Statistical Inference Versus Predictive Modeling in Ordinary Least Squares Regression

Introduction

- There are two reasons to build statistical models: (1) for inference, and (2) for prediction.

- Statistical inference is focused on a set of formal hypotheses, denoted by H_0 for the *null hypothesis* and H_1 for the *alternate hypothesis*, and a test statistic with a known *sampling distribution*. A test statistic will have a specified distribution, e.g. the t-statistic for an OLS regression parameter has a t-distribution with the degrees-of-freedom equal to $n - p$, where p is the number of model parameters or the dimension of the model.

- Predictive modeling is focused on accurately producing an estimated value for the primary quantity of interest or assigning an observation to the correct class (group). Typically, when we use the term "predictive", we are referring to the model's ability to predict future or out-of-sample values, not in-sample values.

The Standard Modeling Process

1. Exploratory Data Analysis: How do our predictor variables relate to the response variable?

2. Model Identification: Which predictor variables should be included in our model?

3. Model Validation: Should we trust our models and the conclusions that we wish to derive from our model?

How we perform the Model Validation step is determined on the prescribed use of the model. Is the model to be used for statistical inference or is it to be used for predictive modeling?

Model Validation for Statistical Inference

- Model validation when the model is to be used for statistical inference is generally referred to as *the assessment of goodness-of-fit*.

- When we fit a statistical model, we have underlying assumptions about the probabilistic structure for that model. All of our statistical inference is derived from those probabilistic assumptions. Hence, if our estimated model, which is dependent upon the sample data, does not conform to these probabilistic assumptions, then our inference will be incorrect.

- When we validate a statistical model to be used for statistical inference, we are validating that the estimated model conforms to these probabilistic assumptions.

- For example in OLS regression we examine the residuals to make sure that they have a normal probability distribution and that they are ho-modscedastic.

Model Validation for Predictive Modeling

- Model validation when the model is to be used for predictive modeling is generally referred to as *the assessment of predictive accuracy*.

- When we fit a statistical model for predictive modeling, we can be much more tolerant of violations of the underlying probabilistic assumptions.

- Our primary interest in predictive modeling is estimating the response variable Y as "accurately" as possible. When validating a predictive model, we tend focus on summary statistics based on the quantity $(Y_i - \widehat{Y}_i)$. Examples include the Mean Absolute Error (MAE) and the Mean Squared Error (MSE).

- The evaluation of predictive models is typically performed through a form of *cross-validation* where the sample is split into a *training sample* and a *testing sample*. In this model validation, the model is estimated on the training sample and then evaluated out-of-sample on the testing sample.

Goodness-Of-Fit Versus Predictive Accuracy

- Goodness-Of-Fit
  1. Goodness-Of-Fit (GOF) is assessed in-sample.
  2. The objective is to confirm the model assumptions.
  3. In OLS regression the GOF is typically assessed using graphical procedures (scatterplots) for the model residuals $e_i = Y_i - \widehat{Y}_i$.

- Predictive Accuracy
  1. Predictive Accuracy (PA) is assessed out-of-sample.
  2. The objective is to measure the error of the predicted values.
  3. In OLS regression PA is typically assessed using error based metrics: Mean Square Error, Root Mean Square Error, and Mean Absolute Error.

Assessing the Goodness-Of-Fit in OLS Regression

- Validate the normality assumption: Produce a Quantile-Quantile Plot (QQ-Plot) of the residuals to compare their distribution to a normal distribution.

- Validate the homoscedasticity assumption (equal variance): Produce a scatterplot of the residuals against each predictor variable. If there is any structure in this plot, then the model will need a transformation of the predictor variable or an additional predictor variable added to the model.

- Interpret the R-Squared measure for your model. Applications tend to have typical ranges for "good" R-Squared values. If Model 1 has a R-Squared of 0.23 and Model 2 has a R-Squared of 0.54, then Model 2 should be prefered to Model 1, provided that Model 2 satisfies the other GOF conditions.

Statistical Inference in OLS Regression

If our Analysis of Goodness-Of-Fit for our OLS regression does not uncover any major violations of the underlying probabilistic assumptions, then we can feel confident in our use of the two primary forms of statistical inference in OLS regression.

- The t-test for the individual model coefficients:

$$H_0 : \beta_i = 0 \quad \text{versus} \quad H_1 : \beta_i \neq 0 \tag{1}$$

  for model coefficient i.

- The test statistic for the corresponding t-test is given by

$$t_i = \frac{\widehat{\beta}_i}{\mathrm{SE}(\widehat{\beta}_i)} \tag{2}$$

  where t_i has degrees of freedom equal to the sample size minus the number of model parameters, i.e. df = n - dim(Model).

Statistical Inference in OLS Regression - Continued

In addition to the "local" tests of a regression effect for the individual predictor variables, we also have a "global" test for a regression effect.

