# **Practical Machine Learning**

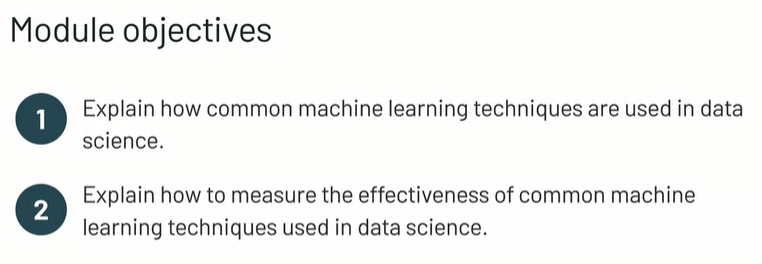
**Learning Objectives**

* Describe common machine learning techniques and how to measure their effectiveness.
* Describe and evaluate linear regression for regression problems
* Describe how to ensure machine learning models generalize to out-of-sample data
* Describe and evaluate logistic regression for classification problems
* Describe the decision tree as a tool for supervised learning problems

# **Regression Modelling**

# **Module and Lesson Introduction**

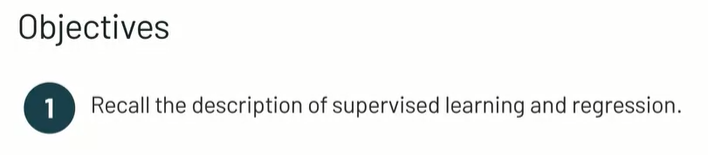
Welcome to another module of data science fundamentals for data analyst. [Veri analisti için başka bir veri bilimi temelleri modülüne hoş geldiniz.] In the previous module, you learned about connecting data science to real-world applications. [Önceki modülde, veri bilimini gerçek dünya uygulamalarına bağlamayı öğrendiniz.] In this module, we'll introduce some key machine learning concepts. [Bu modülde, bazı temel makine öğrenimi kavramlarını tanıtacağız.] We've titled this module practical machine learning and there are four lessons in it. [Bu modüle pratik makine öğrenimi adını verdik ve içinde dört ders var.]



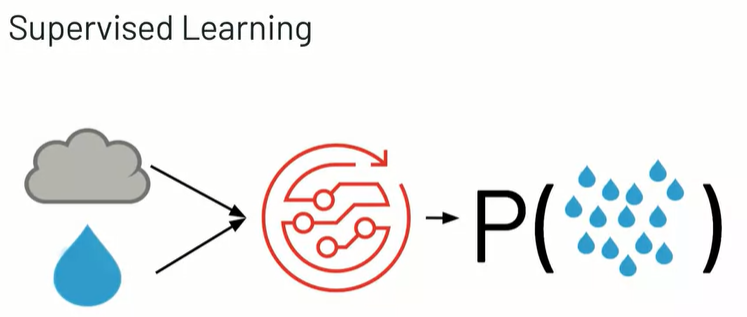
The overall objective for this module is to describe common machine learning techniques and how to measure their effectiveness. [Bu modülün genel amacı, yaygın makine öğrenimi tekniklerini ve bunların etkinliğinin nasıl ölçüleceğini açıklamaktır.] Just like in the previous modules, the videos will focus on concepts, and the activities will allow you to apply what you've learned. [Tıpkı önceki modüllerde olduğu gibi, videolar kavramlara odaklanacak ve aktiviteler öğrendiklerinizi uygulamanıza izin verecek.] Welcome to the module, have fun. [Modüle hoş geldiniz, iyi eğlenceler.]

# **A Review of Supervised Learning and Regression**

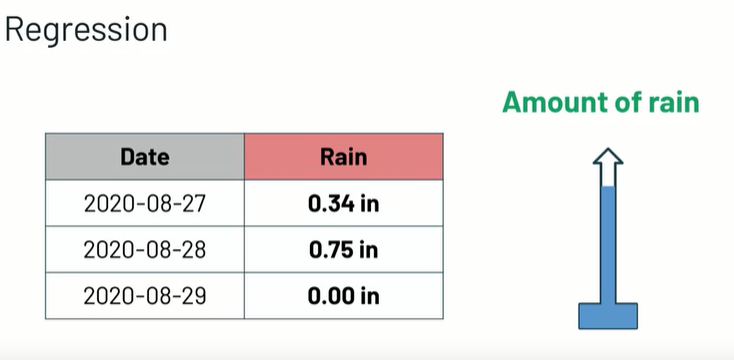
Hello and welcome back. [Merhaba ve tekrar hoş geldiniz.]



In this video, we'll recall the description of supervised learning and regression, that we learned about previously in this course. [Bu videoda, daha önce bu derste öğrendiğimiz denetimli öğrenme ve regresyon açıklamasını hatırlayacağız.]



We previously learned about different types of machine learning. [Daha önce farklı makine öğrenimi türlerini öğrenmiştik.] One application of machine learning is for predicting some outcome. [Makine öğreniminin bir uygulaması, bazı sonuçları tahmin etmek içindir.] For example, predicting whether or not it will rain. [Örneğin, yağmur yağıp yağmayacağını tahmin etmek.] This is a type of supervised machine learning, which is when an algorithm is learning a function that maps a collection of inputs to an output, based on example pairs of inputs and output, which is our training data. [Bu bir tür denetimli makine öğrenimidir; bu, bir algoritmanın, eğitim verilerimiz olan örnek girdi ve çıktı çiftlerine dayalı olarak bir girdi koleksiyonunu bir çıktıya eşleyen bir işlevi öğrendiği zamandır.] In the case of supervised learning, the algorithm is being supervised by certain truth values or labels, for the function output. [Denetimli öğrenme durumunda, algoritma, işlev çıktısı için belirli doğruluk değerleri veya etiketleri tarafından denetlenir.]



Recall that regression is a subset of supervised learning, where the model is trying to predict a quantity or amount, based on continuous label values. [Regresyonun, modelin sürekli etiket değerlerine dayalı olarak bir miktar veya miktarı tahmin etmeye çalıştığı denetimli öğrenmenin bir alt kümesi olduğunu hatırlayın.] For example, the amount of rain on any given date. [Örneğin, herhangi bir tarihteki yağmur miktarı.] This is different than predicting a class label, or a probability of a class label, which is classification. [Bu, bir sınıf etiketini veya sınıflandırma olan bir sınıf etiketi olasılığını tahmin etmekten farklıdır.] Now that we've learned some of the concepts of regression and recapped how regression fits into supervised learning, in the next video, we'll move on to talking about the linear regression algorithm, and some of its key concepts of how it works. [Artık bazı regresyon kavramlarını öğrendiğimize ve regresyonun denetimli öğrenmeye nasıl uyduğunu özetlediğimize göre, bir sonraki videoda lineer regresyon algoritması ve nasıl çalıştığına dair bazı temel kavramları hakkında konuşmaya geçeceğiz.]

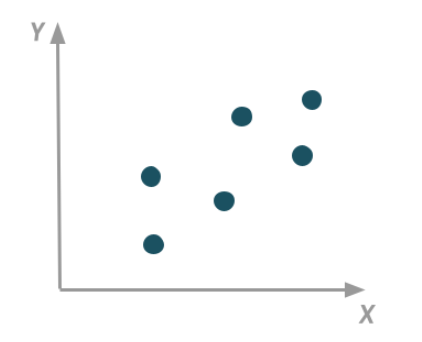
# **An Introduction to Linear Regression**

One of the most common and foundational regression algorithms is **linear regression**. [En yaygın ve temel regresyon algoritmalarından biri doğrusal regresyondur.] In this reading, we’ll describe linear regression as a tool for modeling regression problems. [Bu okumada, doğrusal regresyonu, regresyon problemlerini modellemek için bir araç olarak tanımlayacağız.]

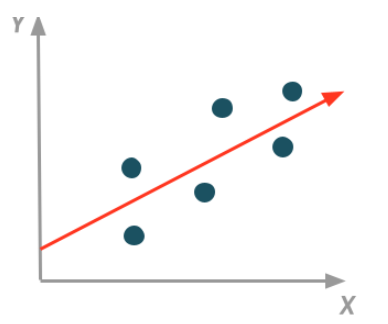
**Simple Linear Regression**

The simplest form of linear regression is aptly named **simple linear regression**. [Doğrusal regresyonun en basit biçimi, uygun bir şekilde basit doğrusal regresyon olarak adlandırılır.] Simply linear regression derives a function to map a single input variable (also called a predictor, independent variable, or feature) to a single output variable (also called a response variable, target variable, or dependent variable). [Basitçe doğrusal regresyon, tek bir girdi değişkenini (aynı zamanda bir tahmin edici, bağımsız değişken veya özellik olarak da adlandırılır) tek bir çıktı değişkenine (yanıt değişkeni, hedef değişken veya bağımlı değişken olarak da adlandırılır) eşlemek için bir işlev türetir.] It’s called linear because the function it derives represents a linear relationship between the input and the output. [Türettiği fonksiyon girdi ve çıktı arasında doğrusal bir ilişkiyi temsil ettiği için doğrusal olarak adlandırılır.]

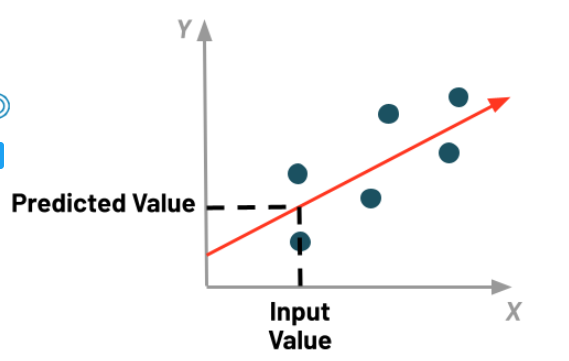
The image below displays points with a single input variable on the X-axis and the corresponding output variable on the Y-axis: [Aşağıdaki resim, X ekseninde tek bir giriş değişkenine ve Y ekseninde karşılık gelen çıkış değişkenine sahip noktaları gösterir:]



Simple linear regression’s objective is to find the **line-of-best fit** between these points:



The predictions of the simple linear regression for each value of the input variable will be the corresponding value of the dependent variable along the line of best fit.



