

Training, Estimation, And Testing Error

Hossein Falaki
Computer Science Department,
University of California, Los Angeles
falaki@cs.ucla.edu

1 Introduction

This note explores the concepts of *training*, *estimation*, and *testing* errors. We will investigate how increasing the complexity of a model (number of explanatory variables) affects these three types of error in regression.

2 Definitions

Suppose observations (Y values) are arbitrary functions of a set of independent variables (i.e., the assumed function is not necessarily linear):

$$Y = f(X) + \epsilon \quad (1)$$

We can still use regression to find a linear estimation for the observations. This model is not *correct*, but may be useful.¹ We continue to assume that observation errors are uncorrelated, zero-mean, and with constant variance:

$$\begin{aligned} E(\epsilon_i) &= 0 \\ \text{Var}(\epsilon_i) &= \sigma^2 \end{aligned}$$

If we fit a linear model on these observations (e.g., using Least Squares) we will face three types of errors. Suppose $\hat{Y} = X\hat{\beta}_{LS}$, where $\hat{\beta}_{LS}$ is the best linear unbiased estimator for Y .

Definition 2.1. *Training Error* is defined as $E(\|Y - \hat{Y}\|)$

Definition 2.2. *Estimation Error* or the *model bias* is defined as $E(\|f - \hat{Y}\|)$.

Definition 2.3. *Testing Error* is defined as $E(\|Y_{new} - \hat{Y}\|)$

¹“Essentially, all models are wrong, but some are useful.” [1]

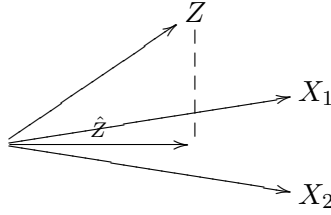
obs.	regressors(X)				f	Y	Y_{new}
1	x_{11}	x_{12}	\dots	x_{1p}	f_1	$y_1 = f_1 + \epsilon_1$	$y_{1,new} = f_1 + \epsilon_{1,new}$
2	x_{21}	x_{22}	\dots	x_{2p}	f_2	$y_2 = f_2 + \epsilon_2$	$y_{2,new} = f_2 + \epsilon_{2,new}$
\vdots	\vdots	\vdots		\vdots	\vdots	\vdots	\vdots
n	x_{n1}	x_{n2}	\dots	x_{np}	f_n	$y_n = f_n + \epsilon_n$	$y_{n,new} = f_n + \epsilon_{n,new}$

Table 1: Assumption in a regression analysis

3 Statistical Analysis

To studying the effect of the model complexity (p) on the three errors we will use the following simple lemmas:

Lemma 3.1. *If X_1, \dots, X_p are orthonormal vectors, the projection of a vector, Z into the space defined by X_1, \dots, X_p is:*



$$\hat{Z} = b_1 X_1 + \dots + b_p X_p$$

where,

$$b_j = \langle Z, X_j \rangle$$

Lemma 3.2. *Assuming $\epsilon \sim N(0, \sigma^2 I_p)$ and $A_{p \times 1}$ a is a constant vector, then:*

$$E(\langle \epsilon, A \rangle) = 0$$

$$Var(\langle \epsilon, A \rangle) = \|a\|^2 \sigma^2$$

Lemma 3.3. *If a and b are two vectors:*

$$\begin{aligned} \|a + b\|^2 &= (a + b)^T (a + b) = (a^T + b^T)(a + b) \\ &= a^T a + b^T b + a^T b + b^T a \\ &= \|a\|^2 + \|b\|^2 + 2 \langle a, b \rangle \end{aligned}$$

Lemma 3.4.

$$\min_{b_1, \dots, b_p} \left\| Z - \sum_{j=1}^p b_j X_j \right\|^2 \geq \min_{b_1, \dots, b_p=1} \left\| Z - \sum_{j=1}^p b_j X_j - b_{p+1} X_{p+1} \right\|^2$$

Without loss of generality we will assume that the explanatory variables form an orthonormal vector.²

$$\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_p X_p$$

where for each j:

$$\hat{\beta}_j = \langle Y, X_j \rangle$$

Substituting Y with its true value ($Y = f + \epsilon$):

