Probability and Statistics

Chapter 8: Simple Linear Regression and Correlation

Dr. Phan Thi Huong

HoChiMinh City University of Technology
Faculty of Applied Science, Department of Applied Mathematics
Email: huongphan@hcmut.edu.vn

HCM city — 2021.

Outline

OUTLINE

Outline

OUTLINE

LEARNING OUTCOMES

After careful study of this chapter, you should be able to do the following:

1. Understand how the method of least squares is used to estimate the parameters in a linear regression model.

LEARNING OUTCOMES

After careful study of this chapter, you should be able to do the following:

1. Understand how the method of least squares is used to estimate the parameters in a linear regression model.
2. Test statistical hypotheses and construct confidence intervals on regression model parameters.

LEARNING OUTCOMES

After careful study of this chapter, you should be able to do the following:

1. Understand how the method of least squares is used to estimate the parameters in a linear regression model.
2. Test statistical hypotheses and construct confidence intervals on regression model parameters.
3. Use the regression model to predict a future observation.

Learning outcomes

After careful study of this chapter, you should be able to do the following:

1. Understand how the method of least squares is used to estimate the parameters in a linear regression model.
2. Test statistical hypotheses and construct confidence intervals on regression model parameters.
3. Use the regression model to predict a future observation.
4. Analyze residuals to determine whether the regression model is an adequate fit to the data or whether any underlying assumptions are violated.

LEARNING OUTCOMES

After careful study of this chapter, you should be able to do the following:

1. Understand how the method of least squares is used to estimate the parameters in a linear regression model.
2. Test statistical hypotheses and construct confidence intervals on regression model parameters.
3. Use the regression model to predict a future observation.
4. Analyze residuals to determine whether the regression model is an adequate fit to the data or whether any underlying assumptions are violated.
5. Apply the correlation model

LEARNING OUTCOMES

After careful study of this chapter, you should be able to do the following:

1. Understand how the method of least squares is used to estimate the parameters in a linear regression model.
2. Test statistical hypotheses and construct confidence intervals on regression model parameters.
3. Use the regression model to predict a future observation.
4. Analyze residuals to determine whether the regression model is an adequate fit to the data or whether any underlying assumptions are violated.
5. Apply the correlation model
6. Use R software to fit simple linear regression models and interpret the output.

Regression analysis and binary data

We are often interested in trying to determine the relationship between a pair of variables. For instances,

Regression analysis and binary data

We are often interested in trying to determine the relationship between a pair of variables. For instances,

- how does the amount of money spent in advertising a new product relate to the first month's sales figures for that product?

REGRESSION ANALYSIS AND BINARY DATA

We are often interested in trying to determine the relationship between a pair of variables. For instances,

- how does the amount of money spent in advertising a new product relate to the first month's sales figures for that product?
- how does the height of a father relate to that of his son?

REGRESSION ANALYSIS AND BINARY DATA

We are often interested in trying to determine the relationship between a pair of variables. For instances,

- how does the amount of money spent in advertising a new product relate to the first month's sales figures for that product?
- how does the height of a father relate to that of his son?
- how does the electrical energy consumption of a house relate to the size of the house?

Regression analysis and binary data

We are often interested in trying to determine the relationship between a pair of variables. For instances,

- how does the amount of money spent in advertising a new product relate to the first month's sales figures for that product?
- how does the height of a father relate to that of his son?
- how does the electrical energy consumption of a house relate to the size of the house?
- ...

Regression analysis and binary data

We are often interested in trying to determine the relationship between a pair of variables. For instances,

- how does the amount of money spent in advertising a new product relate to the first month's sales figures for that product?
- how does the height of a father relate to that of his son?
- how does the electrical energy consumption of a house relate to the size of the house?
- ...

$\Rightarrow$ the relationship between those variables are not deterministic

REGRESSION ANALYSIS AND BINARY DATA

We are often interested in trying to determine the relationship between a pair of variables. For instances,

- how does the amount of money spent in advertising a new product relate to the first month's sales figures for that product?
- how does the height of a father relate to that of his son?
- how does the electrical energy consumption of a house relate to the size of the house?
- ...

⇒ the relationship between those variables are not deterministic
⇒ The collection of statistical tools that are used to model and explore relationships between variables that are related in a nondeterministic manner is called regression analysis.

A simple linear regression model

The case of simple linear regression considers a single predictor variable or independent variable X and a dependent or response variable Y.

A SIMPLE LINEAR REGRESSION MODEL

The case of simple linear regression considers a single predictor variable or independent variable X and a dependent or response variable Y.
Suppose that for a specified value X of the independent variable the value of the response variable Y can be expressed as

$$Y = \beta_0 + \beta_1 x + \varepsilon, \tag{1}$$

where

- β_0, β_1 are unknown parameters and called regression coefficients.

A SIMPLE LINEAR REGRESSION MODEL

The case of simple linear regression considers a single predictor variable or independent variable X and a dependent or response variable Y.
Suppose that for a specified value X of the independent variable the value of the response variable Y can be expressed as

$$Y = \beta_0 + \beta_1 x + \varepsilon, \tag{1}$$

where

- β_0, β_1 are unknown parameters and called regression coefficients.
- ε is called the random error and assumed to be normally distributed with $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}ar(\varepsilon) = \sigma^2$.