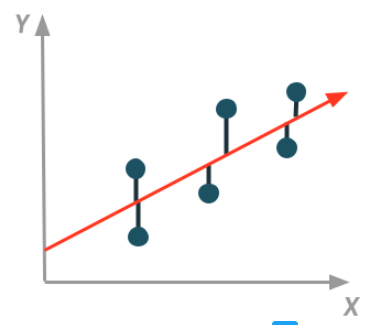
**Computing the Line of Best Fit**

**Sum of Absolute Errors**

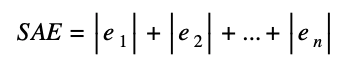
You might be wondering how linear regression determines the location of that line of best fit. [Doğrusal regresyonun bu en uygun doğrunun yerini nasıl belirlediğini merak ediyor olabilirsiniz.]

Recall that our objective with regression is to closely predict the actual output variable values. [Gerileme ile amacımızın gerçek çıktı değişkeni değerlerini yakından tahmin etmek olduğunu hatırlayın.] As a result, a linear regression model is better if its resulting line of best fit (which are the predictions for each value of the input) is close to each of the actual output values. [Sonuç olarak, bir lineer regresyon modeli, ortaya çıkan en iyi uyum çizgisi (girdinin her değeri için tahminlerdir) gerçek çıktı değerlerinin her birine yakınsa daha iyidir.]

There are a lot of different ways to measure the closeness of the line to the actual values. [Doğrunun gerçek değerlere yakınlığını ölçmenin birçok farklı yolu vardır.] For example, we could measure the simple vertical distance between each line and its values – we call these **residuals**. [Örneğin, her bir çizgi ile değerleri arasındaki basit dikey mesafeyi ölçebiliriz – bunlara artıklar diyoruz.]



Residuals that are shorter, or have a length closer to zero, represent a better prediction. [Daha kısa veya sıfıra yakın bir uzunluğa sahip artıklar daha iyi bir tahmin temsil eder.] If most of the residuals are close to zero, then the line is a better fit. [Kalıntıların çoğu sıfıra yakınsa, o zaman çizgi daha uygundur.] We could measure the overall quality of the line of best fit by summing together each of the absolute values of the residuals – we call this the sum of absolute errors. [Artıkların mutlak değerlerinin her birini bir araya toplayarak en uygun çizginin genel kalitesini ölçebiliriz - buna mutlak hataların toplamı diyoruz.]

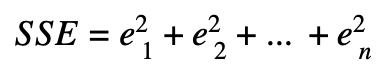


We have to take the absolute value of each residual or error. [Her artık veya hatanın mutlak değerini almalıyız.] Otherwise, the positive and negative residuals would cancel each other out. [Aksi takdirde, pozitif ve negatif artıklar birbirini yok edecektir.]

**Sum of Squared Errors**

While calculating the sum of absolute errors does compute a line of best fit, most linear regression algorithms use another method: the sum of squared errors. [Mutlak hataların toplamını hesaplamak en uygun çizgiyi hesaplarken, çoğu doğrusal regresyon algoritması başka bir yöntem kullanır: karesel hataların toplamı.]

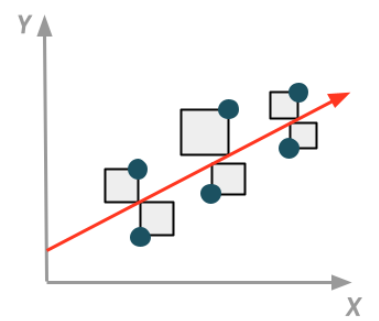
Instead of taking the absolute value of each residual, the sum of squared errors method squares each residual. [Her bir artığın mutlak değerini almak yerine, karesel hataların toplamı yöntemi her bir artığın karesini alır.]



This has two primary impacts: [Bunun iki temel etkisi vardır:]

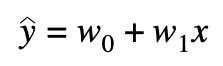
1. Similar to taking the absolute value, each residual will be represented as a positive number. [Mutlak değer almaya benzer şekilde, her artık pozitif bir sayı olarak temsil edilecektir.] This keeps positive and negative errors from canceling each other out. [Bu, pozitif ve negatif hataların birbirini iptal etmesini engeller.]
2. In contrast to taking the absolute value, more extreme errors are highly penalized as a result of squaring. [Mutlak değer almanın aksine, kare almanın bir sonucu olarak daha aşırı hatalar yüksek oranda cezalandırılır.] While this can keep the linear regression algorithm from ignoring poor predictions, it can also make it particularly susceptible to outliers. [Bu, lineer regresyon algoritmasının zayıf tahminleri görmezden gelmesini engellese de, onu özellikle aykırı değerlere karşı hassas hale getirebilir.]

The sum of squared errors method is commonly called the ordinary least squares method because it tries to minimize the area of the sum of squared residuals: [Karesi alınmış hataların toplamı yöntemine, karesi alınmış artıkların toplamının alanını en aza indirmeye çalıştığı için genellikle sıradan en küçük kareler yöntemi denir:]



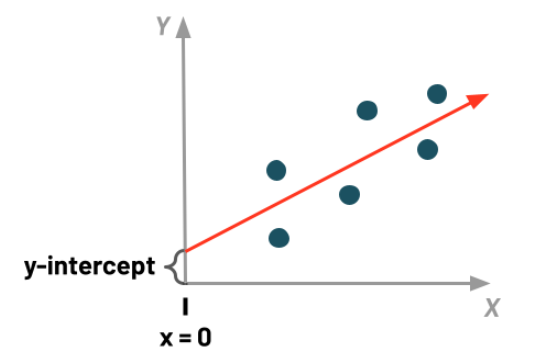
**Model Equation**

This line of best fit can be represented by the following equation:

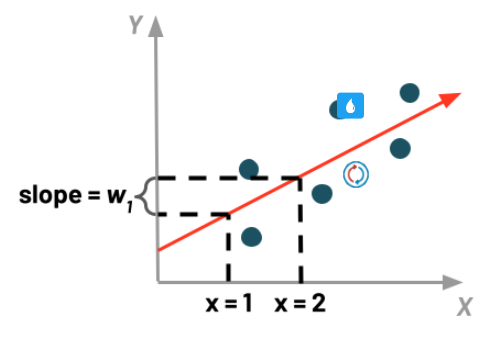


In this equation, ŷ represents the predicted value. This value varies based on the value of our input variable, x.

In addition, w0 represents what’s known as the **y-intercept**. This is the value of the prediction ŷ when the input variable x is equal to 0. It corresponds to the Y-axis value in our linear regression visual:



Next, there is w1 which represents the slope of the line of best fit with respect to input variable x and output variable y. As input variable x increases by one, the value of the prediction ŷ increases by w1. As input variable x decreases by one, the value of the prediction ŷ decreases by w1.

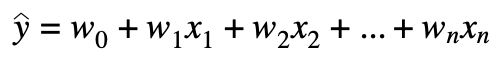


We call this value w1 a **coefficient**.

**Multiple Linear Regression**

So far, we’ve only considered simple linear regression – linear regression models that only have one input variable. But in practice, there are usually a lot of different variables impacting an output variable.

When we want to model this scenario, we use **multiple linear regression**. Multiple linear regression uses multiple input variables to predict a single output variable. Its equation is extended to include more input variables and corresponding coefficients, but it still contains a single prediction and a single y-intercept.



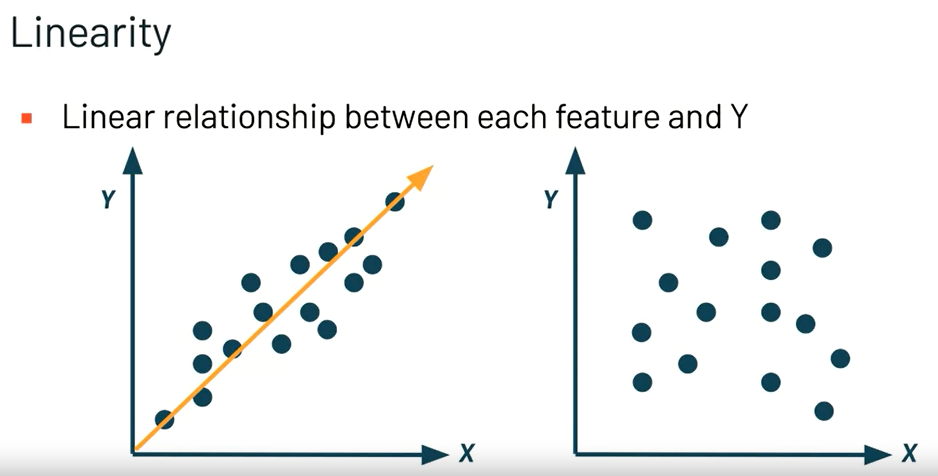
In this equation, each input variable xi has its own coefficient wi – therefore, it has its own linear relationship with the output variable independent from all other input variables, according to the model.

In multiple regression models, these coefficients represent the average effect on the output variable from increasing the respective input variable’s value by one, while holding all other input variables constant. That’s a key distinction from simple linear regression: rather than simply ignoring all other input variables by not including them in the model, multiple linear regression fixes the other input variable values.

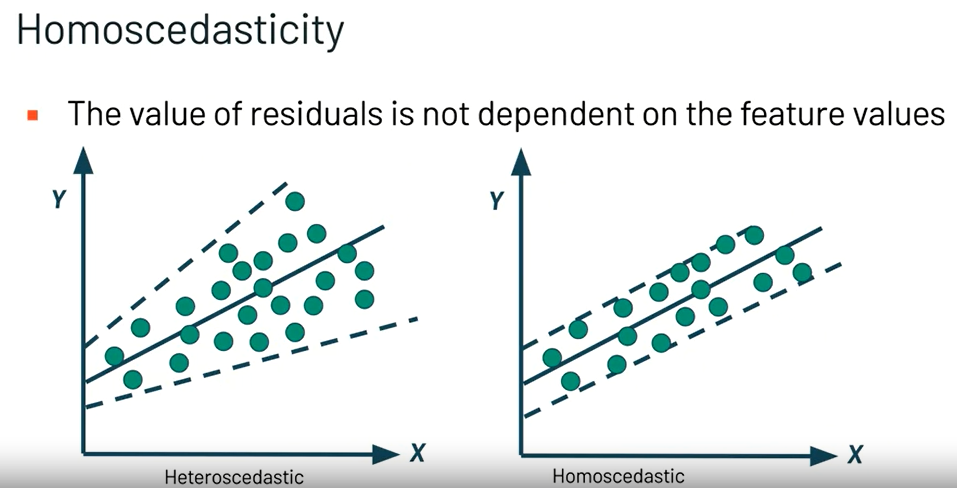
However, this might not actually be the case in the real world, and we’ll cover more on this assumption and others in the next video.

