**Student: Kasia Gorska**

**NBA – Regression Analysis**

1. **After you have run regression, explain why your choice was or was not a good choice.**

I have selected minutes played as an independent variable (x-variable). I believe that it was a good choice because the minutes played can be controlled. Having said so, this independent variable can be used to test the effects on the dependent variable (points scored), since the depended variable is the variable being tested and measured in an experiment.

1. **Do the results have a good `R Square` value?** **Explain what your coefficients mean.**

The correlation coefficient, the multiple R, measures the strength of a linear relationship between two variables. In our example, the multiple R is 0.90 (rounded to 2 digits); therefore, it appears that there is a strong correlation between the independent variable (minutes played) and the dependent variable (points scored).

In our example, R2, the coefficient of determination, is 0.82 (rounded to 2 digits). It means that 81% of our values fit the regression analysis model. In other words, 81% of the dependent variables (y-values) are explained by the independent variables (x-values). The other 19% of the total variation in y-values remains unexplained.

Generally speaking, R-square of 95% or more is considered a good fit. However, it is important to mention that R-square does not demonstrate whether a regression model is accurate. For example, we can have a low R-square value for a good model, or a high R-square value for a model that does not fit the data. A high R-square value does not guarantee that a cause-and-effect relationship exists.

1. **Is your choice statistically reliable?**

It is hard to answer this question definitively because we can never be completely 100% certain that a relationship exists between two variables. There are too many possible errors to be controlled, for example: various problems with reliability and validity, researcher bias, sampling error, etc.

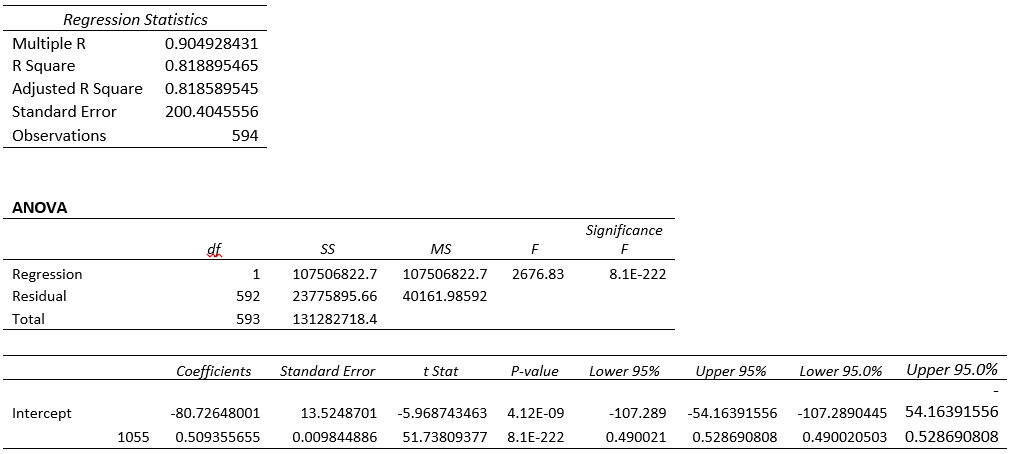
In addition to the above, the p-values for the coefficients indicate whether these relationships are statistically significant. The p-value for each independent variable tests the null hypothesis that the variable has no correlation with the dependent variable. In our example, the p-value is high, therefore, we may want to modify our model, since it appears that some of the variables are not statistically significant and, therefore, they reduce the model’s precision.

1. **Finally, produce a scatter plot and explain your overall analysis.**

**Chart 1. Minutes Played vs. Points Scored.**

**Multiple R = 0.90; R Square = 0.82**

**Table 1. 2017 NBA Season - Regression Analysis.**



As we can see in Chart 1, there is a large positive linear association. Meaning, the points are close to the linear trend line (a linear relationship exists). By looking at our data we can conclude that variable-x and variable-y have a positive association because, in general, as minutes played increase, so do the points scored by the players. When it comes to the strength of the relationship between the two variables, we can observe that a moderate relationship exists between those variables. On the other hand, our model has a high p-value (see Table 1). Generally speaking, high p-values reduce model’s precision, therefore, we may consider modifying our model, since it appears that some of the variables are not statistically significant.