Linear Regression Fundamentals

v1.0

Chandler Zachary

The purpose of this document is to provide a brief introduction to the novice and reminder to myself of the most fundamental intuition of linear regression. These notes assume some upper division math, probability, and statistics. For brevity, I do not give definitions of terms here, and I skip large portions of proofs and computations which can easily be found in suitable texts.

The Probability of Linear Regression

When we perform linear regression, we have measurements generated by two or more random variables, X and Y, that we suspect vary together, and we want to use these data to infer something about the strength of the relationship between these two random variables. We could do that simply by using a correlation coefficient, and that would probably be suitable for plenty of applications. Linear regression does more. It expands the understanding of the relationship by allowing us to specify a *functional* relationship of one variable in terms of the other.

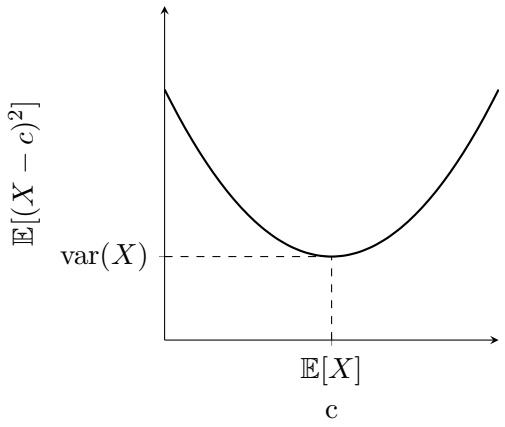

Figure 1: Minimizing mean squared error

Consider the model

$$Y = b_0 + b_1 X + u,$$

where $u \sim N(0, \sigma^2)$, and b_0, b_1 are parameters to be estimated. Linear regression obtains the linear estimator $b_0 + b_1 X$ of Y by solving the minimization problem

$$\min_{b_0, b_1} \quad \mathbb{E}[(Y - b_0 - b_1 X)^2].$$

To find the solutions for b_0, first fix b_1. Then,

$$\begin{aligned} & \frac{d}{db_0}\mathbb{E}[(Y-b_0-b_1X)^2] \overset{FOC}{=} 0 \\ \Rightarrow \quad & 0 = -2\mathbb{E}[(Y-b_0-b_1X)] \\ & \quad = -2(\mathbb{E}[Y]-b_0-b_1\mathbb{E}[X]) \\ \Rightarrow \quad & b_0 = \mathbb{E}[Y]-b_1\mathbb{E}[X] \end{aligned}$$

With this value in hand, b_1 is found by:

$$\begin{aligned} \mathbb{E}[(Y-b_0-b_1X)^2] &= \mathbb{E}[(Y+b_1X-\mathbb{E}[Y]-b_1\mathbb{E}[X])^2] \\ &= \mathbb{E}[((Y-\mathbb{E}[Y])-b_1(X-b_1\mathbb{E}[X]))^2] \\ &= \mathbb{E}[(Y-\mathbb{E}[Y])^2]+b_1^2\mathbb{E}[(X-\mathbb{E}[X])^2]-2b_1\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\ &= \sigma_Y^2+b_1^2\sigma_X^2-2b_1cov(X,Y) \\ \Rightarrow \quad \frac{d}{db_1}\sigma_Y^2+b_1^2\sigma_X^2-2b_1cov(X,Y) &\overset{FOC}{=} 0 \\ = 2b_1\sigma_X^2-2cov(X,Y) & \\ \Rightarrow \quad b_1 = \frac{cov(X,Y)}{\sigma_X^2} &= \rho\frac{\sigma_Y}{\sigma_X} \end{aligned}$$

Thus, the least mean squares linear estimator of Y given X is

$$\mathbb{E}[Y]+\frac{cov(X,Y)}{\sigma_X^2}(X-\mathbb{E}[X]),$$

and the mean squared estimation error is

$$\mathbb{E}[(Y-\hat{b}_0-\hat{b}_1X)^2] = (1-\rho^2)\sigma_Y^2.$$

The Statistics of Linear Regression

$$y \sim N(b_0+b_1x_i, \sigma^2)$$

$$\hat{b}_0 = \bar{y}-\hat{b}_1\bar{x}, \quad \hat{b}_1 = \frac{\sum_i x_iy_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2}$$

$$\sigma_{\hat{b}_1}^2 = \frac{\sigma^2}{\sum(x_i-\bar{x})^2}, \quad \sigma_{\hat{\boldsymbol{b}}}^2 = \boldsymbol{\sigma}^2(\boldsymbol{x}^T\boldsymbol{x})^{-1}$$

