International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Diagnosis for Leverage and Influence in Mixture Experiments for Engine Management System

Chetan[1], **Ratna Raj Laxmi**[2]

[1]Research Scholar, Department of Statistics, M.D. University, Haryana, India

[2]Professor, Department of Statistics, M.D. University, Haryana, India

Abstract: ***Engine management system not only involved the manufacturing of engines but also involved the science which provides efficient performance of any engine. The mixture experiments play an important role especially when our aim is to increase the octane rating of the engine. These experiments involved the empirical prediction of the responses to any mixture of the components when the response depends only on proportions of the fuel components but not on the total amount of fuel mixture. In this paper, we present the study to show the importance of these experiments in engine management system and diagnosis techniques for leverage and influential points for these experiments.***

Keywords: Mixture Experiments, Leverage Points, Influential Points, Least Square Method, Octane Rating, etc.

1. Introduction

Since the landmark articles by Scheffe (1958, 1963), researchers have proposed many criteria for evaluating and comparing mixture experiments. The purpose of these designs is the empirical prediction of the response to any mixture of the components when the response depends only on the proportions of the components but not on the total amount of mixture. In this type of experiment, the quality of the end product depends on the relative proportions of the components in the mixture. If we denote the number of components in the mixture by q and the proportion contributed by x_i for the i^{th} component $(i = 1,2, \dots \dots, q)$, the following constraints apply to the mixture component proportions

$$0 \leq x_i \leq 1$$

$$\sum_{i=1}^{q} x_i = 1$$

The experimental region defined by these by constraints is a $(q-1)$ dimensional simplex. Often further constraints are imposed on the mixture components that in a design with the shape of a simples. When additional constraints are imposed on the component proportional in the form of lower and upper bounds

$$0 < L_i \leq x_i \leq U_i < 1$$

or as the linear multicomponent constraints,

$$C_j \leq A_{1j}x_1 + A_{2j}x_2 + \dots \dots \dots + A_{qj}x_q \leq D_j$$

where the A_{ij}, C_j and D_j are scalar constants, these additional constraints may alter the shape of the experimental region from that of a simplex to one of an irregularly shaped convex polyhedron inside the simplex.

Mixture model forms most commonly used fitting mixture data are the canonical polynomials introduced by Scheffe (1958, 1963). The first degree or liner blending model is

$$Y = \sum_{i=1}^{q} \beta_i x_i + \varepsilon$$

where$Y$ represents the observed value of the response, the coefficient β_i is the expected response to component i and ε is the random error in the observed response value having expectation zero and variance σ^2. The second-degree or binary nonlinear blending model is

$$Y = \sum_{i=1}^{q} \beta_i x_i + \sum_{i<j}\sum \beta_{ij} x_i x_j + \varepsilon$$

whereβ_{ij} is a measure of nonlinear blending of components i and j.

These mixture models play an important role in engine management system especially when we talk about the octane rating. It is a standard measure of the performance of a motor or aviation fuel. The higher the octane number the more compression the fuel can withstand before detonating. The motor octane rating or motor octane number (MON) is determined at 900 rpm engine speed instead of 600 rpm for research octane number (RON). They are determined by running the fuel in a test engine with a variable compression ratio under controlled conditions and comparing the results with these for mixtures of iso-octane. Depending on the composition of the fuel, the MON of a modern engine gasoline will be about 8 to 12 octane lower than the RON, but there is no direct link between RON and MON. These engines specifications typically require both a minimum RON and minimum MON.

Cornell (2002) mentioned the data for motor octane ratings from 12 different blends in an effort to determine the effects of the following gasoline blending components with the specified ranges

Straight run (x_1): $0 \leq x_1 \leq 0.21$
Reformate (x_2): $0 \leq x_2 \leq 0.02$
Thermally cracked naphtha (x_3): $0 \leq x_3 \leq 0.12$
Catalytically cracked naphtha (x_4): $0 \leq x_4 \leq 0.62$
Polymer (x_5): $0 \leq x_5 \leq 0.12$
Alkylate (x_6): $0 \leq x_6 \leq 0.74$
Natural Gasoline (x_7): $0 \leq x_7 \leq 0.08$

