Chapter 6: Regression

By Joan Llull[*]

Probability and Statistics.
QEM Erasmus Mundus Master. Fall 2016

Main references:

— Goldberger: 13.1, 14.1-14.5, 16.4, 25.1-25.4, (13.5), 15.2-15.5, 16.1, 19.1

I. Classical Regression Model

A. *Introduction*

In this chapter, we are interested in estimating the conditional expectation function $\mathbb{E}[Y|X]$ and/or the optimal linear predictor $\mathbb{E}^*[Y|X]$ (recall that they coincide in the case where the conditional expectation function is linear). The generalization of the result in Chapter 3 about the optimal linear predictor for the case in which Y is a scalar and X is a vector is:

$$\mathbb{E}^*[Y|X] = \alpha + \beta' X \quad \Rightarrow \quad \begin{array}{c} \beta = [\text{Var}(X)]^{-1} \text{Cov}(X, Y) \\ \alpha = \mathbb{E}[Y] - \beta' \mathbb{E}[X]. \end{array} \tag{1}$$

Consider the bivariate case, where $X = (X_1, X_2)'$. It is interesting to compare $\mathbb{E}^*[Y|X_1]$ and $\mathbb{E}^*[Y|X_1, X_2]$. Let $\mathbb{E}^*[Y|X_1] = \alpha^* + \beta^* X_1$ and $\mathbb{E}^*[Y|X_1, X_2] = \alpha + \beta_1 X_1 + \beta_2 X_2$. Thus:

$$\mathbb{E}^*[Y|X_1] = \mathbb{E}^*[\mathbb{E}^*[Y|X_1, X_2]|X_1] = \alpha + \beta_1 X_1 + \beta_2 \mathbb{E}^*[X_2|X_1]. \tag{2}$$

Let $\mathbb{E}^*[X_2|X_1] = \gamma + \delta X_1$. Then:

$$\mathbb{E}^*[Y|X_1] = \alpha + \beta_1 X_1 + \beta_2(\gamma + \delta X_1) \quad \Rightarrow \quad \begin{array}{c} \beta^* = \beta_1 + \delta\beta_2 \\ \alpha^* = \alpha + \gamma\beta_2. \end{array} \tag{3}$$

This result tells us that the effect of changing variable X_1 on Y is given by a direct effect (β_1) and an indirect effect through the effect of X_1 on X_2 and X_2 on Y. For example, consider the case in which Y is wages, X_1 is age, and X_2 is education, with $\beta_1, \beta_2 > 0$. If we do not include education in our model, then we could obtain a β_1^* that is negative, as older individuals may have lower education.

B. *Ordinary Least Squares*

Consider a set of observations $\{(y_i, x_i) : i = 1, ..., N\}$ where y_i are a scalars, and x_i are vectors of size $K \times 1$. Using the analogy principle, we can propose a natural

[*] Departament d'Economia i Història Econòmica. Universitat Autònoma de Barcelona. Facultat d'Economia, Edifici B, Campus de Bellaterra, 08193, Cerdanyola del Vallès, Barcelona (Spain). E-mail: joan.llull[at]movebarcelona[dot]eu. URL: http://pareto.uab.cat/jllull.

estimator for α and β:[1]

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{(a,b)} \frac{1}{N} \sum_{\text{i=1}}^{\text{N}} (y_i - a - b' x_i)^2. \tag{4}$$

This estimator is called ***Ordinary Least Squares***. The solution to the above problem is:

$$\hat{\beta} = \left[\sum_{\text{i=1}}^{\text{N}} (x_i - \bar{x}_N)(x_i - \bar{x}_N)' \right]^{-1} \sum_{\text{i=1}}^{\text{N}} (x_i - \bar{x}_N)(y_i - \bar{y}_N),$$
$$\hat{\alpha} = \bar{y}_N - \hat{\beta}' \bar{x}_N. \tag{5}$$

Note that the first term of $\hat{\beta}$ is a $K \times K$ matrix, while the second is a $K \times 1$ vector.

