Section 10: Advanced Regression + Geometry of Least Squares

TF: Reagan Rose (rrose@g.harvard.edu)

November 18, 2015

Motivation

The methods we've used so far in this class provide techniques for analyzing data in *one dimension*. But in statistics, we're never interested in analyzing just one observation of data or just one variable. Statistics and data are inherently *multidimensional*. To deal with this, instead of applying our one-dimensional techniques to every single line of data or every single variable of interest, we use **multivariate notation** and **linear algebra** to make our lives easier and do analyses quickly and easily!

Review of Linear Algebra

First let's review some basic concepts from linear algebra.

Vectors

If we have two vectors p-dimensional vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, then we have the following **vector operations**:

- **Dot product:** $\boldsymbol{u} \cdot \boldsymbol{v} = \boldsymbol{u}^T \boldsymbol{v} = \sum_{i=1}^p u_i v_i$
- **Norm**: $||\boldsymbol{u} - \boldsymbol{v}||^2 = (\boldsymbol{u} - \boldsymbol{v}) \cdot (\boldsymbol{u} - \boldsymbol{v}) = ||\boldsymbol{u}||^2 + ||\boldsymbol{v}||^2 - 2\boldsymbol{u} \cdot \boldsymbol{v}$

and **vector properties:**

- **orthogonality**: $\boldsymbol{u} \perp \boldsymbol{v}$ if $\boldsymbol{u} \cdot \boldsymbol{v} = 0$.
- **decomposition**: for an arbitrary vector $\boldsymbol{y} \in \mathbb{R}^p$, $\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{y}^{\perp}$

Let's call $\hat{\boldsymbol{y}}$ the "best linear predictor of y", where "best" means that

$$||\boldsymbol{y} - \hat{\boldsymbol{y}} < ||\boldsymbol{y} - \boldsymbol{v}||^2$$

for every vector $\boldsymbol{v} \neq \hat{\boldsymbol{y}}$ in a subspace $V \subseteq \mathbb{R}^p$. Equivalently, we can write

$$\arg\ \min_{v \in V} ||\boldsymbol{y} - \boldsymbol{v}||^2 = \hat{\boldsymbol{y}}.$$

In other words, $\hat{\boldsymbol{y}}$ is the "closest" vector to $\boldsymbol{y}$ in the subspace V.

<u>Matrices</u>

Matrices can be tricky to work with. So let's review some basic concepts and properties. Below is a list of important **matrix operations** that you'll need to remember when working with regression formulae. To get started, let a be an arbitrary $p \times 1$ vector, b an arbitrary $1 \times b$ vector, and M an arbitrary $p \times p$ matrix. The following properties hold

- **Vector/Matrix equivalence**: Every vector is a matrix. The converse is not true.
- **Matrix multiplication:** Matrices can be multiplied if the *inner dimensions* match (i.e. $(p \times 1) \times (1 \times p)$ can be multiplied and will be a $p \times p$ matrix. $(p \times 1) \times (n \times 1)$ cannot be multiplied). Order does matter in matrix multiplication (i.e.$A \times B \neq B \times A$), and multiplication is done using *row by column* logic.
- **Addition/Subtraction**: Matrices of the same dimension can be added and subtracted pointwise
- **Column space**: $C(X) =$ Column space of $\boldsymbol{X} = \{\boldsymbol{X}a :$ all vectors $a \in \mathbb{R}^p\}$
- **Inverse**: Only square matrices can be inverted. If $\boldsymbol{M} = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}$ then

$$\boldsymbol{M}^{-1} = \frac{1}{m_{11}m_{22} - m_{12}m_{21}} \begin{pmatrix} m_{21} & -m_{12} \\ -m_{21} & m_{11} \end{pmatrix}$$

Multiple Regression Framework

Now let's apply these linear algebra concepts to multiple linear regression. Recall that the general framework for regression assumes that we have data:

$$(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)$$

where $Y_i, i = 1, \ldots, n$ is the outcome variable for observation i and $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})$ is a vector of p predictor variables for observation i. Usually we let $X_{i1} = 1$ for all i to represent the intercept term. This then allows us to write the **multiple regression model**

$$Y_i = \boldsymbol{X_i}\boldsymbol{\beta} + \epsilon_i,$$

where as usual $\epsilon_i \sim^{iid} \mathcal{N}(0, \sigma^2)$ represents the "noise" in the model and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ is the coefficient vector. We can then compile this into a compact, matrix-form equation as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is a vector of outcomes,

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \ldots \boldsymbol{X}_n \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{pmatrix}$$

is a matrix of predictors and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n) \sim \mathcal{N}_p(\boldsymbol{0}, \sigma^2 I_p)$ is a random vector of "noise" drawn from a *multivariate normal distribution.* We call $\boldsymbol{X}$ the **model matrix**, which is an extremely important concept in the theory of linear models.

