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.

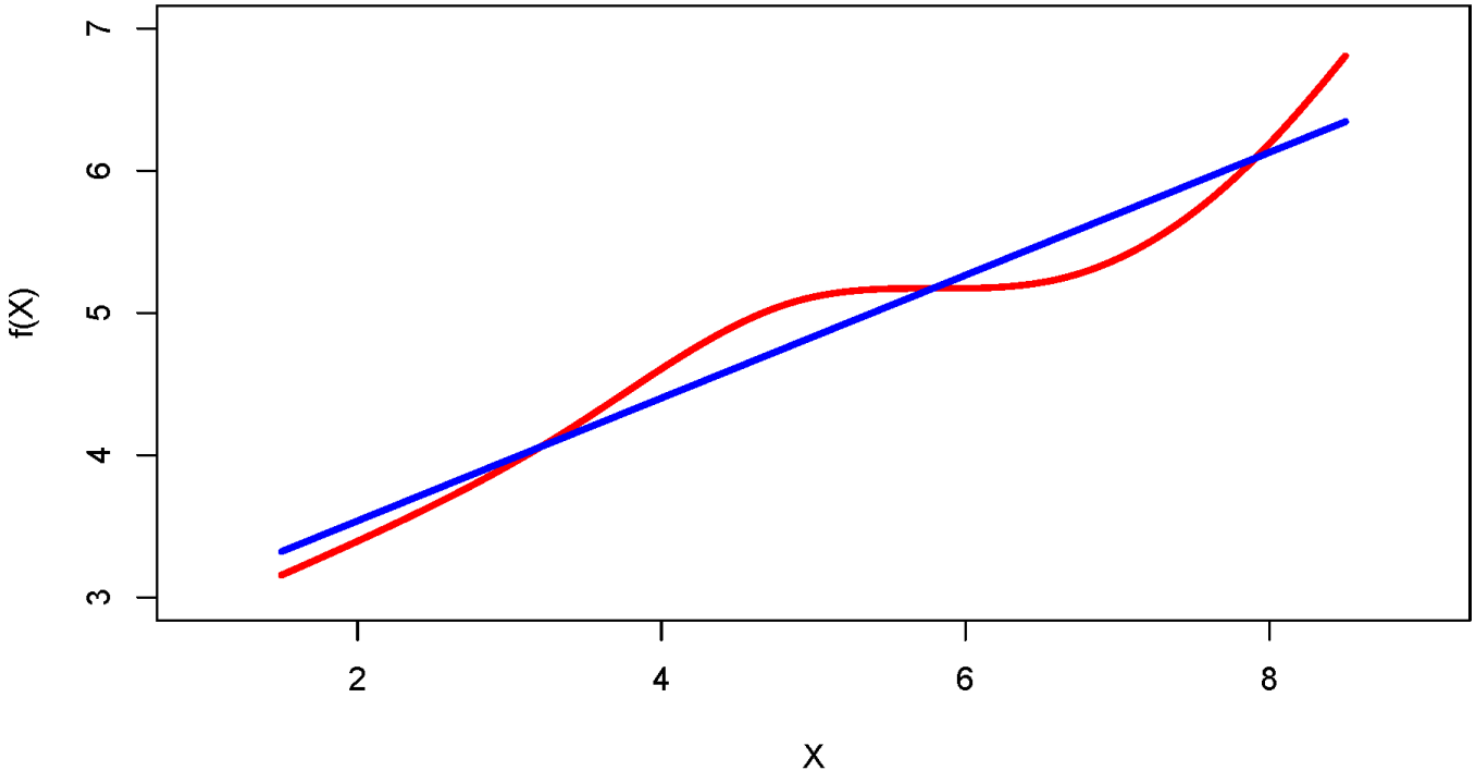
Linear Regression

❑ 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.
- ……

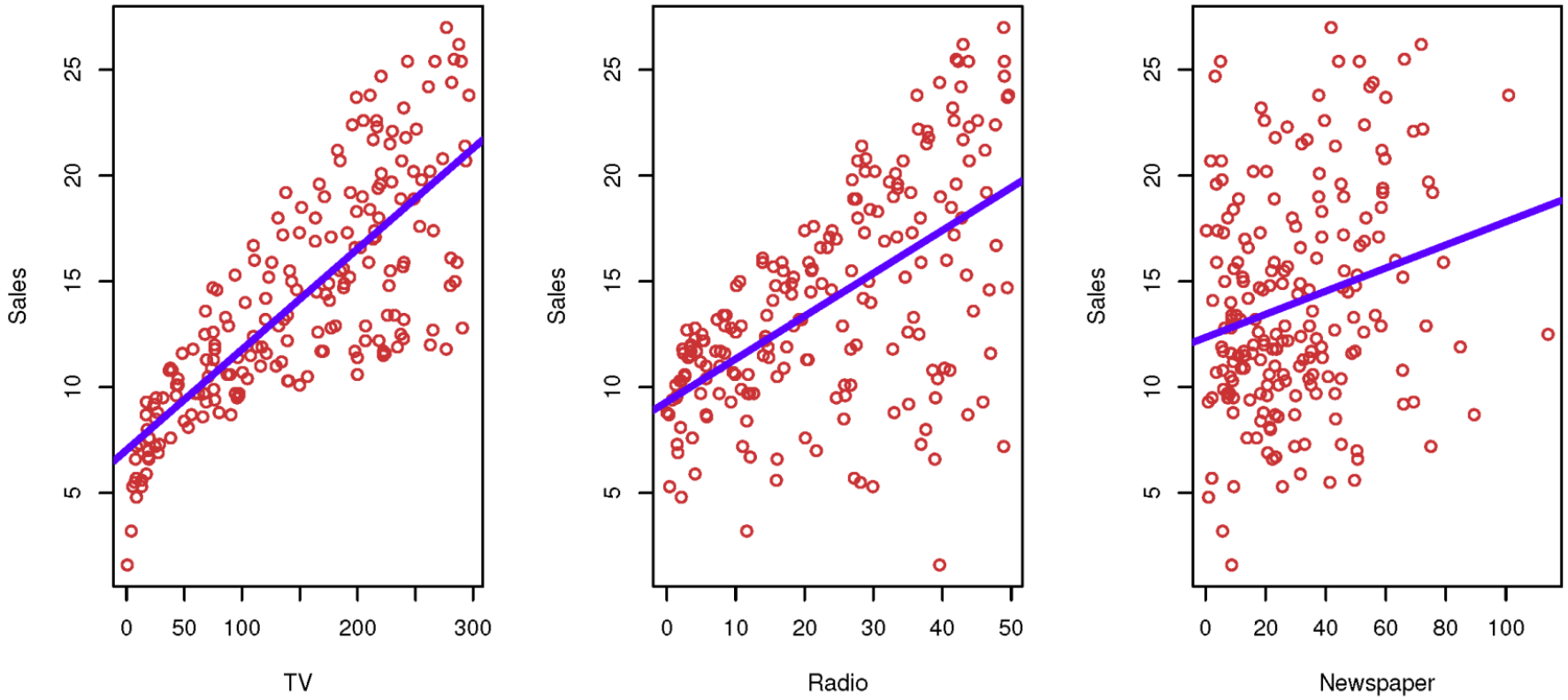
Linear Regression

- 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 $X = 0$)