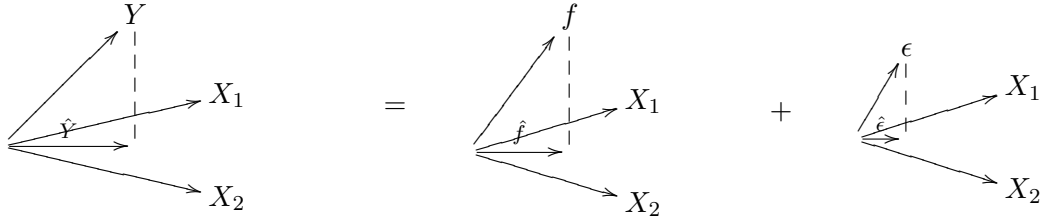
$$\begin{aligned} \hat{\beta}_j &= \langle Y, X_j \rangle = \langle f + \epsilon, X_j \rangle \\ &= \langle f, X_j \rangle + \langle \epsilon, X_j \rangle = \beta_{j,best} + \delta_j \end{aligned}$$

Therefore:

$$\begin{aligned} \hat{f} &= \beta_{1,best} X_1 + \dots + \beta_{p,best} X_p \\ \hat{\epsilon} &= \delta_1 X_1 + \dots + \delta_p X_p \end{aligned}$$

\hat{f} and $\hat{\epsilon}$ are the projection of f and ϵ into the space of X_1, \dots, X_p , therefore:

$$\hat{Y} = \hat{f} + \hat{\epsilon} \quad (2)$$



And similarly

$$\hat{\beta}_j = \beta_{j,best} + \delta_j \quad (3)$$

In Equation 3 $\beta_{j,best}$ s are fixed and δ_j s are random (due to random noise). Since $E(\delta_j) = 0$ and $Var(\delta_j) = \sigma^2$:

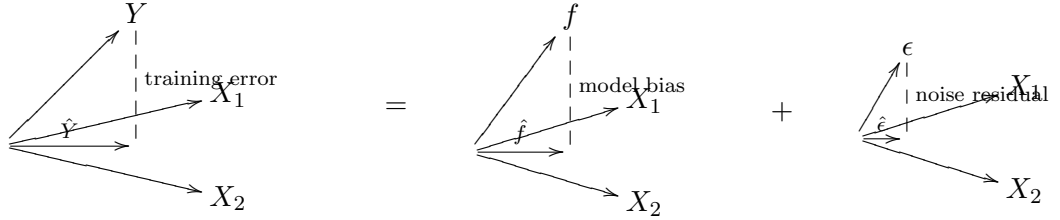
$$\begin{aligned} E(\hat{\beta}_j) &= \beta_{j,best} \\ Var(\hat{\beta}_j) &= \sigma^2 \end{aligned}$$

²If this assumption does not hold, we can find p orthonormal vectors in the space defined by $X_{n \times p}$.

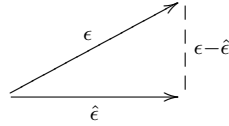
3.1 Training Error

We can write the training error as following:

$$\begin{aligned}
 E(\|Y - \hat{Y}\|^2) &= E(\|(f - \hat{f}) + (\epsilon - \hat{\epsilon})\|^2) \\
 &= E(\underbrace{\|f - \hat{f}\|^2}_{\text{fixed}} + \|\epsilon - \hat{\epsilon}\|^2 + 2 \underbrace{\langle f - \hat{f}, \epsilon - \hat{\epsilon} \rangle}_0) \\
 &= \|f - \hat{f}\|^2 + E(\|\epsilon - \hat{\epsilon}\|^2) \\
 &= \|f - \hat{f}\|^2 + E(\|\epsilon\|^2) - E(\|\hat{\epsilon}\|^2) \tag{4.1} \\
 &= \|f - \hat{f}\|^2 + n\sigma^2 - p\sigma^2 \tag{4}
 \end{aligned}$$



The equation 4.1 holds because ϵ , $\hat{\epsilon}$, and $\epsilon - \hat{\epsilon}$ form a right-angled triangle:



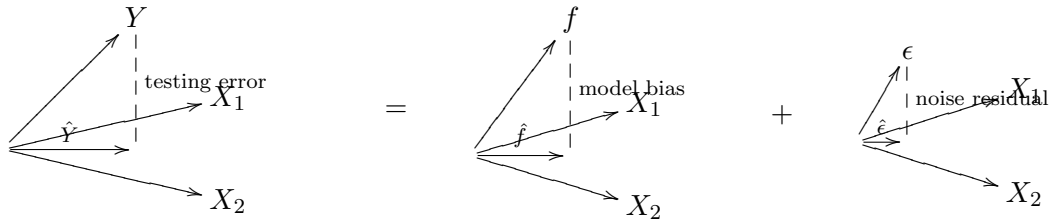
As we increase p (complexity of the model) both the model bias and the noise fitting residual decrease (Lemma 3.4), therefore training error decreases with increasing model complexity.

