statistical learning

Dawei Wang

February 5, 2022

An overview of statistical learning

Statistical learning refers to a vast of tools for understanding data. These tools can be classified as *supervised* or *unsupervised*. Broadly speaking, supervised statistical learning involves building a statistical model for predicting, or estimating, an *output* based on one or more *inputs*. With unsupervised statistical learning, there are inputs but no supervising output.

Regression problem

Classification problem

Clustering problem

Statistical learning

What Is Statistical Learning

Suppose that we observe a quantitative response Y and p different predictors, $X_1, X_2, \cdots, X_p$. We assume that there is some relationship between Y and $X = (X_1, X_2, \cdots, X_p)$, which can be written in the very general form

$$Y = f(X) + \varepsilon$$

Here f is some fixed but unknown function of $X_1, X_2, \cdots, X_p$, and ε is a random *error term*, which is independent of X and has mean zero. In this formulation, f represents the *systematic* information that X provides about Y.

Why Estimate f?

There are two main reasons that we may wish to estimate for f: *prediction* and *inference*. We discuss each in turn.

Prediction

In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained. In this setting, since the error term averages to zero, we can predict Y using

$$\hat{Y} = \hat{f}(X)$$

Where $\hat{f}$ represents our estimates for f, and $\hat{Y}$ represents the resulting prediction for Y.

The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities, which we will call the *reducible error* and the *irreducible error*. In general, $\hat{f}$ will not be a perfect estimate for f, and this inaccuracy will introduce some error. This error is *reducible* because we can potentially improve the accuracy of $\hat{f}$ by using the most appropriate statistical learning technique to estimate f. However, even if it were possible to form a perfect estimate for f, so that our estimated response took the form $\hat{Y} = f(X)$, our prediction would still have some error in it! This is because Y is also a function of ε, which, by definition, cannot be predicted using X. Therefore, variability associated with ε also affects the accuracy of our predictions. This is known as the *irreducible* error, because no matter how well we estimate f, we cannot reduce the error introduced by ε.

Why is the irreducible error larger than zero? The quantity ε may contain unmeasured variables that are useful in predicting Y : since we dont measure them, f cannot use them for its prediction. The quantity ε may also contain unmeasurable variation.

Consider a given estimate $\hat{f}$ and a set of predictors X, which yields the prediction $\hat{Y} = \hat{f}(X)$. Assume for a moment that both $\hat{f}$ and X are fixed.

Then, it is easy to show that

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \varepsilon - \hat{f}(X)]^2 \\ &= [f(X) - \hat{f}(X)]^2 + Var(\epsilon) \end{aligned}$$

where $E(Y - \hat{Y})^2$ represents the average, or expected value, of the squared difference between the predicted and actual value of Y, and $Var(\varepsilon)$ represents the variance associated with the error term ε.

The focus of is on techniques for estimating f with the aim of minimizing the reducible error. It is important to keep in mind that the irreducible error will always provide an upper bound on the accuracy of our prediction for Y. This bound is almost always unknown in practice.

Inference

We are often interested in understanding the way that Y is affected as $X_1, \cdots, X_p$ change. In this situation we wish to estimate f, but our goal is not necessarily to make predictions for Y. We instead want to understand the relationship between X and Y, or more specifically, to understand how Y changes as a function of $X_1, \cdots, X_p$. Now $\hat{f}$ cannot be treated as a black box, because we need to know its exact form. In this setting, one may be interested in answering the following questions:

1. Which predictors are associated with the response?

2. What is the relationship between the response and each predictor?

3. Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

Depending on whether our ultimate goal is prediction, inference, or a combination of the two, different methods for estimating f may be appropriate.

For example, linear models allow for relatively simple and interpretable inference, but may not yield as accurate predictions as some other approaches. In contrast, some of the highly non-linear approaches that we discuss in the later chapters of this book can potentially provide quite accurate predictions for Y, but this comes at the expense of a less interpretable model for which inference is more challenging.

How Do We Estimate f?

Broadly speaking, most statistical learning methods for this task can be characterized as either *parametric* or *non-parametric.* We now briefly discuss these two types of approaches.

Parametric methods

Parametric methods involve a two-step model-based approach.

