

# Linear Regression and Least Squares Estimation

Guy Lebanon

April 25, 2007

Linear regression is probably the most popular model for predicting a RV  $Y$  based on RVs  $X_1, \dots, X_{p-1}$ . Specifically, the linear regression model is a conditional model for  $Y|X$  that assumes

$$Y = \beta_0 + \sum_{i=1}^{p-1} \beta_i X_i + \epsilon, \quad \text{or with setting } X_0 \equiv 1, \quad Y = \sum_{i=0}^{p-1} \beta_i X_i + \epsilon = \langle \beta, X \rangle + \epsilon \quad (1)$$

where  $\epsilon$  is a RV with mean 0 that is independent of  $X_1, \dots, X_{p-1}$ . We will usually use the second form of (1) since it will simplify the notation. An equivalent formulation of (1) that is sometime used to define linear regression is  $\mathbf{E}(Y|X) = \langle \beta, X \rangle$ . In this note we will assume that  $\epsilon \sim N(0, \sigma^2)$  and hence the linear regression assumption translates to  $Y|X \sim N(\langle \beta, X \rangle, \sigma^2)$ . The parameter vector  $\beta$  defines a linear hyperplane in  $\mathbb{R}^p$  (or subspace if we use  $X_0 \equiv 1$ ) which describes the mean response  $Y$  given  $X$ .

The variables  $X$  most commonly represent numeric ordinal quantities. However, they may represent binary categorical predictor, in which case  $X_i \in \{0, 1\}$ . To represent a more general categorical variable taking values in  $\{1, \dots, c\}$  (such as color) we use  $c - 1$  indicator variables such as  $X_i = 1_{\{i\}}$  for  $i = 1, \dots, c - 1$  (the last class is automatically chosen if all variables above are “turned off” or set to 0. The variables  $X_i$  may be dependent or related to each other or may represent nonlinear dependencies, for example  $X_1 = X'_1, X_2 = (X'_2)^2, X_3 = X'_1 X'_2$ . The above situation, called polynomial regression, can still be studied using the framework of (1) since we primarily look at the conditional distribution and we still have  $Y|X \sim N(\langle \beta, X \rangle, \sigma^2)$ . Furthermore, we can also be primarily interested in a transformed response variable  $W$  for which  $Y = f(W)$  linearly regresses  $X$ , for example  $W = \exp(\langle \beta, X \rangle + \epsilon)$ . On the other hand, linear regression cannot handle arbitrary nonlinear combinations of predictors and parameters for example  $Y = \beta_1 \exp(\beta_2 X) + \epsilon$ .

The data in regression analysis is usually multiple iid samples  $(X^{(i)}, Y^{(i)}) \sim p_{Y|X} p_X, i = 1, \dots, n$  where  $p_{Y|X}$  is defined by (1) and  $p_X$  is an arbitrary marginal. In the case of observational data  $p_X$  corresponds to nature and in the case of experimental data  $p_X$  corresponds to the experimental design. We represent the conditional relationship between the training data  $(X^{(i)}, Y^{(i)}), i = 1, \dots, n$  in matrix form  $\mathbf{Y} = \mathbf{X}\beta + \epsilon$  where  $\mathbf{Y} = (Y^{(1)}, \dots, Y^{(n)}) \in \mathbb{R}^{n \times 1}, \mathbf{X} \in \mathbb{R}^{n \times p}$  is a matrix whose rows are  $X^{(i)}$ , and  $\epsilon \sim N(0, \sigma^2 I)$ . The matrices  $\mathbf{X}, \mathbf{Y}, \epsilon$  corresponding to the  $n$  training set instances appear in bold face to avoid confusion with the random variables  $X, Y, \epsilon$ . Thus  $\mathbf{Y}|\mathbf{X} \sim N(\mathbf{X}\beta, \sigma^2 I) = p_{Y|X}^n (Y_1 = Y^{(1)}, \dots, Y_n = Y^{(n)} | X_1 = X^{(1)}, \dots, X_n = X^{(n)})$ .

The standard way of obtaining the parameter vector  $\beta$  is by minimizing the sum of square deviations of the observations from the model predictions

$$\hat{\beta} = \arg \min_{\beta} \text{SSE}(\beta) \quad \text{where} \quad \text{SSE}(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|^2 = \sum_{i=1}^n (Y^{(i)} - \langle \beta, X^{(i)} \rangle)^2$$

which is equivalent to the maximum conditional likelihood estimator  $\hat{\beta} = \arg \max_{\beta} \prod p(Y^{(i)}|X^{(i)})$ . The minimization above is that of a quadratic function and has a closed form expression, derived below. Differ-

entiating the SSE criteria with respect to  $\beta$  and setting it to 0 gives the set of normal equations

$$\begin{aligned}\nabla_{\beta} \text{SSE}(\beta) = 0 &\Leftrightarrow \sum_i (Y^{(i)} - \langle \beta, X^{(i)} \rangle) X_j^{(i)} = 0 \quad \forall j \quad \text{or} \quad \mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{Y} \\ &\Rightarrow \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.\end{aligned}$$

In the special case when the columns  $u_1, \dots, u_n$  of  $\mathbf{X}$  are orthogonal the components of the least squares projection  $\hat{\beta}$  become the standard orthogonal basis projections  $\hat{\beta}_j = \frac{\langle u_j, \mathbf{Y} \rangle}{\|u_j\|^2}$ .

