

REGRESSION

12.1 Simple Linear Regression Model

12.2 Fitting the Regression Line

12.3 Inferences on the Slope Parameter

Suppose we have a data set with the following paired observations:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

Example:

$$\begin{aligned}x_i &= \text{height of person } i \\y_i &= \text{weight of person } i\end{aligned}$$

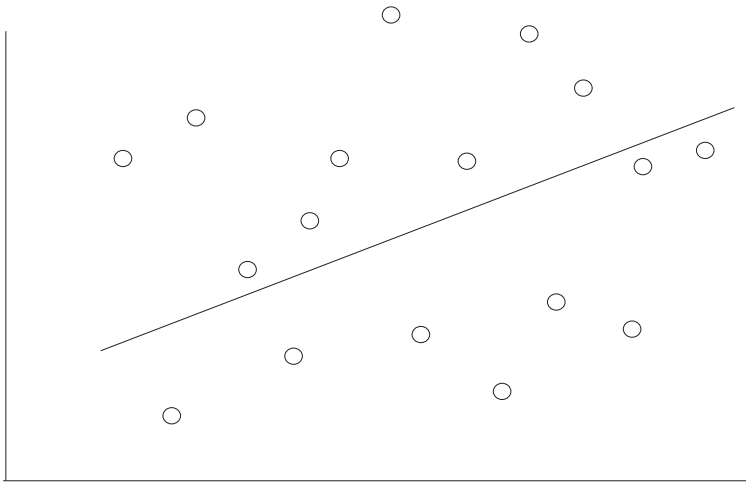
Can we make a model expressing y_i as a function of x_i ?

Estimate y_i for fixed x_i . Let's model this with the simple linear regression equation,

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

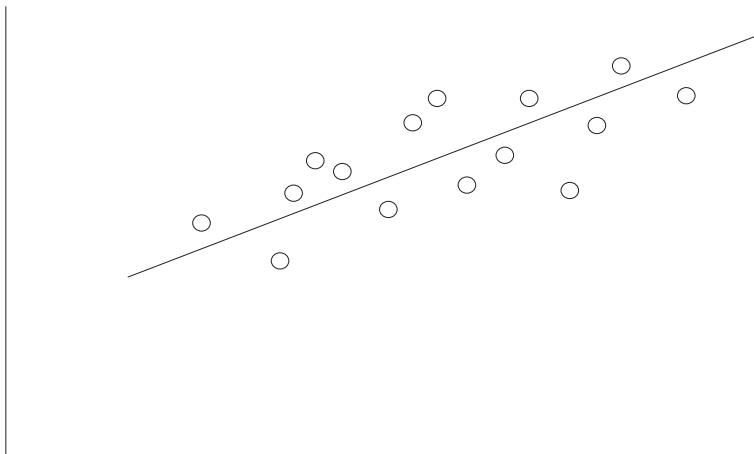
where β_0 and β_1 are unknown constants and the error terms are usually assumed to be

$$\begin{aligned} \varepsilon_1, \dots, \varepsilon_n &\stackrel{iid}{\sim} N(0, \sigma^2) \\ \Rightarrow y_i &\sim N(\beta_0 + \beta_1 x_i, \sigma^2). \end{aligned}$$