A SIMPLE LINEAR REGRESSION MODEL

A simple linear regression model given in the equation (1) states that mean of the random variable Y is related to x by the following straight-line relationship:

$$\mathbb{E}[Y|x] = \beta_0 + \beta_1 x,$$

where β_0 and β_1 are respectively the intercept and the slope of the straight-line.

Assumptions of the error term

Given n pairs of observations (x_1, y_1), (x_2, y_2),... , (x_n, y_n) which are collected from a random sample of size n, the equation (1) indicates

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1,\, 2,\, \ldots, n$$

. A simple linear regression requires

ASSUMPTIONS OF THE ERROR TERM

Given n pairs of observations (x_1, y_1), (x_2, y_2),... , (x_n, y_n) which are collected from a random sample of size n, the equation (1) indicates

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1,\ 2,\ \ldots, n$$

. A simple linear regression requires

- The error terms ε_i are mutually independent.

ASSUMPTIONS OF THE ERROR TERM

Given n pairs of observations (x_1, y_1), (x_2, y_2),... , (x_n, y_n) which are collected from a random sample of size n, the equation (1) indicates

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1,\ 2,\ \ldots, n$$

. A simple linear regression requires

- The error terms ε_i are mutually independent.
- $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ or $Y \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$.

A SCATTER DIAGRAM FOR PAIR DATA

A SCATTER DIAGRAM FOR PAIR DATA

How might an observed dataset be a candidate for a simple linear regression model?
Let's consider a simple example of how the speed of a car affects its stopping distance, that is, how far it travels before it comes to a stop.

A SCATTER DIAGRAM FOR PAIR DATA

How might an observed dataset be a candidate for a simple linear regression model?
Let's consider a simple example of how the speed of a car affects its stopping distance, that is, how far it travels before it comes to a stop.

The `cars` dataset contains 50 observations of two variables `speed`(mph) and `dist` (ft).

A scatter diagram for pair data

How might an observed dataset be a candidate for a simple linear regression model?
Let's consider a simple example of how the speed of a car affects its stopping distance, that is, how far it travels before it comes to a stop.

The `cars` dataset contains 50 observations of two variables `speed`(mph) and `dist` (ft).

	speed	dist
1	4.00	2.00
2	4.00	10.00
3	7.00	4.00
4	7.00	22.00
5	8.00	16.00
...	...	...
48	24.00	93.00
49	24.00	120.00
50	25.00	85.00

A scatter diagram for pair data

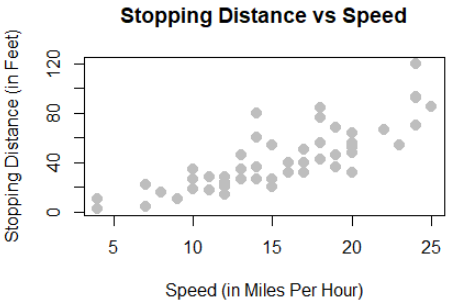


FIGURE 1: The scatter diagram of the cars dataset.

A scatter diagram for pair data

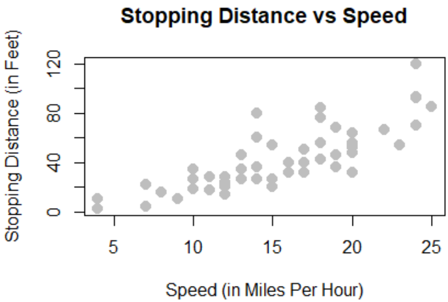


FIGURE 1: The scatter diagram of the `cars` dataset.

⇒ A scatter diagram of the observed dataset can give us an suggestion of a linear regression model.

ESTIMATING THE REGRESSION PARAMETERS

ESTIMATING THE REGRESSION PARAMETERS

- Let $\hat{\beta}_0$, $\hat{\beta}_1$ are respectively estimates of β_0 and β_1.

Estimating the regression parameters

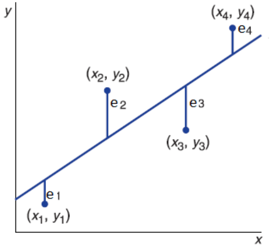
- Let $\hat{\beta}_0$, $\hat{\beta}_1$ are respectively estimates of β_0 and β_1.
- The fitted regression line is given by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

ESTIMATING THE REGRESSION PARAMETERS

- Let $\hat{\beta}_0$, $\hat{\beta}_1$ are respectively estimates of β_0 and β_1.
- The fitted regression line is given by

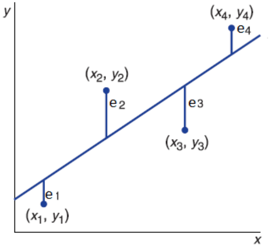
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$



Estimating the regression parameters

- Let $\hat{\beta}_0, \hat{\beta}_1$ are respectively estimates of β_0 and β_1.
- The fitted regression line is given by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

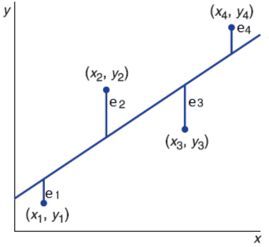


The residual $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$ describes the error in the fit of the model to the ith observation y_i.

