLS estimator is the most efficient unbaiased linear estimator (BLUE).

Estimating the variance of $\hat{\beta}_1$

- Our formula for $\mathbb{V}\left[\hat{\beta}_1\right]$ requires σ_x^2 and σ_ε^2
- When they are unknown they can be estimated from our data:

$$\hat{\sigma}_x^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$
$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-1}\sum e_i^2 = \frac{1}{n-1}\text{RSS}$$

- Likewise, we can estimate the variance of $\hat{\beta}_1$

$$\hat{\sigma}_{\beta_1}^2 = \frac{1}{n} \cdot \frac{\text{RSS}}{\sum(x_i - \bar{x})^2}$$

Some additional considerations

- The LLN implies that $\hat{\beta}_0$ and $\hat{\beta}_1$ are consistent
- The CLT implies that the distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ is approximately normal for large samples
- We often do inference assuming that:

$$\hat{\beta}_1 \sim N\left(\beta_1 \, , \, \frac{1}{n} \cdot \frac{\text{RSS}}{\sum(x_i - \bar{x})^2}\right)$$

- Without homoskedasticity, we need to adjust our estimation of $\mathbb{V}\left[\,\hat{\beta}_1\,\right]$
- Some of the classical assumptions are sufficient but not necessary

[0]

Inference

- Inference refers to deriving information from the data
- In statistics, inference takes the form of hypothesis testing

- Today we will focus on significance testing
- We wish to determine whether the data conclusively suggests that x has a positive (negative) effect on y

- We will also establish confidence sets for our estimates and our predictions

Significance

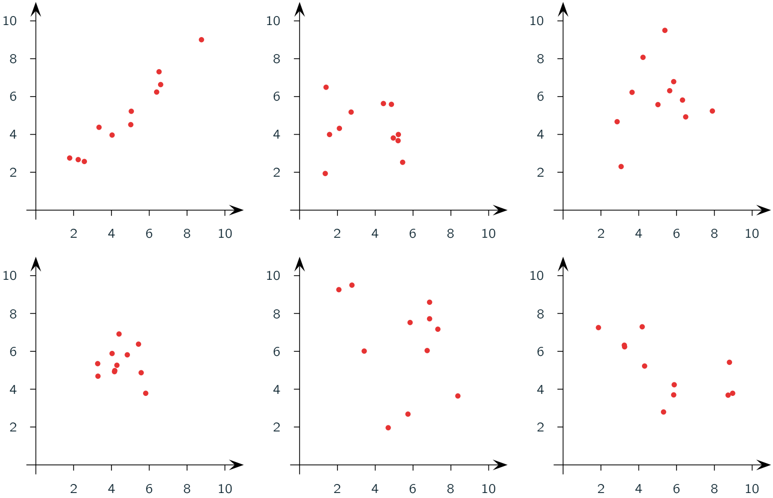
- Suppose that we obtain a positive OLS slope coefficient $\hat{\beta}_1 > 0$
- This does not guarantee that there is a positive relation, i.e. $\beta_1 > 0$
- Another possibility is that $\beta_1 = 0$ and the positive estimate comes from samling error

- We say that $\hat{\beta}_1$ is significant if the data decisively suggests that $\hat{\beta}_1 \neq 0$
- Formally, want to test hypothesis of the form

$$\mathscr{H}_0\colon\ \beta_1 \neq 0 \qquad \text{vs.} \qquad \mathscr{H}_1\colon\ \beta_1 = 0$$

