Statistical Learning for Data Science

Lecture 07

唐晓颖

电子与电气工程系

南方科技大学

March 13, 2023

Linear Regression

- A simple approach of supervised learning.

- A useful tool for predicting a quantitative response.

- Although it may seem overly simplistic, it is extremely useful both conceptually and practically.

- Many fancy statistical learning approaches can be seen as generalizations or extensions of linear regression.

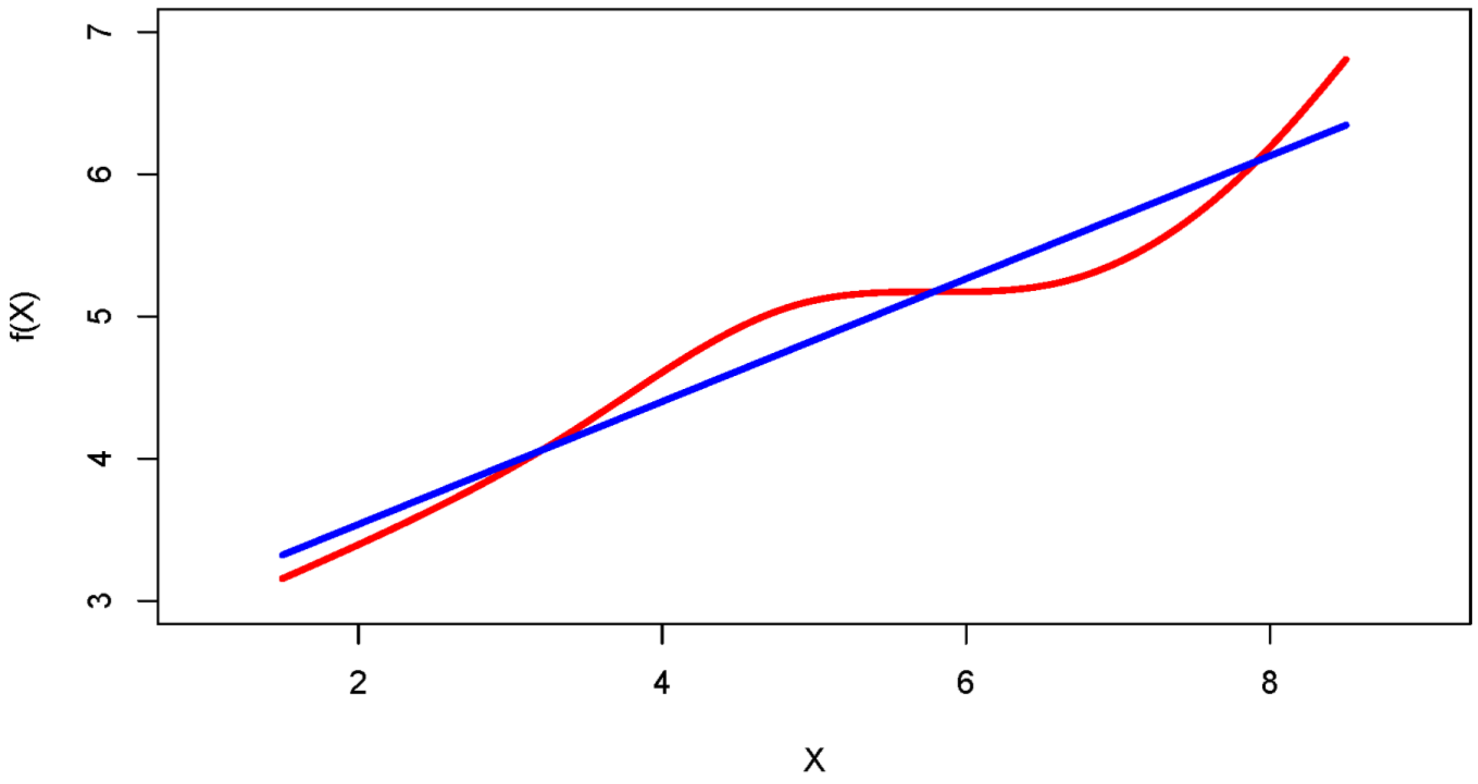
Linear Regression

❑ Assumptions of Linear Regression:

- The Independent variables should be linearly related to the dependent variables.
- Every feature in the data is Normally Distributed.
- There should be little or no multi-collinearity in the data.
- There should be little or no Auto-Correlation in the data.
- ……