# **Linear Regression Assumptions**

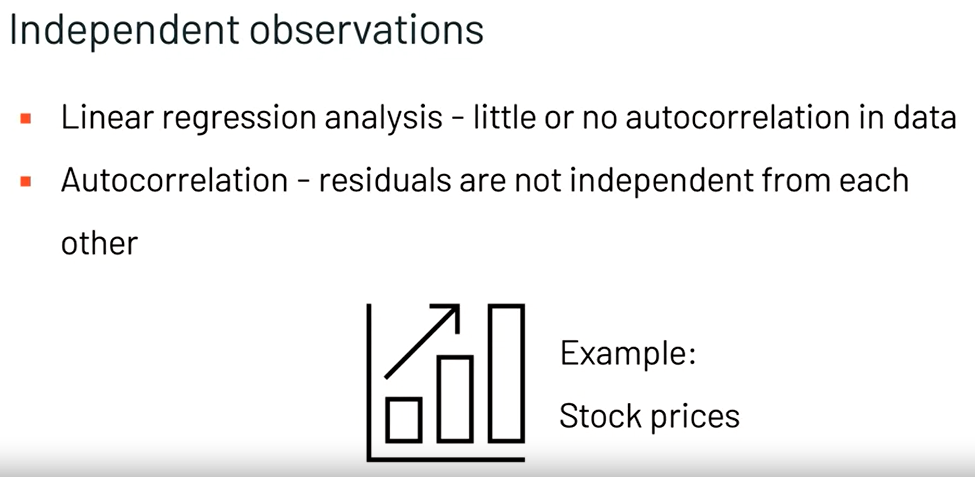
In this video, we'll describe some of the assumptions and limitations that are associated with the linear regression model. [Bu videoda, lineer regresyon modeliyle ilişkili bazı varsayımları ve sınırlamaları açıklayacağız.]



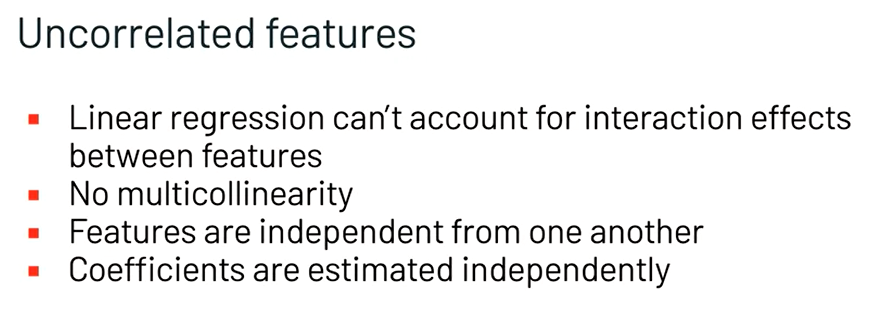
In a linear regression model, there are five key assumptions that we need to make sure are true before building up regression model. [Doğrusal bir regresyon modelinde, regresyon modelini oluşturmadan önce doğru olduğundan emin olmamız gereken beş temel varsayım vardır.] First, the linear regression needs the relationship between the independent and dependent variables to be linear. [İlk olarak, doğrusal regresyon, bağımsız ve bağımlı değişkenler arasındaki ilişkinin doğrusal olmasını gerektirir.] In other words, we want to see that there's actually a relationship and it's not just random. [Başka bir deyişle, aslında bir ilişki olduğunu ve bunun rastgele olmadığını görmek istiyoruz.] In the graph on the left, we see that there is clearly a linear relationship, and we can fit a line through that. [Soldaki grafikte açıkça doğrusal bir ilişki olduğunu görüyoruz ve bunun üzerinden bir doğru sığdırabiliriz.] Whereas the graph on the right, there doesn't seem to be any linear relationship between the independent and dependent variables or each feature and why. [Sağdaki grafik ise, bağımsız ve bağımlı değişkenler veya her bir özellik ve neden arasında doğrusal bir ilişki yok gibi görünüyor.] It's also important to check for outliers when we are checking for this relationship, because linear regression is sensitive to outlier effects. [Doğrusal regresyon, aykırı değer etkilerine duyarlı olduğundan, bu ilişkiyi kontrol ederken aykırı değerleri kontrol etmek de önemlidir.] The best way that we can test the linearity assumption is as we show on the slide with the scatter plots. [Doğrusallık varsayımını test etmenin en iyi yolu, dağılım grafikleriyle slaytta gösterdiğimiz gibidir.]



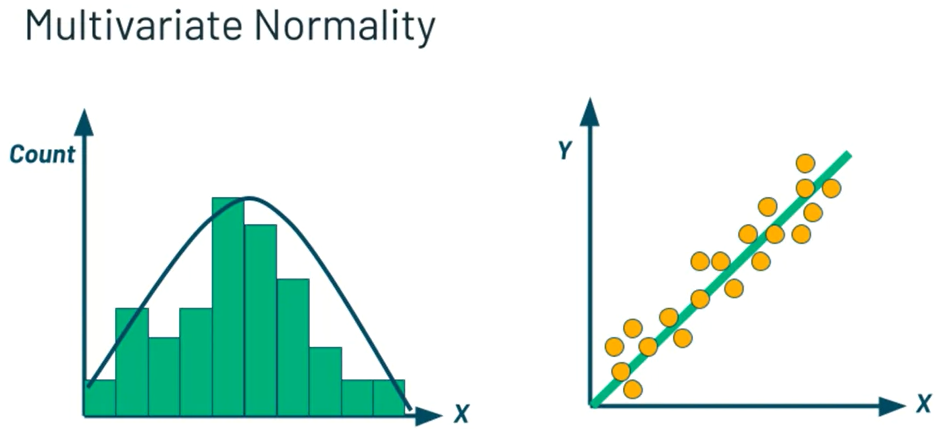
The next assumption of linear regression models is called homoscedasticity. [Doğrusal regresyon modellerinin bir sonraki varsayımına homoskedastisite denir.] Try saying that three times first. [Bunu önce üç kez söylemeyi dene.] This means that the value of the residuals is not dependent on the feature values. [Bu, artıkların değerinin özellik değerlerine bağlı olmadığı anlamına gelir.] We see the chart on the left shows a heteroscedastic relationship. [Soldaki grafiğin heteroskedastik bir ilişki gösterdiğini görüyoruz.] The green dots here indicate or represent the residuals, while the line in the middle shows a regression fitted to them. [Buradaki yeşil noktalar artıkları gösterir veya temsil eder, ortadaki çizgi ise onlara uyan bir gerilemeyi gösterir.] We see that as our independent variable X increases, the residuals also get larger. [Bağımsız değişkenimiz X arttıkça artıkların da büyüdüğünü görüyoruz.] Our model does better at predicting smaller values for X than it does for predicting larger values of X. [Modelimiz, X için daha küçük değerleri tahmin etmede, X'in daha büyük değerlerini tahmin etmekten daha iyidir.] This means that it is not homoscedastic and we can't use a linear regression. [Bu, homoskedastik olmadığı ve doğrusal bir regresyon kullanamayacağımız anlamına gelir.] When we see the chart on the right, we see that the residuals stay the same throughout different values of X, which means it's homoscedastic. [Sağdaki grafiği gördüğümüzde, X'in farklı değerleri boyunca artıkların aynı kaldığını görüyoruz, bu da homoskedastik olduğu anlamına geliyor.]



Another assumption of linear regression is that there's little or no autocorrelation in the data, meaning that the observations are independent of each other. [Doğrusal regresyonun bir başka varsayımı, verilerde çok az otokorelasyon olması veya hiç olmamasıdır, bu da gözlemlerin birbirinden bağımsız olduğu anlamına gelir.] Autocorrelation occurs when the residuals are not independent from each other. [Artıklar birbirinden bağımsız olmadığında otokorelasyon oluşur.] For instance, this typically occurs with stock prices, where the price is not independent from the previous price. [Örneğin, bu tipik olarak, fiyatın önceki fiyattan bağımsız olmadığı hisse senedi fiyatlarında ortaya çıkar.]



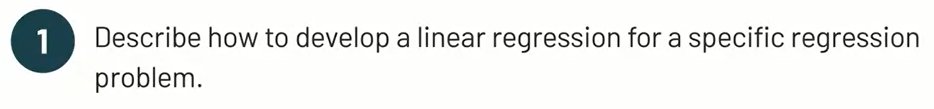
Another assumption for linear regression is that the features are uncorrelated. [Doğrusal regresyon için başka bir varsayım, özelliklerin ilişkisiz olmasıdır.] The linear regression model applies coefficient estimates independently from one another. [Doğrusal regresyon modeli, katsayı tahminlerini birbirinden bağımsız olarak uygular.] It can't account for interaction effects or correlation between features. [Özellikler arasındaki etkileşim etkilerini veya korelasyonu açıklayamaz.] This is a more complex multi-feature kind of relationship. [Bu, daha karmaşık, çok özellikli bir ilişki türüdür.] Linear regression assumes that there's little to none of this type of multicollinearity. [Doğrusal regresyon, bu tür çoklu bağlantının çok az olduğunu veya hiç olmadığını varsayar.] In other words, all the features are independent from one another, and the coefficients are estimated independently by the model. [Başka bir deyişle, tüm özellikler birbirinden bağımsızdır ve katsayılar model tarafından bağımsız olarak tahmin edilir.] We can test for multicollinearity with three central criteria. [Çoklu doğrusallığı üç merkezi kriterle test edebiliriz.] One is doing a correlation matrix. [Biri bir korelasyon matrisi yapıyor.] Another is called tolerance, where the tolerance measures the influence of one independent variable on all of the other independent variables. [Bir diğerine tolerans denir; burada tolerans, bir bağımsız değişkenin diğer tüm bağımsız değişkenler üzerindeki etkisini ölçer.] There's also something called the Variance Inflation Factor or VIF. [Ayrıca Varyans Enflasyon Faktörü veya VIF adı verilen bir şey var.] That's another indication that multicollinearity may be present. [Bu, çoklu bağlantının mevcut olabileceğinin bir başka göstergesidir.]



