Sains Malaysiana 50(12)(2021): 3733-3744
http://doi.org/10.17576/jsm-2021-5012-22

Optimised Reduction of Surgical Gloves Pinholes using Forward Search Method

(Penurunan Teroptimum Lubang Jarum pada Sarung Tangan Pembedahan menggunakan Kaedah Pencarian Terkehadapan)

AZIMAH AHMAD, NUR ANISAH MOHAMED @ A. RAHMAN* & ZAHARAH WAHID

ABSTRACT

This research investigates the factors that affect the existence of pinholes in surgical gloves during the manufacturing process. Since eight factors affect the existence of pinholes in surgical gloves, a two-level fractional factorial design 2^{8-4} was used to study the main effects and the first-order interactions of the multiple variables. Multiple linear regressions are used to model the data. This paper also examines the presence of influential points in the data using the influential measures in linear regression such as Cook's Distance, DFFITS, DFBETAS, Studentized Residual, Standardized Residual, Hadi's measure, and the robust forward search. The impact of influential points is further assessed through deletion of potential influential points and model selection using adjusted R2, information criterion, and stepwise selection to see whether these influential points significantly improved the existing model.

Keywords: Fractional factorial design; influential points; multiple linear regression; robust forward search; stepwise selection

ABSTRAK

Kertas penyelidikan ini mengkaji faktor-faktor ketika proses pembuatan yang menyebabkan kewujudan liang jarum pada sarung tangan pembedahan. Oleh kerana terdapat lapan faktor yang menyebabkan kewujudan liang jarum pada sarung tangan pembedahan, reka bentuk dua-peringkat faktorial pecahan 2^{8-4} digunakan untuk mengkaji kesan-kesan utama dan interaksi peringkat pertama pelbagai pemboleh ubah ini. Regresi linear berganda dengan terma interaksi digunakan untuk memodelkan data. Kertas penyelidikan ini turut mengkaji kewujudan titik-titik berpengaruh dalam data dengan menggunakan ukuran berpengaruh dalam regresi linear seperti Jarak Cook, DFFITS, DFBETAS, reja piawai, reja studentized, Ukuran Hadi dan aturan Pencarian Kedepan yang mantap. Impak titik-titik berpengaruh ini terus dinilai melalui penghapusan titik-titik berpengaruh yang berpotensi dan pemilihan model menggunakan R^2 yang disesuaikan, kriteria informasi dan pemilihan bertahap untuk melihat sama ada titik-titik berpengaruh dapat meningkatkan model sedia ada dengan ketara.

Kata kunci: Pemilihan bertahap; pencarian kedepan mantap; reka bentuk faktorial pecahan; regresi linear berganda; titik-titik berpengaruh

INTRODUCTION

A pinhole defect in surgical rubber gloves may pose higher risks of infection in both healthcare workers and patients. Gessler et al. (2011) believed that gloves are a potential source of contamination due to their thin and highly flexible nature which may lead to rupture or puncture. Thus, glove sensitivity test procedures are developed to detect pinholes in gloves. Malaysian Rubber Glove Manufacturers' Association (MARGMA) and the Malaysian Rubber Board (MRB) have formulated Standard Malaysian Glove (SMG) certification requiring manufacturers to produce gloves with defective pinholes rate known as Acceptance Quality Level (AQL) of 1.5 (Ong et al. 2001). This requirement follows the European Standards EN-455-125 (Patel et al. 2003).

The manufacturing process usually involves several factors. For example, in Wahid (1998), there are eight factors were identified which are curing temperature profile (A), latex temperature in dip tank (B), oven temperature before coagulation (C), % of calcium nitrate (D), humidity (E), % of calcium carbonate (F), oven temperature before latex dip (G) and PH of latex compound

(H). For a process with eight factors, it is necessary to conduct 28 or 256 experiments for a complete factorial matrix. Due to the consideration of limited resources and the costs of running all experiments, it will be advantageous to use the fractional factorial design to reduce the number of runs required.

According to the Malaysian Rubber Export Promotion Council (MREPC), Malaysia is a leading rubber glove producer in the world that supplies more than 50% of the world's demand for medical gloves (MREPC 2020a, 2020b). In 2019, Malaysia exported RM 1.57 billion worth of surgical gloves to meet the global demand. Hence, it is of interest for researchers to conduct a study relating to pinholes in surgical gloves to improve the standard and quality of surgical gloves produced. Lowering the 'curing temperature profile' and 'latex temperature in dip tank' while increasing the 'oven temperature before latex dip' and high humidity with low latex temperature will minimize the pinhole defect (Wahid 1998; Wahid & Tham 2012). In both of the studies, the two levels saturated fractional factorial design was used to explain the controllable factors that significantly affect the pinhole defect in surgical gloves and multiple regression techniques are applied to the data. Using the Wilcoxon signed-rank tests, Tan et al. (2020) have found the temperature of latex and oven temperature after coagulant dip will affect the pinhole's defect. However, these studies have excluded former oven temperature, humidity, calcium carbonate, and latex pH from the examined factors. Jirasukprasert et al. (2014) have optimized the oven's temperature and conveyor's speed using six sigma methods to reduce defects such as holes and stains in rubber gloves.

Materials and Methods

Fractional Factorial Design

The fractional factorial design is a modified standard factorial design that provides information on main effects and low-order interactions without the need to conduct a full factorial design. It is advantageous to use fractional factorial design if we have limited resources, note that the high order interactions are not significant, and plan to do a screening experiment, i.e., intend to find significant factors (Oehlert 2000). A 2^k factorial can be confounded into 2^q blocks of size 2^{k-q}. A 2^{k-q} fractional factorial consists of k factors each at two levels that use 2^{k-q} experimental units and factor level combination. Fractional factorials are categorized according to their resolution that shows the types of effects that are aliased. A design with resolution R will have no j factors interaction that is aliased to fewer than R-j factors interaction (Oehlert 2000).

