STAC67H: Regression Analysis

Fall, 2014

Instructor: Jabed Tomal

Department of Computer and Mathematical Sciences
University of Toronto Scarborough
Toronto, ON
Canada

September 10, 2014

Simple Linear Regression Model:

Distribution of Error Unspecified

A simple linear regression model is defined as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \ ; \ i = 1, 2, \cdots, n.$$

Simple Linear Regression Model:

Distribution of Error Unspecified

A simple linear regression model is defined as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \ ; \ i = 1, 2, \cdots, n.$$

Where,

1. Y_i is the value of the response/outcome variable in the ith trial,

Simple Linear Regression Model:

Distribution of Error Unspecified

A simple linear regression model is defined as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \ ; \ i = 1, 2, \cdots, n.$$

Where,

1. Y_i is the value of the response/outcome variable in the ith trial,
2. β_0 and β_1 are the regression coefficients (parameters),

Simple Linear Regression Model:

Distribution of Error Unspecified

A simple linear regression model is defined as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \ ; \ i = 1, 2, \cdots, n.$$

Where,

1. Y_i is the value of the response/outcome variable in the ith trial,
2. β_0 and β_1 are the regression coefficients (parameters),
3. X_i is the value of the predictor variable in the ith trial (known constant),

Simple Linear Regression Model:

Distribution of Error Unspecified

A simple linear regression model is defined as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \ ; \ i = 1, 2, \cdots, n.$$

Where,

1. Y_i is the value of the response/outcome variable in the ith trial,
2. β_0 and β_1 are the regression coefficients (parameters),
3. X_i is the value of the predictor variable in the ith trial (known constant),
4. ϵ_i is a random error term with

Simple Linear Regression Model:

Distribution of Error Unspecified

A simple linear regression model is defined as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \ ; \ i = 1, 2, \cdots, n.$$

Where,

1. Y_i is the value of the response/outcome variable in the ith trial,
2. β_0 and β_1 are the regression coefficients (parameters),
3. X_i is the value of the predictor variable in the ith trial (known constant),
4. ϵ_i is a random error term with
   - mean $E(\epsilon_i) = 0$,

Simple Linear Regression Model:

Distribution of Error Unspecified

A simple linear regression model is defined as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \ ; \ i = 1, 2, \cdots, n.$$

Where,

1. Y_i is the value of the response/outcome variable in the ith trial,
2. β_0 and β_1 are the regression coefficients (parameters),
3. X_i is the value of the predictor variable in the ith trial (known constant),
4. ϵ_i is a random error term with
   - mean $E(\epsilon_i) = 0$,
   - variance $Var(\epsilon_i) = \sigma^2$ and

Simple Linear Regression Model:

Distribution of Error Unspecified

A simple linear regression model is defined as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \ ; \ i = 1, 2, \cdots, n.$$

Where,

1. Y_i is the value of the response/outcome variable in the ith trial,
2. β_0 and β_1 are the regression coefficients (parameters),
3. X_i is the value of the predictor variable in the ith trial (known constant),
4. ϵ_i is a random error term with
   - mean $E(\epsilon_i) = 0$,
   - variance $Var(\epsilon_i) = \sigma^2$ and
   - ϵ_i and ϵ_j are uncorrelated so that their covariance is zero, i.e., $Cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$.

Simple Linear Regression Model:

This regression model is called

1. *Simple*: As there is only one predictor variable.

Simple Linear Regression Model:

This regression model is called

1. *Simple*: As there is only one predictor variable.
2. *linear in the parameters*: no parameter appears as an exponent or is multiplied or divided by another parameter.

Simple Linear Regression Model:

This regression model is called

1. *Simple*: As there is only one predictor variable.
2. *linear in the parameters*: no parameter appears as an exponent or is multiplied or divided by another parameter.
3. *linear in the predictor variable*: the variable appears only in the first power.

Simple Linear Regression Model:

This regression model is called

1. *Simple*: As there is only one predictor variable.
2. *linear in the parameters*: no parameter appears as an exponent or is multiplied or divided by another parameter.
3. *linear in the predictor variable*: the variable appears only in the first power.

Simple Linear Regression Model:

This regression model is called

1. *Simple*: As there is only one predictor variable.
2. *linear in the parameters*: no parameter appears as an exponent or is multiplied or divided by another parameter.
3. *linear in the predictor variable*: the variable appears only in the first power.

