MA 575 Linear Models:

Cedric E. Ginestet, Boston University

Mean Squared Error and Comparing Models

Week 3, Lecture 2

1 Mean Squared Error (MSE)

Recall that the MSE is a theoretical quantity, which is defined as the expectation of the squared distance between the estimator and the target parameter,

$$\text{MSE}(\widehat{\beta}, \beta) := \mathbb{E}[(\widehat{\beta} - \beta)^2 | X].$$

This criterion should be contrasted with the RSS encountered earlier in the course. The RSS pertains to **model estimation**, since we are already assuming a given model for some particular data set; and it suffices to estimate the specific values of our estimators for the unknown parameters.

The MSE combines the previous two criteria, on the unbiasedness and the variance of $\widehat{\beta}$, through the following decomposition:

$$\begin{aligned}\mathbb{E}[(\widehat{\beta} - \beta)^2 | X] &= \mathbb{E}[(\widehat{\beta} - \mathbb{E}[\widehat{\beta}|X] + \mathbb{E}[\widehat{\beta}|X] - \beta)^2 | X] \\ &= \mathbb{E}\big[(\widehat{\beta} - \mathbb{E}[\widehat{\beta}|X])^2 | X\big] + 2\mathbb{E}\big[(\widehat{\beta} - \mathbb{E}[\widehat{\beta}|X])(\mathbb{E}[\widehat{\beta}|X] - \beta) | X\big] + \mathbb{E}\big[(\mathbb{E}[\widehat{\beta}|X] - \beta)^2 | X\big].\end{aligned}$$

Here, the cross-product can be seen to cancel out, since the second term in this cross-product does not depend on Y, it follows that we obtain,

$$\begin{aligned}\mathbb{E}[(\widehat{\beta} - \mathbb{E}[\widehat{\beta}|X])(\mathbb{E}[\widehat{\beta}|X] - \beta) | X] &= (\mathbb{E}[\widehat{\beta}|X] - \beta)\mathbb{E}[(\widehat{\beta} - \mathbb{E}[\widehat{\beta}|X]) | X] \\ &= (\mathbb{E}[\widehat{\beta}|X] - \beta)(\mathbb{E}[\widehat{\beta}|X] - \mathbb{E}[\widehat{\beta}|X]) = 0.\end{aligned}$$

Thus, the MSE admits the following decomposition, into a variance and a bias term:

$$\text{MSE}(\widehat{\beta}, \beta) = \mathbb{V}\text{ar}[\widehat{\beta}|X] + b^2(\widehat{\beta}),$$

where the bias of $\widehat{\beta}$ is defined as follows,

$$b^2(\widehat{\beta}) := \Big(\mathbb{E}[\widehat{\beta}|X] - \beta\Big)^2.$$

Other estimators can be considered such as in **ridge regression** and or in the **lasso**, for which a bias can be introduced, which will result in a lower variance, and an overall lower MSE. In particular,

1. In ridge regression, we consider a set of estimators, which are biased, but remain linear.
2. Whereas in using the lasso, we consider a set of estimators, which are unbiased and unlinear.

2 Comparing Models

2.1 Difference of Model-specific RSSs

As an alternative to the simple regression model, we may consider the following candidate model, composed of a single y-intercept,

$$\mathbb{E}[Y|X = x] = \beta_0.$$

This simpler model yields the following RSS,

$$\text{RSS}_1(\beta_0) := \sum_{i=1}^{n}(y_i - \beta_0)^2,$$

which is minimized by the sample mean,

$$\widehat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i =: \bar{y}.$$

Thus, it appears logical to consider the difference

$$\text{SSreg} := \text{RSS}_1(\widehat{\beta}_0) - \text{RSS}_2(\widehat{\beta}_0, \widehat{\beta}_1),$$

and to evaluate whether or not this difference is statistical important, in some precise sense. However, how shall we normalize this difference? But how large is a large difference?