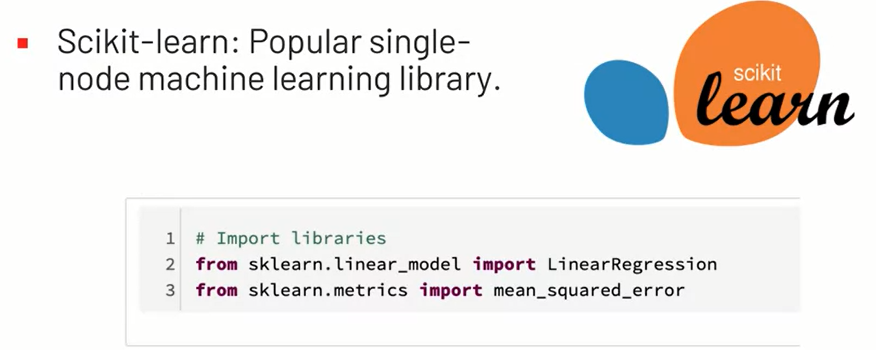
Another assumption for linear regression is that all variables need to be multivariate normal, which means that each feature itself has a normal distribution. [Doğrusal regresyon için başka bir varsayım, tüm değişkenlerin çok değişkenli normal olması gerektiğidir; bu, her özelliğin kendisinin normal bir dağılıma sahip olduğu anlamına gelir.] We can check this with the histogram or a Q-Q plot. [Bunu histogram veya Q-Q grafiği ile kontrol edebiliriz.] Normality can also be checked with a goodness of fit test. [Normallik, uyum iyiliği testi ile de kontrol edilebilir.] If we see that the features are not distributed normally, we can fix that using a nonlinear transformation, like a log transformation, which might fix the issue. [Özelliklerin normal bir şekilde dağıtılmadığını görürsek, log dönüşümü gibi doğrusal olmayan bir dönüşüm kullanarak sorunu çözebiliriz.] You've now learned the basics of the linear regression algorithm, and then the key assumptions going into it. [Artık lineer regresyon algoritmasının temellerini ve ardından buna ilişkin temel varsayımları öğrendiniz.] In the next video, we'll walk through an example of how you would apply linear regression modeling to a real-world problem. [Bir sonraki videoda, gerçek dünya problemine lineer regresyon modellemesini nasıl uygulayacağınıza dair bir örnek üzerinden gideceğiz.]

# **Applying Linear Regression**

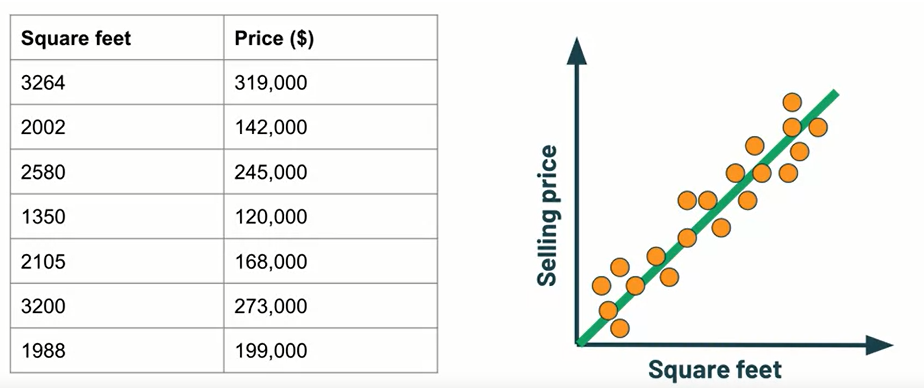
In the previous videos, we learned what linear regression is and the key assumptions for applying linear regression to data. [Önceki videolarda, doğrusal regresyonun ne olduğunu ve verilere doğrusal regresyon uygulamak için temel varsayımları öğrendik.] In this video we'll go through an example of how to develop a linear regression model for a specific problem or question. [Bu videoda, belirli bir problem veya soru için lineer regresyon modelinin nasıl geliştirileceğine dair bir örnek üzerinden gideceğiz.]



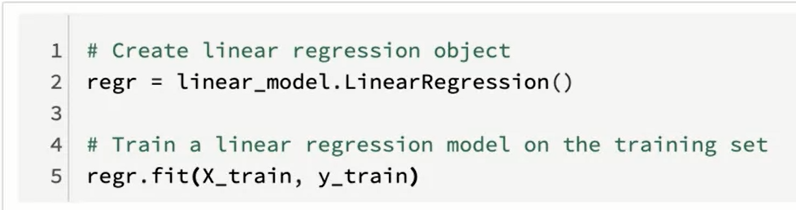
In our example, we want to predict the selling price of a house based solely on the house's square footage. [Örneğimizde, bir evin satış fiyatını yalnızca evin metrekaresine göre tahmin etmek istiyoruz.] We will use a data set that consists of a number of past sales and includes the square footage of the house and the price it's sold for. [Bir dizi geçmiş satıştan oluşan ve evin metrekaresini ve satıldığı fiyatı içeren bir veri seti kullanacağız.] This data set will be used to train a linear regression model so that when we have a new house and we know the square footage of that house, we can predict hopefully with some degree of accuracy, how much that house will sell for. [Bu veri seti lineer bir regresyon modeli eğitmek için kullanılacak, böylece yeni bir evimiz olduğunda ve o evin kare görüntülerini bildiğimizde, o evin ne kadara satacağını bir dereceye kadar doğrulukla tahmin edebiliriz.]



We can use Scikit-learn, a popular machine learning library for many machine learning applications. [Birçok makine öğrenimi uygulaması için popüler bir makine öğrenimi kitaplığı olan Scikit-learn'i kullanabiliriz.] In this case will import the linear regression model and the mean squared error metric as you can see in the code block on the slide. [Bu durumda, slayttaki kod bloğunda görebileceğiniz gibi, doğrusal regresyon modelini ve ortalama kare hata metriğini içe aktaracaktır.]



The first step is to identify what our input features and our output label target are. [İlk adım, girdi özelliklerimizin ve çıktı etiketi hedefimizin ne olduğunu belirlemektir.] In this case, the input is the square footage and the output is the price. [Bu durumda girdi metrekare, çıktı ise fiyattır.] We can create a scatter plot with the regression line to determine if there appears to be a linear relationship between the input and the output, and as we see on this simplified graph, there is obviously a very clear linear relationship between those two. [Girdi ve çıktı arasında doğrusal bir ilişki olup olmadığını belirlemek için regresyon çizgisiyle bir dağılım grafiği oluşturabiliriz ve bu basitleştirilmiş grafikte gördüğümüz gibi, açıkçası bu ikisi arasında çok net bir doğrusal ilişki vardır.]



Next, we want to fit a linear regression model on our data. [Ardından, verilerimize doğrusal bir regresyon modeli uydurmak istiyoruz.] This is also called training the model where we teach the model based on our input data, how to develop the function that will identify the relationship between the input and the output. [Buna, girdi verilerimize dayanarak modeli öğrettiğimiz, girdi ile çıktı arasındaki ilişkiyi belirleyecek fonksiyonun nasıl geliştirileceğini öğrettiğimiz modelin eğitimi de denir.] And as you see in the code block on the side, we do this by creating a linear regression object, calling the linear regression model. [Ve yandaki kod bloğunda gördüğünüz gibi bunu lineer bir regresyon nesnesi oluşturarak, lineer regresyon modelini çağırarak yapıyoruz.] And then we fit it on our X-train and our y- train training data set. [Sonra onu X-treni ve y-treni eğitim veri setimize sığdırırız.]



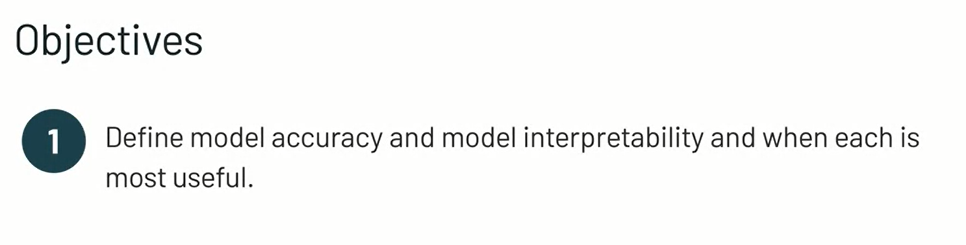
Once we fit or trained the model on our training data set, we can view the coefficients that the model identifies. [Modeli eğitim veri setimize sığdırdığımızda veya eğittiğimizde, modelin tanımladığı katsayıları görebiliriz.] The coefficients pertained to each of the input features in are really simple example here, there's only one feature square footage, so we would only have that one single coefficient. [Girdi özelliklerinin her birine ilişkin katsayılar burada gerçekten basit bir örnektir, yalnızca bir özellik kare görüntüsü vardır, bu nedenle yalnızca tek bir katsayıya sahip oluruz.] In this case, we would get a positive number for this coefficient, which tells us that as square footage increases, the selling price also increases. [Bu durumda, bize metrekare arttıkça satış fiyatının da arttığını söyleyen bu katsayı için pozitif bir sayı alırız.] We'll go into more detail later on the coefficients and how we can interpret them. [Daha sonra katsayılar ve bunları nasıl yorumlayabileceğimiz üzerinde daha fazla ayrıntıya gireceğiz.]

# **Linear Regression Lab 1**

Next, you'll complete a lab on linear regression using Databricks.

To do so, log in to your Databricks Community Edition account and locate your DSFDA folder. Then, open Module 5, then click on "5.1.1 Lab - Linear Regression 1" to open your notebook.

