Mahidol University
Faculty of Medicine Ramathibodi Hospital
Section for Clinical Epidemiology and Biostatistics

Correlation and Simple Linear Regression

Sasivimol Rattanasiri, Ph.D

Section for Clinical Epidemiology and Biostatistics

Ramathibodi Hospital, Mahidol University

E-mail: sasivimol.rat@mahidol.ac.th

Outline

- Correlation analysis
  - Estimation of correlation coefficient
  - Hypothesis testing
- Simple linear regression analysis

Correlation analysis

- Estimation of correlation coefficient
  - Pearson's correlation coefficient.
  - Spearman's correlation coefficient.
- Hypothesis testing

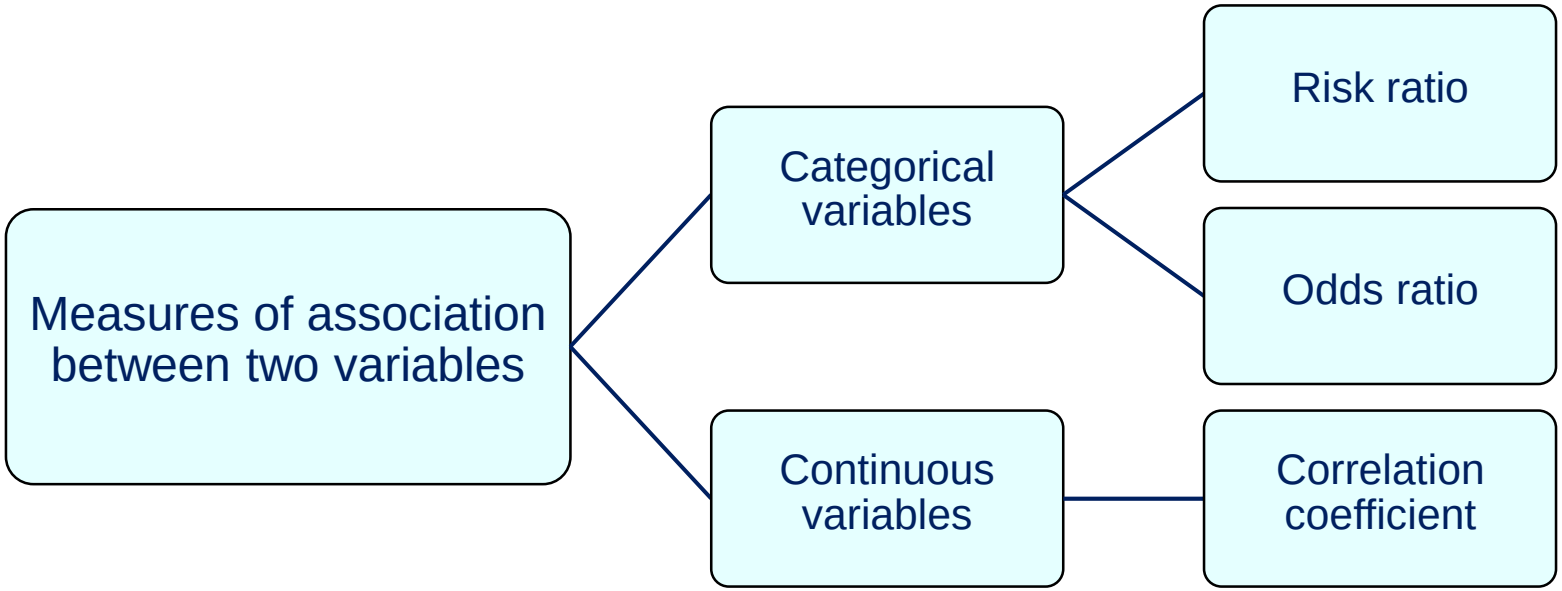


Figure 1. Flow chart for measures of the strength of association

For examples

- The correlation between age and percentage of body fat.
- The correlation between cholesterol level and systolic blood pressure (SBP).
- The correlation between age and bone mineral density (BMD).

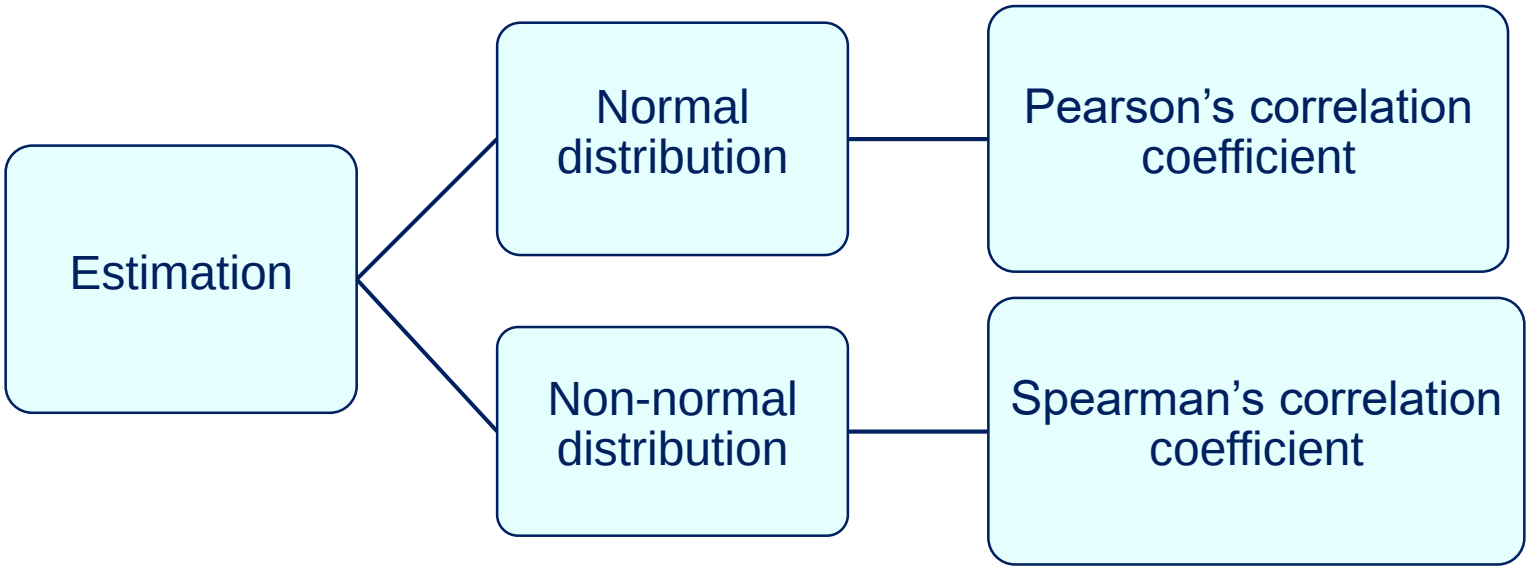


Figure 2. Flow chart for estimation of correlation coefficient based upon the distribution of data

Pearson's correlation coefficient

The sample correlation coefficient (r) can be defined as:

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

where,

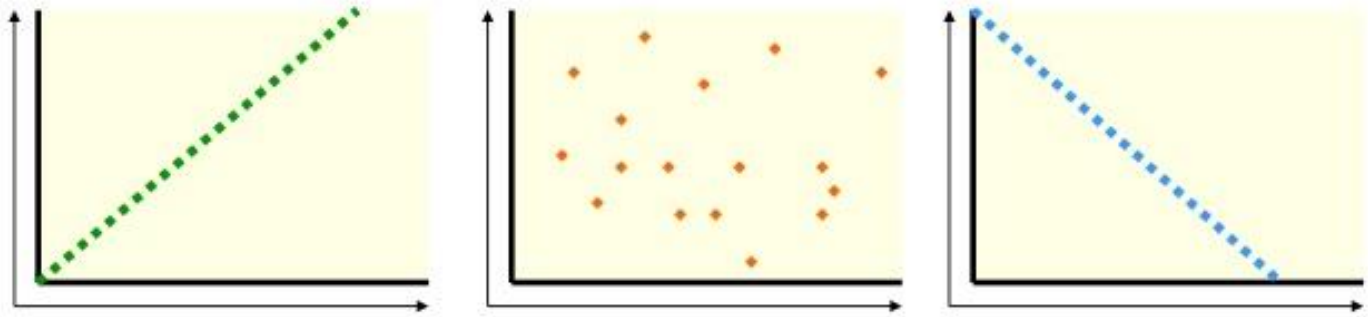
x_i and y_i are observations of the variables X and Y,

$\bar{x}$ and $\bar{y}$ are sample means of the variables X and Y.

Interpretation

- Correlation coefficient lies between -1 to +1.
- If correlation coefficient is equal -1 or +1, it indicates that there is perfect linear association between two continuous variables.
- If the correlation coefficient is near 0, it indicates that there is no linear association between two continuous variables. However, a nonlinear relationship may exist.

Correlation



Perfect positive correlation (+1.00)

No relationship (0.00)

Perfect negative correlation (-1.00)

Scatterplots, showing patterns of correlations

Interpretation

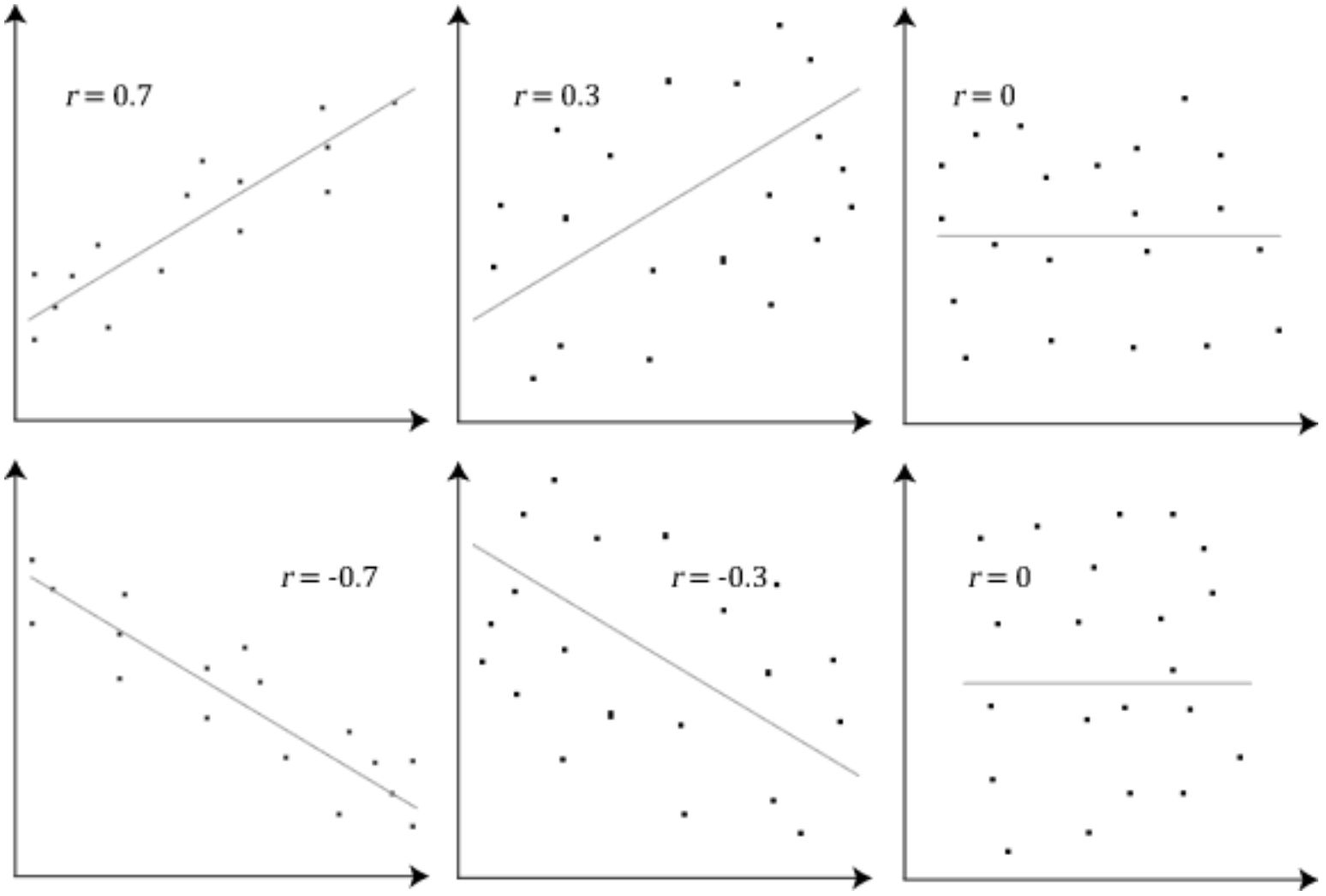
- If r>0, the correlation between two continuous variables is positive.
- It means that if as a value of one variable increases, a related value of the another variable increases, whereas as a value of one variable decreases, a related value of the another variable decreases.

Interpretation

- If $r<0$, the correlation between two continuous variables is negative.
- It means that if as a value of one variable increases, a related value of the another variable decreases, whereas as a value of one variable decreases, a related value of the another variable increases.

Interpretation

- Guideline for correlation coefficient:
  - If it is near -0.3 or +0.3, it indicates that there is weak linear association.
  - If it is near -0.5 or +0.5, it indicates that there is moderate linear association.
  - If it is near -0.7 or +0.7, it indicates that there is strong linear association.


r = 0.7
r = 0.3
r = 0
r = -0.7
r = -0.3
r = 0

Limitations of the sample Pearson's correlation coefficient

- The correlation coefficient is the measure of the strength of the *linear* relationship between two continuous variables.
- The Pearson's correlation coefficient is not a valid measure, if they have a nonlinear relationship.

Class example I

- Researcher wanted to assess the correlation between BMI and SBP in 30 subjects.