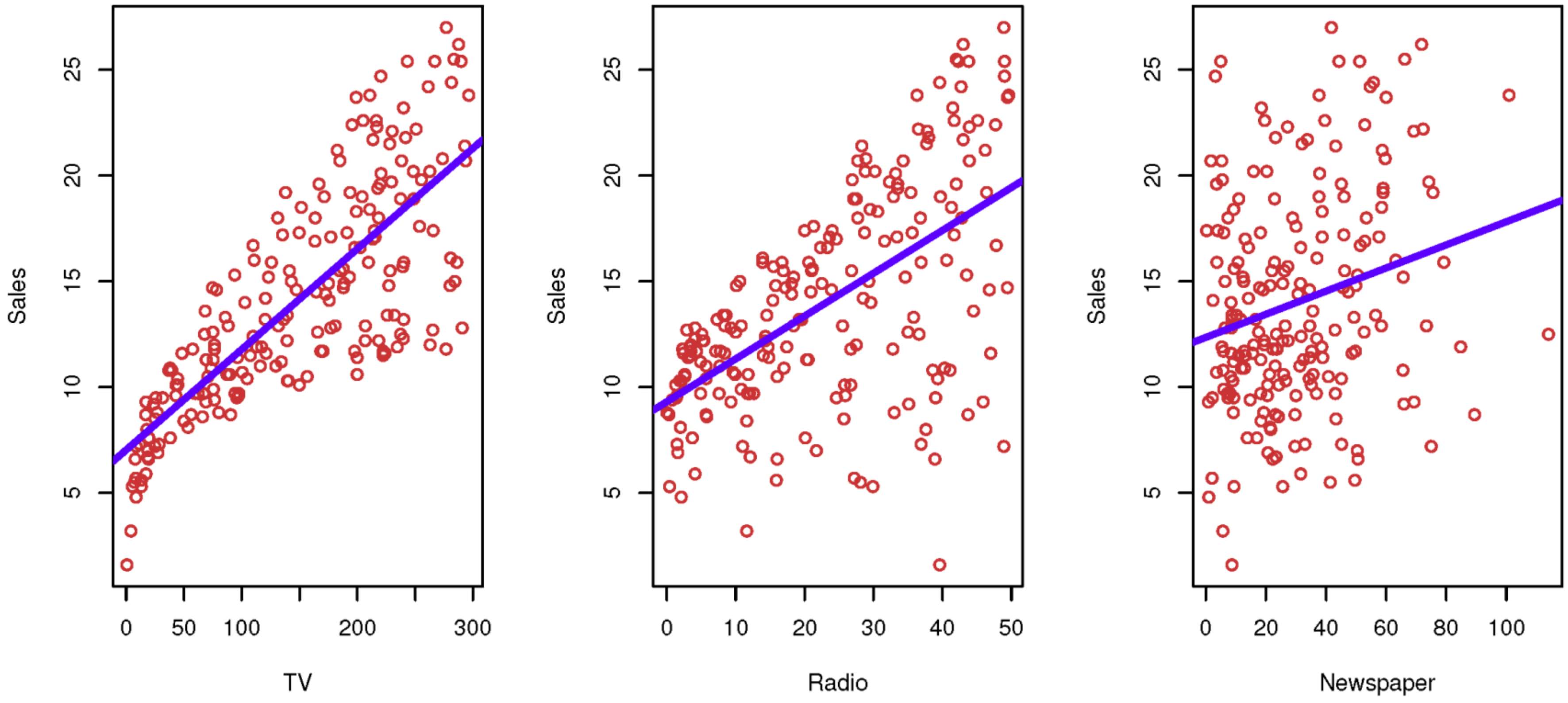
Linear Regression

- True regression functions are never linear.



Linear Regression

❑ Advertising data



Linear Regression

❑ Suggest a marketing plan for next year that will result in high product sales based on the advertising data

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media? (an interaction effect)

Linear regression can be used to answer each of these questions!

Simple Linear Regression

❑ Predict a quantitative response on the basis of a single predictor variable

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

For example,

$$\text{sales} = \beta_0 + \beta_1 \text{TV} + \varepsilon$$

β_0 : intercept (the expected value of Y when $X = 0$)

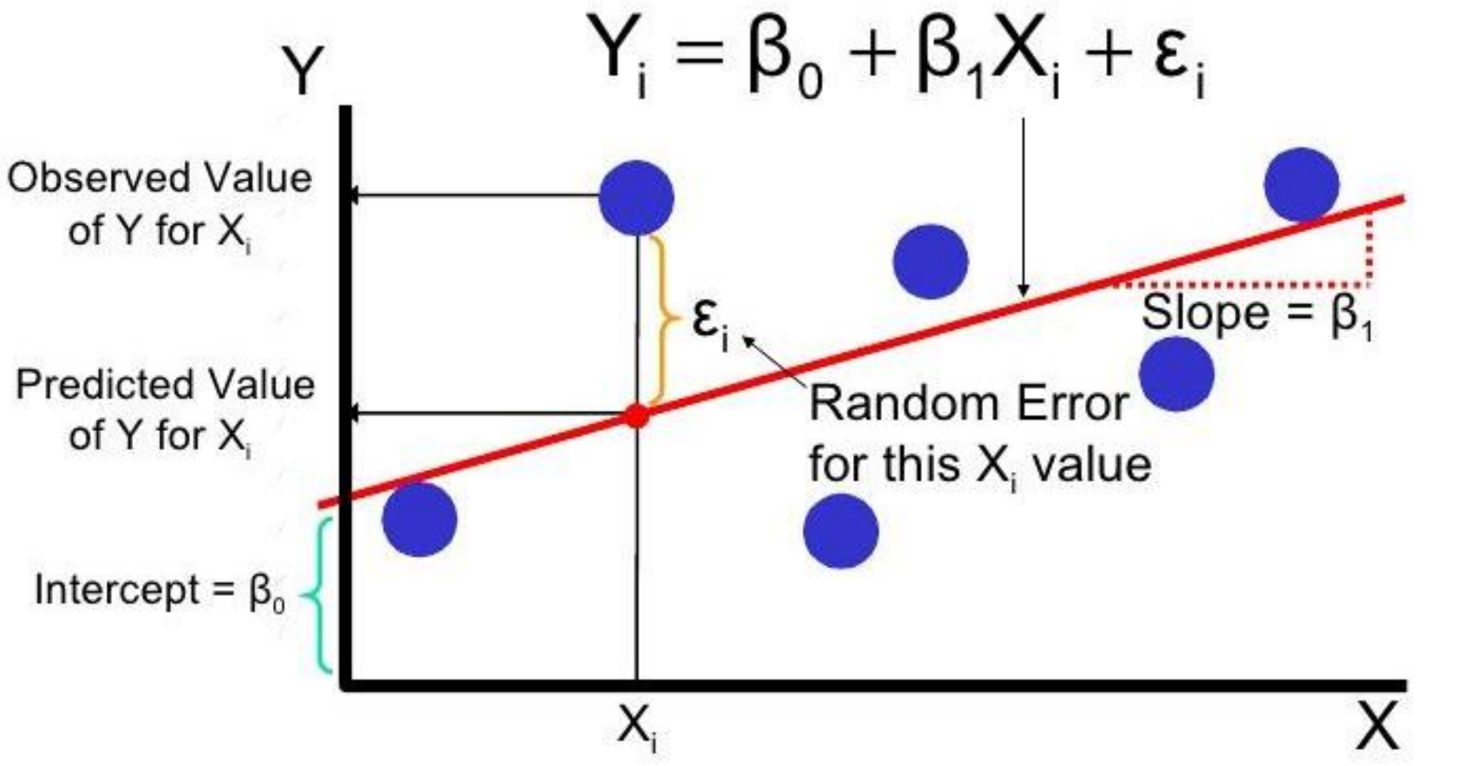
β_1 : slope (the average increase in Y associated with a one-unit increase in X)

(β_0, β_1): coefficients or parameters

ε : error term

Simple Linear Regression

- ❑ Predict a quantitative response on the basis of a single predictor variable



$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
Y
Observed Value of Y for X_i
Predicted Value of Y for X_i
Intercept = β_0
ε_i
Random Error for this X_i value
Slope = β_1
X_i
X