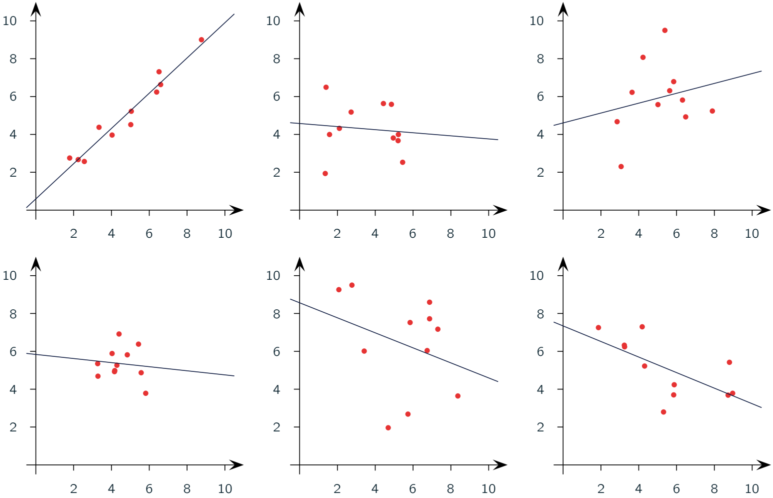
Example: significance

Realized samples



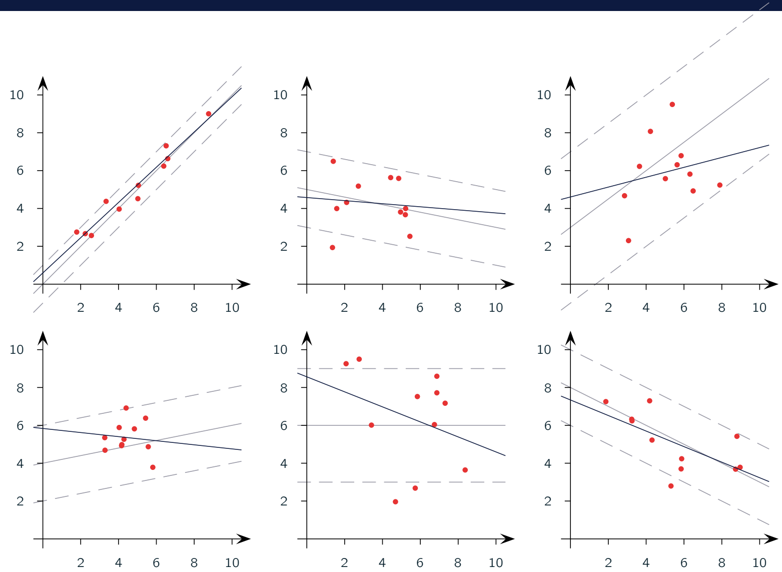
Example: significance

Estimated models



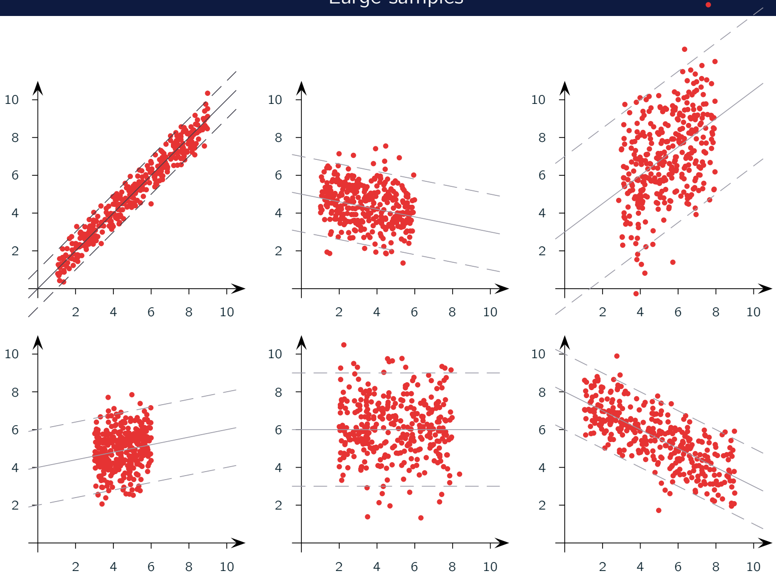
Example: significance

True models



Example: significance

Large samples



t–statistic

- Suppose that we want to test for:

$$\mathscr{H}_0\colon\ \beta_1 \neq 0 \qquad \text{vs.} \qquad \mathscr{H}_1\colon\ \beta_1 = 0$$