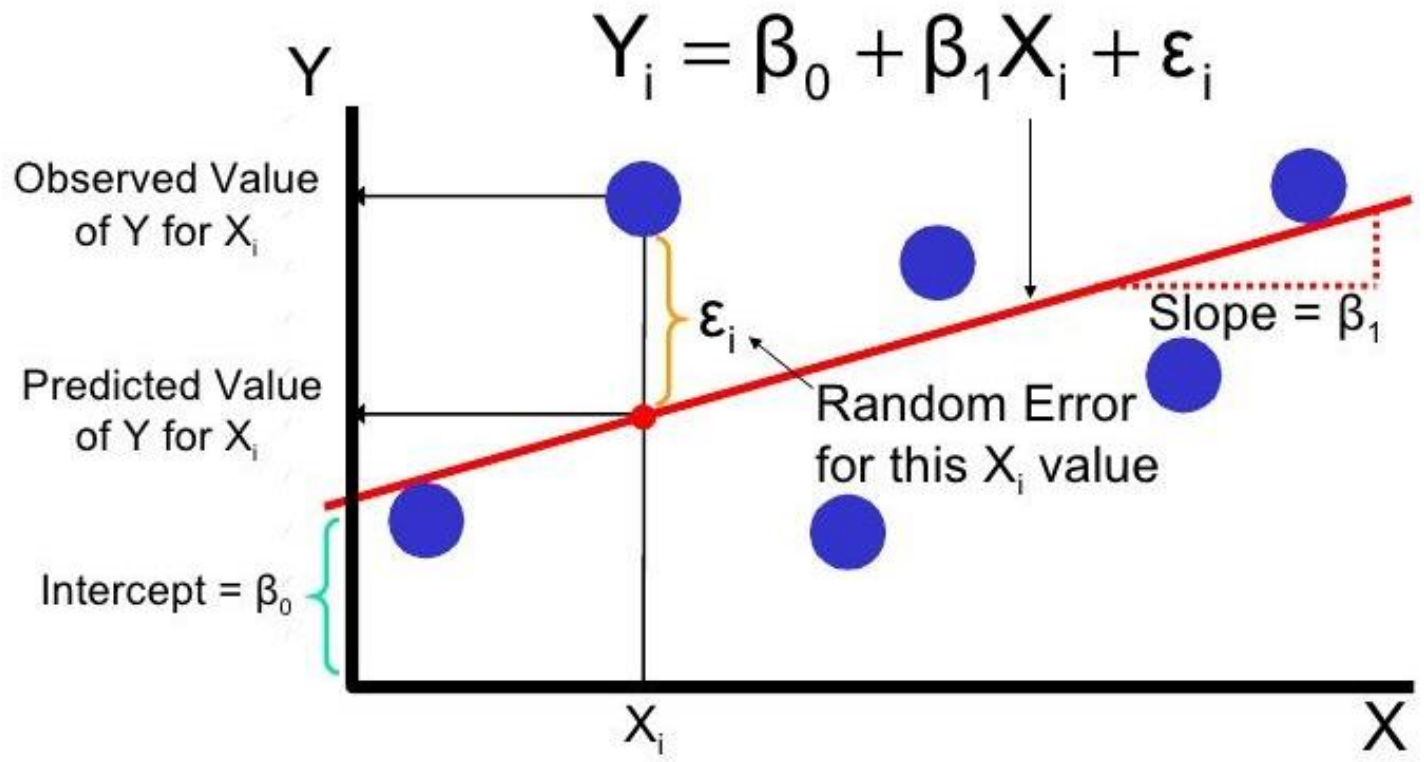
β_1 : slope (the average increase in Y associated with a one-unit increase in X)

(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



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 $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$