3.2 Testing Error

Now assume we used the same model to predict the values of test data ($Y_{new} = f + \epsilon_{new}$).

$$\begin{aligned}
 E(\|Y_{new} - \hat{Y}\|^2) &= E(\|(f + \epsilon_{new}) - (\hat{f} + \hat{\epsilon})\|^2) \\
 &= \|f - \hat{f}\|^2 + E(\|\epsilon_{new} - \hat{\epsilon}\|^2) \\
 &= \|f - \hat{f}\|^2 + E(\|\epsilon\|^2 + \|\hat{\epsilon}\|^2 - 2 \langle \epsilon_{new}, \hat{\epsilon} \rangle) \\
 &= \|f - \hat{f}\|^2 + E(\|\epsilon\|^2) + E(\|\hat{\epsilon}\|^2) - 2 \underbrace{E(\langle \epsilon_{new}, \hat{\epsilon} \rangle)}_0 \\
 &= \|f - \hat{f}\|^2 + E(\|\epsilon\|^2) + E(\|\epsilon_{new}\|^2) \\
 &= \|f - \hat{f}\|^2 + n\sigma^2 + p\sigma^2 \tag{5}
 \end{aligned}$$

$E(\langle \epsilon_{new}, \hat{\epsilon} \rangle) = 0$ because according to the assumptions observation errors are uncorrelated, therefore ϵ_{new} and ϵ are uncorrelated and as a result ϵ_{new} and $\hat{\epsilon}$ are also uncorrelated.



Equation 5 indicates that as the complexity of the model (p) increases, the testing error increases because the noise residual ($n\sigma^2 + p\sigma^2$) increases.

4 Discussion and Example

We just discovered a deep statistical phenomenon. By increasing the complexity of the model we start modeling the noise, and thus decreasing the training error beyond the model bias. But because noise is completely random, this will haunt us when we try to predict future values, and the testing error increases as in Figure 4.

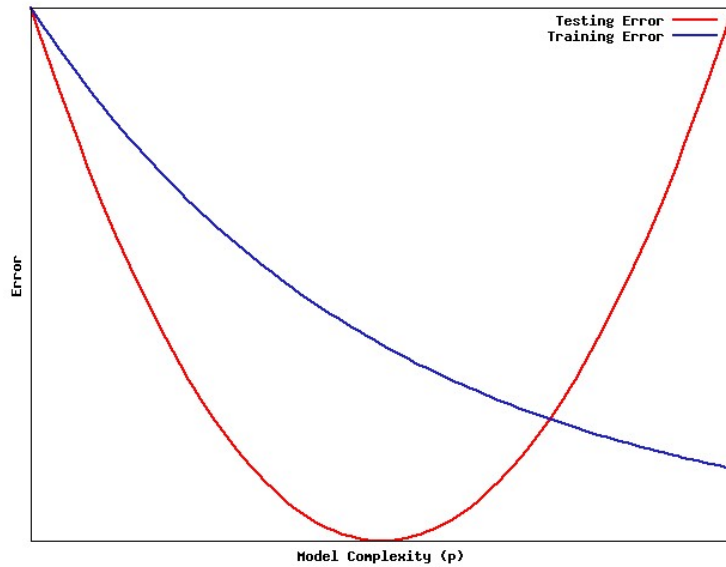


Figure 1: Training error and testing error as the model complexity increases

To understand this point better, consider an extreme case where the model is simply zero. In other words:

$$Y = 0 + \epsilon$$

In this case we can reach zero training error by setting $p = n$, but at the testing stage we will get:

$$E(\|Y_{new} - \hat{Y}\|^2) = n\sigma^2 + p\sigma^2 = 2n\sigma^2$$

Whereas by not using any models, $p = 0$, we could get testing error $n\sigma^2$.

This example indicates that by using a complex enough model we can model an absolutely random phenomenon (e.g., results of flipping fair coins), and get zero training error. But this does not indicate that the model is useful, because its testing error is higher than the variance of the data.

5 Acknowledgment

This note is based on Prof. Ying Nian Wu's lectures on Theoretical Statistics at UCLA.

References

- [1] Box, George E. P.; Norman R. Draper *Empirical Model-Building and Response Surfaces*, p. 424 Wiley. ISBN 0471810339.