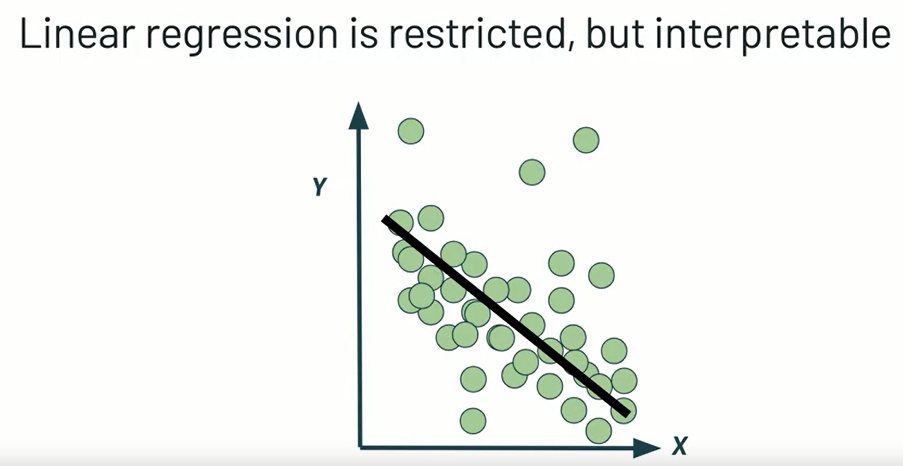
# **Accuracy and Interpretability**



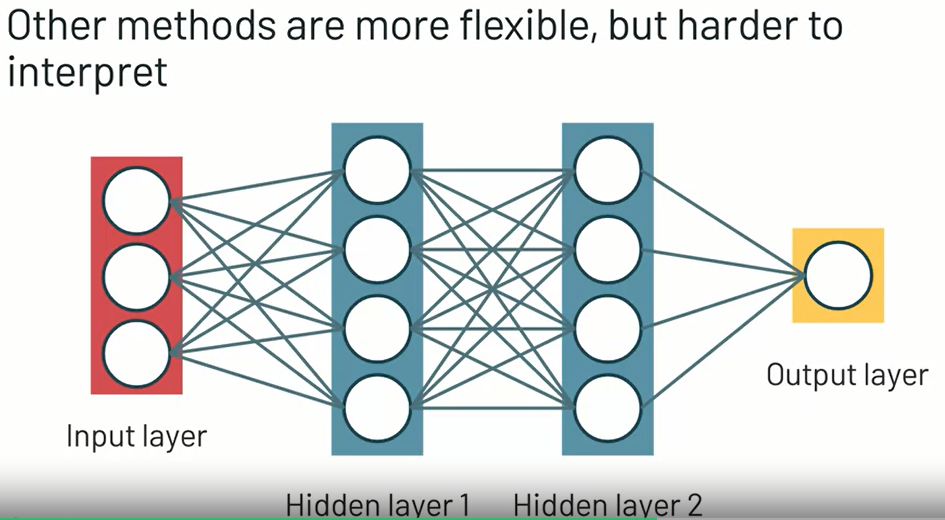
Now that you've learned a bit about how to actually build a linear regression model for a real problem, we're going to talk about some specifics about Machine Learning models in general. [Artık gerçek bir problem için bir lineer regresyon modelinin nasıl oluşturulacağını biraz öğrendiğinize göre, genel olarak Makine Öğrenimi modelleri hakkında bazı özelliklerden bahsedeceğiz.] In this video, we'll talk about model accuracy and model interpretability and understand the trade-offs that come with those two things. [Bu videoda, model doğruluğu ve model yorumlanabilirliği hakkında konuşacağız ve bu iki şeyle birlikte gelen ödünleşimleri anlayacağız.]



Generally, machine-learning models can fall anywhere on a spectrum from very restrictive or interpretable, to very flexible or accurate. [Genel olarak, makine öğrenimi modelleri, çok kısıtlayıcı veya yorumlanabilirden çok esnek veya doğruya kadar bir yelpazede herhangi bir yere düşebilir.] Models will allow us to either understand their predictions and the features that they select, or have more accuracy. [Modeller ya onların tahminlerini ve seçtikleri özellikleri anlamamızı ya da daha fazla doğruluk elde etmemizi sağlayacaktır.] There's usually a trade-off where you have to prioritize one or the other. [Genellikle birine veya diğerine öncelik vermeniz gereken bir takas vardır.] Generally, the more complex models, like the one you see on the right, that are more flexible, will perform better and have higher accuracy but they're difficult to interpret. [Genel olarak, sağda gördüğünüz gibi daha karmaşık modeller daha esnektir, daha iyi performans gösterir ve daha yüksek doğruluğa sahiptir ancak yorumlanmaları zordur.] That means that it's difficult to see what the model is actually doing in terms of features that are important or how features are correlated with the target or the outcome. [Bu, önemli olan özellikler açısından modelin gerçekte ne yaptığını veya özelliklerin hedef veya sonuçla nasıl ilişkilendirildiğini görmenin zor olduğu anlamına gelir.] But on the other hand, on the far left, more restrictive models are more interpretable because you can see exactly what the coefficients are, for example, of a feature, and you can understand how the models ranking different features in terms of importance for predicting the outcome. [Ancak öte yandan, en solda, daha kısıtlayıcı modeller daha yorumlanabilir çünkü örneğin bir özelliğin katsayılarının tam olarak ne olduğunu görebilir ve modellerin tahmin için önem açısından farklı özellikleri nasıl sıraladığını anlayabilirsiniz. sonuç.] But these more restrictive models generally might not be as accurate or perform as well as the more flexible models. [Ancak bu daha kısıtlayıcı modeller genellikle daha esnek modeller kadar doğru veya iyi performans göstermeyebilir.] Choosing which type that you want to use, involves both an understanding of the trade-off, and knowing what the goal of the model is, so what your actual data science objective is for your project. [Hangi türü kullanmak istediğinizi seçmek, hem ödünleşimin anlaşılmasını hem de modelin amacının ne olduğunu, dolayısıyla projeniz için gerçek veri bilimi hedefinizin ne olduğunu bilmeyi içerir.] If your goal is more to determine which feature or features are most important for determining an outcome, you would probably use a more restrictive model like a linear regression, because it will give you more interpretability. [Amacınız bir sonucu belirlemek için hangi özelliğin veya özelliklerin daha önemli olduğunu belirlemekse, muhtemelen size daha fazla yorumlanabilirlik sağlayacağı için doğrusal regresyon gibi daha kısıtlayıcı bir model kullanırsınız.] On the other hand, if accuracy is your main goal, and you really don't care that much about understanding the feature importances or other details, you just want to get the best predictions possible, then you would try to have a more flexible model that will give you that better accuracy but less interpretability. [Öte yandan, ana hedefiniz doğruluk ise ve özelliklerin önemini veya diğer ayrıntıları anlamakla gerçekten o kadar ilgilenmiyorsanız, yalnızca mümkün olan en iyi tahminleri elde etmek istiyorsanız, o zaman daha esnek bir modele sahip olmaya çalışırsınız. bu size daha iyi doğruluk, ancak daha az yorumlanabilirlik sağlayacaktır.]

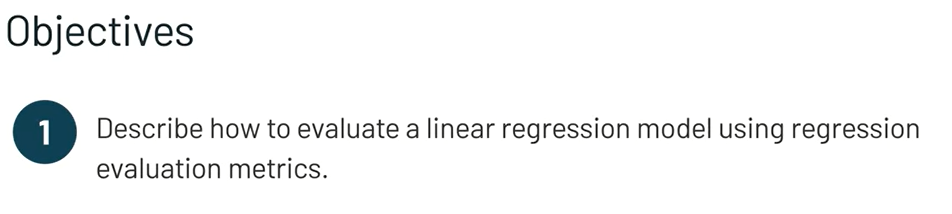


Linear regression is a very restricted and highly interpretable model, so it falls on the left side of that spectrum that we just saw. [Doğrusal regresyon çok kısıtlı ve oldukça yorumlanabilir bir modeldir, dolayısıyla az önce gördüğümüz spektrumun sol tarafında yer alır.] Linear regression allows us to look at the model's predictions, and the coefficients for the features in real terms. [Doğrusal regresyon, modelin tahminlerine ve özelliklerin katsayılarına gerçek terimlerle bakmamızı sağlar.] So if we're looking at sale prices and square footage, we can actually see which square footages where most related to certain sales prices or if we have more features than just square footage, we can see how the model ranks these different features. [Yani satış fiyatlarına ve metrekarelere bakarsak, aslında hangi metrekarelerin belirli satış fiyatlarıyla en çok ilişkili olduğunu görebiliriz veya sadece metrekareden daha fazla özelliğe sahipsek, modelin bu farklı özellikleri nasıl sıraladığını görebiliriz.] It's easily understandable. [Kolay anlaşılır.] We could see that in actual square footage terms or dollar amounts. [Bunu gerçek metrekare cinsinden veya dolar miktarlarında görebiliriz.] The drawback to a model like this though, is that the predictive power is somewhat limited. [Böyle bir modelin dezavantajı, tahmin gücünün bir şekilde sınırlı olmasıdır.] That means that the accuracy might not be as good as you could get with a different type of model. [Bu, doğruluğun farklı bir model türüyle elde edebileceğiniz kadar iyi olmayabileceği anlamına gelir.] If you're more concerned about how the model is going to perform on unseen new data, you might choose a different model. [Modelin görünmeyen yeni veriler üzerinde nasıl performans göstereceğiyle daha fazla ilgileniyorsanız, farklı bir model seçebilirsiniz.] One thing that's a common approach, is to start with a linear regression model, and that can give you a good idea of your features and the coefficients. [Yaygın bir yaklaşım olan bir şey, lineer bir regresyon modeli ile başlamaktır ve bu size özellikleriniz ve katsayılarınız hakkında iyi bir fikir verebilir.] It might not give you the accuracy you're looking for, but then you can go on and try a different type of model with the same data, and that way you can have both the understanding and the eventual higher accuracy from trying a different type of model. [Size aradığınız doğruluğu vermeyebilir, ancak daha sonra aynı verilerle farklı bir model türünü deneyebilir ve bu şekilde farklı bir tür denemekten hem anlayışa hem de nihai olarak daha yüksek doğruluğa sahip olabilirsiniz. modeli.]

