

STAT 331 - Applied Linear Models

Cameron Roopnarine

Last updated: September 9, 2020

Contents

Contents	1
0.1 Simple linear regression	2

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

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

- 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 $x_j = 0$.
 - β_1, \dots, β_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?

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

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

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

LECTURE 2 | 2020-09-09

0.1 Simple linear regression

A linear model with response variable (Y) and *one* explanatory variable (x); that is,

$$\bar{Y} = \beta_0 + \beta_1 x + \varepsilon$$

Data consists of pairs (x_i, y_i) where $i = 1, \dots, n$.

Before fitting any model, we might

- make a scatterplot to visualize if there is a linear relationship between x and y
- calculate *correlation*

If X and Y are random variables, then

$$\rho = \text{Corr}[X, Y] = \frac{\text{Cov}[X, Y]}{sd(x)sd(y)}$$

Based on (x_i, y_i) , the estimated correlation is

$$\begin{aligned} r &= \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \end{aligned}$$

Measures the strength and direction of *linear* relationship between x and y .

- $|r| \approx 1$ strong linear relationship
- $|r| \approx 0$ lack of linear relationship
- $r > 0$ positive relationship
- $r < 0$ negative relationship
- $-1 \leq r \leq 1$

But does not tell us how to predict Y from x . To do so, we need to estimate β_0 and β_1 .

For data (x_i, y_i) for $i = 1, \dots, n$, the simple linear regression model is

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

Assume $\varepsilon \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ Therefore, $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. In other words, $\mathbf{E}[Y_i] = \mu_i = \beta_0 + \beta_1 x_i$ and $\text{Var}[Y_i] = \sigma^2$. Note that the Y_i 's are independent, but they are *not* independently distributed.

Use the *Least Squares* (LS) to estimate β_0 and β_1 .

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2 = S(\beta_0, \beta_1)$$

LS is equivalent to MLE when ε_i 's are iid and Normal.

Taking partial derivatives:

$$\begin{aligned} \frac{dS}{d\beta_0} &= 2 \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)] (-1) \\ \frac{dS}{d\beta_1} &= 2 \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)] (-x_i) \end{aligned}$$

Now,

$$\frac{dS}{d\beta_0} = 0 \iff \sum_{i=1}^n y_i - n\beta_0 - \beta_1 \sum_{i=1}^n x_i = 0 \iff \beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\begin{aligned} \frac{dS}{d\beta_1} = 0 &\stackrel{\text{plug } \beta_0}{\iff} \sum_{i=1}^n [y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i] x_i = 0 \\ &\iff \sum_{i=1}^n x_i (y_i - \bar{y}) - \beta_1 \sum_{i=1}^n x_i (x_i - \bar{x}) = 0 \\ &\iff \beta_1 = \frac{\sum_{i=1}^n x_i (y_i - \bar{y})}{\sum_{i=1}^n x_i (x_i - \bar{x})} \end{aligned}$$

We can also show that

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Notationally, we use hats to show that they are estimates

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

Call $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ the **fitted values** and $e_i = y_i - \hat{\mu}_i$ the **residual**.