$$\hat{\sigma}_{\hat{b}_1}^2 = \frac{\hat{\sigma}^2}{\sum(x_i-\bar{x})^2}, \quad \hat{\sigma}_{\hat{\boldsymbol{b}}}^2 = \hat{\boldsymbol{\sigma}}^2(\boldsymbol{x}^T\boldsymbol{x})^{-1}$$

$$t = \frac{\hat{b}-b}{\hat{\sigma}_{\hat{b}}} \sim t_{n-k}, \quad \hat{b} \pm t_{\alpha,k}\hat{\sigma}_{\hat{b}}$$

$$F = \frac{\sum_i(\hat{y}_i-\bar{y})^2/(m-1)}{\sum_i \hat{e}_i^2/(n-m)} \sim F_{m-1,n-m}, \quad \frac{\left(\left(\sum_i(\hat{y}_i-\bar{y})^2\right)_r - \left(\sum_i(\hat{y}_i-\bar{y})^2\right)_u \Big/ k}{\left(\sum_i(\hat{y}_i-\bar{y})^2\right)_u/(n-m)} \sim F_{k,n-m}$$

The Geometry of Linear Regression

Given a matrix system

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{b},$$

where $\boldsymbol{x}$ is $n \times k$ and $\boldsymbol{y} \in \mathbb{R}^n$, a least squares solution of $\boldsymbol{y} = \boldsymbol{x}\boldsymbol{b}$ is $\hat{\boldsymbol{b}} \in \mathbb{R}^k$ such that

$$||\boldsymbol{y} - \boldsymbol{x}\hat{\boldsymbol{b}}||^2 \leq ||\boldsymbol{y} - \boldsymbol{x}\boldsymbol{b}||^2, \quad \forall \boldsymbol{b} \in \mathbb{R}^k.$$

We might also write that $\hat{\boldsymbol{b}}$ solves

$$\min_{\boldsymbol{b}} \quad ||\boldsymbol{y} - \boldsymbol{x}\boldsymbol{b}||^2.$$

In this framework, n is the number of observations, and k is the number of regressors. The solution to the minimization problem is:

$$\begin{aligned} \mathbf{0} &\overset{FOC}{=} -2\boldsymbol{x}^T(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{b}) \\ \mathbf{0} &= -\boldsymbol{x}^T\boldsymbol{y} + \boldsymbol{x}^T\boldsymbol{x}\boldsymbol{b} \\ \hat{\boldsymbol{b}} &= (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y} \end{aligned}$$

The matrix $\boldsymbol{x}$ is composed of column vectors of data, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k$, where each $\boldsymbol{x}_i$ is some variable that contributes to explaining the outcome of interest. The column vector $\boldsymbol{y}$ is a vector of observed measurements for the outcome of interest. If sampling and experimentation have truly enforced randomization, then each $\boldsymbol{x}_i$ will be orthogonal to the others, thus contributing *independent* information.

The least squares solution $\hat{\boldsymbol{b}}$ defines a vector of weights $(\boldsymbol{x}_i^T\boldsymbol{x}_i)^{-1}\boldsymbol{x}_i^T\boldsymbol{y}$ such that

$$\hat{\boldsymbol{y}} = \frac{\boldsymbol{x}_1^T\boldsymbol{y}}{\boldsymbol{x}_1^T\boldsymbol{x}_1}\boldsymbol{x}_1 + \cdots + \frac{\boldsymbol{x}_k^T\boldsymbol{y}}{\boldsymbol{x}_k^T\boldsymbol{x}_k}\boldsymbol{x}_k.$$

This vector defines an orthogonal projection $\hat{\boldsymbol{y}}$ of $\boldsymbol{y}$ onto a subspace of $\mathbb{R}^k$, where k is the number of independent explanatory variables. This orthogonal projection $\hat{\boldsymbol{y}}$ is the sum of projections of $\boldsymbol{y}$ onto one-dimensional subspaces that are each mutually orthogonal. In this way, $\hat{\boldsymbol{b}}$ "constructs" $\hat{\boldsymbol{y}}$ from the columns of $\boldsymbol{x}$. The distance $\hat{\boldsymbol{e}} = ||\boldsymbol{y} - \hat{\boldsymbol{y}}||$ is called the residual, and least squares finds the $\hat{\boldsymbol{b}}$ that corresponds the the minimum residual such that $\boldsymbol{y} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{e}}$.