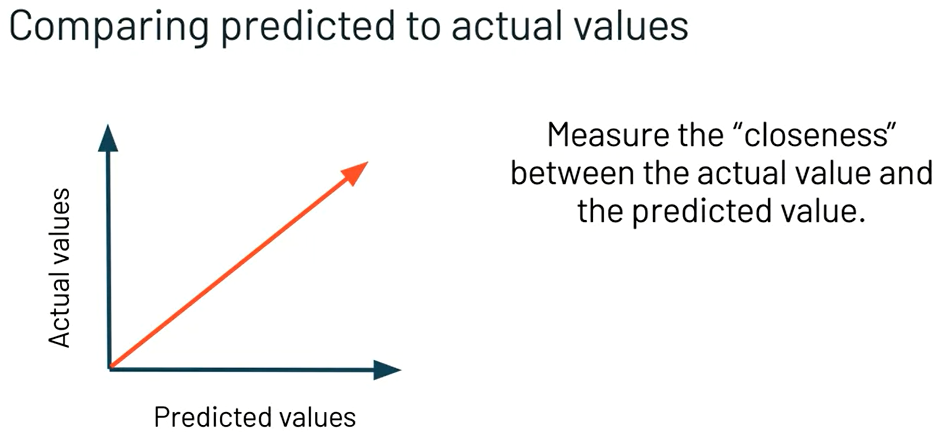


There's other machine learning methods that are a lot more flexible but harder to interpret. [Çok daha esnek ancak yorumlanması daha zor olan başka makine öğrenimi yöntemleri de var.] A deep learning neural network model is the extreme example of this, where the model is very flexible and is often very accurate and powerful, making good predictions, and dealing with new unseen data, but it's very complex and very difficult, if not impossible, to interpret anything about how the model is performing its calculations. [Bir derin öğrenme sinir ağı modeli, modelin çok esnek olduğu ve genellikle çok doğru ve güçlü olduğu, iyi tahminlerde bulunduğu ve yeni görünmeyen verilerle uğraştığı, ancak imkansız değilse de çok karmaşık ve çok zor olduğu bunun en uç örneğidir. modelin hesaplamalarını nasıl yaptığı hakkında herhangi bir şeyi yorumlamak için.] What we're looking at on the slide, is a basic illustration of a neural network, where we have an input layer on the left, which is our features, and the output layer on the right, which is the target. [Slaytta baktığımız şey, bir sinir ağının temel bir gösterimidir, burada solda özelliklerimiz olan bir girdi katmanımız ve sağda hedef olan çıktı katmanımız vardır.] But the intermediate steps between these two, are these jumble of multiple connections between the layers, so we have numerous hidden layers between the input and the output, and this is where the model becomes what's known as a black box model. [Ancak bu ikisi arasındaki ara adımlar, katmanlar arasındaki bu karmakarışık çoklu bağlantılardır, yani girdi ve çıktı arasında çok sayıda gizli katmana sahibiz ve bu, modelin kara kutu modeli olarak bilinen hale geldiği yerdir.] Because we can't really examine what each of these hidden layers is doing, or how it's making its decisions; so really low interpretability but high-accuracy. [Çünkü bu gizli katmanların her birinin ne yaptığını veya kararlarını nasıl verdiğini gerçekten inceleyemiyoruz; çok düşük yorumlanabilirlik ama yüksek doğruluk.] If you understand that trade off and you decide that fits your objective, then this might be a model you would try for your problem. [Bu takası anlarsanız ve amacınıza uygun olduğuna karar verirseniz, bu, probleminiz için deneyeceğiniz bir model olabilir.] Deep learning and neural networks are beyond the scope of this course, but they're just a common example of that opposite end of the accuracy and interpretability trade-off. [Derin öğrenme ve sinir ağları bu kursun kapsamı dışındadır, ancak bunlar yalnızca doğruluk ve yorumlanabilirlik değiş tokuşunun karşıt ucunun yaygın bir örneğidir.] Now that you have an understanding of the interpretability versus accuracy, or restrictive versus flexible trade-off with Machine Learning models, we'll go into the next video, where we'll learn about different regression metrics, how to evaluate regression models, and understand the different types of metrics you can use to evaluate your model. [Artık, Makine Öğrenimi modelleriyle yorumlanabilirliğe karşı doğruluk veya kısıtlayıcıya karşı esnek değiş tokuşa ilişkin bir anlayışa sahip olduğunuza göre, farklı regresyon metriklerini, regresyon modellerinin nasıl değerlendirileceğini ve nasıl değerlendirileceğini öğreneceğimiz bir sonraki videoya geçeceğiz. modelinizi değerlendirmek için kullanabileceğiniz farklı metrik türlerini anlayın.]

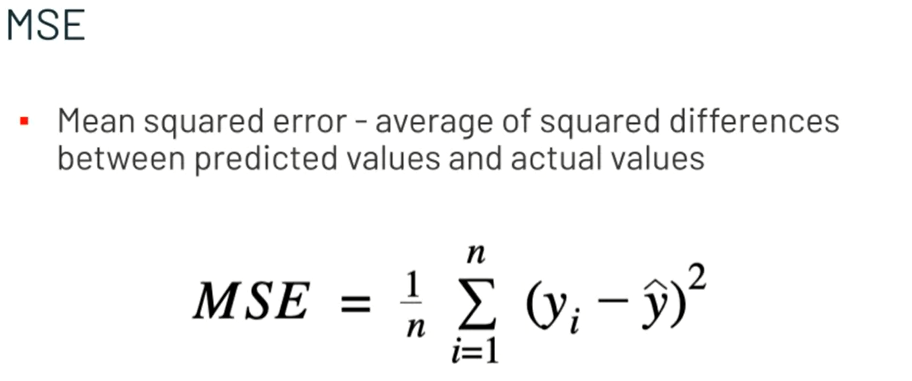
# **Regression Evaluation**



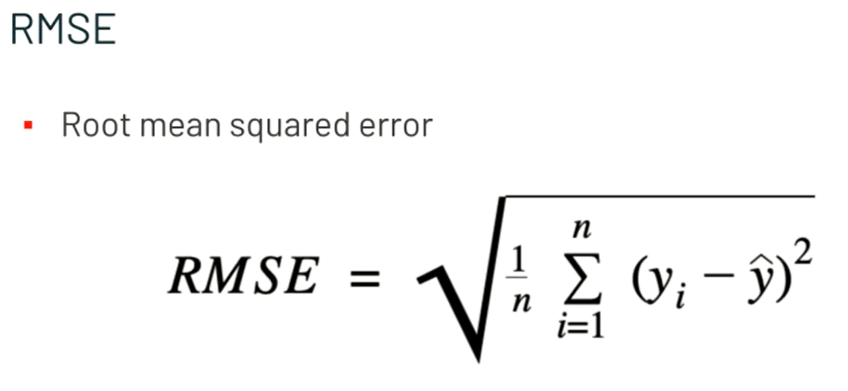
In this video, we're going to be talking about some of the techniques you can use to evaluate a regression model and understand how well it's performing. [Bu videoda, bir regresyon modelini değerlendirmek ve ne kadar iyi performans gösterdiğini anlamak için kullanabileceğiniz bazı tekniklerden bahsedeceğiz.] We'll take a look at some of the most commonly used regression evaluation metrics and understand what they tell you about a linear regression model. [En sık kullanılan regresyon değerlendirme metriklerinden bazılarına göz atacağız ve size lineer regresyon modeli hakkında ne söylediklerini anlayacağız.]



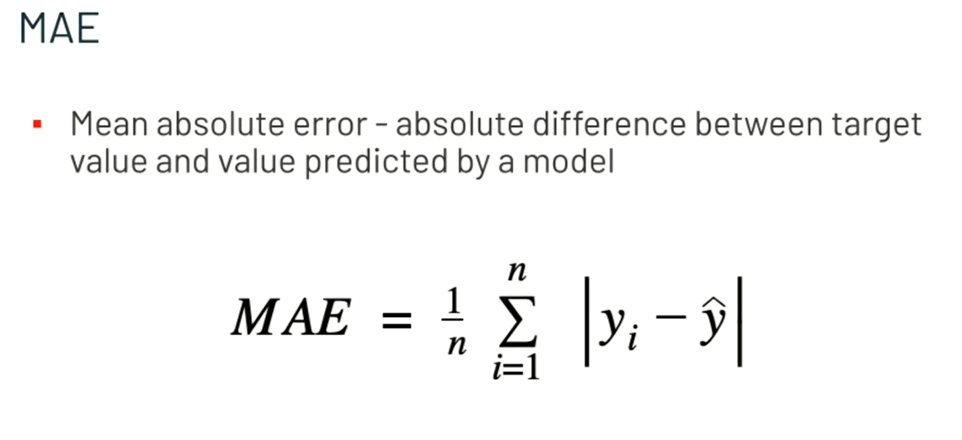
When we fit a linear regression model to data, we want to compare the predicted outcome to the actual values for that outcome. [Verilere doğrusal bir regresyon modeli uydurduğumuzda, tahmin edilen sonucu bu sonuç için gerçek değerlerle karşılaştırmak istiyoruz.] Going back to our house prices example, for each data point, which indicates a square footage, we would compare the price that the model predicted to the actual sales price. [Ev fiyatları örneğimize geri dönersek, bir metrekareyi belirten her veri noktası için, modelin öngördüğü fiyatı gerçek satış fiyatıyla karşılaştırırdık.] We talked earlier about this concept of minimizing the residuals. [Kalıntıları en aza indirme kavramından daha önce bahsetmiştik.] This means that a good model will have little distance between the predicted and actual values so the residuals will be small and that would indicate that the actual values and the predicted values are quite close. [Bu, iyi bir modelin tahmin edilen ve gerçek değerler arasında çok az mesafeye sahip olacağı ve dolayısıyla artıkların küçük olacağı anlamına gelir ve bu, gerçek değerler ile tahmin edilen değerlerin oldukça yakın olduğunu gösterir.] There's various ways to measure how close the model got to the actual values. [Modelin gerçek değerlere ne kadar yaklaştığını ölçmenin çeşitli yolları vardır.]