C. Algebraic Properties of the OLS Estimator

Let us introduce some compact notation. Let $\delta \equiv (\alpha, \beta')'$ be the parameter vector, let $y = (y_1, ..., y_N)'$ be the vector of observations of Y, and let $W = (w_1, ..., w_N)'$ such that $w_i = (1, x_i')'$ be the matrix (here we are using capital letters to denote a matrix, not a random variable) of observations for the remaining variables. Then:

$$\hat{\delta} = \arg\min_d \sum_{\text{i=1}}^{\text{N}} (y_i - w_i' d)^2 = \arg\min_d (y - Wd)'(y - Wd). \tag{6}$$

And the solution is:

$$\hat{\delta} = \left(\sum_{\text{i=1}}^{\text{N}} w_i w_i' \right)^{-1} \sum_{\text{i=1}}^{\text{N}} w_i y_i = (W'W)^{-1} W' y. \tag{7}$$

Let us do the matrix part in detail. First note:

$$(y - Wd)'(y - Wd) = y'y - y'Wd - d'W'y + d'W'Wd$$
$$= y'y - 2d'W'y + d'W'Wd. \tag{8}$$

The last equality is obtained by observing that all elements in the sum are scalars. The first order condition is:

$$-2W'y + 2(W'W)\hat{\delta} = 0,$$
$$W'y = (W'W)\hat{\delta}, \tag{9}$$
$$\hat{\delta} = (W'W)^{-1} W' y.$$

[1] To avoid complications with the notation below, in this chapter we follow the convention of writing the estimators as a function of realizations (y_i, x_i) instead of doing it as functions of the random variables (Y_i, X_i).

Note that we need $W'W$ to be full rank, such that it can be inverted. This is to say, we require ***absence of multicollinearity***.

D. Residuals and Fitted Values

Recall from Chapter 3 the prediction error $U \equiv y - \alpha - \beta' X = y - (1, X')\delta$. In the sample, we can define an analogous concept, which is called the ***residual***: $\hat{u} = y - W\hat{\delta}$. Similarly, we can define the vector of ***fitted values*** as $\hat{y} = W\hat{\delta}$. Clearly, $\hat{u} = y - \hat{y}$. Some of their properties are useful:

1) $W'\hat{u} = 0$. This equality comes trivially from the derivation in (9): $W'\hat{u} = W'(y - W\hat{\delta}) = W'y - (W'W)\hat{\delta} = 0$. Looking at these matrix multiplications as sums, we can observe that they imply $\sum_{\mathrm{i=1}}^{\mathrm{N}} \hat{u}_i = 0$, and $\sum_{\mathrm{i=1}}^{\mathrm{N}} x_i \hat{u}_i = 0$. Interestingly, these are sample analogs of the population moment conditions satisfied by U.
2) $\hat{y}'\hat{u} = 0$ because $\hat{y}'\hat{u} = \hat{\delta} W'\hat{u} = \hat{\delta} \cdot 0 = 0$.
3) $y'\hat{y} = \hat{y}'\hat{y}$ because $y'\hat{y} = (\hat{y} + \hat{u})'\hat{y} = \hat{y}'\hat{y} + \hat{u}'\hat{y} = \hat{y}'\hat{y} + 0 = \hat{y}'\hat{y}$.
4) $\iota' y = \iota'\hat{y} = N\bar{y}$, where ι is a vector of ones, because $\iota'\hat{u} = \sum_{\mathrm{i=1}}^{\mathrm{N}} \hat{u}_i = 0$, and $\iota' y = \iota'\hat{y} + \iota'\hat{u}$.

E. Variance Decomposition and Sample Coefficient of Determination

Following exactly the analogous arguments as in the proof of the variance decomposition for the linear prediction model in Chapter 3 we can prove that:

$$y'y = \hat{y}'\hat{y} + \hat{u}'\hat{u} \quad \text{and} \quad \widehat{\mathrm{Var}}(y) = \widehat{\mathrm{Var}}(\hat{y}) + \widehat{\mathrm{Var}}(\hat{u}), \tag{10}$$

where $\widehat{\mathrm{Var}}(z) \equiv N^{-1} \sum_{\mathrm{i=1}}^{\mathrm{N}} (z - \bar{z})^2$ To prove the first, we simply need basic algebra:

$$\hat{u}'\hat{u} = (y - \hat{y})'(y - \hat{y}) = y'y - \hat{y}'y - y'\hat{y} + \hat{y}'\hat{y} = y'y - \hat{y}'\hat{y}. \tag{11}$$

The last equality is obtained following the result $y'\hat{y} = \hat{y}'\hat{y}$ obtained in item 3) from the list above. To prove the second equality in (10), we need to recall from Chapter 4 that we can write $\sum_{\mathrm{i=1}}^{\mathrm{N}} (y - \bar{y})^2 = (y - \iota\bar{y})'(y - \iota\bar{y})$. And now, we can operate:

$$(y - \iota\bar{y})'(y - \iota\bar{y}) = y'y - \bar{y}\iota' y - y'\iota\bar{(y)} + \bar{y}^2 \iota'\iota = y'y - N\bar{y}^2. \tag{12}$$

Given the result in item 4) above, we can conclude that $(\hat{y} - \iota\bar{y})'(\hat{y} - \iota\bar{y}) = \hat{y}'\hat{y} - N\bar{y}^2$. Thus:

$$N\widehat{\mathrm{Var}}(\hat{u}) = \hat{u}'\hat{u} = y'y - \hat{y}'\hat{y} = y'y - N\bar{y}^2 - (\hat{y}'\hat{y} - N\bar{y}^2) = N\widehat{\mathrm{Var}}(y) - N\widehat{\mathrm{Var}}(\hat{y}), \tag{13}$$

completing the proof.

