chalkdust

Chalkdust Dissertation Prize 2025

The Model Olympics

Using Multiple-Response Regression

By Raul Unnithan

July 14, 2025

1 Introduction

Imagine you are deciding where to put your money. One investment promises maximum profits, while another promises to protect the planet by scoring highly on sustainability measures. Traditionally, financial models predict only one of these outcomes. But in a world where investors increasingly value both, is there a way to predict them simultaneously?

This was the challenge I explored in my final-year project. My goal was to find which type of multiple-response regression (MRR) model best predicts the profitability and sustainability of equity funds, using data about fund size, costs and strategy.

Why not just use standard (single-response) regression for each response? Let us take an example where we try to predict some students' maths and science scores based on their hours studied and hours spent on past papers. We know there might be a relationship between these scores, as students who are good at maths often excel at science, since both rely on similar analytical skills. However, building two separate models ignores potential correlations between each score.

MRR, on the other hand, cleverly accounts for this correlation, and it extends single-response regression by modelling both outcomes together. Mathematically, instead of predicting a single outcome, it predicts a matrix of outcomes, each column representing one response variable. This captures the correlation between outcomes, that is, how changes in one outcome relate to changes in the other.

2 Example

Let us work through an example to demonstrate how the simplest case of multiple-response regression, multiple-response linear regression, works. Continuing with our previous example of maths and science scores, let us say we have the following dataset:

X_1: Hours Studied	X_2: Hours Spent on Papers	Y_1: Math Scores	Y_2: Science Scores
5	2	78	80
7	3	85	79
8	4	88	88
3	1	65	70
10	5	92	74

Table 1: Study Time vs. Exam Scores

Using the formula for Pearson's Moment Correlation Coefficient, the correlation between the responses, maths and science scores, is:

$$r = \frac{\sum_i (y_{1,i} - \bar{y}_1)(y_{2,i} - \bar{y}_2)}{\sqrt{\sum_i (y_{1,i} - \bar{y}_1)^2 \ \sum_i (y_{2,i} - \bar{y}_2)^2}} = \frac{151.4}{\sqrt{449.2 \times 184.8}} \approx 0.526 (3\text{dp}).$$

Note: it is important to check for correlation because if the responses have zero correlation, they are independent. This means that you can just build separate regression models for each response.

Now, let us apply the simplest case of MRR: multiple response linear regression, to this example. The general form of multiple-response linear regression is similar to standard linear regression, and it is as follows:

$$\mathbf{Y}_{(n\times m)} = \mathbf{X}_{(n\times p)}\, \mathbf{B}_{(p\times m)} + \mathbf{E}_{(n\times m)}, \tag{1}$$

where n is the number of observations, m is the number of responses, and p is the number of predictors, **including the column of ones**. Meanwhile, $\mathbf{Y}$ is the response variable matrix, $\mathbf{X}$ is the design matrix, $\mathbf{B}$ is the coefficient matrix, and $\mathbf{E}$ is the error variable matrix.

Like with single-response linear regression, the first step in this modelling approach is to decide what predictors we want to evaluate fit with. Let us use both predictors here. Next, (remembering the column of ones) calculate the coefficient matrix, using the least squares approximation, as:

$$\begin{aligned}\hat{\mathbf{B}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} \\ &= \left(\begin{bmatrix}1 & 5 & 2\\ 1 & 7 & 3\\ 1 & 8 & 4\\ 1 & 3 & 1\\ 1 & 10 & 5\end{bmatrix}^\top \begin{bmatrix}1 & 5 & 2\\ 1 & 7 & 3\\ 1 & 8 & 4\\ 1 & 3 & 1\\ 1 & 10 & 5\end{bmatrix}\right)^{-1} \begin{bmatrix}1 & 5 & 2\\ 1 & 7 & 3\\ 1 & 8 & 4\\ 1 & 3 & 1\\ 1 & 10 & 5\end{bmatrix}^\top \begin{bmatrix}78 & 80\\ 85 & 79\\ 88 & 88\\ 65 & 70\\ 92 & 74\end{bmatrix} \\ &= \begin{bmatrix}55.4732 & 63.5268\\ 2.8661 & 1.2946\\ -0.0268 & 1.2946\end{bmatrix}\end{aligned}$$

Note: a hat has been placed over the coefficient matrix here, as this method simply provides an estimate for the coefficient matrix.