Multiple Linear Regressions

Suppose that y is a single dependent variable or response variable that depends on independent or regressor variables such as $x_1, x_2, ..., x_k$. A multiple linear regressions model is a mathematical model that characterizes the relationship between these variables. The equation for multiple linear regression models with k independent variables is:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon_i \qquad (1)$$

where $i = 1, 2, \ldots, n$. The parameter β is the regression coefficient where β_0 is the intercept term, k refers to the number of independent variables and ϵ is the error term which is also known as the residuals. Multiple linear regression techniques can also be used to analyze a model with interaction term such as the following first- order model in two variables where interaction term is added:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon_i. \qquad (2)$$

If we let $x_1 x_2 = x_3$ and $\beta_{12} = \beta_3$ in (2), we will get

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon_i. \qquad (3)$$

The method of least squares will be used to estimate the regression coefficients β_k. Provided there are observations, then (1) can be written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i. \qquad (4)$$

We can write (4) in matrix notation as $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, where,

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix}, = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_k \end{bmatrix}, \text{ and } \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{bmatrix}.$$

Outlier

An outlier is a data point whose response y, does not follow the general trend of the rest of the data. The outlier may represent data-recording error or a poor approximation of the actual model (Montgomery 2009). Standardized residuals and studentized residuals are some of the methods that can be utilized to detect outliers.

STANDARDIZED RESIDUALS

Standardized residual is defined as ordinary residual divided by an estimate of its standard deviation:

$$d_i = \frac{e_i}{\hat{\sigma}}$$

where the residual e_i, is the difference between the actual observation y_i and the corresponding estimated value $\hat{y}_i$ (i.e., $y_i - \hat{y}_i$) and the estimate of standard error $s, \hat{\sigma} = \sqrt{MS_E}$. The Mean Squared Error MS_E is the estimate of variance, $\hat{\sigma}^2$. The standardized residual has mean zero and approximately unit variance that is useful in detecting outliers. Any observation with standardized residual outside the interval $-2 \leq d_i \leq 2$ is considered as an outlier that should be carefully examined (Montgomery 2009).

STUDENTIZED RESIDUALS

Studentized residual is defined as:

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}$$

for $i = 1, 2, \ldots, n$. If the form of the model is correct, the studentized residual has variance $V(r_i) = 1$ regardless of the location of x_i. Any studentized residual observation outside the interval $-3 \leq r_i \leq 3$ is considered an outlier..

LEVERAGE

An observation with extreme predictor x a value is considered as high leverage. Leverage of the i^{th} case is defined as:

$$h_{ii} = [\boldsymbol{H}]_{ii} = \boldsymbol{x}_i^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_i,$$

where H is the i^{th} diagonal element of the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}$. Any observation with leverage higher than $\frac{2k}{n}$ (where k is the number of independent variables) will be considered as high leverage (Montgomery 2009).

INFLUENTIAL POINTS

A data point is considered influential if it unduly influences any part of regression analysis, such as the predicted responses, the estimated slope coefficients, or the hypothesis test results. Outliers and high leverage data points have the potential to be influential, but we generally have to investigate to determine whether they are influential. It is noted that influential measures are employed to detect influential points.

INFLUENTIAL MEASURES

Many types of influential measures are useful in identifying influential observations. The influential measures employed in this study are Cook's Distance, DFFITS, DFBETAS, and Hadi's Measure.

COOK'S DISTANCE

Cook (1977) introduced Cook's Distance that is made up of components that show how good the model fits the i^{th} observation of y_i and how far is the observation from the rest (Montgomery 2009). The Cook's Distance is defined as:

$$D_i = \frac{r_i^2}{k}\frac{h_{ii}}{(1 - h_{ii})},$$

where $i = 1, 2, \ldots, n$.

Note that Cook's Distance is expressed as a function of studentized residual and the leverage where k is the number of independent variables. Montgomery (2009) highlighted that any observation which $D_i > 1$ is considered as an influential point. Cook (1977) pointed that the cut-off for Cook's Distance is $D_i > 4/n$ and data point with large D_i strongly influences the fitted value of the model.

DFFITS

The difference in fits for observation i, denoted DFFITS_i, is defined as:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}} = r_{i(i)}\sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

where $\hat{y}_{i(i)}$ is the estimated value of y without observation i $\hat{\sigma}_{(i)}$ is the estimated standard error without a point i, h_{ii} is the leverage, and $r_{i(i)}$ is the studentized residual without a point i. According to Belsley et al. (1980), an observation is deemed influential if the absolute value of its DFFITS value is greater than $2\frac{\sqrt{(k+1)}}{(n-k-1)}$, where, n is the number of observations and k is the number of predictor terms.

DFBETAS

Belsley et al. (1980) suggested using the DFBETAS statistics to measure the changes in each regression coefficient. It is calculated by deleting the i^{th} observation. is defined in the following equation:

$$DFBETAS_{ij} = \frac{\beta_j - \beta_j(i)}{\hat{\sigma}(i)\sqrt{(\boldsymbol{X}^T\boldsymbol{X})_{jj}^{-1}}}$$

where $\beta_j(i)$ is the regression coefficient computed without the i^{th} observation; $\hat{\sigma}_{(i)}$ is the estimated standard error without the i^{th} observation; $\mathbf{X}$ is (n x k) matrix of the independent variables; and $\mathbf{X^T X_{jj}}^{-1}$ is the $(jj)^{th}$ element of $(\mathbf{X^T X})^{-1}$. There are different DFBETAS plots for each term in the model. Usually, observations with a high value of DFBETAS are considered influential points. A general cut-off value of 2 indicates influential points and $\frac{2}{\sqrt{n}}$ is a size-adjusted cut-off.

