ON THE PREDICTION ERROR

Leandro da Silva PEREIRA[1]
Lucas Monteiro CHAVES[2]
Devanil Jaques de SOUZA[3]

▪ ABSTRACT: The theory of model prediction error is presented in details from the point of view of geometric constructions. It is expected that this approach can be a possible pedagogical tool in the treatment of the subject. Although the focus is essentially conceptual, all algebraic passages is developed in order to facilitate a greater understanding for the reader. Two elementary examples are presented.

▪ KEYWORDS: Model fitting; double expectation; covariance; geometry.

1 Introduction

A statistical model should, like almost every scientific procedure, have one eye on the past and two eyes on the future. Once a model has been fitted to a data set, this certainly describes well the past. However, it will also describe well the future? For example, when modeling influenza data that occurred last winter, will the model be able to predict well the number of cases in the next winter? This is called by Statisticians model predictive capacity, this being perhaps the most important characteristic of a model. In spite of this, in the textbooks we find several techniques on how to adjust models to data and how to measure the quality of these adjustments, but in general, little attention is given to measuring the capacity of these models to predict future data. This fact is not justified since the formalization of the predictive capacity of a model is not mathematically more complicated than

[1] Universidade Tecnológica Federal do Paraná - UTFPR, Departamento de Matemática, CEP: 86812-460, Apucarana, PR, Brazil, E-mail: *leandropereira@utfpr.edu.br*
[2] Universidade Federal de Lavras - UFLA, Departamento de Ciências Exatas, CEP: 37200-900, Lavras, MG, Brazil, E-mail: *lucas@ufla.br*
[3] Universidade Federal de Lavras - UFLA, Departamento de Estatística, CEP: 37200-900, Lavras, MG, Brazil, E-mail: *devaniljaques@ufla.br*

the adjustment techniques. This paper intends to develop this approach in a trivial, yet exhaustive way, explaining all the mathematical passages in order to facilitate the reader's understanding, having been motivated by the student's difficulties in understand the prediction error formula in Efron(2004).

2 Methods

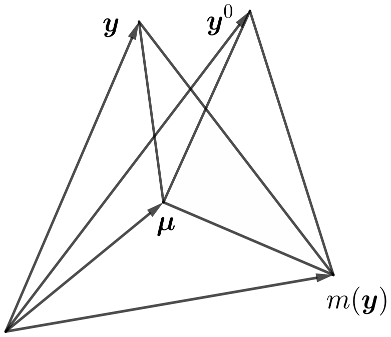

Figure 1 - Geometry of the prediction error.

Let $\boldsymbol{y} = (y_1, ..., y_n)$ be a random vector with mean vector $\boldsymbol{\mu} = (\mu_1, ..., \mu_n) = E[\boldsymbol{y}]$. When a data vector $\boldsymbol{y}$ is observed, some adjustment technique is adopted and then a model is proposed and expressed in the form $\hat{\boldsymbol{\mu}} = m(\boldsymbol{y})$, where m is a function of $\mathbb{R}^n \to \mathbb{R}^n$. How can we assess the predictive capacity of $\boldsymbol{\mu} = \hat{m}(y)$? The question is, if a new vector $\boldsymbol{y}^0$ is observed, how close to $\hat{\mu} = m(\mathbf{y})$ will this vector be? Such question does not cover the entire prediction problem because there is still the problem that the data vector used in the adjustment was the realization of a random vector. Then it is necessary that the whole process be randomized:

The data $\boldsymbol{y}$ are observed → the estimative $\hat{\boldsymbol{\mu}} = \mathrm{m}(\boldsymbol{y})$ is obtained → new data vector $\boldsymbol{y}^0$ of the same random phenomenon is observed → the square of deviations $\left\|\boldsymbol{y}^0 - \mathrm{m}(\boldsymbol{y})\right\|^2$ is then calculated.

So, if this process is repeated several times, what is the mean of the sum of squares of deviations? It is necessary to formalize this procedure in terms of mathematical expectations.

