ALLISON O'HAIR: --previous video, we used data in linear regression to show that if a team scores at least 135 more runs than their opponent throughout the regular season, then we predict that that team will win at least 95 games and make the playoffs. This means that we need to know how many runs a team will score, which we will show can be predicted using batting statistics, and how many runs a team will allow, which we will show can be predicted using fielding and pitching statistics.

Let's start by creating a linear regression model to predict runs scored. The Oakland A's are interested in answering the question, how does a team score more runs? They discovered that two baseball statistics were significantly more important than anything else. These were on-base percentage, or OBP, which is the percentage of time a player gets on base, including walks, and slugging percentage, or SLG, which measures how far a player gets around the bases on his turn and measures the power of a hitter.

Most teams and people in baseball focused instead on batting average, or BA, which measures how often a hitter gets on base by hitting the ball. This focuses on hits instead of walks. The Oakland A's claimed that on-base percentage was the most important. Slugging percentage was important. And batting average was overvalued. Let's see if we can use linear regression in R to verify which baseball statistics are more important to predict runs.

In the previous video, we created her dataset in R and called it Moneyball. Let's take a look at the structure of our data again. Our dataset includes many variables including runs scored, or RS, on-base percentage, OBP, slugging percentage, SLG, and batting average, BA. We want to see if we can use linear regression to predict runs scored using the three hitting statistics, on-base percentage, slugging percentage, and batting average.

So let's build a linear regression equation using the LM function to predict runs scored using the independent variables OBP, SLG, and BA. Our dataset is Moneyball. If we look at the summary of our regression equation, we can see that all of our independent variables are highly significant. And our R squared is 0.93. If we look at our coefficients, we see that the coefficient for batting average is negative. This says that all other things being equal, a team with a higher batting average will score fewer runs. But this is a bit counterintuitive.

What's going on here is a case of multi-collinearity. These three hitting statistics are highly

correlated. So it's hard to interpret our model. Let's see if we can remove batting average, the variable with the negative coefficient and the least significance, to see if we can improve the interpretability of our model.

So let's rerun our regression equation without the variable BA. If we look at the summary of our new regression equation, we can see that our variables are again highly significant. The coefficients are all positive, as we would expect them to be. And our R squared is now 0.9296, or about the same as before.

So this model is simpler with one fewer variable, but has about the same predictive ability as a three variable model. You can experiment and see that if we had taken out on-base percentage or slugging percentage instead of batting average, our R squared would have gone down more. If we look at the coefficients, we can see that the coefficient for on-base percentage is significantly higher than the coefficient for slugging percentage. Since these variables are on about the same scale, this tells us that on-base percentage is worth more than slugging percentage. So using linear regression we're able to verify the claims made in Moneyball that batting average is overvalued, on-base percentage is the most important, and slugging percentage is important.

We can create a very similar model to predict runs allowed or opponent runs. This model uses pitching statistics, opponents on-base percentage, or OOBP, and opponents slugging percentage, or OSLG. The statistics are computed in the same way as on-base percentage and slugging percentage, but use the actions of the opposing batters against our team's pitcher and fielders. Using our dataset in R, we can build a linear regression model to predict runs allowed using opponents on-base percentage and opponents slugging percentage. This is, again, a very strong model with an R squared value of 0.91. And both variables are significant.

In the next video, we'll show how we can apply these models to predict whether or not a team will make the playoffs.