Lecture 04
Applications and Intro to Multivariate Regression

14 September 2015

Taylor B. Arnold
Yale Statistics
STAT 312/612

Yale

Goals for today

1. Galton’s heights data
2. multivariate regression: normal equations

Galton Heights Application

Multivariate Regression Models

The multivariate linear regression model is, on the surface, only a slight generalization of the simple linear regression model:

$$y_i = x_{1,i}\beta_1 + x_{2,i}\beta_2 + \cdots + x_{1,p}\beta_p + \epsilon$$

The multivariate linear regression model is, on the surface, only a slight generalization of the simple linear regression model:

$$y_i = x_{1,i}\beta_1 + x_{2,i}\beta_2 + \cdots + x_{1,p}\beta_p + \epsilon$$

The statistical estimation problem now becomes one of estimating the p components of the multivariate vector β.

A sample can be re-written in terms of the vector x_i (the vector of covariates for a single observation):

$$y_i = x_i^t \beta + \epsilon$$

In matrix notation, we can write the linear model simultaneously for all observations:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{2,1} & \cdots & x_{p,1} \\ x_{1,2} & \ddots & & x_{p,2} \\ \vdots & & \ddots & \vdots \\ x_{1,n} & x_{2,n} & \cdots & x_{p,n} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

In matrix notation, we can write the linear model simultaneously for all observations:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{2,1} & \cdots & x_{p,1} \\ x_{1,2} & \ddots & & x_{p,2} \\ \vdots & & \ddots & \vdots \\ x_{1,n} & x_{2,n} & \cdots & x_{p,n} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Which can be compactly written as:

$$Y = X\beta + \epsilon$$

In matrix notation, we can write the linear model simultaneously for all observations:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{2,1} & \cdots & x_{p,1} \\ x_{1,2} & \ddots & & x_{p,2} \\ \vdots & & \ddots & \vdots \\ x_{1,n} & x_{2,n} & \cdots & x_{p,n} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Which can be compactly written as:

$$Y = X\beta + \epsilon$$

Note: we use the transpose for $x_i^t\beta$ but not for $X\beta$!

For reference, note the following equation

$$Y = X\beta + \epsilon$$

Yields these dimensions:

$$\begin{aligned} Y &\in \mathbb{R}^{n} \\ X &\in \mathbb{R}^{n \times p} \\ \beta &\in \mathbb{R}^{p} \\ \epsilon &\in \mathbb{R}^{n} \end{aligned}$$

Vector Norms

When working with vectors and matricies, it will be helpful to represent certain quantities by norms. The p-norm of a vector is given by:

$$||x||_p^p = \sum_{i=1}^{n} |x_i|^p$$

Vector Norms

When working with vectors and matricies, it will be helpful to represent certain quantities by norms. The p-norm of a vector is given by:

$$||x||_p^p = \sum_{i=1}^{n} |x_i|^p$$

In particular, the squared 2-norm yields the sum of squares of a vector.

Vector Norm Properties

The following properties are true of all vector norms, for a scalar α and vectors v_1 and v_2.

$$
\begin{aligned}
||\alpha v_1|| &= |\alpha| \cdot ||v_1|| \\
||v_1 + v_2|| &\leq ||v_1|| + ||v_2||
\end{aligned}
$$

p-Norm Properties

p-norms have several additional properties that we will find useful.

p-Norm Properties

p-norms have several additional properties that we will find useful.

Define q such that:

$$\frac{1}{p} + \frac{1}{q} = 1$$

The q-norm and p-norm are then said to be *dual* to one another.

p-Norm Properties

p-norms have several additional properties that we will find useful.

Define q such that:

$$\frac{1}{p} + \frac{1}{q} = 1$$

The q-norm and p-norm are then said to be *dual* to one another.

Notice that the 2-norm is dual to itself.

p-Norm Properties, cont.

Hölder's inequality then yields

$$|v_1^t v_2| \leq ||v_1||_p ||v_2||_q$$

p-Norm Properties, cont.

Hölder's inequality then yields

$$|v_1^t v_2| \leq ||v_1||_p ||v_2||_q$$

As a special case, the Cauchy–Schwarz inequality gives that:

$$|v_1^t v_2|^2 \leq ||v_1||_2^2 ||v_2||_2^2$$

p-Norm Properties, cont.

Finally, and of most importance for us today, note that the squared 2-norm is exactly equal to the self inner product:

$$||v_1||_2^2 = v_1^t v_1$$

Least squares (again)

To estimate the least squares solution, which is again the MLE for independent normal errors, we see that:

$$\widehat{\beta} \in \underset{b \in \mathbb{R}^p}{\arg\min} \left\{ ||Y - X\beta||_2^2 \right\}$$

Least squares (again)

To estimate the least squares solution, which is again the MLE for independent normal errors, we see that:

$$\widehat{\beta} \in \underset{b \in \mathbb{R}^p}{\arg\min} \left\{ ||Y - X\beta||_2^2 \right\}$$

Now using vector norms to denote the sum of squares.

