Statistics 151a - Linear Modelling: Theory and Applications

Adityanand Guntuboyina
Department of Statistics
University of California, Berkeley

20 January 2015

The Regression Problem

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

The dependent variable (also known as the response variable) is denoted by y.

The independent (or explanatory variables) are denoted by $x_1, \ldots, x_p$.

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.

Data Examples

1. Bodyfat Data
   (`http://lib.stat.cmu.edu/datasets/bodyfat`).
2. Boston Housing Data
3. Savings Ratio Data
4. Car Seat Position Data
5. Tips Data
6. Frogs Data
7. Email Spam Data

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 values of the variable y are $y_1, \ldots, y_n$ and are collected in the column vector $Y = (y_1, \ldots, y_n)^T$.

The values of the explanatory variables are collected in an $n \times p$ matrix denoted by X. The (i, j)th entry of this matrix is x_{ij} and it denotes the value of the jth variable x_j for the ith subject.

Linear Models

Linear models provide an important tool for solving regression problems.

They have a great number of diverse applications.

They also have a rich mathematical structure.

The Linear Model

$y_1, \ldots, y_n$ are assumed to be random variables (think of the BodyFat dataset where the response variable cannot be accurately measured because of measurement error). But x_{ij} are assumed to be non-random.

The mean of y_i is a linear combination of $x_{i1}, \ldots, x_{ip}$:

$$\mathbb{E}y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}.$$

Also the variance of y_i is the same and equals σ^2 for each i. The different y_is are uncorrelated.

Sometimes, it is also assumed that $y_1, \ldots, y_n$ are jointly normal.

The Linear Model (continued)

If e_i denotes $y_i - \mathbb{E}y_i$, then the model can be written succinctly as

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i \qquad \text{for } i = 1, \ldots, n \qquad (1)$$

where $e_1, \ldots, e_n$ are uncorrelated random variables with mean zero and variance σ^2.

Because e_i has mean zero, it can be considered noise.

The equation (1) therefore says that the value of the response variable for the ith subject equals a linear combination of its explanatory variables give or take some noise. Hence the name linear model.

Linear Model in Matrix Notation

Let β denote the column vector $(\beta_1, \ldots, \beta_p)^T$ and e denote the column vector $(e_1, \ldots, e_n)^T$.

The linear model can be written even more succinctly as:

$$Y = X\beta + e$$

where $\mathbb{E}e = 0$ and $Cov(e) = \sigma^2 I$.

$Cov(e)$ is an $n \times n$ matrix whose (i, j)th entry denotes the covariance between e_i and e_j. I denotes the identity matrix.

Parameters of the Linear Model

The numbers $\beta_1, \ldots, \beta_p$ and σ^2 are the parameters in this model.

β_j can be 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 are kept constant.

It is very important to note that the interpretation of β_j depends not just on the jth explanatory variable but also on all the other explanatory variables in the model.

Linear Models and Regression Analysis

Linear models can be used to achieve the two main objectives in a regression problem: prediction and understanding the relationship between response and explanatory variables.

For prediction, suppose a new subject comes along the values of the explanatory variables for whom are $x_1 = \lambda_1, \ldots, x_p = \lambda_p$ respectively. What then would be a reasonable prediction of her response?

The linear model says that the mean of his response is $\beta_1\lambda_1 + \cdots + \beta_p\lambda_p = \lambda^T\beta$ where $\lambda = (\lambda_1, \ldots, \lambda_p)^T$.

Thus for the prediction problem, we need to learn how to estimate $\lambda^T\beta$ as λ varies.

Estimation

Our first step will be study estimation of β (and σ^2) with special emphasis on estimation of $\lambda^T\beta$.

A very beautiful mathematical theory of Best Linear Unbiased Estimation can be constructed for estimation of $\lambda^T\beta$.

Under a joint-normality assumption on $y_1, \ldots, y_n$, we also study usual estimation methods such as MLE, UMVUE and Bayes estimation.

Inference

To answer questions on the relations between the explanatory and the response variable, we need to test hypotheses of the form $H_0 : \beta_j = 0$.

A beautiful theory of hypothesis testing for the linear model under the additional assumption of joint-normality.

We study this theory in detail after estimation.

Is the linear model a good model?