Similar to the population case described in Chapter 3, this result allows us to write the ***sample coefficient of determination*** as:

$$R^2 \equiv 1 - \frac{\sum_{\mathrm{i=1}}^{\mathrm{N}} u_i^2}{\sum_{\mathrm{i=1}}^{\mathrm{N}} (y_i - \bar{y})^2} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{N}} (\hat{y}_i - \bar{y})_i^2}{\sum_{\mathrm{i=1}}^{\mathrm{N}} (y_i - \bar{y})^2} = \frac{\widehat{\mathrm{Var}}(\hat{y})}{\widehat{\mathrm{Var}}(y)} = \frac{[\widehat{\mathrm{Cov}}(y, \hat{y})]^2}{\widehat{\mathrm{Var}}(\hat{y})\widehat{\mathrm{Var}}(y)} = \rho_{y,\hat{y}}^2. \tag{14}$$

The last equality is obtained by multiplying and dividing by $\hat{y}'\hat{y}$, and using that $\hat{y}'\hat{y} = y'\hat{y}$ as shown above.

F. Assumptions for the Classical Regression Model

So far we have just described algebraic properties of the OLS estimator as an estimator of the parameters of the linear prediction of Y given X. In order to use the OLS estimator to obtain information about $\mathbb{E}[Y|X]$, we require additional assumptions. This extra set of assumptions constitute what is known as the ***classical regression model***. These assumptions are:

- **Assumption 1 (linearity+strict exogeneity)**: $\mathbb{E}[y|W] = W\delta$, which is equivalent to say $\mathbb{E}[y_i|x_1, ..., x_N] = \alpha + x_i'\beta$, or to define $y \equiv W\delta + u$ where $\mathbb{E}[u|W] = 0$. There are two main conditions embedded in this assumption. The first one is ***linearity***, which implies that the optimal linear predictor and the conditional expectation function coincide. The second one is that $\mathbb{E}[y_i|x_1, ..., x_N] = \mathbb{E}[y_i|x_i]$, which is called ***(strict) exogeneity***. Exogeneity implies that $\mathrm{Cov}(u_i, x_{kj}) = 0$ and $\mathbb{E}[u_i|W] = 0$. To prove it, note that $\mathbb{E}[u_i] = \mathbb{E}[\mathbb{E}[u_i|W]] = \mathbb{E}[\mathbb{E}[y_i - \alpha - x_i'\beta|W]] = \mathbb{E}[\mathbb{E}[y_i|W] - \alpha - x_i'\beta] = 0$, and, hence, $\mathrm{Cov}(u_i, x_{kj}) = \mathbb{E}[u_i x_{kj}] = \mathbb{E}[x_{kj}\,\mathbb{E}[u_i|W]] = 0$. This assumption is satisfied by an i.i.d. random sample:

$$\begin{aligned} f(y_i|x_1, ..., x_N) &= \frac{f(y_i, x_1, ..., x_N)}{f(x_1, ..., x_N)} = \frac{f(y_i, x_i)f(x_1)...f(x_{i-1})f(x_{i+1})...f(x_N)}{f(x_1)...f(x_N)} \\ &= \frac{f(y_i, x_i)}{f(x_i)} = f(y_i|x_i), \end{aligned} \tag{15}$$

  which implies that $\mathbb{E}[y_i|x_1, ..., x_N] = \mathbb{E}[y_i|x_i]$. This is not satisfied, for example, by time series data: if $x_i = y_{i-1}$ (that is, a regressor is the lag of the dependent variable), as $\mathbb{E}[y_i|x_1, ..., x_N] = \mathbb{E}[y_i|x_i, x_{i+1} = y_i] = y_i \neq \mathbb{E}[y_i|x_i]$.