HADI'S MEASURE

Hadi (1992) introduced another influential measure with the following formula to measure the influence of the i^{th} observation:

$$H_i = \frac{h_{ii}}{1-h_{ii}} + \frac{k+1}{1-h_{ii}}\frac{a_i^2}{1-a_i^2}, \qquad i = 1,2,\dots,n,$$

where $a_i^2 = \frac{e_i^2}{SSE}$ is the square of the i^{th} normalized residual; h_{ii} is the leverage of the ith case; and k is the number of independent variables. An influential point will have large values of H_i.

FORWARD SEARCH METHOD

The objective of the forward search is to find observations that are different in the data and to determine the effect of this observation on inferences made about a model (Atkinson et al. 2012). Initially, the small size of subset, m_0 from n observations is robustly selected. For example, take $m_0 = k$, the number of parameters in the regression model. Subsequently, a sample of 1,000 subsets to each of which regression is fitted by least- squares and the median of the n squared residuals need to be computed (Atkinson & Riani 2000). The starting subset will be $S^*(m_0)$ that provides the smallest median squared residual. Then, a larger subset is considered by ordering the n squared residuals from the least-squares fit to the subset $S^*(m)$ of m observations and using the $m + 1$ observations with the smallest squared residuals to form $S^*(m+1)$ (Atkinson & Riani 2000). As a result, a series of parameter estimates for $k \leq m \leq n$ is obtained. The parameter estimates are consistent as m increases if there is no outlier. At the end of the search, observations that are remote from the fitted model will be included. These points can be possible outliers or influential points. Since our data is quite small, we can search over all subsets of size k of the n observations. In this forward search regression, the progress of S^2, the estimate of the error variance, can also be monitored (Atkinson & Riani 2000). The forward search function for the linear model can be obtained using the forward package in R.

THE DATA AND MISSING VALUE TREATMENT

The surgical gloves data were obtained from Wahid (1998) in a locally owned Malaysian rubber glove company for her doctoral dissertation. The Surgical Gloves data consists of 32 datasets with 15 independent variables (Variables A-H and 7 interaction terms AB, AE, BD, BE, BF, BG, and CD) and 3 response variables (Y1-Y3). In this paper, we focus on the response variable, Y3 (Mean of Pinholes). Hence, the other response variables Y1 and Y2 were excluded from this study.

There are four missing values in the response variable of Y3 due to run number 2, 5, 7, and 8 of the second replicate (or observation number 18, 21, 23, and 24) of the experiment, suspected the water to the coagulant tank promotes the stirrer speed and create bubbles in the tank. These four missing values are imputed using Multivariate Imputation via Chained Equation (MICE) - MIDAStouch package in R statistical software. MICE have emerged as a method of addressing missing data and the missing data are Missing at Random (MAR). In MAR, the probability that a value is missing depends on the observed value only (Azur et al. 2011). Missing values are imputed based on the observed values for a given observation and the relations observed in the data for other observations, assuming the observed variables are included in the imputation model (Schafer & Graham 2002). MIDAS, an acronym for Multiple Imputation using Distance Aided Selection of Donors, is a method described in Siddique and Belin (2008) which can handle a variety of data types and has desirable features of Bayesian approaches such as the ability to reflect parameter uncertainty, handle missing covariate values, incorporate all available information into the imputation model and allow the user to impute non-ignorable missing values. MIDAStouch is a new Predictive Mean Matching (PMM) procedure proposed by Gaffert et al. (2016) which explains that MIDAStouch is the best way in doing Multiple PMM imputations. MIDAStouch improved the MIDAS method by estimating a value of K in the data rather than using a fixed K and apply piece-wise linear function (PLF) correction to the total variance.

RESULTS AND DISCUSSION

As mentioned earlier, there are four missing values in this surgical gloves data. Since there are only 32 observations, we impute the missing values using the MICE-MIDAStouch in R software since the imputation is better than removing missing values (Lüdtke et al. 2017). The comparisons of F-statistics and $\overline{R^2}$ between the full linear regression model with removed missing values and full linear regression model with imputed missing values

using MICE-MIDAStouch, MICE-PMM, and MICE-mean, it is concluded that the full model with imputed missing values using MICE-MIDAStouch gives a better fit model. Hence, the comparisons of the statistical summary of these 4 models are shown in Table 1.

TABLE 1. Comparison of full model with and without imputation

Linear Models	Full model with removed missing values	Full model with imputation using MICE MIDAStouch	Full model with imputation using MICE mean	Full model with imputation using MICE PMM
Res Standard Error	0.07395 on 12 degrees of freedom	0.07943 on 16 degrees of freedom	0.0849 on 16 degrees of freedom	0.0834 on 16 degrees of freedom
Multiple R2	0.7227	0.6999	0.5128	0.5825
	0.3761	**0.4186**	0.056	0.191
F Statistics	2.085 on 15 and 12 DF,	**2.488 on 15 and 16 DF,**	1.123 on 15 and 16 DF	1.488 on 15 and 16 DF,
p-value	0.1028	0.0401	0.4093	0.2194

The full model with imputed missing values using MICE-MIDAStouch will be the null model. The null model is a multiple linear regression with interaction term that has the following formula:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_{12} x_1 x_2 + \beta_{15} x_1 x_5 + \beta_{24} x_2 x_4 + \beta_{25} x_2 x_5 + \beta_{26} x_2 x_6 + \beta_{27} x_2 x_7 + \beta_{34} x_3 x_4 + \epsilon \quad (5)$$

where y=Y3, x_1=A, x_2=B, x_3=C, x_4=D, x_5 = E, x_6=F, x_7=G and x_8=H.