The response values and the component setting are presented in the Table 1.1

Table 1.1: Gasoline Motor Octane Ratings

S.No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
1	0.00	0.23	0.00	0.00	0.00	0.74	0.03	98.70
2	0.00	0.10	0.00	0.00	0.12	0.74	0.04	97.80
3	0.00	0.00	0.00	0.10	0.12	0.74	0.04	96.60
4	0.00	0.49	0.00	0.00	0.12	0.37	0.02	92.00
5	0.00	0.00	0.00	0.62	0.12	0.18	0.08	86.60
6	0.00	0.62	0.00	0.00	0.00	0.37	0.01	91.20
7	0.17	0.27	0.10	0.38	0.00	0.00	0.08	81.90
8	0.17	0.19	0.10	0.38	0.02	0.06	0.08	83.10
9	0.17	0.21	0.10	0.38	0.00	0.06	0.08	82.40
10	0.17	0.15	0.10	0.38	0.02	0.10	0.08	83.20
11	0.17	0.36	0.12	0.25	0.00	0.00	0.06	81.40
12	0.00	0.00	0.00	0.25	0.00	0.37	0.08	88.10

This situation truly defines the mixture experiment under the restricted region (with lower and upper bound restrictions). These types of experiments are very common in the field of engine management especially when our aim is to increase the octane rating of the engine. The octane rating of the engine is increased by using proper blending of the components. Therefore, it is mandatory that these blending should be proper. When, in statistics, we talk about the proper blending of component, it means, they should not be leverage and/or influential point.

2. Diagnostics for Leverage and Influence

As we know that each observation in the sample has the same weight in determining the outcome. In the regression situation, this is not the case. The location of observation in x-space may affect the regression coefficients. Therefore, our basic aim is too focused on these observations so that our fitted model should be adequate. It is generally observed that outliers are often identified by unusually large residuals and that these observations can also affected the regression results. It is important to use diagnostics for leverage and influence in conjunction with the residual analysis techniques. Sometimes we find that a regression coefficient may have a sign that does not make engineering or scientific sense, a regressor known to be important may be statistically insignificant, or a model that fits the data well and that is logical form an application-environment perspectives may produce poor predictions. These situations may be the result of one or perhaps a few influential observations. Due to the importance of leverage and influential observations we use different diagnostics here.

For understanding the basic behind the leverage and influence point consider the following fingers.

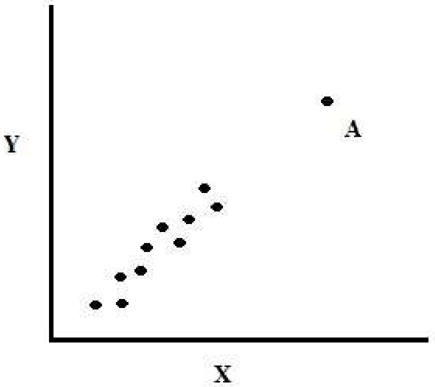

Figure 1: An example of leverage point

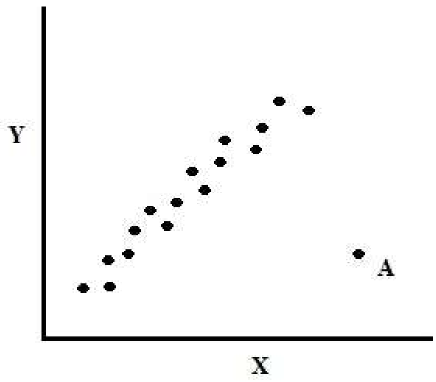

Figure 2: An example of an influential observation

The point labeled A in Figure 1 is remote in x-space from the rest of the sample, but it lies almost on the regression line passing through the rest of the sample points. This is an example of a leverage point i.e. it has an unusual x-value and may control certain model properties. This point does not affect the estimates of the regression coefficients but it certainly will have a dramatic effect on the model summary statistics such as R^2 and the standard error of the regression coefficient. But the point labeled A in Figure 2, this point has a moderately unusual x-coordinate and the y value is unusual as well. This is an influence point, i.e. it has a noticeable impact on the model coefficients in that it pulls the regression model in its direction. In an extreme case, the parameter estimates may depend more on the influential subset of points then on majority of the data. This is obviously an undesirable situation we would like for a regression model to be representative of all of the sample observations not an artifact of few. Consequently, we would like to find these influential points are indeed bad values then they should be eliminated from the sample. On the other hand, there may be nothing wrong with these points, but if they control key model properties we would like to know it as it could affect the end use of the regression model.