- **Assumption 2 (homoskedasticity)**: $\mathrm{Var}(y|W) = \sigma^2 I_N$. This assumption implies (along with the previous one) that $\mathrm{Var}(y_i|x_1, ..., x_N) = \mathrm{Var}(y_i|x_i) =$

σ^2 and $\mathrm{Cov}(y_i, y_j|x_1, ..., x_N) = 0$ for all $i \neq j$:

$$\begin{aligned}\mathrm{Var}(y_i|x_i) &= \mathrm{Var}(\mathbb{E}[y_i|x_1, ..., x_N]|x_i) + \mathbb{E}[\mathrm{Var}(y_i|x_1, ..., x_N)|x_i] \\ &= \mathrm{Var}(\mathbb{E}[y_i|x_i]|x_i) + \mathbb{E}[\sigma^2|x_i] = 0 + \sigma^2 = \sigma^2. \qquad (16)\end{aligned}$$

We could also check as before that an i.i.d. random sample would satisfy this condition.

II. Statistical Results and Interpretation

A. Unbiasedness and Efficiency

In the classical regression model, $\mathbb{E}[\hat{\delta}] = \delta$:

$$\mathbb{E}[\hat{\delta}] = \mathbb{E}[\mathbb{E}[\hat{\delta}|W]] = \mathbb{E}[(W'W)^{-1}W'\,\mathbb{E}[y|W]] = \mathbb{E}[\delta] = \delta, \qquad (17)$$

where we crucially used the Assumption 1 above. Similarly, $\mathrm{Var}(\hat{\delta}|W) = \sigma^2(W'W)^{-1}$:

$$\mathrm{Var}(\hat{\delta}|W) = (W'W)^{-1}W'\,\mathrm{Var}(y|W)W(W'W)^{-1} = \sigma^2(W'W)^{-1}, \qquad (18)$$

where we used Assumption 2. Note that $\mathrm{Var}(\hat{\delta}) = \sigma^2\,\mathbb{E}[(W'W)^{-1}]$:

$$\mathrm{Var}(\hat{\delta}) = \mathrm{Var}(\mathbb{E}[\hat{\delta}|W]) + \mathbb{E}[\mathrm{Var}(\hat{\delta}|W)] = 0 + \sigma^2\,\mathbb{E}[(W'W)^{-1}]. \qquad (19)$$

The first result that we obtained indicates that OLS gives an unbiased estimator of δ under the classical assumptions. Now we need to check how good is it in terms of efficiency. The ***Gauss-Markov Theorem*** establishes that OLS is a BLUE (best linear unbiased estimator). More specifically, the theorem states that in the class of estimators that are conditionally unbiased and linear in y, $\hat{\delta}$ is the estimator with the minimum variance.

To prove it, consider an alternative linear estimator $\tilde{\delta} \equiv Cy$, where C is a function of the data W. We can define, without loss of generality, $C \equiv (W'W)^{-1}W' + D$, where D is a function of W. Assume that $\tilde{\delta}$ satisfies $\mathbb{E}[\tilde{\delta}|W] = \delta$ (hence, $\tilde{\delta}$ is another linear unbiased estimator). We first check that $\mathbb{E}[\tilde{\delta}|W] = \delta$ is equivalent to $DW = 0$:

$$\begin{aligned}&\mathbb{E}[\tilde{\delta}|W] = \mathbb{E}[\delta + (W'W)^{-1}W'u + DW\delta + Du|W] = (I + DW)\delta \\ &\qquad (I + DW)\delta = \delta \Leftrightarrow DW = 0, \qquad (20)\end{aligned}$$

given that $\mathbb{E}[Du|W] = D\,\mathbb{E}[u|W] = 0$. An implication of this is that $\tilde{\delta} = \delta + Cu$, since $DW\delta = 0$. Hence:

$$\begin{aligned}\mathrm{Var}(\tilde{\delta}|W) &= \mathbb{E}[(\tilde{\delta} - \delta)(\tilde{\delta} - \delta)'|W] = \mathbb{E}[Cuu'C'|W] = C\,\mathbb{E}[uu'|W]C' = \sigma^2 CC' \\ &= (W'W)^{-1}\sigma^2 + \sigma^2 DD' = \mathrm{Var}(\hat{\delta}|W) + \sigma^2 DD' \geq \mathrm{Var}(\hat{\delta}|W). \qquad (21)\end{aligned}$$