$$y = \beta_0 + \beta_1 x$$

with "high" σ^2

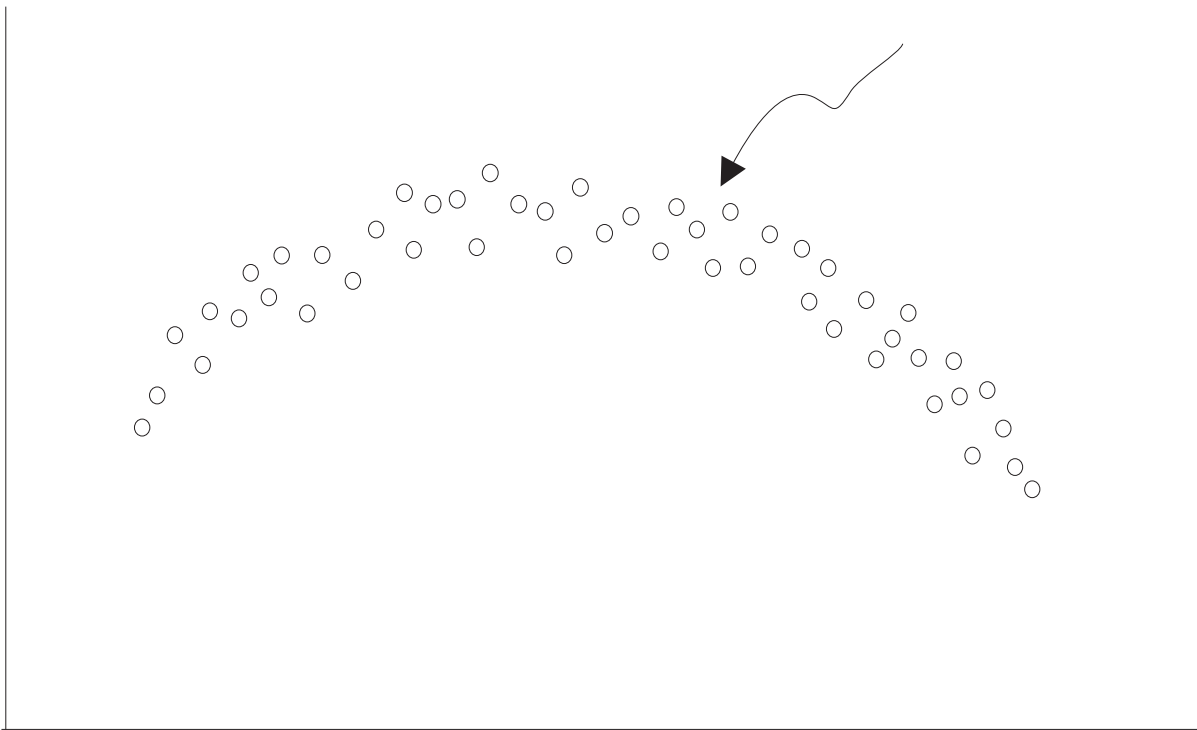


$$y = \beta_0 + \beta_1 x$$

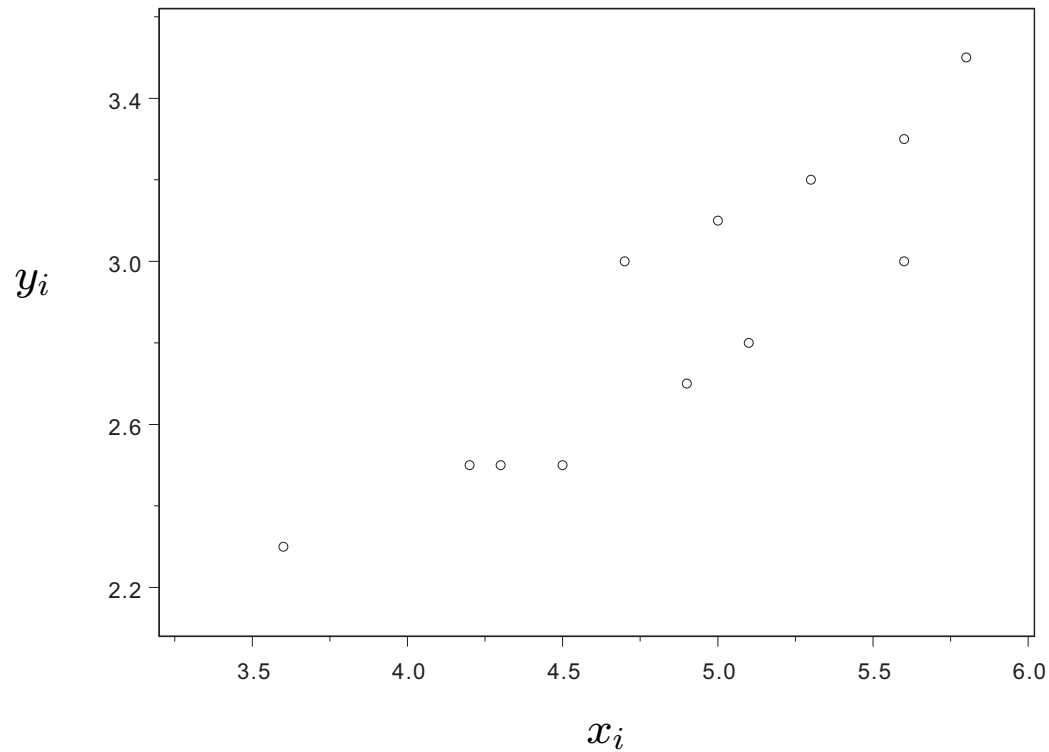
with "low" σ^2

Warning! Look at data before you fit a line to it:

doesn't look very linear!



	x_i Production (\$ million)	y_i Electric Usage (million kWh)
Jan	4.5	2.5
Feb	3.6	2.3
Mar	4.3	2.5
Apr	5.1	2.8
May	5.6	3.0
Jun	5.0	3.1
Jul	5.3	3.2
Aug	5.8	3.5
Sep	4.7	3.0
Oct	5.6	3.3
Nov	4.9	2.7
Dec	4.2	2.5



Great... but how do you fit the line?

Fit the regression line $y = \beta_0 + \beta_1 x$ to the data

$$(x_1, y_1), \dots, (x_n, y_n)$$

by finding the “best” match between the line and the data. The “best” choice of β_0, β_1 will be chosen to minimize

$$Q = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 = \sum_{i=1}^n \varepsilon_i^2.$$

This is called the least square fit. Let's solve...

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum (y_i - (\beta_0 + \beta_1 x_i)) = 0$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum x_i (y_i - (\beta_0 + \beta_1 x_i)) = 0$$

$$\Leftrightarrow \sum y_i = n\beta_0 + \beta_1 \sum x_i$$

$$\sum x_i y_i = -2 \sum x_i (y_i - (\beta_0 + \beta_1 x_i)) = 0$$

After a little algebra, get

$$\hat{\beta}_1 = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \text{where } \bar{y} \equiv \frac{1}{n} \sum y_i \text{ and } \bar{x} \equiv \frac{1}{n} \sum x_i.$$

Let's introduce some more notation:

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$$

$$= \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$$

$$= \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

These are called “sums of squares.”