As remote points potentially have disproportionate impact on the parameter estimates, standard error, predicted value and model summary statistics. The hat matrix

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

plays an important role in identifying influential observations. The hat matrix $\boldsymbol{H}$ determines the variances and covariance of $\hat{y}$and $\boldsymbol{e}$, since $\mathrm{var}(\hat{y}) = \sigma^2\boldsymbol{H}$and $\mathrm{var}(\boldsymbol{e}) = \sigma^2(\boldsymbol{I}-\boldsymbol{H})$. The elements h_{ii} of the matrix $\boldsymbol{H}$, which may be given as

$$h_{ii} = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$$

where$\boldsymbol{x}_i'$is the i^{th} row of the $\boldsymbol{X}$ matrix. The hat matrix diagonal is a standardized measure of the distance of the i^{th} observation from the center of the x-space. Thus, large hat diagonals reveal observations that are potentially influential because they are remote in x-space from the rest of the sample. It turns out that the average size of the hat diagonal is $\bar{h} = \frac{p}{n}$, because $\sum_{i=1}^{n} h_{ii} = rank(\boldsymbol{H}) = rank(\boldsymbol{X}) = p$ and traditionally assume that any observation for which the hat diagonal exceeds twice the average $\frac{2p}{n}$ is remote enough from the rest of the data to be considered a leverage point.

Cook (1977, 1979) has suggested a way to measure influence using a measure of the squared distance between the least-squares estimate based on all n points $\widehat{\boldsymbol{\beta}}$ and the estimate observed by deleting the i^{th} point, say $\widehat{\boldsymbol{\beta}}_{(i)}$. This distance measure can be expressed as

$$D_i(\boldsymbol{X}'\boldsymbol{X}, pMS_{Res}) \equiv D_i = \frac{\left(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}\right)' \boldsymbol{X}'\boldsymbol{X}\left(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}\right)}{pMS_{Res}}$$

$$= \frac{r_i^2}{p} \frac{Var(\hat{y}_i)}{Var(e_i)} = \frac{r_i^2}{p} \frac{h_{ii}}{(1 - h_{ii})}$$

Apart from the constant p, D_i is the product of the square of the i^{th}studentized residual and$\frac{h_{ii}}{(1-h_{ii})}$. This ration can be shown to be the distance from the vector $\boldsymbol{x}_i$to the centroid of the remaining data. Thus, D_i is made up of a component that measures how well the model fits the i^{th} observation y_i and a component that measures how far that point is from the rest of the data.

The magnitude of D_i is usually assessed by comparing it to $F_{\alpha,p,n-p}$. If $D_i = F_{\alpha,p,n-p}$, then deleting point i would move $\widehat{\boldsymbol{\beta}}_{(i)}$boundary of an approximate 50% confidence region for$\boldsymbol{\beta}$ based on the complete data set. This is a large displacement and indicates that the least-squares estimate is sensitive to i^{th} data point. Since $F_{\alpha,p,n-p} \cong 1$, we usually consider points for which $D_i > 1$ to be influential.

As Cook's distance measure is a deletion diagnostic, that is, it measures the influence of the i^{th} observation if it is removed from the sample. In similar scenario Belsley, kuh and Welsch (1980) introduced another measure of deletion influence of i^{th} observation on the predicted or fitted values and given by

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S_{(i)}^2 h_{ii}}} = \left(\frac{h_{ii}}{1 - h_{ii}}\right)^{1/2} t_i$$

where$\hat{y}_{(i)}$ is the fitted value of y_i obtained without the use of the i^{th} observation. The denominator is just standardization, since $\mathrm{Var}(\hat{y}_i) = \sigma^2 h_{ii}$. Thus $DFFITS_i$ is the number of standard deviations that the fitted value $\hat{y}_i$ changes if observation i is removed. In fact $DFFITS_i$ is nothing but the value of R- student (t_i) multiplied by the leverage of the i^{th} observation$\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$. If the data point has high leverage then h_{ii} will be close to unity. However, if $h_{ii} \cong 0$, the effect of R- student will be moderated. Similarly a near-zero R-student combined with a high-leverage point could produce a small value of $DFFITS_i$. Thus, $DFFITS_i$ is affected by both leverage and prediction error. Belsley, kuh and Welsch (1980) suggested that any observation for which $|DFFITS_i| > 2\sqrt{p/n}$ warrants attention.