Note: A model that is linear in the parameters and in the predictor variable is also called a *first-order model*.

Simple Linear Regression Model:

Important Features of the Model

1. The response Y_i is the sum of two components: (1) the constant regression term $\beta_0 + \beta_1 X_i$, and (2) the random term ϵ_i. Hence, Y_i is also random.

Simple Linear Regression Model:

Important Features of the Model

1. The response Y_i is the sum of two components: (1) the constant regression term $\beta_0 + \beta_1 X_i$, and (2) the random term ϵ_i. Hence, Y_i is also random.
2. Since $E\{\epsilon_i\} = 0$, $E\{Y_i\} = \beta_0 + \beta_1 X_i$. Thus the response Y_i comes from a probability distribution whose mean is $E\{Y_i\} = \beta_0 + \beta_1 X_i$. Hence, the regression function is $E\{Y|X\} = \beta_0 + \beta_1 X$.

Simple Linear Regression Model:

Important Features of the Model

1. The response Y_i is the sum of two components: (1) the constant regression term $\beta_0 + \beta_1 X_i$, and (2) the random term ϵ_i. Hence, Y_i is also random.
2. Since $E\{\epsilon_i\} = 0$, $E\{Y_i\} = \beta_0 + \beta_1 X_i$. Thus the response Y_i comes from a probability distribution whose mean is $E\{Y_i\} = \beta_0 + \beta_1 X_i$. Hence, the regression function is $E\{Y|X\} = \beta_0 + \beta_1 X$.
3. The response Y_i exceeds or falls short of the value of the regression function by the error term amount ϵ_i.

Simple Linear Regression Model:

Important Features of the Model

1. The response Y_i is the sum of two components: (1) the constant regression term $\beta_0 + \beta_1 X_i$, and (2) the random term ϵ_i. Hence, Y_i is also random.
2. Since $E\{\epsilon_i\} = 0$, $E\{Y_i\} = \beta_0 + \beta_1 X_i$. Thus the response Y_i comes from a probability distribution whose mean is $E\{Y_i\} = \beta_0 + \beta_1 X_i$. Hence, the regression function is $E\{Y|X\} = \beta_0 + \beta_1 X$.
3. The response Y_i exceeds or falls short of the value of the regression function by the error term amount ϵ_i.
4. $Var(\epsilon_i) = \sigma^2$ implies $Var(Y_i) = \sigma^2$, a constant regardless of the level of the predictor variable X.

Simple Linear Regression Model:

Important Features of the Model

1. The response Y_i is the sum of two components: (1) the constant regression term $\beta_0 + \beta_1 X_i$, and (2) the random term ϵ_i. Hence, Y_i is also random.
2. Since $E\{\epsilon_i\} = 0$, $E\{Y_i\} = \beta_0 + \beta_1 X_i$. Thus the response Y_i comes from a probability distribution whose mean is $E\{Y_i\} = \beta_0 + \beta_1 X_i$. Hence, the regression function is $E\{Y|X\} = \beta_0 + \beta_1 X$.
3. The response Y_i exceeds or falls short of the value of the regression function by the error term amount ϵ_i.
4. $Var(\epsilon_i) = \sigma^2$ implies $Var(Y_i) = \sigma^2$, a constant regardless of the level of the predictor variable X.
5. The error terms ϵ_i and ϵ_j are uncorrelated, so are the responses Y_i and Y_j for $i \neq j$.

Simple Linear Regression Model:

Important Features of the Model

In summary, the presented regression model implies that the responses Y_i come from probability distributions whose means are $E\{Y_i\} = \beta_0 + \beta_1 X_i$ and whose variances are σ^2, the same for all levels of X. Further, any two responses Y_i and Y_j are uncorrelated.

Simple Linear Regression Model:

Meaning of Regression Parameters

1. **The slope coefficient** β_1: It indicates the change in the mean of the probability distribution of Y per unit increase in X.

Simple Linear Regression Model:

Meaning of Regression Parameters

1. **The slope coefficient** β_1: It indicates the change in the mean of the probability distribution of Y per unit increase in X.
2. **The intercept of the regression line** β_0: When the spope of the model includes $X = 0$, β_0 gives the mean of the probability distribution of Y at $X = 0$. When the scope of the model does not cover $X = 0$, β_0 does not have any particular meaning as a separate term in the regression model.