ESTIMATING THE REGRESSION PARAMETERS

- Let $\hat{\beta}_0$, $\hat{\beta}_1$ are respectively estimates of β_0 and β_1.
- The fitted regression line is given by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$



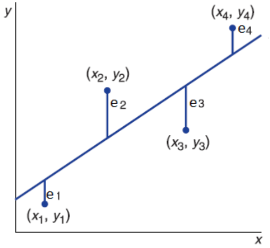
The residual $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$ describes the error in the fit of the model to the ith observation y_i.

- The key concept: An optimized fitted regression line should be "close to the observed data".

ESTIMATING THE REGRESSION PARAMETERS

- Let $\hat{\beta}_0$, $\hat{\beta}_1$ are respectively estimates of β_0 and β_1.
- The fitted regression line is given by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$



The residual $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$ describes the error in the fit of the model to the ith observation y_i.

- The key concept: An optimized fitted regression line should be "close to the observed data".
- $\hat{\beta}_0$ and $\hat{\beta}_1$ will be found by the least-square method.

ESTIMATING THE REGRESSION PARAMETERS

DEFINITION

For a dataset of n observations (x_1, y_1), ..., (x_n, y_n), the sum of squares for errors is defined by

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

Estimating the regression parameters

Definition

For a dataset of n observations (x_1, y_1), ..., (x_n, y_n), the sum of squares for errors is defined by

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

The least-square method aims to find the estimates $\hat{\beta}_0$, and $\hat{\beta}_1$ by minimizing SSE. Those estimates are called least squares estimates.

Estimating the regression parameters

Theorem

The least squares estimates of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \dfrac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \dfrac{\left(\sum_{i=1}^n x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}}, \qquad \text{and} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Estimating the regression parameters

Theorem

The least squares estimates of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \dfrac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \dfrac{\left(\sum_{i=1}^n x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}}, \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where S_{xx} and S_{xy} are defined by

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}, \qquad \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}, \qquad \bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

Estimating the regression parameters

Example 1

A large midwestern bank is planning on introducing a new word processing system to its secretarial staff. To learn about the amount of training that is needed to effectively implement the new system, the bank chose eight employees of roughly equal skill. These workers were trained for different amounts of time and were then individually put to work on a given project. The following data indicate the training times and the resulting times (both in hours) that it took each worker to complete the project.

Training time($= x$)	22	18	30	16	25	20	10	14
Time to complete project ($= Y$)	18.4	19.2	14.5	19.0	16.6	17.7	24.4	21.0

ESTIMATING THE REGRESSION PARAMETERS

EXAMPLE 1 (CONTINUED)

(A) What is the estimated regression line?

(B) Predict the amount of time it would take a worker who receives 28 hours of training to complete the project.

(C) Find the residual e_i of an observation $(x_i, y_i) = (22, 18.4)$.

Solution:

Analysis of variance

Analysis of variance

- $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = S_{yy}$ is the total sum of squares. SST measure the total variation of y_i which is the variation of y_i compared to the average value $\bar{y}$.

Analysis of variance

- $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = S_{yy}$ is the total sum of squares. SST measure the total variation of y_i which is the variation of y_i compared to the average value $\bar{y}$.
- $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 S_{xy}$ is the regression sum of squares. SSR Measure the variation of y_i resulted by different values of x.

Analysis of variance

- $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = S_{yy}$ is the total sum of squares. SST measure the total variation of y_i which is the variation of y_i compared to the average value $\bar{y}$.
- $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 S_{xy}$ is the regression sum of squares. SSR Measure the variation of y_i resulted by different values of x.
- $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the error sum of squares. SSE Measure the variation of y_i arisen by error aspects.

ANALYSIS OF VARIANCE

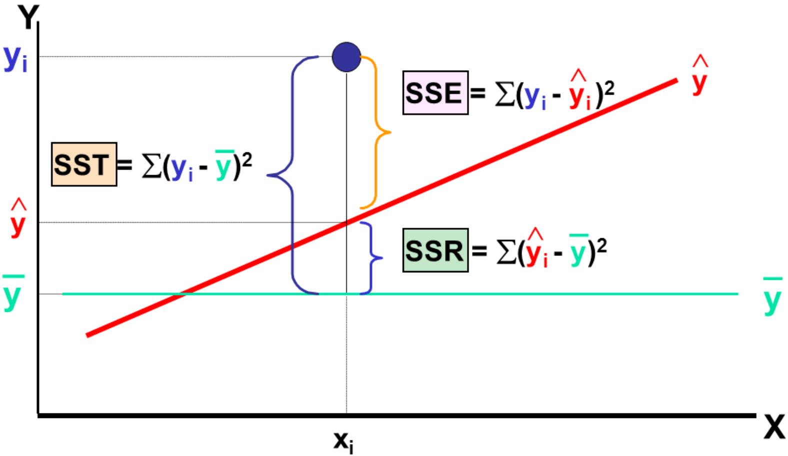
- $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = S_{yy}$ is the total sum of squares. SST measure the total variation of y_i which is the variation of y_i compared to the average value $\bar{y}$.
- $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 S_{xy}$ is the regression sum of squares. SSR Measure the variation of y_i resulted by different values of x.
- $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the error sum of squares. SSE Measure the variation of y_i arisen by error aspects.