Since we have two random vectors $\boldsymbol{y}$ and $\boldsymbol{y}^0$, it is necessary for a proper definition of prediction error to take expectation in relation to each of these random vectors, that is, the double expectation $E\left[E_0\left[\left\|\boldsymbol{y}^0 - \mathrm{m}(\boldsymbol{y})\right\|^2\right]\right]$, where $E[\,]$ is the expectation with respect to the vector $\boldsymbol{y}$ and $E_0[\,]$ the expectation with respect to

$\boldsymbol{y}^0$ (MOOD; GRAYBILL and BOES, 1974; CASELLA and BERGER, 2002). In order to have an intuition in the calculation of this double expectation, a little of geometry will be used.

The excess of algebraic expressions is due to the fact that all passages was explicited, what is expected to facilitate the understanding.

We have three random vectors $\boldsymbol{y}$, $\mathrm{m}(\boldsymbol{y})$, $\boldsymbol{y}^0$ and a parametric vector $\boldsymbol{\mu}$. Two triangles are then defined by $\{\boldsymbol{y},\ \boldsymbol{\mu},\ \mathrm{m}(\boldsymbol{y})\}$ and $\{\boldsymbol{y}^0,\ \boldsymbol{\mu},\ m(\boldsymbol{y})\}$.

Applying the law of cosines to the dashed triangle $\{\boldsymbol{y}^0,\ \boldsymbol{\mu},\ m(\boldsymbol{y})\}$ in Figure 1, we have:

$$\begin{aligned}\left\|\boldsymbol{y}^0-\mathrm{m}(\boldsymbol{y})\right\|^2 &= \left\|\boldsymbol{y}^0-\boldsymbol{\mu}\right\|^2+\|m(\boldsymbol{y})-\boldsymbol{\mu}\|^2-2\left\|\boldsymbol{y}^0-\boldsymbol{\mu}\right\|\|m(\boldsymbol{y})-\boldsymbol{\mu}\|\cos(\theta)\\ &= \left\|\boldsymbol{y}^0-\boldsymbol{\mu}\right\|^2+\|m(\boldsymbol{y})-\boldsymbol{\mu}\|^2-2\left\langle\boldsymbol{y}^0-\boldsymbol{\mu},m(\boldsymbol{y})-\boldsymbol{\mu}\right\rangle.\end{aligned}$$

Taking the expectation of the previous equation in relation to $\mathbf{y}^0$ follows

$$\begin{aligned}E_0\left[\left\|\boldsymbol{y}^0-m(\boldsymbol{y})\right\|^2\right] &= E_0\left[\left\|\boldsymbol{y}^0-\boldsymbol{\mu}\right\|^2+\|m(\boldsymbol{y})-\boldsymbol{\mu}\|^2-2\left\langle\boldsymbol{y}^0-\boldsymbol{\mu},m(\boldsymbol{y})-\boldsymbol{\mu}\right\rangle\right]\\ &= E_0\left[\left\|\boldsymbol{y}^0-\boldsymbol{\mu}\right\|^2\right]+E_0\left[\|m(\boldsymbol{y})-\boldsymbol{\mu}\|^2\right]-2E_0\left[\left\langle\boldsymbol{y}^0-\boldsymbol{\mu},m(\boldsymbol{y})-\boldsymbol{\mu}\right\rangle\right].\end{aligned}\tag{1}$$

Since $m(\boldsymbol{y})-\boldsymbol{\mu}$ does not depend on $\boldsymbol{y}^0$ and the expectation is taken in relation to $\boldsymbol{y}^0$, then:

$$-2E_0\left[\left\langle\boldsymbol{y}^0-\boldsymbol{\mu},m(\boldsymbol{y})-\boldsymbol{\mu}\right\rangle\right]=-2\left\langle E_0\left[\boldsymbol{y}^0-\boldsymbol{\mu}\right],\left[m(\boldsymbol{y})-\boldsymbol{\mu}\right]\right\rangle.$$