The subspace of $\mathbb{R}^k$ is the column space of $\boldsymbol{x}$. The vector $\boldsymbol{y}$ does not fit into this column space, so we must find a vector that does. The minimization solution identifies the vector of weights $\hat{\boldsymbol{b}}$ that forms $\hat{\boldsymbol{y}}$ in the column space of $\boldsymbol{x}$. This $\hat{\boldsymbol{y}}$ is orthogonal to the error $\boldsymbol{e}$. Since $\hat{\boldsymbol{y}}$ is in the column space of $\boldsymbol{x}$, it must be true that $\hat{\boldsymbol{y}}^T\boldsymbol{e} = 0$ and $\boldsymbol{x}^T\boldsymbol{e} = 0$.

The Calculus of Linear Regression

Consider the relationship

$$y_i = a_0 + a_1 x_{1_i} + u_i.$$

In this model, the effect of x_1 on y can be found using the first derivative:

$$\frac{dy}{dx_1} = a_1.$$

This can be interpreted as the amount of change in y for a very small change in x_1. If we have reason to believe that some other factor contributes to variation in y, we can extend the model:

$$y_i = b_0 + b_1 x_{1_i} + b_2 x_{2_i} + e_i.$$

Here, we model the effect of x_2 on y in addition to x_1. This allows us to do something very important with partial derivatives: we can isolate the partial effects of x_1 and x_2 holding the other constant. These are called *partial effects*, and they are, respectively:

$$\frac{\partial y}{\partial x_1} = b_1, \quad \frac{\partial y}{\partial x_2} = b_2.$$

Here, b_1 is the change in y given a very small change in x_1, holding constant (also, "controlling for") x_2, and b_2 is the change in y given a very small change in x_2, holding constant x_1. For any given level of x_1, x_2 the slopes b_2, b_1 give us the change in y. This is what we mean by "holding constant" or "controlling for" the other variable. By modeling explanatory variables this way, we have taken advantage of additional variation we believe to be present in the data.

If we think x_1 and x_2 change together, we must take this a step further. This implies a composite function, $y = f(x_1, x_1(x_2))$, and the total derivative with respect to x_1 is:

$$\frac{dy}{dx_1} = \frac{\partial y}{\partial x_1} + \frac{\partial y}{\partial x_2}\frac{\partial x_2}{\partial x_1}.$$

This illuminates the relationship between a_1 and b_1. Notice:

$$a_1 = \frac{dy}{dx_1} = \frac{\partial y}{\partial x_1} + \frac{\partial y}{\partial x_2}\frac{\partial x_2}{\partial x_1} = b_1 + b_2\frac{\partial x_2}{\partial x_1}.$$

We say b_1 is the effect of x_1 on y, conditioned on x_2, whereas a_1 is the unconditioned effect of x_1 on y. The importance of this relationship is that changing estimates of $\hat{a}_1$ when new regressors are added (i.e. $\hat{a}_1 \neq \hat{b}_1$) indicate that the new regressor is correlated with both y and x_1.

This discussion is incomplete without considering the residuals in the context of additional regressors. Intuitively, one should recognize that $\sum_i \hat{e}_i^2 \leq \sum_i \hat{u}_i^2$ with equality if $b_2 = 0$. The proof is a straightforward substitution. Since we're considering the case where x_1 is a function of x_2, we can model the relationship:

$$x_{1_i} = d_1 x_{2_i} + r_{1_i}.^1$$

We can use the relationship

$$x_{1_i} = \hat{x}_{1_i} + \hat{r}_{1_i}$$

and the substitution:

$$\hat{x}_{1_i} = \hat{d}_1 x_{2_i} = \frac{cov(x_1, x_2)}{var(x_2)} x_{2_i},$$

[1] I've demeaned the model here.

to note that the regression of x_1 on x_2 isolates everything about x_2 that explains x_1, leaving the unexplained stuff in $\hat{r}_1$. This observation allows us to arrive at the relationship

$$\hat{b}_1 = \frac{\sum_i \hat{r}_{1_i} y_i}{\sum_i \hat{r}_{1_i}^2}.$$

This equation says that the coefficient estimate is equal to the covariance of y and everything about x_1 not explained by x_2 scaled by the variance of everything about x_1 not explained by x_2 *when we've controlled for the effect of* x_2 (i.e. - removed the effect of x_2). The intuition for this can be seen by rearranging the relationship:

$$b_1 = a_1 - b_2 \frac{\partial x_2}{\partial x_1}.$$

I've omitted a rather lengthy proof here for the sake of brevity, so the reader should take some time to work through it to understand the relationships.