Thus, we have a fundamental identity

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

$$SST = SSR + SSE$$

Estimating the variance



COEFFICIENT OF DETERMINATION

DEFINITION

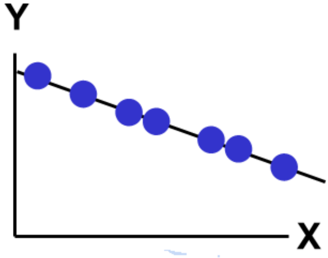
The coefficient of determination is the proportion of variation in the response variables that is explained by the different values of independent variable compared to the total variation. That is computed by

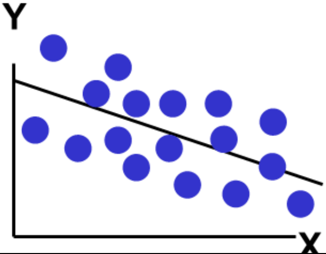
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{3}$$

Note that $0 \leq R^2 \leq 1$.

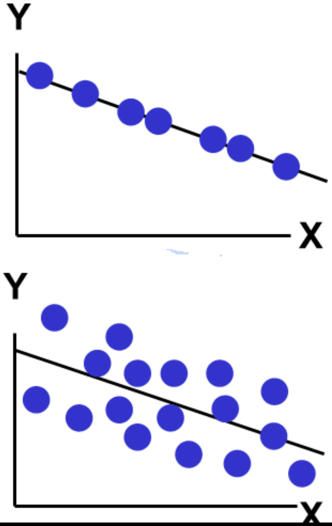
Coefficient of determination

Coefficient of determination





COEFFICIENT OF DETERMINATION



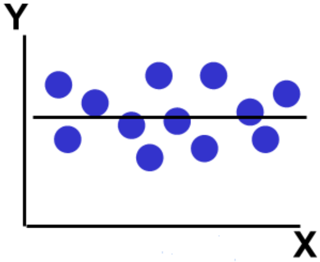
A value of R^2 near 1 indicates that most of the variation of the response data is explained by the different values of independent variable. In other word, a the linear regression model is explaining well the relationship between Y and x.

COEFFICIENT OF DETERMINATION

A value of R^2 near 0 indicates that little of the variation is explained by the different values of x or only a little portion of pair (Y_i, x_i) has linear correlation.

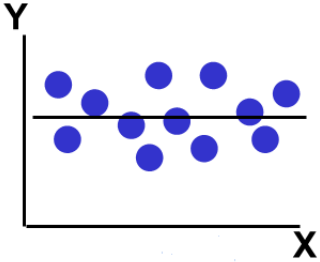
COEFFICIENT OF DETERMINATION

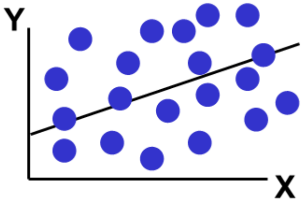
A value of R^2 near 0 indicates that little of the variation is explained by the different values of x or only a little portion of pair (Y_i, x_i) has linear correlation.



COEFFICIENT OF DETERMINATION

A value of R^2 near 0 indicates that little of the variation is explained by the different values of x or only a little portion of pair (Y_i, x_i) has linear correlation.





COEFFICIENT OF DETERMINATION

EXAMPLE 4

A new-car dealer is interested in the relationship between the number of salespeople working on a weekend and the number of cars sold. Data were gathered for six consecutive Sundays:

Number of salespeople	5	7	4	2	4	8
Number of cars sold	22	20	15	9	17	25

(A) Determine the estimated regression line.

(B) What is the coefficient of determination?

(C) How much of the variation in the number of automobiles sold is explained by the number of salespeople?

(D) Test the null hypothesis that the mean number of sales does not depend on the number of salespeople working.

Estimating the variance

Considering the simple linear model: $Y_i = \beta_0 + x_i\beta_1 + \varepsilon_i,\ i = 1, \ldots, n$

Estimating the variance

Considering the simple linear model: $Y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$, $i = 1, \ldots, n$
The ith error term $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

ESTIMATING THE VARIANCE

Considering the simple linear model: $Y_i = \beta_0 + x_i\beta_1 + \varepsilon_i,\ i = 1,\ldots,n$
The ith error term $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.
How would we estimate σ^2?

ESTIMATING THE VARIANCE

Considering the simple linear model: $Y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$, $i = 1, \ldots, n$
The ith error term $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.
How would we estimate σ^2?

THEOREM

The mean squares error (MSE) of a simple linear regression is defined by

$$MSE = \frac{SSE}{n-2}.$$

The mean squares error is an unbiased estimate of σ^2, that is

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2}$$

Proof:

ESTIMATING THE VARIANCE

Estimating the variance

- A more convenient computing formula for SSE is

$$SSE = SST - \hat{\beta}_1 S_{xy}.$$

- The standard error of $\hat{\sigma}^2$ is

$$SE(\hat{\sigma}^2) = \sqrt{\frac{SSE}{n-2}}$$

ESTIMATING THE VARIANCE

- A more convenient computing formula for SSE is

$$SSE = SST - \hat{\beta}_1 S_{xy}.$$

- The standard error of $\hat{\sigma}^2$ is

$$SE(\hat{\sigma}^2) = \sqrt{\frac{SSE}{n-2}}$$

- $SE(\hat{\sigma}^2)$ indicates the variation of the observed data y_i compared to the fitted linear regression line.

ESTIMATING THE VARIANCE

EXERCISE 2

The following data give, for certain years between 1982 and 2002, the percentages of British women who were cigarette smokers.

