## Measures of Predictive Performance

In the last module, we learned about how to get the best linear prediction for a given data. The next step is to check how good the BLP / regression model is?

In this section we have two goals:

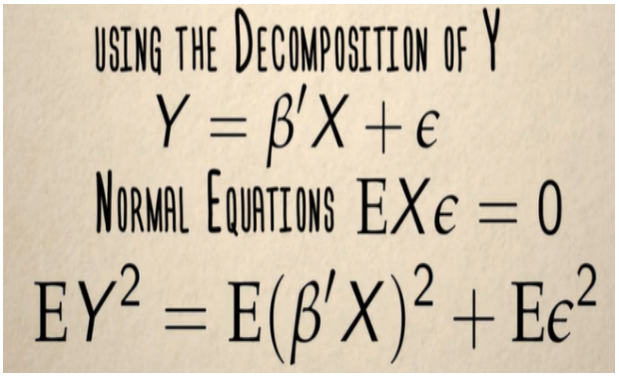
1. Understand the analysis of variance (ANOVA) in the population and in the sample.
2. Learn to assess the out-of-sample(test data) predictive performance of the sample linear regression.

Let's begin with the population case. In the normal equations that we derived in the previous segment, we can decompose the variation of Y into the sum of explained variation and residual variation, as shown in the formula **Y=**$\beta' X + \epsilon$**,**

We next define the population Mean Squared Prediction Error (or MSE), the expectation of $\epsilon$ squared where $\epsilon$ is the residual and is given by $\epsilon$ = **Y – β'X**

Mean Squared error $E\epsilon 2$ = **E(Y – β'X)$^2$ =** $\frac{1}{N}$**Σ( (Y – β'X)$^2$).**

Now, we also learned in the previous module that β'**X 'is the explained part,** while **epsilon($\epsilon$) residuals are also called the unexplained part** by our regression equation, therefore we can also define a metric to assess the quality of the regression line as R2, where the **R2 is the ratio of explained variation to the total variation.** In other words, the population R2 is the proportion of variation of Y explained by the BLP and as such, it is bounded *below by 0 and above by 1*.

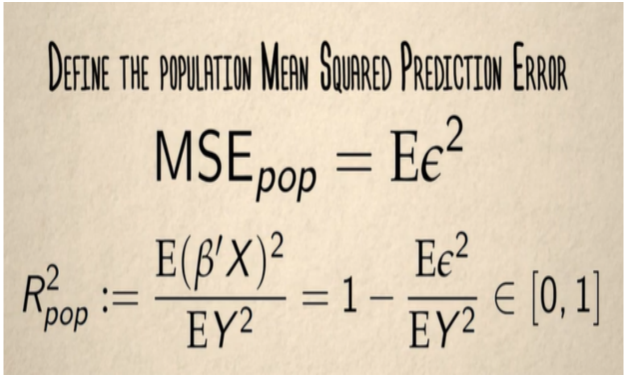


USING THE DECOMPOSITION OF Y

$$Y = \beta' X + \epsilon$$

NORMAL EQUATIONS $EX\epsilon = 0$

$$EY^2 = E(\beta' X)^2 + E\epsilon^2$$



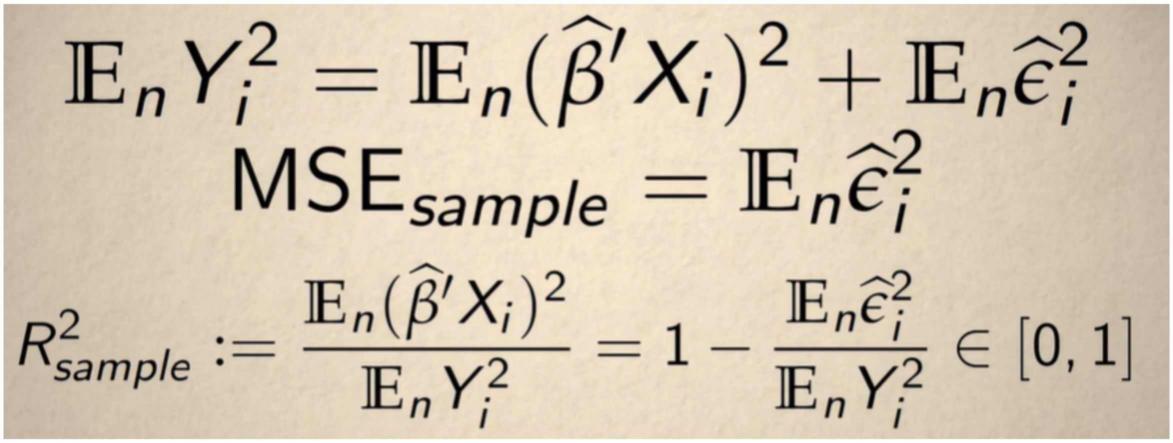
DEFINE THE POPULATION MEAN SQUARED PREDICTION ERROR

$$MSE_{pop} = E\epsilon^2$$

$$R^2_{pop} := \frac{E(\beta' X)^2}{EY^2} = 1 - \frac{E\epsilon^2}{EY^2} \in [0,1]$$

