Simple linear regression: I. introduction

Miaoyan Wang

Department of Statistics
UW Madison

Simple linear regression

References:

- Chapter 2 in JF (Julian J. Faraway)
- Chapter 2.1-2.9, 2.11 in RC (Ronald Christensen)

Both textbooks are available in Canvas/files/textbook/

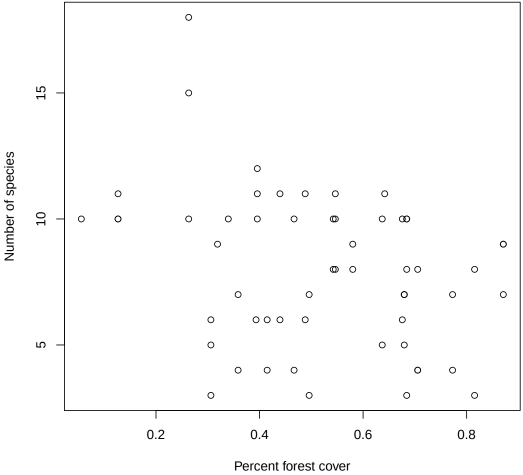
Example: Wetland Species Richness

- A study was performed on insect species richness in 58 wetlands in Ontario, Canada.
- The goal of the study was to determine the relationship between forest density around the wetland and insect species richness.
- The investigators sample insects in each wetland and then recorded the number of species present in each sample.
- The percent forest cover within a 1500-meter buffer around the wetland was also recorded, among other wetland characteristics.

Example: Wetland Species Richness

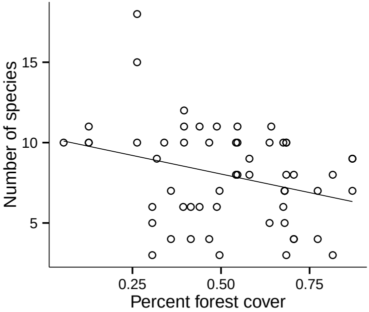
wetland	y	x	wetland	y	x
1	10	0.056	30	5	0.637
2	8	0.546	31	6	0.488
3	10	0.637	32	9	0.580
4	8	0.815	33	4	0.705
5	10	0.676	34	11	0.439
6	9	0.871	35	8	0.705
7	4	0.467	36	5	0.680
8	3	0.684	37	10	0.396
9	3	0.496	38	10	0.467
10	4	0.415	39	5	0.306
11	7	0.680	40	10	0.684
12	7	0.773	41	6	0.415
13	9	0.319	42	10	0.684
14	10	0.127	43	10	0.340
15	3	0.306	44	7	0.871
16	6	0.676	45	9	0.871
17	8	0.684	46	7	0.680
18	10	0.546	47	18	0.263
19	10	0.542	48	12	0.396
20	15	0.263	49	6	0.306
21	11	0.488	50	4	0.359
22	7	0.359	51	6	0.439
23	7	0.680	52	8	0.542
24	6	0.393	53	4	0.705
25	4	0.773	54	11	0.127
26	3	0.815	55	7	0.496
27	11	0.642	56	10	0.263
28	8	0.580	57	10	0.127
29	11	0.396	58	11	0.546

Example: Wetland Species Richness



Specific Goals

- To describe the relationship between the percent forest cover (x) and the number of species (y).
- To estimate or predict the number of species for a given percent forest cover.
- Q: How to account for uncertainty in the fitted line and variation?



Modeling Idea

- Model y by a random variable Y.
- Regard x as fixed, or condition on x (x could be modeled by a random variable X.)
- Consider the model of Y conditional on $X = x$:

$$E(Y|X = x) = \beta_0 + \beta_1 x.$$

- β_0, β_1 are fixed unknown parameters (i.e., the intercept and slope) characterizing the relationship between X and Y.

Simple Linear Regression Model

The formal simple linear regression (SLR) model for the data (x_i, y_i) is:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

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

- Y_i is the ith **response variable**.
- X_i is the ith **explanatory variable** (also called predictors, covariates).
- ε_i is the ith **random error** term.
- The random errors follow a normal distribution with mean zero and variance σ^2 and are independent of each other.
- That is, $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$.
  iid = independently and identically distributed

Features of Simple Linear Regression Model

Under the SLR model for the data (x_i, y_i):

- Simple one explanatory variable only
- Linear parameters enter the model linearly.
- Regression Galton: taller fathers tend to have shorter sons; regression toward the mean
- Randomness Q: What kind of distribution does Y_i have?
- Independence The random errors are independent and thus the response variables are (conditionally) independent.
  Q: What kind of independence?
  Q: What kind of dependence?
- The model parameters are: $\beta_0, \beta_1, \sigma^2$.

Model Assumptions

- A straight line relationship between the response variable Y and the explanatory variable X:

$$E(Y_i|X_i) = \beta_0 + \beta_1 x_i.$$

- Equal variance:

$$Var(Y_i|X_i) = \sigma^2.$$

- Independence (conditional on X_i, X_i'):

$$Cov(Y_i, Y_{i'}) = 0 \quad \text{for} \quad i \neq i'.$$

- Normal distribution:

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

Equivalent Model Assumptions

Equivalently, the assumptions are

- A straight line relationship between the response variable Y and the explanatory variable X:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{where} \quad E(\varepsilon_i) = 0$$

- Equal variance:

$$Var(\varepsilon_i) = \sigma^2.$$

- Independence:

$$Cov(\varepsilon_i, \varepsilon_{i'}) = 0 \quad \text{for} \quad i \neq i'.$$

- Normal distribution:

$$\varepsilon_i \sim N(0, \sigma^2).$$

Q: ε_i are iid. How about Y_i? iid? Not iid? It depends?

Model Parameters

- The model parameters are β_0, β_1, and σ^2 (population parameters).
- β_0 and β_1: **regression coefficients**.
- β_0: **intercept**.
  When the model scope includes $x = 0$, β_0 can be interpreted as the mean of Y at $x = 0$.
- β_1: **slope**.
  Interpreted as the change in the mean of Y per unit increase in x.
- σ^2: **error variance**, sometimes written as σ_ε^2 or $\sigma_{Y|x}^2$.

Q: How to estimate the model parameters based on data?

Estimation of Model Parameters

- Our goal is to estimate these model parameters by estimators $\hat{\beta}_0, \hat{\beta}_1$, and $\hat{\sigma}^2$, based on data.
- Two methods:
  - Least squares (LS).
  - Maximum likelihood (ML).