Multiple Linear Regression

Authors: Xuxian Chen, Kevin Cheng, Sujay Chebbi. (PDF)

Following the assumptions of Multiple Linear Regression (MLR), we continued to cover several regression concepts, including Mean Square Error(MSE), Maximum Likeliihood Estimate(MLE), Adjusted R-square, Regression Excel Output, etc.

1. R-square is usually lower in the test set than in the train set, if you have a large enough data set.
2. When using a more complex model, the r-square naturelly goes up, and the adjusted r-square is to normalize that.
3. You cannot compare the Mean Square Error(MSE) of different model because the scale.
4. Colinearity Problem: dependencies between the X's inflate standard errors. When there is multicolinearity, one features's value contrain another, leading to a higher uncertaincy in prediction.
5. Onehot coding: creating dummy variables for categorical feature. When we are adding the parameters, we need to consider if the parameter creates values.
6. Sanity check if the linear model is reasonable: Check residual plot.

   Professor also introduced the COVID-19 models presented in FiveThirtyEight webpage for us to look at.

A Deeper Look into Standard Error of MLR

- Standard Error Formulae

$$s_{b_j}^2 = \frac{S^2}{(N-1)\,(\text{Variation in } X_j \text{ not associated with other X's})}$$

  >1. S^2 is the variance of the output noise >2. $S_{b_j}^2$ is a standard error on the weight (β) of feature j >3. Variation in X_j not associated with other X's is the important part of the formular against colinearity

- $S_{b_j}^2$ helps us to determine whether the feature significant or not (I can be used to reject Hypothese that b_j equal zero); S^2 helps us to determine the model output(y's) precision.

- With a smaller standard error, the output precision is higher.

- Standard error is also valid for both linear and nonlinear regression models, which is convenient if you need to compare the fit between both types of models, and it is why some people prefer it over R-square

Residual Plot

Key ideas: A random pattern of residuals supports a linear model.

That if a pattern is dedected, very often it means that there is some transformation can be added to our models, in another word, there could be a more accurate model for the data.

"P < 0.05" Might Not Mean What You Think: American Statistical Association Clarifies P Values

Whole article: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017929/

Extract:

Test your knowledge: Which of the following is true?

P > 0.05 is the probability that the null hypothesis is true.

1. 1 minus the P value is the probability that the alternative hypothesis is true.
2. A statistically significant test result (P ≤ 0.05) means that the test hypothesis is false or should be rejected.
3. A P value greater than 0.05 means that no effect was observed.

If you answered "none of the above," you may understand this slippery concept better than many researchers.

Some thought-provoking statement from the article:

> The ASA panel defined the P value as "the probability under a specified statistical model that a statistical summary of the data (for example, the sample mean difference between two compared groups) would be equal to or more extreme than its observed value."
>
> "You can fish through a sea of data and find one positive finding and then convince yourself that even before you started your study that would have been the key hypothesis and it has a lot of plausibility to the investigator."
>
> "If success is defined based on passing some magic threshold, biases may continue to exert their influence regardless of whether the threshold is defined by a P value, Bayes factor, false-discovery rate, or anything else,"
>
> "Consult a statistician when writing a grant application rather than after the study is finished; limit the number of hypotheses to be tested to a realistic number that doesn't increase the false discovery

rate; be conservative in interpreting the data; don't consider P = 0.05 as a magic number; and whenever possible, provide confidence intervals."