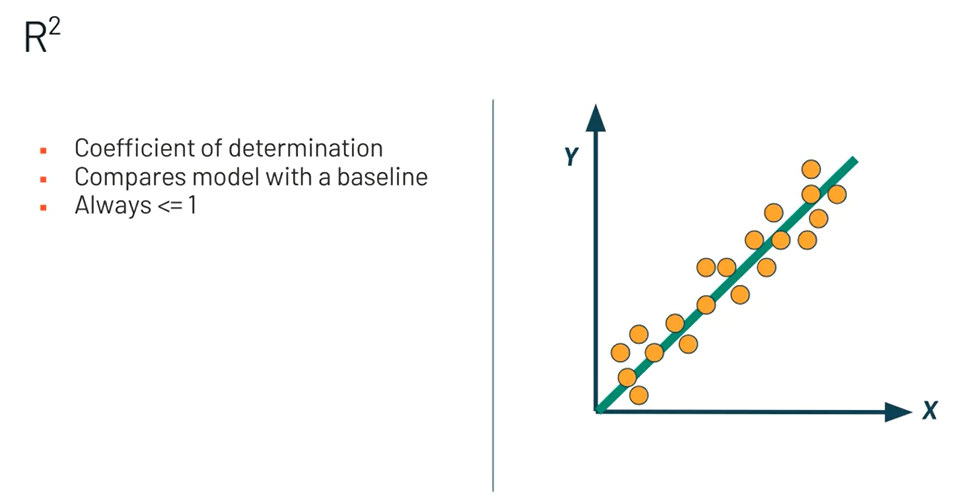
One such metric is MSE and this is very commonly used and it is the mean squared error. [Böyle bir metrik MSE'dir ve bu çok yaygın olarak kullanılır ve ortalama kare hatasıdır.] This measures the average of the squared differences between the predicted values and the actual values for the entire dataset. [Bu, tüm veri kümesi için tahmin edilen değerler ile gerçek değerler arasındaki kare farklarının ortalamasını ölçer.] Because it squares the difference between the predicted and actual values, in other words, it squares the residuals, the MSE actually penalizes larger errors. [Tahmin edilen ve gerçek değerler arasındaki farkı karelediğinden, diğer bir deyişle artıkların karesini aldığından, MSE aslında daha büyük hataları cezalandırır.] If you want to train a model and make sure to penalize large errors, you could use MSE as your metric to evaluate how well the model does. [Bir modeli eğitmek ve büyük hataları cezalandırdığınızdan emin olmak istiyorsanız, modelin ne kadar iyi olduğunu değerlendirmek için ölçümünüz olarak MSE'yi kullanabilirsiniz.] The possible values for MSE can range from zero to infinity and it's indifferent to the direction of the errors. [MSE için olası değerler sıfırdan sonsuza kadar değişebilir ve hataların yönüne kayıtsızdır.] Whether they're greater than or lesser than, if the residual is positive or negative, the MSE is going to be indifferent to that direction and lower values are better. [Daha büyük veya daha küçük olsunlar, kalıntı pozitif veya negatifse, MSE bu yöne kayıtsız kalacaktır ve daha düşük değerler daha iyidir.] An ideal MSE would be close to zero. [İdeal bir MSE sıfıra yakın olacaktır.] Now, I don't say an ideal MSE is equal to zero because you are evaluating this MSE on a training data set and so if the MSE is zero, which means there are no errors at all, the model predicted with 100 percent accuracy, all of the data points. [Şimdi, ideal bir MSE'nin sıfıra eşit olduğunu söylemiyorum çünkü bu MSE'yi bir eğitim veri setinde değerlendiriyorsunuz ve bu nedenle MSE sıfırsa, yani hiç hata yok, model yüzde 100 doğrulukla tahmin edildi, tüm veri noktaları.] It means that your model is probably overfitting to the training data. [Bu, modelinizin muhtemelen eğitim verilerine fazla uyduğu anlamına gelir.]



MSE is difficult to interpret in real terms though, because it gives you a relative number, you can't really evaluate what that means in terms of, for example, square footage or dollar amounts. [MSE'nin gerçek terimlerle yorumlanması zordur, çünkü size göreceli bir sayı verir, bunun ne anlama geldiğini, örneğin, kare görüntüleri veya dolar miktarları açısından gerçekten değerlendiremezsiniz.] That's why the root mean squared error or RMSE is often used and by taking the root of the MSE, it ends up having the same units as the dependent variable. [Bu nedenle, kök ortalama kare hatası veya RMSE sıklıkla kullanılır ve MSE'nin kökü alınarak, bağımlı değişkenle aynı birimlere sahip olur.] In our example again, the units for the RMSE would be actual dollar amounts. [Yine örneğimizde, RMSE için birimler gerçek dolar tutarları olacaktır.] You would see how far off your model is on average in dollars for predicting a sales price. [Bir satış fiyatını tahmin etmek için modelinizin ortalama olarak dolar cinsinden ne kadar uzakta olduğunu görürsünüz.] This means that there's no absolute good or bad threshold for an RMSE value but you can define it based on your dependent variable. [Bu, bir RMSE değeri için mutlak iyi veya kötü bir eşik olmadığı, ancak bunu bağımlı değişkeninize göre tanımlayabileceğiniz anlamına gelir.] If we're talking about how sales prices maybe being off by a few thousand dollars or even a few tens of thousands of dollars might not be too bad. [Satış fiyatlarının nasıl birkaç bin dolar, hatta birkaç on binlerce dolar düştüğünden bahsediyorsak, o kadar da kötü olmayabilir.] But if we're talking about something on a different scale, like the cost of groceries, thousands of dollars off would mean your model is pretty terrible. [Ancak, bakkaliye maliyeti gibi farklı bir ölçekten bahsediyorsak, binlerce dolar indirim, modelinizin oldukça korkunç olduğu anlamına gelir.] This is really dependent on what you're model is being used for. [Bu gerçekten modelinizin ne için kullanıldığına bağlıdır.] The RMSE also penalizes errors since it squares the residual values before taking the square root. [RMSE, karekök almadan önce artık değerlerin karesini aldığı için hataları da cezalandırır.]

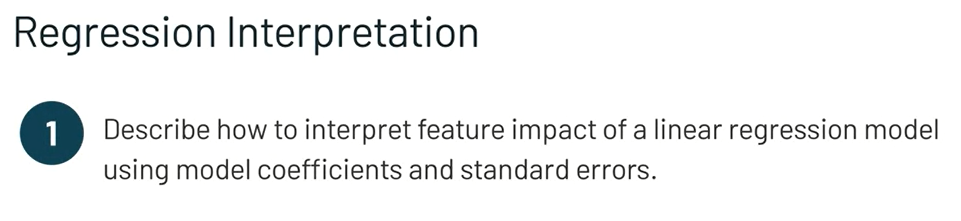


Mean absolute error, or MAE, is the absolute difference between the target value and the value predicted by the model. [Ortalama mutlak hata veya MAE, hedef değer ile model tarafından tahmin edilen değer arasındaki mutlak farktır.] The MAE is more robust to outliers and it doesn't penalize the errors as extremely as MSE or RMSE. [MAE aykırı değerlere karşı daha dayanıklıdır ve hataları MSE veya RMSE kadar aşırı derecede cezalandırmaz.] MAE is a linear score, which means that all of the individual differences are weighted equally. [MAE, tüm bireysel farklılıkların eşit ağırlıkta olduğu anlamına gelen doğrusal bir puandır.] It's not suitable for applications where you would want to pay more attention to the outliers, since it's robust to outliers. [Aykırı değerlere karşı dayanıklı olduğundan, aykırı değerlere daha fazla dikkat etmek isteyeceğiniz uygulamalar için uygun değildir.] Similar to MSE, the values for MAE can range from zero to infinity and it's also indifferent to the direction of the residuals. [MSE'ye benzer şekilde, MAE değerleri sıfırdan sonsuza kadar değişebilir ve aynı zamanda artıkların yönüne de kayıtsızdır.] Also like MSE, the lower value, the better for MAE. [Ayrıca MSE gibi, değer ne kadar düşükse MAE için o kadar iyidir.]