Note that $\boldsymbol{X\beta}$ is matrix multiplication, so we can think of this as

$$\boldsymbol{X\beta} = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \beta_1 \begin{pmatrix} X_{11} \\ \vdots \\ X_{n1} \end{pmatrix} + \ldots + \beta_p \begin{pmatrix} X_{1p} \\ \vdots \\ X_{np} \end{pmatrix} = \beta_1 \boldsymbol{X_{\cdot 1}} + \ldots + \beta_p \boldsymbol{X_{\cdot p}}.$$

The Least Squares Problem

Note that we can define the **residuals** as:

$$e_i = Y_i - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}$$

in matrix/vector form, this becomes

$$\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X\beta}.$$

The **least squares problem** then can be expressed as:

$$min_\beta ||\boldsymbol{e}||^2 = min_\beta ||\boldsymbol{Y} - \boldsymbol{X\beta}||^2.$$

Differentiating this and setting equal to zero yields the least squares estimator

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y}.$$

Geometry of Least Squares

A important decomposition used in linear regression is

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{y}^\perp,$$

where $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the predicted value of $\boldsymbol{y}$ and $\boldsymbol{y}^\perp = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the orthogonal component of $\boldsymbol{y}$. Thus, somewhat obviously:

$$\boldsymbol{Y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{e}.$$

But this is important because of what we know about the Pythagorean Theorem: when $\boldsymbol{u} \perp \boldsymbol{v}$, $||\boldsymbol{u} + \boldsymbol{v}||^2 = ||\boldsymbol{u}||^2 + ||\boldsymbol{v}||^2$. Thus we have

$$\boldsymbol{X}\hat{\boldsymbol{\beta}} \perp \boldsymbol{e} \implies ||\boldsymbol{Y}||^2 = ||\boldsymbol{X}\hat{\boldsymbol{\beta}}||^2 + ||\boldsymbol{e}||^2$$

so that if everything is centered around 0, we have derived:

$$SST = SSR + SSE.$$

Also recall from linear algebra that if we consider $\boldsymbol{x} - \bar{\boldsymbol{x}}$ and $\boldsymbol{Y} - \bar{\boldsymbol{Y}}$ as vectors, and consider the angle, θ, between them, then we can write:

$$cos(\theta) = \frac{(\boldsymbol{x} - \bar{\boldsymbol{x}}) \cdot (\boldsymbol{Y} - \bar{\boldsymbol{Y}})}{||\boldsymbol{x} - \bar{\boldsymbol{x}}||||\boldsymbol{Y} - \bar{\boldsymbol{Y}}||}$$

then plugging in what we know about these terms will tell us that

$$cos(\theta) = \frac{S_{XY}}{S_X S_Y} = \hat{\beta}_1 \frac{S_X}{S_Y} = R.$$

So **the angle between the (centered) predictors and the (centered) outcome vector tells us exactly the correlation coefficient!**. It is also possible to prove that

$$cos^2(\theta) = R^2,$$

where R^2 is the coefficient of determination for the regression.This makes sense, since when $\boldsymbol{Y}$ and $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ are close, then $\boldsymbol{X}$ gives us a lot of information to get a good estimate for $\boldsymbol{Y}$, so $cos(\theta)$ is high, and so is R^2.

Exercises

Question 1. Regression 3 Ways
Suppose we have the following dataset, with one outcome variable Y, one predictor variable X and two observations:

$$\boldsymbol{Y} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 2.5 \\ 1.1 \end{pmatrix}.$$

(a) Find β_0 and β_1 for the regression model for this data: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, using the **univariate formulas** you know.

(b) Solve for $\hat{\boldsymbol{\beta}} = (\beta_0, \beta_1)^T$ **by hand** using the **matrix/vector formulas** you know.

(c) Now solve for β_0 and β_1 using the **lm function** in R.

(d) Finally, solve for $\hat{\boldsymbol{\beta}} = (\beta_0, \beta_1)^T$ using **matrix multiplication** in R.

Question 2. Download the HEIGHT.CSV dataset we worked with last week. We're interested in modeling child height based on all of the other variables.

(a) Construct the model matrix $\boldsymbol{X}$ for this model.

(b) Use matrix multiplication to calculate the regression coefficients.

(c) Use matrix multiplication to find the standard errors of the regression coefficients, and use these to find the t-statistics for each of the coefficients.

(d) Report the final model, including only the significant predictors.