Residuals : Introduction

Samahriti Mukherjee
Roll No.: BS2003
B.Stat 3rd year
Indian Statistical Institute

1 Introduction

We begin with asking about the **Goodness of Fit** of the Linear Model. We have learned that given a particular form of the linear model, i.e., given a designed matrix, how to find the best estimator for $\hat{\vec{\beta}}$ & $\hat{\sigma}^2$.

But the question is whether the best estimators are enough or not.

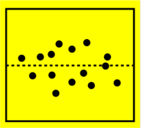
Naturally your first expectation would be to have some measure such that once you have fitted the model, you compute that measure (something like an ideal number) and compare it with a threshold. If our measure exceeds the threshold then we say that our estimator is good. If not, i.e., if the number is below the threshold then we say our estimator is bad, or the other way round, i.e., if our measure exceeds the threshold then we say that our estimator is bad. If not, i.e., if the number is below the threshold then we say our estimator is good. Unfortunately you can never capture the nuance of a linear model in terms of a single summary statistic.

2 Residuals

The best technique always remains to look at the residuals, i.e., the residual vector.

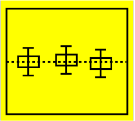
When we say that we want to look at the residuals, we mean plotting the residuals against whatever comes to our mind.

We might plot them according to the order of observation, i.e., if the ith residual is called e_i, we can plot the e_i's against the i's. Always attach them against some covariates, or we may plot them against any factor input that we have. Whenever you make any plot of the residuals make sure that you draw the zero length, i.e., the horizontal line through 0. You should get the ideal estimation when you are plotting the residuals either against the observation number or against any covariate.



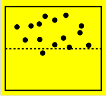
In the left figure, you can see that all the points are more or less patternlessly scattered around the horizontal line through 0 symmetrically, which is our desired thing. Any departure from this will make us suspect that our model is not a good fit. Ofcourse you should also make sure that these points are not too far away from 0, in which case you will have very large error. But the main point is **you should not have any systematic departure from 0**. Even if you have some large departure from 0 that is possibly allowable but no systematic departure is allowed.

If you are plotting the residuals against some factor input then your plot should be like this where each factor (or it can be a factor combination) will have a scatterplot (or a boxplot) . All these boxplots should be more or less centered about the 0 line and they should have similar amount of scatter.

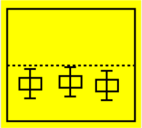


If they have different scatters then you should suspect that homoscedasticity is not a good assumption.

3 Some Examples where Things can go Wrong



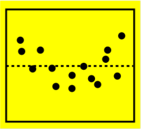
Suppose we are plotting against observation order or covariant and our figure is like the figure in the left, i.e. all the residuals are more or less above the 0 line. So they are not symmetrically around 0. This happens when there is no intercept term. You should rectify it by putting an intercept term.



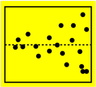
If you are plotting against some factor input in that case you might get all the boxplots consistently below or consistently above of the horizontal line through 0. This also suggests that you should put an intercept term.

If there is a systematic departure from 0, i.e. on an average they are against 0 but initially the residuals are positive, then they are negative and after that they are again positive.

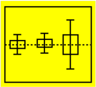
Again the systematic departure is typically same when you are working with a radiation setup. This is a continuous thing covariate. Possibly including higher powers of the covariates will solve the problem. You must make sure that including higher power term is really justified in terms of the domain of the problem.



If you are seeing the case where you are plotting against the observation order then the observations are most likely not i.i.d. There is something which is changing over time. So when you are taking the first or last few measurements you have done error in the positive direction. When you are taking the middle measurements you have done error in the negative direction. So possibly, you are having a time series effect in your setup.



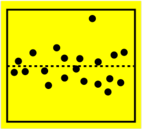
If you see a pattern like the figure in the left, then this indicates **Heteroscedasticity**. So things have a fanning out effect or a fanning in effect depends on how you look at it. There are various ways by which you can tackle it namely by going into **Generalised Least Squares** etc.



A similar pattern is observed if you plot the residuals against factors. You will see some boxes which are tall and some boxes which are thin. All boxplots are more or less around 0, but this indicates that you have possible heteroscedasticity.

Now we consider the case where you may have an outlier like the figure in the left.

Often you may want to ignore the outlier as it is just one point, or you may just throw away your model.



Finding the reason behind why this is an outlier might be the most important part of the particular statistical report instead of throwing the model (which should be the general idea), because it often turns out that some special feature is related with the outlier which may open up a new area of research.

So these are the various ways by which you can detect lack of fit of your model by looking at the residual plot and this is by all means the most important way by which you can detect the lack of fit.