2.2 F-test

The F-test for regression is defined using the following formulae for comparing model M_1 with model M_2,

$$F := \frac{\frac{\text{RSS}_1 - \text{RSS}_2}{\text{df}_1 - \text{df}_2}}{\frac{\text{RSS}_2}{\text{df}_2}}.$$

Since the degrees of freedom for these quantities are $\text{df}_1 = n - p_1$ and $\text{df}_2 = n - p_2$, respectively, we obtain,

$$F := \frac{\frac{\text{RSS}_1 - \text{RSS}_2}{p_2 - p_1}}{\frac{\text{RSS}_2}{n - p_2}}.$$

For simple regression, this gives $\text{df}_1 = n - 1$ and $\text{df}_2 = n - 2$, such that

$$F := \frac{(\text{RSS}_1 - \text{RSS}_2)/1}{\text{RSS}_2\,/(n-2)},$$

In your textbook, it is shown, using equation (2), that this formula can be re-written in this manner,

$$F := \frac{(\text{SYY} - \text{RSS})/1}{\widehat{\sigma}^2} = \frac{\text{SSreg}}{\widehat{\sigma}^2}. \tag{1}$$

That is, the F-statistic is simply defined as the **re-scaled** version of the difference $\text{SSreg} := \text{SYY} - \text{RSS}$. Therefore, we are here interested in conducting the following hypothesis test,

$$\begin{aligned} \text{H}_0 &: \quad \mathbb{E}[Y|X = x] = \beta_0, \\ \text{H}_1 &: \quad \mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x, \end{aligned}$$

We wish to test whether $\mathbb{E}[Y|X = x]$ is *constant* as x varies.

If the error terms are additionally assumed to be **iid** realizations from a normal distribution, such that

$$E_i \overset{\text{iid}}{\sim} N(0, \sigma^2),$$

then it can be shown that the F-statistic in equation (1) follows an F-distribution, which is denoted by

$$F \sim F(p_2 - p_1, n - p_2),$$

which follows from the fact that we are here considering a ratio of two **independent** random variables that both have a χ^2**-distribution** with respective degrees of freedom $p_2 - p_1$, and $n - p_2$. Then, the null hypothesis, H_0, is **rejected at level** α, if the p-value that one obtains is less than α.

2.3 Coefficient of Determination (R^2)

The coefficient of determination measures the **percentage of variance explained**. From the definition of SSreg, we have

$$\text{SSreg} = \text{RSS}(\widehat{\beta}_0) - \text{RSS}(\widehat{\beta}_0, \widehat{\beta}_1)$$

Dividing both sides by SYY, we obtain

$$\frac{\text{SSreg}}{\text{SYY}} = \frac{\text{RSS}(\widehat{\beta}_0)}{\text{SYY}} - \frac{\text{RSS}(\widehat{\beta}_0, \widehat{\beta}_1)}{\text{SYY}},$$

which simplifies to the so-called **coefficient of determination**, or R^2,

$$R^2 := \frac{\text{SSreg}}{\text{SYY}} = 1 - \frac{\text{RSS}(\widehat{\beta}_0, \widehat{\beta}_1)}{\text{SYY}}.$$

In the homework this week, you will be asked to show that,

$$\text{RSS}(\widehat{\beta}_0, \widehat{\beta}_1) = \text{SYY} - \frac{\text{SXY}^2}{\text{SXX}}. \tag{2}$$

Therefore, when comparing the two models based on (β_0) and (β_0, β_1), the SSreg can be re-written as

$$\begin{aligned} \text{SSreg} &= \text{RSS}(\widehat{\beta}_0) - \text{RSS}(\widehat{\beta}_0, \widehat{\beta}_1) \\ &= \text{SYY} - \left(\text{SYY} - \frac{\text{SXY}^2}{\text{SXX}} \right) = \frac{\text{SXY}^2}{\text{SXX}}. \end{aligned}$$

Then, the R^2 can be re-expressed as

$$R^2 := \frac{\text{SXY}^2}{\text{SYY} \cdot \text{SXX}} = \frac{[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})]^2}{[\sum_{i=1}^n (y_i - \bar{y})][\sum_{i=1}^n (x_i - \bar{x})]} = \widehat{r}^2_{xy}.$$

We will see, later in the course, that this relationship between the R^2 and the correlation coefficient still holds, when considering multiple regression, although in this case we need to consider the **multiple correlation** coefficient. The R^2, just like the correlation coefficient is therefore a **scale-free** measure of the strength of the relationship between Y and X.

