Lecture 1

August 23, 2018

The Regression Problem

- This class deals with the regression problem where the goal is to understand the relationship between a response variable and one or more explanatory variables.

The Regression Problem

- This class deals with the regression problem where the goal is to understand the relationship between a response variable and one or more explanatory variables.
- The response variable (also known as the dependent variable) is denoted by y and the explanatory variables (also known as independent variables or features) are denoted by $x_1, \dots, x_p$.

The Regression Problem

- This class deals with the regression problem where the goal is to understand the relationship between a response variable and one or more explanatory variables.
- The response variable (also known as the dependent variable) is denoted by y and the explanatory variables (also known as independent variables or features) are denoted by $x_1, \dots, x_p$.
- When y is a categorical or discrete variable, this problem is also called the classification problem.

The Regression Problem

- This class deals with the regression problem where the goal is to understand the relationship between a response variable and one or more explanatory variables.
- The response variable (also known as the dependent variable) is denoted by y and the explanatory variables (also known as independent variables or features) are denoted by $x_1, \ldots, x_p$.
- When y is a categorical or discrete variable, this problem is also called the classification problem.
- For example, y might be a person's hourly wage, x_1 is the number of years of education, and x_2 is the number of years of work experience.

Regression Data

- We will have n subjects and data on the variables (y and $x_1, \ldots, x_p$) are collected from each of these subjects.

Regression Data

- We will have n subjects and data on the variables (y and $x_1, \ldots, x_p$) are collected from each of these subjects.
- The value of the response for the i-th subject is denoted by y_i. The value of the explanatory variable x_j for the ith subject is denoted by x_{ij}.

Regression Data

- We will have n subjects and data on the variables (y and $x_1, \ldots, x_p$) are collected from each of these subjects.
- The value of the response for the i-th subject is denoted by y_i. The value of the explanatory variable x_j for the ith subject is denoted by x_{ij}.
- The observations $(y_i, x_{i1}, \ldots, x_{ip})$ corresponding to the i-th subject are assumed to be independent for $i = 1 \ldots, n$.

Objectives of Regression

- There are two main objectives in a regression problem:

Objectives of Regression

- There are two main objectives in a regression problem:
  1. To predict the response variable based on the explanatory variables.

Objectives of Regression

- There are two main objectives in a regression problem:
  1. To predict the response variable based on the explanatory variables.
  2. To identify which among the explanatory variables are related to the response variable and to explore the forms of these relationships.

Objectives of Regression

- There are two main objectives in a regression problem:
  1. To predict the response variable based on the explanatory variables.
  2. To identify which among the explanatory variables are related to the response variable and to explore the forms of these relationships.
- Mathematically, the objective of regression analysis is to understand the conditional distribution of y given $x_1, \ldots, xp$. Often one is only interested in the conditional mean: $\mathbb{E}(y|x_1, \ldots, x_p)$.

Objectives of Regression

- There are two main objectives in a regression problem:
  1. To predict the response variable based on the explanatory variables.
  2. To identify which among the explanatory variables are related to the response variable and to explore the forms of these relationships.
- Mathematically, the objective of regression analysis is to understand the conditional distribution of y given $x_1, \ldots, xp$. Often one is only interested in the conditional mean: $\mathbb{E}(y|x_1, \ldots, x_p)$.
- It must be noted that the conditional mean $\mathbb{E}(y|x_1, \ldots, x_p)$ is not a single number but a function of $x_1, \ldots, x_p$. It is therefore called the regression function.

Objectives of Regression

- There are two main objectives in a regression problem:
  1. To predict the response variable based on the explanatory variables.
  2. To identify which among the explanatory variables are related to the response variable and to explore the forms of these relationships.
- Mathematically, the objective of regression analysis is to understand the conditional distribution of y given $x_1, \dots, xp$. Often one is only interested in the conditional mean: $\mathbb{E}(y|x_1, \dots, x_p)$.
- It must be noted that the conditional mean $\mathbb{E}(y|x_1, \dots, x_p)$ is not a single number but a function of $x_1, \dots, x_p$. It is therefore called the regression function.
- For example, in the wage example, learning this conditional mean would involve learning all the values of $\mathbb{E}(y|x_1, \dots, x_p)$.

Objectives of Regression

- There are two main objectives in a regression problem:
  1. To predict the response variable based on the explanatory variables.
  2. To identify which among the explanatory variables are related to the response variable and to explore the forms of these relationships.
- Mathematically, the objective of regression analysis is to understand the conditional distribution of y given $x_1, \dots, xp$. Often one is only interested in the conditional mean: $\mathbb{E}(y|x_1, \dots, x_p)$.
- It must be noted that the conditional mean $\mathbb{E}(y|x_1, \dots, x_p)$ is not a single number but a function of $x_1, \dots, x_p$. It is therefore called the regression function.
- For example, in the wage example, learning this conditional mean would involve learning all the values of $\mathbb{E}(y|x_1, \dots, x_p)$.
- In general, the regression function $\mathbb{E}(y|x_1, \dots, x_p)$ would involve an infinite number of parameters!

