**MATH 243 - Project Three**

[00:00:00.39] SPEAKER: Hi, class. In this video, we will go over project three, which is in module seven of the course. So let's first open up the template for the project. And then we will go over the Python script in Codio. So in Module Seven folder, if we scroll down under seven dash four Project Three Submission. We'll click on Requirements and Rubric. And then scroll down until you see Summary Report Word Document File link here.

[00:00:32.69] So if you click this, this will open up the template. And I have already opened this up here. And so you will notice that the project starts with, again, an introduction to be sure to provide an introduction to this project. And so there are a bunch of questions here. What is the data set that you're exploring? How will your results be used? And what type of analyzes will you be running in this project?

[00:01:00.83] Now, in this project, you will be creating some regression models. And you will also be studying relationship between variables using scatterplot, as well as correlation coefficient. And then in step two, there are a couple of variables that are new here. So we will study these variables in the Python script. But be sure to read what these variables are. And then explain them in the data preparation step here.

[00:01:44.69] And so the first analysis in the project is a scatterplot and correlation for the total number of wins and average relative skill. So here we're just studying the relationship between total number of wins and the average relative skill. And the idea here is obviously we expect that if the average relative skill is high, then the total number of wins should be higher. And if the average relative skill is low, then the expectation is that the total number of wins will be lower.

[00:02:23.81] So let's open up the Python script in Codio. This is in step two in the Python script. So we will run this step and see how we can answer these questions here. So here's the Python script. So let's open up the script here by clicking this Project Three Jupyter Script. This will open up our Python script for this project. And be sure to go over this table. Read the introduction to this project.

[00:02:58.77] So here we have total number of wins, average points scored in a regular season, average skill, average relative skill of each team. And then you have these two variables that are new, average point differential and average relative skill differential.

[00:03:17.04] Average points differential is the average point differential between the team and their opponents in the regular season. And then the average\_elo\_differential is the average relative skill differential between the team and their opponents in a regular season. So be sure to go over this. And explain this in the data preparation step.

[00:03:39.24] So here's step one. We will run step one here to prepare our data set. So once this prints here, the data set is ready. And so here we can see we have the teams here, the average points scored by the team, average points by the opponents, average relative skill, average relative skill of the opponent.

[00:04:08.20] So the average differential is simply the difference between the two relative skill averages here. And then average-- sorry, this is the average points differential. That's the difference in the average points between the team and the opponent. And then you have the average relative skill differential here and then the total number of wins.

[00:04:32.71] So this is your data set. This is kind of like a summary data of the raw data that you've been using in the first two projects. All right, so here's this second step where we will create a scatterplot and correlation for the total number of wins and average relative skill. So you'll have to do some edits just like before. So replace this with the name of the data frame used in this project.

[00:04:58.36] Now, if you scroll up, the dataframe is nba\_wins\_df. So I'm going to highlight this. Control-C to copy it. And then I will make this edit here and here as well. Now replace relative skill with the name of the variable for average relative skill. So that here is avg\_elo\_n. And this will be in quotes. So 'avg\_elo\_n.'

[00:05:39.44] And then for the number of wins, the variable name is total\_wins. So replace wins with the name of the variable for the total number of wins. So again, single quotes, 'total\_wins'. And then I have to make the same edit here. So I'm just going to copy and paste it.

[00:06:08.23] Yes, this here. This step calculates the correlation coefficient and its p-value. And it tells us whether the coefficient is statistically significant, meaning it's statistically significantly different than 0. And I'll make this edit here. Whoops. Here.

[00:06:32.46] So that should be it. Now let's run this step. So here we have the total number of wins by average relative skill. And just like we expected, if the average relative skill is higher, you are expected to win more games. If it's lower, then obviously, the total number of wins is lower.

[00:06:56.29] And the correlation coefficient is 0.9. So it's very strongly positively correlated. These two variables are. And the p-value is 0. And that means this correlation coefficient is statistically significant. So this should help you answer these questions here. So again, be sure to provide your analysis for each of these-- addressing each of these questions. So make sure that you read these questions thoroughly. And then everything is here that will help you answer those questions.

