# R1: Overview of regression

Yaroslav Mukhin

Feb 8, 2015

Abstract. We consider the *population* and put the linear model in perspective of its probabilistic framework. The goal is to clarify the target of OLS estimation.

Let $(Y, X)$ be $L^2(\Omega, \mathcal{F}, P)$ random variables. All we could possibly want to know about these quantities is contained in the joint distributions $F_{Y,X}$. This can be thought of the following collection of probabilities

$$\{P\{(Y, X) \in A\} \ ; \ A \text{ rectangle in } \mathbb{R}^{p+1}\}.$$

However our interest is usually asymmetric in that we are primarily interested in the random variable $Y$, and use random vector $X$ only as a tool to reduce uncertainty about $Y$. In this case, we don't need to know probabilities of events that allow for uncertainty in both $Y$ and $X$, but only those events that prescribe a fixed value $x$ to $X$, whatever that value might happen to be. This information is contained in the collection of conditional distributions $F_x$ [1] that describe uncertainty about $Y$, when $X$ is known to have realized as $x$. Equivalently,

$$\{P(Y \in A | X = x) \ ; \ A \text{ rectangle in } \mathbb{R}^1, \ x \in \mathbb{R}^p\}.$$

A distribution of $Y$ for each value $x$ is a lot of information. The regression function $f(x)$ that assigns to each $x$ the mean of the conditional distribution $F_x$ is a nice way to summarize and visualize the dependence of randomness in $Y$ on $X$: [2]

$$f(x) \coloneqq \mathrm{E}[Y | X = x].$$

There is a very nice geometric interpretation of regression that relies on the notion of orthogonality in linear spaces. If we decompose

$$L^2(\Omega, \mathcal{F}, P) = L^2(\Omega, \sigma(X), P) \oplus L^2(\Omega, \sigma(X), P)^{\perp},$$

then regression $\mathrm{E}[Y|X]$ is the orthogonal projection of $Y$ onto $L^2(\Omega, \sigma(X), P)$:

$$Y \equiv \mathrm{E}[Y|X] + \varepsilon,$$

where the residual $\varepsilon$ is orthogonal to *all* functions of $X$: for every $Z \in L^2(\Omega, \sigma(X), P)$ we have $\mathrm{E}\varepsilon Z = 0$. We make one more use of orthogonality to express the regression function as an infinite vector of coordinates. Let $\{\psi_j\}_{j \geq 1}$ be an orthonormal basis for $L^2(\Omega, \sigma(X), P)$ chosen in the way that $\psi_1, \ldots, \psi_p$ span all linear combinations of random variables $X_1, \ldots, X_p$, and decompose $f(X)$ into its Fourier series

$$f(X) = \sum_{j=1}^{p} \alpha_j \psi_j + \sum_{j > p} \alpha_j \psi_j. \tag{1}$$

If we think of $\psi$ as appropriately normalized polynomials in $X$, then the first $p$ terms is the linear part of $f$ and the rest of the series is the nonlinear part of the regression function. Finally, when we estimate a regression, we project $Y$ only onto the first $p$ terms of above decomposition which

[1] As a side remark, it is at all not straightforward to give meaning to the above set. The difficulty stems from the fact that the collection of events we must assign conditional probabilities to is uncountable , whereas measure theory allows for only countable operations. The construction is called regular conditional probabilities and is possible only under topological conditions that allow to reconcile this gap.

[2] Continuing with the side remark, conditional probabilities are defined in terms of the conditional expectation, but this construction does not guaranty that $A \mapsto P(Y \in A | X = x)$ is a probability measure for any given $x$.

can be called the linear regression function.

