1 Introduction: Linear Regression Models

1.1 Statistical Models

- A ***statistical model***:
  - ▷ Specifies the distribution of a (random) ***response variable*** $\boldsymbol{Y}$.
  - ▷ Describes a functional relationship between the **mean** (expected value) of Y and $\boldsymbol{p-1}$ ***predictor variables*** $\boldsymbol{X_1, X_2, \ldots, X_{p-1}}$ and $\boldsymbol{p}$ unknown ***parameters*** $\boldsymbol{\beta_0, \beta_1, \ldots, \beta_{p-1}}$:

$$\mu \;=\; \mathrm{E}(Y) \;=\; f(X_1, X_2, \ldots, X_{p-1};\; \beta_0, \beta_1, \ldots, \beta_{p-1})\,.$$

(The predictors are sometimes called ***covariates*** if they're numerical and ***factors*** if they're categorical.)

1.2 Linear Regression Models

- The simplest example of a statistical model is the ***simple linear regression model***, which has only **one predictor** variable X and **two parameters**.

Simple Linear Regression Model:

$$Y_i \;=\; \beta_0 + \beta_1 X_i + \epsilon_i\,,$$

where for $i = 1, 2, \ldots, n$:

- ▷ Y_i is the response for the ith individual.
- ▷ X_i is the value of a (numerical) predictor variable X for the ith individual.
- ▷ β_0 represents the intercept of the true regression line.
- ▷ β_1 represents the slope of the true regression line.
- ▷ ϵ_i is a random ***error***, with

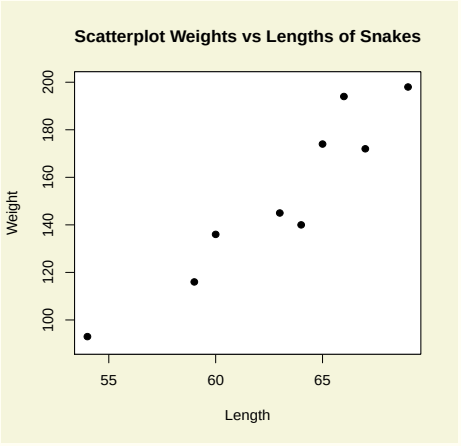
$$\epsilon_i \;\sim\; N(0,\, \sigma^2)\,,$$

and the ϵ_i's are assumed to be independent of each other, and therefore uncorrelated.

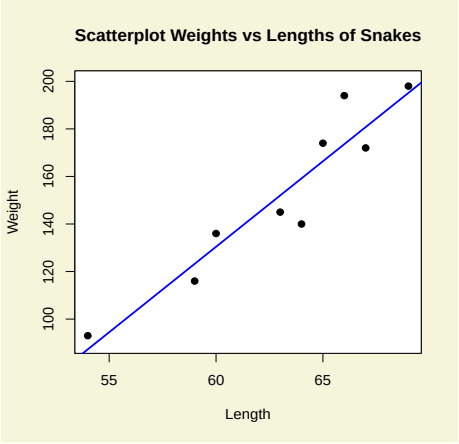
Example 1.1 The data and scatterplot below show the lengths (X) and weights (Y) of nine female snakes.

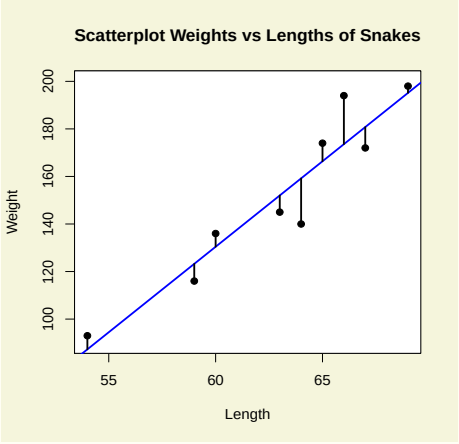
Lengths and Weights of Female Snakes

Snake	Length (cm)	Weight (g)
1	60	136
2	69	198
3	66	194
4	64	140
5	54	93
6	67	172
7	59	116
8	65	174
9	63	145



We can fit a line to the points using the method of *least squares* described later. This so-called *fitted regression line* is shown in the plots below.





- For the *simple linear regression model*:
  - ▷ The data are stored in the format:

Observation	Predictor Variable X	Response Variable Y
1	X_1	Y_1
2	X_2	Y_2
$\vdots$	$\vdots$	$\vdots$
n	X_n	Y_n

- The observed values $X_1, X_2, \ldots, X_n$ of the **predictor** variable are usually considered to be **non-random** (i.e. hand-picked), so we don't have to model variation in these values.