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

$$RSS = e_1^2 + e_2^2 + ... + e_n^2$$

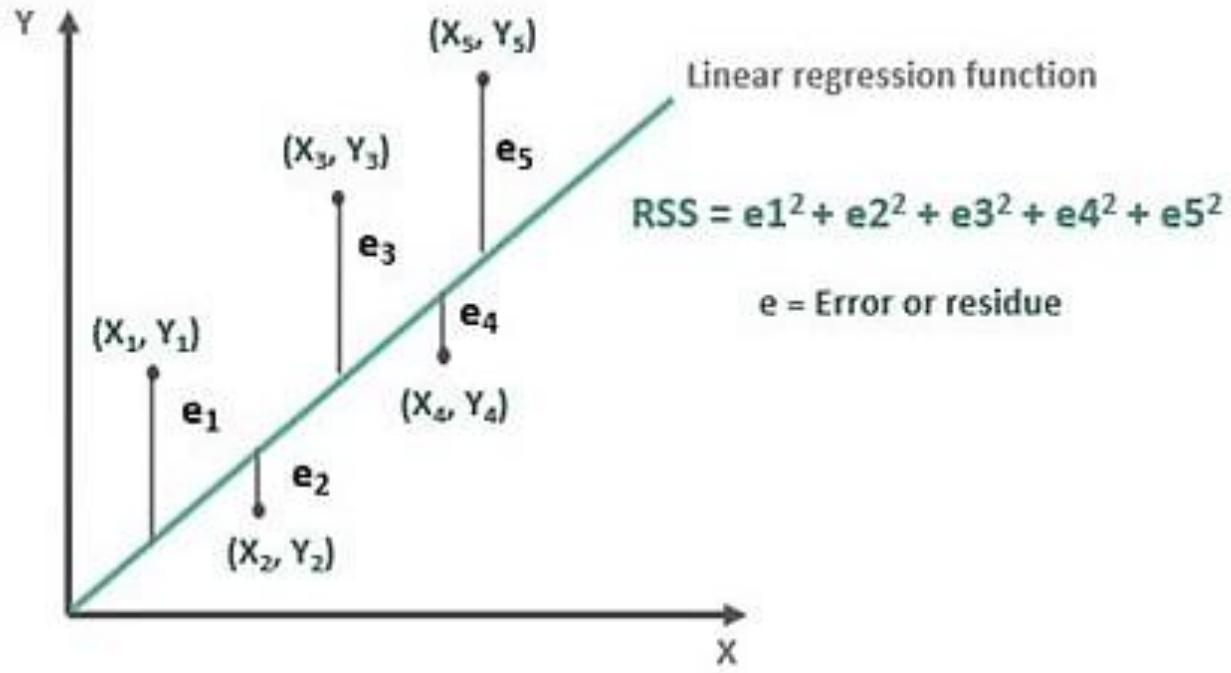
Or

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

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

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

公式推导：

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

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)

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

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

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

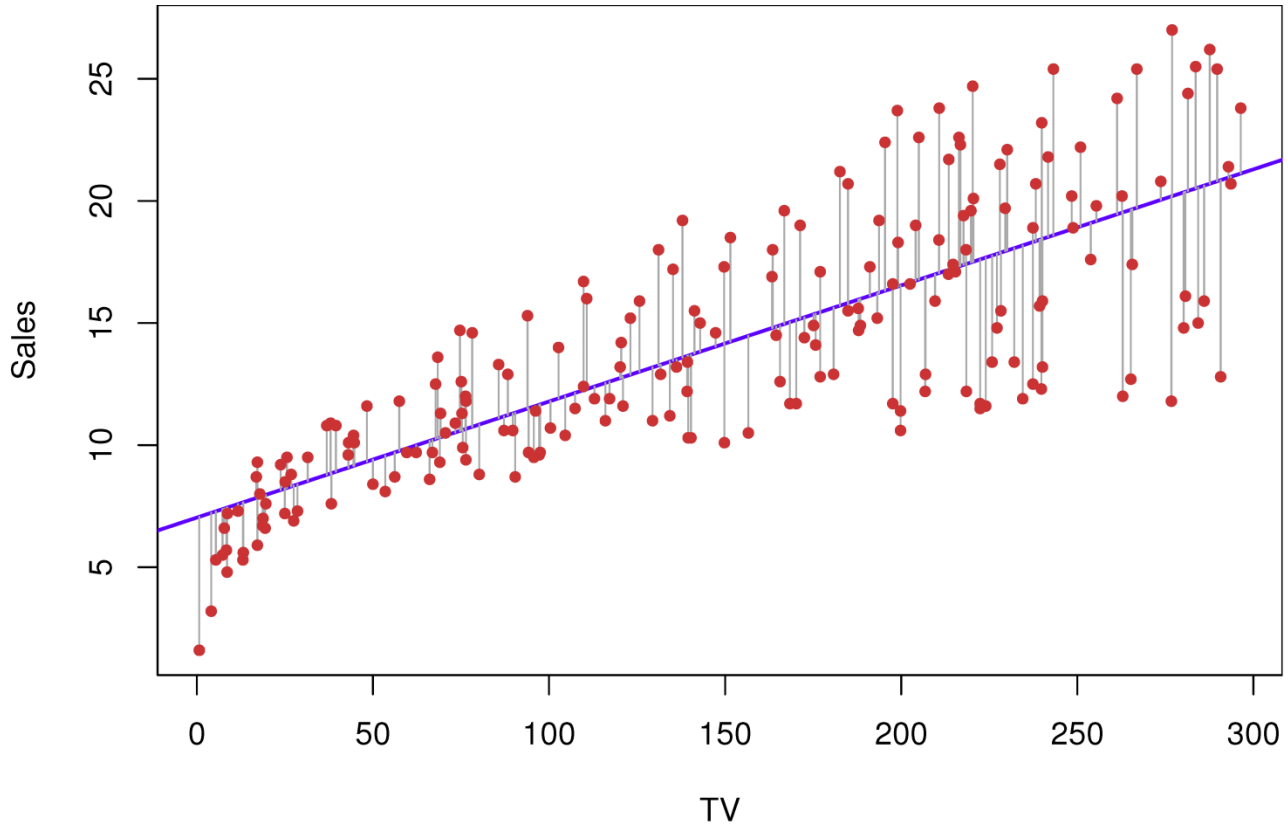
$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum x_i y_i - \sum \bar{y} x_i - \sum \bar{x} y_i + \sum \bar{x}\bar{y}}{\sum x_i^2 - 2\sum \bar{x} x_i + \sum \bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

