Using EDA, ANOVA and Regression to Optimise some Microbiology Data.

Neil Binnie
Auckland University of Technology

*Journal of Statistics Education* Volume 12, Number 2 (2004),
ww2.amstat.org/publications/jse/v12n2/datasets.binnie.html

Copyright © 2004 by Neil Binnie, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

**Key Words:** Analysis of variance; Exploratory data analysis; Interactions; Multiple regression; Optimisation; Outlier; Polynomial regression.

**Abstract**

Bacteria are cultured in medical laboratories to identify them so patients can be treated correctly. The tryptone dataset contains measurements of bacteria counts following the culturing of five strains of *Staphylococcus aureus*. It also contains the time of incubation, temperature of incubation and concentration of tryptone, a nutrient. The question is whether the conditions recommended in the protocols for the culturing of these strains are optimal. The task is to find the incubation time, temperature and tryptone concentration that optimises the growth of this bacterium. Students may explore these data at several levels. Graphical methods can be used to investigate the relationship between the variables. ANOVA can be used with one-way, two-way and factorial models with interactions, to identify significant factors. Multiple polynomial regression methods can be used to model the data, with optimal conditions estimated by partial differentiation.

**1. Introduction**

A person may have a boil or an infected wound from an operation. If it is not responding to antibiotics, they will often be required to supply a swab from the wound for a medical laboratory to investigate. A sterile cotton bud swab is wiped across the wound and wetted with the pus that contains myriads of bacteria cells. The cotton bud is then swirled in a salt enrichment broth and incubated for a period of time. This broth may then be diluted and a drop wiped across the surface of a solid nutritive medium so that bacteria are transferred to the medium surface. The laboratory cultures the sample on a variety of media one of which is suitable for the growth of Methicillin resistant *Staphylococcus aureus* (MRSA). The aim is to culture the offending bacterium quickly so that it can be identified and susceptibility testing performed so that a suitable antibiotic can be used. For details on culturing techniques, see *Bacteriological Analytical Manual Online (2001).*
A Bachelor of Applied Science (BAppSc) student at Auckland University of Technology (AUT) was working in a medical laboratory where they had a very poor record of recovering MRSA. They suggested a project to identify the optimal conditions for culturing this bacterium. For the laboratory, they need to be confident that, if the bacterium is present, they are providing it with the best conditions for its growth. The medical laboratory cultures a sample on a nutrient gel, and want to observe whether the bacterium is present or not. If the bacterium is present, it appears on the culture and is visible to the naked eye. No counting is done, as the task of the medical laboratory is simply to detect its presence.

For this experiment, five strains of MRSA were obtained from Environmental Scientific Research (ESR), Wellington, NZ. These were maintained on human blood agar and subcultured onto fresh blood agar every few days and incubated for 24 hours. Salt enrichment broth was dispensed in 5 mL amounts and inoculated with the desired strain of MRSA by touching an individual colony of MRSA from the blood plate with a wire loop and placing the loop into the broth making sure the sample is removed. A barely visible amount of the colony, much smaller than a pinhead, will contain millions of cells. The broth was incubated for the time and temperature specified by the design of the experiment. Each broth was diluted 1000 times and one micro litre of this dilution transferred onto plate count agar and incubated for 24 hours. MRSA colonies were then counted manually. The number of bacteria in the broth was then estimated from the colonies counted and the dilution.

MRSA is a bacterium, whose strains are resistant to penicillin and sensitive only to the expensive antibiotic vancomycin. There are at least eight strains of this bacterium and five were available from ESR. This bacterium is cultured in a 1.5% (by weight) sodium chloride enrichment broth. The salt in the broth inhibits the growth of most normal flora that may be present at the site from which the specimen was collected. This ensures that it will be easier to identify the Staphylococcus aureus if it is present.

Tryptone is a source of amino acids derived from casein (a protein derived from milk), which is included in the broth as a nutrient. It contains nitrogen, in a form readily available to the bacteria, which encourages growth. The current protocol requires this to be included in the broth at a concentration of 1.0% (by weight) and cultured at 35 degrees Celsius for 24 hours. The goal is to decide if these are the best conditions for culturing the MRSA.

