Quiz #3

Dustin Leatherman

March 3, 2019

1

In the chimpanzee learning times case study, why is the learning time data not analyzed as a 10-group analysis of variance with four observations on each sign?

The Chimp itself was a block in this study. Each cell consists of a particular Chimp and a particular Word. If Chimp was not a block, then the analysis could be done as the question suggests.

2

In the chimpanzee learning times case study, what would have to be true to support the inference that these learning time patterns are representative of learning times for these signs in a wider population of chimpanzees?

These chimpanzees would have to be randomly selected from a larger group of chimpanzees.

3

In the chimpanzee learning times case study, why were Tukey-Kramer intervals used in the comparison of sign means?

Since there were 10 Choose 2 signs, the Tukey-Kramer intervals are helpful for making larger numbers of comparisons.

4

Having no replicates means that the model that is developed will fit the exact data. Data is variable so the model has the potential to be wrong when applied to other datasets.

5

Below is a plot of the isotopic composition of structural bone carbonate (X) and the isotopic composition of the coexisting calcite cements (Y) in 18 bone samples from a specimen of the dinosaur Tyrannosaurus rex. Evidence that the mean of Y is positively associated with X was used in an argument that the metabolic rate of this dinosaur resembled warm-blooded more than cold-blooded animals. (Data from R. E. Barrick and W. J. Showers, "Thermophysiology of Tyrannosaurus rex: Evidence from Oxygen Isotopes," Science 265 (1994): 222–24).

a

Comment on the effects on the p-value fo significance of regression and on R-squared of deleting the case with the smallest values of X and the two cases with the smallest values of X.

The p-value remains low for each of the three datasets which means there is definitely a relationship between the isotopic composition of Carbonate and Calcite cements. As the leftmost x values are removed, the estimated slope and the R^2 value decreases, and the standard error increases. Since there are no other data points on the leftside of the graph, these two x-values are influential and greatly affect the model. The R^2 value decreases because there is less data in sparser areas to drive the calculation of the slope. If those values are removed, then the model explains less of the variability found in the dataset.

b

R^2 changes drastically with the removal of the two x-values because they provide anchors for the regression line. Without those points, the regression line could be any positive slope.

c

Below are the plots of the case influence statistics. Comment on any interesting cases.

The leftmost x-values have high leverage which indicates that these may be influential. As seen in (a), these two points affect the slope so they are influential. While case 1 has a higher Cook's Distance than the other cases, it does not exceed 1 which is a rule of thumb for indicating whether or not a data point is influential. Case 2 has a low Cook's Distance, meaning that if it were dropped, the slope wouldn't change much. The absolute Studentized Residuals does not exceed 2 but it does appear to have a parabolic shape in the residiual plot when it should not have a shape.

d

Below are the plots of the case influence statistics with the smallest X deleted. Comment on the differences in the two sets of case statistics.

The Leverage and Cook's Distance for case 2 have now increased and somewhat match case 1 as described in (c). The key difference is that the Cook's Distance for case 2 has increased to ~2.5 in this plot whereas it was <1 for case 1 in the first plot. The Studentized Residuals have changed a little but still fall within an absolute value of 2 meaning. The high leverage and high Cook's Distance indicate that this is potentially influential and the high Studentized Residual means that it could be a potential outlier.

e

Why may pairs of influential observations not be found with the usual case influential statistics?

The usual case influential statistics convey how the model changes if a given value is substituted. If there are pairs of influential observations and one is taken away, then the model is not affected as much since there is still an outlier observation to keep the regression somewhat consistent.

f

What might one conclude about the influence of the two unusual observations in this data set?

These two observations are influential since they affect the slope in a meaningful way. However, given the sparsity of the data and the reduction in R^2 after removal, the model with the full dataset is preferred unless more data is retrieved.