The F-value and probability of F-value for each factor in the null model can be found in the Analysis of Variance (ANOVA) table. A variable with high F-value is considered to be a significant variable. From the analysis, the variable G gives the highest F-value of 10.41 followed by B(7.86), CD(4.38), AB(4.02), D(3.52) & BE(3.04). Hence, we can conclude that these seven variables are significant.

The response variable Y3 is a continuous variable whilst all other independent variables are categorical. The independent variables (A-H) are represented symbolically by a low level (-1) and a high level (+1) as presented in Table 2, which indicate the high and low level of each variable.

TABLE 2. Level of independent variables

Variable	Definition	Level	
		Low (-1)	High (+1)
A	curing temperature profile	80 °C	95 °C
		100 °C	110 °C
		115 °C	125 °C
		120 °C	130 °C
		130 °C	150 °C
B	latex temperature in dip tank	25-26 °C	29-30 °C
C	oven temperature before coagulation	75-80 °C	90-95 °C
D	calcium nitrate (%)	7.0-8.0%	11.0-12.0%
E	humidity	Low	High
F	calcium carbonate (%)	2.5-3.5%	4.5-5.5%
G	oven temperature before latex dip	170-180 °C	190-200 °C
H	PH of latex compound	10.1-10.5	10.6-10.9

A factorial design experiment with eight independent variables required 2^8=256 runs, where it is very costly and time-consuming. Since the resources are limited, it is more economical and timesaving to run a fractional factorial design that only used 16 runs. Hence, the L16 array is the suitable orthogonal array to be used for this experimental design. The fractional factorial design enables us to obtain information on main effects and low-order interactions without having to do the full factorial design. Hence, a 2^8 factorial design can be confounded into sixteen (2^4) blocks of size 16(2^4), which means that the 2^{8-4} fractional factorial consists of 16 factors at every two levels. The order of these 16 trials were run and randomized completely to reduce the effects of factors that were not controlled in the experimentation (Wahid & Tham 2012). These 16 trials run is replicated twice which means 32 experiments were conducted for the mean of the pinholes. Table 3 shows the fractional factorial design of the data. Please note that observation numbers 18, 21, 23, and 24 were imputed using MICE-MIDAStouch. (-1) in a column means that the factor listed above the column should be adjusted at its low-level while (1) indicates that the factor should be adjusted at its high level in that particular run. For example, all factors in the first trial run are set at low levels while in the second trial run, factors A, E, F, and G are set at high levels and factors B, C, D, and H are set at low levels.

TABLE 3. 32 Experiment runs on factors affecting mean of pinholes

Obs	A	B	C	D	E	F	G	H	Y3
1	-1	-1	-1	-1	-1	-1	-1	-1	0.2236
2 aefg	1	-1	-1	-1	1	1	1	-1	0.1000
3 befh	-1	1	-1	-1	1	1	-1	1	0.2000
4 abgh	1	1	-1	-1	-1	-1	1	1	0.2000
5 cegh	-1	-1	1	-1	1	-1	1	1	0.0000
6 acfh	1	-1	1	-1	-1	1	-1	1	0.2445
7 bcfg	-1	1	1	-1	-1	1	1	-1	0.1414
8 abce	1	1	1	-1	1	-1	-1	-1	0.4243
9 dfgh	-1	-1	-1	1	-1	1	1	1	0.2828
10 adeh	1	-1	-1	1	1	-1	-1	1	0.1000
11 bdeg	-1	1	-1	1	1	-1	1	-1	0.1732
12 abdf	1	1	-1	1	-1	1	-1	-1	0.2445
13 cdef	-1	-1	1	1	1	1	-1	-1	0.1000
14 acdg	1	-1	1	1	-1	-1	1	-1	0.0000
15 bcdh	-1	1	1	1	-1	-1	-1	1	0.2236
16 abcdefgh	1	1	1	1	1	1	1	1	0.1732
17	-1	-1	-1	-1	-1	-1	-1	-1	0.2000
18 aefg	1	-1	-1	-1	1	1	1	-1	0.0000
19 befh	-1	1	-1	-1	1	1	-1	1	0.2000
20 abgh	1	1	-1	-1	-1	-1	1	1	0.1732
21 cegh	-1	-1	1	-1	1	-1	1	1	0.2236
22 acfh	1	-1	1	-1	-1	1	-1	1	0.2828
23 bcfg	-1	1	1	-1	-1	1	1	-1	0.2444
24 abce	1	1	1	-1	1	-1	-1	-1	0.4223
25 dfgh	-1	-1	-1	1	-1	1	1	1	0.0000
26 adeh	1	-1	-1	1	1	-1	-1	1	0.2000
27 bdeg	-1	1	-1	1	1	-1	1	-1	0.1000
28 abdf	1	1	-1	1	-1	1	-1	-1	0.2445
29 cdef	-1	-1	1	1	1	1	-1	-1	0.1732
30 acdg	1	-1	1	1	-1	-1	1	-1	0.1000
31 bcdh	-1	1	1	1	-1	-1	-1	1	0.1000
32 abcdefgh	1	1	1	1	1	1	1	1	0.2236

We used statistical software Minitab to perform the fractional factorial design of the null model. From Figure 1, we conclude that the factors B and G are significant based on the Normal Plot while Pareto Chart shows the factors CD, AB, D, and BE are the next significant factors after G and B.