Furthermore, $E_0\left[\boldsymbol{y}^0-\boldsymbol{\mu}\right]=0$. Thus, the previous equation is

$$-2E_0\left[\left\langle\boldsymbol{y}^0-\boldsymbol{\mu},m(\boldsymbol{y})-\boldsymbol{\mu}\right\rangle\right]=-2\left\langle 0,m(\boldsymbol{y})-\boldsymbol{\mu}\right\rangle=0.$$

Then we can write:

$$E_0\left[\left\|\boldsymbol{y}^0-m(\boldsymbol{y})\right\|^2\right]=E_0\left[\left\|\boldsymbol{y}^0-\boldsymbol{\mu}\right\|^2\right]+E_0\left[\|m(\boldsymbol{y})-\boldsymbol{\mu}\|^2\right].$$

It is very interesting to note that in relation to mathematical expectation, the Pythagorean theorem holds true, i. e., on average the dashed triangle is a right triangle. Also, note that no distributional hypothesis for the random vector $\boldsymbol{y}^0$ has been assumed, not even the hypothesis of symmetry. This fact goes beyond our intuition and we will leave this question to the reader, how to explain this fact better?

Finally, $E_0\left[\|m(\boldsymbol{y})-\boldsymbol{\mu}\|^2\right] = \|m(\boldsymbol{y})-\boldsymbol{\mu}\|^2$, because $m(\boldsymbol{y})-\boldsymbol{\mu}$ does not depend on $\boldsymbol{y}^0$. Thus,

$$E_0\left[\left\|\boldsymbol{y}^0-m(\boldsymbol{y})\right\|^2\right]=E_0\left[\left\|\boldsymbol{y}^0-\boldsymbol{\mu}\right\|^2\right]+\|m(\boldsymbol{y})-\boldsymbol{\mu}\|^2.$$

and therefore the prediction error is given by

$$E\left[E_0\left[\left\|\boldsymbol{y}^0 - m(\boldsymbol{y})\right\|^2\right]\right] = E\left[E_0\left[\left\|\boldsymbol{y}^0 - \boldsymbol{\mu}\right\|^2\right]\right] + E\left[\|m(\boldsymbol{y}) - \boldsymbol{\mu}\|^2\right]. \qquad (2)$$

Noting that $E_0\left[\left\|\boldsymbol{y}^0 - \boldsymbol{\mu}\right\|^2\right]$ is a population quantity which does not depend on $\boldsymbol{y}$, it follows that $E\left[E_0\left[\left\|\boldsymbol{y}^0 - \boldsymbol{\mu}\right\|^2\right]\right] = E_0\left[\left\|\boldsymbol{y}^0 - \boldsymbol{\mu}\right\|^2\right]$, and the equation (2) can be written as

$$E\left[E_0\left[\left\|\boldsymbol{y}^0 - m(\boldsymbol{y})\right\|^2\right]\right] = E_0\left[\left\|\boldsymbol{y}^0 - \boldsymbol{\mu}\right\|^2\right] + E\left[\|m(\boldsymbol{y}) - \boldsymbol{\mu}\|^2\right].$$

Since $\boldsymbol{y}^0$ and $\boldsymbol{y}$ can be considered the same random vector, $E_0\left[\left\|\boldsymbol{y}^0 - \boldsymbol{\mu}\right\|^2\right] = E\left[\|\boldsymbol{y} - \boldsymbol{\mu}\|^2\right]$, and then

$$E\left[E_0\left[\left\|\boldsymbol{y}^0 - \boldsymbol{\mu}(\boldsymbol{y})\right\|^2\right]\right] = E\left[\|\boldsymbol{y} - \boldsymbol{\mu}\|^2\right] + E\left[\|\boldsymbol{m}(\boldsymbol{y}) - \boldsymbol{\mu}\|^2\right]. \qquad (3)$$

Now, by applying the law of cosines in the dashed triangle $\{\boldsymbol{y},\, \boldsymbol{\mu},\, m(\boldsymbol{y})\}$ (Figure 1) we have:

$$\|\boldsymbol{y} - m\,(\boldsymbol{y})\|^2 = \|\boldsymbol{y} - \boldsymbol{\mu}\|^2 + \|m\,(\boldsymbol{y}) - \boldsymbol{\mu}\|^2 - 2\,\langle \boldsymbol{y} - \boldsymbol{\mu},\, m\,(\boldsymbol{y}) - \boldsymbol{\mu}\rangle \qquad (4)$$

