Multiple Linear Regression

Round 2, Quant Club Recruitment 2024-25

Siddharth Sharma

September 5, 2024

1 Introduction

This document is about *linear regression*, a very simple approach for supervised learning. In particular, linear regression is a useful tool for predicting a quantitative response. Simple linear regression, as we have seen in our Probability & Statistics course, is a useful approach for predicting a response on the basis of a single predictor variable. However, in practice we often have more than one predictor. This is where *multiple* linear regression enters the picture. Here, we will attempt to use our understanding of simple linear regression and extend it to some of the key ideas underlying the multiple linear regression model.

It is highly recommended to revise simple linear regression before moving onto the following content. You may refer to material used in PnS or you can use this document.

2 Multiple Linear Regression

We consider the problem of regression when the study variable depends on more than one explanatory or independent variables (also called predictors), called a multiple linear regression model. This model generalizes the simple linear regression in two ways. It allows the mean function $E(y)$ to depend on more than one explanatory variables and to have shapes other than straight lines, although it does not allow for arbitrary shapes.

A natural way to try to do this would be to simply fit a simple linear regression model for the dependent variable onto each independent variable. However, this approach is not suitable because of two reasons. Firstly, it is unclear as to how a single prediction can be made given different *levels* or *classes* of the independent variables. Secondly, since the individual regressions do not take into account the other explanatory (independent) variables, important information about the relation (say, correlation) amongst the independent variables is lost

- resulting in a possibly very horrible model that fails to capture any sensible information.

A better approach is to extend the simple linear regression model so that it can directly accommodate multiple predictors. We do this by giving each predictor a separate slope coefficient (think, *weight*) in a single model. In general, suppose that we have p distinct predictors. Then the multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon \tag{1}$$

where X_j represents the jth predictor and β_j quantifies the association between that variable and the response. We interpret β_j as the average effect on Y of a one unit increase in X_j , holding all other predictors fixed. ϵ denotes the random error component reflecting the difference between the observed and fitted linear relationship.

2.1 Setting up the model

Let an experiment be conducted n times, and the data is obtained as follows:

Observation Number	**Response** y	**Explanatory Variables** $X_1, X_2, \ldots, X_k$
1	y_1	$x_{11}, x_{12}, \ldots, x_{1k}$
2	y_2	$x_{21}, x_{22}, \ldots, x_{2k}$
$\vdots$	$\vdots$	$\vdots$
n	y_n	$x_{n1}, x_{n2}, \ldots, x_{nk}$

We assume the model, same as (1):

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

for each observation. The n-tuples of observations are also assumed to follow the same model, so they satisfy the following system of equations:

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \epsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \epsilon_2 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \epsilon_n
\end{aligned}$$

These n equations can be written in matrix form as:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

or more generally:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2}$$

where $\mathbf{y} = (y_1, y_2, \dots, y_n)^\top$ is an $n \times 1$ vector of observations, $\mathbf{X}$ is an $n \times (k+1)$ matrix of explanatory variables (with a column of ones for the intercept), $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)^\top$ is a $(k+1) \times 1$ vector of coefficients, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of errors.

2.2 Estimating the parameters

This subsection assumes a decent base in Linear Algebra (Mathematics-II). A complete understanding at the first reading is not expected. However, you are expected to have some outline idea of how we arrived at the main results.

Let $\mathcal{B}$ be the set of all possible vectors $\boldsymbol{\beta}$. If there is no further information, then $\mathcal{B}$ is the k-dimensional real Euclidean space. The objective is to find a vector $\boldsymbol{b} = (b_1, b_2, \dots, b_k)'$ from $\mathcal{B}$ that minimizes the sum of squared deviations of $\boldsymbol{\epsilon}_i$, i.e.,

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{3}$$

for given $\mathbf{y}$ and $\mathbf{X}$.

A minimum will always exist as $S(\boldsymbol{\beta})$ is a real-valued, convex, and differentiable function. We can write (3) as

$$S(\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta}.$$

Now, differentiate $S(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}^T\mathbf{y}.$$

Set the derivative to zero:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 \quad \Rightarrow \quad \mathbf{X}^T\mathbf{X}\boldsymbol{b} = \mathbf{X}^T\mathbf{y},$$

where the following result is used:

If $f(\mathbf{z}) = \mathbf{z}^T A\mathbf{z}$ is a quadratic form, where $\mathbf{z}$ is an $m \times 1$ vector and A is any $m \times m$ symmetric matrix, then $\frac{\partial f(\mathbf{z})}{\partial \mathbf{z}} = 2A\mathbf{z}$.