[00:07:39.08] Now in the next step, we are going to create a simple linear regression model using the two variables. So remember that our regression model has a response variable, which is what we are trying to predict. And then we have predictors which are the independent variables that we're using to predict the response variable.

[00:08:04.14] Now, in a simple linear regression model, we only have one predictor. So here in this case, we're trying to predict total number of wins using average relative skill. So average relative skill is the predictor. And the total number of wins is the response variable.

[00:08:22.72] Now, the first bullet here, in general, how is a simple linear regression model used to predict the response variable using the predictor? So be sure to just explain the theory behind this. So I will help you address this question. The next question here is, what is the equation for your model? And then what are the overall-- what are the results of the overall F-test?

[00:08:46.86] So this question and the ones that follow, you'll have to run step three in the Python script to address those questions. So let's go in step three. So here's the step. Now we have to make two edits. So notice here, replace RESPONSE\_VARIABLE with the variable name that is being predicted. So what are we trying to predict? We're trying to predict total\_wins or total wins. And the variable name there is total\_wins.

[00:09:20.99] Now, here, this is important, it says, do not enclose this variable in quotes. So for example, if the variable name is var1, then replace this with var1 without the single quotes. So this is what we're trying to edit. The response variable will be 'total\_wins'. And then the same for predictor without the single quotes.

[00:09:52.76] So we need, for the predictor, it will be average relative skill. And the average relative skill variable name is 'avg\_elo\_n'. So let's make that edit here. That should be it. So let's now run this step. So here's the output. So you can use this output to address the questions.

[00:10:25.11] Here you can see that we have the R-squared is given here. F-statistic is given here. The p-value for the F-statistic is here. And remember, one of the questions was write the equation, the model equation. So you have to write the model equation using the coefficients that are given here.

[00:10:52.08] So I will let you write that. So just be sure that-- when you're writing the equation, be sure that you have the total number of wins and an average relative skill in the right position in the equation. So again, average relative skill is the predictor. And total number of wins is the response variable.

[00:11:27.57] These coefficients here should allow you to write that equation. Now, in the next step, we will create a scatterplot and get the correlation for the total number of wins and average points scored. So we just created a scatterplot in one of the previous steps for total number of wins and average relative skill. Now we're going to do the same the total number of wins but with average points scored.

[00:11:55.86] So step four in the Python script. So let's go to step four. And here, so we have to make these edits. Pretty much all edits are the same as step two, except we have to use average point scored now for the second variable. So if we scroll back to step two, this part here will be the same total number of wins.

[00:12:29.22] So I will copy paste that here and also here. And then I will make this edit for this dataframe name here as well. And now we need the variable name for average point scored. So if we scroll back to this table, we can see that the average point scored is 'avg\_pts'. So that is what we're going to type here in single quotes.

[00:13:08.93] And then the same thing here. And that should be it. If we run this, so we get our scatterplot as well as the correlation coefficient. So you can see that the correlation coefficient is moderately correlated. So it's not a strong correlation like in step two. But this moderate correlation is still statistically significant because the p-value is 0.

[00:13:41.42] So this should help you answer the questions here in this section of the template. The next section in the template is a multiple regression model now where we're predicting the total number of wins using average point scored and average relative skill. So essentially, we're just adding average point scored to the previous model. And that makes it a multiple regression model. It's not a simple linear regression model anymore.

[00:14:15.01] So again, first, you have to explain just in general, how is a multiple linear regression model used to predict the response variable? So just be sure to read the ZyBook section on this. And that will help you answer this question. And then again, you have to get the equation for your model and answer these questions. So this is in step five of the Python script.

[00:14:37.96] So if we go back here, this is step five. Now, again, similar to step three, you have to make these edits. And remember to not include single quotes this time. So the response variable is total\_wins. The first predictor variable here is the average relative skill or avg\_elo\_n. And then the second variable is for the points scored. So that would be avg\_pts.