The mle estimator  $\hat{\beta}$  leads to the predicted values associated with an arbitrary  $X$ :

$$\hat{Y} = \langle \hat{\beta}, X \rangle = \langle (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}, X \rangle. \quad (2)$$

It is interesting to examine the predictions given by  $\hat{\beta}$  to the training data  $\mathbf{X}$ :

$$\hat{\mathbf{Y}} = \mathbf{X} \hat{\beta} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{H} \mathbf{Y} \quad \text{for} \quad \mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T. \quad (3)$$

We assume from here on that the training data  $\mathbf{X}$  is fixed and examine various distributions related to  $\hat{\beta}$ , conditioned on  $\mathbf{X}$ . Since the least squares estimation of  $\hat{\beta}$  reconstructs  $\mathbf{Y}$  as a linear combination of the columns of  $\mathbf{X}$ ,  $\hat{\beta}$  is the projection of  $\mathbf{Y}$  on the column space of  $\mathbf{X}$ . The projection is represented by the projection matrix  $\mathbf{H}$  (symmetric and idempotent i.e.  $\mathbf{H}^T = \mathbf{H}$ ,  $\mathbf{H}^2 = \mathbf{H}$ ). If  $\mathbf{X}^T \mathbf{X}$  is not invertible, a generalized inverse can be used in calculating  $\hat{\beta}$ . There are many generalized inverses, but they all lead to the same least squares estimate  $\hat{\beta}$  which is uniquely defined as subspace projection. The projection onto the complement subspace represented by the projection matrix  $\mathbf{I} - \mathbf{H}$  leads to the concept of residuals  $\mathbf{e} \stackrel{\text{def}}{=} \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H}) \mathbf{Y}$ .

By the linearity of  $\hat{\beta}$  in the random variables  $\mathbf{Y}$ , we are able to study distributions related to  $\hat{\beta}$ . Since

$$\begin{aligned}\mathbf{E}(\hat{\beta}) &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{E}(\mathbf{Y}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \beta = \beta \\ \text{Var}(\hat{\beta}) &= \text{Var}((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \text{Cov}(\mathbf{Y}) (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \sigma^2 \mathbf{I} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}\end{aligned}$$

by linearity of  $\hat{\beta}$  in  $\mathbf{Y}$  we have that (conditioned on  $\mathbf{X}$ )

$$\hat{\beta} \sim N(\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}). \quad (4)$$

The following theorem provides a strong motivation for using  $\hat{\beta}$ . It can be strengthened to show that  $\hat{\beta}$  is the UMVUE.

**Theorem 1** (Gauss-Markov).  *$\hat{\beta}$  is BLUE (best linear unbiased estimator) i.e. among all unbiased linear estimators it has the smallest variance.*

*Proof.* We already know that  $\hat{\beta}$  is linear and unbiased. Let  $\tilde{\beta} = \mathbf{A} \mathbf{Y}$  be a any other unbiased linear estimator of  $\beta$ . Then to prove the theorem we need to show that  $\text{Var}(\tilde{\beta}) - \text{Var}(\hat{\beta})$  is positive semidefinite. Since

$$\tilde{\beta} = \mathbf{A} \mathbf{Y} = (\mathbf{D} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) (\mathbf{X} \beta + \epsilon) = (\mathbf{D} \mathbf{X} + \mathbf{I}) \beta + (\mathbf{D} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \epsilon$$

(for some matrix  $\mathbf{D}$ ) for  $\tilde{\beta}$  to be unbiased we must have  $\mathbf{D} \mathbf{X} = 0$ . We then have

$$\begin{aligned}\text{Var}(\tilde{\beta}) &= \mathbf{E}(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^T = \mathbf{E}(\mathbf{D} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \epsilon \epsilon^T (\mathbf{D}^T + \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}) \\ &= \sigma^2 (\mathbf{D} \mathbf{D}^T + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} + \mathbf{D} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{D}^T) \\ &= \sigma^2 \mathbf{D} \mathbf{D}^T + \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \Rightarrow \text{Var}(\tilde{\beta}) - \text{Var}(\hat{\beta}) \quad \text{is positive semi-definite.}\end{aligned}$$

□