Simple Linear Regression Model:

Alternative Versions of Regression Model

1. An alternative version of our regression model is as following

$$Y_i = \beta_0 X_0 + \beta_1 X_i + \epsilon_i \qquad \text{where } X_0 = 1.$$

This version of the model associates an X variable with each regression coefficient.

Simple Linear Regression Model:

Alternative Versions of Regression Model

1. An alternative version of our regression model is as following

$$Y_i = \beta_0 X_0 + \beta_1 X_i + \epsilon_i \quad \text{where } X_0 = 1.$$

This version of the model associates an X variable with each regression coefficient.

2. Another version of our regression model could be

$$Y_i = \beta_0^* + \beta_1 (X_i - \bar{X}) + \epsilon_i \quad \text{where } \beta_0^* = \beta_0 + \beta_1 \bar{X}.$$

Here, the interpretation of the regression intercept is different than before.

Simple Linear Regression Model:

Data for Regression Analysis

Ordinarily, we do not know the values of the regression coefficients β_0 and β_1, and we need to estimate them from relevant data.

Simple Linear Regression Model:

Data for Regression Analysis

Ordinarily, we do not know the values of the regression coefficients β_0 and β_1, and we need to estimate them from relevant data.

- **Observational Data**: Obtained from nonexperimental studies. Such studies do not control the explanatory or predictor variable(s) of interest.

Simple Linear Regression Model:

Data for Regression Analysis

Ordinarily, we do not know the values of the regression coefficients β_0 and β_1, and we need to estimate them from relevant data.

- **Observational Data**: Obtained from nonexperimental studies. Such studies do not control the explanatory or predictor variable(s) of interest.
  - Example: A company official wished to study the relation between *age of employee* (X) and *number of days of illness last year* (Y). Data obtained from personal record and the explanatory variable, *age*, was not controlled.

Simple Linear Regression Model:

Data for Regression Analysis

Ordinarily, we do not know the values of the regression coefficients β_0 and β_1, and we need to estimate them from relevant data.

- **Observational Data**: Obtained from nonexperimental studies. Such studies do not control the explanatory or predictor variable(s) of interest.
  - Example: A company official wished to study the relation between *age of employee* (X) and *number of days of illness last year* (Y). Data obtained from personal record and the explanatory variable, *age*, was not controlled.
  - A major limitation of observational data is that they often do not provide adequate information about cause-and-effect relationships.

Simple Linear Regression Model:

Data for Regression Analysis

- **Experimental Data**: Obtained from controlled experiment.

Simple Linear Regression Model:

Data for Regression Analysis

- **Experimental Data**: Obtained from controlled experiment.
  - Example: An insurance company wishes to study the relation between *productivity of its analysts* in processing claims and *length of training*. Each group of 3 employees, selected at random from a total of 9 employees, are trained for two, three and five weeks. The productivity of the analysts are observed for the next 10 weeks.

Simple Linear Regression Model:

Data for Regression Analysis

- **Experimental Data**: Obtained from controlled experiment.
  - Example: An insurance company wishes to study the relation between *productivity of its analysts* in processing claims and *length of training*. Each group of 3 employees, selected at random from a total of 9 employees, are trained for two, three and five weeks. The productivity of the analysts are observed for the next 10 weeks.
  - Here, control is exercised over the explanatory variable, *length of training*.

Simple Linear Regression Model:

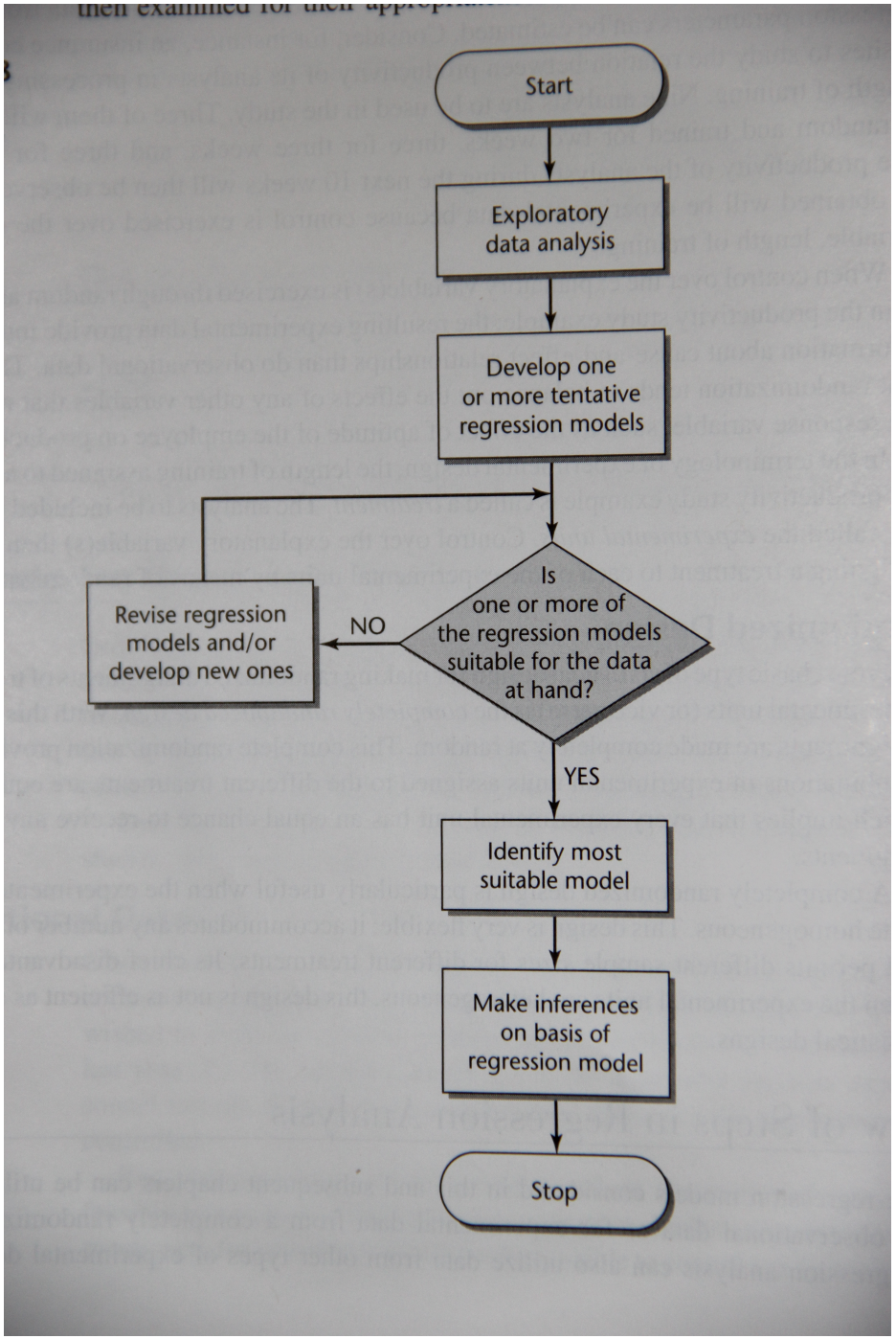
Data for Regression Analysis

- **Experimental Data**: Obtained from controlled experiment.
  - Example: An insurance company wishes to study the relation between *productivity of its analysts* in processing claims and *length of training*. Each group of 3 employees, selected at random from a total of 9 employees, are trained for two, three and five weeks. The productivity of the analysts are observed for the next 10 weeks.
  - Here, control is exercised over the explanatory variable, *length of training*.
  - Provide much stronger information about cause-and-effect relationships than do observational data.

Simple Linear Regression Model:

Overview of Steps in Regression Analysis

The following flowchart shows a typical strategy for regression analysis. In the usual situation, where we do not have adequate knowledge to specify the appropriate regression model in advance, the first step is an exploratory study of the data.


