Dolphin Longevity Predict Report

Zhaoying Meng

October 11, 2020

**Background**

This data set contains two variables that have been assigned to dolphins based on their most recent physical examinations: longevity (a health index based on a physical examination of the dolphin, 100 = maximum possible value) and IQ (measured via a series of validated tests performed by a scientist). The set also contains two additional variables: a close relationships index (higher scores indicate the dolphin has formed one or more close relationships with other dolphins) and a social integration score (higher scores indicate a dolphin is more socially active within its pod). These variables have been calculated using analysis of the GPS tracking data of the dolphins over the same period. Create a model for the longevity score based on the other variables provided. If a reliable model could be created, this could help marine biologists use GPS data to identify dolphins with higher health risks and could also help extend the reach of research dollars by potentially helping scientists reduce their experimental costs of performing physical examinations of dolphins in the wild.

**Relationships**

Chart, scatter chart

Description automatically generated

1. Longevity and close relationships

The table indicates the correlation coefficients of longevity and close relationships is 0.9267 which is greater than 0, so there may be a positive relationship between them. That means dolphins which have higher scores in having more close relationships with other dolphins have a long longevity. The scatter plot illustrates the same situation. In addition, because 0.9267 is greater than 0.6, they have a strong relationship.

1. Longevity and social integration

The table indicates the correlation coefficients of longevity and social integration is 0.8713 which is positive and greater than 0.6, so there may be a positive strong relationship between them. That means dolphins which have higher scores in having more socially active within its pod have a long longevity. The scatter plot is consistent with what is expected.

1. Longevity and IQ

The table indicates the correlation coefficients of longevity and IQ is -0.0433 which is almost approaches 0, so there may be not a linear relationship between them. It cannot be seen from the scatter plot that the dolphin's IQ score has any obvious relationship with its longevity.

**Linear regression model**

**Table

Description automatically generatedChart, scatter chart

Description automatically generated**

By observing the data in question 1 above, the correlation coefficients between longevity and close relationships is the biggest which is 0.9267. That means close relationships has the greatest correlation with longevity. So, I create a linear regression model for them.

Let *x* represent the scores of close relationships.

The model equation: ŷ = -37.7773 + 1.4080 x

Coefficient of determination is 0.8588 showing 85.88% of the variability in the sample values of longevity can be explained by the linear regression model between longevity and close relationships.

**Hypothesis test**

Claim: the estimated regression parameter *β1* is 0. Ho: μ = 0; Ha: μ ≠ 0

Because the P-value associated with the estimated regression parameter b1 is close to 0 (< 0.0001) less than the level of significance (0.05), the conclusion is REJECT THE NULL *β1* = 0. We can be 95% confident in this model that there is a relationship between the score on longevity and close relationships.

**Leverage points or outliers**

**Chart, scatter chart

Description automatically generated Table

Description automatically generated**

**Chart, scatter chart

Description automatically generated Table

Description automatically generated**

The linear regression model equation:

Above: ŷ = -37.7773 + 1.4080 x (with the leverage point)

Below: ŷ = -41.1935 + 1.4669 x (without the leverage point)

First, [according](javascript:;) [to](javascript:;) [the](javascript:;) [definition](javascript:;), a data point has high leverage if it has "extreme" predictor *x* values. I observe the data set. When the score of close relationships is 100 which is very different with other scores of close relationships.

At the same time, when I observe the residual graph, there is a high leverage point at about 100 above the x-axis which is much larger than other values. I think this point may have an impact on the predicted value, so I try to delete this point in the data set. Then I rerun the Linear Regression for the data set and get the data comparison and regression model above.

For the first regression, when the score of close relationships is 100, the residual is -9 (94 - 103) and r-squared values is 0.8588. For the second, the residual is -11 (94 - 105) and r-squared values is 0.8487. I think the first equation is still the best choice, but the leverage point is worth concerning.

**Quadratic and a cubic regression mode**

This is the residuals and statistical data for the linear regression model.

**Chart, scatter chart

Description automatically generated Table

Description automatically generated**

This is the residuals and statistical data for the quadratic regression model.

**Chart, scatter chart

Description automatically generated Table

Description automatically generated**

This is the residuals and statistical data for the cubic regression model.