Five strains of MRSA were used so that the conditions for optimum growth could be investigated for each strain separately. These strains were WSPP1 MRSA, WSPP2 MRSA, Akh2 MRSA, Phage pattern 52/52A/79/..., and Phage pattern 29/52/77/+. Because of these curious names they are simply referred to in the data as 1, 2, 3, 4 and 5 respectively. Once the optimal conditions are found for each strain, the conditions could then be chosen that would be the best compromise to enhance the growth of whatever strain happened to be present in any given sample.

2. Data Source

The data was collected by Gavin Cooper at the Auckland University of Technology and is used with his permission. This was done for a project for his Bachelor of Applied Science entitled "The Efficiency of the Recovery of Methicillin Resistant Staphylococcus aureus from Salt Enrichment Broth".

3. Description of the Data

There are 30 cases in this data set. For each case the first five observations are the bacteria COUNTS for the five strains of MRSA. A value of 62 stands for 62 million CFU/mL. (CFU = Colony Forming Unit each of which originated from one bacterium) The TIME is the incubation time, which takes the values 24 or 48 hours. The TEMPERATURE is the incubation temperature, which takes the values 27, 35 and
43 degrees Celsius and the CONCENTRATION is the Tryptone concentration, which takes the values 0.6, 0.8, 1.0, 1.2 and 1.4 percent by weight. This is a factorial design with no replications.

4. Pedagogical Uses

The relationships between the variables can be first explored graphically to get some feeling for the data. First concentrate on one strain - five groups in a class could each investigate a different strain, for example. Comments below relate to Strain one, and Minitab graphs and analysis tools are referred to. The aim is to find how the Counts depend on Time, Temperature and Concentration. The dependence of Count with each predictor variable can be investigated separately using dot-plots, box-plots and scatter-plots. Interaction plots, main effects plots and multiple scatter plots give more insight into these relationships.

These graphs will give rise to fruitful discussion and estimates of the optimal conditions for growth of the bacteria. For Count1 (the count for strain 1) they show that 48 hours give higher growth than 24 hours, optimal temperature is about 35 degrees and optimal concentration is about 1.2%. A plot of Count1 by Temperature shows an unusually high value for 27 degrees. This point, number 9, gives the opportunity to discuss outliers and what the experimenter may do with this case.

Because the predictor variables are all measured, but each has only a few discrete values, both ANOVA and Regression analysis may be used to investigate this data.

The graphs suggest that each of the predictor variables is significant for Count1. First students could use a 2-sample t-test and one-way ANOVA to test this. Time and Temperature are found to be significant factors at the 0.05 level. Note that $\alpha = 0.05$ is used for the remainder of this paper.

Two-way ANOVA with Time and Concentration shows Concentration as well as Time is a significant factor. This is a useful opportunity to discuss the increased power of this test compared with the one-way ANOVA. The other combinations of two variables can be investigated.

Including all three factors in a General Linear Model, as an additive model, indicates that all are significant. Tukey pairwise comparisons show that the Temperature 27 degrees gives significantly lower values of Count1 than the other temperatures, and that Concentration 1.2 gives significantly higher Count1 values for all other concentrations except 1.4.

Interactions can also be included in this model. For Count1 there is a significant interaction between Time and Temperature and between Time and Concentration. It is worth looking at the interaction plots and describing what these mean in the context of the problem. For Count1 the effect of Temperature is less critical when the culture is incubated for 48 hours. Note that this data includes the outlier that was discussed above, so this may not be true if this value were removed or changed. The effect of Concentration is dependent on Time, with the peak at 1.2% being much more pronounced at 48 hours than at 24 hours. The Ryan-Joiner Normality test on the residuals confirms the residuals may be normally distributed.