ID	BMI (x)	SBP (y)	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
1	18.63	102	-3.82	-17	64.94	14.59	289
2	20.97	110	-1.48	-9	13.32	2.19	81
3	25.88	120	3.43	1	3.43	11.76	1
4	24.51	106	2.06	-13	-26.78	4.24	169
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
30	20.43	98	-2.02	-21	42.42	4.08	441
	$\bar{x}$ =22.45	$\bar{y}$ = 119			591.94	269.87	8646

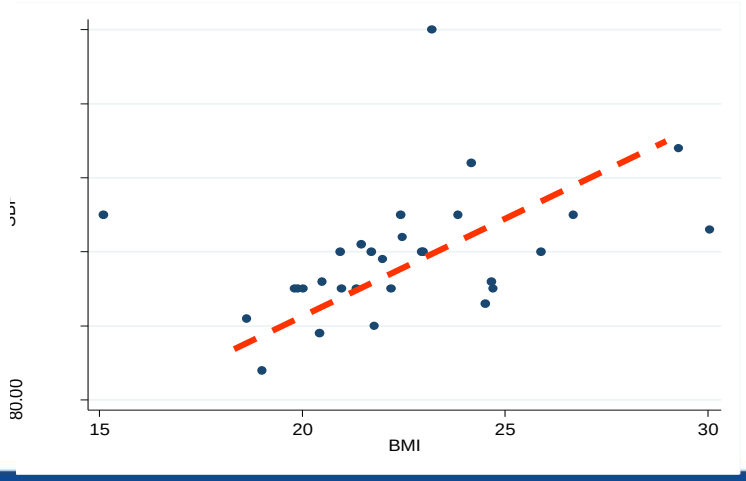
The Pearson's correlation coefficient can be defined as:

$$r = \frac{\sum_{i=1}^{30}(x_i - 22.45)(y_i - 119.0)}{\sum_{i=1}^{30}\sqrt{(x_i - 22.45)^2(y_i - 119.0)^2}}$$

$$= 0.3875$$

Interpretation

- The correlation coefficient is equal to 0.3875.
- This means that there is linear correlation between BMI and SBP.
- This correlation is positive, that is SBP tends to be higher for higher BMI which is confirming the visual impression.



Hypothesis testing

The hypothesis testing for the correlation coefficient is performed under the null hypothesis that there is no linear correlation between two variables.

$$H_0 : \rho = 0$$

Statistical test

Statistical test of hypothesis testing for the correlation coefficient can be defined as

$$t = \frac{r\sqrt{n-2}}{1-r^2}$$

This test has the *t* distribution with n-2 degrees of freedom.

Steps for Hypothesis Testing

1. Generate the null and alternative hypothesis

Null hypothesis:

H_0: There is no linear correlation between BMI and SBP,

or

H_0: The correlation coefficient is equal to zero.

Steps for Hypothesis Testing

1. Generate the null and alternative hypothesis

Alternative hypothesis:

H_a: There is linear correlation between BMI and SBP,

or

H_a: The correlation coefficient is not equal to zero.

Steps for Hypothesis Testing

2. Select appropriate statistical test

```
. pwcorr bmi sbp1,obs sig

             |      bmi      sbp1
-------------+------------------
         bmi |   1.0000
             |
             |       30
             |
        sbp1 |   0.3875   1.0000
             |   0.0344
             |       30        30
```

Correlation coefficient

P value

Steps for Hypothesis Testing

3. Draw a conclusion

- The *p* value for this example is 0.0344 which is less than the level of significance.
- Thus, we reject the null hypothesis and conclude that there is linear correlation between BMI and SBP.

Spearman's rank correlation coefficient

- If normality assumption of Pearson's correlation is violated, Spearman's rank correlation will be applied.
- Spearman's correlation is a statistical measure of strength of monotonic relationship between two variables. It is a nonparametric statistic.
- It is used to assess the correlation between two variables when either or both of the variables do not have a normal distribution.

Monotonic function

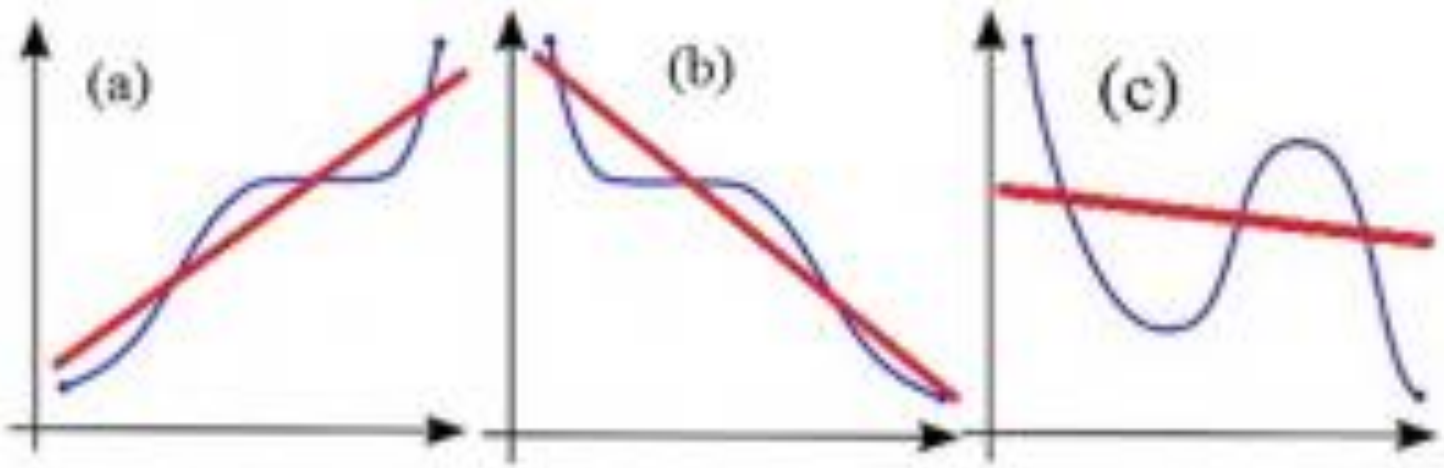


Figure 3. (a) is monotonically increasing, (b) is monotonically decreasing, and (c) is non-monotonic function

Spearman's rank correlation coefficient

Spearman's rank correlation coefficient can be calculated as Pearson's correlation coefficient for the ranked values of two variables

$$r_s = \frac{\sum_{i=1}^{n}(x_{ri} - \bar{x}_r)(y_{ri} - \bar{y}_r)}{\sqrt{\sum_{i=1}^{n}(x_{ri} - \bar{x}_r)^2 \sum_{i=1}^{n}(y_{ri} - \bar{y}_r)^2}}$$

Class example II

- Researchers wanted to assess the correlation between age and amount of calcium intake in 80 adults.
- The calcium intake data did not have a normal distribution.