The R^2, however, cannot be used for **model comparison**. Indeed, as we continue to add parameters the R^2 carries on increasing, without including any penalty for model complexity. The **adjusted** R^2 tries to control for this added complexity, and this quantity is automatically reported in R, but this type of correction is not widely accepted in the literature.

2.4 F-statistic and R^2

These two statistics have identical numerators, but different denominators,

$$F = \frac{\text{SSreg}}{\text{RSS}_2/(n-2)}, \qquad \text{and} \qquad R^2 = \frac{\text{SSreg}}{\text{RSS}_1}.$$

Since we have more parameters in RSS_2 than in RSS_1, it follows that we necessarily have $\text{RSS}_1 \geq \text{RSS}_2$; and therefore the R^2 is comprised between 0 and 1.

3 Degrees of Freedom

Why are we dividing RSS_1 by $n-1$, and RSS_2 by $n-2$? Can we give a substantive meaning to these choices of normalizing constants? In some precise sense, these degrees of freedom are quantifying the space of the domains of RSS_1 and RSS_2, and therefore capture a measure of the flexibility of these two statistics.

3.1 Constraints

The number of degrees of freedom of any statistic (i.e. functions of the data) is computed using the general language of **systems of linear equations**. In this theory, we may have the following system of m equations (or **constraints**) with 2 **coefficients**,

$$\begin{aligned} a_{11}\theta_1 + a_{12}\theta_2 &= b_1 \\ a_{21}\theta_1 + a_{22}\theta_2 &= b_2 \\ \vdots \qquad & \qquad \vdots \\ a_{m1}\theta_m + a_{m2}\theta_m &= b_m, \end{aligned}$$

If $m < 2$, then this system is *under-determined*, and if $m > 2$, then this system if *over-determined*. When we wish to determine the degrees of freedom of a particular statistic, such as $\text{RSS}_1(\widehat{\beta}_0)$, for instance,

$$\text{RSS}_1(\widehat{\beta}_0) = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0)^2,$$

we are, in fact, asking: What is the domain of that statistic? In order to answer this question, however, we need to consider the steps that were required to produce this particular quantity. Treating RSS_1 as a random variable, we have minimized

$$\widehat{\beta}_0(Y_1, \ldots, Y_n) = \underset{\beta_0 \in \mathbb{R}^1}{\text{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0)^2.$$

Now, differentiating with respect to $\widehat{\beta}_0$, we have $-2\sum_{i=1}^{n}(Y_i - \beta_0) = 0$. This yields the so-called **normal equation**,

$$\sum_{i=1}^{n} Y_i = n\beta_0.$$

Now, this can be regarded as a simple **system of linear equations** with a single equation and n coefficients. Here, the coefficients are the Y_i's, which are constrained by this particular equality. Thus, the degree of freedom of the statistic RSS_1 is the **dimension** of the **solution set** obtained when trying to solve the system of normal equations, *with respect to* the Y_i's.

Similarly, we may consider the system of linear equations that constrain the values taken by RSS_2,

$$-2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i) = 0,$$
$$-2\sum_{i=1}^{n} x_i(Y_i - \beta_0 - \beta_1 x_i) = 0.$$

These equations can be re-written in their **normal** forms,

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i,$$
$$\sum_{i=1}^{n} x_i Y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2,$$

where $\sum x_i$, $\sum Y_i$, $\sum x_i Y_i$ and $\sum x_i^2$ are **sufficient statistics** for (β_0, β_1). That is, these quantities contain all the required information from this data set about (β_0, β_1). Thus, the degrees of freedom of a particular model can be defined as the **number of data points**, minus the number of constraints, provided by the **normal equations**.

3.2 Estimator for σ^2

When considering the model RSS_2, we have the following estimator

$$\widehat{\sigma}^2 := \frac{\text{RSS}(\widehat{\beta}_1, \widehat{\beta}_2)}{n-2},$$

and this estimator can be shown to be **unbiased** with respect to σ^2,

$$\mathbb{E}[\widehat{\sigma}^2 | X] = \mathbb{V}\text{ar}[E_i | X] = \mathbb{V}\text{ar}[Y_i - \beta_0 - \beta_1 | X] = \sigma^2.$$

Thus, when computing the F-statistic, we select this particular denominator because this yields an unbiased estimator of σ^2.