- Recall that approximately $\hat{\beta}_1 \sim N\left(\beta_1, \hat{\sigma}^2_{\hat{\beta}_1}\right)$

- Therefore, under the null hypothesis:

$$t = \frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)} \sim N(0,1)$$

- We can use this statistic to test our hypothesis
- If t is far away from 0, then $\mathscr{H}_o$ is likely to be false
- Rule of thumb: 2 standard deviations $\sim$ 96% significance

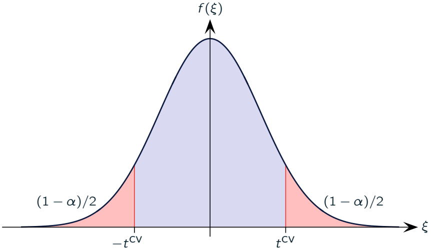
Significance

$$\mathscr{H}_0\colon\ \beta_1 = 0 \quad \text{vs.} \quad \mathscr{H}_1\colon\ \beta_i \neq 0$$

$$t = \frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)}$$

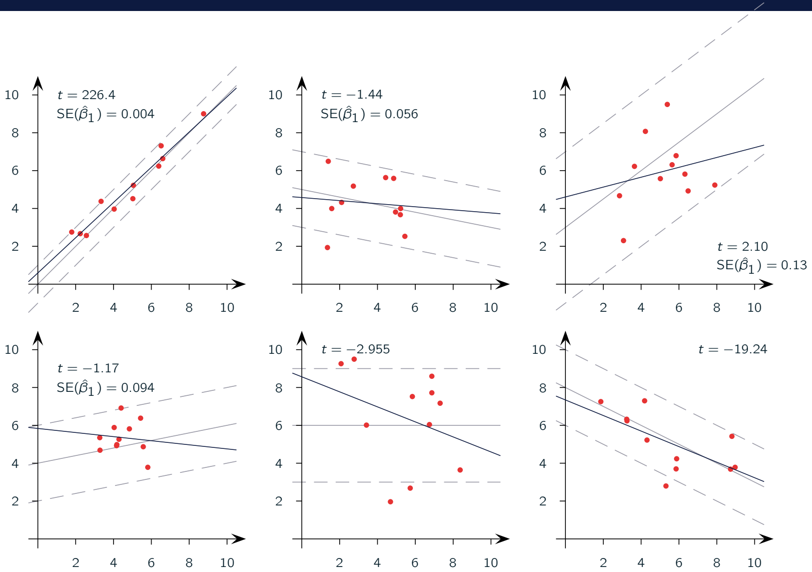
- Under $\mathscr{H}_0$ the asymptotic distribution of t is $N(0,1)$
- A test of significance α is to reject $\mathscr{H}_0$ if:

$$|t| > t^{\mathrm{cv}} = \Phi^{-1}\big((1-\alpha)/2\big)$$



Example: significance

True models



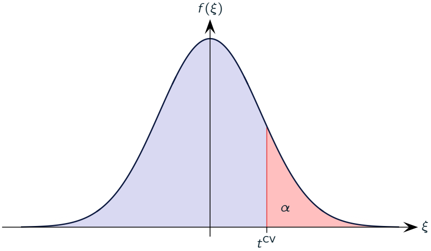
One sided hypothesis

$$\mathscr{H}_0\colon\ \beta_1 \leq b \quad \text{vs.} \quad \mathscr{H}_1\colon\ \beta_i > b$$

$$t = \frac{\hat{\beta}_1 - b}{\mathrm{SE}(\hat{\beta}_1)}$$

- Under $\mathscr{H}_0$ the asymptotic distribution of t is $N(0, 1)$
- A test of significance α is to reject $\mathscr{H}_0$ if:

$$t > t^{\mathrm{cv}} = \Phi^{-1}(\alpha)$$



Regression output

- Most linear regression software will report:
  - Estimate $\hat{\beta}_1$
  - Standard error for the estimate $\text{SE}(\hat{\beta}_1)$
  - t-statistic value t
  - p-value
  - Confidence interval $\hat{\beta}_1 \pm 1.96\text{SE}(\hat{\beta}_1)$
  - Normal and adjusted R^2