Simple Linear Regression

❑ Predict a quantitative response on the basis of a single predictor variable

Given some estimates $(\hat{\beta}_0, \hat{\beta}_1)$ for the model coefficients, we predict future response (sales) using $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Indicates a prediction of Y on the basis of $X = x$

Note: *We use a hat symbol to denote the estimated value for an unknown parameter or coefficient, or to denote the predicted value of the response.*

Simple Linear Regression

❑ Estimating the coefficients by minimizing the *least squares*

Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, then $e_i = y_i - \hat{y}_i$ represents the *i*-th *residual*

Define the residual sum of squares (RSS) as

$$RSS = e_1^2 + e_2^2 + \ldots + e_n^2$$

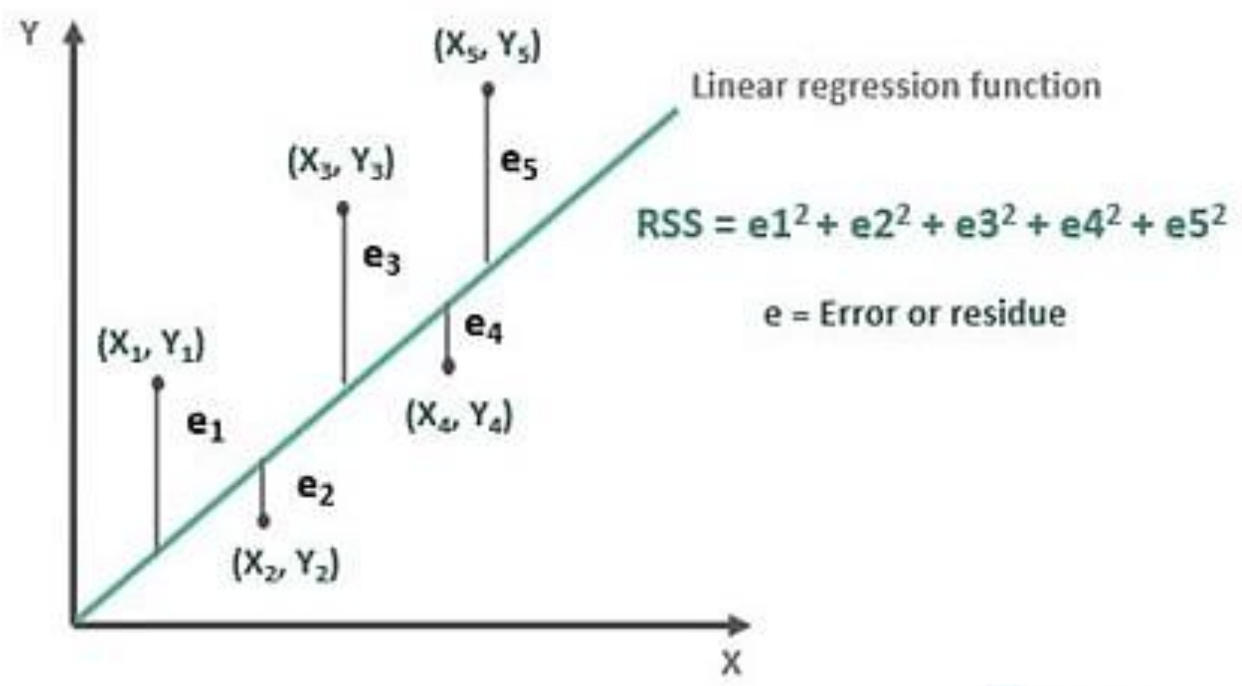
Or

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

Simple Linear Regression

❑ Estimating the coefficients by minimizing the *least squares*

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$



Simple Linear Regression

❑ Estimating the coefficients by minimizing the *least squares*

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

The *least squares* approach chooses $(\hat{\beta}_0, \hat{\beta}_1)$ to minimize the RSS

