Understanding the Link Between Mean Squared Error (MSE) and Entropy: From Univariate to Multivariate Models

Explanation based on concepts discussed in
Olsen, Pérez Ruiz, and Wiśniowski

by Diego Pérez Ruiz

May 9, 2025

Abstract

This document aims to provide an explanation of the relationship between the Mean Squared Error (MSE), a common metric for prediction accuracy, and entropy, a measure of uncertainty from information theory. We begin with the simpler univariate case involving a single predictor and then extend the concepts to the more complex multivariate case with multiple predictors. The discussion is based on the theoretical framework presented in the working paper *Children's Work and Child Labour: Prevalence Rates and The Importance of Plural Causality.*

1 Introduction: Prediction Error and Uncertainty

In statistical modeling, one of the main objective is to construct models that accurately predict an outcome variable. The Mean Squared Error (MSE) serves as a common metric for this, quantifying the average squared discrepancy between observed and predicted values. On the other hand, information theory offers tools such as entropy to measure the inherent uncertainty or randomness associated with variables. This document explores the connection between these two domains: specifically, how reducing the uncertainty about an outcome variable (i.e., lowering its conditional entropy given predictors) is linked to reducing the prediction error (i.e., lowering MSE).

2 The Univariate Case: A Single Predictor Variable

We first consider the scenario in which we want to predict a response variable Y using a single predictor variable X.

2.1 The Fundamental Prediction Model

The relationship between Y and X is typically modeled as:

$$Y = f(X) + \epsilon, \tag{1}$$

where:

- Y represents the actual observed value of the response variable.
- X is the observed value of the predictor variable.
- $f(X)$ denotes our prediction function. In an ideal setting, $f(X)$ is the conditional expectation of Y given X:
$$f(X) = E[Y|X]. \tag{2}$$
This conditional expectation provides the *best guess* for Y when X is known, in terms of minimising squared error.
- ϵ is the error term (or residual), capturing the portion of Y not explained by our prediction $f(X)$. Thus, $\epsilon = Y - f(X)$.

2.2 Measuring Prediction Error: Mean Squared Error (MSE)

The MSE quantifies the average squared difference between the actual values Y and the model's predictions $f(X)$:

$$\text{MSE} = E[(Y - f(X))^2] \tag{3}$$

By substituting $\epsilon = Y - f(X)$, the MSE can also be expressed as the expected squared error:

$$\text{MSE} = E[\epsilon^2] \tag{4}$$

If $f(X) = E[Y|X]$ (that is, the prediction is optimal in the least-squares sense), then the expected error $E[\epsilon]$ is zero. In this case, the MSE is equivalent to the variance of the error term, $\text{MSE} = \text{Var}(\epsilon)$. A lower MSE means that the predictions of our model are, on average, closer to the true observed values.

2.3 Quantifying Uncertainty: Entropy and Information

Entropy, a core concept of information theory, measures the uncertainty or randomness inherent in a variable.

- $H(Y)$: The entropy of Y. This quantifies the total uncertainty about Y *before* considering any information from X.
- $H(Y|X)$: The conditional entropy of Y given X. This measures the *remaining uncertainty* about Y *after* X has been observed. For discrete variables, it is defined as:
$$H(Y|X)_{\text{discrete}} = -\sum_{x,y} p(x,y) \log p(y|x) = -E[\log p(Y|X)] \tag{5}$$
while for continuous variables, where $p(x,y)$ is the joint probability density function and $p(y|x)$ is the conditional probability density function, the conditional differential entropy is defined as:
$$h(Y|X)_{\text{continuous}} = -\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x,y) \log p(y|x)\, dx\, dy = -E[\log p(Y|X)] \tag{6}$$

(This definition, particularly the expectation form, is analogous to Equation 2 in the Olsen et al. paper, which uses $p(x, y)$ and $p(y|x)$ generally). A highly informative predictor X will lead to a low $H(Y|X)$ (or $h(Y|X)$ for continuous cases), while a poor predictor will result in $H(Y|X)$ remaining high.