[00:15:24.77] And you can write these variable names. So you can write average points first and then avg\_elo\_n. The order really does not matter. All right. So we have the edits in. Let's run this step. So here's the output. Again, you have your R scored, F-statistic, the probability, the p-value of the F-test. And then you have your coefficients here. And this here will help you write the multiple regression model for the total number of wins.

[00:16:01.91] So for example, the intercept for the model is minus 152 here, 0.5736. The coefficient for avg\_elo\_n in the model is 0.1055. The coefficient for average points is 0.3497. So this will help you write the equation. And then you can answer these questions here.

[00:16:26.62] Now, clearly you can see for the F-test, the p-value is 4.418 to the negative 243, which essentially is just 0.0. So the p-value is 0. So the model is statistically significant. And then you just have to answer these questions.

[00:16:42.61] One important item I wanted to clarify here is that you have to address the overall F-test, as well as the individual T-tests. So for the overall F-test, the F-statistic is here. And its p-value is here. For the individual T-tests, where we're testing each predictor, whether it's statistically significant in the model or not, the t value for the test is given here in this column.

[00:17:11.71] And the p-value of that individual T-test is given here. So, for example, for avg\_elo\_n, the p-value is 0. And for average points scored, the p-value is also 0. It means that each of these variables are statistically significant in the model.

[00:17:31.89] Now, here's the last model in the project. And this one is also a multiple regression model. But this time, we're adding two more predictors. So we're predicting the total number of wins using average point scored, average relative skill, and average points differential, as well as average relative skill differential. So there are four variables now, four predictors trying to predict the total number of wins.

[00:18:02.11] Now, again, to run this step, you have to run step six in the Python script. You have to write the equation and then address the rest of the questions. So let's go to step six. So here is step six. Now, in step six, you can see that you have to write the model yourself.

[00:18:22.93] So the easiest thing to do here is if you go back to step five, you can copy these steps here. Let's copy them here. Now, we already have half the model written here. We have the total number of wins using predictors, avg\_elo\_n and average\_pts. So those are these two variables. We just have to add these to predictors into this equation. So we have two of them.

[00:19:00.79] So let's first add avg\_pts\_differential. So this is the variable name, avg\_pts\_differential. So I'm going to copy, paste it. And then do the same for avg\_elo\_differential. Copy and paste. And that should be it. We should be able to run this model. I'll just change the variable name here to model3, since this is our third model. So it's not confusing.

[00:19:34.50] If we run it, here's the output. So now again, you have your overall F-test is provided here with its p-value. And you have the coefficients for each of the four predictors, as well as the intercept is given here. And the p-value for the T-test, for individual T-test is provided here.

[00:20:02.58] So using this data, you should be able to explain if the model is statistically significant based on the overall F-test. And then, which of the four predictors are statistically significant based on the individual T-tests. So you can address those questions.

[00:20:26.49] And then you have you have to do some predictions. So for example, what is the predicted total number of wins for a team that is averaging 75 points with a relative skill of this, average points differential of minus 5 and average relative differential of minus 30.

[00:20:45.87] Now, these, you can get. But first, what you have to do is write the equation. So the equation in this scenario, for example, would be total\_wins equals 34.5753 minus 0.0134 avg\_elo\_n plus 0.2597 avg\_pts plus 1.6206 avg\_pts\_differential plus 0.0525 avg\_elo\_differential.

[00:21:26.49] And then you have these values provided for the predictors. So you will plug in these values into the model and do the arithmetic to get the total number of wins. So that's fairly straightforward. And then you have a second prediction that you have to do here. So again, ZyBook does a good job of explaining this step as well. So be sure to read the section ZyBook so you can see how you can get these predictions.

[00:22:04.82] So that concludes this project. And then you have to provide a brief conclusion to briefly summarize your findings and what is the practical importance of the analysis that you performed. Also address which model would you actually use. Out of the three models that you have, which one would you actually use if you had to pick one?