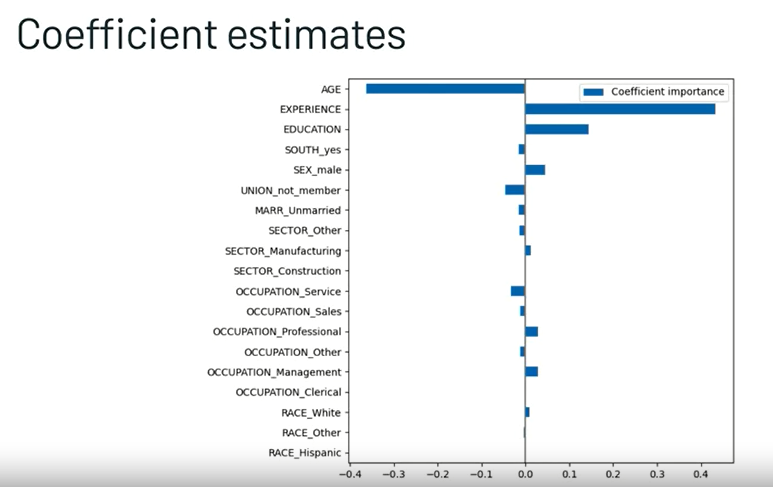


R-squared is very commonly used and talked about and it's also called the coefficient of determination. [R-kare çok yaygın olarak kullanılır ve hakkında konuşulur ve aynı zamanda belirleme katsayısı olarak da adlandırılır.] It's another metric that we use to evaluate the performance of a regression model. [Bir regresyon modelinin performansını değerlendirmek için kullandığımız başka bir ölçümdür.] The R-squared metric helps us to compare our current model with a constant baseline. [R-kare metriği, mevcut modelimizi sabit bir temel ile karşılaştırmamıza yardımcı olur.] It tells us how much our model is better from a constant baseline. [Bize modelimizin sabit bir taban çizgisinden ne kadar daha iyi olduğunu söyler.] This baseline is chosen by taking the mean of the data and drawing a line at the mean. [Bu taban çizgisi, verilerin ortalaması alınarak ve ortalamada bir çizgi çizilerek seçilir.] R-squared is a scale-free score and that implies that it doesn't matter whether the predicted values are too large or too small. [R-kare, ölçeksiz bir puandır ve bu, tahmin edilen değerlerin çok büyük veya çok küçük olmasının önemli olmadığı anlamına gelir.] The R-squared will always be less than or equal to one. [R-kare her zaman bire eşit veya küçük olacaktır.] Another way to think about R-Squared is that the number indicates the percentage of the response variable variation that's explained by the model. [R-Kare hakkında düşünmenin başka bir yolu, sayının model tarafından açıklanan yanıt değişkeni varyasyonunun yüzdesini göstermesidir.] It explains how much of the changes in the dependent variable are explained by a model fit to the independent variable. [Bağımlı değişkendeki değişikliklerin ne kadarının bağımsız değişkene uygun bir modelle açıklandığını açıklar.] Zero percent would indicate that the model explains none of that variability of the response data and an R-squared of 100 percent indicates that the model explains all of the variability. [Yüzde sıfır, modelin yanıt verilerinin bu değişkenliğinin hiçbirini açıklamadığını gösterir ve yüzde 100'lük bir R-kare, modelin tüm değişkenliği açıkladığını gösterir.] In general, the higher the R-squared value, the better the model fits your data. [Genel olarak, R-kare değeri ne kadar yüksek olursa, model verilerinize o kadar iyi uyar.] Next, you'll confirm your understanding of regression evaluation metrics by completing a brief knowledge check. [Ardından, kısa bir bilgi kontrolünü tamamlayarak regresyon değerlendirme metriklerini anladığınızı onaylayacaksınız.]

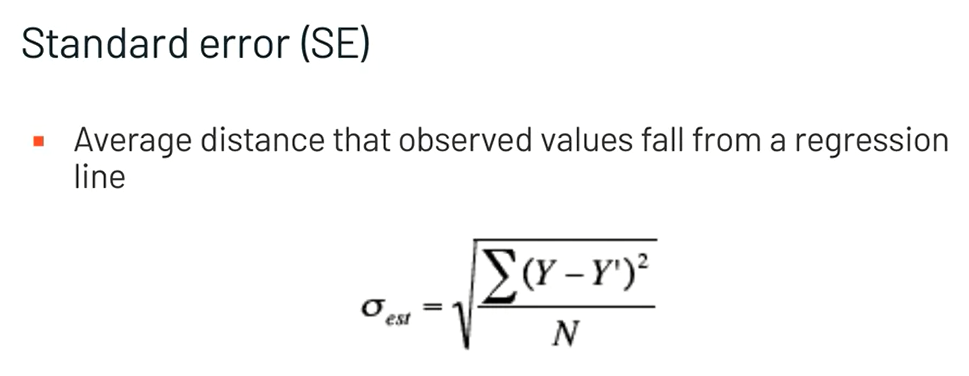
# **Regression Interpretation**



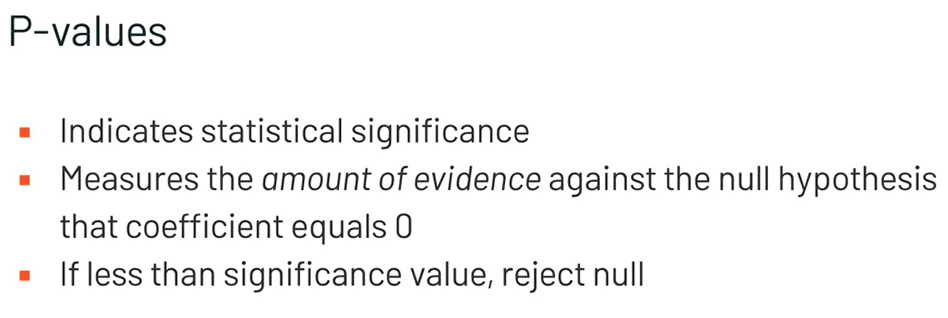
In the final video of this lesson we'll be looking at some of the ways that you can interpret and evaluate the information that a regression model gives you. [Bu dersin son videosunda, bir regresyon modelinin size verdiği bilgileri yorumlayıp değerlendirebileceğiniz bazı yollara bakacağız.] As we saw previously, we can extract certain values and information after fitting a linear regression model on some data. [Daha önce gördüğümüz gibi, bazı verilere doğrusal bir regresyon modeli yerleştirdikten sonra belirli değerleri ve bilgileri çıkarabiliriz.]



We showed how to extract the coefficients after fitting a model, there's a coefficient for each feature and these coefficients tell us how important each feature is relative to other features. [Bir model uydurduktan sonra katsayıların nasıl çıkarılacağını gösterdik, her özellik için bir katsayı var ve bu katsayılar bize her bir özelliğin diğer özelliklere göre ne kadar önemli olduğunu söylüyor.] The larger the coefficient, the more important that feature is in determining the target value. [Katsayı ne kadar büyükse, hedef değeri belirlemede o özellik o kadar önemlidir.] In other words, the more highly correlated that feature is with the target. [Başka bir deyişle, o özellik hedefle ne kadar yüksek düzeyde ilişkiliyse o kadar yüksek korelasyona sahiptir.] A positive coefficient means there's a positive correlation between the future and the target while a negative coefficient means that there's a negative correlation between the two. [Pozitif bir katsayı, gelecek ile hedef arasında pozitif bir korelasyon olduğu anlamına gelirken, negatif bir katsayı, ikisi arasında negatif bir korelasyon olduğu anlamına gelir.] It's important to note that if features have very different scales, for example, if we have one feature that's measured in millions and another only ranges from 1-3. [Unutulmamalıdır ki, özellikler çok farklı ölçeklere sahipse, örneğin milyonlarca ölçülen bir özelliğimiz varsa ve bir diğeri yalnızca 1-3 arasında değişiyorsa.] The model will wait the larger numbers more heavily and downplay or ignore the importance of the smaller ones. [Model, daha büyük sayıları daha yoğun bir şekilde bekleyecek ve daha küçük sayıların önemini küçümseyecek veya görmezden gelecektir.] And this is only because they're on different scales so we can address this issue using an important technique known as feature scaling. [Ve bunun tek nedeni, farklı ölçeklerde olmalarıdır; bu nedenle, özellik ölçekleme olarak bilinen önemli bir teknik kullanarak bu sorunu çözebiliriz.] There are several methods for performing feature scaling. [Özellik ölçekleme gerçekleştirmek için birkaç yöntem vardır.] The most common of these is standardization, where we make all of the features have a mean of zero and a standard deviation of one. [Bunlardan en yaygın olanı, tüm özelliklerin ortalamasının sıfır ve standart sapmasının bir olduğu standartlaştırmadır.] We're not really going to get much more into talking about feature scaling right now, but it's just important to note when you're thinking about deriving coefficient estimates that you need to keep in mind the scale of your different features. [Şu anda özellik ölçeklendirme hakkında daha fazla konuşmayacağız, ancak farklı özelliklerinizin ölçeğini aklınızda tutmanız gereken katsayı tahminleri türetmeyi düşündüğünüzde şunu not etmek önemlidir.]



Another concept that we should talk about is the standard error of the regression or the SE. [Bahsetmemiz gereken bir diğer kavram, regresyon veya SE'nin standart hatasıdır.] It's also known as the standard error of the estimate. [Tahminin standart hatası olarak da bilinir.] It's basically the standard deviation of the residuals. [Temelde artıkların standart sapmasıdır.] It represents the average distance that the observed values fall from the regression line. [Gözlenen değerlerin regresyon çizgisinden düştüğü ortalama mesafeyi temsil eder.] And it's convenient because it tells us how wrong the model is using units of the response variable. [Ve kullanışlıdır çünkü bize modelin yanıt değişkeninin birimlerini ne kadar yanlış kullandığını söyler.] Smaller values are generally better because it indicates that the predicted values are closer to the actual values. [Daha küçük değerler genellikle daha iyidir çünkü tahmin edilen değerlerin gerçek değerlere daha yakın olduğunu gösterir.] Sometimes the standard error of the regression can be more important or informative than just assessing the R squared. [Bazen regresyonun standart hatası, sadece R kareyi değerlendirmekten daha önemli veya bilgilendirici olabilir.] Because it tells us a little bit more about how our model is performing. [Çünkü bize modelimizin nasıl performans gösterdiği hakkında biraz daha bilgi veriyor.]



In linear regression a P value indicates whether the relationship between an independent variable and the dependent variable is statistically significant while controlling for the other variables in the model. [Doğrusal regresyonda bir P değeri, modeldeki diğer değişkenler kontrol edilirken bağımsız bir değişken ile bağımlı değişken arasındaki ilişkinin istatistiksel olarak anlamlı olup olmadığını gösterir.] The P value for each variable measures the amount of evidence against the null hypothesis that the coefficient is zero. [Her değişken için P değeri, katsayının sıfır olduğu sıfır hipotezine karşı kanıt miktarını ölçer.] If the p value is less than your predetermined significance level, which is often 0.05, you reject the null hypothesis and conclude that the parameter or the coefficient doesn't equal to zero. [p değeri önceden belirlenmiş anlamlılık düzeyinizden küçükse, ki bu genellikle 0,05'tir, boş hipotezi reddedersiniz ve parametrenin veya katsayının sıfıra eşit olmadığı sonucuna varırsınız.] In other words, you would say that changes in the independent variable are related to changes in the dependent variable so that you are saying you're observing some effect or something taking place. [Başka bir deyişle, bağımsız değişkendeki değişikliklerin bağımlı değişkendeki değişikliklerle ilgili olduğunu söylersiniz, böylece bir etki veya meydana gelen bir şey gözlemlediğinizi söylersiniz.]

# **Linear Regression Lab 2**

Next, you'll complete a lab on linear regression using Databricks.

To do so, log in to your Databricks Community Edition account and locate your DSFDA folder. Then, open Module 5, then click on "5.1.2 Lab - Linear Regression 2" to open your notebook.