Chapter 1

Introduction

Regression analysis is a method for investigating the functional relationship between two types of variables.

- Response variable (also called dependent variable): Y

- Explanatory variable (also called predictor, regressor, independent variable, covariate): X or $X_1, \ldots, X_k$

Suppose that we observe variables (Y, X), where Y is the response variable and X is predictor. Now, we wish to quantify the relationship between Y and X. Then, the regression model is

$$Y = f(X) + \epsilon,$$

where ϵ is a random noise (or error) with zero mean. We seek a unknown function f for modeling the relationship between X and Y and predicting Y given value of the input X.

For our class,

- we assume the specific form of function f. However, in general,

  $$f(x) = \mathrm{E}[Y|X = x],$$

  which is the pointwise solution of

  $$\min_f \mathrm{E}|Y - f(X)|^2.$$

  This function is often called **the regression of Y on X**.

Possible relationship between response variable and predictors

1. Linear relationship:

$$\left.\begin{aligned} \mathrm{E}[Y|X=x] &= \beta_0 + \beta_1 x \\ \mathrm{E}[Y|X=x] &= \beta_0 + \beta_1 (1/x) \\ \mathrm{E}[Y|X=x] &= \beta_0 + \beta_1 \sqrt{x} \end{aligned}\right\}\text{Simple linear regression}$$

$$\left.\begin{aligned} \mathrm{E}[Y|X=x] &= \beta_0 + \beta_1 x + \beta_2 x^2 \\ \mathrm{E}[Y|X_1=x_1,\ldots,X_k=x_k] &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_k \end{aligned}\right\}\text{Multiple linear regression}$$

2. Nonlinear relationship:
$\mathrm{E}[Y|X=x] = \frac{\beta_0}{1+\exp(\beta_1 x)}$

Note: "Linear relationship" means that the unknown parameters are linearly related to a response variable. We may use transformations to make the relationship linear.

- we assume that the predictor variable is fixed without loss of generality. In fact, we can deal with a random predictor X by conditioning X.
  - Experimental study

    e.g.) Grain production and plant density: The treatment design consisted of five plant densities (10,20,30,40,50). Each of the five treatments was assigned randomly to three field plots. The resulting grain yields are recorded.
  - Observational study

    e.g.) Rocket propellant data in the below.

Uses of regression analysis

- Data description

  Engineers and scientists frequently use equations to summarize or describe a set of data (e.g., investigate the relationship between Y and the X's; i.e., see how Y changes with the X's). Regression analysis is helpful in developing such equations. For example, we may collect a considerable amount of rocket propellant data, and a regression model would probably be a much more convenient and useful summary of those data than a table or even a graph.

- Parameter estimation

  The equations to describe the data include parameters. Sometimes parameter estimation problems can be solved by regression methods.

- Prediction and estimation

  After the data are used to fit a model relating Y to the predictors, we may wish to predict the value of Y at specific value of X. For example, we may wish to predict delivery time for a specified number of cases of soft drinks to be delivered. These predictions may be helpful in planning delivery activities such as routing and scheduling or in evaluating the productivity of delivery operations.

- Control

  Regression models may be used for control purposes. For example, a chemical engineer could use regression analysis to develop a model relating the tensile strength of paper to the hardwood concentration in the pulp. This equation could then be used to control the strength to suitable values by varying the level of hardwood.

Examples

- Rocket Propellant Data (Ex 2.1 in Montgomery et al.)

  A rocket motor is manufactured by bonding an igniter propellant and a sustainer propellant together inside of a metal housing. The shear strength of the bond between the two types of propellant is an important quality characteristic. It is suspected that shear strength is related to the age in weeks of the batch of sustainer propellant. Twenty observations on shear strength and the age of the corresponding batch of propellant have been collected and a part of the data is shown in the below table.

Observation	Shear Strength (psi)	Age of Propellant (weeks)
1	2158.70	15.50
2	1678.15	23.75
3	2316.00	8.00
4	2061.30	17.00
5	2207.50	5.50
⋮	⋮	⋮

  a. How are shear strength and propellant age related? How can we model how shear strength and propellant age are related?

  b. How well can we predict a shear strength of the bond between two types of propellant from knowing the age of sustainer propellant?

  c. What can we say about the relationship between shear strength and propellant age in the population?

  d. Which observations are unusual?