Taking the expectation, we have as result:

$$E\left[\|\boldsymbol{y} - \boldsymbol{\mu}\|^2\right] + E\left[\|m(\boldsymbol{y}) - \boldsymbol{\mu}\|^2\right] = E\left[\|\boldsymbol{y} - m(\boldsymbol{y})\|^2\right] + 2E\left[\langle \boldsymbol{y} - \boldsymbol{\mu}, m(\boldsymbol{y}) - \boldsymbol{\mu}\rangle\right],$$

Therefore, by substituting the last result in (3) it follows that:

$$E\left[E_0\left[\left\|\boldsymbol{y}^0 - m(\boldsymbol{y})\right\|^2\right]\right] = E\left[\|\boldsymbol{y} - m(\boldsymbol{y})\|^2\right] + 2E\left[\langle \boldsymbol{y} - \boldsymbol{\mu}, m(\boldsymbol{y}) - \boldsymbol{\mu}\rangle\right].$$

It can be seen that

$$
\begin{aligned}
E\left[\langle \boldsymbol{y}-\boldsymbol{\mu}, m(\boldsymbol{y})-\boldsymbol{\mu}\rangle\right] &= E\left[\langle \boldsymbol{y}-\boldsymbol{\mu}, m(\boldsymbol{y})-E\left[m(\boldsymbol{y})\right]+E\left[m(\boldsymbol{y})\right]-\boldsymbol{\mu}\rangle\right]\\
&= E\left[\langle \boldsymbol{y}-\boldsymbol{\mu}, m(\boldsymbol{y})-E\left[m(\boldsymbol{y})\right]\rangle\right]+E\left[\langle \boldsymbol{y}-\boldsymbol{\mu}, E\left[m(\boldsymbol{y})\right]-\boldsymbol{\mu}\rangle\right]\\
&= E\left[\langle \boldsymbol{y}-\boldsymbol{\mu}, m(\boldsymbol{y})-E\left[m(\boldsymbol{y})\right]\rangle\right]+\langle E\left[\boldsymbol{y}-\boldsymbol{\mu}\right], E\left[m(\boldsymbol{y})\right]-\boldsymbol{\mu}\rangle\\
&= E\left[\langle \boldsymbol{y}-\boldsymbol{\mu}, m(\boldsymbol{y})-E\left[m(\boldsymbol{y})\right]\rangle\right]+\langle 0, E\left[m(\boldsymbol{y})\right]-\boldsymbol{\mu}\rangle\\
&= E\left[\langle \boldsymbol{y}-\boldsymbol{\mu}, m(\boldsymbol{y})-E\left[m(\boldsymbol{y})\right]\rangle\right]\\
&= E\left[\sum_{i=1}^{n}\left(\boldsymbol{y}_i-\boldsymbol{\mu}_i\right)\left(m(\boldsymbol{y})_i-E[m(\boldsymbol{y})]_i\right)\right]\\
&= \sum_{i=1}^{n} E\left[\left(\boldsymbol{y}_i-\boldsymbol{\mu}_i\right)\left(m(\boldsymbol{y})_i-E[m(\boldsymbol{y})]_i\right)\right]\\
&= \sum_{i=1}^{n} \operatorname{cov}\left(\boldsymbol{y}_i, m(\boldsymbol{y})_i\right)\\
&= \operatorname{cov}(\boldsymbol{y}, m(\boldsymbol{y})).
\end{aligned}
$$

Therefore:

$$
E\left[E_0\left[\left\|\boldsymbol{y}^0-m(\boldsymbol{y})\right\|^2\right]\right]=E\left[\|\boldsymbol{y}-m(\boldsymbol{y})\|^2\right]+2\operatorname{cov}(\boldsymbol{y}, m(\boldsymbol{y})). \tag{5}
$$

This formula was presented in Efron(2004) without many details. The prediction error is then the error of the model adjustment added with a penalty related to the covariance between the data and the estimator. Note that if the covariance between the data vector $\boldsymbol{y}$ and the model $m(\boldsymbol{y})$ is large, this means a certain instability of the model in the sense that a larger variation in the data also implies a large variation in the model, which is not a desirable feature of a model . Therefore, the predictive capacity is a trade-off between the expectation of fit error and the variability of the model with the data.

