# Spring 2015 Statistics 151a (Linear Models) : Lecture Two

Aditya Guntuboyina

22 January 2015

## 1 Recap

### 1.1 The Regression Problem

There is a response variable $y$ and $p$ explanatory variables $x_1, \ldots, x_p$. The goal is understand the relationship between $y$ and $x_1, \ldots, x_p$.

There are $n$ subjects and data is collected on the variables from these subjects.

Data on the response variable is $y_1, \ldots, y_n$ and is represented by the column vector $Y = (y_1, \ldots, y_n)^T$ (the $T$ here stands for transpose).

Data on the $j$th explanatory variable $x_j$ is $x_{1j}, x_{2j}, \ldots, x_{nj}$. This data is represented by the $n \times p$ matrix $X$ whose $(i,j)$th entry is $x_{ij}$. In other words, the $i$th row of $X$ has data collected from the $i$th subject and the $j$th column of $X$ has data for the $j$th variable.

### 1.2 The Linear Model

1. $y_1, \ldots, y_n$ are assumed to be random variables but $x_{ij}$ are assumed to be non-random.
2. It is assumed that

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

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

In matrix notation, the second assumption above can be written as

$$Y = X\beta + e \qquad \text{with } \mathbb{E}e = 0 \text{ and } Cov(e) = \sigma^2 I_n$$

where $I_n$ denotes the $n \times n$ identity matrix. $\beta = (\beta_1, \ldots, \beta_p)^T$ and $e = (e_1, \ldots, e_n)^T$.

If $Z_1, \ldots, Z_n$ are random variables with $Z = (Z_1, \ldots, Z_n)^T$, then $Cov(Z)$ denotes the $n \times n$ matrix whose $(i,j)$th entry denotes the covariance between $Z_i$ and $Z_j$. In particular, the $i$th diagonal entry of $Cov(Z)$ would denote the variance of the random variable $Z_i$. Therefore, $Cov(e) = \sigma^2 I_n$ is a succinct way of saying that the covariance between $e_i$ and $e_j$ would equal 0 when $i \neq j$ and $\sigma^2$ when $i = j$.

## 2 The Intercept Term

Among other things, the linear model stipulates that

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

for $i = 1, \ldots, n$. This implies that when the values of the explanatory variables $x_{i1}, \ldots, x_{ip}$ are all equal to 0, then $\mathbb{E}y_i = 0$. This is of course not always a reasonable assumption. One therefore modifies the linear model slightly by stipulating that

$$\mathbb{E}y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}. \tag{1}$$

Now $\mathbb{E}y_i$ does not have to be zero when all the explanatory variables take on the value zero. The term $\beta_0$ above is known as the *intercept* term. Usually, in linear models, one **always** includes the intercept term by default.

If we let $x_0$ denote the "variable" which always takes the value 1, then (1) can be written as

$$\mathbb{E}y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

Therefore this model with the intercept term is just the same as the previous linear model with this additional variable along with the $p$ explanatory variables.

With the intercept term, one can write the linear model in matrix form as

$$Y = X\beta + e \qquad \text{with } \mathbb{E}e = 0 \text{ and } Cov(e) = \sigma^2 I_n$$

where $X$ denotes the $n \times (p+1)$ matrix whose first column consists of all ones and the rest of the columns correspond to the values of the $p$ explanatory variables and $\beta = (\beta_0, \ldots, \beta_p)^T$.

When $p = 1$ i.e., when there is only one explanatory variable, this linear model (with the intercept term) is called the *simple linear regression model*: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

From now on, we will always consider linear models with the intercept term which means that the first column of $X$ (which is an $n \times (p+1)$ matrix) is always the vector of ones and $\beta$ is a vector of length $p + 1$.

## 3 Estimation in the Linear Model

The quantities $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ and $\sigma^2 > 0$ are *parameters* in the linear model. These need to be estimated from the data. The process of estimating the parameters is also referred to as fitting the linear model to data.

Let us first focus on the estimation of $\beta$.

The idea behind the linear model is that one tries to explain the response value $y_i$ via the linear combination $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. It makes sense, therefore, to estimate $\beta$ by the **minimizer** of the sum of squares

$$S(\beta) := \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip}\right)^2.$$

Using matrix notation, this can be written as

$$S(\beta) = ||Y - X\beta||^2.$$

The norm $||x||$ of a vector $x = (x_1, \ldots, x_n)^T$ is defined as $||x|| := \sqrt{x_1^2 + \cdots + x_n^2}$. Note the equality $||x||^2 = x^T x$. Using this, we can write

$$S(\beta) = (Y - X\beta)^T (Y - X\beta) = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta.$$

This can be minimized via calculus. Take partial derivatives with respect to $\beta_i$ for $i = 0, 1, \ldots, p$ and equate them to 0. It is easy to check that

$$\nabla S(\beta) = 2X^T X\beta - 2X^T Y.$$

where

$$\nabla S(\beta) = \left( \frac{\partial S(\beta)}{\partial \beta_1}, \ldots, \frac{\partial S(\beta)}{\partial \beta_p} \right).$$

denotes the gradient of $S(\beta)$ with respect to $\beta = (\beta_1, \ldots, \beta_p)^T$. It follows therefore that the minimizer of $S(\beta)$ satisfies the equality

$$X^T X\beta = X^T Y. \tag{2}$$

This gives $p$ linear equations for the $p$ unknowns $\beta_1, \ldots, \beta_p$. This important set of equations are called *normal equations*. Their solution, denoted by $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ gives an estimate of $\beta$ called the least squares estimate. If the values of the $p$ explanatory variables for a subject are $\lambda_1, \ldots, \lambda_p$, then the estimate of his mean response is given by $\hat{\beta}_0 + \lambda_1 \hat{\beta}_1 + \cdots + \lambda_p \hat{\beta}_p$.