Where $EY^2$ is the variance of Y and is given by: $\frac{1}{N}\Sigma(Y - \overline{Y})^2$, $\overline{Y}$ is the average value of Y.

***The analysis of variance*** *in the sample proceeds analogously. We simply replaced population expectations with empirical expectations. Using the decomposition of Yi into the explained and unexplained part and the normal equations for the sample, we can decompose the sample variation of Yi into the sum of explained variation and residual variation. The former is given by the sample variance of the sample best linear predictor and the latter is given by the sample variance of the residual.*
*We can define the sample MSE as the sample variance of the residuals. And we can define the sample R2 as the ratio of the explained to the total variation in the sample.*

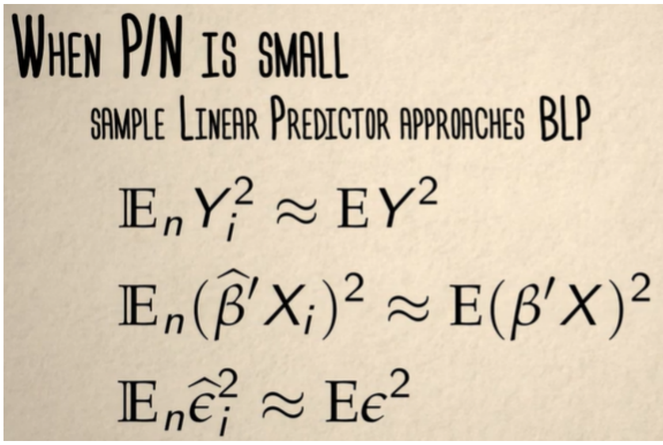


$$\mathbb{E}_n Y_i^2 = \mathbb{E}_n(\widehat{\beta}' X_i)^2 + \mathbb{E}_n \widehat{\epsilon}_i^2$$

$$\text{MSE}_{sample} = \mathbb{E}_n \widehat{\epsilon}_i^2$$

$$R^2_{sample} := \frac{\mathbb{E}_n(\widehat{\beta}' X_i)^2}{\mathbb{E}_n Y_i^2} = 1 - \frac{\mathbb{E}_n \widehat{\epsilon}_i^2}{\mathbb{E}_n Y_i^2} \in [0, 1]$$

We know that when p/n is small, the sample linear predictors get really close to the best linear predictor, thus, when p/n is small, we expect that sample averages of $E_n Y^2 i$, $E_n(\widehat{\beta}' Xi)^2$ and $E_n \widehat{\epsilon}^2$ be close to the population averages of EY2, $E(\beta' Xi)^2$ and $E\epsilon^2$



WHEN P/N IS SMALL

SAMPLE LINEAR PREDICTOR APPROACHES BLP

$$\mathbb{E}_n Y_i^2 \approx \mathrm{E} Y^2$$

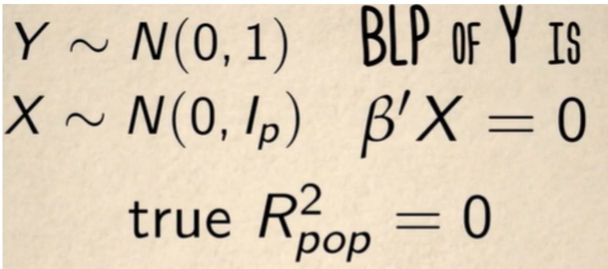
$$\mathbb{E}_n(\widehat{\beta}' X_i)^2 \approx \mathrm{E}(\beta' X)^2$$

$$\mathbb{E}_n \widehat{\epsilon}_i^2 \approx \mathrm{E}\epsilon^2$$

So, in this case, the sample R2 and the sample MSE will be close to the true quantities – the population R2 and the MSE. When p/n is not small, we expect the discrepancy between the two sets of measures can be substantial, and the sample R2 and the sample MSE are not good measures of the predictive ability.