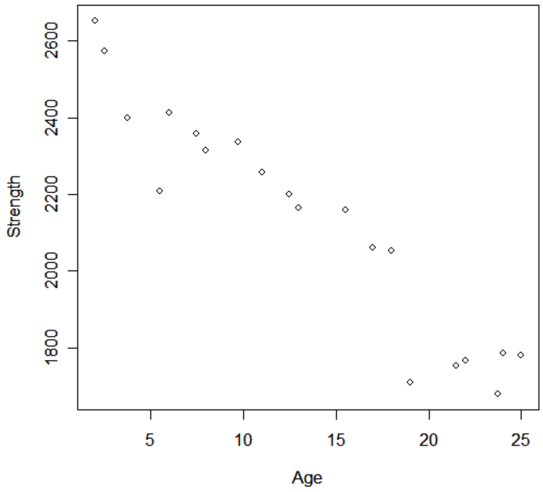


Figure 1.1: Scatter plot of shear strength against propellant age.

- Delivery Time Data (Ex 3.1 in Montgomery et al.):

  A soft drink bottler is analyzing the vending machine service routes in his distribution system. He is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet. This service activity includes stocking the machine with beverage products and minor maintenance or housekeeping. Two most important variables affecting the delivery time are the number of cases of product stocked and the distance walked by the route driver. For the study, twenty five observations were collected on delivery time.

Observation	Delivery Time (min)	Number of Cases	Distance (ft)
1	16.68	7	560
2	11.50	3	220
3	12.03	3	340
4	14.88	4	80
5	13.75	6	150
⋮	⋮	⋮	⋮

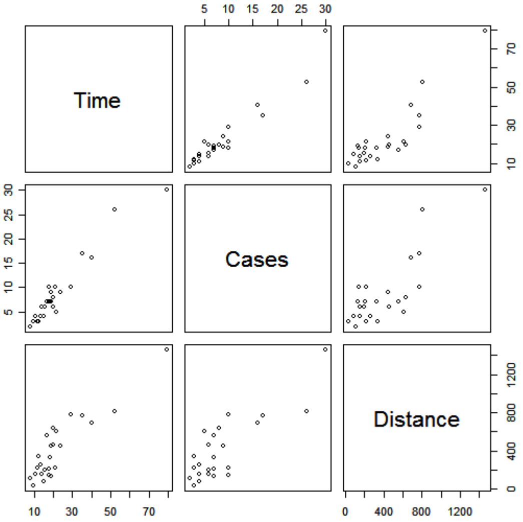


Figure 1.2: Scatter plot matrix for the delivery time data.

- Nutritional Requirement Data:

  Thirty one individuals were randomly assigned to one of three nitrogen intake levels (by controlling diet) and in each case nitrogen balance was measured. The data consist of caloric input (kcal), nitrogen intake (ni), an adjusted intake measure based on the protein source used (niq) and nitrogen balance (balance). The objective of the study is to develop a model for how nitrogen balance relates to intake and then determine an individual's nitrogen intake requirement at which the expected balance is 0.

Observations	kcal	ni	niq	balance
1	49.50	31.60	30.70	−22.70
2	50.60	32.40	31.40	−18.70
3	46.80	31.80	30.80	−17.40
4	46.00	31.90	30.90	−20.90
5	48.90	32.00	31.00	−21.60
⋮	⋮	⋮	⋮	⋮

Source of data: Kishi et al. (1978) Requirement and utilization of egg protein by Japanese young men with marginal intakes of energy. *Journal of Nutrition.*

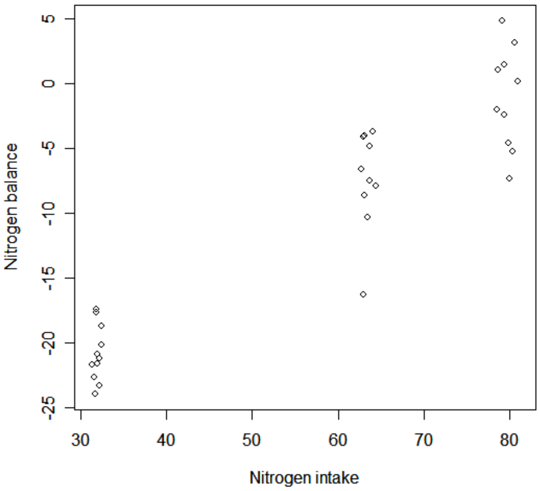


Figure 1.3: Scatter plot of nitrogen balance against intake.