Start
Exploratory data analysis
Develop one or more tentative regression models
Is one or more of the regression models suitable for the data at hand?
NO
Revise regression models and/or develop new ones
YES
Identify most suitable model
Make inferences on basis of regression model
Stop

Simple Linear Regression Model:

Estimation of Regression Function

We have data for n observations or trials (X_i, Y_i) ; $i = 1, 2, \cdots, n$. In simple notations, we write

$$(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n).$$

Simple Linear Regression Model:

Estimation of Regression Function

The simple linear regression model can be rearranged to get the error term for the *i*th observation or trial as following

$$\epsilon_i = Y_i - \beta_0 - \beta_1 X_i \quad ; \; i = 1, 2, \cdots, n.$$

Simple Linear Regression Model:

Estimation of Regression Function

The simple linear regression model can be rearranged to get the error term for the *i*th observation or trial as following

$$\epsilon_i = Y_i - \beta_0 - \beta_1 X_i \quad ; \; i = 1, 2, \cdots, n.$$

We want to estimate β_0 and β_1 such that each of the ϵ_i is as small as possible.

Simple Linear Regression Model:

Method of Least Squares

The method of least squares requires that we consider the sum of the n squared deviations of Y_i from its expected value:

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Simple Linear Regression Model:

Method of Least Squares

The method of least squares requires that we consider the sum of the n squared deviations of Y_i from its expected value:

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

According to the method of least squares, the estimators of β_0 and β_1 are those values b_0 and b_1, respectively, that minimize the criterion Q for the given sample observations $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$.

Simple Linear Regression Model:

Method of Least Squares

The values of b_0 and b_1 that minimize Q for any particular set of sample data are given by the following simultaneous equations:

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_i.$$

$$\sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2.$$

The equations are called *normal equations*, and b_0 and b_1 are called *point estimators* of β_0 and β_1, respectively.

Simple Linear Regression Model:

Method of Least Squares

The normal equations can be solved simultaneously for b_0 and b_1:

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2},$$

and

$$b_0 = \bar{Y} - b_1\bar{X},$$

where $\bar{X}$ and $\bar{Y}$ are the means of the X_i and the Y_i observations, respectively.

Simple Linear Regression Model:

Exercise 1.21 Airfreight breakage. A substance used in biological and medical research is shipped by airfreight to users in cartons of $1,000$ ampules. The data below, involving 10 shipments, were collected on the number of time the carton was transferred from one aircraft to another over the shipment route (X) and the number of ampules found to be broken upon arrival (Y). Assume that the first-order regression model is appropriate.

i:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
X_i:	1	0	2	0	3	1	0	1	2	0
Y_i:	16	9	17	12	22	13	8	15	19	11

(a) Obtain the estimated regression function. Plot the estimated regression function and the data. Does a linear regression function appears to give a good fit here?

Simple Linear Regression Model:

In this problem $n = 10$, $\bar{X} = 1$, $\bar{Y} = 14.2$, $\sum_{i=1}^{n} X_i^2 = 20$, and $\sum_{i=1}^{n} X_i Y_i = 182$. Hence,

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \frac{182 - 10 \times 1 \times 14.2}{20 - 10 \times (1)^2} = 4.0,$$

and

$$b_0 = \bar{Y} - b_1 \bar{X} = 14.2 - 4.0 \times 1 = 10.2.$$

Simple Linear Regression Model:

The estimated regression function is

$$\widehat{E(Y|X)} = 10.2 + 4.0X.$$

Simple Linear Regression Model:

The estimated regression function is

$$\widehat{E(Y|X)} = 10.2 + 4.0X.$$

We interpret the estimated parameters as following:

- $b_1 = 4.0$: with the increase of 1 transfers from one aircraft to another over the shipment route, the mean number of broken ampules increases to 4.0.

Simple Linear Regression Model:

The estimated regression function is

$$\widehat{E(Y|X)} = 10.2 + 4.0X.$$

We interpret the estimated parameters as following:

- $b_1 = 4.0$: with the increase of 1 transfers from one aircraft to another over the shipment route, the mean number of broken ampules increases to 4.0.
- $b_0 = 10.2$: given the number of transfers being 0, the mean number of broken ampules is 10.2.

