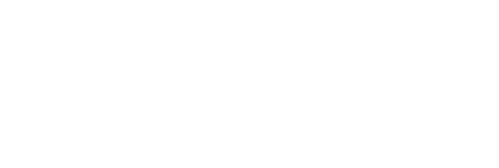


**ASSIGNMENT COVER SHEET**



Department of Mathematics and Statistics

# Student Name : Student ID : Unit Name : Lecturer’s Name : Due Date : Date Submitted : DECLARATION

I have read and understood Curtin’s policy on plagiarism, and, except where indicated, this assignment is my own work and has not been submitted for assessment in another unit or course. I have retained a copy of the assignment for my own records.

# \_ [Signature of student]

For Lecturer’s Use only

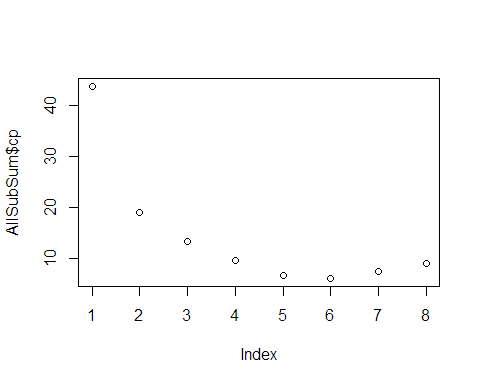
# Overall Mark: out of a total of Percentage:

Lecturer’s Comments:

# Use the Answer Guide below to answer each questions

1. Use an appropriate screening design to identify the significant input variables. Clearly specify the design used, the data simulated, subsequent analysis and significant variables.

The design used for this section was a Plackett Burman design, The reasoning behind this is due to using a run size of 12 which is a multiple of 4 but not a factor of 2, otherwise a fractional factorial design could have been used.

## X1 X2 X3 X4 X5 X6 X7 X8 outMtx  
## 1 0.15 50000 63070 1100 116.0 820 1120 9855 193.435074  
## 2 0.05 50000 115600 900 116.0 820 1680 9855 8.078586  
## 3 0.15 100 115600 1100 63.1 820 1680 12045 186.291376  
## 4 0.05 50000 63070 1100 116.0 700 1680 12045 36.255496  
## 5 0.05 100 115600 900 116.0 820 1120 12045 39.604517  
## 6 0.05 100 63070 1100 63.1 820 1680 9855 31.204480  
## 7 0.15 100 63070 900 116.0 700 1680 12045 135.532142  
## 8 0.15 50000 63070 900 63.1 820 1120 12045 74.123560  
## 9 0.15 50000 115600 900 63.1 700 1680 9855 84.565843  
## 10 0.05 50000 115600 1100 63.1 700 1120 12045 65.803385  
## 11 0.15 100 115600 1100 116.0 700 1120 9855 235.642365  
## 12 0.05 100 63070 900 63.1 700 1120 9855 33.046273

The simulated responses were generated by 2 levels being either +1 or -1, according to the values the +1 were converted into their maximum corresponding values, and the -1 were converted into their minimum corresponding values for all 8 variables. From these newly set values the corresponding waterflow rate was calculated and analysis was done as to which of the variables were significant.

The analysis carried out was ANOVA as well as stepwise regression. From ANOVA there 4 variables which were found to significant, with X1 and X4 having the highest significance at 0.002 and 0.014 respectively. From stepwise results, for all directions of stepwise regression, 7 variables were found to be of significance. The model being fitted contained initially contained no variables and was lowered accordingly to AIC (Akaike Information Criterion).

As stepwise appeared to make better predictions of the relevance of variables we decided to use the 6 variables found from forward stepwise regression as the relevant variables, from the 7 provided. This is due to two reasons, when plotting out Mallows CP, 6 variables appears to be the most optimal variable selection possible and when carrying out stepwise the difference between 7 and 6 variable was an AIC difference of 0.07 which is relatively insignificant.