Two important questions arise are: (1) **Does there exist a solution to the normal equations** and (2) **If yes, then is the solution unique?**

The answer to the first question is **yes**. The normal equations always have a solution. The reason is the following: $X^T Y$ lies in the column space of $X^T$. Further, the column spaces of $X^T$ and $X^T X$ are identical and thus $X^T Y$ can always be written as $X^T X u$ for some vector $u$.

The answer to the second question is **yes if $X^T X$ is invertible** and **no if $X^T X$ is not invertible**.

Do the normal equations (2) admit a *unique* solution? Answer: **No in general. Yes if $X^T X$ is invertible**.

If $X^T X$ is invertible, then the solution to the normal equations is given by $\hat{\beta} = (X^T X)^{-1} X^T Y$. The estimate of the linear function $\lambda^T \beta$ for a vector $\lambda = (\lambda_1, \ldots, \lambda_p)^T$ is then given by $\lambda^T (X^T X)^{-1} X^T Y$.

If $X^T X$ is not invertible, then the normal equations have many solutions. In this case, how does one estimate $\beta$? Here, it actually turns out that the vector $\beta$ **cannot** be estimated. This is explained next.

# 4 When $X^T X$ is not necessarily invertible

The vector $\beta$ cannot be estimated in this case.

Observe first that $X^T X$ being invertible is equivalent to the rank of $X$ being equal to $p + 1$. Thus when $X^T X$ is not invertible, the rank of $X$ is strictly smaller than $p + 1$. In other words, some column of $X$ is a linear combination of the rest of the columns of $X$ i.e., at least one of explanatory variables is redundant in the sense that it can be written as a linear combination of the other explanatory variables.

Let us consider an example here. Suppose $p = 2$ and that the two explanatory variables $x_1$ and $x_2$ are actually the same i.e., $x_{i1} = x_{i2}$ for each $i = 1, \ldots, n$. It should be clear then that the rank of $X$ is at most 2. The linear model can then be written as

$$y_i = \beta_0 + (\beta_1 + \beta_2)\, x_{i1} + \epsilon_i$$

for $i = 1, \ldots, n$. It should be clear that from these observations, the parameters $\beta_1$ and $\beta_2$ **cannot** be estimated. On the other hand, $\beta_1 + \beta_2$ can be estimated.

Thus when $X^TX$ is not invertible, the parameter vector $\beta$ cannot be estimated while certain special linear combinations can be estimated.

**It can be shown that a linear combination $\lambda^T\beta$ can be estimated if and only if $\lambda$ lies in the column space of $X^T$. This is equivalent to saying that $\lambda$ lies in the column space of $X^TX$ because the column spaces of $X^T$ and $X^TX$ are always equal**.

In the example just discussed, the vector $(0,1,1)^T$ is in the column space of $X^T$ which implies that $\beta_1+\beta_2$ is estimable. On the other hand, the vector $(0,1,0)^T$ is not in the column space of $X^T$ which implies that $\beta_1$ is not estimable.

When $X^TX$ is invertible, then the column space of $X^T$ contains all $(p+1)$ dimensional vectors and then every linear combination of $\beta$ is estimable.

# 5 Least Squares Estimates

Consider the normal equations $X^TX\beta = X^TY$. Let $\hat{\beta}_{ls}$ denote any solution (it is unique only if $X^TX$ is invertible).

Let $\lambda^T\beta$ be estimable (i.e., $\lambda$ lies in the column space of $X^T$ or equivalently the column space of $X^TX$). Then estimate $\lambda^T\beta$ by $\lambda^T\hat{\beta}_{ls}$. This is called the least squares estimate of $\lambda^T\beta$.

**Result 5.1.** *If $\lambda^T\beta$ is estimable, then $\lambda^T\hat{\beta}_{ls}$ is the same for every solution $\hat{\beta}_{ls}$ of the normal equations. In other words, the least squares estimate of $\lambda^T\beta$ is unique.*

*Proof.* Since $\lambda^T\beta$ is estimable, the vector $\lambda$ lies in the column space of $X^TX$ and hence $\lambda = X^TXu$ for some vector $u$. Therefore,

$$\lambda^T\hat{\beta}_{ls} = u^TX^TX\hat{\beta}_{ls} = u^TX^TY$$

where the last equality follows from the fact that $\hat{\beta}_{ls}$ satisfies the normal equations. Since $u$ only depends on $\lambda$, this proves that $\lambda^T\hat{\beta}_{ls}$ does not depend on the particular choice of the solution $\hat{\beta}_{ls}$ of the normal equations. □

Thus when $\lambda^T\beta$ is estimable, it is estimated by the least squares estimate $\lambda^T\hat{\beta}_{ls}$ (which is uniquely defined). When $\lambda^T\beta$ is not estimable, it of course does not make sense to try to estimate it.