- Italian Restaurant Menu Pricing Data

Suppose that your client, who plans to create a new Italian restaurant in Manhattan, wants to determine the price of the restaurant's dinner menu such that it is competitively positioned with other Italian restaurants in the target area. Data from surveys of customers of 168 Italian restaurants in the target area are available. The data are in the form of the average of customer views on

Y = Cost = the price (in US dollars) of dinner (including 1 drink and a tip),
X_1 = Food = customer rating of the food (out of 30),
X_2 = Décor = customer rating of the decor (out of 30),
X_3 = Service = customer rating of the service (out of 30),
X_4 = Location = 1 (0) if the restaurant is east (west) of Fifth Avenue.

Case	Restaurant	Cost	Food	Decor	Service	East
1	Daniella Ristorante	43	22	18	20	0
2	Tello's Ristorante	32	20	19	19	0
3	Biricchino	34	21	13	18	0
4	Bottino	41	20	20	17	0
5	Da Umberto	54	24	19	21	0
6	Le Madri	52	22	22	21	0
7	Le Zie	34	22	16	21	0
8	Pasticcio	34	20	18	21	1
9	Belluno	39	22	19	22	1
10	Cinque Terre	44	21	17	19	1
11	Fino	45	19	17	20	1
12	Marchi's	47	21	19	21	1
13	Nicola Paone	52	21	19	20	1
14	Notaro	35	19	17	19	1
15	Rossini's	47	20	18	21	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Source of data: *Zagat Survey 2001: New York City Restaurants*, Zagat, New York.

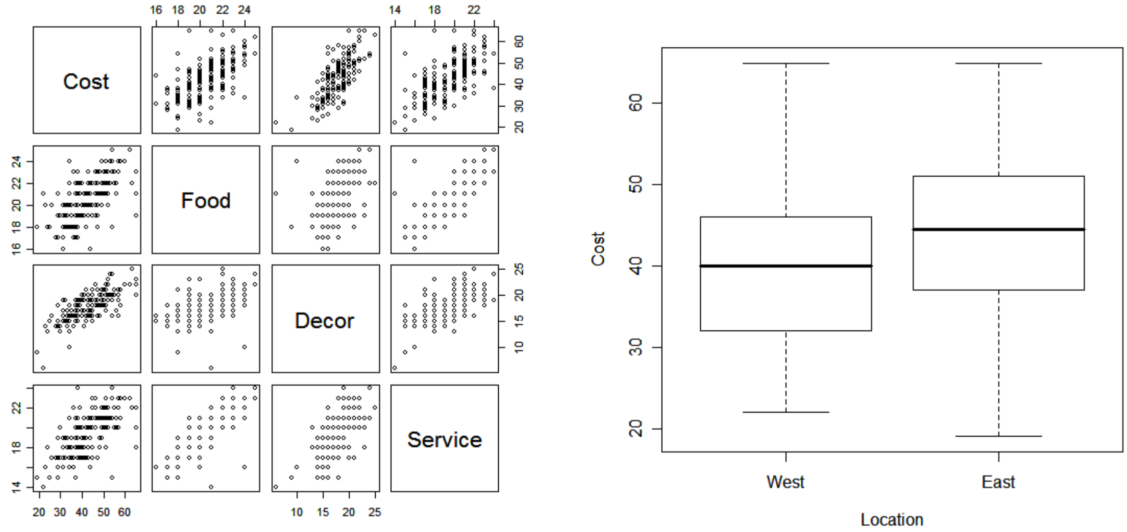
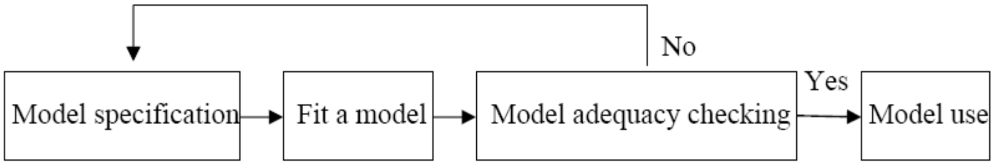


Figure 1.4: Scatterplot matrix for Italian restaurant menu pricing data and Box plots of cost for location.

Regression analysis process



Course contents

- Simple linear regression
- Multiple linear regression
- Model diagnostics
- Transformations and weighted least squares
- Variable selection.