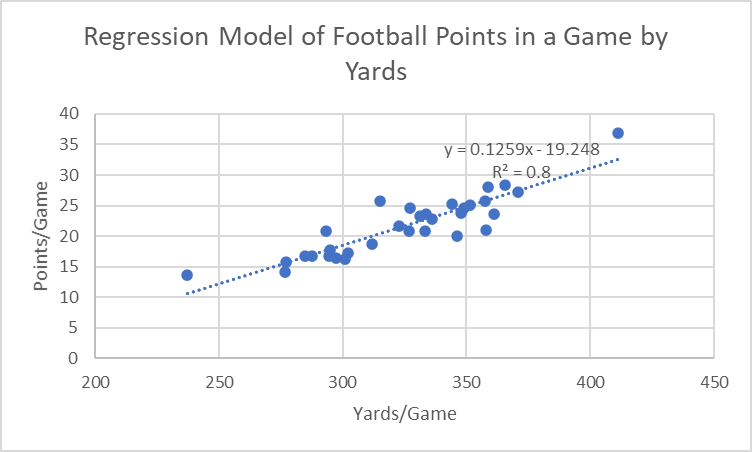
**BIS 305**

# Assignment 7

**Due 11/7/22**

11. The Excel file National Football League provides various data on professional football for one season.

a. Construct a scatter diagram for points/game and yards/game in the Excel file. Does there appear to be a linear relationship?



There appears to be a positive linear relationship between yards ran and points scored. The more Yards, the more points scored

b. Use the Regression tool to develop a model for predicting points/game as a function of yards/game. Explain the statistical significance of the model and the R2 value.

Using the information highlighted, I can understand that the regression equation is

y= -19.099 + .125x. For everyone extra yard ran, we can expect the overall score of a game to increase by .125

The

|  |  |
| --- | --- |
| SUMMARY OUTPUT |  |
|  |  |
| *Regression Statistics* | |
| Multiple R | 0.893509319 |
| R Square | 0.798358903 |
| Adjusted R Square | 0.791405762 |
| Standard Error | 2.298671233 |
| Observations | 31 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 1 | 606.6955934 | 606.6955934 | 114.819888 | 1.34531E-11 |
| Residual | 29 | 153.2327937 | 5.283889438 |  |  |
| Total | 30 | 759.9283871 |  |  |  |

.

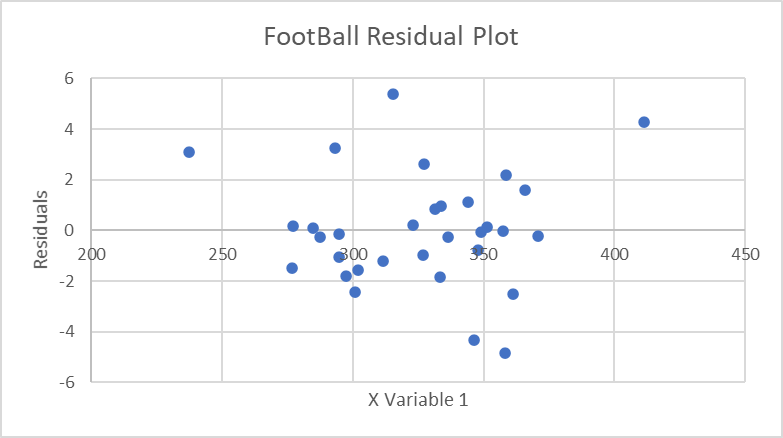
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -19.09923409 | 3.818775109 | -5.001403212 | 2.52676E-05 | -26.90950614 | -11.28896205 | -26.90950614 | -11.28896205 |
| 344.1 | 0.125319367 | 0.011695253 | 10.71540424 | 1.34531E-11 | 0.101399888 | 0.149238846 | 0.101399888 | 0.149238846 |