3 Examples

The simplest case occurs when we want to calculate the prediction error in the situation where $\boldsymbol{y}=(y_1,\ldots,y_n)$ is an i.i.d sample. We want to estimate $\boldsymbol{\mu}=E[\boldsymbol{y}]$. In this case, we will use as estimator $\hat{\boldsymbol{\mu}}=\mathrm{m}(\boldsymbol{y})=\bar{\boldsymbol{y}}=(\bar{y},\ldots,\bar{y})$, where $\bar{y}=\frac{1}{n}\sum\limits_{i=1}^{n}y_i$.

The prediction error of this estimator in relation to a new observation $\boldsymbol{y}^0$ is given by

$$
\begin{aligned}
E\left[E_0\left[\left\|\boldsymbol{y}^0-\bar{\boldsymbol{y}}\right\|^2\right]\right] &= E\left[\left\|\boldsymbol{y}-\bar{\boldsymbol{y}}\right\|^2\right]+2\mathrm{cov}(\boldsymbol{y},\bar{\boldsymbol{y}})\\
&= \sum_{i=1}^{n} E\left[\left(\mathrm{y}_i-\bar{\mathrm{y}}\right)^2\right]+2\sum_{i=1}^{n}\mathrm{cov}(\mathrm{y}_i,\bar{\mathrm{y}})\\
&= \sum_{i=1}^{n} E\left[\left(\mathrm{y}_i-\mu+\mu-\bar{\mathrm{y}}\right)^2\right]+2\sum_{i=1}^{n}\mathrm{cov}\left(\mathrm{y}_i,\sum_{j=1}^{n}\frac{1}{n}\mathrm{y}_j\right)\\
&= \sum_{i=1}^{n} E\left[\left(\mathrm{y}_i-\mu\right)^2+2(\mathrm{y}_i-\mu)(\mu-\bar{\mathrm{y}})+\left(\mu-\bar{\mathrm{y}}\right)^2\right]+2\sum_{i=1}^{n}\mathrm{cov}\left(\mathrm{y}_i,\frac{1}{n}\mathrm{y}_i\right)\\
&= \sum_{i=1}^{n}\left\{E\left[\left(\mathrm{y}_i-\mu\right)^2\right]+2E\left[(\mathrm{y}_i-\mu)(\mu-\bar{\mathrm{y}})\right]+E\left[\left(\mu-\bar{\mathrm{y}}\right)^2\right]\right\}+\frac{2}{n}\sum_{i=1}^{n}\mathrm{var}(\mathrm{y}_i)\\
&= \sum_{i=1}^{n}\left\{\sigma^2+2E\left[(\mathrm{y}_i-\mu)(\mu-\bar{\mathrm{y}})\right]+\frac{1}{n}\sigma^2\right\}+2\sigma^2\\
&= \sum_{i=1}^{n}\left\{\sigma^2+2E\left[(\mathrm{y}_i-\mu)(\mu-\frac{1}{n}\sum_{j=1}^{n}\mathrm{y}_j)\right]+\frac{1}{n}\sigma^2\right\}+2\sigma^2
\end{aligned}
$$

