Notes

Adam Kaderbhai

Outliers: Leverage, Discrepancies and Influence

In a two variable dataset individual points may be unusual in their x-values, y-values or both. An outlier is a point with high discrepancy the one whose y-value is far from the general trend of the data.

Observations with unususal x-values have greater potential to affect the fit of the model

Such observations are said to have high leverage and this is only dependent on the x-value

An observation that changes the fit of the model is said to have high influence

For an observation to be influential it has to have high discrepancy and leverage

These ideas can be quantified:

- Discrepancy can be measured by studentized residuals
- Leverage can be measured by hat values
- Influence can be measured by cook's distance

Questions:

1.) We say "An outlier is a point with high discrepancy the one whose y-value is far from the general trend of the data". It could also be x-value?

2.) So once an outlier it is a point with both high discrepancy and leverage and hence influential?

R^2 The Coefficient of determination

R^2 is a measure of spread of observations about a regression line or other statistical model. It represents the fraction of the variance of the response variable that can be attributed to changes in the explanatory variable

For linear models R^2 is always between 0 and 1. Larger values indcate that the data is tightly packed around the regression line.

In a least-squares regression model with a single explanatory variable the coefficient of determination is the square of the sample coefficient

$R^2 = r^2$

In a linear model with multiple explanatory variables R^2 is equal to the square of the correlation between observed and fitted y-values

- R^2 it measures the proportion of the variability in y that can be explained by variability in x. By definition it is naturally lower for datasets with high $\text{Var}(y_i)$
- Models with R^2 are not neccesarily good.
- Models with low %R^2$ are not neccsarily bad