Scatter Plot:

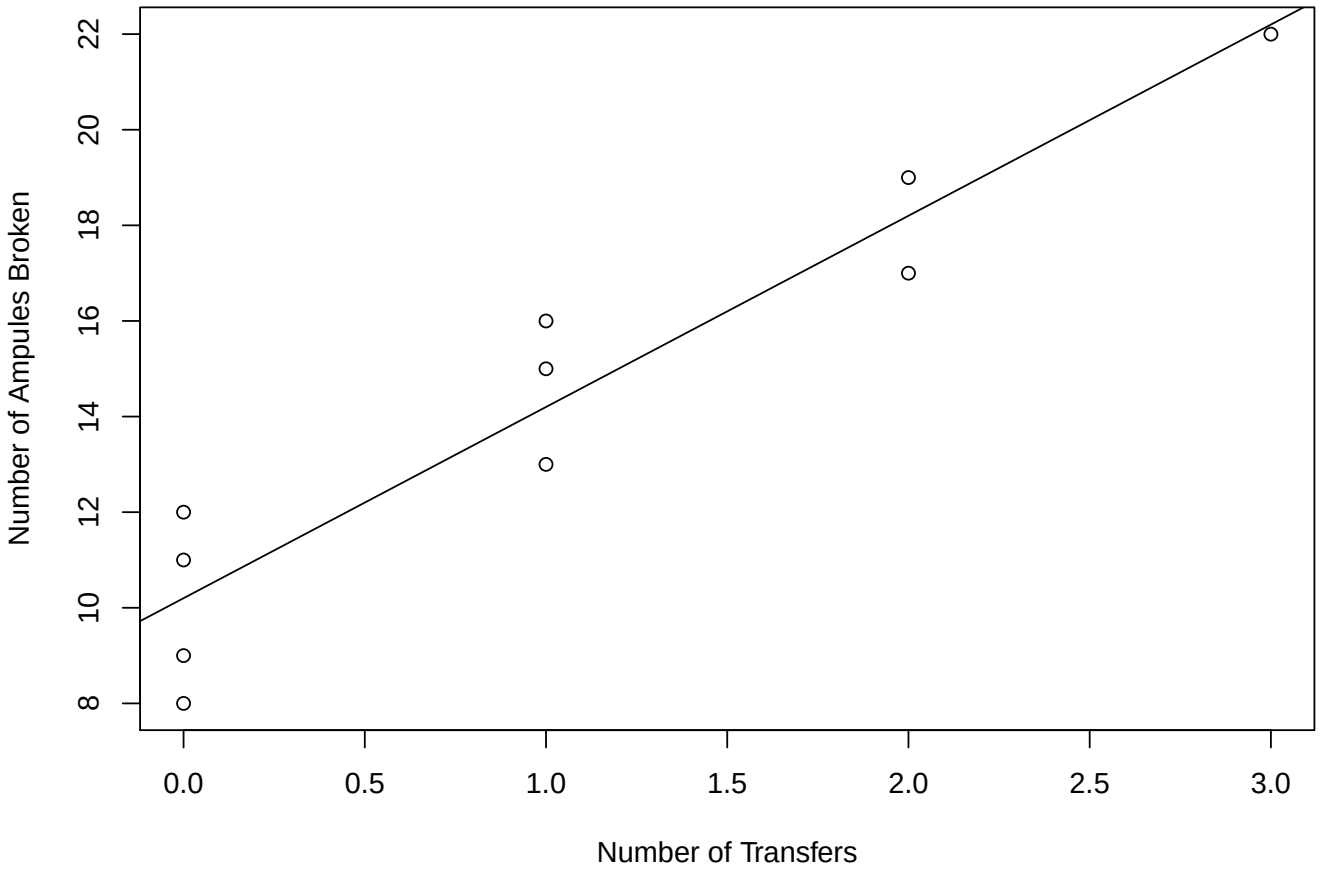


Figure: Plot of the data with the estimated regression function.

Properties of Least Squares Estimators:

Gauss-Markov Theorem: Under the conditions of the stated regression model, the least squares estimators b_0 and b_1 are unbiased and have minimum variance among all unbiased linear estimators.

Properties of Least Squares Estimators:

1. The *Gauss-Markov Theorem* states that b_0 and b_1 are unbiased estimators of β_0 and β_1, respectively. Hence,

$$E\{b_0\} = \beta_0 \qquad E\{b_1\} = \beta_1$$

so that neither estimator tends to overestimate or underestimate systematically.