1. First, we make an assumption about the functional form, or shape, of f.

2. After a model has been selected, we need a procedure that uses the training data to *fit* or *train* the model.

The model-based approach just described is referred to as parametric; it reduces the problem of estimating f down to one of estimating a set of parameters. Assuming a parametric form for f simplifies the problem of estimating f because it is generally much easier to estimate a set of parameters than it is to fit an entirely arbitrary function f.The potential disadvantage of a parametric approach is that the model we choose will usually not match the true unknown form of f. If the chosen model is too far from the true f, then our estimate will be poor. We can try to address this problem by choosing *flexible* models that can fit many different possible functional forms for f. But in general, fitting a more flexible model requires estimating a greater number of parameters. These more complex models can lead to a phenomenon known as *overfitting* the data, which essentially means they follow the errors, or *noise*, too closely. These issues are discussed throughout this book.

Non-parametric methods

Non-parametric methods do not make explicit assumptions about the functional form of f. Instead they seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly. Such approaches can have a major advantage over parametric approaches: by avoiding the assumption of a particular functional form for f, they have the potential to accurately fit a wider range of possible shapes for f. Any parametric approach brings with it the possibility that the functional form used to estimate f is very different from the true f, in which case the resulting model will not fit the data well. In contrast, non-parametric approaches completely avoid this danger, since essentially no assumption about the form of f is made. But non-parametric approaches do suffer from a major disadvantage: since they do not reduce the problem of estimating f to a small number of parameters, a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

The Trade-off Between Prediction Accuracy and Model Interpretability

Of the many methods that we examine in this book, some are less flexible, or more restrictive, in the sense that they can produce just a relatively small range of shapes to estimate f.

One might reasonably ask the following question: *why would we ever choose to use a more restrictive method instead of a very flexible approach?*

There are several reasons that we might prefer a more restrictive model. If we are mainly interested in inference, then restrictive models are much more interpretable.

We have established that when inference is the goal, there are clear advantages to using simple and relatively inflexible statistical learning methods. In some settings, however, we are only interested in prediction, and the interpretability of the predictive model is simply not of interest.

In this setting, we might expect that it will be best to use the most flexible model available. Surprisingly, this is not always the case! We will often obtain more accurate predictions using a less flexible method. This phenomenon, which may seem counterintuitive at first glance, has to do with the potential for overfitting in highly flexible methods.

Supervised versus Unsupervised Learning

Most statistical learning problems fall into one of two categories: supervised or unsupervised.

For each observation of the predictor measurement(s) $x_i, i = 1, \cdots, n$ there is an associated response measurement yi. We wish to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference).

In contrast, unsupervised learning describes the somewhat more challenging situation in which for every observation $i = 1, \cdots, n$, we observe a vector of measurements x_i but no associated response y_i.In this setting, we are in some sense working blind; the situation is referred to as *unsupervised* because we lack a response variable that can supervise our analysis. What sort of statistical analysis is possible? We can seek to understand the relationships between the variables or between the observations. One statistical learning tool that we may use in this setting is *cluster analysis*, or clustering. The goal of cluster analysis is to ascertain, on the basis of $x_1, \cdots, x_n$, whether the observations fall into relatively distinct groups.

Assessing Model Accuracy

There is no free lunch in statistics: no one method dominates all others over all possible data sets. On a particular data set, one specific method may work best, but some other method may work better on a similar but different data set. Hence it is an important task to decide for any given set of data which method produces the best results. Selecting the best approach can be one of the most challenging parts of performing statistical learning in practice. In this section, we discuss some of the most important concepts that arise in selecting a statistical learning procedure for a specific data set.

Measuring the Quality of Fit

In order to evaluate the performance of a statistical learning method on a given data set, we need some way to measure how well its predictions actually match the observed data. That is, we need to quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation. In the regression setting, the most commonly-used measure is the mean squared error (MSE), given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

where $\hat{f}(x_i)$ is the prediction of $\hat{f}$ gives for the ith observation. The MSE will be small if the predicted responses are very close to the true responses, and will be large if for some of the observations, the predicted and true responses differ substantially.

The MSE is computed using the training data that was used to fit the model, and so should more accurately be referred to as the *training MSE*. But in general, we do not really care how well the method works on the training data. Rather, *we are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data.*

Suppose that we fit our statistical learning method on our training observations $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ and we obtain the estimate $\hat{f}$. We can then compute $\hat{f}(x_1), \hat{f}(x_2), \cdots, \hat{f}(x_n)$. If these are approximately equal to $y_1, y_2, \cdots, y_n$, then the training MSE is small. However, we are really not interested in whether$\hat{f}(x_i) \approx y_i$; instead, we want to know whether $\hat{f}(x_0)$ is approximately equal to y_0, where (x_0, y_0) is a *previously unseen test observation not used to train the statistical learning method.* We want to choose the method that gives the lowest *test* MSE, as opposed to the lowest training MSE. In other words, if we had a large number of test observations, we could compute

$$Ave(\hat{f}(x_0) - y_0)^2$$

the average squared prediction error for these test observations (x_0, y_0). Wed like to select the model for which the average of this quantity—the test MSE—is as small as possible.