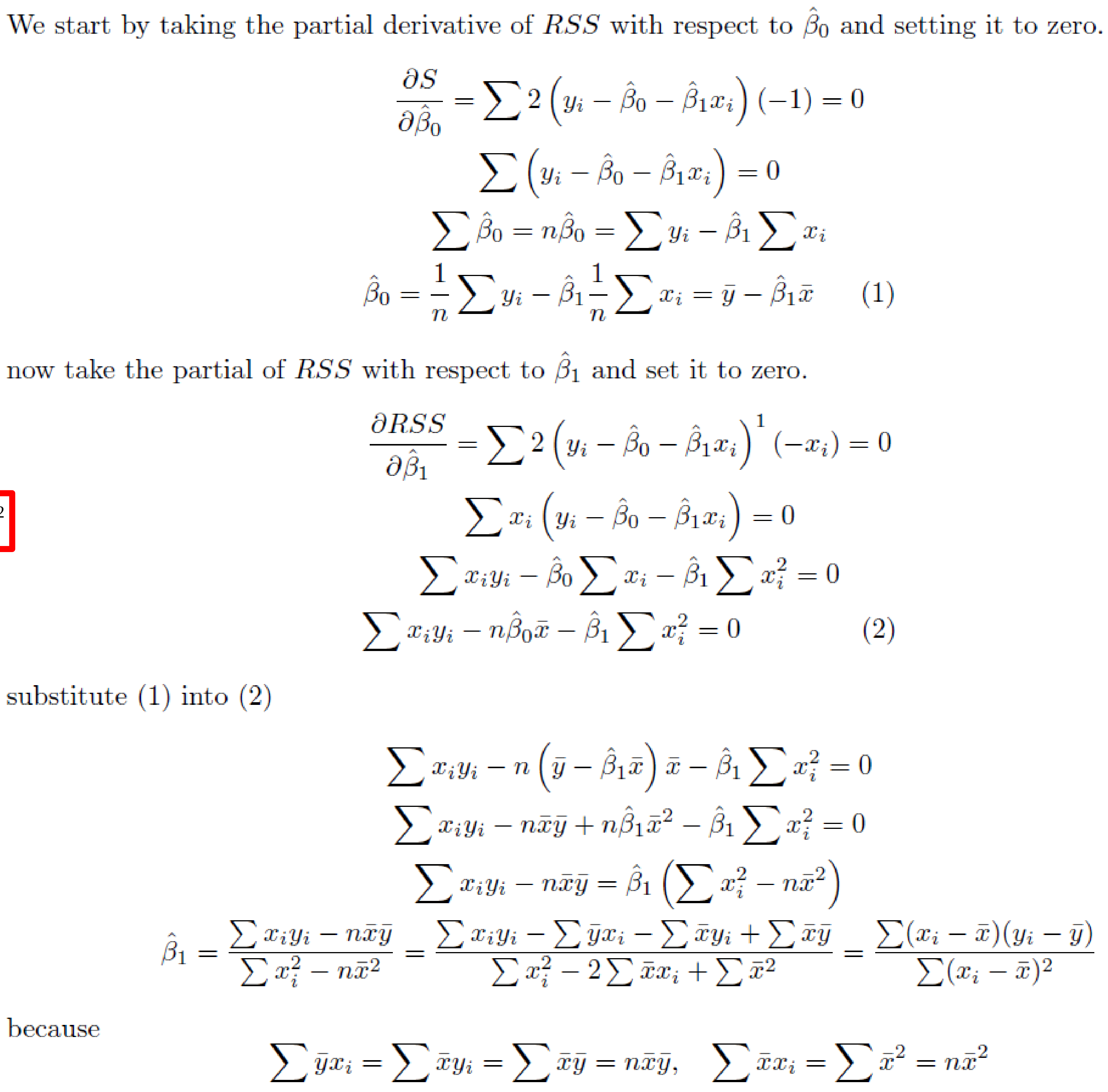
⬇

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$: the sample means

公式推导：

$$\boxed{RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2}$$



We start by taking the partial derivative of RSS with respect to $\hat{\beta}_0$ and setting it to zero.

$$\frac{\partial S}{\partial \hat{\beta}_0} = \sum 2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)(-1) = 0$$

$$\sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

$$\sum \hat{\beta}_0 = n\hat{\beta}_0 = \sum y_i - \hat{\beta}_1 \sum x_i$$

$$\hat{\beta}_0 = \frac{1}{n}\sum y_i - \hat{\beta}_1 \frac{1}{n}\sum x_i = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (1)$$

now take the partial of RSS with respect to $\hat{\beta}_1$ and set it to zero.

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = \sum 2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^1 (-x_i) = 0$$

$$\sum x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

$$\sum x_i y_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2 = 0$$

$$\sum x_i y_i - n\hat{\beta}_0 \bar{x} - \hat{\beta}_1 \sum x_i^2 = 0 \qquad (2)$$

substitute (1) into (2)

$$\sum x_i y_i - n\left(\bar{y} - \hat{\beta}_1 \bar{x}\right)\bar{x} - \hat{\beta}_1 \sum x_i^2 = 0$$

$$\sum x_i y_i - n\bar{x}\bar{y} + n\hat{\beta}_1 \bar{x}^2 - \hat{\beta}_1 \sum x_i^2 = 0$$

$$\sum x_i y_i - n\bar{x}\bar{y} = \hat{\beta}_1 \left(\sum x_i^2 - n\bar{x}^2\right)$$

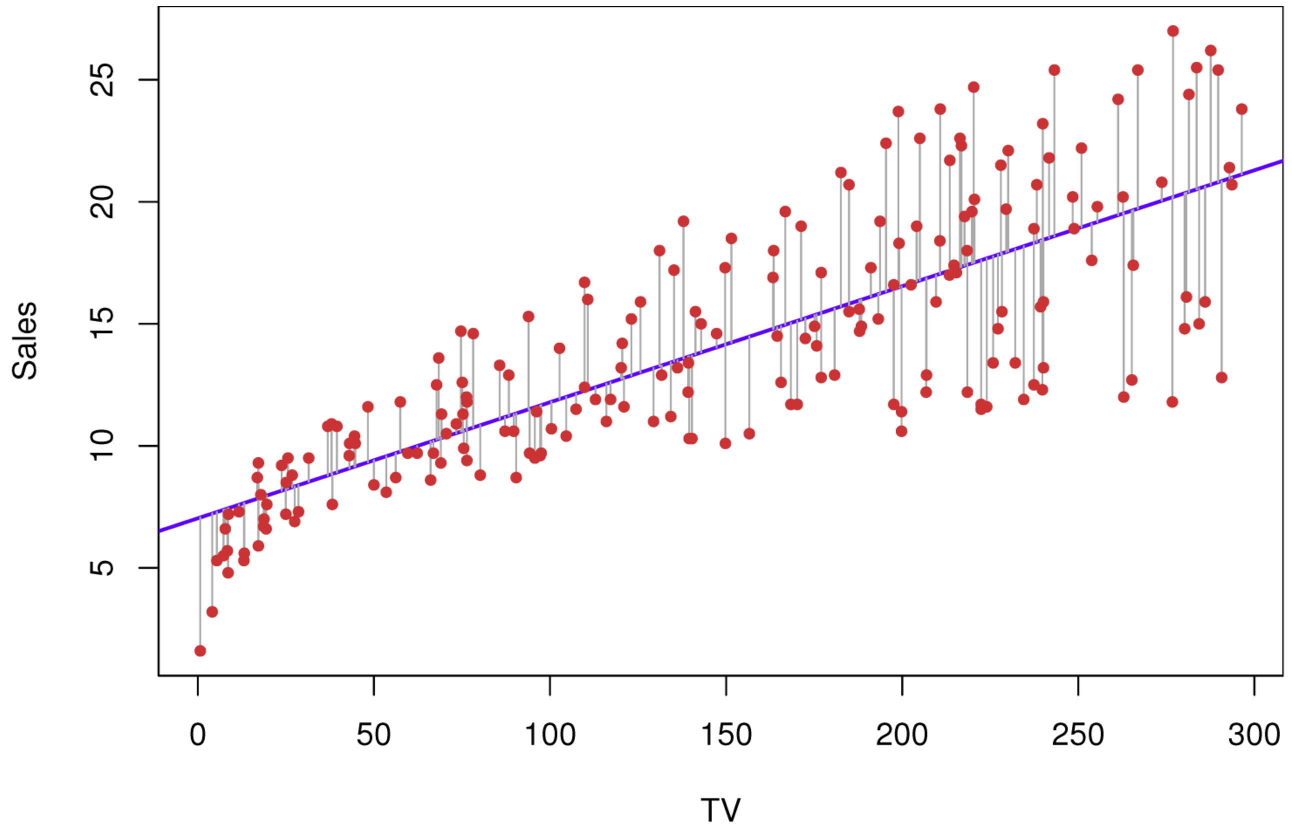
$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum x_i y_i - \sum \bar{y} x_i - \sum \bar{x} y_i + \sum \bar{x}\bar{y}}{\sum x_i^2 - 2\sum \bar{x} x_i + \sum \bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

because

$$\sum \bar{y} x_i = \sum \bar{x} y_i = \sum \bar{x}\bar{y} = n\bar{x}\bar{y}, \quad \sum \bar{x} x_i = \sum \bar{x}^2 = n\bar{x}^2$$