  (But all of the regression procedures in this class *can* still be carried out when the X's are random.).
- The **variance** σ^2 of the error term ϵ is **constant** (doesn't depend on X).
- The **mean** of the **response** variable Y is a **linear** function of X:

$$\mu_i \;=\; \mathrm{E}(Y_i) \;=\; \beta_0 + \beta_1 X_i \,,$$

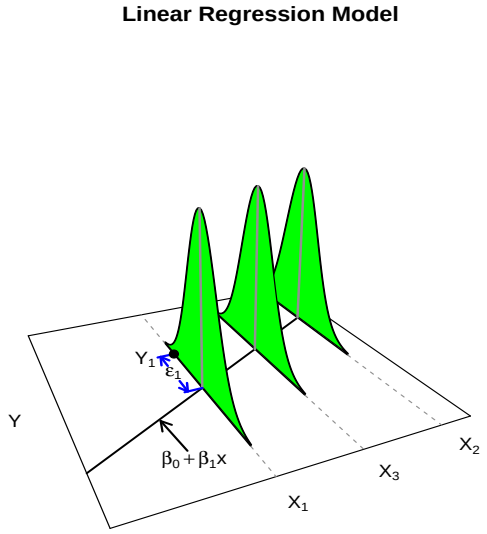
  but its **variance** is **constant** (doesn't depend on X):

$$\mathrm{Var}(Y_i) \;=\; \mathrm{Var}(\beta_0 + \beta_1 X_i + \epsilon_i) \;=\; \mathrm{Var}(\epsilon_i) \;=\; \sigma^2 \,.$$

- The observed **responses** $Y_1, Y_2, \ldots, Y_n$ are **independent** and **normally distributed**:

$$Y_i \;\sim\; N(\beta_0 + \beta_1 X_i,\; \sigma^2) \,.$$

- $\mathrm{E}(Y)$ is also linear in the *parameters* β_0 and β_1 (i.e. it can be written as a linear combination of β_0 and β_1).



- The ***multiple linear regression model*** is an extension of the simple linear regression model that includes $p-1$ predictor variables.

Multiple Linear Regression Model:

$$Y_i \;=\; \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{ip-1} + \epsilon_i,$$

where for $i = 1, 2, \ldots, n$:

- ▷ Y_i is the response for the ith individual.
- ▷ $X_{i1}, X_{i2}, \ldots, X_{ip-1}$ are the values of $p-1$ (numerical) predictor variables for the ith individual.
- ▷ β_0 represents an intercept term of the true regression model mean response.
- ▷ $\beta_1, \ldots, \beta_{p-1}$ represent coefficients of the true regression model mean response.
- ▷ ϵ_i is a random ***error***, with

$$\epsilon_i \;\sim\; N(0,\, \sigma^2)\,,$$

and the ϵ_i's are assumed to be independent of each other, and therefore uncorrelated.

Example 1.2 Streams and lakes contain dissolved oxygen that supports fish and other aquatic life. But organic pollutants consume dissolved oxygen when they chemically degrade via oxidation.

The chemical oxygen demand (COD) of a water supply is the amount of oxygen that would be needed to chemically degrade via oxidation the organic compounds contained in the water. It's used an indirect measure of organic pollution.

The data below are COD measurements in stormwater runoff for each of $n = 18$ rainfall events in the Pear River Delta, South China. Also reported are the rain depth and the length of the antecedent dry period.

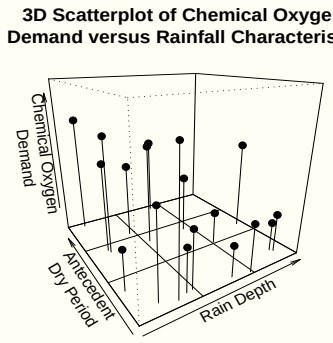
Chemical Oxygen Demand in Stormwater Runoff

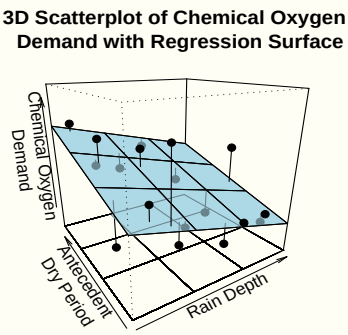
Rainfall Date	COD (mg/L)	Rain Depth (mm)	Antecedent Dry Period (days)
9/25/05	296	4.1	0.83
2/26/06	256	9.0	7.83
3/22/06	518	6.7	0.43
4/06/06	451	3.6	6.38
4/23/06	469	2.9	10.57
5/17/06	323	3.4	6.66
6/15/06	161	9.6	0.73
7/15/06	336	28.6	4.87
7/16/06	119	19.6	0.16
7/25/06	379	10.9	6.22
7/26/06	75	27.2	0.98
4/06/06	177	18.5	6.35
4/23/06	295	17.6	10.57
5/02/06	93	8.5	0.68
5/06/06	29	26.8	1.17
5/10/06	46	20.3	3.81
5/31/06	37	16.3	0.38
6/12/06	77	1.3	2.54

We'll consider **COD** to be the **response variable** and investigate its relationship to the **two predictors**, **rain depth** and **antecedent dry period**:

Y = COD
X_1 = Rain Depth
X_2 = Antecedent Dry Period

We can inspect this relationship visually using a **three-dimensional scatterplot** as shown below left.