Now, substitute this into our standard formula for predicting responses in linear regression:

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X}\hat{\mathbf{B}} \\ &= \begin{bmatrix}1 & 5 & 2\\ 1 & 7 & 3\\ 1 & 8 & 4\\ 1 & 3 & 1\\ 1 & 10 & 5\end{bmatrix} \begin{bmatrix}55.4732 & 63.5268\\ 2.8661 & 1.2946\\ -0.0268 & 1.2946\end{bmatrix} = \begin{bmatrix}69.7361 & 74.2376\\ 75.4454 & 77.1214\\ 78.2850 & 79.7106\\ 63.5757 & 68.7062\\ 83.6599 & 83.1808\end{bmatrix}.\end{aligned}$$

This set of predictions can be evaluated in various ways. I used a metric called the average normalised root mean squared error (ANRMSE), normalising by the standard deviation. This measures how large a model's prediction errors are on average, relative to the scale of the actual data. Larger discrepancies between predictions and actual values are penalised more heavily, making it easy to compare accuracy across different variables or models. Mathematically, the ANRMSE is defined as:

$$\text{ANRMSE} = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ij}-\hat{y}_{ij})^2}}{s_j}\right),$$

where m and n are the same as in Equation 1, y_{ij} is the observed value for response j and observation i, $\hat{y}_{ij}$ is the predicted value, and s_j is the sample standard deviation of the jth response given by:

$$s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_{ij}-\bar{y}_j)^2}.$$

ANRMSE values can be interpreted as follows: an excellent model has a 0–0.5 ANRMSE, a good model has a 0.5–0.6 ANRMSE, a satisfactory model has a 0.6–0.7 ANRMSE, and an unsatisfactory model has an ANRMSE greater than 0.7.

After extensive substitution and simplification, the ANRMSE for multiple-response linear regression here is 0.835 (3dp). This indicates that this model is unsatisfactory for our sample dataset.
Note: we can expect this as linear modelling does not adapt to the data as well as other models.

3 Multiple-Response Regression in Equity Funds

A tale of two scores: Return on Equity and Sustainability Score

My project extended the example in Section 2 to a cleaned dataset of over 1,200 equity funds. An equity fund is an investment fund that buys shares in companies. In other words, it pools together money from lots of investors to purchase stocks, aiming to grow the money over time. For each fund, I had 12 predictor variables, such as fund size, which measures the total assets in the fund, and dividend yield, which measures how much profit is paid out to investors.
Ultimately, I wanted to predict two outcomes: Return on Equity (ROE), a measure of profitability, and Sustainability Score, a measure of how well a fund mitigates environmental and social risks.

MRR was appropriate here due to the negative correlation between ROE and Sustainability Score. An important caveat here is that a lower Sustainability Score means an equity fund has better sustainability. Therefore, increasing ROE leads to improved sustainability.

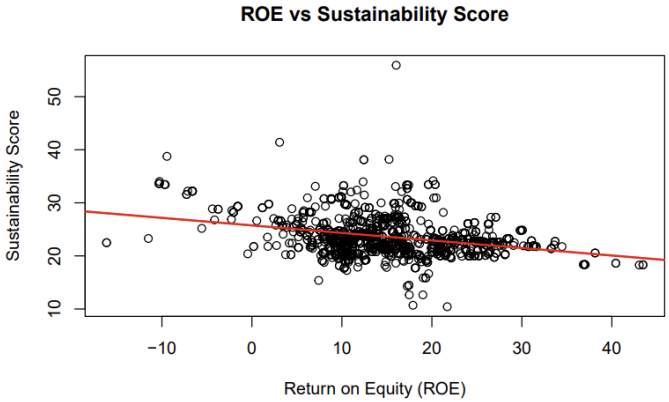

Figure 1: ROE and Sustainability Score Correlation

The model Olympics

My project also evaluated and compared different MRR models on the dataset. I used several methods: multiple-response linear regression, shrinkage methods, random forests and extreme gradient boosting. Note: for more information about these models, please see my dissertation (link below).

Each model generated its predictions for ROE and Sustainability Score for each fund differently, but I evaluated all of said predictions using the ANRMSE. The winner? Extreme gradient boosting, which outperformed the simpler models by capturing complex non-linear relationships in the dataset.

I was thinking of an Olympics-type podium of the top 3 performing models?

Why it matters?

Why should we care about modelling two responses together? Because real-world decisions rarely hinge on just one outcome. Investors want both profitability and responsibility. Doctors want treatments that extend life and improve quality. Mathematicians can help by building tools that reveal the interplay between goals, so we do not have to choose blindly.

In my project, multiple-response regression brought this trade-off into focus. By modelling both outcomes together, it showed that high returns often came with improved sustainability.

4 Further Reading and Visual Ideas

If you would like to explore this topic, look up: the concept of covariance matrices, Ridge and Lasso Regression and ensemble methods like Random Forests and Extreme Gradient Boosting.