**Chart, scatter chart

Description automatically generated Table

Description automatically generated**

By observing the residuals for the linear regression model, we can see the mean of the residuals is not zero at several values of the independent variable close relationships. Residuals for relatively small or relatively large values of close relationships is more positive, while the values in the middle is more negative. There appears to be a curvilinear relationship between longevity and close relationships (smiley face). This suggests the relationship between the independent variable close relationships and the dependent variable longevity may be nonlinear, and so a simple linear regression model may not be appropriate.

Then we observe the residuals for the quadratic regression model, its shape is similar to the residuals for the linear regression model. There appears to be a curvilinear relationship between longevity and close relationships (smiley face). So, a quadratic regression model may not be appropriate, too.

In the last, the residuals for the cubic regression model is much better. It is averaging close to zero and a consistent variance. I think the cubic regression model seems to have the best fit statistically.

In addition, according to statistics, the coefficient of determination for cubic regression model is r-squared values = 91.8%, for quadratic regression model is 87.99%, for linear regression model is 85.88%. So, the cubic regression model explains the most variation in the sample values of longevity in the three-regression model. So, the cubic regression model is superior to the quadratic regression model and the linear regression model.

**Multivariate linear regression model**

Chart, scatter chart

Description automatically generatedTable

Description automatically generated

Let *x*1 represent the scores of close relationships.

Let *x*2 represent the scores of social integrations.

Let *x*3 represent the scores of IQ.

The model equation: ŷ = -65.2028 + 1.0396 *x*1 + 0.7417 *x*2 – 0.1109 *x*3

The coefficient of determination is 0.8588 in simple linear regression model, while the coefficient of determination is 0.8889 in the multivariate linear regression model. It shows 88.89% of the variability in the sample values of longevity can be explained by the multivariate linear regression model. It explains over 3% more variation than did the simple linear regression. The multiple linear regression model improves on the linear regression model in question 2.

**95% confident**

* 1. Claim: the estimated regression parameter β1 is 0. Ho: μ = 0; Ha: μ ≠ 0

Because the P-value associated with the estimated regression parameter b1 is close to 0 (<0.0001) less than the level of significance (0.05), the conclusion is REJECT THE NULL β1 = 0. We can be 95% confident in this model that there is a relationship between the score on longevity and close relationships.

* 1. Claim: the estimated regression parameter β2 is 0. Ho: μ = 0; Ha: μ ≠ 0

Because the P-value associated with the estimated regression parameter b2 is 0.0253 less than the level of significance (0.05), the conclusion is REJECT THE NULL β2 = 0. We can be 95% confident in this model that there is a relationship between the score on longevity and social integrations.

* 1. Claim: the estimated regression parameter β3 is 0. Ho: μ = 0; Ha: μ ≠ 0

Because the P-value associated with the estimated regression parameter b3 is 0.2979 greater than the level of significance (0.05), the conclusion is FAIL TO REJECT THE NULL β3 = 0. We cannot be 95% confident in this model that there is a relationship between the score on longevity and the scores of IQ. That means there is not a relationship between longevity and the scores of IQ at the 0.05 level of significance.

**Final multivariate model**

Because there is not a relationship between longevity and the scores of IQ, we should remove the variable IQ. Then I get the final multivariate model.

Table

Description automatically generated

Let *x*1 represent the scores of close relationships.

Let *x*2 represent the scores of social integrations.

The model equation: ŷ = -78.0949 + 1.0178 *x*1 + 0.7760 *x*2

Compare the simple linear regression model (in question 2) to the final multivariate linear regression model, their coefficient of determination is 0.8588 and 0.8843. The final multivariate linear regression model explains about 3% more variation than did the simple linear regression. So, it improves on the linear regression model, too.

Compare the multivariate linear regression model (in question 6) to the final multivariate linear regression model, their coefficient of determination is 0.8889 and 0.8843. They have the same level to explain variation in the sample values of longevity. I think the final multivariate linear regression model have the best fit statistically, because it deletes the irrelevant independent variable IQ making the model simpler, and at the same time has a similar fit with the multivariate linear regression model (in question 6).