For example, *when p is equal to n, we can have sample MSE equal to 0 and sample R2 equal 1 no matter what the population MSE or R2 is.*

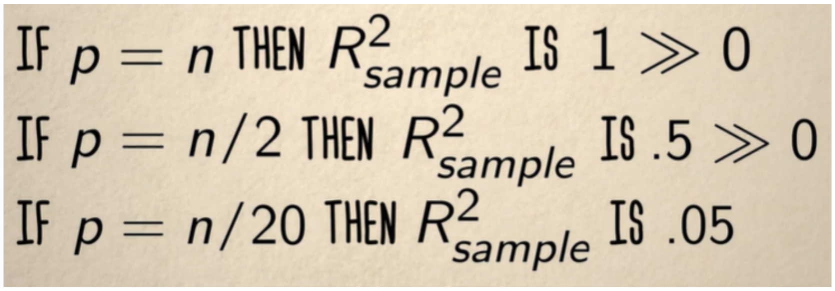
The following simulation example will support our reasoning. In this example, Y and X are statistically independent and generated from the normal distributions with mean 0 and variance 1. This means that the true linear predictors of Y given X are simply 0 and the true R2 is also zero.



$$Y \sim N(0,1) \qquad \text{BLP OF } Y \text{ IS}$$
$$X \sim N(0, I_p) \qquad \beta' X = 0$$
$$\text{true } R^2_{pop} = 0$$

Ip represents the standard deviation of the independent features.

Suppose the number of observations is n, the number of regressors is p. If p = n, then the typical sample R2 will be 1, which is very far away from the true number of zero. This is an example of extreme over-fitting. If p = n/2, then the typical sample R2 is about half, which is still far off from the truth. If p = n/20, then the typical sample R2 is about .05, which is no longer far off from the truth.



$$\text{IF } p = n \text{ THEN } R^2_{sample} \text{ IS } 1 \gg 0$$
$$\text{IF } p = n/2 \text{ THEN } R^2_{sample} \text{ IS } .5 \gg 0$$
$$\text{IF } p = n/20 \text{ THEN } R^2_{sample} \text{ IS } .05$$

For the simulation data, when p=n  sample R2 =1 but the true R2 is zero which is very far from the true value. When p=n/2 we have n observations and n/2 features and the true R^2 is 0.5 which is very far from the true value. So these values are observations from the above simulation data. There's no proof for this. If we want to replicate the results you need to follow a similar simulation and try it out.

The key takeaway is Adding more independent variables or predictors to a regression model tends to increase the R-squared value, which tempts makers of the model to add even more variables. Adjusted R-squared is used to determine how reliable the correlation is and how much it is determined by the addition of independent variables. So, we should not look at R^squared as it increases with the number of features. We should look at Adjusted R^square which explains whether the added variable is important to the model or not.

$$R^2_{adjusted} := 1 - \frac{n}{n-p} \frac{\mathbb{E}_n \widehat{\epsilon}_i^2}{\mathbb{E}_n Y_i^2}$$

$$\mathrm{MSE}_{adjusted} = \frac{n}{n-p} \mathbb{E}_n \widehat{\epsilon}_i^2$$

A more universal way to measure out-of-sample performance is to perform data splitting:

1. use a random part of data for estimating/training the prediction rule,

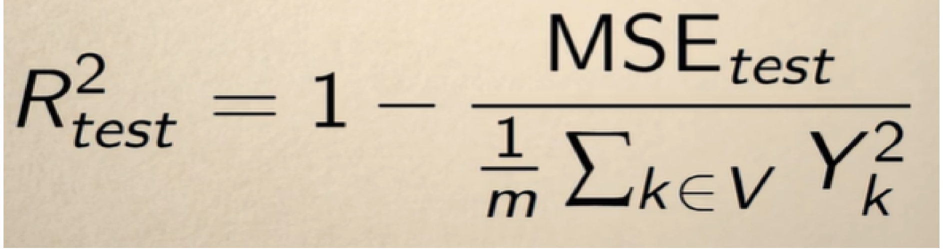
2. use the other part to evaluate the quality of the prediction rule, recording out-of-sample mean squared error and R2.

The part of data used for estimation is called the training sample.

The part of data used for evaluation is called the testing or validation sample.

Suppose we use n observations for training and m for testing/validation. **We use the training sample to compute $\hat{\beta}$.** Let V denote the indexes of the observations in the test sample. Then the out-of-sample/test mean squared error and R2 are:

$$\mathrm{MSE}_{test} = \frac{1}{m} \sum_{k \in V} (Y_k - \widehat{\beta}' X_k)^2$$



$$R^2_{test} = 1 - \frac{\mathrm{MSE}_{test}}{\frac{1}{m}\sum_{k\in V} Y_k^2}$$