Table 1. Rank data for assessing the correlation between age and percentage of fat

Subject	Age	Rank_age	% Fat	Rank_fat
1	23	1.5	9.5	2
2	23	1.5	27.9	7
3	27	3.5	7.8	1
4	27	3.5	17.8	3
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
15	58	15.5	33.0	13
16	58	15.5	33.8	14
17	60	17	41.1	17
18	61	18	34.5	15

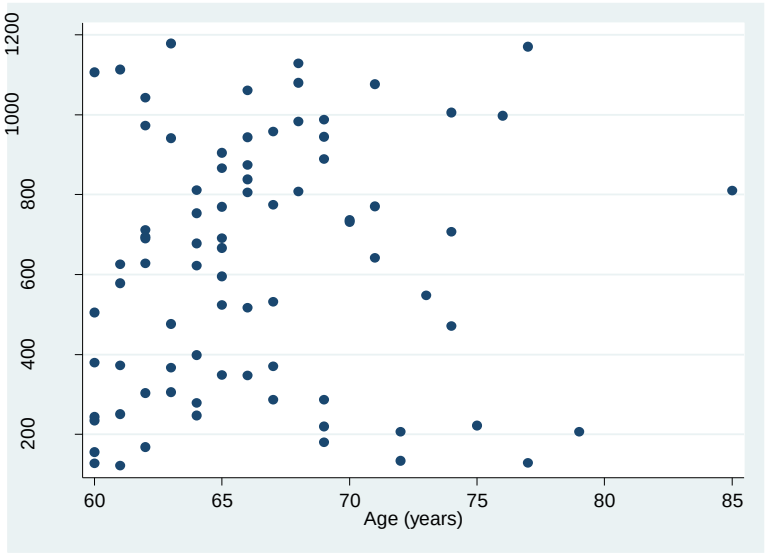
Spearman's rank correlation coefficient can be calculated as:

$$r_s = \frac{\sum_{i=1}^{80}(x_{ri} - 40.5)(y_{ri} - 40.5)}{\sqrt{\sum_{i=1}^{80}(x_{ri} - 40.5)^2 \sum_{i=1}^{80}(y_{ri} - 40.5)^2}}$$

$$= 0.17$$

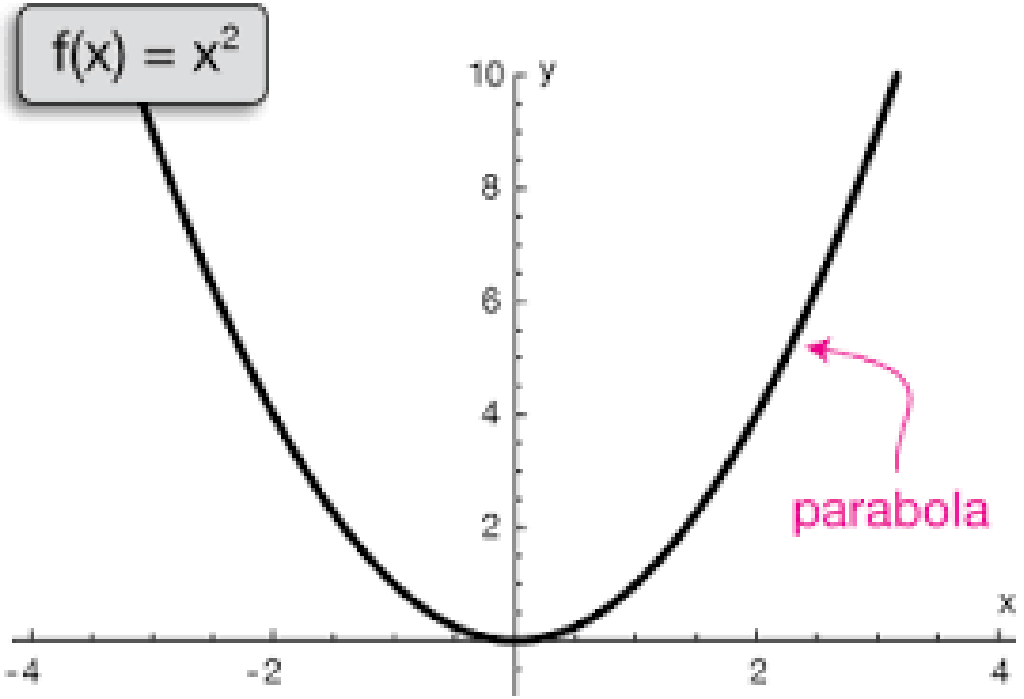
Interpretation

- The Spearman's rank correlation coefficient between age and amount of calcium intake is equal to 0.17.
- These results suggest a weak correlation between age and calcium intake.



Advantages of the Spearman's rank correlation coefficient

- It is appropriate for both continuous and discrete variables, including ordinal variable.
- There is no requirement of normality.
- If Spearman's rank correlation is equal to zero, it does not imply there is no relationship between two variables.



There is a perfect quadratic relationship between two variables, the Spearman's rank correlation coefficient is equal to zero.

Hypothesis testing

The hypothesis testing for the Spearman's correlation coefficient is performed under the null hypothesis that there is no correlation between two variables.

$$H_0 : \rho_s = 0$$

Statistical test

The statistical test for hypothesis testing of the Spearman's correlation coefficient can be defined as:

$$t_s = \frac{r_s \sqrt{n-2}}{1-r_s^2}$$

Steps for Hypothesis Testing

1. Generate the null and alternative hypothesis

Null hypothesis:

H_0: There is no correlation between age and calcium intake,

or

H_0: The correlation coefficient is equal to zero.

Steps for Hypothesis Testing

1. Generate the null and alternative hypothesis

Alternative hypothesis:

H_a: There is correlation between age and calcium intake,

or

H_a: The correlation coefficient is not equal to zero.

Steps for Hypothesis Testing

2. Select appropriate statistical test

```
. spearman ca_intake age
 Number of obs =       80
Spearman's rho =       0.1655
Test of Ho: caintake and age are independent
    Prob > |t| =       0.1423
```

Steps for Hypothesis Testing

3. Draw a conclusion

- The spearman's correlation coefficient is 0.17. These results suggest a weak. correlation between age and calcium intake.
- The *p* value is equal to 0.14, so we fail to reject the null hypothesis and conclude that there is non-monotonic correlation between age and calcium intake.

Simple linear regression analysis

I. Simple linear regression model

II. Fitting linear regression model

III. Coefficient of determination (r^2)

IV. Assumption checking

V. Estimation of mean predicted values

VI. Estimation of individual predicted values

Main objectives of applying regression analysis

Regression analysis is the statistical method which is used to

- ✓ Study the relationship between two or more variables.
- ✓ Predict the unobserved values from the relationship that we know.

When the linear regression should be applied?

Linear regression analysis is the statistical method which is used to predict outcome from other predictors.

- Outcome: only continuous variable.
- Predictors: either continuous or categorical variables.

When the linear regression should be applied?

- Outcome of interest is measured once per subject.
- Study design can be any observational studies (case-control, cohort, cross-sectional) or RCT.

Example

- Researchers may be interested in predicting the change in SBP which is related to a given change in BMI.
- Clearly, correlation analysis does not carry out this purpose; it just indicates the strength of the correlation as a single number.