Then,

$$\begin{aligned}
E\left[E_0\left[\left\|\boldsymbol{y}^0-\bar{\boldsymbol{y}}\right\|^2\right]\right] &= \sum_{i=1}^{n}\left\{\sigma^2+2E\left[(\mathrm{y}_i-\mu)(\mu-\frac{1}{n}\sum_{j\neq i}\mathrm{y}_j-\frac{1}{n}\mathrm{y}_i)\right]+\frac{1}{n}\sigma^2\right\}+2\sigma^2\\
&= \sum_{i=1}^{n}\left\{\sigma^2+2E\left[(\mathrm{y}_i-\mu)(\mu-\frac{1}{n}\sum_{j\neq i}\mathrm{y}_j)-(\mathrm{y}_i-\mu)\frac{1}{n}\mathrm{y}_i\right]+\frac{1}{n}\sigma^2\right\}+2\sigma^2\\
&= \sum_{i=1}^{n}\left\{\sigma^2+2E\left[\mathrm{y}_i-\mu\right]E\left[\mu-\frac{1}{n}\sum_{j\neq i}\mathrm{y}_j\right]-\frac{2}{n}E\left[(\mathrm{y}_i-\mu)\mathrm{y}_i\right]+\frac{1}{n}\sigma^2\right\}+2\sigma^2\\
&= \sum_{i=1}^{n}\left\{\sigma^2+0-\frac{2}{n}\left(E\left[\mathrm{y}_i^2\right]-\frac{1}{n}E\left[(\mathrm{y}_i\mu)\right]\right)+\frac{1}{n}\sigma^2\right\}+2\sigma^2\\
&= \sum_{i=1}^{n}\left\{\sigma^2-\frac{2}{n}\left(\sigma^2+\mu^2\right)-\frac{2}{n}\mu^2+\frac{1}{n}\sigma^2\right\}+2\sigma^2\\
&= \sum_{i=1}^{n}\left\{\sigma^2-\frac{2}{n}\sigma^2+\frac{1}{n}\sigma^2\right\}+2\sigma^2\\
&= \sum_{i=1}^{n}\left\{\frac{\sigma^2(n-1)}{n}\right\}+2\sigma^2\\
&= \sigma^2(n-1)+2\sigma^2\\
&= \sigma^2(n+1).
\end{aligned}$$

Figure 2 describes such a situation.

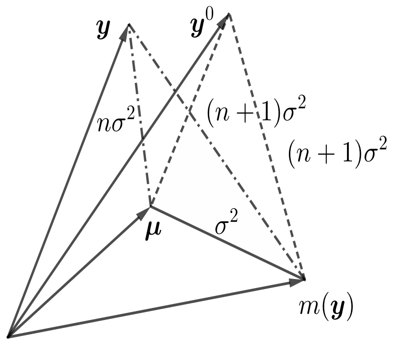

Figure 2 - Geometry when m$(\boldsymbol{y}) = \bar{\boldsymbol{y}}$.

The previous example is a particular case of the more general situation in which we want to calculate the prediction error of the least squares estimator in linear regression $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $rank(X) = p$ and $\boldsymbol{\mu} = E[\boldsymbol{y}] = X\boldsymbol{\beta}$. As a predictor of a new observation $\boldsymbol{y}^0$ we will use $m(\mathbf{y}) = X(X'X)^{-1}X'\mathbf{y} = H\mathbf{y}$ (RENCHER; SHAALJE, 2008). Observing that $H\mu = \mu$ then:

$$\begin{aligned}
E\left[E_0\left[\left\|\boldsymbol{y}^0 - H\boldsymbol{y}\right\|^2\right]\right] &= E\left[\left\|\boldsymbol{y} - H\boldsymbol{y}\right\|^2\right] + 2\text{cov}\left(\boldsymbol{y}, H\boldsymbol{y}\right) \\
&= E\left[\langle \boldsymbol{y} - H\boldsymbol{y}, \boldsymbol{y} - H\boldsymbol{y}\rangle\right] + 2\sum_{i=1}^{n}\text{cov}\left(y_i, \left(H\boldsymbol{y}\right)_i\right) \\
&= E\left[\langle \left(I-H\right)\boldsymbol{y}, \left(I-H\right)\boldsymbol{y}\rangle\right] + 2\sum_{i=1}^{n}\text{cov}\left(y_i, \sum_{s=1}^{n} h_{is}y_s\right) \\
&= E\left[\boldsymbol{y'}(I-H)'\left(I-H\right)\boldsymbol{y}\right] + 2\sum_{i=1}^{n} h_{ii}\text{cov}\left(y_i, y_i\right) \\
&= E\left[\boldsymbol{y'}(I-H)^2\boldsymbol{y}\right] + 2\sigma^2\sum_{i=1}^{n} h_{ii} \\
&= E\left[\boldsymbol{y'}\left(I-H\right)\boldsymbol{y}\right] + 2\sigma^2 tr(H) \\
&= \sigma^2 tr\left(I-H\right) + \boldsymbol{\mu'}\left(I-H\right)\boldsymbol{\mu} + 2\sigma^2 tr(H) \\
&= n\sigma^2 + \sigma^2 tr(H) \\
&= (n+p)\sigma^2
\end{aligned}$$