Properties of Least Squares Estimators:

2. The *Gauss-Markov Theorem* states that the estimators b_0 and b_1 are more precise (i.e., their sampling distributions are less variable) than any other estimators belonging to the class of unbiased estimators that are linear functions of the observations $Y_1, Y_2, \cdots, Y_n$.

Properties of Least Squares Estimators:

3. The estimators b_0 and b_1 are linear functions of the observations $Y_1, Y_2, \cdots, Y_n$.

Properties of Least Squares Estimators:

The estimator b_1 can be expressed as

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sum_{i=1}^{n} k_i Y_i,$$

which is a linear combination of $Y_1, Y_2, \cdots, Y_n$. Here, $k_i = (X_i - \bar{X}) / \sum_{i=1}^{n}(X_i - \bar{X})^2$.

Properties of Least Squares Estimators:

The estimator b_1 can be expressed as

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sum_{i=1}^{n} k_i Y_i,$$

which is a linear combination of $Y_1, Y_2, \cdots, Y_n$. Here, $k_i = (X_i - \bar{X}) / \sum_{i=1}^{n}(X_i - \bar{X})^2$.
The estimator b_0 can be expressed as

$$b_0 = \sum_{i=1}^{n} k_{0i} Y_i,$$

which is a linear combination of $Y_1, Y_2, \cdots, Y_n$. Here, $k_{0i} = (1/n) - \bar{X} k_i$.

Point Estimation of Mean Response:

Given sample estimators b_0 and b_1 of the parameters in the regression function:

$$E\{Y|X\} = \beta_0 + \beta_1 X$$

we estimate the regression function as follows

$$\hat{Y} = b_0 + b_1 X$$

where $\hat{Y}$ (read Y hat) is the value of the estimated regression function at the level X of the predictor variable.

Point Estimation of Mean Response:

In our *airfreight breakage* exercise (1.21), the estimated regression function is

$$\hat{Y} = 10.2 + 4.0X.$$

Point Estimation of Mean Response:

For the cases in the study, we will call $\hat{Y}_i$

$$\hat{Y}_i = 10.2 + 4.0X_i \quad ; \; i = 1, 2, \cdots, n$$

the *fitted value* for the *i*th case.

Point Estimation of Mean Response:

For the cases in the study, we will call $\hat{Y}_i$

$$\hat{Y}_i = 10.2 + 4.0X_i \quad ; \; i = 1, 2, \cdots, n$$

the *fitted value* for the *i*th case.
Note: The *fitted* value $\hat{Y}_i$ is to be viewed in distinction to the *observed value* Y_i.

Point Estimation of Mean Response:

In our *airfreight breakage* exercise (1.21), the *fitted* values are in the third row of the following table

i:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
X_i:	1	0	2	0	3	1	0	1	2	0
Y_i:	16	9	17	12	22	13	8	15	19	11
$\hat{Y}_i$:	14.2	10.2	18.2	10.2	22.2	14.2	10.2	14.2	18.2	10.2

Point Estimation of Mean Response:

In our *airfreight breakage* exercise (1.21), the *fitted* values are in the third row of the following table

i:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
X_i:	1	0	2	0	3	1	0	1	2	0
Y_i:	16	9	17	12	22	13	8	15	19	11
$\hat{Y}_i$:	14.2	10.2	18.2	10.2	22.2	14.2	10.2	14.2	18.2	10.2

Note: Remember the fitted values are the points in your estimated regression function.

Point Estimation of Mean Response:

Alternative Model

When the alternative regression model

$$Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \epsilon_i$$

is in concern, the lease squares estimator b_1 of β_1 remains the same as before.

Point Estimation of Mean Response:

Alternative Model

When the alternative regression model

$$Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \epsilon_i$$

is in concern, the lease squares estimator b_1 of β_1 remains the same as before.
The lease squares estimator of $\beta_0^* = \beta_0 + \beta_1 \bar{X}$ becomes

$$b_0^* = b_0 + b_1 \bar{X} = \bar{Y}. \text{ How?}$$

Point Estimation of Mean Response:

Alternative Model

When the alternative regression model

$$Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \epsilon_i$$

is in concern, the lease squares estimator b_1 of β_1 remains the same as before.
The lease squares estimator of $\beta_0^* = \beta_0 + \beta_1 \bar{X}$ becomes

$$b_0^* = b_0 + b_1 \bar{X} = \bar{Y}. \text{ How?}$$

Hence, the estimated regression function for alternative model is

$$\hat{Y} = \bar{Y} + b_1(X - \bar{X}).$$

Point Estimation of Mean Response:

In our *airfreight breakage* exercise (1.21), the *estimated* regression function of the alternative model is

$$\hat{Y} = 14.2 + 4(X - \bar{X}),$$

which gives the same fitted values.