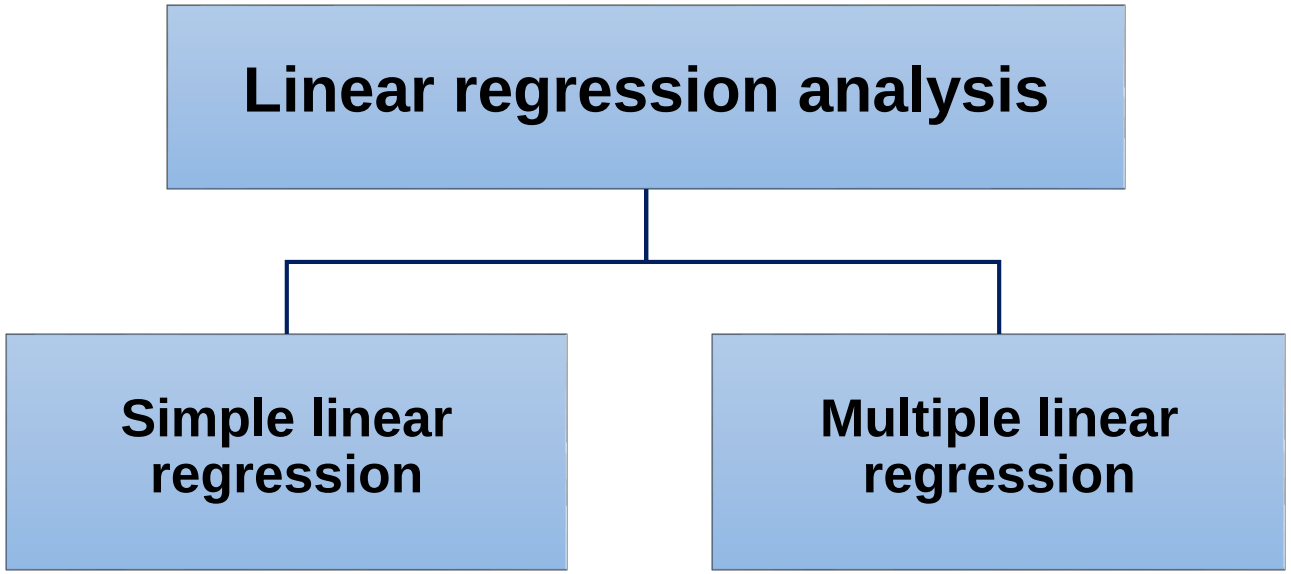


Figure 4. Types of linear regression analysis

Simple linear regression

Researchers would like to predict the change of SBP which is related to a given change in BMI.

- The outcome is SBP.
- The predictor is BMI.

Multiple linear regression

Researchers wanted to predict the change of SBP which is related to a given change in BMI and age.

- Outcome is SBP.
- First predictor is BMI.
- Second predictor is age.

Simple linear regression analysis

I. **Simple linear regression model**

II. Fitting linear regression model

III. Coefficient of determination (r^2)

IV. Assumption checking

V. Estimation of mean predicted values

VI. Estimation of individual predicted values

I. Linear regression model

The full linear regression model for the population can take the following form:

$$Y = \alpha + \beta X + \varepsilon$$

where,

- **ε** refers to error or residual,
- **α** and β are regression coefficients,
- **α** refers to intercept and
- **β** refers to slope of the regression line .

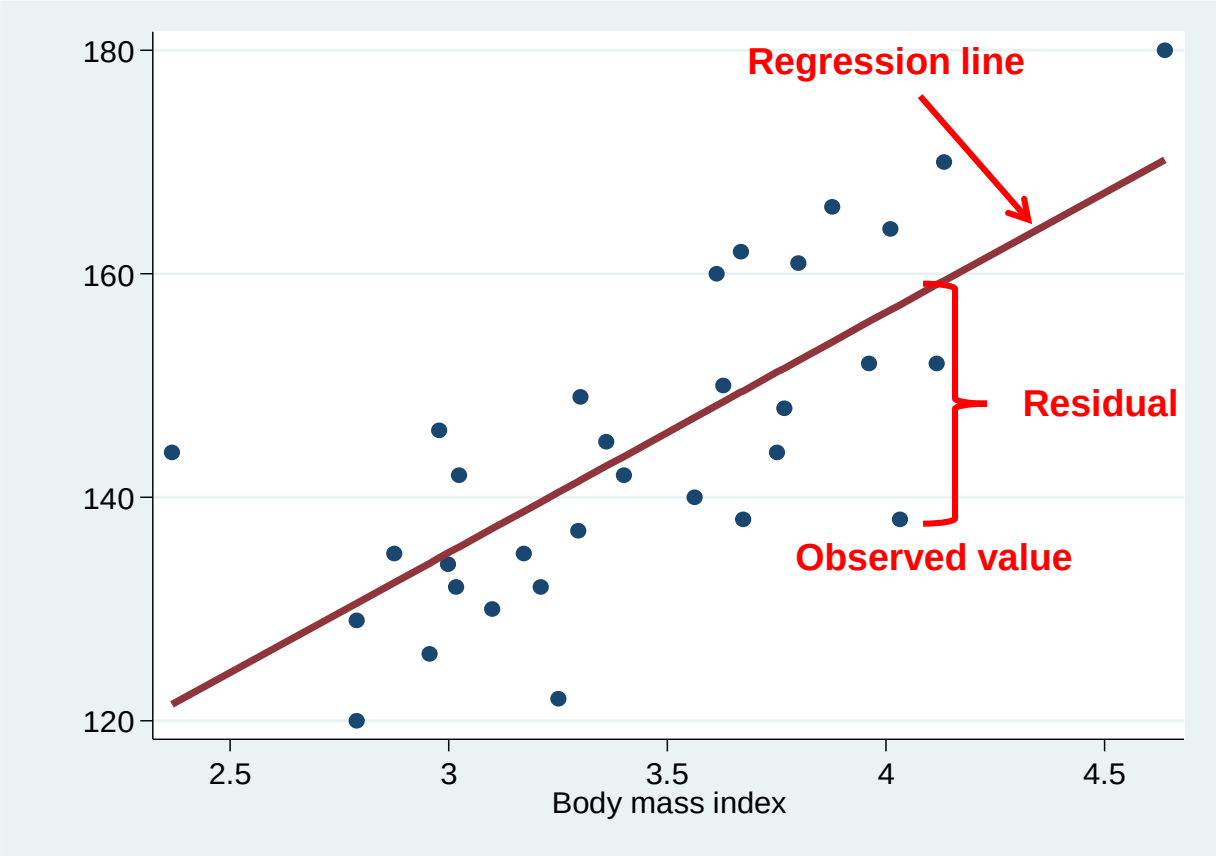


Figure 5. Scatter plot of relationship between SBP and BMI with regression line

I. Linear regression model

The sample linear regression model is defined as:

$$\hat{y}_i = a + bx_i + e_i$$

where,

- y_i is the observed values of outcome variable.
- $\hat{y}_i$ is the predicted value of y_i for a particular value of x_i
- $e_i = y_i - \hat{y}_i$ refer to residual
- **a** and **b** refer to regression coefficient

I. Linear regression model

$$\hat{y}_i = a + bx_i + e_i$$

- The y-intercept is defined as the mean value of dependent variable Y when independent variable X is equal to zero.
- The y-intercept has no practical meaning because the independent variable cannot be anywhere near zero, for example, blood pressure, weight, or height.

I. Linear regression model

$$\hat{y}_i = a + bx_i + e_i$$

- The slope is interpreted as the change in dependent variable Y which is related to a one unit change in independent variable X.
- The y-intercept and slope are called regression coefficients which need to be estimated.

Simple linear regression analysis

I. Simple linear regression model

II. Fitting linear regression model

III. Coefficient of determination (r^2)

IV. Assumption checking

V. Estimation of mean predicted values

VI. Estimation of individual predicted values

II. Fitting linear regression model

- A method of least squares is used to determine the best fitting straight line to a set of data.
- This method produces the line that minimizes the distance between the observed and predicted values.

Method of least squares

The residual can be defined as:

$$e_i = y_i - \hat{y}_i$$

To find the line that best fits to the set of data, we minimize the sum of the squares of the residuals, which can be defined as:

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

Method of least squares

The slope (b) in the simple linear regression line which gives minimum residual sum of squares, can be defined as:

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

$$= \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

$$\therefore \quad b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Method of least squares

From the regression model:

$$y = a + bx$$

The intercept (a) in the regression line which gives minimum residual sum of squares can be defined as:

$$a = \bar{y} - b\bar{x}$$

Class example III

Researchers wanted to predict the change in SBP which is related to a given change in BMI.

