

# Linear Regression for Quantitative Finance Interviews

Edmund A. Berry

March 5, 2015

## 1 Introduction

We begin by defining terms.

Consider some variable you are trying to understand,  $y$ . You have a dataset that contains  $n$  data points, and each data point,  $i$  has its own value of  $y$ ,  $y_i$ . You describe this collection of values for  $y$  in a vector representation, as follows:

$$y = \langle y_1, y_2, \dots, y_n \rangle \quad (1)$$

You decide to try to predict the value of this variable as a linear combination of a series of  $k$  predictor variables,  $X$ .  $X$  is best described as an  $n \times k$  matrix, where each of the  $n$  rows corresponds to a single data point, and each of the  $k$  columns corresponds to a single predictor variable. You can refer to a single row (i.e. a single data point with  $k$  variables) as  $X_i$ . These rows can also be described in a vector representation, as follows:

$$X_i = \langle X_{i1}, X_{i2}, \dots, X_{ik} \rangle \quad (2)$$

A linear combination of  $k$  predictor variables requires  $k$  coefficients. You call these  $k$  coefficients  $\beta$ , and you also describe them using a vector representation:

$$\beta = \langle \beta_1, \beta_2, \dots, \beta_k \rangle \quad (3)$$

Call your final prediction  $\hat{y}$ . The linear combination of your  $X$  predictor variables with  $\beta$  coefficients is defined in vector notation as:

$$\hat{y} = X\beta \quad (4)$$

Equivalently, you can write this using subscripts:

$$\hat{y}_i = \sum_{j=1}^k X_{ij} \cdot \beta_j \quad (5)$$

Your estimate  $\hat{y}$  might not be perfectly equal to  $y$ , so you should define some residual variable,  $\epsilon$  to describe the difference:

$$\epsilon_i = y_i - \sum_{j=1}^k X_{ij} \cdot \beta_j \quad (6)$$

This allows you to finally describe the true variable,  $y$ , in terms of your predictor variables:

$$y_i = \sum_{j=1}^k X_{ij} \cdot \beta_j + \epsilon_i \quad (7)$$

Equivalently, in the vector representation:

$$y = X\beta + \epsilon \quad (8)$$

The goal of the regression is to pick some vector of coefficients  $\beta$  (length  $k$ ) such that the residuals of our prediction,  $\epsilon$  (length  $n$ ), are minimized.

## 2 Regression

There are many metrics to use to determine whether a given vector of coefficients,  $\beta$  minimizes  $\epsilon$ . By far the most common is the sum of the difference of squares,  $S$ , which is defined as follows:

$$S(\beta) = \sum_{i=1}^n \left| y_i - \sum_{j=1}^k X_{ij} \cdot \beta_j \right|^2 \quad (9)$$

Equivalently, in the vector representation:

$$S(\beta) = \|y - X\beta\|^2 \quad (10)$$

One nice feature of the sum of the difference of squares is that the minimization problem has a unique solution, as long as the  $k$  columns of matrix  $X$  are linearly independent. This solution is derived below.

We begin by redefining  $S$  in terms of the residuals,  $\epsilon$ :

$$S(\beta) = \sum_{i=1}^n \epsilon_i^2 \quad (11)$$

Equivalently, in the vector representation:

$$S(\beta) = \|\epsilon\|^2 \quad (12)$$

If we think of  $\beta$  as a surface, then  $S(\beta)$  is minimized when the gradient vector of  $S(\beta)$  has magnitude zero. There is a geometric argument for this: if the gradient is not zero, then there is a direction that we can move such that the gradient may be reduced. The gradient,  $\nabla S(\beta)$ , is a vector of length  $k$ , where each element of the vector is defined as a partial derivative:

$$\frac{\partial S}{\partial \beta_j} = 2 \sum_{i=1}^n \epsilon_i \frac{\partial \epsilon_i}{\partial \beta_j} \quad (13)$$

Recalling the definition of  $\epsilon$  from Equation 6, the partial derivate is easy to determine:

$$\frac{\partial \epsilon_i}{\partial \beta_j} = -X_{ij} \quad (14)$$

This allows us to completely determine each element of the gradient of  $S(\beta)$ :

$$\frac{\partial S}{\partial \beta_j} = -2 \sum_{i=1}^n \epsilon_i \cdot X_{ij} \quad (15)$$

$$\frac{\partial S}{\partial \beta_j} = -2 \sum_{i=1}^n \left( y_i - \sum_{l=1}^k X_{il} \cdot \beta_l \right) \cdot X_{ij} \quad (16)$$

We define  $\hat{\beta}$  as the value of  $\beta$  that minimizes  $S(\beta)$  such that  $\frac{\partial S}{\partial \beta_j} = 0$  for all  $j$  between 1 and  $k$ :

$$-2 \sum_{i=1}^n \left( y_i - \sum_{l=1}^k X_{il} \cdot \hat{\beta}_l \right) \cdot X_{ij} = 0 \quad (17)$$

We can divide out a factor of 2 and rearrange this. For all  $j$  between 1 and  $k$ :

$$\sum_{i=1}^n \sum_{l=1}^k X_{ij} X_{il} \hat{\beta}_l = \sum_{i=1}^n X_{ij} y_i \quad (18)$$

Equivalently, in vector notation:

$$(X^T X) \hat{\beta} = X^T y \quad (19)$$

Solving for  $\hat{\beta}$ :

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (20)$$

### 3 Explained variance

We define the number of degrees of freedom,  $\text{ndof}$ , to be the difference between the number of entries in the dataset,  $n$ , and the number of discriminating variables,  $k$ :

$$\text{ndof} = n - k \quad (21)$$

Note that in classical linear regression  $n$  must be greater than  $k$ , or else the dataset could be fit perfectly. With this in mind, we can define the residual standard deviation,  $\hat{\sigma}$ , as follows:

$$\hat{\sigma} = \sqrt{\sum_{i=1}^n \frac{\epsilon_i^2}{(n - k)}} \quad (22)$$

Here,  $\hat{\sigma}$  summarizes the scale of the residuals. You can think of this as a measure of the average distance that the linear model output falls from the true value.

If we call the true standard deviation of the data  $s_y$ , then we can define  $R^2$  such that:

$$R^2 = 1 - \frac{\hat{\sigma}^2}{s_y^2} \quad (23)$$