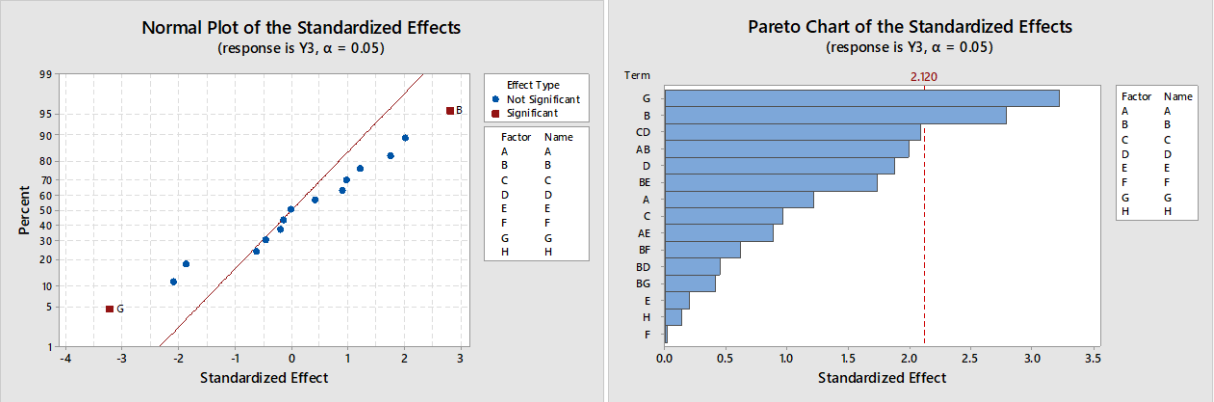

FIGURE 1. Normal Plot & Pareto Chart of the Standardized Effects for Y3

The stepwise selection was conducted suggested the following model (Model 1) as the best model. Model 1 has a linear regression formula as follows:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \epsilon_i \quad (6)$$

where y = Y3 is the response; x_1= A, x_2= B, x_3= D, x_4= G, x_5 = AB, x_6= BE and x_7 = CD are the predictors or independent variables and is the intercept.

In Table 4, the ANOVA table of the null model where the factors B, D, G, AB, and CD are significant variables because they have a smaller p-value that is lesser than 0.10. We use these significant variables B, D, G, AB, and CD to construct Model 2 that has the following linear regression formula:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon_i \quad (7)$$

where y_i= Y3 is the outcome, x_1= B, x_2= D, x_3= G, x_4= AB and x_5 = CD are the predictors or independent variables and β_0 is the intercept. The comparison of the statistical summary of the Null Model, Model 1, and Model 2 is shown in Table 4.

TABLE 4. Comparison of Null Model, Model 1 and Model 2

Linear Models	Null Model	Model 1	Model 2
ResidualStandard Error	0.07943 on 16 degrees of freedom	0.0699 on 24 degrees of freedom	0.07491 on 26 degrees of freedom
Multiple R2	0.6999	0.6514	0.5663
	0.4186	**0.5497**	0.4829
F Statistics	2.488 on 15 and 16 DF,	6.406 on 7 and 24 DF	**6.79** on 5 and 26 DF,
p-value	0.0401	0.0002574	0.0003588
AIC	-59.48	**-70.68**	-67.69
BIC	-34.56	**-57.49**	-57.43

Based on Table 4, Model 1 is found to be the best-fitted model with the highest $\overline{R^2}$ value and the lowest AIC and BIC values.

We further our investigation to diagnostic checking on Model 1 including using influential measures to find any influential points, leverage, or outliers. Figure 2 shows that Standardized Residual & Cook's Distance detects observation numbers 5, 9, and 25 as possible outliers. The second plot (top right) in Figure 2 shows that the Studentized Residual does not detect any possible outlier with the threshold (±3). If we reduce the threshold (±2), the same points highlighted in the first plot which are observation number 5, 9, and 25 will be detected as outliers. The third plot (bottom left) in Figure 2 highlighted there are no points with high leverage in the data. Based on the third plot, we can imply that there is a possibility that there is no influential point in Model 1 since there is no high leverage point.