We can fit a plane to the points using the method of *least squares* described later. This so-called *fitted multiple regression plane* is shown in the plot above right.

- For the *multiple regression model*:
  - ▷ The data are stored in the format:

Individual Observation	1st Predictor Variable X_1	2nd Predictor Variable X_2	$\cdots$	$p-1$st Predictor Variable X_{p-1}	Response Variable Y
1	X_{11}	X_{12}	$\cdots$	X_{1p-1}	Y_1
2	X_{21}	X_{22}	$\cdots$	X_{2p-1}	Y_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
n	X_{n1}	X_{n2}	$\cdots$	X_{np-1}	Y_n

  - ▷ The observed values

$$\begin{matrix} X_{11}, & X_{12}, & \ldots, & X_{1p-1} \\ X_{21}, & X_{22}, & \ldots, & X_{2p-1} \\ & & \vdots & \\ X_{n1}, & X_{n2}, & \ldots, & X_{np-1} \end{matrix}$$

    of the **predictor** variables are usually considered to be **non-random** (i.e. hand-picked), so we don't have to model variation in these values.
  - ▷ The **variance** σ^2 of the error term ϵ is **constant** (doesn't depend on the predictors $X_1, X_2, \ldots, X_{p-1}$).

▷ The **mean** of the **response** variable Y is a **linear** function of the predictors $X_1, X_2, \ldots, X_{p-1}$:

$$\mu_i \;=\; \mathrm{E}(Y_i) \;=\; \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip-1} \qquad (1)$$

but its **variance** is **constant** (doesn't depend on $X_1, X_2, \ldots, X_{p-1}$):

$$\mathrm{Var}(Y_i) \;=\; \mathrm{Var}(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip-1} + \epsilon_i) \;=\; \mathrm{Var}(\epsilon_i) \;=\; \sigma^2\,.$$

▷ The observed **responses** $Y_1, Y_2, \ldots, Y_n$ are **independent** and **normally distributed**:

$$Y_i \;\sim\; N(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{ip-1},\; \sigma^2)\,.$$

▷ The mean response (1) is also **linear** in the ***parameters*** $\beta_0, \beta_1, \ldots, \beta_p$ (i.e. it can be written as a linear combination of $\beta_0, \beta_1, \ldots, \beta_p$).

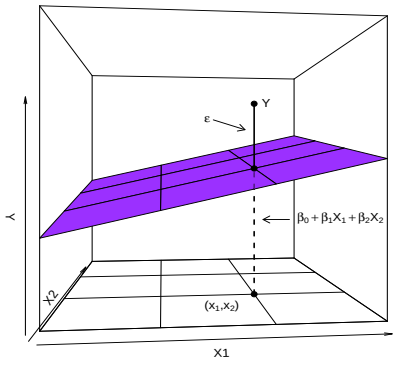


Figure 1: Depiction of the multiple regression model $Y = \beta_0 + \beta_1\, X_1 + \beta_2\, X_2 + \epsilon$.

2 Simple Linear Regression

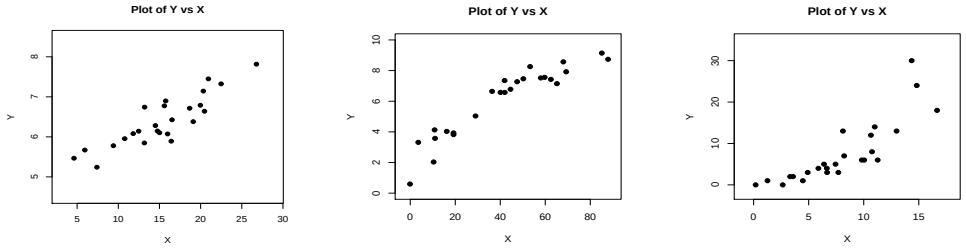
2.1 Preliminary Analysis

- **Plot the data first**.

  ▷ If linear with constant variance, proceed with simple linear regression analysis.

  ▷ If nonlinear but with constant variance, consider transforming X to make it linear or using a different model (e.g. polynomial regression).