Observations

- t-tests do not test validity
- t-tests do not test importance
- Confidence is not probability

Regression output

- Most linear regression software will report:
  - Estimate $\hat{\beta}_1$
  - Standard error for the estimate $\text{SE}(\hat{\beta}_1)$
  - t-statistic value t
  - p-value
  - Confidence interval $\hat{\beta}_1 \pm 1.96\text{SE}(\hat{\beta}_1)$
  - Normal and adjusted R^2

Observations

- t-tests do not test validity
- t-tests do not test importance
- Confidence is not probability

Regression output

- Most linear regression software will report:
  - Estimate $\hat{\beta}_1$
  - Standard error for the estimate $\mathrm{SE}(\hat{\beta}_1)$
  - t-statistic value t
  - p-value
  - Confidence interval $\hat{\beta}_1 \pm 1.96\mathrm{SE}(\hat{\beta}_1)$
  - Normal and adjusted R^2

Observations

- t-tests do not test validity
- t-tests do not test importance
- Confidence is not probability

Regression output

- Most linear regression software will report:
  - Estimate $\hat{\beta}_1$
  - Standard error for the estimate $\mathrm{SE}(\hat{\beta}_1)$
  - t-statistic value t
  - p-value
  - Confidence interval $\hat{\beta}_1 \pm 1.96\mathrm{SE}(\hat{\beta}_1)$
  - Normal and adjusted R^2

Observations

- t-tests do not test validity
- t-tests do not test importance
- Confidence is not probability

Regression output

- Most linear regression software will report:
  - Estimate $\hat{\beta}_1$
  - Standard error for the estimate $\mathrm{SE}(\hat{\beta}_1)$
  - t-statistic value t
  - p-value
  - Confidence interval $\hat{\beta}_1 \pm 1.96\mathrm{SE}(\hat{\beta}_1)$
  - Normal and adjusted R^2

Observations

- t-tests do not test validity
- t-tests do not test importance
- Confidence is not probability

Regression output

- Most linear regression software will report:
  - Estimate $\hat{\beta}_1$
  - Standard error for the estimate $\text{SE}(\hat{\beta}_1)$
  - t-statistic value t
  - p-value
  - Confidence interval $\hat{\beta}_1 \pm 1.96\text{SE}(\hat{\beta}_1)$
  - Normal and adjusted R^2

  Observations

  - t-tests do not test validity
  - t-tests do not test importance
  - Confidence is not probability

Regression output

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	-2.681508	1.393991	-1.92	0.055	-5.424424	.0614073
x2	-3.702419	.1540256	-24.04	0.000	-4.005491	-3.399348
x3	.1086104	.090719	1.20	0.232	-.0698947	.2871154
_cons	906.7392	28.26505	32.08	0.000	851.1228	962.3555

$$\hat{y} = \underset{(28.27)}{906} \underset{(1.39)}{-2.68}\, x_1 \underset{(0.15)}{-3.70}\, x_2 \underset{(0.09)}{+0.109}\, x_3$$

Prediction intervals

- For predictive purposes we can still generate confidence intervals arround $\hat{y}_i$
- A naive way to do so is to use just the residual variance:

$$y_i \in \left(\hat{y}_i - K \cdot \mathrm{RSS}\ ,\ \hat{y}_i + K \cdot \mathrm{RSS}\right)$$

- This yields the confidence bands in previous figures
- This would be accurate only if $\hat{\beta}_0 = \beta_0$ and $\hat{\beta}_1 = \beta_1$
- One needs to adjust form the variance of the estimators