Applied Linear Algebra in Data Analysis

Introduction to Linear Regression

Sivakumar Balasubramanian

Department of Bioengineering
Christian Medical College, Bagayam
Vellore 632002

What is regression

- Regression is a method to summarize how average values of an *outcome* vary across the levels of one or more *predictor* variables.

- **Linear regression** is a special case of regression where the relationshio between the outcome and the predictor variables is assumed to be linear (affine to be accurate).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_N x_N + \epsilon$$

Why learn regression?

There are several important uses of regresion models:

1. **Prediction**
2. **Exploring associations**
3. **Extrapolation**
4. **Causal inference**

Why learn regression? – Prediction

Prediction: Given a set of predictor variables, we can predict the outcome variable.

- Predicting the level of impairment for a patient at 6 months, given their the initial impairment level in the first week post-stroke, severity of stroke, and other clinical and demographics features.
- Predicting the probability of malignnancy from the radiological features of a tumor.
- Predicting the number a job offers student will have in their final semester based on their GPA, type of their project, and communcation skills.
- Predicting survival time for a patient with a particular type of cancer based on the tumor malignancy, and other clinically relevant features.

Why learn regression? – Exploring associations

Exploring associations: Given a set of predictor variables, we can predict the outcome variable.

- Predicting the level of impairment for a patient at 6 months, given their the initial impairment level in the first week post-stroke, severity of stroke, and other clinical and demographics features.
- Predicting the probability of malignnancy from the radiological features of a tumor.
- Predicting the number a job offers student will have in their final semester based on their GPA, type of their project, and communcation skills.
- Predicting survival time for a patient with a particular type of cancer based on the tumor malignancy, and other clinically relevant features.

Why learn regression? – Extrapolation

Extrapolation: Given a set of predictor variables, we can predict the outcome variable.

- Predicting the level of impairment for a patient at 6 months, given their the initial impairment level in the first week post-stroke, severity of stroke, and other clinical and demographics features.
- Predicting the probability of malignnancy from the radiological features of a tumor.
- Predicting the number a job offers student will have in their final semester based on their GPA, type of their project, and communcation skills.
- Predicting survival time for a patient with a particular type of cancer based on the tumor malignancy, and other clinically relevant features.

Why learn regression? – Causal inference

Causal inference: Given a set of predictor variables, we can predict the outcome variable.

- Predicting the level of impairment for a patient at 6 months, given their the initial impairment level in the first week post-stroke, severity of stroke, and other clinical and demographics features.
- Predicting the probability of malignnancy from the radiological features of a tumor.
- Predicting the number a job offers student will have in their final semester based on their GPA, type of their project, and communcation skills.
- Predicting survival time for a patient with a particular type of cancer based on the tumor malignancy, and other clinically relevant features.

Fitting a Simple Regression Model - With Fake Data

Consider the following simple regression model,

$$y = \beta_0 + \beta_1 \cdot x + \epsilon$$

where,

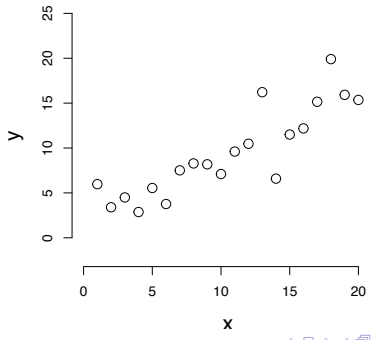
- y is the outcome variable (dependent variable)
- x is the predictor variable (independent variable)
- β_0 is the intercept (the predicted value of y when $x = 0$)
- β_1 is the slope (the change in y for a one-unit change in x)
- ϵ is the error term (the difference between the observed and predicted values of y)

β_0 and β_1 are called the *parameters* of the regression model.

Fitting a Simple Regression Model - With Fake Data

Let's generate fake data for some assumed values of β_0 and β_1, which produces N data points of the form $(x_i, y_i)_{i=1}^{N}$.

```
x <- 1:20
n <- length(x)
b0 <- 0.5
b1 <- 0.8
sigma <- 2.5
y <- b0 + b1 * x + sigma * rnorm(n)
```



Fitting a Simple Regression Model - With Fake Data

We will use the 'statsmodel' library to fit a simple linear regression model to the data we generated.

```
import statsmodels.api as sm

X = sm.add_constant(x)
model = sm.OLS(y, X).fit()
model.summary()
```

The following scatter plot shows an instance of the data generated by the above code.