Simple Linear Regression

- ❑ Estimating the coefficients by minimizing the *least squares*



$$\hat{\beta}_1 = 0.0475$$

$$\hat{\beta}_0 = 7.03$$

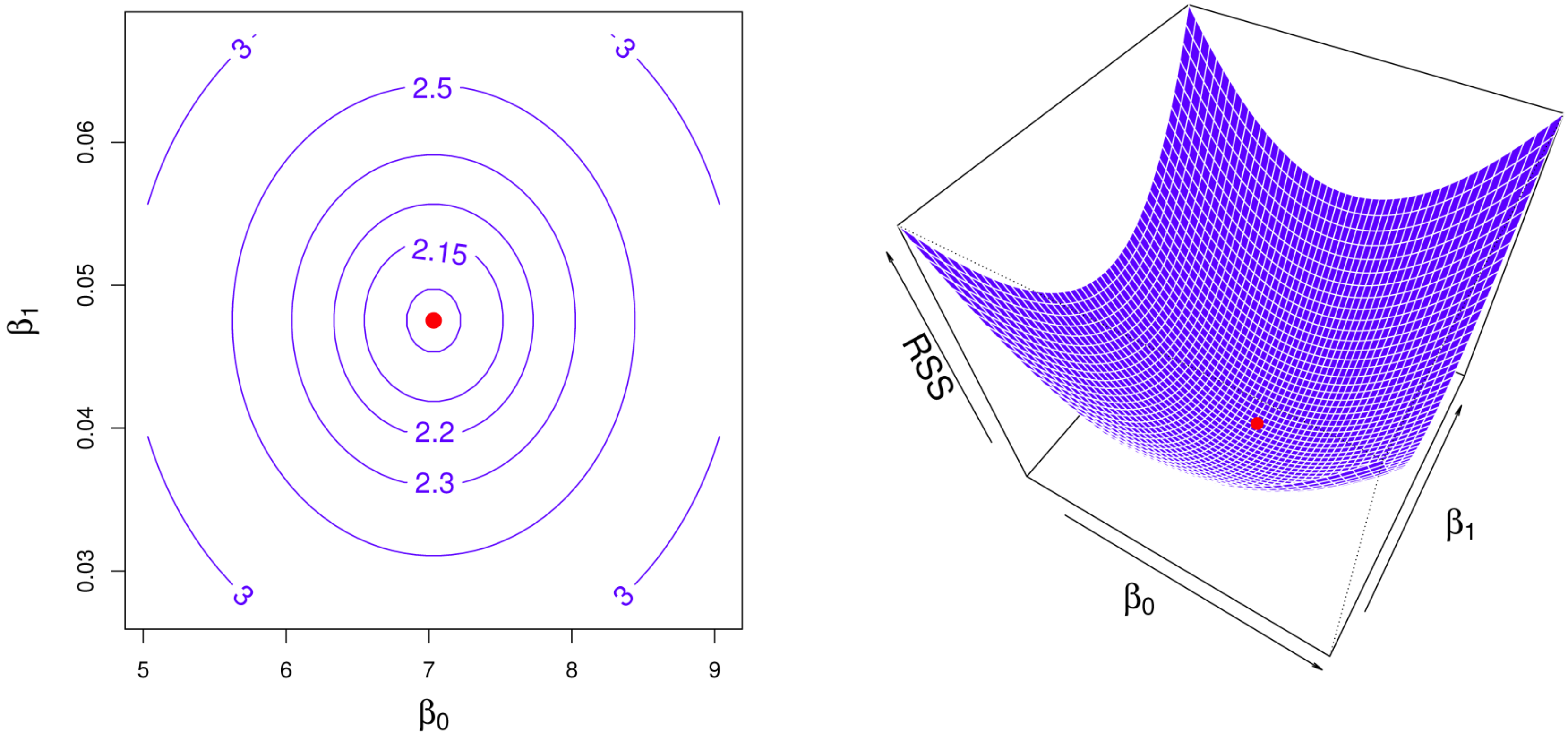
⬇

Conclusion: an addition $1,000 spent on TV advertising is associated with selling approximately 47.5 additional units of the product.

Simple Linear Regression

- Estimating the coefficients by minimizing the *least squares*

RSS as a function of $(\hat{\beta}_0, \hat{\beta}_1)$ in sales-versus-TV; the red dot represents the least squares estimates of the two parameters.



Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

$Y = \beta_0 + \beta_1 X + \varepsilon$ *--population regression line*

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ *--least squares line*

Least squares line: the regression line of sample data calculated by the method of least squares, which is considered as the best estimate of the population regression line.

Population regression line: the true relationship between the independent variable and the dependent variable in the population.

Note: When conducting regression analysis, we need to be cautious about extrapolating sample results to the population, and we need to provide appropriate explanations of limitations and conditions associated with the results.