Fitting simple linear regression by STATA

```
. regress sbp1 bmi

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(1, 28)        =      4.95
       Model |  1298.35408         1  1298.35408   Prob > F        =    0.0344
    Residual |  7347.64592        28  262.415926   R-squared       =    0.1502
-------------+----------------------------------   Adj R-squared   =    0.1198
       Total |        8646        29  298.137931   Root MSE        =    16.199

------------------------------------------------------------------------------
        sbp1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   2.193388    .986084     2.22   0.034     .1734861    4.213289
       _cons |   69.75698   22.33493     3.12   0.004     24.00596     115.508
------------------------------------------------------------------------------
```

Simple linear regression model

The linear regression model for the prediction of the SBP from the BMI can be defined as:

$$\hat{\text{SBP}} = 69.76 + 2.19 \times (\text{BMI})$$

Interpretation

$$\hat{SBP} = 69.76 + 2.19 \times (BMI)$$

- The Y-intercept of the regression line is 69.76, implying that the mean of SBP is equal to 69.76 when the BMI is equal to zero.
- The slope of the regression line is 2.19, implying that for each one-unit increase in BMI, the SBP increases by 2.19 mmHg on average.

CI of regression coefficients

- The uncertainty of the regression coefficients is shown by the CI of the population regression coefficients.
- The *t* distribution is used to estimate the confidence intervals for the regression coefficients.

CI for population intercept

$$a \pm (t_{1-\alpha/2} \times \hat{se}(a))$$

where, $$\hat{se}(a) = s_{y|x} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}}$$

and, $$s_{y|x} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}}$$

CI for population slope

$$b \pm (t_{1-\alpha/2} \times \hat{se}(b))$$

where, $$\hat{se}(b) = \frac{s_{y|x}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

and, $$s_{y|x} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

Fitting simple linear regression by STATA

```
. regress sbp1 bmi

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(1, 28)        =      4.95
       Model |  1298.35408         1  1298.35408   Prob > F        =    0.0344
    Residual |  7347.64592        28  262.415926   R-squared       =    0.1502
-------------+----------------------------------   Adj R-squared   =    0.1198
       Total |        8646        29  298.137931   Root MSE        =    16.199

------------------------------------------------------------------------------
        sbp1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   2.193388    .986084     2.22   0.034     .1734861    4.213289
       _cons |   69.75698   22.33493     3.12   0.004     24.00596     115.508
------------------------------------------------------------------------------
```

Interpretation

- The 95% CI of the y-intercept lies between 24.01 to 115.51.
- The 95% CI of the slope lies between 0.17 to 4.21. It means that for each one-unit increase in BMI, an SBP increases lies between 0.17 to 4.21 mmHg.

Tests of hypotheses on regression coefficients

•The null hypothesis about the slope of regression line can be defined as:

$$H_0 : \beta = 0$$

• If the population slope is equal to zero, it means that there is no linear relationship between predictor and outcome variables.

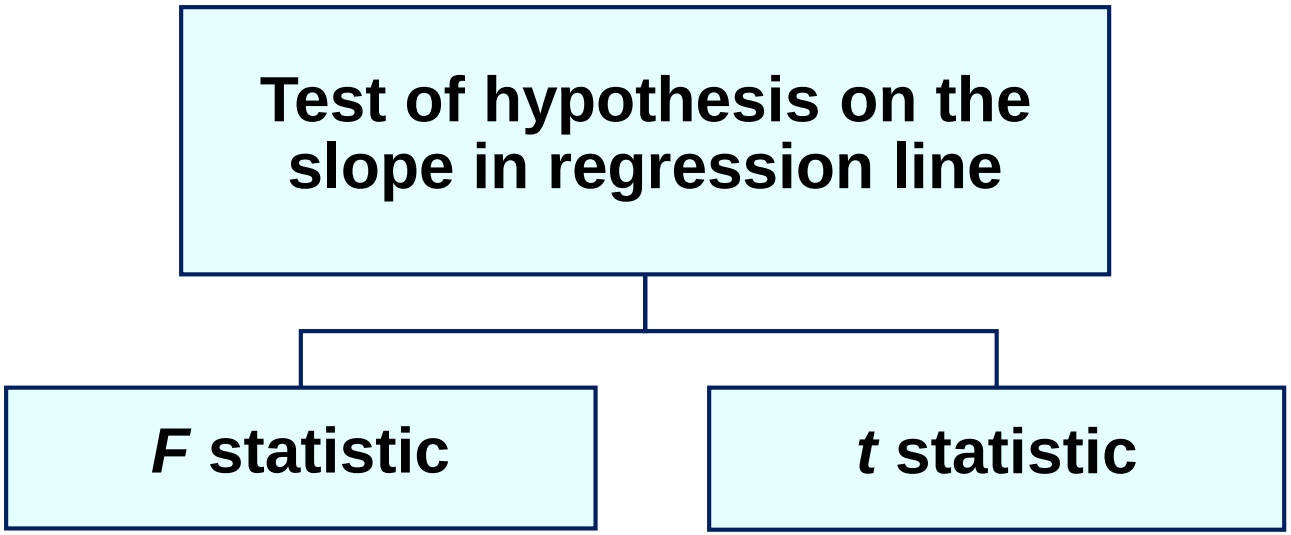


Figure 6. Flow chart for test of hypothesis about the slope in the regression line

Hypothesis testing with F statistic

- The hypothesis testing with *F* statistic is based upon the ANOVA table.
- The principle of this test is to divide the total variation of the observed values of outcome variable (Y) into two components:
  - Explained variation
  - Unexplained variation

Table 2. ANOVA table for regression analysis

Source	SS	df	MS (variance)	F test
Regression	$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$	**k**	**MSR=SSR/df**	**MSR/MSE**
Error	$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	**n-k-1**	**MSE=SSE/df**	
Total	$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	**n-1**		

Hypothesis testing with t statistic

The t statistic can be defined as:

$$t = \frac{b - 0}{\hat{se}(b)}$$

where,

b is un-bias estimator of the slope,

$\hat{se}(b)$ is un-bias estimator of the SE of the slope.

This ratio has a t distribution with n-2 degree of freedom.

Fitting simple linear regression by STATA

```
. regress sbp1 bmi

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(1, 28)        =      4.95
       Model |  1298.35408         1  1298.35408   Prob > F        =    0.0344
    Residual |  7347.64592        28  262.415926   R-squared       =    0.1502
-------------+----------------------------------   Adj R-squared   =    0.1198
       Total |      8646          29  298.137931   Root MSE        =    16.199

------------------------------------------------------------------------------
        sbp1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   2.193388    .986084     2.22   0.034     .1734861    4.213289
       _cons |   69.75698   22.33493     3.12   0.004     24.00596     115.508
------------------------------------------------------------------------------
```

Interpretation

- We reject the null hypothesis and conclude that slope of regression line is not equal to zero.
- It means that there is a statistically significant linear relationship between SBP and BMI - SBP increases as BMI increases.