The diagnostics D_i and $DFFITS_i$ provide insight about the effect of observations on the estimated coefficients $\hat{\beta}_j$ted values $\hat{y}_j$. They do not provide any information about overall precision of estimation. Since it is fairly common practice to use the determinant of the covariance matrix as a convenient scalar measure of precision, called the generalized variance, which is given as

$$GV(\widehat{\boldsymbol{\beta}}) = \left|Var(\widehat{\boldsymbol{\beta}})\right| = \left|\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right|$$

To express the role of the i^{th} observation on the precision of estimation, we could define

$$COVRATIO_i = \frac{\left|\left(\boldsymbol{X}_{(i)}'\boldsymbol{X}_{(i)}\right)^{-1} S_{(i)}^2\right|}{(\boldsymbol{X}'\boldsymbol{X})^{-1} MS_{Res}} = \frac{\left(S_{(i)}^2\right)^p}{MS_{Res}^p}\left(\frac{h_{ii}}{1 - h_{ii}}\right)$$

Clearly if $COVRATIO_i > 1$, the i^{th} observation improves the precision of estimation, while $COVRATIO_i < 1$, inclusion of the i^{th} point degrades precision. Also Belsley, kuh and Welsch (1980) suggest that if $COVRATIO_i > 1 + {}^{3p}/_{n}$ or if $COVRATIO_i < 1 - {}^{3p}/_{n}$, then the i^{th} point should be considered influential. The statistics for the leverage and influential observations for the octane ratings data are given in the Table 2

Table 2: Statistics Obtain for Octane Rating Data

Observations	h_{ii}	D_i	$DFFITS_i$	$COVRATIO_i$
	(a)	*(b)*	*(c)*	*(d)*
1	0.628536	0.265038	1.378809	2.2691961
2	0.472832	0.098846	0.809023	2.6988824
3	0.492866	0.112731	0.868150	2.6197794
4	0.740512	0.650773	1.513754	1.8448953
5	0.961314	38.32691	13.60639	3.9425280
6	0.699843	0.460353	1.487539	1.7998073
7	0.276131	0.031705	0.448251	2.6710209
8	0.254814	0.027668	0.418015	2.6581129
9	0.255736	0.027793	0.418949	2.6623719
10	0.268411	0.030251	0.437618	2.6619701
11	0.900000	0.543591	0.404068	1.9448353
12	0.949004	21.78343	13.81912	4.0579872

3. Conclusion and Discussion

Engine management system not only involved the manufacturing of engines but also involved the science which provides efficient performance of any engine. The performance of the engine can be measured by their octane rating. This rating of engines is totally depends on the combination of fuel which is used to run it. The mixture experiments play an important role especially when our aim is to increase the octane rating. These experiments involved the empirical prediction of the responses to any mixture of the components when the response depends only on proportions of the fuel components but not on the total amount of fuel mixture. Because of our aim, it is mandatory that fuel blending should be proper. In statistics, the blending should be proper, it means, there is no leverage and/or influential points. Due to this reason the diagnostic for these points should be important so that our fitted model should be adequate. The statistics in column b of Table 2 contains the value of Cook's measure. The larger values of the D_i statistics are $D_5$and D_{12}, which indicate that deletion of observation 5 and 12 would move the least square estimate. Therefore, we would conclude that observation 5 and 12 are definitely influential using the cutoff of unity. Inspection of column c also reveals that both points 5 and 12 are influential. Column d contains the values of $COVRATIO_i$ and their cutoff values are 1.75 and 2.75. Note that the values of $COVRATIO_5$ and $COVRATIO_{12}$ exceed these limits. Both observation tends to improve the precision because their$COVRATIO > 1$. Hence we can say that observation 5 and 12 are influential and change our model statistics.

Reference

[1] Belsly, D.A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York.

[2] Cook, R. D. (1977). Detection of influential observation in linear regression, Technometrics, **19**, pp 15-18.

[3] Cook, R. D. (1979). Influential observation in linear regression, Journal of American Statistics Association, **74**, pp 169-174.

[4] Scheffé, H. (1958). Experiments with mixtures. *Journal of the Royal Statistical Society*, **B20(2)**, pp344-360.

[5] Scheffé, H. (1963). Simplex-centroid design for experiments with mixtures. *Journal of the Royal Statistical Society*, **B25(2)**, pp235-263.