Simple Linear Regression

- ❑ Assessing the accuracy of the coefficient estimates

$Y = \beta_0 + \beta_1 X + \varepsilon$ *--population regression line*

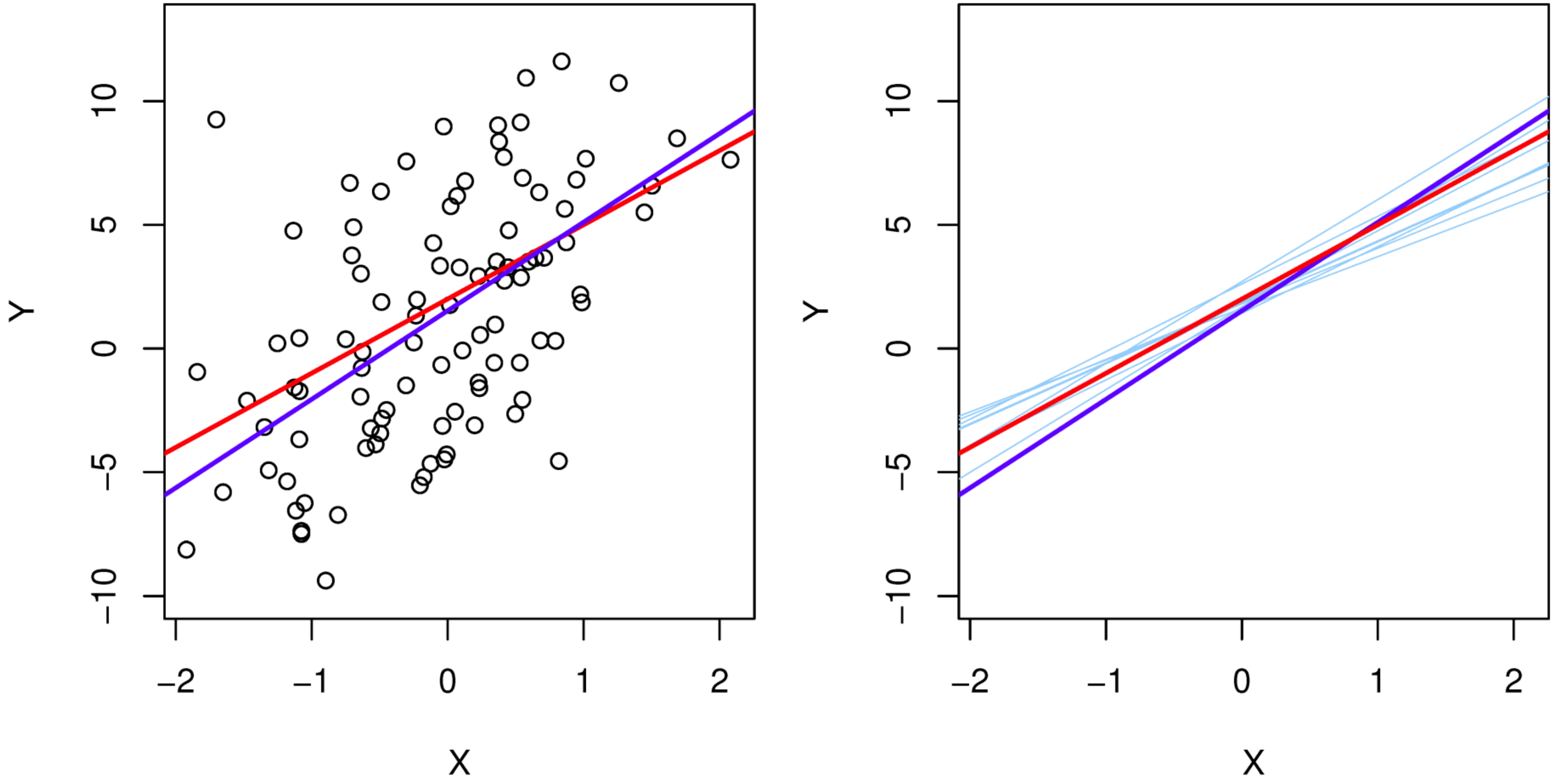
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ *--least squares line*

- ⌖ In real applications, the *least squares line* can be computed from observations, but the *population regression line* is unobserved.

- ⌖ Different data sets that are generated from the same true model will result in different *least square lines*, but the *population regression line* does not change.

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates



--red line: population regression line

--other lines: least squares regression lines (generated from different samples/observations)

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

Why there is a difference between the population line and the least squares line?

An analogy to the estimation of the mean of a random variable:

We are using information from a sample to estimate characteristics of a large population.

Example: we are interested in knowing the population mean μ of some random variable Y . A reasonable estimate is $\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, the sample mean.

The sample mean and the population mean are different, but in general the sample mean will provide a good estimate.

We can obtain different samples/observations, resulting in different sample means.

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

Unbiased

$$E_{\mu}(\hat{\mu}) = \mu$$

On the basis of one particular set of observations, $\hat{\mu}$ might underestimate μ, and on the basis of another particular set of observations, $\hat{\mu}$ might overestimate μ . But if we could average a huge number of estimates obtained from a huge number of sets of observations, then this average would *exactly* equal μ .

An unbiased estimator does not *systematically* over- or under-estimate the true parameter!

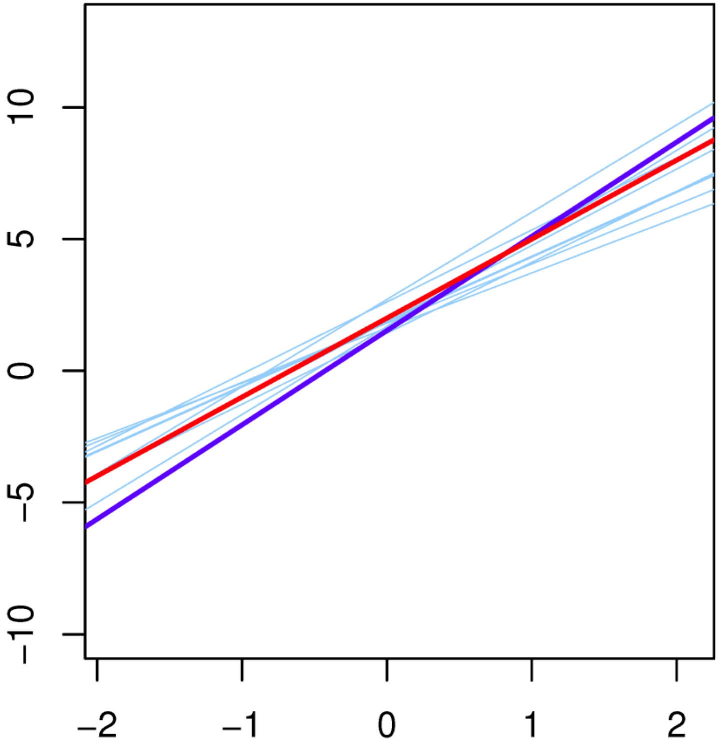
The property of unbiasedness holds for the least squares coefficient estimates.

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

Unbiased

The property of unbiasedness holds for the least squares coefficient estimates.



The average of many least square lines (blue and light blue ones) is very close to the true population regression line (red one)!