Year	1982	1984	1988	1990	1994	1996	1998	2000	2002
Percentage	33.1	31.8	30.4	24.3	26.3	27.7	26.3	25.3	24.8

Treat these data as coming from a linear regression model, with the input being the year and the response being the percentage. Take 1982 as the base year, so 1982 has input value $x = 0$, 1986 has input value $x = 4$, and so on.

(A) Estimate the value of σ^2.

(B) Predict the percentage of British women who smoked in 1997.

Hypothesis tests in Simple Linear Regression

Hypothesis tests in Simple Linear Regression

- Hypothesis tests of β_1 includes the following cases:

$$(a)\begin{cases} H_0 : \beta_1 = b_1 \\ H_1 : \beta_1 \neq b_1 \end{cases} \qquad (b)\begin{cases} H_0 : \beta_1 = b_1 \\ H_1 : \beta_1 < b_1 \end{cases} \qquad (c)\begin{cases} H_0 : \beta_1 = b_1 \\ H_1 : \beta_1 > b_1 \end{cases}$$

where b_1 and a confident level α are given.

HYPOTHESIS TESTS IN SIMPLE LINEAR REGRESSION

- Hypothesis tests of β_1 includes the following cases:

$$(a)\begin{cases} H_0:\beta_1=b_1\\ H_1:\beta_1\neq b_1\end{cases}\qquad (b)\begin{cases} H_0:\beta_1=b_1\\ H_1:\beta_1< b_1\end{cases}\qquad (c)\begin{cases} H_0:\beta_1=b_1\\ H_1:\beta_1> b_1\end{cases}$$

  where b_1 and a confident level α are given.
- An important hypothesis is $\beta_1=0$. Its importance lies in the fact that it is equivalent to stating that a response does not linearly depend on the value of the input; or, in other words, there is no regression on the input value.

HYPOTHESIS TESTS IN SIMPLE LINEAR REGRESSION

THEOREM

Let $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ be a simple linear regression model for a dataset of n independent observations where $\varepsilon_i \sim \mathcal{N}(0,1)$. Considering $\hat{\beta}_0$ and $\hat{\beta}_1$ are respectively the least-square estimates of β_0 and β_1, then

Hypothesis tests in Simple Linear Regression

Theorem

Let $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ be a simple linear regression model for a dataset of n independent observations where $\varepsilon_i \sim \mathcal{N}(0,1)$. Considering $\hat{\beta}_0$ and $\hat{\beta}_1$ are respectively the least-square estimates of β_0 and β_1, then

1. $\hat{\beta}_0$ and $\hat{\beta}_1$ follow normal distribution.

HYPOTHESIS TESTS IN SIMPLE LINEAR REGRESSION

THEOREM

Let $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ be a simple linear regression model for a dataset of n independent observations where $\varepsilon_i \sim \mathcal{N}(0,1)$. Considering $\hat{\beta}_0$ and $\hat{\beta}_1$ are respectively the least-square estimates of β_0 and β_1, then

1. $\hat{\beta}_0$ and $\hat{\beta}_1$ follow normal distribution.
2. The expectation and variance of $\hat{\beta}_0$ and $\hat{\beta}_1$ are respectively

$$\mathbb{E}(\hat{\beta}_0) = \beta_0, \ \mathbb{V}ar(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2, \tag{4}$$

$$\mathbb{E}(\hat{\beta}_1) = \beta_1, \ \mathbb{V}ar(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \tag{5}$$

Proof:

Hypothesis tests in Simple Linear Regression

A hypothesis test of β_1 follows steps belows:

Hypothesis tests in Simple Linear Regression

A hypothesis test of β_1 follows steps belows:

1. State the hypotheses H_0 and H_1.

Hypothesis tests in Simple Linear Regression

A hypothesis test of β_1 follows steps belows:

1. State the hypotheses H_0 and H_1.
2. State the confident level α.

Hypothesis tests in Simple Linear Regression

A hypothesis test of β_1 follows steps belows:

1. State the hypotheses H_0 and H_1.
2. State the confident level α.
3. Compute the test statistic:

$$T_{\beta_1} = \frac{\hat{\beta}_1 - b_1}{SE(\hat{\beta}_1)} \sim t(n-2)$$

where

$$SE(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

Hypothesis tests in Simple Linear Regression

4. Determine the rejected range or compute p-value:

Alternative hypothesis	**rejected range**	p **- value**
$H_1 : \beta_1 \neq b_1$	$\lvert t_{\beta_1} \rvert > t^{n-2}_{\alpha/2}$	$p = 2\mathbb{P}(T_{n-2} \geq \lvert t_{\beta_0} \rvert)$
$H_1 : \beta_1 < b_1$	$t_{\beta_1} < -t^{n-2}_{\alpha}$	$p = \mathbb{P}(T_{n-2} \leq t_{\beta_0})$
$H_1 : \beta_1 > b_1$	$t_{\beta_1} > t^{n-2}_{\alpha}$	$p = \mathbb{P}(T_{n-2} \geq t_{\beta_0})$

5. Conclude whether H_0 is rejected or not.

Hypothesis tests in Simple Linear Regression

A hypothesis test of β_0 follows steps belows:

Hypothesis tests in Simple Linear Regression

A hypothesis test of β_0 follows steps belows:

1. State the hypotheses H_0 and H_1.

Hypothesis tests in Simple Linear Regression

A hypothesis test of β_0 follows steps belows:

1. State the hypotheses H_0 and H_1.
2. State the confident level α.

Hypothesis tests in Simple Linear Regression

A hypothesis test of β_0 follows steps belows:

1. State the hypotheses H_0 and H_1.
2. State the confident level α.
3. Compute the test statistic:

$$T_{\beta_0} = \frac{\hat{\beta}_0 - b_0}{SE(\hat{\beta}_0)} \sim t(n-2)$$

where

$$SE(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

Hypothesis tests in Simple Linear Regression

4. Determine the rejected range or compute p-value:

Alternative hypothesis	**rejected range**	p **- value**
$H_1 : \beta_0 \neq b_0$	$\lvert t_{\beta_0} \rvert > t^{n-2}_{\alpha/2}$	$p = 2\mathbb{P}(T_{n-2} \geq \lvert t_{\beta_0} \rvert)$
$H_1 : \beta_0 < b_0$	$t_{\beta_0} < -t^{n-2}_{\alpha}$	$p = \mathbb{P}(T_{n-2} \leq t_{\beta_0})$
$H_1 : \beta_0 > b_1$	$t_{\beta_0} > t^{n-2}_{\alpha}$	$p = \mathbb{P}(T_{n-2} \geq t_{\beta_0})$

5. Conclude whether H_0 is rejected or not.

CONFIDENCE INTERVALS ON PARAMETERS

THEOREM

CONFIDENCE INTERVALS ON PARAMETERS

THEOREM

Under the assumption that the observations are normally and independently distributed, a $100(1-\alpha)\%$ confidence interval on the slope β_1 in simple linear regression is

$$\hat{\beta}_1 - t_{\alpha/2}^{n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2}^{n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \tag{6}$$

Confidence intervals on parameters

Theorem

Under the assumption that the observations are normally and independently distributed, a $100(1-\alpha)\%$ confidence interval on the slope β_1 in simple linear regression is

$$\hat{\beta}_1 - t_{\alpha/2}^{n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2}^{n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \tag{6}$$

Similarly, a $100(1-\alpha)\%$ confidence interval on the intercept β_0 is

$$\hat{\beta}_0 - t_{\alpha/2}^{n-2}\sqrt{\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)\hat{\sigma}^2} \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2}^{n-2}\sqrt{\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)\hat{\sigma}^2} \tag{7}$$

CONFIDENCE INTERVALS ON PARAMETERS

EXERCISE 3

The following table relates the number of sunspots that appeared each year from 1970 to 1980 to the number of automobile accident deaths during that year. The data for automobile accident deaths are in units of 1000 deaths.

Year	Sunspots	Automobile deaths
70	165	54.6
71	89	53.3
72	55	56.3
73	34	49.6
74	9	47.1
75	30	45.9
76	59	48.5
77	83	50.1
78	109	52.4
79	127	52.5
80	153	53.2

Test the hypothesis that the number of automobile accident deaths is not linearly related to the number of sunspots. Use the 5 percent level of significance.

SAMPLE CORRELATION COEFFICIENT

DEFINITION

Considering a sample of n observations: $(X_i, Y_i), i = 1, \ldots, n$. The sample correlation coefficient r_{XY}, is defined by

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX}SST}} \quad (8)$$

SAMPLE CORRELATION COEFFICIENT

Note that

$$\hat{\beta}_1 = \sqrt{\frac{SST}{S_{XX}}} r_{XY}$$

thus

$$r_{XY}^2 = \hat{\beta}_1^{\,2} \frac{S_{XX}}{SST} = \hat{\beta}_1 \frac{S_{XY}}{SST} = \frac{SSR}{SST}$$

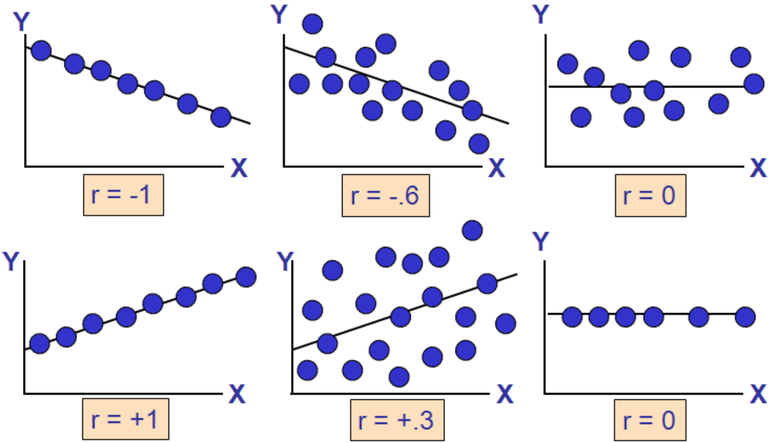
• The coefficient of determination R^2 in a simple linear regression model equals to the square of the sample correlation coefficient.

$$R^2 = r_{XY}^2$$

Sample correlation coefficient

- The range of r_{XY}: $-1 \leq r_{XY} \leq 1$,
- $-1 \leq r_{XY} < 0$: negative correlation. r_{XY} is closer to -1 indicating a stronger negative correlation between X and Y.
- $0 < r_{XY} \leq 1$: positive correlation. r_{XY} is closer to 1 indicating a stronger positive correlation between X and Y.
- r_{XY} is closer to 0 indicating a weak correlation between X and Y. $r_{XY} = 0$: indicating linearly independent between X and Y.