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*



$\hat{\beta}_1 = 0.0475$

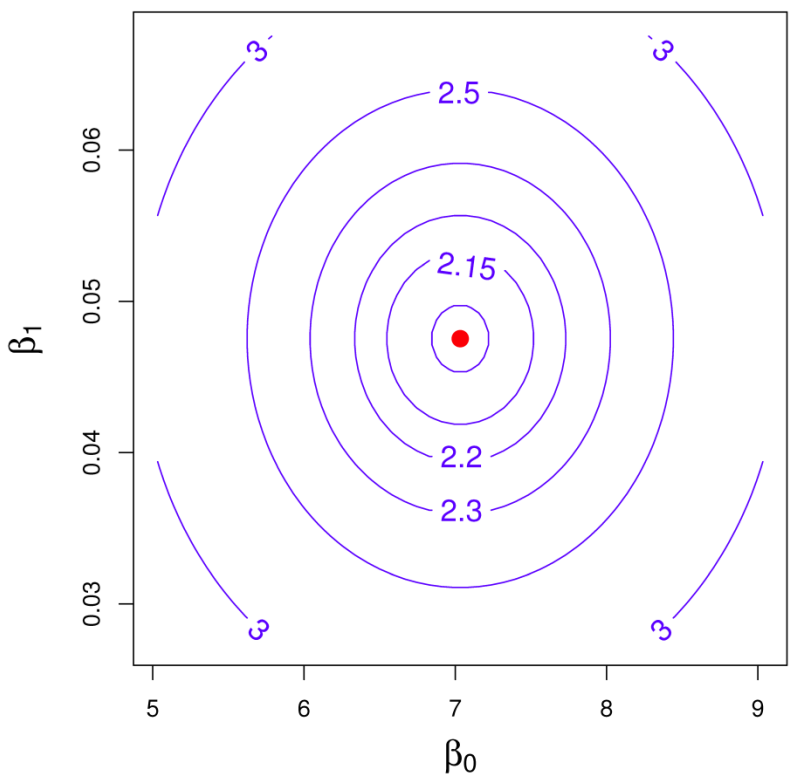
$\hat{\beta}_0 = 7.03$

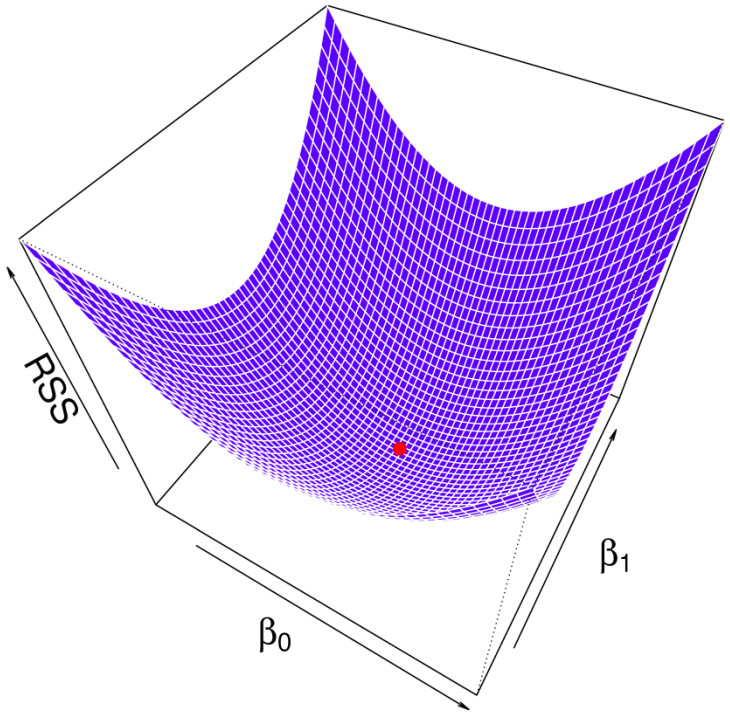
Conclusion: an addition $1,000 spent on TV advertising is associated with selling approximately 47.5 additional units of the product.

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*

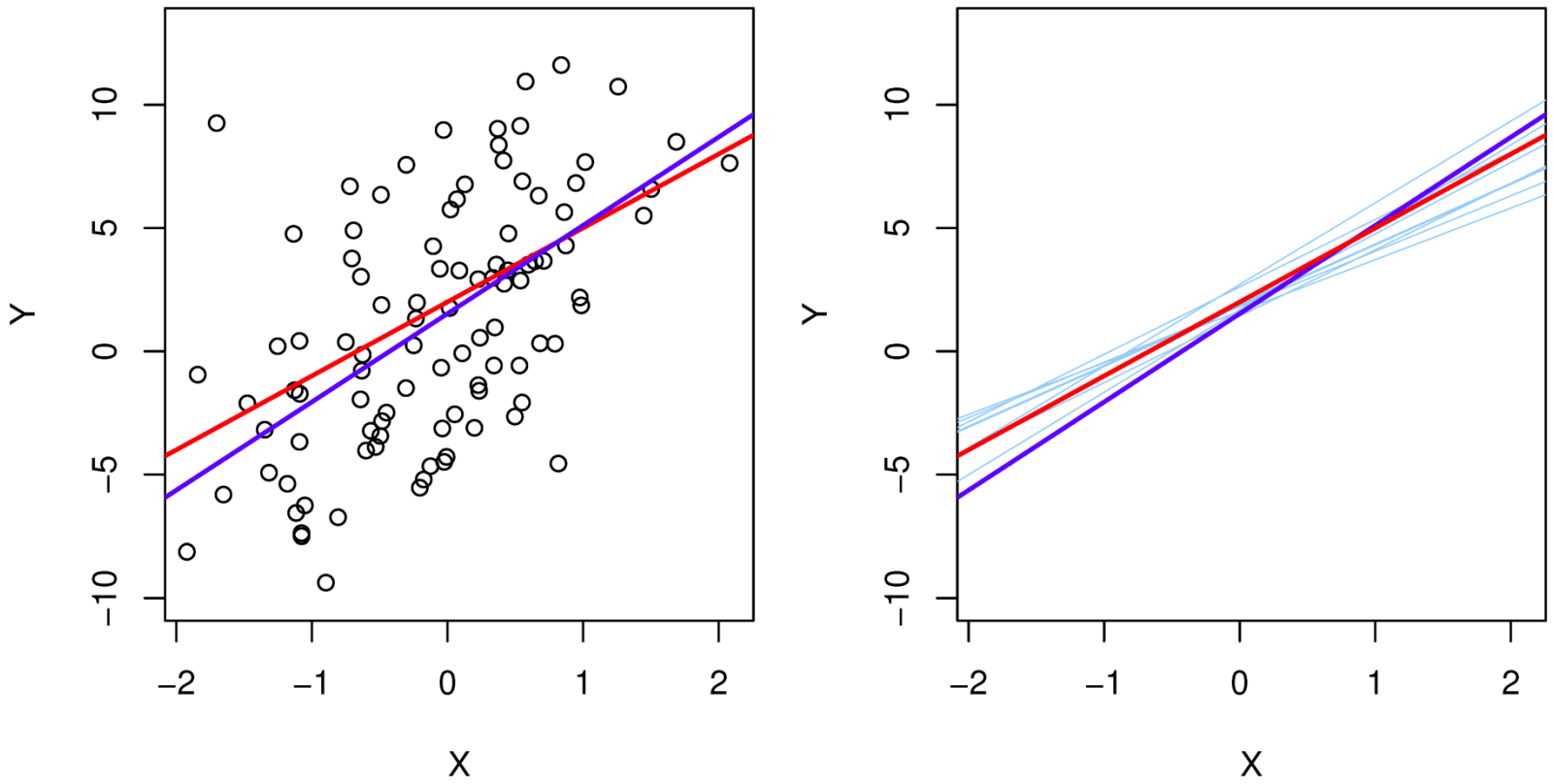
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$ *--least squares line*

- 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!