Then, after a little more algebra, we can write

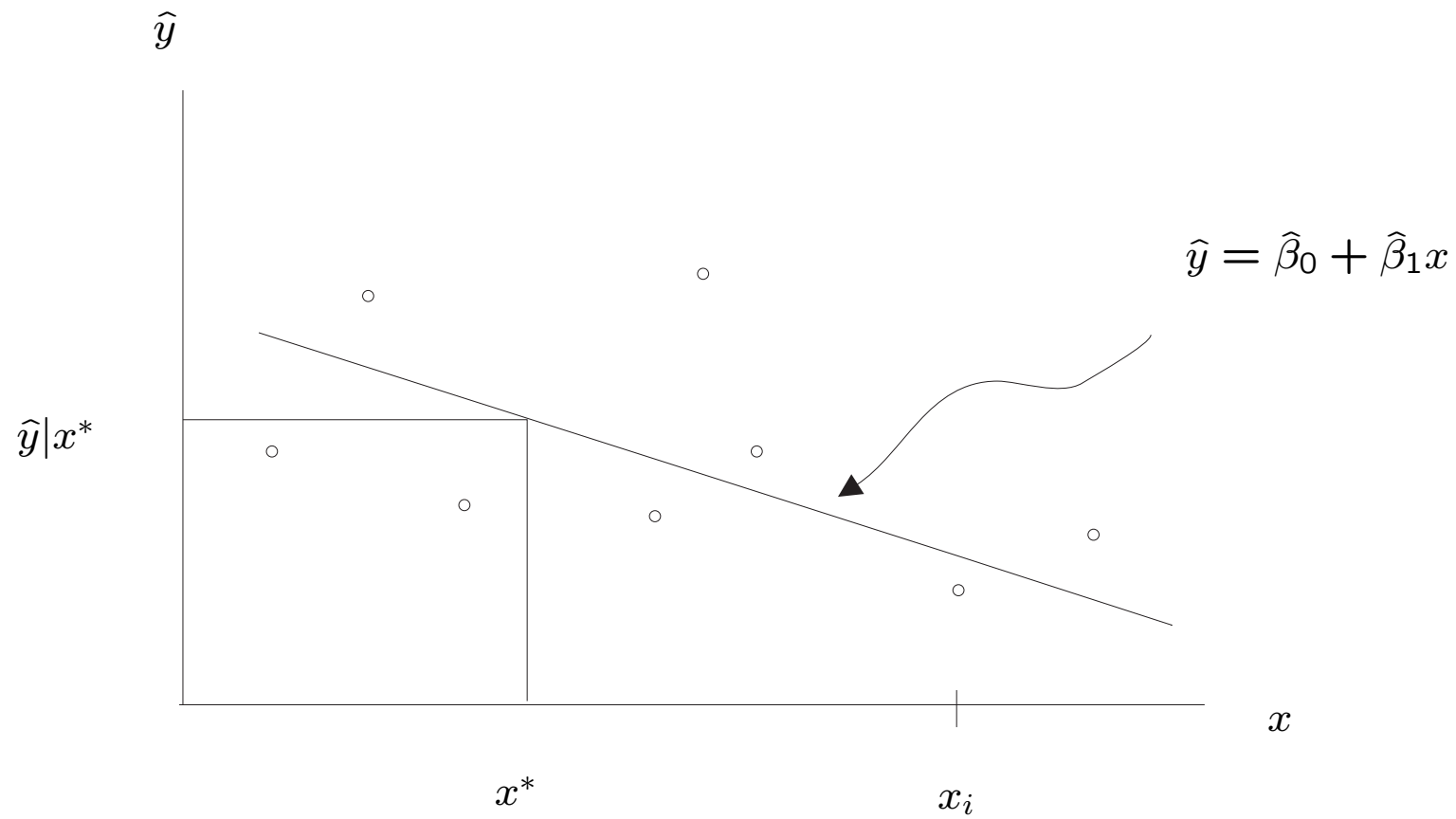
$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Fact: If the ε_i 's are iid $N(0, \sigma^2)$, it can be shown that $\hat{\beta}_0$ and $\hat{\beta}_1$ are the MLE's for β_0 and β_1 , respectively. (See text for easy proof).

Anyhow, the fitted regression line is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Fix a specific value of the explanatory variable x^* , the equation gives a fitted value $\hat{y}|x^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$ for the dependent variable y .



For actual data points x_i , the fitted values are $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

observed values : $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

fitted values : $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Let's estimate the error variation σ^2 by considering the deviations between y_i and \hat{y}_i .

$$\begin{aligned} SSE &= \sum (y_i - \hat{y}_i)^2 = \sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \\ &= \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i. \end{aligned}$$

Turns out that $\hat{\sigma}^2 \equiv \frac{SSE}{n-2}$ is a good estimator for σ^2 .

Example: Car plant energy usage $n = 12$, $\sum_{i=1}^{12} x_i = 58.62$, $\sum y_i = 34.15$, $\sum x_i^2 = 291.231$, $\sum y_i^2 = 98.697$,

$$\sum x_i y_i = 169.253$$

$$\hat{\beta}_1 = 0.49883, \hat{\beta}_0 = 0.4090$$

\Rightarrow fitted regression line is

$$\hat{y} = 0.409 + 0.499x \quad \hat{y}|5.5 = 3.1535$$

What about something like $\hat{y}|10.0$?