- $I(X;Y)$: The mutual information between X and Y. This represents the reduction in uncertainty about Y that is gained by knowing X:

$$I(X;Y) = H(Y) - H(Y|X) \tag{7}$$

Given our model $Y = f(X) + \epsilon$, the uncertainty in Y conditional on X is effectively the uncertainty associated with the error term ϵ (given X):

$$H(Y|X) = H(f(X) + \epsilon|X) = H(\epsilon|X) \tag{8}$$

(Analogous to Equation 5 in the Olsen et al. paper). If ϵ is independent of X, then $H(\epsilon|X)$ simplifies to $H(\epsilon)$, the entropy of the error term itself.

2.4 The Link Between MSE and Conditional Entropy (Univariate)

A key insight is the existence of a lower bound on prediction error, related to the conditional entropy. An important inequality states:

$$\text{MSE} = E[(f(X) - Y)^2] \geq \frac{1}{2\pi e} \exp\{2H(Y|X)\} \tag{9}$$

where e is the base of the natural logarithm. This inequality highlights that the MSE is fundamentally limited by the remaining uncertainty $H(Y|X)$; a lower conditional entropy (meaning that X explains more about Y) is necessary for achieving a lower MSE.

2.4.1 The Special Case of Normal Errors

A frequently assumption in regression is that the error term, ϵ, follows a Normal (Gaussian) distribution, $\epsilon \sim N(0, \sigma^2)$. Under this assumption:

- The MSE is equal to the error variance: $\text{MSE} = E[\epsilon^2] = \sigma^2$.
- The differential entropy of a normally distributed variable $\epsilon \sim N(0, \sigma^2)$ is given by:

$$H(\epsilon) = \frac{1}{2} + \frac{1}{2} \ln(2\pi\sigma^2) = \frac{1}{2} \ln(2\pi e\sigma^2) \tag{10}$$

(Equation 6 in Olsen et al.). If we equate $H(Y|X)$ with $H(\epsilon)$ (assuming errors are independent of X or consistently $N(0, \sigma^2)$ for any X), then:

$$H(Y|X) = \frac{1}{2} \ln(2\pi e\sigma^2) \tag{11}$$

This relationship can be rearranged to express σ^2 (and thus MSE) directly in terms of $H(Y|X)$ more precisely:

$$
\begin{aligned}
2H(Y|X) &= \ln(2\pi e\sigma^2) \\
\exp\{2H(Y|X)\} &= 2\pi e\sigma^2 \\
\sigma^2 &= \frac{1}{2\pi e}\exp\{2H(Y|X)\}.
\end{aligned}
$$

Therefore, for normally distributed errors, the MSE is directly related to the conditional entropy:

$$\text{MSE} = \frac{1}{2\pi e}\exp\{2H(Y|X)\}. \tag{12}$$

This equation explicitly shows that reducing the conditional entropy $H(Y|X)$ (i.e., making the predictor X more informative about Y) results in an exponential decrease in the MSE.

3 The Multivariate Case: Multiple Predictor Variables

Now, we are going to extend to scenarios involving multiple predictor variables, represented by a vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)$.

3.1 The Multivariate Prediction Model

The model takes the form:

$$Y = f(\mathbf{X}) + \epsilon, \tag{13}$$

where:

- $\mathbf{X}$ is the vector of predictor variables.
- $f(\mathbf{X}) = E[Y|\mathbf{X}]$ is the optimal prediction function.
- $\epsilon = Y - f(\mathbf{X})$ remains the error term.
- The MSE is still defined as $\text{MSE} = E[(Y - f(\mathbf{X}))^2] = E[\epsilon^2]$.
- $H(Y|\mathbf{X})$ now represents the conditional entropy of Y given the entire set of predictors $\mathbf{X}$. It quantifies the uncertainty about Y that remains after observing $X_1, \ldots, X_d$.