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

Standard error

The standard error of an estimator reflects how it varies under repeated sampling

$$\begin{cases} \mathrm{Var}(\hat{\beta}_0) = \mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \\ \\ \mathrm{Var}(\hat{\beta}_1) = \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{cases}$$

Proof

With $\sigma^2 = \mathrm{var}(\varepsilon) \xrightarrow{\text{can be empirically estimated}} \hat{\sigma} = \mathrm{RSE} = \sqrt{\mathrm{RSS}/(n-2)}$

there are *n* − 2 degrees of freedom for error (we are losing two degrees of freedom because we are estimating two parameters).

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

Confidence interval (CI)

Definition: a statistical concept used to describe the degree of confidence in an estimate.

Point 1: The width of the confidence interval depends on the characteristics of the sample data and the confidence level. Common confidence levels are 95% and 99%, which mean that there is a 95% or 99% probability that the true value of the population parameter is within the calculated confidence interval when the estimation is repeated.

Point 2: the probability interpretation of the confidence interval does not refer to the probability that the true value of the population parameter is within the interval, since the true value is fixed. Instead, the probability interpretation of the confidence interval is based on the repetition of sampling.

Simple Linear Regression

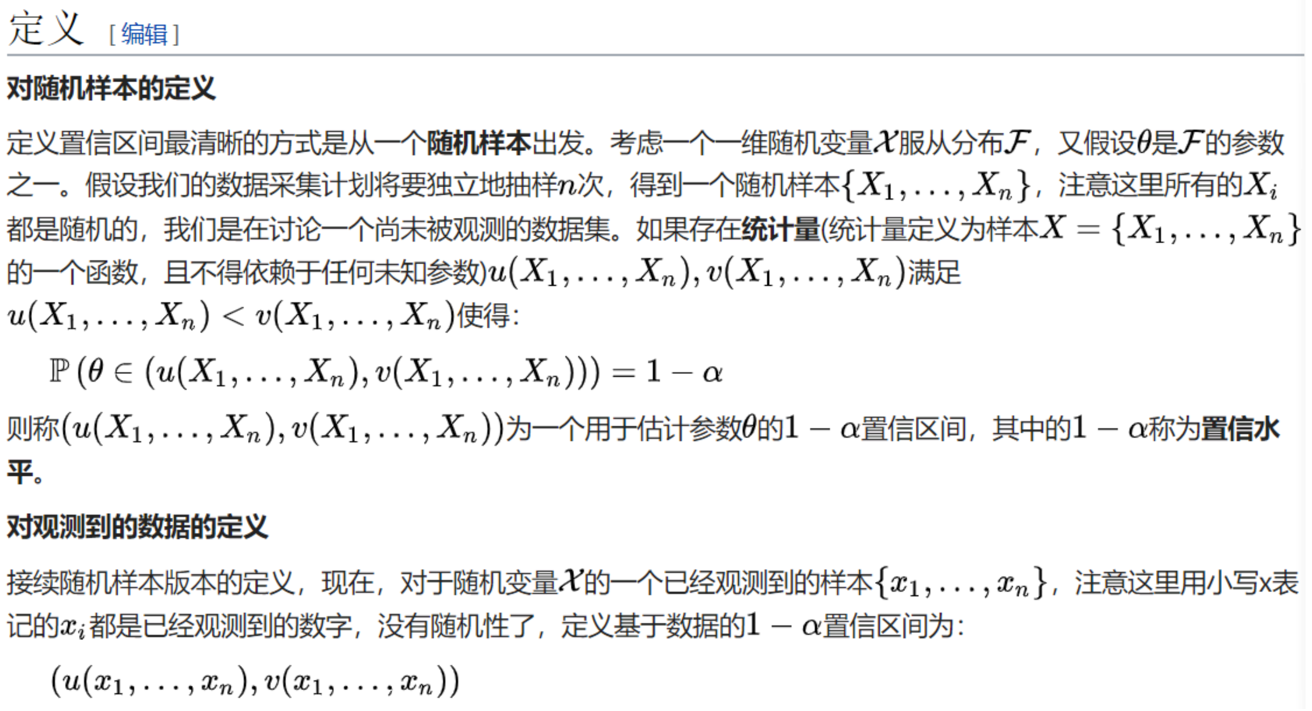
❑ Assessing the accuracy of the coefficient estimates

Confidence interval (CI)

A 95% CI is defined as a range of values such that our estimation procedure will have a 95% probability in terms of providing a range that will contain the true unknown value of the parameter.

For simple linear regression, the 95% CI is approximately take the form:

$$\left\{\begin{array}{l} \left[\hat{\beta}_1 - 2\cdot\mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2\cdot\mathrm{SE}(\hat{\beta}_1)\right] \\ \left[\hat{\beta}_0 - 2\cdot\mathrm{SE}(\hat{\beta}_0),\ \hat{\beta}_0 + 2\cdot\mathrm{SE}(\hat{\beta}_0)\right] \end{array}\right.$$



定义 [编辑]

对随机样本的定义

定义置信区间最清晰的方式是从一个**随机样本**出发。考虑一个一维随机变量$\mathcal{X}$服从分布$\mathcal{F}$，又假设θ是$\mathcal{F}$的参数之一。假设我们的数据采集计划将要独立地抽样n次，得到一个随机样本$\{X_1,\ldots,X_n\}$，注意这里所有的X_i都是随机的，我们是在讨论一个尚未被观测的数据集。如果存在**统计量**(统计量定义为样本$X=\{X_1,\ldots,X_n\}$的一个函数，且不得依赖于任何未知参数)$u(X_1,\ldots,X_n),v(X_1,\ldots,X_n)$满足$u(X_1,\ldots,X_n)<v(X_1,\ldots,X_n)$使得：

$$\mathbb{P}\left(\theta\in(u(X_1,\ldots,X_n),v(X_1,\ldots,X_n))\right)=1-\alpha$$

则称$(u(X_1,\ldots,X_n),v(X_1,\ldots,X_n))$为一个用于估计参数$\theta$的$1-\alpha$置信区间，其中的$1-\alpha$称为**置信水平。**