After get an estimator for β, we have the prediction function $\hat{y} = \tilde{\beta}'x$.

Fixing a value x_0, it is usual to predict the random variable $y = \beta' x_0 + \varepsilon$ by the random variable $\hat{y} = \tilde{\beta}' x_0$. To overcome the high variability of the least square estimator in presence of almost multicollinearity several types of shrinkage estimators was proposed in the literature like: Ridge (HOERL and KENNARD, 1970), LASSO (TIBSHIRANI, 1996) Elastic-Net (ZOU and HASTIE, 2005) and OSCAR (BONDELL and REICH, 2008). Here we will suppose a simple shrinkage of the least square estimator given by $t\hat{\beta}_{ols}, \quad 0 < t < 1$. To predict y^0, the value of the response variable y when the vector of co-variables are x_0, we will use $\hat{y} = t\hat{\beta}'_{ols} x_0$

$$
\begin{aligned}
&E\left[E_0\left[\left(\hat{y}-y^0\right)^2\right]\right] = E\left[E_0\left[\left(\hat{y}-E\left[\hat{y}\right]+E\left[\hat{y}\right]-y^0\right)^2\right]\right] \\
&= E\left[E_0\left[\left(\hat{y}-E\left[\hat{y}\right]\right)^2\right]\right] + 2E\left[E_0\left[\left(\hat{y}-E\left[\hat{y}\right]\right)\left(E\left[\hat{y}\right]-y^0\right)\right]\right] + E\left[E_0\left[\left(E\left[\hat{y}\right]-y^0\right)^2\right]\right] \\
&= E\left[\left(\hat{y}-E\left[\hat{y}\right]\right)^2\right] + 2E\left[\left(\hat{y}-E\left[\hat{y}\right]\right)\left(E\left[\hat{y}\right]-E_0\left[y^0\right]\right)\right] + E_0\left[\left(E\left[\hat{y}\right]-y^0\right)^2\right] \\
&= \operatorname{var}\left[\hat{y}\right] + 2E\left[\left(\hat{y}-E\left[\hat{y}\right]\right)\left(E\left[\hat{\beta}'\right]x_0-\beta' x_0\right)\right] + E_0\left[\left(E\left[\hat{y}\right]-\beta' x_0+\beta' x_0-y^0\right)^2\right] \\
&= \operatorname{var}\left[\hat{y}\right] + 2E\left[\left(\hat{y}-E\left[\hat{y}\right]\right)\left(\beta' x_0-\beta' x_0\right)\right] + E_0\left[\left(E\left[\hat{y}\right]-\beta' x_0\right)^2\right] + \\
&+ 2E_0\left[\left(E\left[\hat{y}\right]-\beta' x_0\right)\left(\beta' x_0-y^0\right)\right] + E_0\left[\left(\beta' x_0-y^0\right)^2\right] \\
&= \operatorname{var}\left[\hat{y}\right] + \left(E\left[\hat{y}\right]-\beta' x_0\right)^2 + 0 + \sigma^2 \\
&= \operatorname{var}\left[\hat{y}\right] + \left(E\left[\hat{y}\right]-\beta' x_0\right)^2 + \sigma^2.
\end{aligned}
$$

Allen (1971) denoted this error of prediction as mean square error of prediction (MSEP). Observe that $\operatorname{var}\left[\hat{y}\right] + \left(E\left[\hat{y}\right]-\beta' x_0\right)^2$ is the mean square error (MSE) of $\hat{y}$ when viewed as an estimator of $\beta' x_0$.