How can we go about trying to select a method that minimizes the test MSE? In some settings, we may have a test data set available that is, we may have access to a set of observations that were not used to train the statistical learning method. We can then simply evaluate test MSE on the test observations, and select the learning method for which the test MSE is smallest. But what if no test observations are available? In that case, one might imagine simply selecting a statistical learning method that minimizes the training MSE. This seems like it might be a sensible approach, since the training MSE and the test MSE appear to be closely related. Unfortunately, there is a fundamental problem with this strategy: there is no guarantee that the method with the lowest training MSE will also have the lowest test MSE. Roughly speaking, the problem is that many statistical methods specifically estimate coefficients so as to minimize the training set MSE. For these methods, the training set MSE can be quite small, but the test MSE is often much larger.

When a given method yields a small training MSE but a large test MSE, we are said to be *overfitting* the data. This happens because our statistical learning procedure is working too hard to find patterns in the training data, and may be picking up some patterns that are just caused by random chance rather than by true properties of the unknown function f. When we overfit the training data, the test MSE will be very large because the supposed patterns that the method found in the training data simply dont exist in the test data. Note that regardless of whether or not overfitting has occurred, we almost always expect the training MSE to be smaller than the test MSE because most statistical learning methods either directly or indirectly seek to minimize the training MSE. Overfitting refers specifically to the case in which a less flexible model would have yielded a smaller test MSE.

The Bias-Variance Trade-Off

The expected test MSE, for a given value x0, can always be decomposed into the sum of three fundamental quantities: the variance of $\hat{f}(x_0)$, the squared bias of $\hat{f}(x_0)$ and the variance of the error term ε. That is,

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\varepsilon)$$

$$\begin{aligned} E[(y - \hat{f})^2] &= E[(f + \varepsilon - \hat{f})^2] \\ &= E[(f + \varepsilon - \hat{f} + E[\hat{f}] - E[\hat{f}])^2] \\ &= (f - E[\hat{f}])^2 + E[\varepsilon^2] + E[(E[\hat{f}] - \hat{f})^2] \\ &= [Bias(\hat{f})]^2 + Var(\varepsilon) + Var(\hat{f}) \end{aligned}$$

What do we mean by the variance and bias of a statistical learning method? Variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set. Since the training data are used to fit the statistical learning method, different training data sets will result in a different $\hat{f}$. But ideally the estimate for f should not vary too much between training sets. However, if a method has high variance then small changes in the training data can result in large changes in $\hat{f}$. In general, more flexible statistical methods have higher variance. Generally, more flexible methods result in less bias.

As a general rule, as we use more flexible methods, the variance will increase and the bias will decrease. The relative rate of change of these two quantities determines whether the test MSE increases or decreases. As we increase the flexibility of a class of methods, the bias tends to initially decrease faster than the variance increases. Consequently, the expected test MSE declines. However, at some point increasing flexibility has little impact on the bias but starts to significantly increase the variance. When this happens the test MSE increases.

The relationship between bias, variance, and test set MSE is referred to as the bias-variance trade-off. Good test set performance of a statistical learning method requires low variance as well as low squared bias. This is referred to as a trade-off because it is easy to obtain a method with extremely low bias but high variance (for instance, by drawing a curve that passes through every single training observation) or a method with very low variance but high bias (by fitting a horizontal line to the data). The challenge lies in finding a method for which both the variance and the squared bias are low.

The Classification Setting

Thus far, our discussion of model accuracy has been focused on the regression setting. But many of the concepts that we have encountered, such as the bias-variance trade-off, transfer over to the classification setting with only some modifications due to the fact that y_i is no longer numerical.

Suppose that we seek to estimate f on the basis of training observations $\{(x_1, y_1), \cdots, (x_n, y_n)\}$, where now $y_1, \cdots, y_n$ are qualitative. The most common approach for quantifying the accuracy of our estimate $\hat{f}$ is the training *errorrate*, the proportion of mistakes that are made if we apply our estimate $\hat{f}$ to the training observations:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

The equation above is referred to as the *training error* rate because it is computed based on the data that was used to train our classifier. As in the regression setting, we are most interested in the error rates that result from applying our classifier to test observations that were not used in training. The *test error* rate associated with a set of test observations of the form (x_0, y_0) is given by

$$Ave(I(y_0 \neq \hat{y}_0))$$

where $\hat{y}_0$ is the predicted class label that results from applying the classifier to the test observation with predictor x_0. A good classifier is one for which the test error is smallest.