The Linear Model for Regression

- Usually the conditional mean is too complicated an object to learn from data. Therefore, one tries to model using fewer parameters.

The Linear Model for Regression

- Usually the conditional mean is too complicated an object to learn from data. Therefore, one tries to model using fewer parameters.

The Linear Model for Regression

- Usually the conditional mean is too complicated an object to learn from data. Therefore, one tries to model using fewer parameters.
- The simplest such model is the linear model.

The Linear Model for Regression

- Usually the conditional mean is too complicated an object to learn from data. Therefore, one tries to model using fewer parameters.
- The simplest such model is the linear model.
- The linear model for regression mainly stipulates that $\mathbb{E}(y|x_1, \ldots, x_p)$ is a linear function of $x_1, \ldots, x_p$. More precisely, one assumes that

$$\mathbb{E}(y|x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p. \tag{1}$$

The Linear Model for Regression

- Usually the conditional mean is too complicated an object to learn from data. Therefore, one tries to model using fewer parameters.
- The simplest such model is the linear model.
- The linear model for regression mainly stipulates that $\mathbb{E}(y|x_1, \ldots, x_p)$ is a linear function of $x_1, \ldots, x_p$. More precisely, one assumes that

$$\mathbb{E}(y|x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p. \tag{1}$$

- Note that this implies that the conditional mean is describable by only p + 1 parameters $\beta_0, \beta_1, \ldots, \beta_p$.

The Linear Model for Regression

- Usually the conditional mean is too complicated an object to learn from data. Therefore, one tries to model using fewer parameters.
- The simplest such model is the linear model.
- The linear model for regression mainly stipulates that $\mathbb{E}(y|x_1, \ldots, x_p)$ is a linear function of $x_1, \ldots, x_p$. More precisely, one assumes that

$$\mathbb{E}(y|x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p. \tag{1}$$

- Note that this implies that the conditional mean is describable by only p + 1 parameters $\beta_0, \beta_1, \ldots, \beta_p$.
- For $j = 1, \ldots, p$, the parameter β_j is interpreted as the increase in the mean of the response variable per unit increase in the value of the jth explanatory variable when all the remaining explanatory variables x_k, $k \neq j$ are kept fixed.

The Linear Model for Regression

- The parameter β_0 is called the intercept.

The Linear Model for Regression

- The parameter β_0 is called the intercept.
- The linear model is often times rewritten in the following way. Suppose that

$$e := y - \mathbb{E}(y|x_1, \ldots, x_p)$$

The Linear Model for Regression

- The parameter β_0 is called the intercept.
- The linear model is often times rewritten in the following way. Suppose that

$$e := y - \mathbb{E}(y|x_1, \ldots, x_p)$$

- Then of course $y = \mathbb{E}(y|x_1, \ldots, x_p) + e$. Because of (1), we can then write

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + e, \text{ with } \mathbb{E}(e|x_1, \ldots, x_p) = 0. \quad (2)$$

The Linear Model for Regression

- The parameter β_0 is called the intercept.
- The linear model is often times rewritten in the following way. Suppose that

$$e := y - \mathbb{E}(y|x_1, \ldots, x_p)$$

- Then of course $y = \mathbb{E}(y|x_1, \ldots, x_p) + e$. Because of (1), we can then write

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + e, \text{ with } \mathbb{E}(e|x_1, \ldots, x_p) = 0. \quad (2)$$

- The random variable e is called the error (interpreted as noise). The equation (2) therefore says that the value of the response variable for *i*-th subject equals a linear combination of its explanatory variable values give or take some noise. Hence the name linear model.

The Linear Model for Regression

- The parameter β_0 is called the intercept.
- The linear model is often times rewritten in the following way. Suppose that

$$e := y - \mathbb{E}(y|x_1, \ldots, x_p)$$

- Then of course $y = \mathbb{E}(y|x_1, \ldots, x_p) + e$. Because of (1), we can then write

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + e, \text{ with } \mathbb{E}(e|x_1, \ldots, x_p) = 0. \quad (2)$$

- The random variable e is called the error (interpreted as noise). The equation (2) therefore says that the value of the response variable for *i*-th subject equals a linear combination of its explanatory variable values give or take some noise. Hence the name linear model.
- Simple linear regression refers to the situation $p = 1$. Here there is only one explanatory variable which is denoted by x

First Part of this class

First Part of this class

1. How to obtain estimates of $\beta_0, \beta_1, \ldots, \beta_p$ from the regression data?

First Part of this class

1. How to obtain estimates of $\beta_0, \beta_1, \ldots, \beta_p$ from the regression data?
2. How to assess the variability of these estimates?