Therefore the significant variables were found to be, Radius of borehole, Radius of influence, Transitivity of upper aquifer, Potentiometric head of upper aquifer, Transmissivity of lower aquifer and Length of borehole. From the ANOVA results the significant variables were, Radius of borehole, Radius of influence, Potentiometric head of upper aquifer, Transmissivity of lower aquifer.

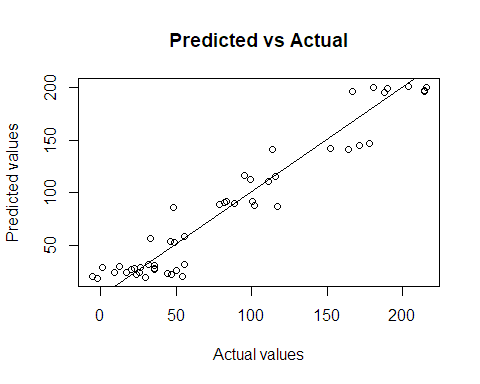
For the D-optimal design two main methods were used, gen.factorial to generate the variables at 3 levels (1,0,-1) and optFederov to select the most optimal rows based on their D-optimal value. 50 rows/runs were used in this experiment due to computer limitations and as it should be enough to provide good results. For this sections two experiments were carried out, with the 4 significant variables from ANOVA for one part and the 6 significant results from stepwise regression for the other.

From the simulated response the corresponding variables of 1,0,-1 were then converted to their max, mean and min values accordingly whilst all other variables were kept at their mean value. Responses were then generated for both the ANOVA and stepwise significant variables (Two different chunks of code used for separate trials of each). Under the assumption that all variables are uniformly distributed between their margins, a new matrix with the same dimensions was produced using these distributions for all 8 variables, this matrix would then be used to compare its responses to that of the model that was created.

When creating the meta model it was necessary to analyse which pair of variables were significant/relevant. To find these the original 4 variables from the ANOVA and 6 variables from the stepwise were fitted to linear models. EDA plots were then created to see the efficiency of the fit. From the residuals vs fitted plot it was clear to tell the model was inefficient due to the curvature present. ANOVA was then carried out on the linear model and variables 1 and 4 were found to be the most significant in that model. A pair was then created between the two and fitted to the linear model giving a much better overall fit.

In the stepwise the same processes were applied and no variables were found to be statistically significant both through ANOVA and summary of the linear model results and in the exploratory plots its clear to see the model isn’t very effective. Although attempts were made to fit a better model with extra terms and/or quadratic terms nothing was able to be compared to the effectiveness of the 4 ANOVA variables.

Both meta models developed don’t use all 8 variables present simply due to not all of them being of enough relevance to be input into the final linear model. As the stepwise variables performed much worse than the ANOVA model variables (found from screening), we will be focusing on the model containing the 4 ANOVA variables as being our main D-optimal model.

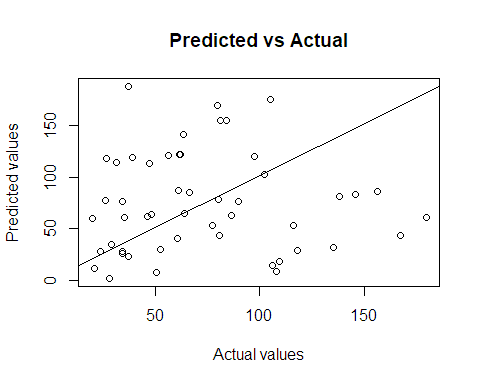
## [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]   
## [1,] 0.05 50000 89335 1110 116.00 760 1400 10950 35.7188582  
## [2,] 0.10 50000 89335 1110 116.00 760 1400 10950 98.9716708  
## [3,] 0.15 50000 89335 1110 116.00 760 1400 10950 214.3965683  
## [4,] 0.05 25050 89335 1110 116.00 760 1400 10950 12.4251768  
## [5,] 0.15 25050 89335 1110 116.00 760 1400 10950 189.9934486  
## [6,] 0.05 100 89335 1110 116.00 760 1400 10950 31.4740592  
## [7,] 0.10 100 89335 1110 116.00 760 1400 10950 94.9539322  
## [8,] 0.15 100 89335 1110 116.00 760 1400 10950 203.9537535  
## [9,] 0.05 50000 89335 1005 116.00 760 1400 10950 17.2433049  
## [10,] 0.15 50000 89335 1005 116.00 760 1400 10950 152.1240618  
## [11,] 0.05 100 89335 1005 116.00 760 1400 10950 22.1262529  
## [12,] 0.15 100 89335 1005 116.00 760 1400 10950 177.8534347  
## [13,] 0.05 50000 89335 900 116.00 760 1400 10950 53.9852933  
## [14,] 0.10 50000 89335 900 116.00 760 1400 10950 45.9528744  
## [15,] 0.15 50000 89335 900 116.00 760 1400 10950 101.5419699  
……