- The Overall F-test for a regression effect:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad \text{versus} \quad H_1 : \beta_i \neq 0 \tag{3}$$

  for some i, i.e. at least one of the predictor variables has an estimated coefficient that is statistically different from zero.

- The test statistic for the Overall F-test is given by

$$F_0 = \frac{SSR/k}{SSE/(n-p)} \tag{4}$$

  which has a F-distribution with $(k, n-p)$ degrees-of-freedom for a regression model with k predictor variables and p total parameters. When the regression model includes an intercept, then $p = k + 1$. If the regression model does not include an intercept, then $p = k$.

Predictive Accuracy in OLS Regression

The two primary metrics for assessing statistical models for out-of-sample predictive accuracy are Mean Square Error and Mean Absolute Error.

- Mean Square Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \tag{5}$$

  Root Mean Square Error (RMSE) is the square root of the MSE, i.e. $RMSE = \sqrt{MSE}$. There is no statistical reason to prefer one measure over the other. However, the RMSE can be used for presentation purposes when the MSE is very small or very large as the square root transformation will increase the small numbers and decrease the large numbers.

- Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \widehat{Y}_i| \tag{6}$$

The Bias-Variance Trade-Off

An interesting and useful property of the Mean Square Error (MSE) is that it can be decomposed into two components: the prediction variance and the square of the prediction bias. This decomposition is referred to as the *Bias-Variance Trade-Off*, and it is referenced throughout predictive modeling, especially in the presentation of concepts from statistical and machine learning.

- Throughout this presentation we have been using the *empirical Mean Square Error* for the predicted values $\widehat{Y}_i$.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \tag{7}$$

- The Bias-Variance Trade-Off is presented from the *theoretical Mean Square Error*

$$MSE = \mathbb{E}(Y_i - \widehat{Y}_i)^2 \tag{8}$$

  where $\mathbb{E}[X]$ denotes the mathematical expectation of X.

Derivation of the Bias-Variance Trade-Off

The derivation of the Bias-Variance Trade-Off is based on a standard algebraic trick of adding zero - $\mathbb{E}(\widehat{Y}_i) - \mathbb{E}(\widehat{Y}_i)$.

$$
\begin{aligned}
MSE &= \mathbb{E}(\widehat{Y}_i - Y_i)^2 && (9)\\
&= \mathbb{E}(\widehat{Y}_i - \mathbb{E}(\widehat{Y}_i) + \mathbb{E}(\widehat{Y}_i) - Y_i)^2 && (10)\\
&= \mathbb{E}[(\widehat{Y}_i - \mathbb{E}(\widehat{Y}_i)) + (\mathbb{E}(\widehat{Y}_i) - Y_i)]^2 && (11)\\
&= \mathbb{E}[(\widehat{Y}_i - \mathbb{E}(\widehat{Y}_i))^2] + 2\mathbb{E}[(\widehat{Y}_i - \mathbb{E}(\widehat{Y}_i))(\mathbb{E}(\widehat{Y}_i) - Y_i)] \\
&\quad +\mathbb{E}[(\mathbb{E}(\widehat{Y}_i) - Y_i)^2] && (12)\\
&= \mathbb{E}[(\widehat{Y}_i - \mathbb{E}(\widehat{Y}_i))^2] + \mathbb{E}[(\mathbb{E}(\widehat{Y}_i) - Y_i)^2] && (13)\\
&= Variance(\widehat{Y}_i) + Bias^2(\widehat{Y}_i) && (14)
\end{aligned}
$$

Note that the quantity $\mathbb{E}[(\widehat{Y}_i - \mathbb{E}(\widehat{Y}_i))(\mathbb{E}(\widehat{Y}_i) - Y_i)] = 0$.

Final Comments on the Bias-Variance Trade-Off

The crux of the Bias-Variance Decomposition is to note that both terms of the decomponsition are non-negative. Hence, we can choose to minimize either the variance or the bias.

- The variance of the predicted value is a measure of the spread of the predicted value from its mean.
- The bias of the predicted value is a measure of the distance from the mean of the predicted value to the target value.

Both of these components are functions of *model complexity*, i.e. the number of parameters in the model. Ideally, you would want have your prediction to be accurate (low bias) and precise (low variance). Bias will decline and variance will increase as the model complexity increases.

The Mean Square Error of the predicted values $\widehat{Y}_i$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \tag{15}$$

should not be confused with the estimate for variance parameter σ^2 in an OLS regression model with the Error Sum of Squares denoted by SSE and p parameters,

$$\widehat{\sigma}^2 = \frac{SSE}{n-p} \tag{16}$$

which is frequently referred to as the *mean square error of the regression* or the *mean square of the residuals*, but is not denoted by MSR as to not be confused with the *mean square of the regression* ($MSR = SSR/k$).

If you are in the context of a fitted OLS regression model, then the term MSE is referring to the estimate $\widehat{\sigma}^2$.