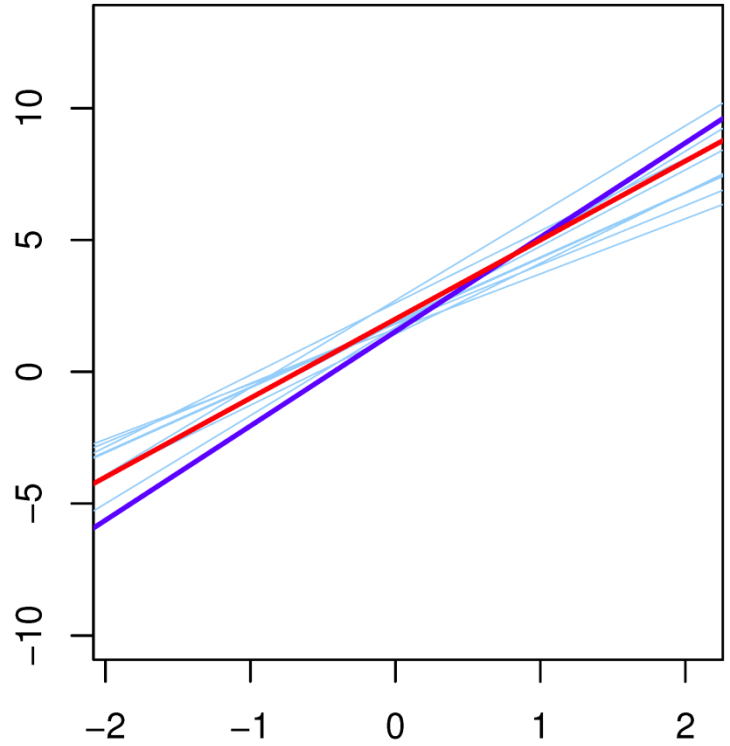
The property of unbiasedness holds for the least squares coefficient estimates.

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

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

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Proof

With $\sigma^2 = \mathrm{var}(\varepsilon)$ —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:

$$\left[\hat{\beta}_1 - 2\cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2\cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

$$\left[\hat{\beta}_0 - 2\cdot \mathrm{SE}(\hat{\beta}_0),\ \hat{\beta}_0 + 2\cdot \mathrm{SE}(\hat{\beta}_0)\right]$$

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)

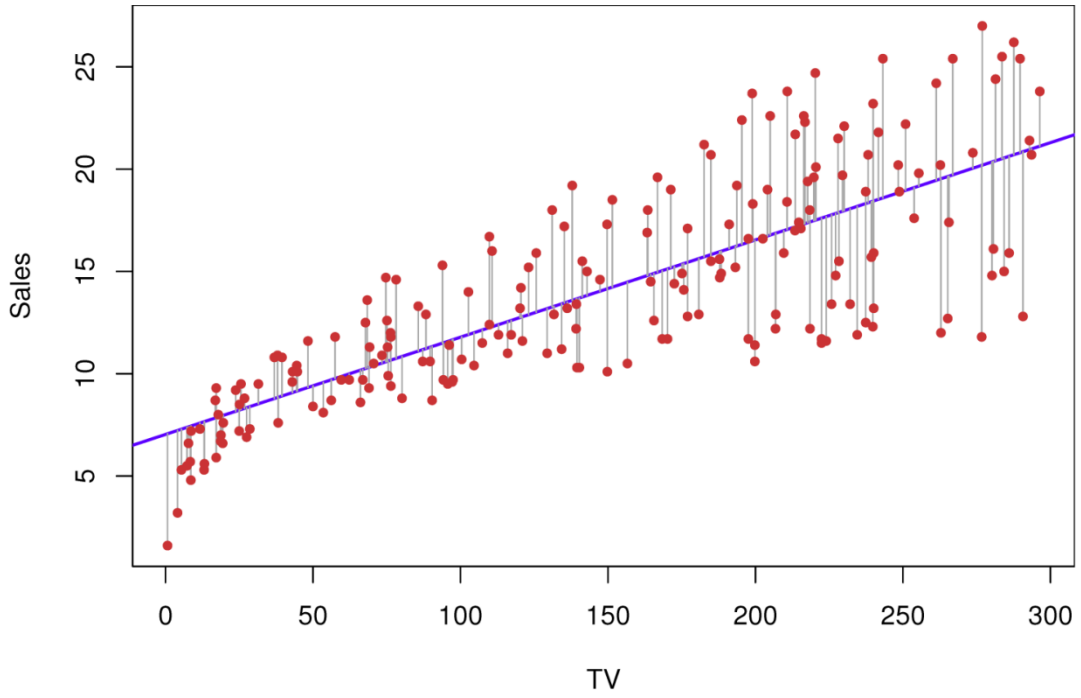
Some misunderstandings on CI:

- 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)



$$\text{CI for } \beta_0: [6.130, 7.935]$$

$$\text{CI for } \beta_1: [0.042, 0.053]$$

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:

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

Simple Linear Regression

- Assessing the accuracy of the coefficient estimates

Hypothesis testing

To test the null hypothesis, we need to determine whether $\hat{\beta}_1$, our estimate for β_1 , is sufficiently far from zero that we can be confident that $\hat{\beta}_1$ is non-zero.

How far is far enough?

It depends on the accuracy of $\hat{\beta}_1$ (depends on $SE(\hat{\beta}_1)$). If $SE(\hat{\beta}_1)$ is small, then even relatively small values of $\hat{\beta}_1$ may provide strong evidence that $\hat{\beta}_1 \neq 0$, and hence there is a relationship between X and Y. In contrast, if $SE(\hat{\beta}_1)$ is large, then $\hat{\beta}_1$ must be large in absolute value in order for us to reject the null hypothesis.

Simple Linear Regression

- Assessing the accuracy of the coefficient estimates

Hypothesis testing

How far is far enough?

t-statistic (to measure the number of standard deviations that $\hat{\beta}_1$ is away from 0)

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

Note: in every Hypothesis Testing case, we need to define a testing statistic!

