Modern Regression

Daniel Lawson University of Bristol

Lecture 02.1.2 (v1.0.2)

Signposting

- The previous section 02.1.1 is about interpretation of Regression in general.
- This lecture contains the mathematical content for Modern Regression - the vectorised version, which uses Matrix algebra.

Notation, Notation, Notation

- There are several choices of convention that we have to make
- Vectors of length k are also matrices, but are they $k \times 1$ or $1 \times k$?
- We use $k \times 1$, i.e. column vectors
- Similarly there are choices about matrix derivatives
- We use derivative with respect to a column vector as a row vector
- Some resources will have everything transposed as a consequence

Linear algebra view of covariance

- The **covariance matrix** of a random variable X
- Where X is a vector-valued RV with length k,
- 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:**

$$\mathrm{X}\mathrm{A} + \mathbf{b} = 0.$$

  - Where X is an n by p matrix of data,
  - A is an p by 1 matrix of coefficients,
  - and $-\mathbf{b}$ is a n-vector of target values.
- However, data are not *usually* generated from a linear model.
- We therefore typically seek the least-bad fit that we can:

$$\mathrm{argmin}||\mathrm{X}\mathrm{A} + \mathbf{b}||_2^2 = \sum_{i=1}^{N}(\mathbf{x}_i\mathrm{A} + b_i)^2$$

  - i.e. we find A and $\mathbf{b}$ such that they minimise the distance (in the squared L_2 norm)
- Linear algebra allows this very effectively!
- 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:

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

- Using the notation that $\mathbf{e}$ is a p by 1 matrix
- The **estimation error** is written in matrix form:

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

- Why? $\mathbf{e}^T\mathbf{e} = \sum_{i=1}^{n} e_i^2$
    - Hence $\mathrm{MSE}(\beta)$ is a 1×1 matrix, i.e. a scalar, and $|\mathrm{MSE}(\beta)| = \mathrm{MSE}(\beta)$.
    - Noticing this sort of thing makes the matrix algebra easier.
- We want to minimise this MSE with respect to the parameters β.

How to do the Matrix Algebra

Lecture 13 of Cosma Shalizi's notes is a really helpful reminder!

- Look at the Matrix Algebra Cheat Sheet - specifically:
  - How does a transpose work?
  - How do you re-order elements?
  - How does a gradient work in linear and quadratic forms?

Minimising MSE

- Taking (vector) derivatives with respect to β:

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

- 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:

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

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

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:

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

- $\mathrm{I} - \mathrm{H}$ 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 an RV $\mathbf{Y}$ (an n-vector):

$$\mathbf{Y}_i = \mathbf{x}_i \beta + \epsilon_i$$

- $\mathbf{x}_i$ is still a vector but *not* a Random Variable!
- ϵ is an $n \times 1$ matrix of RVs with mean $\mathbf{0}$ and covariance $\sigma_s^2 \mathrm{I}$.
- From this it is straightforward to show that the **fitted values are unbiased**:

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

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

Covariance

- Similarly, it is straightforward to show that

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

  using the properties of Variances with the symmetry and idempotency of H.
- In other words, the covariance of the fitted values is determined entirely by the structure of the covariates, via the Hat matrix.

Implications

- Matrix form is a massive simplification of complex algebra
- It is easy to check that e.g. dimensions make sense
- These vector calculations are repeated in many machine-learning methods
- The details change but the principle remains
- Linear-Algebra loss minimisation techniques are extremely important
- They often sit inside a wider argument, e.g. updated conditional on some other parameters

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.
  - Be able to follow the reasoning in a paper where things get complicated.
- Matrix algebra is worth reading up on!
  - Describe it for example in your assessments' reflection.

Signposting

- Make sure to look at **02.1-Regression.R**
- The mathematics behind Modern Regression is analogous to the mathematics underpinning scalable Machine Learning. **It is very important**.
- For accessible material see Cosma Shalizi's Modern Regression Lectures (Lectures 13-14)
- Further reading in chapters 2.3 and 3.2 of The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Friedman, Hastie and Tibshirani)
- Next up: 2.2 Statistical Testing