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \text{ where } S_{xx} = \sum (x_i - \bar{x})^2 \text{ and}$$

$$\begin{aligned} S_{xy} &= \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i - \bar{x})y_i - \bar{y} \sum (x_i - \bar{x}) \\ &= \sum (x_i - \bar{x})y_i \end{aligned}$$

Since the y_i 's are independent with $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ (and the x_i 's are constants), we have

$$\begin{aligned} \mathbb{E}\hat{\beta}_1 &= \frac{1}{S_{xx}} \mathbb{E}S_{xy} = \frac{1}{S_{xx}} \sum (x_i - \bar{x}) \mathbb{E}y_i = \frac{1}{S_{xx}} \sum (x_i - \bar{x}) (\beta_0 + \beta_1 x_i) \\ &= \frac{1}{S_{xx}} \left[\beta_0 \underbrace{\sum (x_i - \bar{x})}_0 + \beta_1 \sum (x_i - \bar{x}) x_i \right] \\ &= \frac{\beta_1}{S_{xx}} \sum (x_i^2 - x_i \bar{x}) = \frac{\beta_1}{S_{xx}} \underbrace{\left(\sum x_i^2 - n\bar{x}^2 \right)}_{S_{xx}} = \beta_1 \end{aligned}$$

$\Rightarrow \hat{\beta}_1$ is an unbiased estimator of β_1 .

Further, since $\hat{\beta}_1$ is a linear combination of independent normals, $\hat{\beta}_1$ is itself normal. We can also derive

$$\text{Var}(\hat{\beta}_1) = \frac{1}{S_{xx}^2} \text{Var}(S_{xy}) = \frac{1}{S_{xx}^2} \sum (x_i - \bar{x})^2 \text{Var}(y_i) = \frac{\sigma^2}{S_{xx}}.$$

Thus, $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$

While we're at it, we can do the same kind of thing with the intercept parameter, β_0 :

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Thus, $E\hat{\beta}_0 = E\bar{y} - \bar{x}E\hat{\beta}_1 = \beta_0 + \beta_1\bar{x} - \bar{x}\beta_1 = \beta_0$ Similar to before, since $\hat{\beta}_0$ is a linear combination of independent normals, it is also normal. Finally,

$$\text{Var}(\hat{\beta}_0) = \frac{\sum x_i^2}{nS_{xx}}\sigma^2.$$

Proof:

$$\begin{aligned}\text{Cov}(\bar{y}, \hat{\beta}_1) &= \frac{1}{S_{xx}} \text{Cov}(\bar{y}, \sum (x_i - \bar{x}) y_i) \\ &= \frac{\sum (x_i - \bar{x})}{S_{xx}} \text{Cov}(\bar{y}, y_i) \\ &= \frac{\sum (x_i - \bar{x}) \sigma^2}{S_{xx} n} = 0\end{aligned}$$

$$\begin{aligned}\Rightarrow \text{Var}(\hat{\beta}_0) &= \text{Var}(\bar{y} - \hat{\beta}_1 \bar{x}) \\ &= \text{Var}(\bar{y}) + \bar{x}^2 \text{Var} \hat{\beta}_1 - 2\bar{x} \underbrace{\text{Cov}(\bar{y}, \hat{\beta}_1)}_0 \\ &= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} \\ &= \sigma^2 \left(\frac{S_{xx} - n\bar{x}^2}{nS_{xx}} \right).\end{aligned}$$

Thus, $\hat{\beta}_0 \sim N(\beta_0, \frac{\sum x_i^2}{nS_{xx}} \sigma^2)$.

Back to $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx}) \dots$

$$\Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim N(0, 1)$$

Turns out:

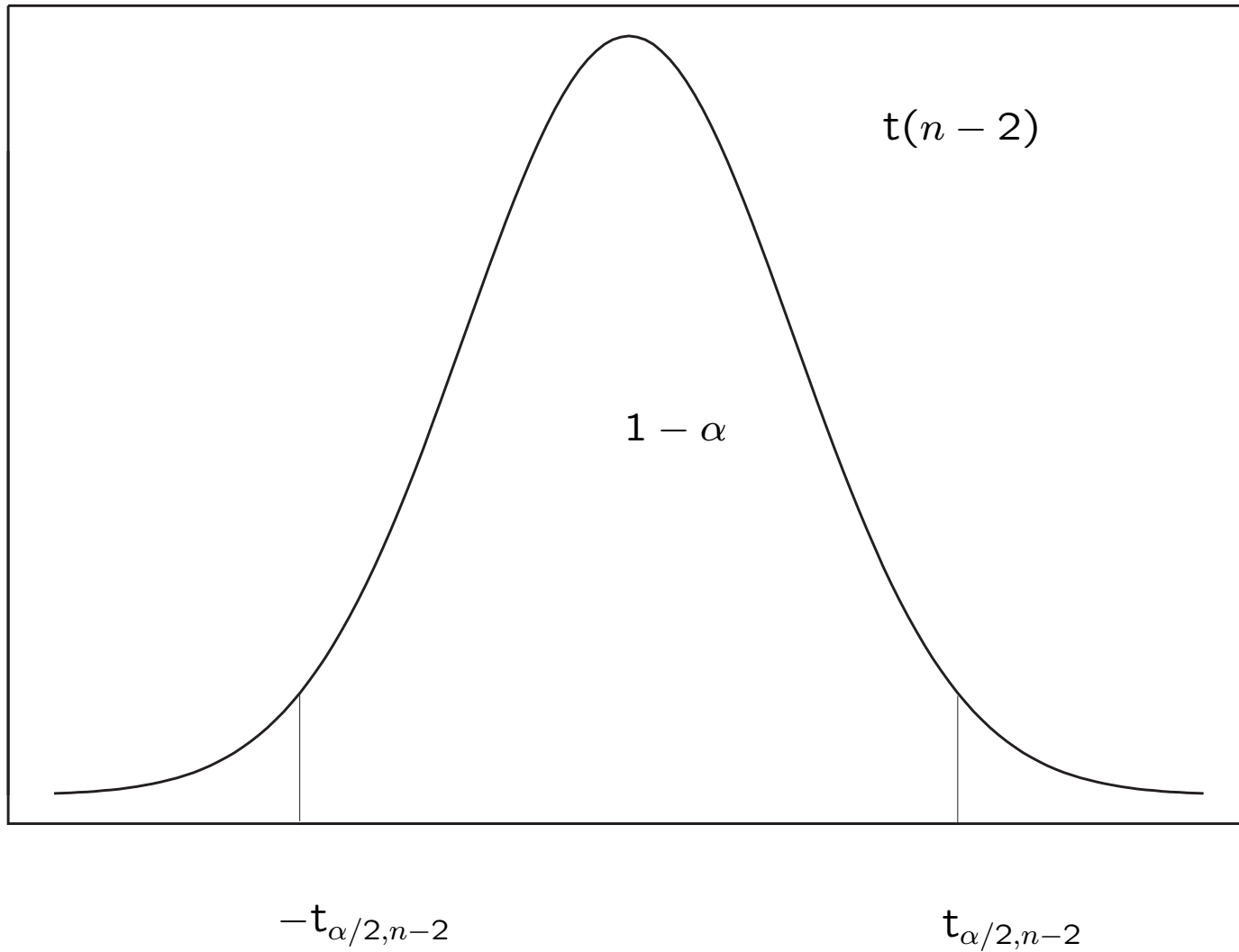
- (1) $\hat{\sigma}^2 = \frac{SSE}{n-2} \sim \frac{\sigma^2 \chi^2(n-2)}{n-2}$;
- (2) $\hat{\sigma}^2$ is independent of $\hat{\beta}_1$.

\Rightarrow

$$\frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{S_{xx}}}}{\hat{\sigma} / \sigma} \sim \frac{N(0, 1)}{\sqrt{\frac{\chi^2(n-2)}{n-2}}} \sim t(n-2)$$

 \Rightarrow

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma} / \sqrt{S_{xx}}} \sim t(n-2).$$



2-sided Confidence Intervals for β_1 :

$$\begin{aligned} 1 - \alpha &= \Pr\left(-t_{\alpha/2, n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \leq t_{\alpha/2, n-2}\right) \\ &= \Pr\left(\hat{\beta}_1 - t_{\alpha/2, n-2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}}\right) \end{aligned}$$

1-sided CI's for β_1 :

$$\begin{aligned} \beta_1 &\in \left(-\infty, \hat{\beta}_1 + t_{\alpha, n-2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}}\right) \\ \beta_1 &\in \left(\hat{\beta}_1 - t_{\alpha, n-2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}}, \infty\right) \end{aligned}$$