The model showed to have quite good performance. The R-squared value was relatively high at 0.9283 and the RMSEP value was also found to be relatively low at 17.692. This in comparison to the 0.1208 and 85.477 results from the 6 variable model explains our choice of linear model to use as a comparison to the kriging model.

The second meta model was constructed based on a space filling design specifically Latin Hypercube Design. Two design matrices were constructed, the first was a 50x4 and the second one was 50x8, the second matrix was constructed to be used as a comparison to the efficiency of the 50x4. All values of the matrices were randomly generated between 0 and 1. The reasoning behind the 50x4 matrix is to use as a comparison to the 4 variable matrix which had the highest performance in the D-optimal design.

Each value of the matrix was then applied into the formula a+((b-a)\*X), where a is the minimum value, b is the maximum value and X is the corresponding value in the matrix. For the 50\*4 matrix, variables 1,2,4 and 5 were used as the corresponding values whereas the 50x8 by all variables were put through the formula. The remaining 4 variables in the 50x4 matrix were set to their mean value and merged with the randomized value matrix. The water flow rate equation was then applied to both matrices to acquire the corresponding values.

To create the models the function buildKriging was used, with the design matrix and corresponding observations (water flow rate) used to build them. Interestingly the results based on the 4 variables shows that those variables were the least important variables in term of building the model. The model which contained all 8 variables contained 3 of the previously mentioned variables as the most important which is quite interesting when comparing both models against one another.

## ColX1 ColX2 ColX3 ColX4 ColX5 ColX6 ColX7 ColX8  
## [1,] 0.08457936 21497.7557 89335 901.6480 115.87354 760 1400 10950  
## [2,] 0.14550773 4692.9707 89335 910.0646 91.31244 760 1400 10950  
## [3,] 0.12122073 19987.9055 89335 1089.9449 94.88280 760 1400 10950  
## [4,] 0.10581723 35032.1539 89335 1048.5860 114.31425 760 1400 10950  
## [5,] 0.05407587 49978.7137 89335 923.6731 72.25693 760 1400 10950  
## [6,] 0.11397813 17610.8691 89335 1009.1902 98.63475 760 1400 10950  
## [7,] 0.09885141 6941.0251 89335 970.1716 99.42017 760 1400 10950  
## [8,] 0.14647120 757.6795 89335 937.2656 106.77471 760 1400 10950  
## [9,] 0.08225324 8849.2037 89335 981.5109 77.19918 760 1400 10950  
## [10,] 0.13191787 18965.0580 89335 905.4599 86.49949 760 1400 10950  
## [11,] 0.07467776 32538.2690 89335 1078.9917 75.29907 760 1400 10950  
## [12,] 0.07624326 25476.0699 89335 1076.0101 64.31246 760 1400 10950  
## [13,] 0.11621888 20257.8789 89335 995.7587 93.12925 760 1400 10950  
## [14,] 0.05086744 39394.4489 89335 914.8255 65.41527 760 1400 10950  
## [15,] 0.13681291 23702.0237 89335 1106.6579 64.03509 760 1400 10950  
…

The predicted values of the 4 variable matrix were quite poor, with most of the predicted values having no relationship to the actual values and having a RMSEP of 62.484 which is quite poor. The fully randomized variable matrix however showed much better results, showing a much straighter line and a low RMSEP of 15.778.