Therefore, $\text{Var}(\hat{\delta}|W)$ is the minimum conditional variance of linear unbiased estimators. Finally, to prove that $\text{Var}(\hat{\delta})$ is the minimum as a result we use the variance decomposition and the fact that the estimator is conditionally unbiased, which implies $\text{Var}(\mathbb{E}[\tilde{\delta}|W]) = 0$. Using that, we obtain $\text{Var}(\tilde{\delta}) = \mathbb{E}[\text{Var}(\tilde{\delta}|W)]$. Hence, proving whether $\text{Var}(\tilde{\delta}) - \text{Var}(\hat{\delta}) \geq 0$, which is what we need to prove to establish that $\text{Var}(\hat{\delta})$ is the minimum for this class of estimators, is the same as proving $\mathbb{E}[\text{Var}(\tilde{\delta}|W) - \text{Var}(\hat{\delta}|W)] \geq 0$. Note that, given a random matrix A, because $Z' \, \mathbb{E}[A]Z = \mathbb{E}[Z'AZ]$ if A is positive semidefinite, $\mathbb{E}[A]$ is also positive semidefinite. Therefore, since we proved that $\text{Var}(\tilde{\delta}|W) - \text{Var}(\hat{\delta}|W) \geq 0$, that is, it is positive semidefinite, then its expectation should be positive semidefinite, which completes the prove.

B. Normal classical regression model

Let us now add an extra assumption:

- **Assumption 3 (normality)**: $y|W \sim \mathcal{N}(W\delta, \sigma^2 I_N)$, that is, we added the normality assumption to Assumptions 1 and 2.

In this case, we can propose to estimate δ by ML (which we know provides the BUE). The conditional likelihood function is:

$$L_N(\delta, \sigma^2) = f(y|W) = (2\pi)^{-\frac{N}{2}} \left(\sigma^{2N}\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(y - W\delta)'(y - W\delta)\right), \quad (22)$$

and the conditional log-likelihood is:

$$\mathcal{L}_{\text{N}}(\delta, \sigma^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(y - W\delta)'(y - W\delta). \quad (23)$$

The first order conditions are:

$$\frac{\partial \mathcal{L}_{\text{N}}}{\partial \delta} = \frac{1}{\sigma^2}W'(y - W\delta) = 0 \quad (24)$$

$$\frac{\partial \mathcal{L}_{\text{N}}}{\partial \sigma^2} = \frac{1}{2\sigma^2}\left(\frac{(y - W\delta)'(y - W\delta)}{\sigma^2} - N\right) = 0, \quad (25)$$

which easily delivers that the maximum likelihood estimator of δ is the OLS estimator, and $\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{N}$. Therefore, we can conclude that, under the normality assumption, the OLS estimator is conditionally a BUE. We could prove, indeed, that $\sigma^2(W'W)^{-1}$ is (conditionally) the Cramer-Rao lower bound. Even though we are not going to prove it (it is not a trivial proof), unconditionally, there is no BUE. To do it, we would need to use the unconditional likelihood $f(y|W)f(W)$ instead of $f(y|W)$ alone.

Regarding $\hat{\sigma}^2$, similarly to what happened with the variance of a random variable, the MLE is biased:

$$\hat{u} = y - W\hat{\delta} = y - W(W'W)^{-1}W'y = (I - W(W'W)^{-1}W')y = My. \tag{26}$$

Similar to what happened in Chapter 5 (check the arguments there to do the proofs), M, which is called the residual maker, is idempotent and symmetric, its rank is equal to its trace, and equal to $N - K$, where K is the dimension of δ (because $\text{tr}(AB) = \text{tr}(BA)$, and hence $\text{tr}(W(W'W)^{-1}W') = \text{tr}(I_K)$), and $MW = 0$. Therefore, $\hat{u} = My = M(W\delta + u) = Mu$. Hence:

$$\hat{u}'\hat{u} = (Mu)'Mu = u'M'Mu = u'Mu = \text{tr}(u'Mu) = \text{tr}(uu'M) = \text{tr}(Muu'), \tag{27}$$

where we used the fact that $u'Mu$ is a scalar (and hence equal to its trace), and some of the tricks about traces used above. Now:

$$\begin{aligned} \mathbb{E}[\hat{u}'\hat{u}|W] &= \mathbb{E}[\text{tr}(Muu')|W] = \text{tr}(\mathbb{E}[Muu'|W]) = \text{tr}(M\,\mathbb{E}[uu'|W]) \\ &= \text{tr}(M\sigma^2 I_N) = \sigma^2\,\text{tr}(M) = \sigma^2(N - K). \end{aligned} \tag{28}$$

Hence, an unbiased estimator is $s^2 \equiv \frac{\hat{u}'\hat{u}}{N-K}$, and, as a result (easy to prove using the law of iterated expectations) an unbiased estimator of the variance of $\hat{\delta}$ is $\widehat{\text{Var}}(\hat{\delta}) = s^2(W'W)^{-1}$.