Ritam Dey

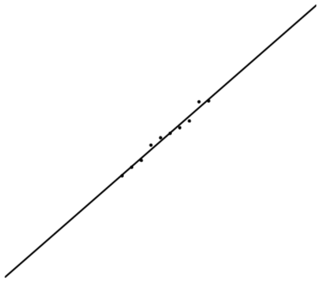
BS1913

Indian Statistical Institute, Kolkata

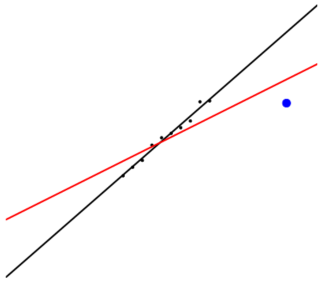
Linear models assignment

1 Preparation of R

We are going to do an influence diagnostic using R i.e. we are interested to find the influential points inside our data. Before that here is a quick reminder of what influential points are and why are they so much important. If we have a cluster of data points and we fit a line passing through the points then most often we see that the points are spread near the line following some distribution. Though there might be cases where we have in our data a cluster of points and a point or few lying far away from the cluster, which makes the regression line not go through either the cluster or the outlying point itself. Had we not included the point in our data, would have nice looking regression line going through the cluster and inference would be very easy. But finding such a point is worth the analysis or not and taking action according to its importance to the inference is the main goal of influence diagnostics.



regression line without outliers



regression line with outlier

Finding outliers in two-dimensional data is quite easy because plotting the data and looking for points beyond the cluster still remains one of the novel approaches. But the problem arises when we consider multidimensional data. But we have automated ways of handling such data. After the detection, necessary steps have to be taken which further depends on the source of the influential point. An unambiguous way of classifying the reasons might be **mistake, misdeed and miscalculation.**

The experimenter might make a mistake in recording the data, so the outlying point is nothing but a typo . In that case, we should be able to detect it and correct it. We might think of this influential diagnostic as a spell checker for the typo in the data.

The next reason is due to the misdeed of the data. What we mean by misdeed is, the data lies outside the population that we are surveying but it somehow came into our sample. Whenever we are carrying out some kind of statistical analysis, we have in our mind some kind of population for which we want our inference to be valid and all the sample points must be inside that population. Occasionally something may enter our sample inadvertently that really does not belong to the population. Consider, for example, we are interested in studying the behaviour of a certain economic class say the lower middle -class in the state of West Bengal, India. Prior knowledge about the population tells us that majority of the lower middle-class students study in government schools or government-sponsored schools. Because an upper-middle class or rich family will find a school with other better facilities along with education and is also able to bear the cost. So we consider all the students studying in government schools or government-sponsored schools as our population and draw our sample from that. Now it is possible to find a student who is not actually belonging to a middle-class family but studying in such a school for some reason which is not our concern. The student is outside our population and we included him in our sample because it went against our general assumptions that financially strong students must go to schools with high admission fees. In that case, we must exclude the student who is an influential observation and consider the rest of the sample.

The third reason for getting an influential observation is that we might have ignored some special properties of the outlier which is still inside the population and that special property makes the observation an extreme case in our population as well as in our sample. To understand the situation, consider we are doing a clinical study and we have taken into account lots of different factors that are going to be of importance. If we find certain characteristics present in the patient, say the patient is pregnant and she is under a particular drug then the effect of the treatment might be drastically different. We had no way of knowing that the combined effect of pregnancy and the drug is going to have an effect on the treatment so we did not include that in the list of factors when building our model. But if the influential observation is very important then our conclusion should be building two different models for two different cases. One case is where the special characteristic does not occur and the other one is when the characteristic occurs. Two different models should be considered for two different cases. In this case, we really do not throw away the data point from our study. Instead, we set the point aside, fit the model to the remaining bulk of the data and then consider that point separately.

So depending on the reason behind the genesis of the influential point we have to act according to the overmentioned way.