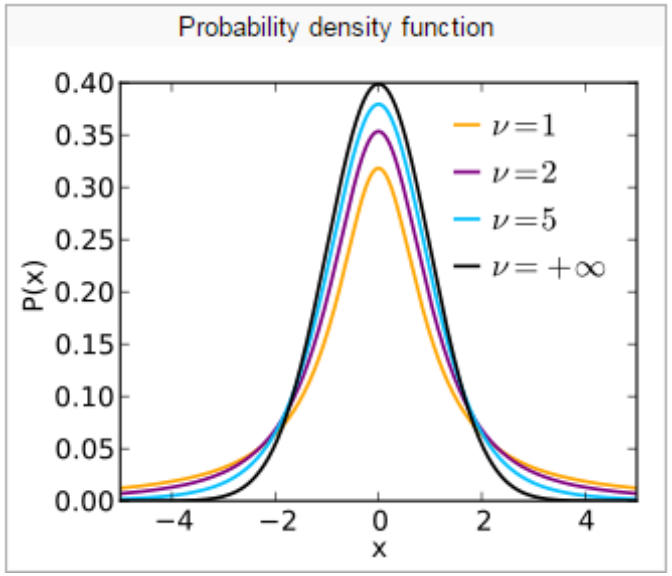
Simple Linear Regression

- Assessing the accuracy of the coefficient estimates

Hypothesis testing

When there is no relationship between *X* and *Y*, namely $\beta_1 = 0$, we have

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)} \sim \text{t-distribution with df } = n - 2$$



p-value: $\Pr(T \geq |t|)$

A small p-value indicates that it is ***unlikely*** to observe such a substantial association between the predictor and the response ***due to chance***, in the absence of any real association between the predictor and the response.

Simple Linear Regression

- Assessing the accuracy of the coefficient estimates

Hypothesis testing

We *reject the null hypothesis* – that is, we declare a relationship to exist between *X* and *Y* – if the p-value is small enough (typically used p-value cutoffs are 0.05 or 0.01).

Results for the advertising data (sales-versus-TV):

	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
Slope	0.0475	0.0027	17.67	< 0.0001

Simple Linear Regression

- Assessing the accuracy of the coefficient estimates

Hypothesis testing

	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
Slope	0.0475	0.0027	17.67	< 0.0001

$\beta_1 \neq 0 \Rightarrow$ there is a (statistically significant) relationship between TV and sales

$\beta_0 \neq 0 \Rightarrow$ In the absence of TV expenditure, sales are (statistically significantly) non-zeros

Simple Linear Regression

- Assessing the accuracy of the coefficient estimates

To quantify the extent to which the model fits the data

Residual Standard Error (RSE)

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

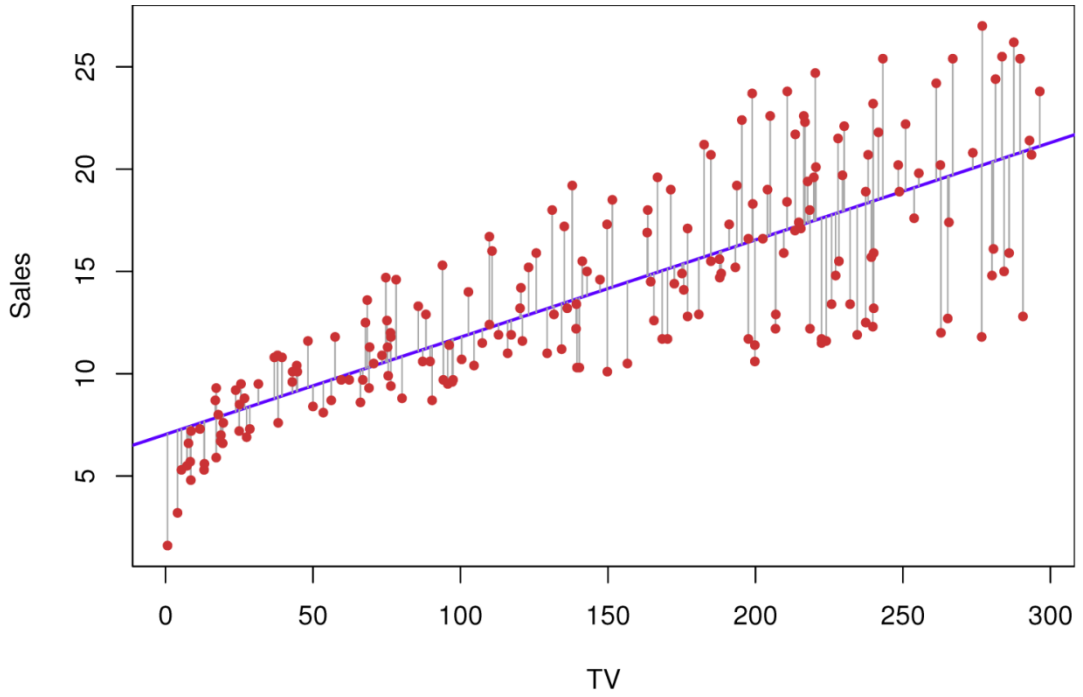
The RSE is an estimate of the standard deviation of ε . Roughly speaking, it is the average amount that the response will deviate from the true regression line.

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

To quantify the extent to which the model fits the data

Residual Standard Error (RSE)



$\mathrm{RSE}: 3.26$ A measure of the *lack of fit* of the model

How to interpret?

- Actual sales in each market deviate from the true regression line by approximately 3,260 units, on average.
- Even if the model were correct and the true values of the unknown coefficients were known exactly, any prediction of sales on the basis of TV advertising would still be off by about 3,260 units on average.

Simple Linear Regression

- Assessing the accuracy of the coefficient estimates

To quantify the extent to which the model fits the data

R^2 *Statistic*

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}, \text{ with TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- TSS measures the total variation in the response *Y*, is the amount of variability inherent in the response before the regression is performed.
- RSS measures the amount of variability that is left unexplained after performing the regression.
- TSS-RSS measures the amount of variability in the response that is explained by performing the regression.
- R^2 measures the proportion of variability in Y that can be explained using X.

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

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

To quantify the extent to which the model fits the data

R^2 *Statistic*

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}, \text{ with } \text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$ 这里面的y拔是样本真实值的平均值

- Ranges between 0 and 1.
- Independent of the scale of Y.
- A number that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression.
- A number near 0 indicates that the regression did not explain much of the variability in the response (the linear model is wrong, the inherent error is high).