对观测到的数据的定义

接续随机样本版本的定义，现在，对于随机变量$\mathcal{X}$的一个已经观测到的样本$\{x_1,\ldots,x_n\}$，注意这里用小写x表记的x_i都是已经观测到的数字，没有随机性了，定义基于数据的$1-\alpha$置信区间为：

$$(u(x_1,\ldots,x_n),v(x_1,\ldots,x_n))$$

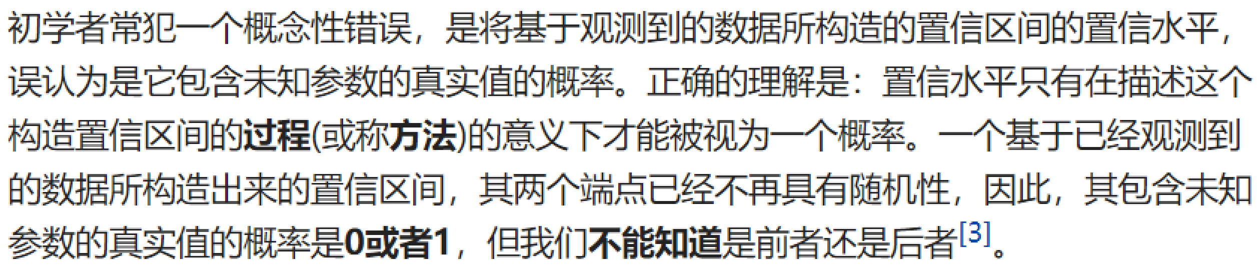
Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

Confidence interval (CI)

Some misunderstandings on CI:

- A 95% confidence interval does not mean that for a given realized interval calculated from sample data there is a 95% probability the population parameter lies within the interval, nor that there is a 95% probability that the interval covers the population parameter. Once an experiment is done and an interval calculated, this interval either covers the parameter value or it does not; it is no longer a matter of probability. **The 95% probability relates to the reliability of the estimation procedure, not to a specific calculated interval**



初学者常犯一个概念性错误，是将基于观测到的数据所构造的置信区间的置信水平，误认为是它包含未知参数的真实值的概率。正确的理解是：置信水平只有在描述这个构造置信区间的**过程**(或称**方法**)的意义下才能被视为一个概率。一个基于已经观测到的数据所构造出来的置信区间，其两个端点已经不再具有随机性，因此，其包含未知参数的真实值的概率是**0或者1**，但我们**不能知道**是前者还是后者[3]。

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

Confidence interval (CI)

Some misunderstandings on CI:

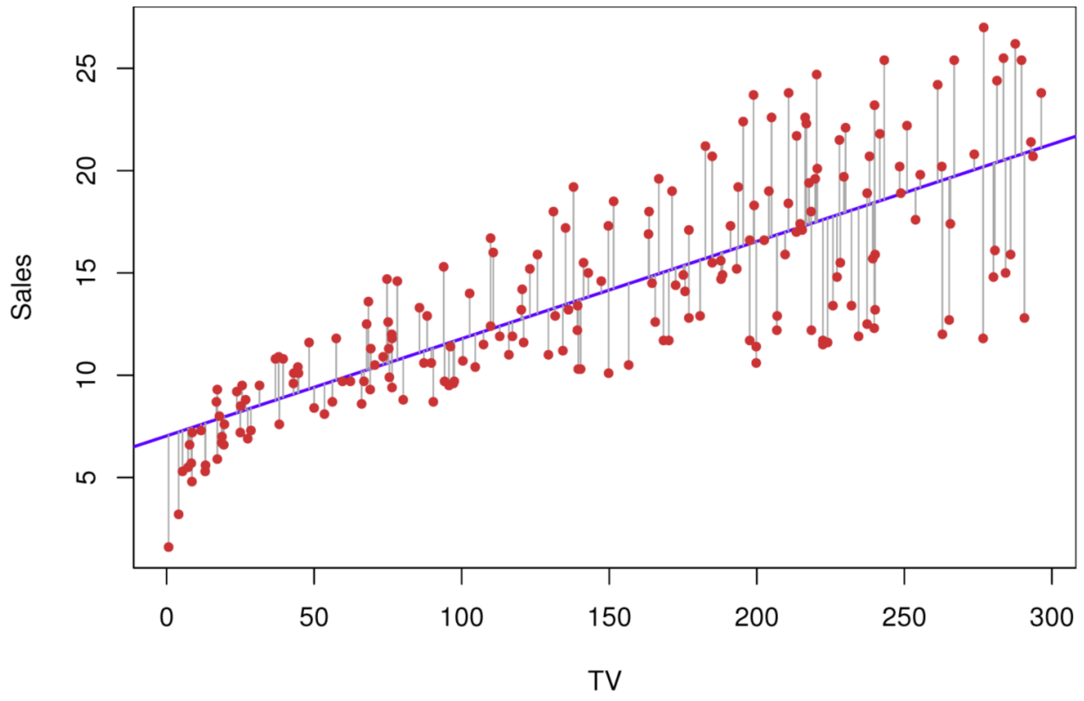
- A 95% confidence interval does not mean that 95% of the sample data lie within the interval.

- A confidence interval is not a range of plausible values for the sample mean, though it may be understood as an estimate of plausible values for the population parameter.

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

Confidence interval (CI)



CI for β_0 : $[6.130, 7.935]$

CI for β_1 : $[0.042, 0.053]$

A general conclusion in statistical consulting:

In the absence of any advertising, sales will, on average, fall somewhere between 6.130 and 7.935 units. Also, for each $1,000 increase in TV advertising, there will be an average increase in sales of between 42 and 53 units.

Simple Linear Regression

- ❑ Assessing the accuracy of the coefficient estimates

Hypothesis testing $Y = \beta_0 + \beta_1 X + \varepsilon$

What is hypothesis testing: In simple linear regression, we need to test the hypothesis of the coefficient to determine whether the independent variable has a significant impact on the dependent variable.

Steps:

1. State the null hypothesis and alternative hypothesis
2. Choose the appropriate test statistic and determine the level of significance (alpha)
3. Collect data and calculate the test statistic
4. Determine the p-value based on the test statistic
5. Compare the p-value to the level of significance (alpha)
6. Make a decision about whether to reject or fail to reject the null hypothesis

Simple Linear Regression

- ❑ Assessing the accuracy of the coefficient estimates

Hypothesis testing $Y = \beta_0 + \beta_1 X + \varepsilon$

Testing the *null hypothesis* of

$H_0:$ There is no relationship between X and Y

Versus the *alternative hypothesis*

$H_a:$ There is some relationship between X and Y

Mathematically:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_a : \beta_1 \neq 0 \end{cases}$$