Sections 1.1, 1.3, 2.1, 2.2.1 of the textbook

Regression Analysis

Often in statistics, we use sample data to investigate the relationship among variables. Finding a mathematical model that best describes the relationship between a variable (**response or dependent** variable) and other variables (**regressors, predictor or independent** variables) whose values we believe can be used to predict that response variables is considered a key task in regression analysis.

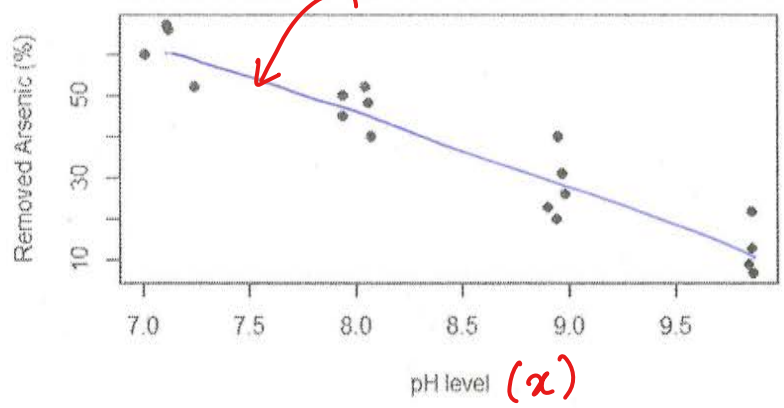
The variable to be predicted is called the **response variable.**

The variable(s) that are used to predict the response variable are called **predictor variable(s).**

Example 1:

Arsenic is found in many ground-waters and some surface waters. Recent health effects research has prompted the Environmental Protection Agency to reduce allowable arsenic levels in drinking water so that many water systems are no longer compliant with standards. This has spurred interest in the development of methods to remove arsenic. The accompanying data on x =pH and y = arsenic removed (%) by a particular process was read from a scatter plot in the article "Optimizing Arsenic Removal During Iron Removal:Theoretical and Practical Considerations" (J. of Water Supply Res. and Tech., 2005: 545–560).

y	x
60	7.01
67	7.11
66	7.12
52	7.24
50	7.94
45	7.94
52	8.04
48	8.05
40	8.07
23	8.90
20	8.94
40	8.95
31	8.97
26	8.98
9	9.85
22	9.86
13	9.86
7	9.87



Scatterplot of arsenic percentage removed and pH level of water

The above scatterplot shows the relationship between the arsenic percentage removed and the pH level of water.

There is a decreasing (negative) approximately linear (because data follow a straight line) relationship.

* Not all the data fall on the line. Therefore, we add random error (ε) in the regression models in the next pages.

Main Goals of Regression Analysis

- Describe the relationship between predictor(s) and response variable
- Based on the identified relationship, predict the y-value for given x-value

Sections 1.1, 1.3, 2.1, 2.2.1 of the textbook

Linear regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \varepsilon$$

random error

y is response variable

$x_1, x_2, \ldots, x_k$ are the predictor variables

$\beta_0, \beta_1, \ldots, \beta_k$ are regression parameters (coefficients)

ε is the error term

This is a **multiple linear regression** model because it has more than one predictor.

The **linear** regression models have linearity in the parameters.

weights of coefficients

For example,

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon = \beta_0 + w_1\beta_1 + w_2\beta_2 + \varepsilon$ This is a linear Function of regression coefficients.

is a **linear** regression model.

However,

$y = \beta^x + \varepsilon$ Not linear in terms of β.

is **NOT a linear** regression model. This is non-linear in β.

To check whether the function is LINEAR in regression coefficients:

* Take partial derivative of function with respect to each coefficient.
* IF it is a function of a regression coefficient, then it is NOT linear in coefficients.