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

To quantify the extent to which the model fits the data

R^2 *Statistic*

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}, \text{ with } \text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

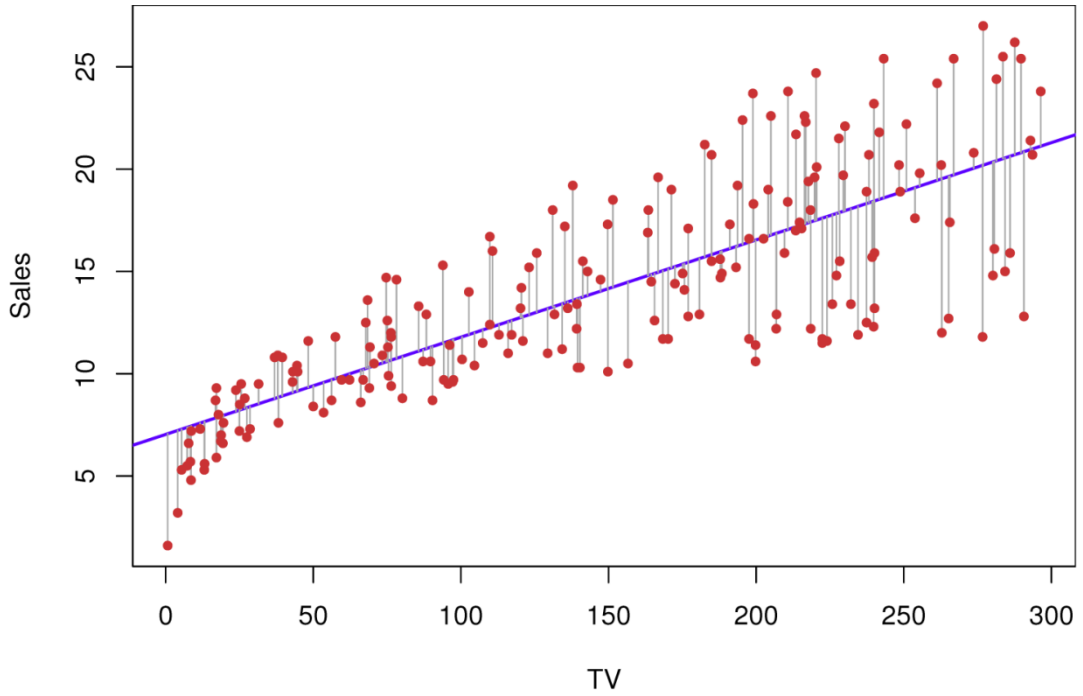
- It can only be used to measure the fitting degree of the linear regression model, not the nonlinear regression model
- It cannot be used to compare the performance of different sets of independent variables in building regression models because different sets of independent variables can have an impact on the R^2 value.

Simple Linear Regression

- Assessing the accuracy of the coefficient estimates

To quantify the extent to which the model fits the data

R^2 *Statistic*



$R^2 : 0.61$

How to interpret?

- Just under two-thirds of the variability in sales is explained by a linear regression on TV.

Simple Linear Regression

- Assessing the accuracy of the coefficient estimates

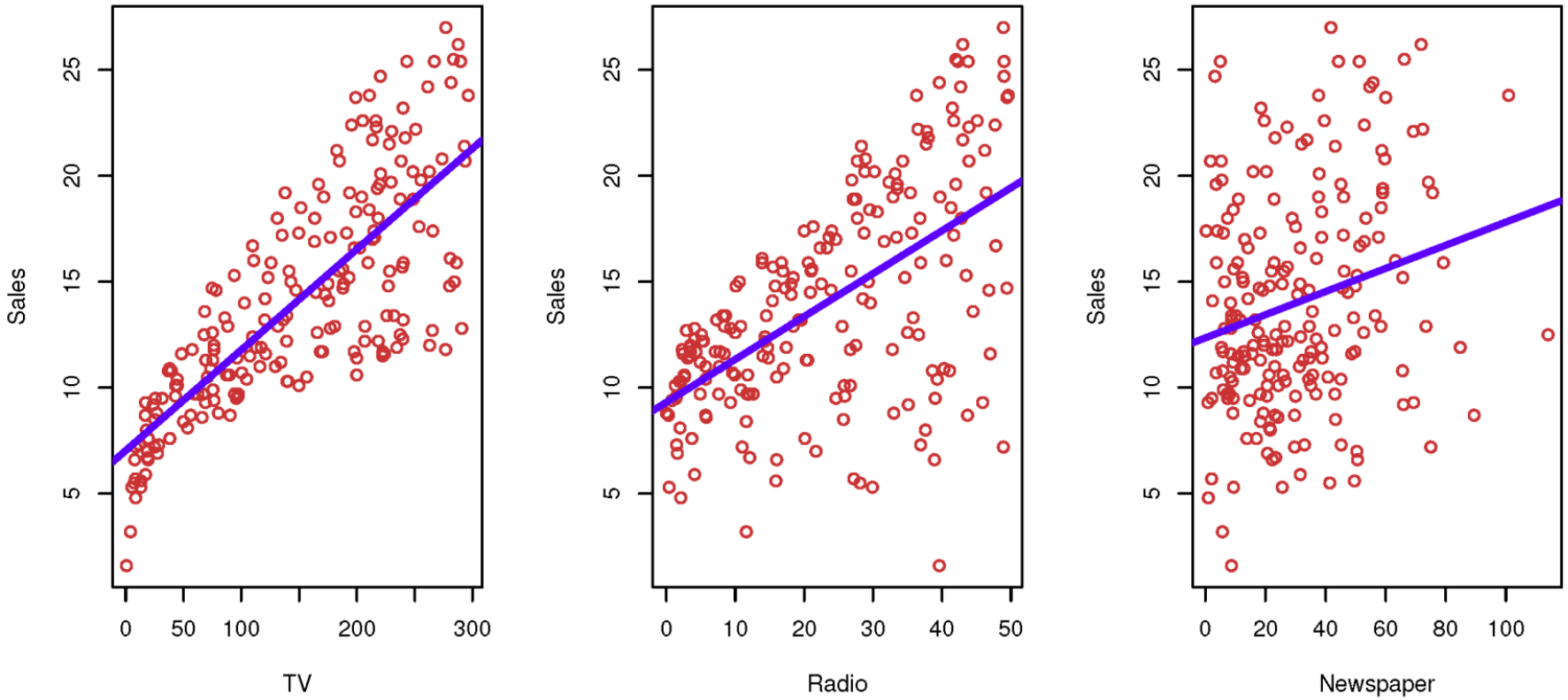
To quantify the extent to which the model fits the data

$$R^2 = [\mathrm{cor}(X, Y)]^2$$

The above equation is only true for simple linear regression, but not for the multiple linear regression case.

