Modern Regression

Daniel Lawson University of Bristol

Lecture 02.1.2 (v1.0.1)

Signposting

- The previous section 02.1.1 is about interpretation of Regression in general.
- This lecture contains the mathematical content for Modern Regression (Matrix representation).

Linear algebra view of covariance

- The **covariance matrix** of a random variable X
- Where X is an $n \times 1$ matrix, i.e. a column vector,
- has entries:

$$\mathrm{Cov}(\mathrm{X})_{ij} = \mathrm{Cov}(\mathrm{X}_i, \mathrm{X}_j) = \mathbb{E}[(\mathrm{X}_i - \mu_i)(\mathrm{X}_j - \mu_j)].$$

- The matrix form for this is:

$$\Sigma = \mathbb{E}[(\mathrm{X} - \mathbb{E}[\mathrm{X}])(\mathrm{X} - \mathbb{E}[\mathrm{X}])^T],$$

- Where $\mu = \mathbb{E}[\mathrm{X}]$.

Linear algebra view of correlation

- Division by standard deviations is required to correctly generalise the **scalar correlation**:

$$\mathrm{Corr}(X, Y) = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}.$$

- The **matrix form** for correlation is:

$$\mathrm{Corr}(\mathrm{X}) = (\mathrm{diag}(\Sigma))^{-1/2}\, \Sigma\, (\mathrm{diag}(\Sigma))^{-1/2}$$

- The matrix inversion is not computationally challenging because it is for a **diagonal matrix**.

Regression is analogous to linear algebra with noise

- Most problems in Linear Algebra can be seen as **solving a system of linear equations:**

$$Ax + b = 0.$$

- However, data are not usually generated from a linear model.
- We therefore typically seek the least-bad fit that we can:

$$\min ||Ax + b||_2^2 = \min \sum_{i=1}^{N} (Ax_i - b)^2$$

  - i.e. we find A and b such that they minimise the distance (in the squared L_2 norm)

- Linear Algebra is therefore a very powerful way to view regression.

Matrix form of least squares

- Consider data X' with p' **features** (columns) and n observations.
- Given the regression problem:

$$\mathbf{y} = \mathrm{X}'\beta' + \mathbf{b} + \mathbf{e}$$

  - to find β' (a matrix dimension $p' \times 1$))
  - and b to minimise 'error':
  - in $e^2 = \sum_{i=1}^{n} \epsilon_i^2$

Matrix form of least squares

- We construct a simpler representation by adding a constant feature:

$$\mathrm{X} = \begin{bmatrix} 1 & \mathrm{X}_{11} & \cdots & \mathrm{X}_{1p'} \\ & & \cdots & \\ 1 & \mathrm{X}_{n1} & \cdots & \mathrm{X}_{np'} \end{bmatrix}$$

  - which has $p = p' + 1$ features.
- We now solve the analogous equation:

$$\mathbf{y} = \mathrm{X}\beta + \mathbf{e}$$

  - which has the same solution but is in a more convenient form.

Mean Squared Error (MSE)

- The **prediction error** is:

$$\mathbf{e}(\beta) = \mathbf{y} - \mathrm{X}\beta$$

- And the **estimation error** can be written:

$$\mathrm{MSE}(\beta) = \frac{1}{n}\mathbf{e}^T\mathbf{e}$$

Minimising MSE

- Taking (vector) derivatives with respect to β:

$$\begin{aligned} \nabla \mathrm{MSE}(\beta) &= \frac{1}{n}(\nabla \mathbf{y}^T \mathbf{y} - 2\nabla \beta^T x^T \mathbf{y} + \nabla \beta^T \mathrm{X}^T \mathrm{X}\beta) && (1) \\ &= \frac{1}{n}(0 - 2x^T \mathbf{y} + 2\mathrm{X}^T \mathrm{X}\beta) && (2) \end{aligned}$$

- which is zero at the optimum $\hat{\beta}$:

$$\mathrm{X}^T \mathrm{X}\hat{\beta} - \mathrm{X}^T \mathbf{y} = 0$$

- with the solution:

$$\hat{\beta} = (\mathrm{X}^T \mathrm{X})^{-1} \mathrm{X}^T \mathbf{y}.$$

- Exercise: For the case $p' = 1$, check that this solution is the same as you can find in regular linear algebra textbooks.

The Hat Matrix

- There is an important and **response independent** quantity hidden in the prediction:

$$\mathrm{H} = \mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T$$

- The fitted values are:

$$\hat{y} = \mathrm{X}\hat{\beta} = \mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathbf{y} = \mathrm{H}\mathbf{y}$$

  - H is dimension $N \times N$
  - H "projects" y into the fitted value space $\hat{y}$

Properties of the Hat Matrix

- **Influence**: $\frac{\partial \hat{y}_i}{\partial y_j} = \mathrm{H}_{ij}$. So H controls how much a change in one observation changes the estimates of each other point.
- **symmetry**: $\mathrm{H}^T = \mathrm{H}$. So influence is symmetric.
- **Idempotency**: $\mathrm{H}^2 = \mathrm{H}$. So the predicted value for any projected point is the predicted value itself.
- You should read up on these and other vector algebra properties.

Residuals and the Hat Matrix

- The residuals can be written:

$$e = y - \mathrm{H}y = (\mathrm{I} - \mathrm{H})\mathbf{y}$$

- $\mathrm{I} - \mathrm{H}$ is also symmetric and idempotent, and can also be interpreted in terms of Influence.
- Because of this,

$$\mathrm{MSE}(\hat{\beta}) = \frac{1}{n}\mathbf{y}^T(1 - \mathrm{H})^T(1 - \mathrm{H})\mathbf{y} = \frac{1}{n}\mathbf{y}^T(1 - \mathrm{H})\mathbf{y}$$

Expectations

- If the data were generated by our model(**!**) then they are described by a random variable $\mathbf{Y}$:

$$\mathbf{Y} = \mathbf{x}\beta + \epsilon$$

- where ϵ is an $n \times 1$ matrix of RVs with mean $\mathbf{0}$ and covariance $\sigma^s\mathrm{I}$.
- From this it is straightforward to show that the **fitted values are unbiased**:

$$\mathbb{E}[\hat{y}] = \mathbb{E}[\mathrm{H}Y] = \mathbf{x}\beta$$

- using the properties of Expectations with the symmetry and idempotency of H.

Covariance

- Similarly, it is straightforward to show that

$$\mathrm{Var}[\hat{y}] = \sigma^2 \mathrm{H}$$

using the properties of Variances with the symmetry and idempotency of H.

Reflection

- By the end of the course, you should:
  - Be able to define **correlation** and **regression** in multivariate context
  - Be able to perform basic calculations using these concepts
  - Be able to extend intuition about their application.
- This is something worth reading up on
  - You should really understand univariate regression

Signposting

- The mathematics behind Modern Regression is entirely analogous to the mathematics of a huge range of advanced, scalable Machine Learning tools.
- This is one of the places you should do your homework.

References

There is a lot more technical detail in Cosma Shalizi's course on **Modern Regression**:
Modern Regression, by Coasma Shalizi
http://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/