We have

$$\operatorname{var}\left[\hat{y}\right] = t^2 x_0'(X'X)^{-1} x_0 \sigma^2$$

and

$$\left(E\left[\hat{y}\right]-\beta' x\right)^2 + \sigma^2 = (t\hat{\beta}_{ols}-\beta)^2 + \sigma^2$$

and then the error of prediction is $t^2 x_0'(X'X)^{-1} x_0 \sigma^2 + (t\beta' x_0 - \beta x_0)^2 + \sigma^2$. It is worth to get the value of t with the minimum prediction error. For this deriving we get,

$$t_{min} = \frac{\left(\beta' x_0\right)^2}{x_0'(X'X)^{-1} x_0 \sigma^2 + \left(\beta' x_0\right)^2}. \tag{6}$$

As an illustrative example consider (RENCHER; SHAALJE, p. 290, 2008): The data contains body fat for a sample of 20 females aged 25-34. The response variable was body fat (y) and two predictor variables were triceps skinfold thickness (x_1) and midarm circumference (x_2). For these data, the matrix $(X'X)^{-1}$ is given by

$$(X'X)^{-1} = \begin{bmatrix} 3.233 & -0.021 & -0.096 \\ -0.021 & 0.003 & -0.002 \\ -0.096 & -0.002 & 0.006 \end{bmatrix},$$

and estimates

$\hat{\beta}'_{ols} = \begin{bmatrix} 6.792 & 1.000 & -0.431 \end{bmatrix}$

$x'_0 = \begin{bmatrix} 1 & 14 & 19 \end{bmatrix}$

$\hat{\sigma}^2 = 6.231.$

Using these estimate in the equation (6), the estimated value for t_{min} is 0.983.

4 Conclusion

It is possible, with difficulty similar to the definition of model adjustment, an understandable mathematical treatment for the prediction error. The use of the triangle in Figure 1 may be a useful tool in understanding the double expectation that defines the prediction error.

The predictive error of a model is evidently a population quantity that needs to be estimated. To obtain an unbiased estimator for the prediction error, we can use for example, the famous Stein's Lemma. But this is beyond the scope of this article.

Acknowledgements

We would like to thank reviewers and editors for their comments and suggestions.

PEREIRA, L. S.; CHAVES, L. M.; SOUZA, D. J. Sobre o erro de predição. *Rev. Bras. Biom.,* Lavras, v.37, n.4, p.435-445, 2019.

- *RESUMO: A teoria do erro de previsão do modelo é apresentada em detalhes sob o ponto de vista das construções geométricas. Espera-se que esta abordagem possa ser uma ferramenta pedagógica possível no tratamento do assunto. Embora o foco seja essencialmente conceitual, todas as passagens algébricas são desenvolvidas a fim de facilitar uma maior compreensão para o leitor. Dois exemplos elementares são apresentados.*

- *PALAVRAS-CHAVE: Ajuste de modelos; esperança dupla; covariancia; geometria.*

References

ALLEN, D. M. Mean square error of prediction as a criterion for selecting variables. *Techometrics*, v.13, n.3, p. 469-475, 1971.

BONDELL, H. D.; REICH, B. J. Simultaneous regression shinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, v.64, p.115-123, 2008.

CASELLA, G.; BERGER, R. L. *Statistical inference*, 2.ed., Pacific Grove: Duxbury. 2002.

EFRON, B. The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, v.99, n.467, p.619-632. 2004.

HOERL, A. E.; KENNARD, R. W. Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, v.12, n.1, p.55-67, 1970.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. , *Introduction to the theory of statistics*, 3.ed., New York: McGraw-Hill. 1974.

RENCHER, A. C.; SCHAALJE, G. B. *Linear models in statistics*, New Jersey: Wiley. 2008.

TIBSHIRANI, R. Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society*, v.58, n.1, p. 267-288, 1996.

ZOU, H; HASTIE, Y. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, v.67, n.2, p.301-320, 2005.

Received on 26.10.2018.

Approved after revised on 03.06.2019.