From this we can deduct that the more variables we have for the kriging model, the more efficient it becomes on the downside the more computational power and time it will take and with potentially multiple factor/level combinations it would realistically be unwise to use so many. For this section we believe the model containing 8 variables is a better performer however it doesn’t perform much better than the D-optimal model containing only 4 variables and one pair of the variables. The RMSEP value difference is only 1.914 and the R-squared value for the D-optimal design being at 0.9283 suggesting the model if fitted nicely and most of the data is accounted for. Therefore based on these aspects, the D-optimal design is the better meta model.

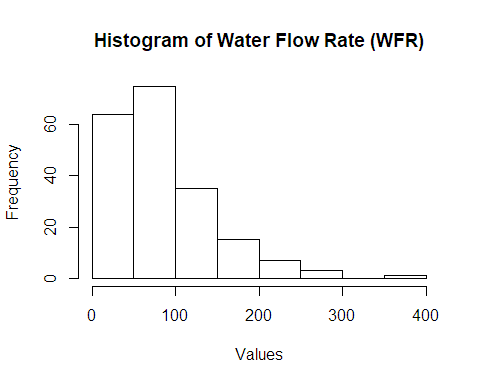
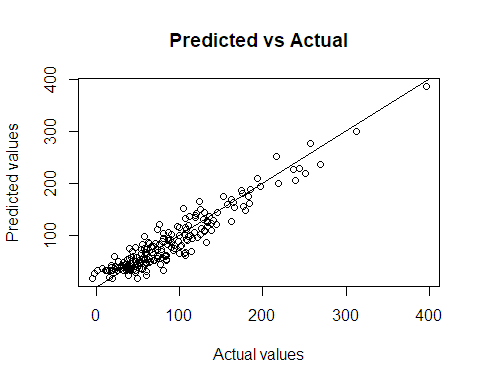
1. Using best meta models in 2, and assuming that each range of input variable provided is the 5th, and 95th percentile respectively, generate the distribution of WFR. Present informative graphs and specify the features of the distribution.

Based on the previously shown results the meta model used was the linear model acquired from the D-optimal design. The reasoning behind this is the fact it only uses 4 variables and performs nearly as well as the 8 variables model tried out from the space filling design and as the model requires less computational power and time it’s sensible to select it as the best meta-model found in part 2.

The assumed distribution for each variable was a normal distribution. The reasoning behind this is that a normal distribution is the easiest to fit for considering the information given (minimum, maximum value allows us to acquire median/mean, and the minimum/maximum value corresponding to the 5th and 95th percentile allows us to easily find the standard deviation using (median-minimum)/qnorm, given qnorm is at mu = 0 and sd = 1). This leaves us with the 4 variables having the parameters mean and standard deviation. (with the means being 0.1, 25050, 1005, 89.55 and the standard deviation being 0.03, 15168.52, 63.84, 16.08 for the respective variables.)

To generate the processes a 50x4 lhd matrix was randomly generated, where all variables are between 0 and 1. Each of these values was then used in the qnorm function, if the value was 0.64, then that would represent the 64th percentile of that variables. This was done for every randomly generated variable of the matrix and applied into the according mean and standard deviation. If it was the 64th percentile for the first variable we would apply qnorm(0.64, X1\_mean, X1\_standard deviation). This was then done for every randomly generated value. From this the results were once again merged into a matrix with all 8 variables, with the other variables being set at their median value.

By applying the same linear model to these results interestingly variables 4 and 5 appear as NA values when carrying out an ANOVA or summary of the model. At the same time however the interaction term between variables 1 and 4 still appears as significant, we are unsure of why this happens. Based on 200 runs, regardless the results provided show an R-squared of 0.896 and an RMSEP of 39.35845 (if no seed is set the values will clearly differ). This model was then used to find the corresponding values of WFR and be able to acquire the distribution of the WFR.



From the graphs above and the results/predictions generated from the model the distribution of the WFR appears to be some form of chi-squared or gamma distribution. It could be interpreted as other distribution types and although we are unsure of how to calculate the correct parameters for it, we believe it to be either a chi-squared or gamma distribution (log-normal could also be argued for).