Table for the association between patient's characteristics and changes of SBP: simple linear regression

Characteristics	Coefficient	95% CI	P value
BMI	2.19	0.17, 4.21	0.034
Age	0.61	0.26, 0.97	0.001
Gender			
Female	5.83	-7.39, 19.06	0.374
Male	0		

Simple linear regression analysis

I. Simple linear regression model

II. Fitting linear regression model

III. Coefficient of determination (r^2)

IV. Assumption checking

V. Estimation of mean predicted values

VI. Estimation of individual predicted values

III. Coefficient of determination (R^2)

- The R^2 represents the part of the total variation of the observed values of Y which is explained by the linear regression model.
- In the other words, the R^2 shows how well the independent variable explains the variation of the dependent variable in the regression model.

III. Coefficient of determination (R^2)

The calculation of the coefficient of determination is based upon the ANOVA table which can be defined as:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{SSR}{SST}$$

Fitting simple linear regression by STATA

```
. regress sbp1 bmi

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(1, 28)        =      4.95
       Model |  1298.35408         1  1298.35408   Prob > F        =    0.0344
    Residual |  7347.64592        28  262.415926   R-squared       =    0.1502
-------------+----------------------------------   Adj R-squared   =    0.1198
       Total |      8646          29  298.137931   Root MSE        =    16.199

------------------------------------------------------------------------------
        sbp1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   2.193388    .986084     2.22   0.034     .1734861    4.213289
       _cons |   69.75698   22.33493     3.12   0.004     24.00596     115.508
------------------------------------------------------------------------------
```

Interpretation

- The R^2 of the relationship between BMI and SBP is equal to 0.1502.
- This implies that 15.02% of the variation among the observed values of the SBP is explained by its linear relationship with BMI.
- The remaining 84.98% (100-15.02%) of variation is unexplained or in the other words, is explained by other variables which are not included in the model.

Simple linear regression analysis

I. Simple linear regression model

II. Fitting linear regression model

III. Coefficient of determination (r^2)

IV. Assumption checking

V. Estimation of mean predicted values

VI. Estimation of individual predicted values

IV. Assumptions of simple linear regression model

1. Linearity

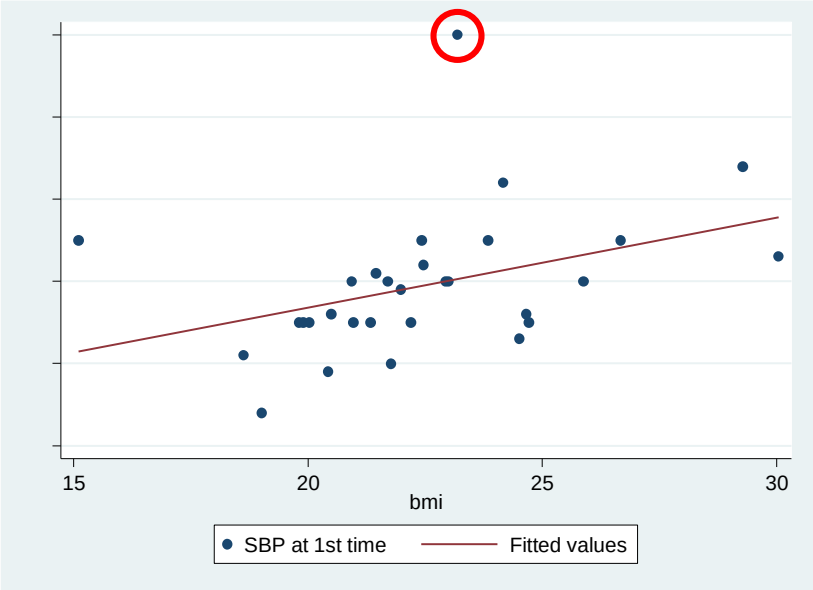
- The relationship between the outcome and predictors should be linear.
- If the relationship between the outcome and predictors is not linear, fitting to the linear regression model is wrong.

Assumptions of simple linear regression model

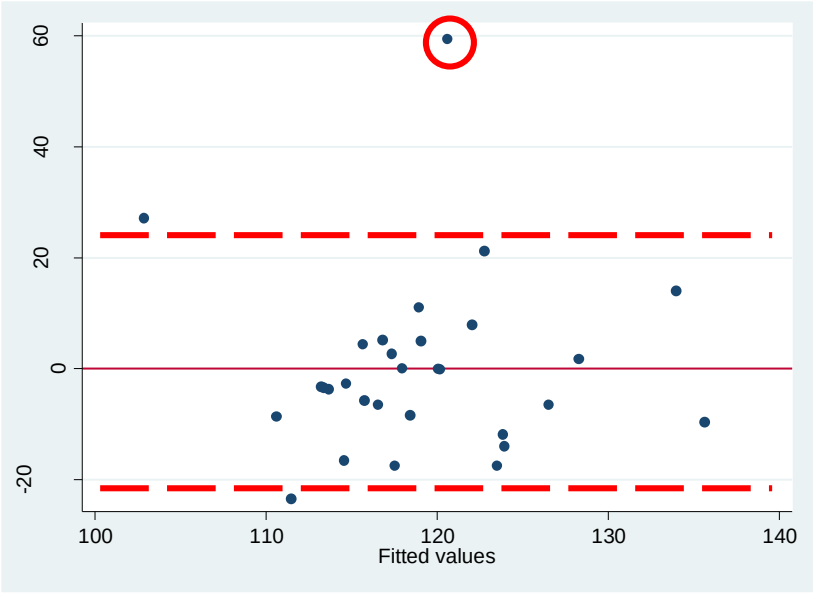
1. Linearity

- Linear regression function can be checked from a residual plot against the predictor, or predicted values, and also from a scatter plot.
- However, the scatter plot is not always as effective as a residual plot.