Residuals:

The *i*th *residual* is the difference between the observed value Y_i and the corresponding *fitted* value $\hat{Y}_i$. This residual is denoted by e_i and is defined in general as follows:

$$e_i = Y_i - \hat{Y}_i.$$

Residuals:

The *i*th *residual* is the difference between the observed value Y_i and the corresponding *fitted* value $\hat{Y}_i$. This residual is denoted by e_i and is defined in general as follows:

$$e_i = Y_i - \hat{Y}_i.$$

For our regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, the residual e_i becomes:

$$e_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1 X_i.$$

Residuals:

- Distinction between the model error term $\epsilon_i = Y_i - E\{Y_i|X_i\}$ and the residual $e_i = Y_i - \hat{Y}_i$: The former involves the vertical deviation of Y_i from the unknown true regression line and hence is unknown. The latter is is the vertical deviation of Y_i from the fitted value $\hat{Y}_i$ on the estimated regression line, and is known.

Residuals:

- Distinction between the model error term $\epsilon_i = Y_i - E\{Y_i|X_i\}$ and the residual $e_i = Y_i - \hat{Y}_i$: The former involves the vertical deviation of Y_i from the unknown true regression line and hence is unknown. The latter is is the vertical deviation of Y_i from the fitted value $\hat{Y}_i$ on the estimated regression line, and is known.
- **Residuals are highly useful for studying whether a given regression model is appropriate for the data at hand.**

Point Estimation of Mean Response:

In our *airfreight breakage* exercise (1.21), the *residual* values are in the fourth row of the following table

i:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
X_i:	1	0	2	0	3	1	0	1	2	0
Y_i:	16	9	17	12	22	13	8	15	19	11
$\hat{Y}_i$:	14.2	10.2	18.2	10.2	22.2	14.2	10.2	14.2	18.2	10.2
e_i:	1.8	-1.2	-1.2	1.8	-0.2	-1.2	-2.2	0.8	0.8	0.8

Properties of Fitted Regression Line:

- The sum of the residuals is zero:

$$\sum_{i=1}^{n} e_i = 0.$$

Properties of Fitted Regression Line:

- The sum of the residuals is zero:

$$\sum_{i=1}^{n} e_i = 0.$$

- The sum of the squared residuals, $\sum_{i=1}^{n} e_i^2$ is the minimum.

Properties of Fitted Regression Line:

- The sum of the residuals is zero:

$$\sum_{i=1}^{n} e_i = 0.$$

- The sum of the squared residuals, $\sum_{i=1}^{n} e_i^2$ is the minimum.
- The sum of the observed values Y_i equals to the sum of the fitted values $\hat{Y}_i$:

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i.$$

Properties of Fitted Regression Line:

- The sum of the weighted residuals is zero when the residual in the ith trial is weighted by the level of the predictor variable in the ith trial:

$$\sum_{i=1}^{n} X_i e_i = 0.$$

Properties of Fitted Regression Line:

- The sum of the weighted residuals is zero when the residual in the ith trial is weighted by the level of the predictor variable in the ith trial:

$$\sum_{i=1}^{n} X_i e_i = 0.$$

- The sum of the weighted residuals is zero when the residual in the ith trial is weighted by the fitted values of the response variable for the ith trial:

$$\sum_{i=1}^{n} \hat{Y}_i e_i = 0.$$

Properties of Fitted Regression Line:

- The sum of the weighted residuals is zero when the residual in the ith trial is weighted by the level of the predictor variable in the ith trial:

$$\sum_{i=1}^{n} X_i e_i = 0.$$

- The sum of the weighted residuals is zero when the residual in the ith trial is weighted by the fitted values of the response variable for the ith trial:

$$\sum_{i=1}^{n} \hat{Y}_i e_i = 0.$$

- The regression always goes through the point $(\bar{X}, \bar{Y})$.