REGRESSION MODELS

- Linear Models
  - **Simple Linear** (only one Predictor variable)

    Model

    $y = \beta_0 + \beta_1 x + \varepsilon$
  - **Multiple Linear** (more than one predictor variable)

    Model

    $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

    $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1 x_2 + \beta_3 x_2^2 + \varepsilon$

    $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$

    $y = \beta_0 + \beta_1 Sin(x_1) + \beta_2 Sin(x_2) + \varepsilon$
- Non-Linear Models
  - **Intrinsically Linear** (can transform to a linear model)

    Model

    $y = \beta_0 x^{\beta_1} + \varepsilon$

    $y = \beta_0 + \beta_1 e^{\beta_2 x} + \varepsilon$

    $y = e^{\beta_0 + \beta_1 x} + \varepsilon$

    $y = e^{\beta} x + \varepsilon$

    etc.
  - **Not Intrinsically Linear** (cannot transform to a linear model)

    Model

    $y = \beta_0 + \beta_1 (\beta_2^x) + \varepsilon$

    $y = \frac{\beta_0}{\beta_1 + \beta_2 e^{\beta_3 x}} + \varepsilon$

    $y = \frac{1}{1 + e^{\beta_0 + \beta_1 x}} + \varepsilon$

    etc.

For example:

$y = e^{\beta_0 + \beta_1 x} + \varepsilon \Rightarrow \ln(y) = \beta_0 + \beta_1 x + \ln(\varepsilon)$

$y^* = \beta_0 + \beta_1 x + \varepsilon^*$ ← This is a linear function of β_0 and β_1

Regression model building process

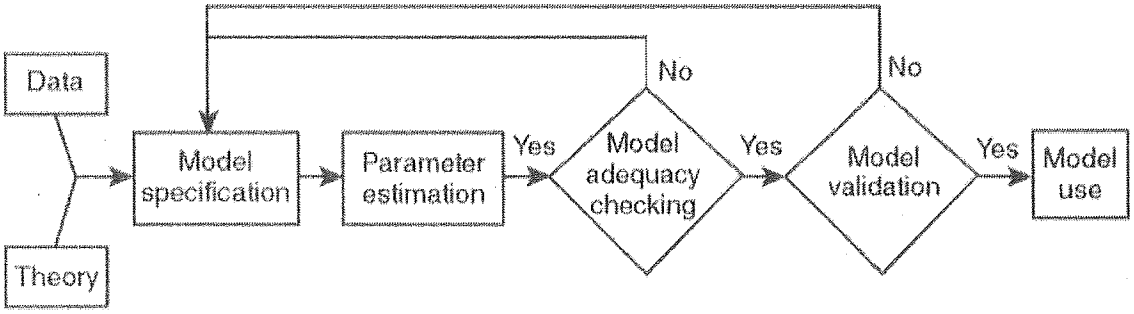


Figure: Regression model building process.

Simple Linear Regression

Only one predictor variable

Relationship is linear (straight line)

The **simple linear** regression model is a regression model that has a **straight line** relationship between the response variable and the predictor variable, and **only one predictor** variable.

Model:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

y is response variable

x is the predictor variable

β_0, β_1 are regression parameters (coefficients)

ε is the random error

β_0 = y-intercept

β_1 = slope

Assumptions:

1. The relationship between the response and predictor is linear.
2. The error term has mean zero. That is, $E(\varepsilon) = 0$
3. The error term has a constant variance for each given x value. That is, $Var(\varepsilon) = \sigma^2$. This is called homoscedasticity (homogeneity of variance)
4. The error terms are uncorrelated (independent). That is, $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$
5. The error terms have a normal distribution.

Errors are independent, and identically distributed $\Rightarrow \varepsilon \overset{iid}{\sim} N(0, \sigma^2)$

With these assumptions about error, the simple linear regression model for n number of observations can be written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{for } i = 1, 2, \ldots, n$$
$$\text{and } \varepsilon_i \sim iid\ N(0, \sigma^2).$$

For example:

i	y	x
1	80	110
2	85	100
3	70	120
4	65	110

$n = 4$

$$y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$$
$$y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2$$
$$y_3 = \beta_0 + \beta_1 x_3 + \varepsilon_3$$
$$y_4 = \beta_0 + \beta_1 x_4 + \varepsilon_4$$