It is possible to show that the test error rate given in is minimized, on average, by a very simple classifier that *assigns each observation to the most likely class, given its predictor values.* In other words, we should simply assign a test observation with predictor vector x_0 to the class j for which

$$Pr(Y = j|X = x_0)$$

is largest.

The Bayes classifier produces the lowest possible test error rate, called the Bayes error rate. Since the Bayes classifier will always choose the class for which the equation above is largest, the error rate at X = x0 will be $1 - max_j Pr(Y = j|X = x_0)$. In general, the overall Bayes error rate is given by

$$1 - E(max_j Pr(Y = j|X))$$

where the expectation averages the probability over all possible values of X.

The Bayes error rate is analogous to the irreducible error, discussed earlier.

In theory we would always like to predict qualitative responses using the Bayes classifier. But for real data, we do not know the conditional distribution of Y given X, and so computing the Bayes classifier is impossible. Therefore, the Bayes classifier serves as an unattainable gold standard against which to compare other methods. Many approaches attempt to estimate the conditional distribution of Y given X, and then classify a given observation to the class with highest estimated probability. One such method is the K-nearest neighbors (KNN) classifier. Given a positive integer K and a test observation x0, the KNN classifier first identifies the K points in the training data that are closest to x0, represented by N0. It then estimates the conditional probability for class j as the fraction of points in N0 whose response values equal j:

$$Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

Finally, KNN applies Bayes rule and classifies the test observation x0 to the class with the largest probability.

The KNN test and training errors as a function of 1/K. As 1/K increases, the method becomes more flexible. As in the regression setting, the training error rate consistently declines as the flexibility increases. However, the test error exhibits a characteristic U-shape, declining at first before increasing again when the method becomes excessively flexible and overfits.

In both the regression and classification settings, choosing the correct level of flexibility is critical to the success of any statistical learning method. The bias-variance tradeoff, and the resulting U-shape in the test error, can make this a difficult task.

Linear Regression

Simple Linear Regression

Simple linear regression lives up to its name: it is a very straightforward approach for predicting a quantitative response Y on the basis of a single predictor variable X. It assumes that there is approximately a linear relationship between X and Y . Mathematically, we can write this linear relationship as

$$Y \approx \beta_0 + \beta_1 X$$

We will sometimes describe the equation above by saying that we are regressing Y on X (or Y onto X).

Once we have used our training data to produce estimates 0 and 1 for the model coefficients, we can predict Y on the basis of a particular value of X by computing

$$\hat{y} = \hat{\beta_0} + \hat{\beta_1} x$$

where $\hat{y}$ indicates a prediction of Y on the basis of X = x. Here we use a hat symbol, ˆ , to denote the estimated value for an unknown parameter or coefficient, or to denote the predicted value of the response.

Estimating the Coefficients

There are a number of ways of measuring closeness. However, by far the most common approach involves minimizing the least squares criterion.

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith *residual* — this is the difference between the ith observed response value and the ith response value that is predicted by our linear model. We define the residual sum of squares (RSS) as

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. Using some calculus, one can show that the minimizers are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Assessing the Accuracy of the Coefficient Estimates

We assume that the true relationship between X and Y takes the form $Y = f(X) + \varepsilon$ for some unknown function f, where ε is a mean-zero random error term. If f is to be approximated by a linear function, then we can write this relationship as

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

The model given by the equation above defines the population regression line, which is the best linear approximation to the true relationship between X and Y.

We can wonder how close $\hat{\beta}_0$ and $\hat{\beta}_1$ are to the true values β_0 and β_1. To compute the standard errors associated with $\hat{\beta}_0$ and $\hat{\beta}_1$, we use the following formulas:

$$SE(\hat{\beta}_0)^2 = \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x}^2)}], \quad SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x}^2)}$$

where $\sigma^2 = Var(\varepsilon)$. For these formulas to be strictly valid, we need to assume that the errors ε_i for each observation are uncorrelated with common variance σ^2.