▷ If nonlinear and with non-constant variance, consider transforming Y (and perhaps also X) or using a different model (e.g. a Poisson regression model).



2.2 Estimation of Parameters

- ***Least squares estimates*** $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$ of the unknown parameters β_0 and β_1 are the values that minimize

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

as a function of β_0 and β_1.

- To find b_0 and b_1, solve the ***normal equations***:

$$\begin{aligned}\frac{\partial Q}{\partial \beta_0} &= -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) \;=\; 0\\ \frac{\partial Q}{\partial \beta_1} &= -2\sum_{i=1}^{n} X_i(Y_i - \beta_0 - \beta_1 X_i) \;=\; 0\end{aligned}$$

- The solution to the normal equations gives the least squares estimates:

Least Squares Estimates of $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$:

$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\\ b_0 &= \bar{Y} - b_1\bar{X}\end{aligned}$$

- The resulting line is called the ***least squares regression line*** (or ***fitted regression line***).

Fitted Regression Line:

$$\hat{Y} \;=\; b_0 + b_1 X \tag{2}$$

- Facts about the least squares estimators:
  - ▷ b_0 and b_1 are ***unbiased*** estimators of β_0 and β_1, i.e.

$$\begin{aligned} \mathrm{E}(b_0) &= \beta_0 \\ \mathrm{E}(b_1) &= \beta_1 \end{aligned}$$

  - ▷ Both b_0 and b_1 are **linear combinations of the Y_i's** since it can be shown that b_1 can be written as

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

    where $k_i = (X_i - \bar{X}) / \sum_i (X_i - \bar{X})^2$, and b_0 can be written as

$$b_0 \;=\; \sum_{i=1}^n c_i Y_i$$

    (find the c_i's yourself!).
  - ▷ Among all unbiased estimators of β_0 and β_1 that are linear combinations of the Y_i's, b_0 and b_1 have the smallest variances (i.e. they're the most precise).
  - ▷ The fitted regression line always **passes through $(\bar{X}, \bar{Y})$** (verify this yourself!).

2.3 Point Estimation of E(Y)

- Recall that for a given value of X,

$$\mathrm{E}(Y) \;=\; \beta_0 + \beta_1 X \,.$$

  Thus plugging X into the fitted regression line

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

  gives a **point estimate of E(Y)** for that X value.

- The ***fitted values*** (or ***predicted values***), denoted $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n$, are the estimates of E(Y_i) for the *observed* predictor values $X_1, X_2, \ldots, X_n$:

Fitted (Predicted) Values:

$$\hat{Y}_i = b_0 + b_1 X_i \qquad \text{for } i = 1, 2, \ldots, n$$

The fitted values **lie on the fitted regression line**.

2.4 Residuals

- The ***residuals***, denoted $\boldsymbol{e_1, e_2, \ldots, e_n}$, are defined as

Residuals:

$$e_i = Y_i - \hat{Y}_i \qquad \text{for } i = 1, 2, \ldots, n$$

The residuals represent **vertical deviations** of the Y_i's **away from the fitted regression line**.

- Properties of residuals and fitted values:
  - $\sum_i \hat{Y}_i = \sum_i Y_i$ and so $\frac{1}{n}\sum_i \hat{Y}_i = \bar{Y}$
  - $\sum_i e_i = 0$
  - $\sum_i X_i e_i = 0$
  - $\sum_i \hat{Y}_i e_i = 0$

2.5 Point Estimation of σ^2

- The ***error sum of squares*** (or ***sum of squared residuals***) is denoted **SSE** and defined as

Error Sum of Squares:

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

- The n residuals squared and summed in SSE are subject to the **two constraints** $\sum_i e_i = 0$ and $\sum_i X_i e_i = 0$, so only $\boldsymbol{n - 2}$ of them are **"free to vary"**.

Thus the ***degrees of freedom*** assosiated with SSE is:

Error Degrees of Freedom:

$$\text{df for SSE} \;=\; n-2$$

The degrees of freedom for SSE is also the **number of observations minus the number of unknown parameters in the model**.

- A ***mean square*** is defined as a sum of squares divided by its degrees of freedom.

  Thus the ***mean squared error***, denoted **MSE**, is:

Mean Squared Error:

$$\text{MSE} \;=\; \frac{\text{SSE}}{n-2}$$

We'll sometimes also denote the mean squared error by $\boldsymbol{s^2}$, i.e. $\boldsymbol{s^2 = \text{MSE}}$.

- It can be shown that the MSE is an **unbiased estimator of $\boldsymbol{\sigma^2}$**, i.e.

$$\text{E(MSE)} \;=\; \sigma^2\,.$$

Thus **we estimate $\boldsymbol{\sigma}$ by $\boldsymbol{\sqrt{\text{MSE}}}$**.