Not always. One can certainly think of situations in which the assumptions do not quite make sense:

- We might believe that $\mathbb{E}y_i$ depends on $x_{i1}, \ldots, x_{ip}$ in a non-linear way (for example, in BodyFat, it might be more sensible to use the square of the neck circumference variable).
- $y_1, \ldots, y_n$ all may not have the same variance (e.g., the measurement error may not be uniform). This is called heteroscedasticity.
- They may not be uncorrelated.
- Joint normality of $y_1, \ldots, y_n$ is sometimes assumed and this can of course be violated.

Is this a good model (continued)?

- If y_i takes only the values 0 and 1, then $\mathbb{E}(y_i) = \mathbb{P}\{y_i = 1\}$. Modelling a probability by a linear combination of variables might not make sense (why?)

- The observations on the explanatory variables x_{ij} also may have measurement errors so that they are non-random.

Diagnostics

Diagnostics indicate whether the assumptions of the linear model are violated or not.

We will spend a lot of time on these diagnostics.

When assumptions of the linear model are violated, more complicated models might be necessary.

Non-linearity

In the regression problem, we have p response variables $x_1, \ldots, x_p$.

But we can create more response variables from these p variables by modifying and combining these p variables.

For example, we might consider
$x_{p+1} = x_1^2, x_{p+2} = \log x_3, x_{p+3} = I\{x_2 > 1\}, x_{p+4} = x_5 x_8$ etc.

The linear model with this new set of variables also works for cases where $\mathbb{E}y_i$ depends non-linearly on the explanatory variables.

In this sense, the linear model can be used to improve itself.

Variable Selection

In a regression problem therefore, we have potentially a large number of explanatory variables to use.

Which of these should we actually use?

This is the problem of variable selection is very important for building a linear model.

We will study this problem in detail.

Heteroscedasticity

Heteroscedasticity refers to the situation when $y_1, \ldots, y_n$ have different variances.

This can be detected by looking at certain plots. There are also many formal tests to check this assumption.

The problem might sometimes be fixed by transforming the response variable. Examples of transformations include $\log y$ or $\sqrt{y}$.

Correlated Errors

This is the situation where $y_1, \ldots, y_n$ are correlated.

This can also be detected by looking at certain plots and formally checked by various tests.

In this case (and in the previous case of heteroscedasticity), a better model would be:

$$Y = X\beta + e$$

with $\mathbb{E}e = 0$ and $Cov(e) = \sigma^2 V$ where V is not necessarily an identity matrix. We study this model as well.

Joint Normality

Joint normality of $y_1, \ldots, y_n$ (or equivalently of $e_1, \ldots, e_n$) is often used to construct hypothesis tests (and confidence intervals) in regression.

This normality can often by checked by various diagnostic plots.

If it is violated, then a simple fix might be to transform the response variable.

One can also rely on asymptotics which say that the tests are still valid as $n \to \infty$ under some conditions on the distribution of $e_1, \ldots, e_n$.

When these conditions are not satisfied, one may use other techniques.

0-1 valued response variables

When the response variable is 0-1 valued, the linear model stipulates that

$$\mathbb{E}y_i = \mathbb{P}\{y_i = 1\} = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

An oddity about this is that the left hand side lies between 0 and 1 while the right hand side need not.

A better model in this case would be:

$$\log \frac{\mathbb{P}\{y_i = 1\}}{1 - \mathbb{P}\{y_i = 1\}} = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

This is the logistic model, a special case of GLM (Generalized Linear Models).

Measurement Errors in Explanatory Variables

If there exist measurement errors in the explanatory variables as well, one needs to use errors-in-variables models.

We may or may not have time to go over these.

Brief Syllabus

- The Linear Model
  1. Estimation
  2. Inference
  3. Diagnostics
  4. Model Building and Variable Selection

- Generalized Linear Models. Essentially the same steps as above.

Prerequisites

1. Linear Algebra: Basic matrix operations, vector subspaces and projections, rank and invertibility of matrices, quadratic forms.

2. Calculus: Derivatives and gradients.

3. Probability: Random variables, probability density and mass functions, Bayes Rule, Expectations, Variances, Covariances, basic probability distributions.

4. Statistics: At least one previous course in statistics required.

5. Programming: R