In general, σ^2 is not known, but can be estimated from the data. This estimate is known as the *residual standard error*, and is given by the formula $RSE = \sqrt{RSS/(n-2)}$. Strictly speaking, when σ^2 is estimated from the data we should write $\hat{SE}(\hat{\beta}_1)$ to indicate that an estimate has been made, but for simplicity of notation we will drop this extra hat.

Standard errors can be used to compute confidence intervals.

Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of

$$H_0 : There\ is\ no\ relationship\ between\ X\ and\ Y$$

versus the *alternative hypothesis*

$$H_A : There\ is\ some\ relationship\ between\ X\ and\ Y$$

Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$

versus:

$$H_A : \beta_1 \neq 0$$

Since if $\beta_1 = 0$ then $Y = \beta_0 + \varepsilon$, and X is not associated with Y. To test the null hypothesis, we need to determine whether $\hat{\beta}_1$ is sufficiently far from zero that we can be confident that β_1 is non-zero.

How far is far enough? This of course depends on the accuracy of $\hat{\beta}_1$ that is, it depends on $SE(\hat{\beta}_1)$. If $SE(\hat{\beta}_1)$ is small, then even relatively small values of $\hat{\beta}_1$ may provide strong evidence that $\beta_1 \neq 0$, and hence that there is a relationship between X and Y. In contrast, if $SE(\hat{\beta}_1)$ is large, then $\hat{\beta}_1$ must be large in absolute value in order for us to reject the null hypothesis. In practice, we compute a t-statistic, given by

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

which measures the number of standard deviations that $\hat{\beta}_1$ is away from 0. If there really is no relationship between X and Y, then we expect that the equation above will have a t-distribution with n 2 degrees of freedom.

Assessing the Accuracy of the Model

Once we have rejected the null hypothesis in favor of the alternative hypothesis, it is natural to want to quantify the extent to which the model fits the data. The quality of a linear regression fit is typically assessed using two related quantities: the residual standard error (RSE) and the R^2 statistic.

Residual Standard Error

Due to the presence of the error terms, even if we knew the true regression line (i.e. even if β_0 and β_1 were known), we would not be able to perfectly predict Y from X. The RSE is an estimate of the standard deviation of ε. Roughly speaking, it is the average amount that the response will deviate from the true regression line. It is computed using the formula

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

The RSE is considered a measure of the lack of fit of the linear model to the data. If the predictions obtained using the model are very close to the true outcome values—that is, if $\hat{y}_i \approx y_i$ for $i = 1, \cdots, n$—then RSE will be small, and we can conclude that the model fits the data very well. On the other hand, if $\hat{y}_i$ is very far from y_i for one or more observations, then the RSE may be quite large, indicating that the model doesnt fit the data well.

R^2 Statistic

The RSE provides an absolute measure of lack of fit of the linear model to the data. But since it is measured in the units of Y, it is not always clear what constitutes a good RSE. The R2 statistic provides an alternative measure of fit. It takes the form of a proportion the proportion of variance explained — and so it always takes on a value between 0 and 1, and is independent of the scale of Y. To calculate R2, we use the formula

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

where $TSS = \sum(y_i - \bar{y})^2$ is the *total sum of squares.* TSS measures the total variance in the response Y , and can be thought of as the amount of variability inherent in the response before the regression is performed. In contrast, RSS measures the amount of variability that is left unexplained after performing the regression. Hence, TSSRSS measures the amount of variability in the response that is explained (or removed) by performing the regression, and R^2 measures the *proportion of variability in Y that can be explained using X.* An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression. A number near 0 indicates that the regression did not explain much of the variability in the response; this might occur because the linear model is wrong, or the inherent error 2 is high, or both.

The R^2 statistic has an interpretational advantage over the RSE, since unlike the RSE, it always lies between 0 and 1. However, it can still be challenging to determine what is a good R^2 value, and in general, this will depend on the application.

The R^2 statistic is a measure of the linear relationship between X and Y. Recall that correlation, defined as

$$Cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y_i})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y_i})^2}}$$

is also a measure of the linear relationship between X and Y. This suggests that we might be able to use $r = Cor(X,Y)$ instead of R^2 in order to assess the fit of the linear model. In fact, it can be shown that in the simple linear regression setting, $R^2 = r^2$. In other words, the squared correlation and the R^2 statistic are identical. However, in the next section we will discuss the multiple linear regression problem, in which we use several predictors simultaneously to predict the response. The concept of correlation between the predictors and the response does not extend automatically to this setting, since correlation quantifies the association between a single pair of variables rather than between a larger number of variables. We will see that R^2 fills this role.

Multiple Linear Regression