|  |  |  |
| --- | --- | --- |
| RESIDUAL OUTPUT |  |  |
|  |  |  |
| *Observation* | *Predicted 25.2* | *Residuals* |
| 1 | 18.62189533 | -2.421895335 |
| 2 | 18.7472147 | -1.547214702 |
| 3 | 15.62676247 | 0.173237533 |
| 4 | 16.60425353 | 0.095746472 |
| 5 | 17.64440427 | 3.255595727 |
| 6 | 24.51190558 | -0.711905578 |
| 7 | 24.92545949 | 0.174540511 |
| 8 | 26.73005837 | 1.669941628 |
| 9 | 24.29886265 | -4.298862654 |
| 10 | 21.36638947 | 0.23361053 |
| 11 | 27.35665521 | -0.156655206 |
| 12 | 22.7073067 | 0.992693305 |
| 13 | 25.8528228 | 2.247177196 |
| 14 | 25.68990763 | 0.010092373 |
| 15 | 15.58916666 | -1.489166657 |
| 16 | 16.93008388 | -0.230083882 |
| 17 | 23.03313705 | -0.233137049 |
| 18 | 32.43208956 | 4.367910435 |
| 19 | 26.16612122 | -2.466121221 |
| 20 | 22.43160409 | 0.868395912 |
| 21 | 17.83238332 | -1.032383324 |
| 22 | 17.84491526 | -0.14491526 |
| 23 | 25.77763118 | -4.777631184 |
| 24 | 21.93032662 | 2.66967338 |
| 25 | 20.40143034 | 5.398569655 |
| 26 | 10.63905166 | 3.060948335 |
| 27 | 24.62469301 | -0.024693008 |
| 28 | 18.18327755 | -1.783277551 |
| 29 | 21.855135 | -0.955135 |
| 30 | 19.96281256 | -1.162812561 |
| 31 | 22.68224282 | -1.782242822 |

17. Use the results for Problem 11 (National Football League) to analyze the residuals to determine if the assumptions underlying the regression analysis are valid. In addition, use the standard residuals to determine if any possible outliers exist.

The residual plot below has a constant variance (No pattern), is approximately normally distributed (with a mean of zero), an is independent of one another. This validates the regression analysis.

I added a residual plot of our data down below. I then added a table that highlights residuals bigger than 3 or less than -3. These residuals are considered outliers.



|  |
| --- |
| *Residuals* |
| 1.12930582 |
| -2.444872576 |
| -1.570761708 |
| 0.163877688 |
| 0.081942455 |
| 3.237062657 |
| -0.761661797 |
| 0.122904067 |
| 1.610100561 |
| -4.347650272 |
| 0.198155426 |
| -0.219345101 |
| 0.951141709 |
| 2.191324487 |
| -0.045019641 |
| -1.498355572 |
| -0.245369289 |
| -0.276170035 |
| 4.282145038 |
| -2.523398344 |
| 0.828097801 |
| -1.051771042 |
| -0.164359955 |
| -4.833142033 |
| 2.63165433 |
| 5.367501745 |
| 3.074265156 |
| -0.074962016 |
| -1.804260612 |
| -0.992812191 |
| -1.191886292 |
| -1.823680464 |

24. Using the data in the Excel file Home Market Value:

Find the best multiple linear regression model for estimating the market value as a function of both the age and size of the house.

**Y = 47331.38 -825.161(House Age) +40.91107 (Square ft)**

Predict the value of a house that is 30 years old and has 1,800 square feet, and one that is 5 years old and has 2,800 square feet.

**Y = 47331.38 -825.161(30) +40.91107 (1800) = $96,216.5**

**Y = 47331.38 -825.161(5) +40.91107 (2800) = $157,757**

36. For the Excel file Auto Survey:

a. Find the best regression model to predict miles/gallon as a function of vehicle age and mileage.



**Y = 36.1809 - 0.76641(vehicle age) - 8.1E-06(Mileage)**

b. Using your result from part a, add the categorical variable purchased to the model. Does this



**Y = 37.25189 + 7.603556 (new or used) - 1.68201(vehicle age) - 1.98E-05(mileage)**

change your result?

By adding New/Used as a dummy variable, it did change my results. It increased the R squared value from 0.22248 to 0.285668.

c. Determine whether any significant interaction exists between vehicle age and purchased variables.



When I multiplied age \* whether it was new or used and then used that as a variable in my multiple regression, I got that the t stat was -2.03144 and the p value is 0.057243. Using an alpha level of .05, because the p value is bigger than .05, there is no significance.