Check assumption of linearity



a) Scatter plot



b) Residual plot

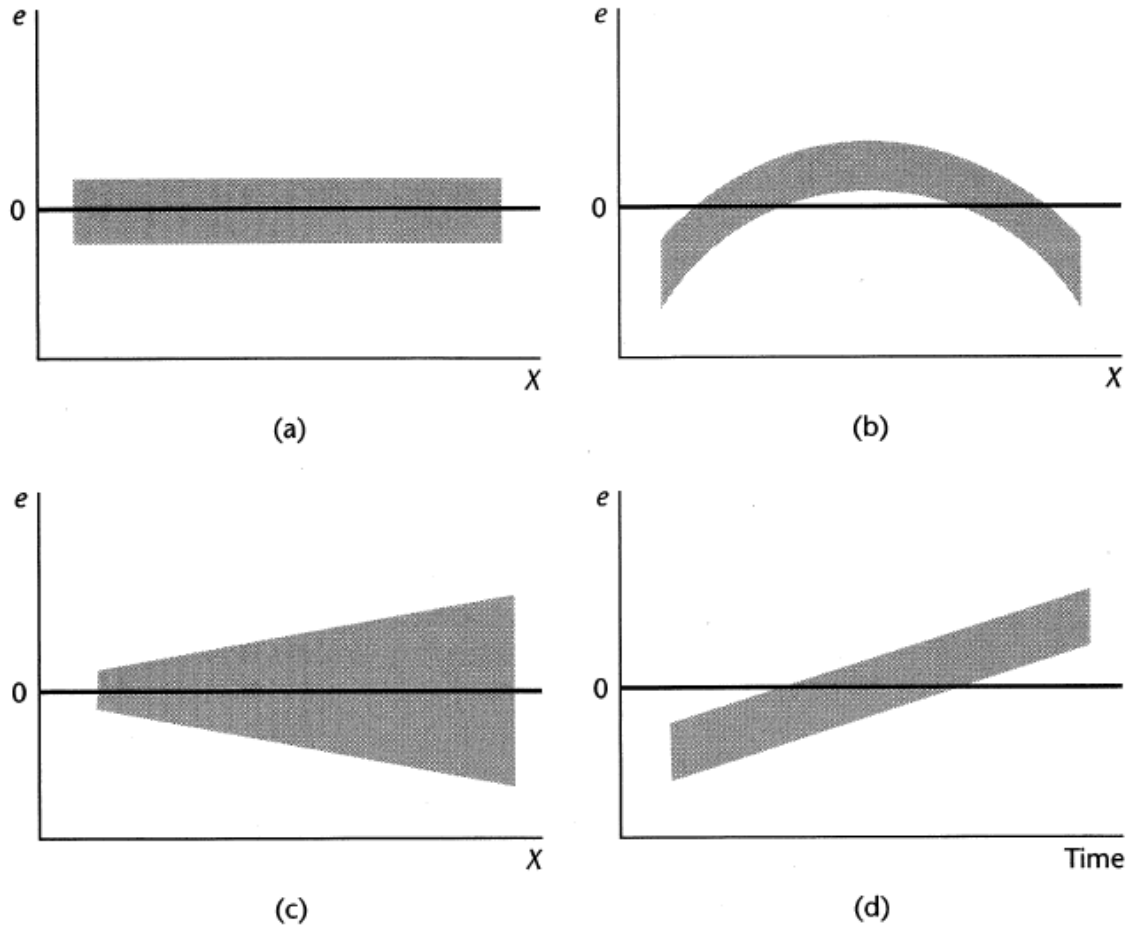


Figure 7. Prototype residual plots

Assumptions of simple linear regression model

2. Normality

- The values of the outcome have a normal distribution for each value of the predictor.
- If this assumption is broken, it will result in an invalid model.

Assumptions of simple linear regression model

2. Normality

- However, if this assumption holds then the residuals should also have a normal distribution.
- This assumption needs to be checked after fitting the simple linear regression model.

Steps for checking normality assumption

- Investigate the characteristics of the distribution of the residuals such as skewness, or kurtosis.
- Create a normal plot of the residuals after fitting the linear regression model.
- Test the hypothesis about normality of residuals by using the Shapiro-Wilk test.

Assumptions of simple linear regression model

3. Homoskedasticity

- The variability of the outcome is the same for each value of the predictor.
- Alternatively, the variance of residuals should be the same for each value of the predictor.
- Standard errors are biased when heteroskedasticity is present. This leads to bias in test statistics and CIs.

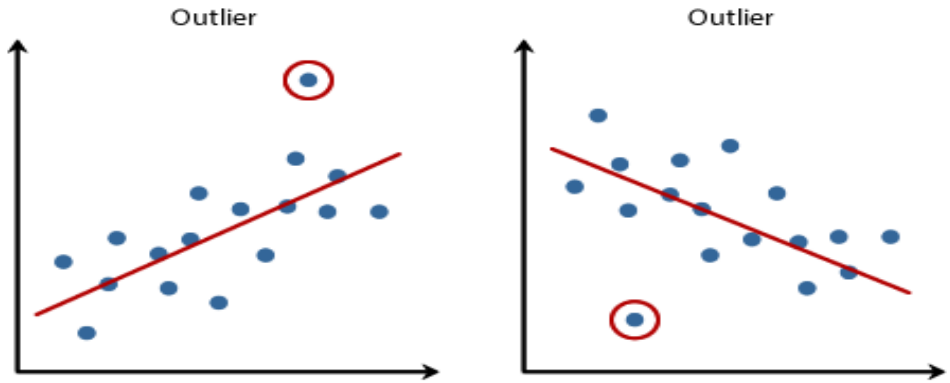
Steps for checking homoscedasticity assumption

- Create a plot of the residuals against the independent variable x or predicted values of the dependent variable y.
- Test hypothesis about homoskedasticity by using the Breusch-Pagan and Cook-Weisberg test .

Assumptions of simple linear regression model

4. Outliers

- The outliers are data values that differ greatly from the majority of a set of data.
- These values fall outside of an overall trend that is present in the data.



Assumptions of simple linear regression model

4. Outliers

- The outliers can result in an invalid model.
- If the outliers are present, data should be checked to make sure that there is no error during data entry, or no error due to measurement.
- If the error came from measurement, that observation must be omitted.

Steps for checking outliers

- Outliers can be checked by plot of the standardized residuals against values of the independent variable.
- The standardized residual can be defined as:

$$z_i = \frac{e_i}{S}$$

where, S is the SD of residuals which can be defined as:

$$S = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2}$$


4
2
0
-2
021
053
013
090
006
127
033
104
117
094
060
087
098
096
120
026
154
051
121
013
115
138
15
20
25
30
bmi

Criteria for identify outliers

- If the residuals lie a long distance from the rest of the observations, usually four or more standard deviation from zero, they are outliers which will affect the validity of the model.
- However, an observation with an standardized residual that is larger than 3 (in absolute value) is generally deemed an outlier.

Simple linear regression analysis

I. Simple linear regression model

II. Fitting linear regression model

III. Coefficient of determination (r^2)

IV. Assumption checking

V. Estimation of mean predicted values

VI. Estimation of individual predicted values

V. Estimation of mean predicted value

- The predicted mean value of y for any specific value of x can be estimated by using the linear regression model.
- For example, the predicted mean value of SBP for all subjects with BMI of 21.45 years (x=21.45) can be defined as:

$$\hat{y} = 69.76 + (2.19 \times 21.45) = 116.81\ \text{mmHg}$$

Interpretation

- This mean predicted value can be interpreted as a point estimator of the mean value of SBP for BMI=21.45
- Thus, we can state that, on average, the SBP is equal to 116.81 mmHg for all subjects whose BMI=21.45.

CI of mean predicted value

The CI for predicted mean can be defined as:

$$\hat{y}_i \pm (t_{1-\alpha/2} \times \hat{se}(\hat{y}_i))$$

where,

$$\hat{se}(\hat{y}_i) = s_{y|x} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

$t_{1-\alpha/2}$ is the appropriate value from the t distribution with n-2 degrees of freedom associated with a confidence of 100x(1-α)%.

Simple linear regression analysis

I. Simple linear regression model

II. Fitting linear regression model

III. Coefficient of determination (r^2)

IV. Assumption checking

V. Estimation of mean predicted values

VI. Estimation of individual predicted values

VI. Estimation of individual predicted value

- We predict an individual value of y for a new member of population.
- The predicted individual value of y is identical to the predicted mean value of y, but CI of the predicted individual value is much wider than CI of the predicted mean value.

CI of individual predicted value

The CI for predicted individual value of y can be defined as:

$$\tilde{y}_i \pm (t_{1-\alpha/2} \times \hat{se}(\tilde{y}_i))$$

where,

$$\hat{se}(\tilde{y}_i) = s_{y|x} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

$t_{1-\alpha/2}$ is the appropriate value from the t distribution with n-2 degrees of freedom associated with a confidence of 100x(1-α)%.