It will be helpful to re-write the sum of squares as: in β:

$$||Y - X\beta||_2^2 = (Y - X\beta)^t(Y - X\beta)$$

It will be helpful to re-write the sum of squares as: in β:

$$
\begin{aligned}
||Y - X\beta||_2^2 &= (Y - X\beta)^t (Y - X\beta) \\
&= (Y^t - \beta^t X^t)(Y - X\beta)
\end{aligned}
$$

It will be helpful to re-write the sum of squares as: in β:

$$\begin{aligned}
||Y - X\beta||_2^2 &= (Y - X\beta)^t (Y - X\beta) \\
&= (Y^t - \beta^t X^t)(Y - X\beta) \\
&= Y^t Y - Y^t X\beta - \beta^t X^t Y + \beta^t X^t X\beta
\end{aligned}$$

It will be helpful to re-write the sum of squares as: in β:

$$\begin{aligned} ||Y - X\beta||_2^2 &= (Y - X\beta)^t (Y - X\beta) \\ &= (Y^t - \beta^t X^t)(Y - X\beta) \\ &= Y^t Y - Y^t X\beta - \beta^t X^t Y + \beta^t X^t X\beta \\ &= Y^t Y - 2Y^t X\beta + \beta^t X^t X\beta \end{aligned}$$

Normal Equations

In order to find the minimum of the sum of squares, we take the gradient with respect to β and set it equal to zero.

Recall that, for a vector a and symmetric matrix A :

$$\nabla_\beta a^t\beta = a$$
$$\nabla_\beta \beta^t A\beta = 2A\beta$$

Normal Equations

In order to find the minimum of the sum of squares, we take the gradient with respect to β and set it equal to zero.

Recall that, for a vector a and symmetric matrix A :

$$\begin{aligned} \nabla_\beta a^t \beta &= a \\ \nabla_\beta \beta^t A \beta &= 2A\beta \end{aligned}$$

This gives the gradient of the sum of squares as:

$$\begin{aligned} \nabla_\beta ||Y - X\beta||_2^2 &= \nabla_\beta \left(Y^t Y - 2Y^t X\beta + \beta^t X^t X \beta \right) \\ &= 2X^t X\beta - 2X^t y \end{aligned}$$

Setting this equal to zero gives a set of p equations called the normal equations:

$$X^t X \widehat{\beta} = X^t y$$

Maximum or Minimum?

To determine whether the normal equations give a local minimum, maximum, or saddle point, we can calculate the Hessian matrix.

Maximum or Minimum?

To determine whether the normal equations give a local minimum, maximum, or saddle point, we can calculate the Hessian matrix. This is a $p \times p$ matrix giving every combination of the second partial derivatives:

$$Hf(\beta) = \begin{pmatrix} \frac{\partial^2 f}{\partial \beta_1 \partial \beta_1} & \frac{\partial^2 f}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 f}{\partial \beta_1 \partial \beta_p} \\ \frac{\partial^2 f}{\partial \beta_2 \partial \beta_1} & \ddots & & \frac{\partial^2 f}{\partial \beta_2 \partial \beta_p} \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 f}{\partial \beta_n \partial \beta_1} & \frac{\partial^2 f}{\partial \beta_n \partial \beta_2} & \cdots & \frac{\partial^2 f}{\partial \beta_n \partial \beta_p} \end{pmatrix}$$

If the Hessian is positive definite ($x^t H x \geq 0$) at a critical point, then the corrisponding point is a minimum of the

Looking at the gradiant of the sum of squares:

$$\nabla_{\beta}||Y - X\beta||_2^2 = 2X^tX - 2X^ty$$

Looking at the gradiant of the sum of squares:

$$\nabla_\beta ||Y - X\beta||_2^2 = 2X^tX - 2X^ty$$

We can see that the Hessian is simply:

$$H_\beta ||Y - X\beta||_2^2 = 2X^tX$$

Looking at the gradiant of the sum of squares:

$$\nabla_\beta ||Y - X\beta||_2^2 = 2X^tX - 2X^ty$$

We can see that the Hessian is simply:

$$H_\beta ||Y - X\beta||_2^2 = 2X^tX$$

Why is this positive definite?

Looking at the gradiant of the sum of squares:

$$\nabla_\beta ||Y - X\beta||_2^2 = 2X^tX - 2X^ty$$

We can see that the Hessian is simply:

$$H_\beta ||Y - X\beta||_2^2 = 2X^tX$$

Why is this positive definite?

$$\begin{aligned} v^t \left(2X^tX\right) v &= 2\left(v^tX^tXv\right) \\ &= 2||Xv||_2^2 \\ &\geq 0 \end{aligned}$$

Back to the normal equations themselves, notice that if the matrix X^tX is invertable, we can 'solve' the normal equations as:

$$\begin{aligned} X^tX\widehat{\beta} &= X^ty \\ \widehat{\beta} &= (X^tX)^{-1}X^ty \end{aligned}$$

Back to the normal equations themselves, notice that if the matrix $X^t X$ is invertable, we can 'solve' the normal equations as:

$$\begin{aligned} X^t X \widehat{\beta} &= X^t y \\ \widehat{\beta} &= (X^t X)^{-1} X^t y \end{aligned}$$

This is not a good way to solve the normal equations numerically, but for deriving theoretical results about the least squares estimator this form will be very useful.

Matricies and Model Frames in R