Since it is assumed that $\text{rank}(\mathbf{X}) = k$ (full rank), $\mathbf{X}^T\mathbf{X}$ is positive definite, and the unique solution of the normal equation is

$$\boxed{\boldsymbol{b} = \hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}} \tag{4}$$

which is termed the ordinary least squares estimator (OLSE) of $\boldsymbol{\beta}$, or exactly what we were looking for to have an estimate of the parameters of our model.

Note that since

$$\frac{\partial^2 S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = 2\mathbf{X}^T\mathbf{X}$$

is at least non-negative definite, $\boldsymbol{b}$ indeed minimizes $S(\boldsymbol{\beta})$.

Exercise. When would $\mathbf{X}$ not be full rank (in terms of predictors)? What would be the nature of $\mathbf{X}^T\mathbf{X}$ be in that case? How does that affect the solution $\boldsymbol{b}$, and can you think of some resolution in such a case?

In practice, most programming languages and software packages have libraries that compute $\boldsymbol{b}$ directly.

2.3 Model Assumptions

Up to now, we have made minimal assumptions about the true distribution of the data. In order to pin down the sampling properties of $\hat{\beta}$, we now consider some assumptions:

OLS1: Linearity

$$y_i = x_i^T\beta + u_i \quad \text{and} \quad \mathbb{E}[u_i] = 0$$

This assumption means that the functional relationship between the dependent variable y_i and explanatory variables x_i is linear in parameters, that the error term u_i (residual) enters additively, and that the parameters are constant across individuals i.

OLS2: Independence

$$\{x_i, y_i\}_{i=1}^n \quad \text{i.i.d. (independent and identically distributed)}$$

Here we (trivially) assume that the observations are independently and identically distributed. This assumption is practically guaranteed by random sampling.

OLS3: Exogeneity

(a) $u_i|x_i \sim N(0, \sigma_i^2)$

(b) $u_i \perp\!\!\!\perp x_i$ (independent)

(c) $\mathbb{E}[u_i|x_i] = 0$ (mean independent)

(d) $\text{Cov}(x_i, u_i) = 0$ (uncorrelated)

OLS3a assumes that the error term u_i is normally distributed conditional on the explanatory variables. OLS3b means that the error term is independent of the explanatory variables. OLS3c states that the mean of the error term is independent of the explanatory variables. OLS3d means that the error term and the explanatory variables are uncorrelated. Can you work out which assumption implies which?

OLS4: Error Variance

(a) $\mathbb{V}[u_i|x_i] = \sigma^2 < \infty$ (homoscedasticity)

(b) $\mathbb{V}[u_i|x_i] = \sigma_i^2 = g(x_i) < \infty$ (conditional heteroscedasticity)

OLS4a (homoscedasticity) means that the variance of the error term is constant. OLS4b (conditional heteroscedasticity) allows the variance of the error term to depend on the explanatory variables.

OLS5: Identifiability

$$\mathbb{E}[x_i x_i^T] = Q_{XX} \quad \text{is positive definite and finite}$$
$$\text{rank}(X) = k + 1 < n$$

OLS5 assumes that the regressors are not perfectly collinear, i.e., no variable is a linear combination of the others. For example, there can only be one constant. Intuitively, OLS5 means that every explanatory variable adds additional information. OLS5 also assumes that all regressors (except the constant) have strictly positive variance both in expectations and in the sample, and do not have too many extreme values. (Refer to exercise question posed at the end of the previous section)

2.4 Model Fitting

Similar to the discussion for simple linear regression we can design hypotheses and define test statistics to check whether our model is truly well-defined at some statistical significance. An analogous treatment follows in the case of multiple linear regression, which we skip here. However, it is highly recommended that the reader think about how the null and alternate hypotheses will be defined (Hint: it might be brought up in the interview). We now turn towards the measures of model fit. As in the case of simple linear regression, the two most common numerical measures are RSE and R^2.

2.4.1 R^2 Statistic

The goodness-of-fit of an OLS regression can be measured using the coefficient of determination, R^2, which is given by:

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}} = \frac{\text{SSE}}{\text{SST}} \tag{5}$$

where $\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares, $\text{SSR} = \sum_{i=1}^{n} \hat{u}_i^2$ is the residual sum of squares, and $\text{SSE} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is called the explained sum of squares.

Recall that in simple regression, R^2 is the square of the correlation between the response variable and the explanatory variable. In multiple linear regression, it turns out that R^2 equals $\text{Cor}(Y, \hat{Y})^2$, the square of the correlation between the response variable Y and the fitted values $\hat{Y}$ from the linear model.