Here, the underlying principle persists: a more informative set of predictors $\mathbf{X}$ will lead to a lower $H(Y|\mathbf{X})$ and, consequently, a lower expected MSE.

3.2 Linking MSE and Conditional Entropy in the Multivariate OLS Context

When employing Ordinary Least Squares (OLS) regression with multiple predictors, the primary objective is still to find the linear combination of predictors that minimises the sum of squared residuals.

The resulting OLS Mean Squared Error (MSE_{OLS}) serves as an estimate of the variance of the error term ϵ in the model

$$Y = \beta_0 + \mathbf{X}\boldsymbol{\beta} + \epsilon.$$

The formula for calculating the MSE_{OLS} remains structurally identical to the univariate case:

$$\text{MSE}_{\text{OLS}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \tag{14}$$

Here, $\hat{Y}_i$ are the predictions from the multivariate OLS model ($\hat{Y}_i = \hat{\beta}_0 + \sum_{j=1}^{d} \hat{\beta}_j X_{ij}$), and e_i are the corresponding residuals. An unbiased estimator for the true error variance σ^2 is often used: $\hat{\sigma}^2 = \frac{1}{n-d-1}\sum e_i^2$, where d is the number of predictors (and $d+1$ is the number of estimated parameters including the intercept).

Note that while potential complexities exist, such as correlations among the predictor variables X_j (multicollinearity), primarily influence:

- The stability and interpretability of individual regression coefficient estimates ($\hat{\beta}_j$).
- The standard errors associated with these coefficient estimates.

However, these correlations do not alter the fundamental definition of the overall model's MSE as the average squared residual. Furthermore, the relationship between this MSE and the entropy of the residuals (if normally distributed) remains consistent.

If we assume that the residuals, e_i, from the multivariate OLS model are approximately normally distributed (which is expected if the true errors ϵ_i are themselves normal, i.e., $\epsilon_i \sim N(0, \sigma^2)$), then the connection to entropy remains the same as in the univariate scenario. The conditional entropy $H(Y|\mathbf{X})$ for the OLS model is effectively the entropy of its residuals, which we can denote as
$H(\text{residuals}_{\text{OLS}})$.

$$\hat{H}(\text{residuals}_{\text{OLS}}) = \frac{1}{2}\ln(2\pi e \cdot \text{MSE}_{\text{OLS}}). \tag{15}$$

Consequently, the MSE_{OLS} can be expressed in terms of this estimated residual entropy:

$$\text{MSE}_{\text{OLS}} = \frac{1}{2\pi e}\exp\{2\hat{H}(\text{residuals}_{\text{OLS}})\}. \tag{16}$$

It is important to emphasise that this relationship connects the overall fit of the model (as measured by MSE_{OLS}) to the uncertainty captured in the model residuals.

Crucially, this direct mathematical link between OLS MSE and residual entropy relies on the assumption that the model's residuals are normally distributed. While the predictor variables $\mathbf{X}$ determine the predicted values $\hat{Y}$ and thus the residuals, their own distribution or correlation structure does not directly enter this specific MSE-entropy formula for OLS.

Specific MSE formulas, such as those found in the Olsen et al. paper that explicitly incorporate predictor variances and correlations into the MSE definition, typically describe different theoretical constructions (e.g., particular forms of heteroscedasticity), and are distinct from this general OLS MSE formulation.

4 Conclusions

The theoretical framework connecting predictive accuracy (MSE) with informational content (conditional entropy $H(Y|\mathbf{X})$) provides valuable relationships. For standard OLS regression, especially when errors are assumed to be normally distributed, the MSE can be directly related to the entropy of the model residuals.

More specialised MSE formulas found in the literature often arise from models with more complex assumptions about error structures or causal mechanisms.

References

[1] Olsen, W., Pérez Ruiz, D. A., & Wiśniowski, A. (2022). *Children's Work and Child Labour: Prevalence Rates and The Importance of Plural Causality.* (Working Paper).