Multiple Linear Regression

- In practice, we often have more than one predictor.



Multiple Linear Regression

- Suppose we have p distinct predictors

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

  - X_j represents the *j*th predictor.
  - β_j quantifies the association between X_j and the response (the average effect on the response of a one unit increase in the predictor, *holding all other predictors fixed*).

- Advertising data

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \varepsilon$$

Multiple Linear Regression

- Interpreting regression coefficients

  - The ideal scenario is when the predictors are uncorrelated (a balanced design):
    - Each coefficient can be estimated and tested separately.
    - Interpretations such as, “a unit change in X_j is associated with a β_j change in Y, while *all the variables stay fixed*”, are possible.

  - Correlations among predictors cause problems:
    - The variance of all coefficients tends to increase, sometimes dramatically
    - Interpretations become hazardous --- when X_j changes, everything else changes.

  - *Claims of causality* should be avoided for observational data.

Multiple Linear Regression

- The woes of (interpreting) regression coefficients
  - A regression coefficient β_j estimates the expected change in Y per unit change in X_j

    with all other predictors held fixed. But predictors usually change together!
  - Example: *Y* = number of tackles by a football player in a season; *W* and *H* are his weight and height. Fitted regression model is $\hat{Y} = b_0 + 0.5W - 0.1H$. How do we interpret $\hat{\beta}_2 = -0.1 < 0$?

Multiple Linear Regression

- Two quotes by famous statisticians

"Essentially, all models are wrong, but some are useful"

– by George Box

"The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively"

– by Fred Mosteller and John Tukey, paraphrasing George Box

Multiple Linear Regression

- Estimating the Regression Coefficients

  - Given estimates $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$, we can make predictions using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p$$

  - Similar to simple linear regression, we estimate $\beta_0, \beta_1, ..., \beta_p$ as the values that minimize the RSS

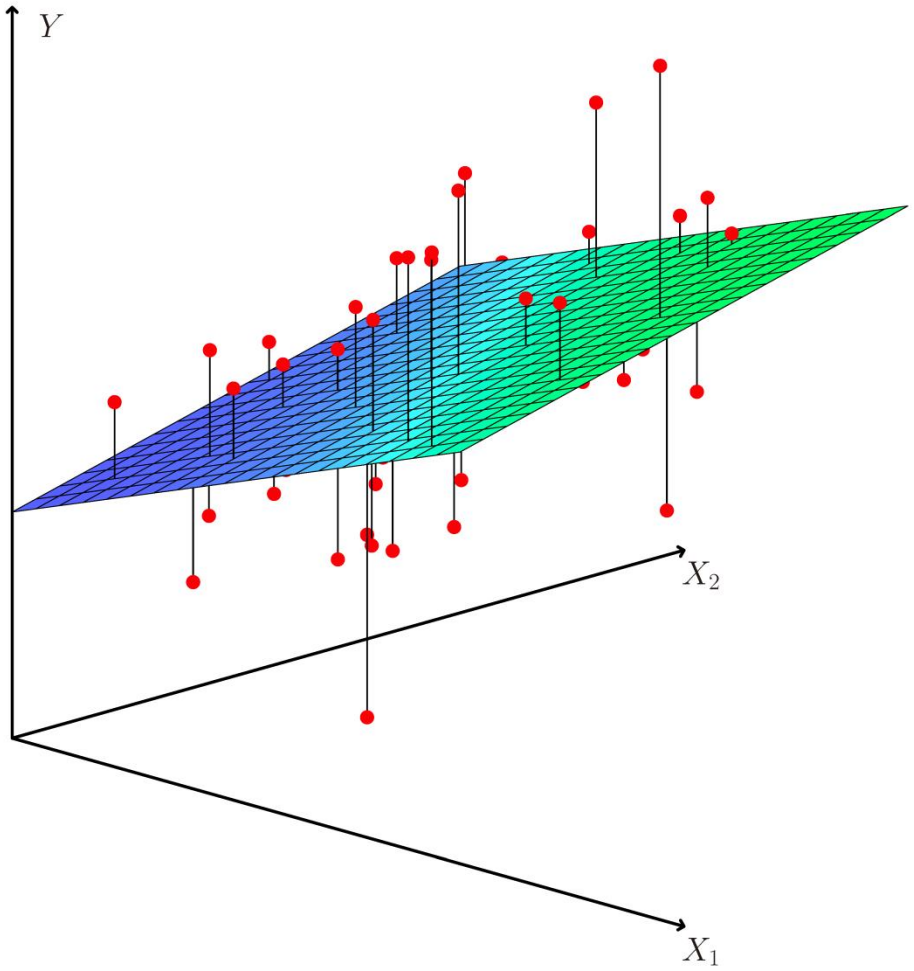
$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - ... - \hat{\beta}_p x_{ip})^2$$

  - The values $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ that minimize the above RSS are the multiple least squares regression coefficient estimates.

Multiple Linear Regression

- Estimating the Regression Coefficients



An example with two predictors and one response. In this case, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation and the plane.

Linear Regression

- Simple Linear Regression & Multiple Linear Regression

	Simple Linear Regression	Multiple Linear Regression
Independent Variables	Only one	Multiple
Model form	$Y = \beta 0 + \beta 1X$	$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta pXp$
Objective	Describing the relationship between two variables and predicting the dependent variable	Describing the relationship between multiple variables and predicting the dependent variable
Evaluation metrics	Residual sum of squares (SSE), coefficient of determination (R^2), T statistic, standard error (SE), etc.	Residual sum of squares (SSE), coefficient of determination (R^2), F statistic, standard error (SE), etc.
Model fitting	Least squares method	Least squares method

Linear Regression

- Simple Linear Regression & Multiple Linear Regression

  - *Multiple Linear Regression: can be considered as an extension of simple linear regression*
  - *Both are used to describe the relationship between independent variables and dependent variables, but multiple linear regression can handle the influence of multiple independent variables on the dependent variable*
  - *Multiple linear regression can be applied to more practical problems*