First Part of this class

1. How to obtain estimates of $\beta_0, \beta_1, \ldots, \beta_p$ from the regression data?
2. How to assess the variability of these estimates?
3. How does one perform inference on $\beta_0, \beta_1, \ldots, \beta_p$?

First Part of this class

1. How to obtain estimates of $\beta_0, \beta_1, \ldots, \beta_p$ from the regression data?
2. How to assess the variability of these estimates?
3. How does one perform inference on $\beta_0, \beta_1, \ldots, \beta_p$?

- The second and third questions above become tractable by adding more assumptions to the linear model. For the second question, one assumes that the conditional variance of y given $x_1, \ldots, x_p$ is constant denoted by σ^2, i.e,

$$\text{var}(y|x_1, \ldots, x_p) = \sigma^2,$$

  for all values of $x_1, \ldots, x_p$. This assumption is referred to as homoskedasticity.

First Part of this class

1. How to obtain estimates of $\beta_0, \beta_1, \ldots, \beta_p$ from the regression data?
2. How to assess the variability of these estimates?
3. How does one perform inference on $\beta_0, \beta_1, \ldots, \beta_p$?

- The second and third questions above become tractable by adding more assumptions to the linear model. For the second question, one assumes that the conditional variance of y given $x_1, \ldots, x_p$ is constant denoted by σ^2, i.e,

$$\text{var}(y|x_1, \ldots, x_p) = \sigma^2,$$

  for all values of $x_1, \ldots, x_p$. This assumption is referred to as homoskedasticity.
- For the third question (inference), it is common to assume that the conditional distribution of y given $x_1, \ldots, x_p$ is normal.

First Part of this class

- Another topic we will spend time on is model selection.

First Part of this class

- Another topic we will spend time on is model selection.
- A more modern topic in regression is the situation when the number of variables p is large compared to n. This leads to interesting challenges in estimation and inference. Time permitting, I will cover some aspects of this.

First Part of this class

- Another topic we will spend time on is model selection.
- A more modern topic in regression is the situation when the number of variables p is large compared to n. This leads to interesting challenges in estimation and inference. Time permitting, I will cover some aspects of this.
- In the second part of the class, we shall study logistic regression and more generally generalized linear models. In the third part of the class, we shall cover principal component analysis.

First Part of this class

- Another topic we will spend time on is model selection.
- A more modern topic in regression is the situation when the number of variables p is large compared to n. This leads to interesting challenges in estimation and inference. Time permitting, I will cover some aspects of this.
- In the second part of the class, we shall study logistic regression and more generally generalized linear models. In the third part of the class, we shall cover principal component analysis.
- The set up for logistic regression is the same as that of linear regression except that the response variable y is assumed to be 0-1 valued i.e., y is a Bernoulli random variable. In this case, it is easy to see that

$$\mathbb{E}(y|x_1, \ldots, x_p) = \mathbb{P}(y = 1|x_1, \ldots, x_p)$$

- The set up for logistic regression is the same as that of linear regression except that the response variable y is assumed to be 0-1 valued i.e., y is a Bernoulli random variable.

- The set up for logistic regression is the same as that of linear regression except that the response variable y is assumed to be 0-1 valued i.e., y is a Bernoulli random variable.
- In this case, it is easy to see that

$$\mathbb{E}(y|x_1, \dots, x_p) = \mathbb{P}(y = 1|x_1, \dots, x_p)$$

and therefore the conditional mean always lies in the interval $[0, 1]$. As a result, the linear model (1) for the conditional mean may not make much sense.

- The set up for logistic regression is the same as that of linear regression except that the response variable y is assumed to be 0-1 valued i.e., y is a Bernoulli random variable.
- In this case, it is easy to see that

$$\mathbb{E}(y|x_1, \ldots, x_p) = \mathbb{P}(y = 1|x_1, \ldots, x_p)$$

  and therefore the conditional mean always lies in the interval $[0, 1]$. As a result, the linear model (1) for the conditional mean may not make much sense.
- A more natural way of modeling the conditional mean for 0-1 responses is to require that:

$$\mathbb{E}(y|x_1, \ldots, x_p) = g(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)$$

  for some function g mapping $\mathbb{R}$ to the interval $[0, 1]$. A popular choice for g is $g(x) = e^x/(1 + e^x)$ which gives rise to the logistic regression model. We will go over estimation and inference for $\beta_0, \beta_1, \ldots, \beta_p$ in logistic regression as well.