One important property of the fitted linear model is that it maximizes this correlation among all possible linear models.

In this case too, R^2 lies by definition between 0 and 1, and it reports the fraction of the sample variation in y that is explained by the regressors x.

The value of R^2 increases by construction with the addition of regressors, even if those regressors are irrelevant. Therefore, R^2 is not a reliable criterion for selecting regressors. A better alternative is the adjusted R^2, which is given by:

$$\text{adj. } R^2 = 1 - \frac{n-1}{n-k-1} \cdot \frac{\text{SSR}}{\text{SST}} \tag{6}$$

where n is the number of observations, and k is the number of regressors in the model.

2.4.2 RSE

The Residual Standard Error in the case of multiple regression can be written as:

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{n-k-1}} \tag{10}$$

3 Model Selection

A good model selection technique will balance goodness of fit with simplicity, which means under a certain performance requirement set by researchers, the best model should be as simple as possible. There are two main reasons we may not be completely satisfied with the least square estimate as in Eq. (4).

- The first is *prediction accuracy*: the least squares estimates often have low *bias* but large *variance*. Prediction accuracy can sometimes be improved

by shrinking or setting some coefficients to zero. By doing so we sacrifice a little bit of bias to reduce the variance of the predicted values, and hence may improve the overall prediction accuracy.

- The second reason is *interpretation*. With a large number of predictors, we often would like to determine a smaller subset that exhibit the strongest effects. In order to get the "big picture," we are willing to sacrifice some of the small details.

In our multiple regression model, we have k regressors. We now aim to possible simplify the model by choosing $l < k$ regressors while maintaining goodness-of-fit. One straightforward strategy could be to use the individual p-values and refit the model with only significant terms. However, defining appropriate p-values is not an easy task. A few model selection strategies are outlined here:

- *Best subset regression* finds for each $l \in \{0, 1, 2, \ldots, k\}$ the subset of size l that gives the smallest residual sum of squares.

- Rather than search through all possible subsets (which becomes infeasible for k much larger than 40), we can seek a good path through them. *Forward-stepwise selection* starts with the intercept, and then sequentially adds into the model the predictor that most improves the fit.

- Other general strategies involve comparing R^2 and adj.R^2, reducing collinearity of predictors, using transformations.

4 Suggested Further Reading

Following are few topics you might wish to explore further.

- Gauss Markov Theorem
- Orthogonal projection interpretation of OLS estimate
- Regression through successive orthogonalisation
- F-statistic for model utility testing
- Qualitative predictors

5 Code Implementation Task

You will now implement a multiple linear regression model to a practical dataset to predict a target variable based on multiple predictors. The dataset for this task is the `Boston Housing` dataset; it can be found in the same drive folder as this document.

The dataset has columns corresponding to:

- *crim*: Per capita crime rate by town.
- *zn*: Proportion of large residential lots (over 25,000 sq. ft.).
- *indus*: Proportion of non-retail business acres per town.
- *Chas*: Binary variable indicating if the property is near Charles River (1 for yes, 0 for no).
- *nox*: Concentration of nitrogen oxides in the air.
- *rm*: Average number of rooms per dwelling.
- *age*: Proportion of old owner-occupied units built before 1940.
- *dis*: Weighted distances to Boston employment centers.
- *rad*: Index of accessibility to radial highways.
- *tax*: Property tax rate per $ 10,000.

The steps to complete the task are:

1. **Data Preparation:**
   - Load the dataset and perform an initial exploration to understand its structure and contents.
   - Clean the data as you deem appropriate.
   - Split the dataset into training and testing subsets (e.g., 80% training, 20% testing).
2. **Model Implementation:**
   - Implement multiple linear regression model using `scikit-learn` in Python (or any other library in your preferred choice of programming language).
   - Train the model using the training dataset.
   - Predict the target variable on the testing dataset.

3. **Model Evaluation:**

   - Evaluate the model performance using appropriate metrics, such as R-squared, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).
   - You may consider using various models (see: Ridge, Lasso, Best subset etc.), and choose the best model according to you.

4. **Model Diagnostics** (Optional)

   - Check for multicollinearity among predictors using Variance Inflation Factor (VIF) or correlation matrices.
   - Perform hypothesis testing on the coefficients to assess their significance.
   - Assess the model assumptions (linearity, homoscedasticity, normality of residuals) and identify any violations.

In the interview that shall follow this task submission, you will be expected to answer some theoretical questions from the content in this document and also questions related to your implementation.

Have fun!