In the above discussion we made extensive use of orthogonality, yet regression coefficients are traditionally defined though a minimization problem. In fact, the two definitions are equivalent! This is a very important and useful characterization of orthogonal decomposition. First observe that eq. (1) is another orthogonal decomposition:

$$L^2(\Omega, \sigma(X), P) = \text{span}(X_1, \ldots, X_p) \oplus \text{span}(\psi_j)_{j>p}$$

where we have replaced the first $p$ terms of the orthonormal basis with the original linear terms that are not necessarily orthogonal. In this new basis, the linear part of the regression function has coordinates $\beta$. We will see how this effects the coefficients of linear regression shortly. Define the linear regression coefficient

$$\mathrm{b} \coloneqq \arg\min_{b \in \mathbb{R}^p} \mathrm{E}[f(X) - b'X]^2 \tag{2}$$

and residual

$$\epsilon(\mathrm{b}) \coloneqq f(X) - \mathrm{b}'X.$$

**Claim**: $b$ solves (2) iff $\epsilon(b)$ is orthogonal to $\text{span}(X_1, \ldots, X_p)$. This follows immediately from the following decomposition of the norm of $f(X)$: [3]

$$\|f(X) - b'X\|^2 = \|f(X) - \beta'X\|^2 + 2\langle f(X) - \beta'X, \beta'X - b'X\rangle + \|\beta'X - b'X\|^2 \tag{3}$$

First recall that $\beta$ is defined as the orthogonal projection of $f(X)$ onto $\text{span}(X_1, \ldots, X_n)$, so that the residual $f(X) - \beta'X$ is orthogonal to $\text{span}(X_1, \ldots, X_p)$ by construction. Thus, the middle term in eq. (3) is zero for any choice of coefficient $b$ because $\beta'X - b'X \in \text{span}(X_1, \ldots, X_p)$, and we see that $\|\epsilon(b)\| \geq \|\epsilon(\beta)\|$. Conversely, if $b$ solves (2), decomposition eq. (3) implies that $\beta'X = b'X$ (the middle term is still zero by definition of $\beta$!) and therefore $\epsilon(b) = \epsilon(\beta)$ is orthogonal to $\text{span}(X_1, \ldots, X_p)$.

Putting all of the above together, we have produced the following orthogonal decomposition of random variable $Y$

$$Y = \beta'X + \sum_{j>p} \alpha_j \psi_j + \varepsilon. \tag{4}$$

The target of linear regression is the first term on the right. It is characterized by minimizing the size of the residual $\epsilon + \varepsilon$ and, equivalently, by producing a residual that is orthogonal to all of $\text{span}(X_1, \ldots, X_p)$.

Above we used $X_1, \ldots, X_p$ as basis to define linear regression coefficient $\beta$. This choice of basis is natural because it is how we observe data, but does not immediately allow us to interpret coefficients $\beta$. If we partition $X = (D, W)$ and $\beta = (\beta_1, \beta_2)$, we can project orthogonally any variable of interest

[3] We use $\|Z\|^2$ to denote $\mathrm{E}Z^2$, and $\langle Y, Z\rangle$ to denote $\mathrm{E}[YZ]$.

onto span$(W)$:

$$\begin{aligned} Y &= W'\gamma_{YW} + \widetilde{Y}, \quad \widetilde{Y} \text{ orthogonal to span}(W) \\ D &= W'\gamma_{DW} + \widetilde{D}, \quad \widetilde{D} \text{ orthogonal to span}(W) \\ \epsilon + \varepsilon &= 0 + \epsilon + \varepsilon \\ W &= W + 0, \end{aligned}$$

third line above follows by recalling that $\epsilon$ and $\varepsilon$ are orthogonal to span$(X)$. Now substitute above decomposition with respect to subspace spanned by $W$ into the regression equation (4):

$$W'\gamma_{YW} + \widetilde{Y} = \beta_1' W'\gamma_{DW} + \beta_1'\widetilde{D} + \beta_2' W + \epsilon + \varepsilon$$

rearrange

$$\widetilde{Y} = \beta_1'\widetilde{D} + \left[\beta_1' W'\gamma_{DW} + \beta_2' W + \epsilon + \varepsilon - W'\gamma_{YW}\right]$$

and note that the term in brackets is orthogonal to span$(\widetilde{D})$! By our equivalence result it follows that $\beta_1$ is the projection coefficient of $\widetilde{Y}$ onto span$(\widetilde{D})$. So we can think of $\beta_1$ as the linear term of regression of $\left[Y - \text{ linear part in } W\right]$ on $\left[D - \text{ linear part in } W\right]$.

# 14.382 Econometrics
Spring 2017