Sample correlation coefficient



Analysis of residuals: assessing the model.

- Analysis of Residuals is used to assess the assumptions of simple linear regression models.

ANALYSIS OF RESIDUALS: ASSESSING THE MODEL.

- Analysis of Residuals is used to assess the assumptions of simple linear regression models.
- The assumptions of simple linear regression models:

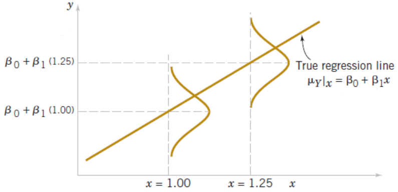
Analysis of residuals: assessing the model.

- Analysis of Residuals is used to assess the assumptions of simple linear regression models.
- The assumptions of simple linear regression models:
  - The linear relationship of Y and x: $Y = \beta_0 + \beta_1 x + \epsilon$ where β_0 and β_1 are the regression coefficients such that given x we have $\mathbb{E}(Y|x) = \beta_0 + \beta_1 x$.

Analysis of residuals: assessing the model.

- Analysis of Residuals is used to assess the assumptions of simple linear regression models.

- The assumptions of simple linear regression models:
  - The linear relationship of Y and x: $Y = \beta_0 + \beta_1 x + \epsilon$ where β_0 and β_1 are the regression coefficients such that given x we have $\mathbb{E}(Y|x) = \beta_0 + \beta_1 x$.
  - Constant variation: The variance σ^2 of Y is invariant for all value of x, e.i. $Var(Y|x) = \sigma^2$.



Analysis of residuals: assessing the model.

ANALYSIS OF RESIDUALS: ASSESSING THE MODEL.

- Normal distribution: $Y|x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$.

Analysis of residuals: assessing the model.

- Normal distribution: $Y|x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$.
- Independence: the observations of Y are independent.

ANALYSIS OF RESIDUALS: ASSESSING THE MODEL.

- Normal distribution: $Y|x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$.
- Independence: the observations of Y are independent.

$\Rightarrow$ to test the normality, we use the Normal probability plot (Q-Q plot) of the residuals or the standardized residuals.

ANALYSIS OF RESIDUALS: ASSESSING THE MODEL.

- Normal distribution: $Y|x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$.
- Independence: the observations of Y are independent.

$\Rightarrow$ to test the normality, we use the Normal probability plot (Q-Q plot) of the residuals or the standardized residuals.

$\Rightarrow$ to test the linearity, independence, and constant variances we use the scatter plot of the residuals or the standardized residuals.

Analysis of residuals: assessing the model.

The QQ-plot

Analysis of residuals: assessing the model.

The QQ-plot

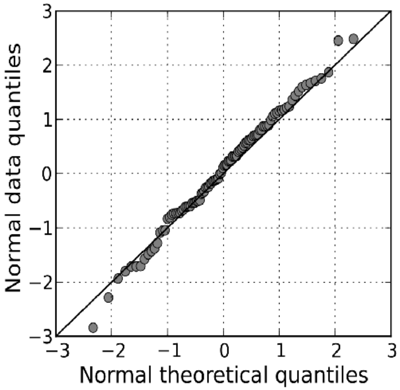
- A Q–Q plot is a plot of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions.

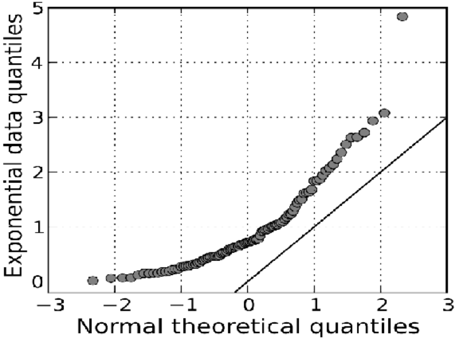
Analysis of residuals: assessing the model.

The QQ-plot

- A Q–Q plot is a plot of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions.
- The points plotted in a Q–Q plot are always non-decreasing when viewed from left to right. If the two distributions being compared are identical, the Q–Q plot follows the 45 deg line $y = x$

Analysis of residuals: assessing the model.





ANALYSIS OF RESIDUALS: ASSESSING THE MODEL.

STANDARDIZED RESIDUALS

The standardized residuals are defined as

$$E_i = \frac{Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)}{\sqrt{SSE/(n-2)}},\ i = 1, 2, \ldots, n$$

Analysis of residuals: assessing the model.

Standardized residuals

The standardized residuals are defined as

$$E_i = \frac{Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)}{\sqrt{SSE/(n-2)}}, \; i = 1, 2, \ldots, n$$

When the simple linear regression model is correct, the standardized residuals are approximately independent standard normal random variables. Thus,

Analysis of residuals: assessing the model.

Standardized residuals

The standardized residuals are defined as

$$E_i = \frac{Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)}{\sqrt{SSE/(n-2)}}, \ i = 1, 2, \ldots, n$$

When the simple linear regression model is correct, the standardized residuals are approximately independent standard normal random variables. Thus,

- they should be randomly distributed about 0 with about 95 percent of their values being between -2 and $+2$ (since $P(-1.96 < Z < 1.96) = 0.95$);

ANALYSIS OF RESIDUALS: ASSESSING THE MODEL.