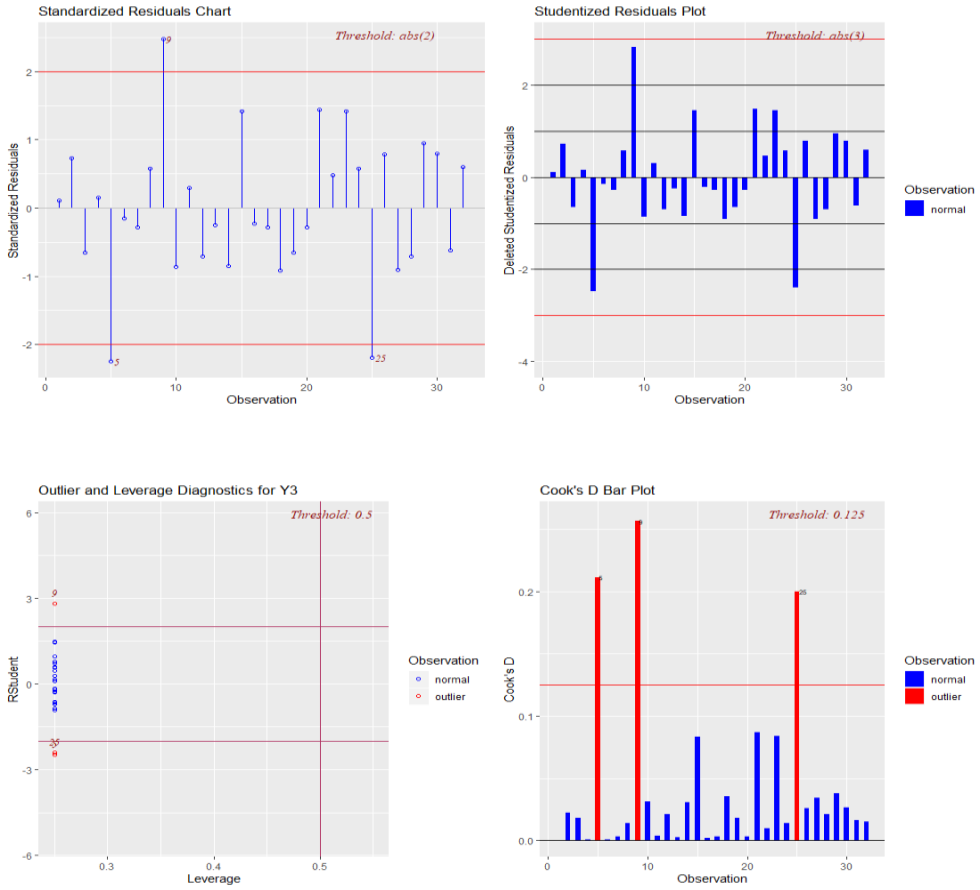

FIGURE 2. Standardized Residuals, Studentized Residuals, Outlier, Leverage and Cook's Distance for Model 1

Figure 3 shows that DFBETAS of the intercept and all factors in Model 1 also concluded that points number 5, 9, and 25 are possible influential points as shown in the first and second plot of Figure 3. The same observations number 5, 9, and 25 were detected by the DFFITS method and Hadi's measure as shown in the third and fourth plots of Figure 3.

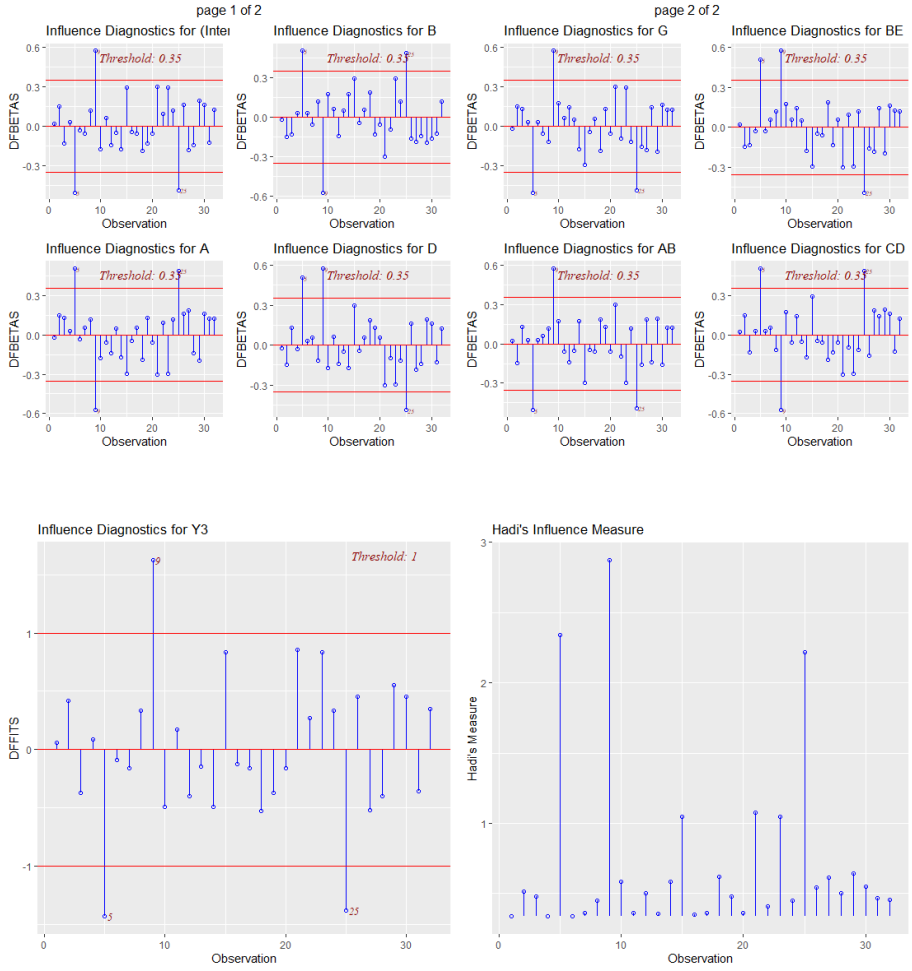

FIGURE 3. DFBETAS, DFFITS and Hadi's Measure of Model 1

Our investigation continues with the last 5 units that are included in the Forward Search method which are observations number 26, 27, 15, 25, and 5. The Scaled Residuals in Figure 4 pointed out that in addition to potential influential points that are observations number 5, 9, and 25 detected by the influential measures of linear regression, the Forward Search method also detects observation number 15 as a potential influential point as shown in the forward plot of least- squares residuals scaled by the final estimate of σ. Note that the last three units included in the Forward Search method, which are observations number 15, 25, and 5 are among the potential influential points shown in the Scaled Residuals in Figure 4.

