STAT 331 - Applied Linear Models

Cameron Roopnarine

Last updated: September 8, 2020

Contents

Lecture 1 | 2020-09-08

Regression model infers the relationship between:

- Response (dependent) variable: variable of primary interest, denoted by a capital letter such as Y.
- Explanatory (independent) variables: (covariates, predictors, features) variables that potentially impact response, denoted $(x_1, x_2, \dots, x_p)$.

Alligator data:

- length (m) Y
- male/female (categorical, 0 or 1) x_1

Mass in stomach:

- fish x_2
- invertebrates x_3
- reptiles x_4
- birds x_5
- other x_6

We imagine we can explain Y in terms of $(x_1, \dots, x_p)$ using some function so that $Y = f(x_1, \dots, x_p)$.

In this course, we will be looking at linear models.

Linear regression model assumes that

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

- Y value of response
- $x_1, \dots, x_p$ values of p explanatory variables (assumed to be fixed constants)
- $\beta_0, \beta_1, \dots, \beta_p$ model parameters
  - β_0 intercept, expected value of Y when all $x_j = 0$.
  - $\beta_1, \dots, \beta_p$ quantify effect on x_j on Y, $j = 1, \dots, p$
  - ε random error "all models are wrong, but some are useful"

Assume $\varepsilon \sim N(0, \sigma^2)$. In general, the model will not perfectly explain the data.

Q: What is the distribution of Y under these assumptions?

$\mathbf{E}[Y] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

$\mathbf{Var}[Y] = \mathbf{Var}[\varepsilon] = \sigma^2$.

$Y \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \sigma^2)$