STANDARDIZED RESIDUALS

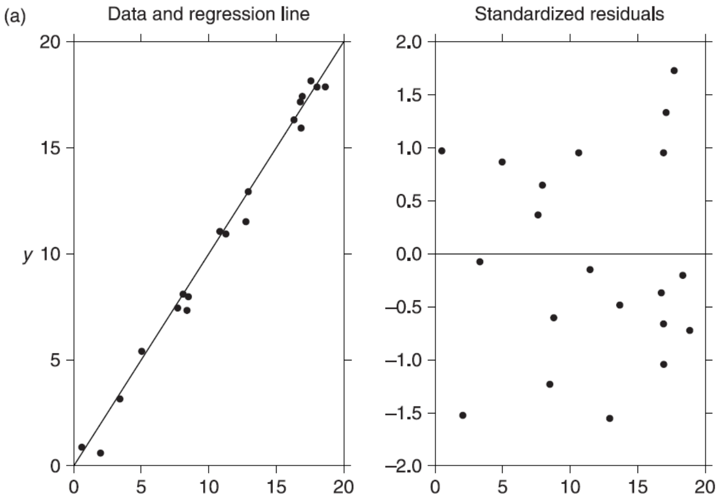
The standardized residuals are defined as

$$E_i = \frac{Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)}{\sqrt{SSE/(n-2)}}, \ i = 1, 2, \ldots, n$$

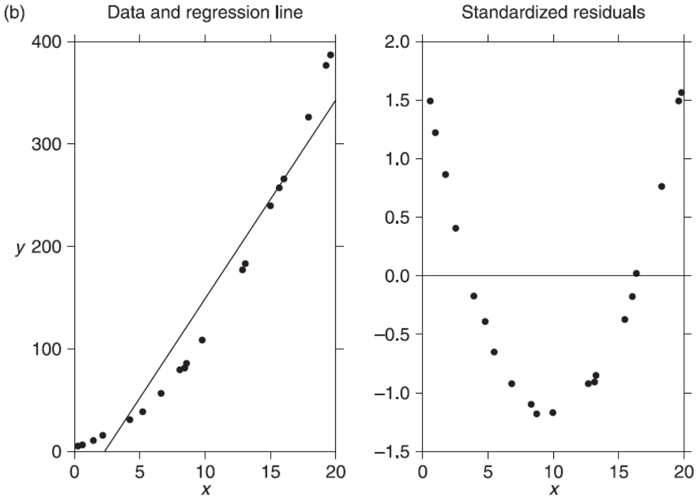
When the simple linear regression model is correct, the standardized residuals are approximately independent standard normal random variables. Thus,

- they should be randomly distributed about 0 with about 95 percent of their values being between -2 and $+2$ (since $P(-1.96 < Z < 1.96) = 0.95$));
- their scatter plot should not indicate any distinct pattern.

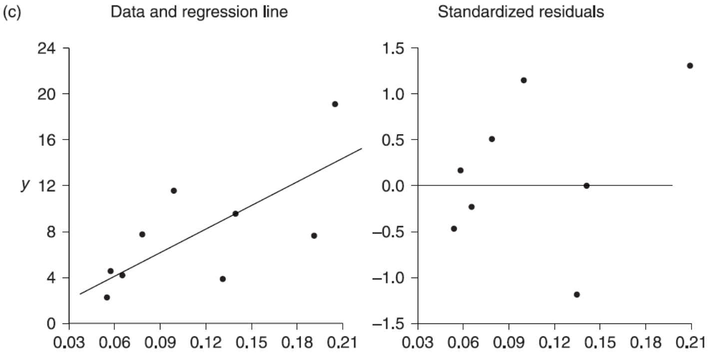
Analysis of residuals: assessing the model.



Analysis of residuals: assessing the model.



Analysis of residuals: assessing the model.



ABUSES OF REGRESSION

Regression is widely used and frequently misused; we mention several common abuses of regression briefly here

ABUSES OF REGRESSION

Regression is widely used and frequently misused; we mention several common abuses of regression briefly here

- Regression relationships are valid for values of the regression variable only within the range of the original data. $\Rightarrow$ be careful with extrapolates.

Abuses of regression

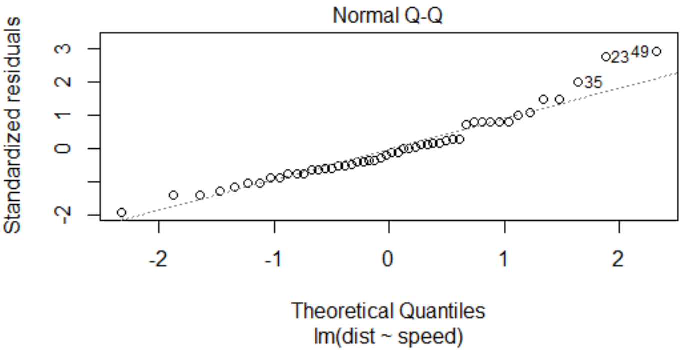
Regression is widely used and frequently misused; we mention several common abuses of regression briefly here

- Regression relationships are valid for values of the regression variable only within the range of the original data. ⇒ be careful with extrapolates.
- It's hard to define what level of R^2 is appropriate to claim the model fits well. Essentially, it will vary with the application and the domain studied.

Interpreting R results

```
> M<- lm(dist ~ speed, data = cars)
> summary(M)

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,	Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

INTERPRETING R RESULTS



INTERPRETING R RESULTS