The scaled Coefficient in Figure 4 provides evidence that there is a changed relationship in the estimated coefficients of Model 1. For example, β_2 and β_4 are initially stable, but they start to diverge after and respectively. In Figure 4, we can see that Cook's Distance value and S^2 value are increased when $m = 30$, 31, and 32, respectively. This indicates the values increases when the last 3 observations (5, 15 and 25) were included in the last step of the search.

We removed the potential influential points detected by the influential measures, i.e., observations number 5, 9 and 25 and define it as Model 1A. Another model by removing potential influential points detected by Forward Search method, i.e., observations number 5, 9, 15, and 25 from Model 1 and define it as Model 1B. The comparison of the statistical summary of Model 1, Model 1A and Model 1B is shown in Table 5.

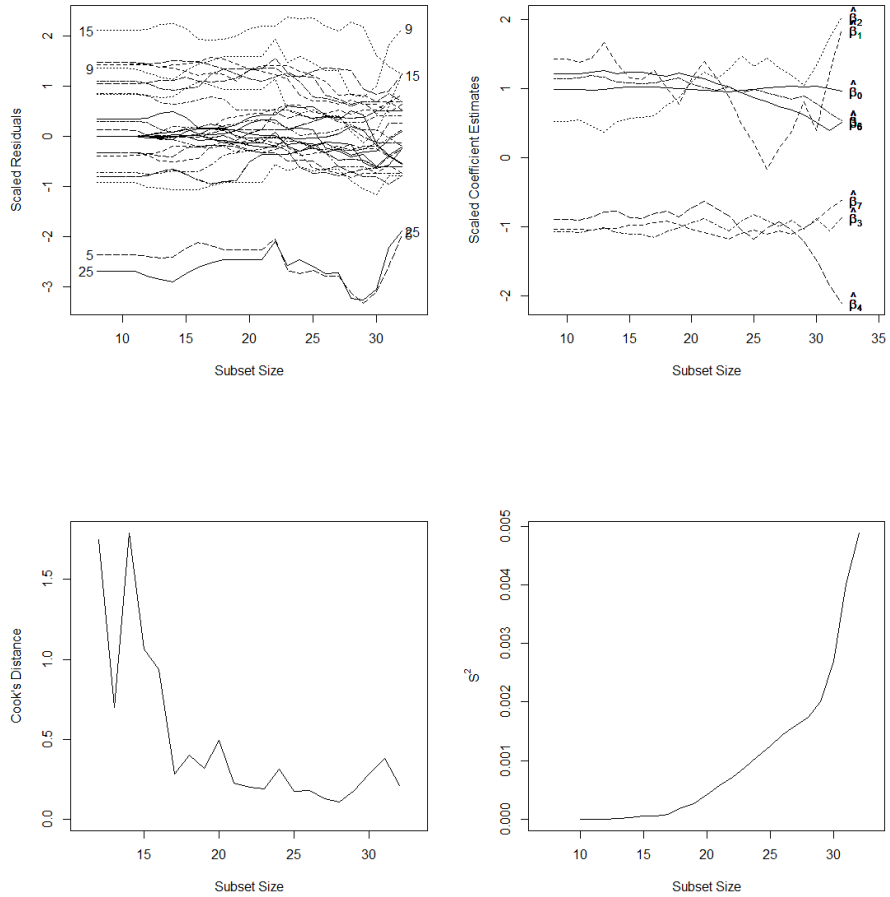

FIGURE 4. Scaled Residuals, Scaled Coefficient, Cook's Distance and S^2 of Forward Search on Model 1

TABLE 5. Comparison of Model 1, Model 1A and Model 1B

Linear Models	Model 1	Model 1A	Model 1B
Residual Standard Error	0.0699 on 24 degrees of freedom	0.04959 on 21 degrees of freedom	0.04323 on 20 degrees of freedom
Multiple R2	0.6514	0.8009	0.8552
$\overline{R^2}$	0.5497	0.7345	**0.8045**
F Statistics	6.406 on 7 and 24 DF	12.07 on 7 and 21 DF	**16.87** on 7 and 20 DF
p-value	0.0002574	4.166e-06	4.048e-07
AIC	-70.68	-83.29	**-87.87**
BIC	-57.49	-70.98	**-75.88**

Based on Table 5, we found Model 1B is the most reliable model since it has the highest $\overline{R^2}$ and F-statistics value and the lowest AIC and BIC values. This also indicates that observations numbers 5, 9, 15, and 25 are influential points. Hence, the final model is:

$$\widehat{Y3} = 0.184 + 0.012A + 0.025B - 0.034D - 0.031G + 0.042AB + 0.026BE - 0.044CD, \quad (8)$$

where $\widehat{Y3}$ is the estimated outcome; x_1= B, x_2= D, x_3= G, x_4= AB and x_5 = CD are the predictors. The residuals and coefficients of equation (8) are in Tables 6 and 7.

TABLE 6. Residuals of the Final Model

Min	1Q	Median	3Q	Max
-0.06	-0.03	-0.00	0.03	0.07

TABLE 7. Coefficients of the Final Model

	Estimate	Standard Error	T-Value
Intercept	0.184	0.009	21.653
A	0.012	0.009	1.370
B	0.025	0.009	2.858
D	-0.034	0.008	-4.128
G	-0.031	0.009	-3.534
AB	0.042	0.009	4.836
BE	0.026	0.008	3.061
CD	-0.044	0.009	-4.975

The $\widehat{Y3}$ is optimized when the process is set at high curing temperature (A), low latex temperature (B), high oven temperature before coagulation (C), high volume of calcium nitrate (D), high humidity (E), and high oven temperature before latex dip (G).