Another approach is to analyse all 5 strains together. Stack the data and create a new predictor variable 'Strain'. Then use ANOVA with Strain, Temperature, Time and Concentration as predictors including two and three-way interactions. Time*Temperature and Time*Concentration are still significant interactions. However, in the context of this problem, only one of these strains is likely to be present at one time, and the information needed is the optimal conditions for each strain separately.
Since all the predictor variables are continuous, regression is another tool to build a model of this situation. The result of the graphical analysis and the ANOVA suggest that there are non-linear relationships between the Count1 and both Temperature and Concentration, and that Time is a significant predictor.

For each strain, develop a multiple polynomial regression model that best fits the data. There is scope for variation in the models here, as decisions will be made about outliers and the degree of the polynomials chosen. This will lead to fruitful discussion about what is the ‘best’ model. It will be of interest later to see whether these choices have much effect on the optimum conditions calculated. Note that this is a fairly small dataset to be performing multiple regression on, but it is still useful as a teaching example.

In modelling Count1, the residual plots for Temperature and Concentration indicate quadratic and cubic terms respectively. When a quadratic term for Temperature and quadratic and cubic terms for Concentration are included, the studentized residual for point 9 is 3.3. Although this point is at the lowest temperature, it has the highest Count. This seems to be an anomaly, so this point may be deleted and a new model developed. Alternately, the value 263 could be replaced by 163, as this looks like the value it was likely to be. This may be the preferred action since 263 appears to be an inputting error, and changing its value, as opposed to deleting it, will preserve the balanced nature of the data. The option of changing data should generate interesting discussion. Here we would like to be able to look in the researcher’s notebook to see the value that was originally written down!

Bearing in mind the interaction terms in the ANOVA model and the interactions seen in the interaction plots, some interaction terms could be included in the regression model. For Count1, Time*Temperature is a significant predictor when added to the model above. If additionally Count1 for point nine is changed to 163, then R-Sq improves from 75.8% to 86.9%. The Ryan-Joiner test on the studentized residuals indicates normality is a reasonable assumption. Residual plots are satisfactory; perhaps a quartic term in Concentration could be introduced.

Partial differentiation can be used on each model to estimate the optimum temperature and concentration. For Count1 several possible models all give optimum Temperature in the range 37.2 – 37.3 and optimum Concentration 1.24. It is clear that 48 hours is always better than 24 hours. It is interesting to note that human body temperature is 37 degrees and that this is optimal for this bacteria growth also. In fact because of this many bacteria are routinely cultured at 37 degrees in the laboratory.

The results for the five strains can be compared, to recommend the conditions that will be best for culturing a sample when it is unknown which of the five strains may be present. Small deviations from the optimal conditions are not a serious problem, as the curves are fairly flat at the peaks so Count values will still be close to optimal. The recommended culturing conditions can now be compared to the original protocol.

It must be noted that it may not be practicable to wait 48 hours for the culture to grow in the broth, since it is important to identify a treatment as quickly as possible so the patient will not deteriorate even further. The data for 24 hours on its own could be investigated to find the optimal Temperature and Concentration if Time was to be limited to 24 hours.

Predicted Count values and prediction intervals for the Counts can be obtained for the optimum conditions decided. These could be compared with those of the original protocol conditions of 24 hours at 35 degrees with a 1.0% tryptone concentration. This would give a quantitative way of deciding how much has been gained by changing the protocol.
5. Getting the Data

The file Tryptone.dat.txt is a text file containing the raw data. The file Tryptone.txt is a documentation file containing a brief description of the dataset.

References

Cooper, G. (1999), "The Efficiency of the Recovery of Methicillin Resistant Staphylococcus aureus from Salt Enrichment Broth," Bachelor of Applied Science project, Department of Applied Science, Auckland University of Technology.

_Bacteriological Analytical Manual Online, 2001._ [www.cfsan.fda.gov/~ebam/bam-toc.html](http://www.cfsan.fda.gov/~ebam/bam-toc.html)

---

Neil Binnie  
Department of Applied Mathematics  
Auckland University of Technology  
Private Mail Bag 92006  
Auckland 1020  
New Zealand  
[neil.binnie@aut.ac.nz](mailto:neil.binnie@aut.ac.nz)