Conclusion

This study aimed to explain the relationship between the number of pinhole defects in surgical gloves and the factors cause during the manufacturing of surgical gloves. A two-level fractional factorial design 2^{8-4} was used to study the main effects and the first-order interactions of the multiple variables. Multiple linear regressions with interaction terms are used to model the data. Imputation using MICE-MIDAStouch is employed to provide values for missing data. The preliminary test shows that the model with imputation using MIDAStouch gives a better fit than the model without any imputation. Stepwise linear regression is then employed to find a model with the best fit.

We further investigate the presence of influential points in the linear model using influential measures in linear regression such as Cook's Distance, DFFITS, DFBETAS, Studentized Residual, Standardised Residual, and Hadi's Measure. The robust Forward Search is also conducted and applies to the model to find possible influential points. The impact of influential points is further assessed through deletion of potential influential points to see whether influential points that significantly change the fit of the model exist. We found that the Forward Search method provides better identification

of influential points. The final model from this study deduced that that the estimated $\widehat{Y3}$, is minimized when the manufacturing process of surgical gloves is set at high curing temperature (A), low latex temperature (B), high oven temperature before coagulation (C), high volume of calcium nitrate (D), high humidity (E) and high oven temperature before latex dip (G).

ACKNOWLEDGEMENTS

The authors would like to thank the University of Malaya for funding this research under a research grant (Grant No: BK023-2014).

REFERENCES

Atkinson, A.C. & Riani, M. 2000. *Robust Diagnostic Regression Analysis*. New York: Springer Science & Business Media.

Atkinson, A.C., Riani, M. & Cerioli, A. 2012. Problems and challenges in the analysis of complex data: Static and dynamic approaches. In *Advanced Statistical Methods for the Analysis of Large Data-Sets,* edited by Di Ciaccio, A., Coli, M. & Angulo Ibanez, J.M. Berlin, Heidelberg: Springer-Verlag. pp. 145-157.

Azur, M.J., Stuart, E.A., Frangakis, C. & Leaf, P.J. 2011. Multiple imputations by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research* 20(1): 40-49.

Belsley, D.A., Kuh, E. & Welsch, R. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* New York: John Wiley & Sons.

Cook, R.D. 1977. Detection of influential observation in linear regression. *Technometrics* 19(1): 15-18.

Gaffert, P., Meinfelder, F. & Bosch, V. 2016. https://www.uni-bamberg.de/fileadmin/uni/fakultaeten/sowi_lehrstuehle/statistik/Personen/Dateien_Florian/properPMM.pdf. Towards and MI-proper predictive mean matching.

Gessler, A., Stärk, A., Sigwarth, V. & Moirandat, C. 2011. How risky are pinholes in gloves? A rational appeal for the integrity of gloves for isolators. *PDA Journal of Pharmaceutical Science and Technology* 65(3): 227-241.

Hadi, A.S. 1992. A new measure of overall potential influence in linear regression. *Computational Statistics & Data Analysis* 14(1): 1-27.

Jirasukprasert, P., Garza-Reyes, J.A., Kumar, V. & Lim, M.K. 2014. A six sigma and DMAIC application for the reduction of defects in a rubber gloves manufacturing process. *International Journal of Lean Six Sigma* 5(1): 2-21.

Lüdtke, O., Robitzsch, A. & Grund, S. 2017. Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods* 22(1): 141-165.

Malaysian Rubber Export Promotion Council (MREPC). 2020a. *Malaysia's Exports of Selected Rubber Products.* http://www.mrepc.com/industry/malaysia_export.php.

Malaysian Rubber Export Promotion Council (MREPC). 2020b. *Standard Malaysian Glove: Product Specifications*. http://www.smgonline.biz/products_specifications.html.

Montgomery, D.C. 2009. *Introduction to Statistical Quality Control.* 6th Ed. New York: John Wiley & Sons.

Oehlert, G.W. 2000. *A First Course in Design and Analysis of Experiments.* New York: Freeman & Company.

Ong, E.L., Lai, P.F., Lim, C.L. & Ng, K.P. 2001. *Standard Malaysian Glove Scheme: Technical Requirements*. Asian Gloves and Dipped Goods Directory, Rubber Asia Publication. p. 59.

Patel, H., Fleming, G. & Trevor Burke, F. 2003. A preliminary report on the incidence of pre-existing pinhole defects in nitrile dental gloves. *British Dental Journal* 195(9): 509-512.

Schafer, J.L. & Graham, J.W. 2002. Missing data: Our view of the state of the art. *Psychological Methods* 7(2): 147-177.

Siddique, J. & Belin, T. 2008. Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in Medicine* 27(1): 83-102.

Tan, A.H., Cham, C.L. & Lim, E.H.Y. 2020. Analysis and prediction of glove quality based on manufacturing factors. In *2020 IEEE International Conference on Power and Energy (PECon)*. IEEE. pp. 420-425.

Wahid, Z. 1998. Potential for process improvement of the rubber glove manufacturing industrial case study process: An industry case study. University of Newcastle. Ph.D. Thesis (Unpublished).

Wahid, Z. & Tham, M.T. 2012. Single array and response modeling of robust design experiments. *International Journal of Applied Physics and Mathematics* 2(5): 359-361.

Azimah Ahmad & Nur Anisah Mohamed @ A.Rahman*
Institute of Mathematical Sciences, Faculty of Science
University of Malaya
50603 Kuala Lumpur, Federal Territory
Malaysia

Zaharah Wahid
Kulliyyah of Engineering
International Islamic University Malaysia
53100 Gombak, Selangor Darul Ehsan
Malaysia

*Corresponding author; email: nuranisah_mohamed@um.edu.my

Received: 18 September 2020
Accepted: 6 April 2021