Subsequent Health Assessment of Volunteers after Kexing Covid-19 Vaccine inoculation

Hao Sun(1003989053)

2020/12/21

1.Abstract

Nowadays, more than 150 vaccine candidates are in development, among them, 45 vaccines candidates are undergoing the Phase III in clinical trails(1). Due to the reason that every vaccine before licensed should meet its two standards safety and efficacy(4). This project is to investigate the reasons to influence the efficacy which represented by the monoclonal antibody(mAb) concentration and the safety which represented by the side effect in volunteers who inoculated by Kexing vaccine. Specifically, this vaccine in the phase III stage have been inoculated to 743 volunteers between 18 to 59 years old and 576 volunteers of them inoculated the second boost(5). The data sets in this project contains many aspects of information of volunteers and which variables may contribute the efficacy and side effect of this vaccine, for example, how many boosters of the vaccine, the age, the cholesterol levels, the uric acid, the BMI and sex. The project uses multiple linear regression models to statistically analysis which variables contributes to the efficacy and safety of this vaccine. From the results, how many times of booster and the BMI level of volunteers are the largest influence to the efficacy of this vaccine. And the cholesterol and BMI levels significantly influence the side effect after inoculation.

2.Key words

Vaccine, Efficacy, Safety(Side Effect), Multiple Linear Regression, Backward Elimination with AIC, Adjusted R Square, Residual Plot Analysis, Cook's Distance

3.Introduction

Nowadays, worldwide scientists have devoted themselves to the understand and research of SARS-COV-2 and achieved significant results(2). For example, 45 candidate vaccines are going through clinical trails in human. One of the Covid-19 candidate vaccine is named Kexing which is undergoing the phase III (final stage) clinical trails(3). How is the subsequent health assessment of these volunteers in phase III clinical trail?

Vaccine absolutely is the best method to thoroughly clear the COVID-19 pandemic which is a newly emerging infectious disease begins in late 2019 in Wuhan, China(2). A high efficacy vaccine could active both innate and adaptive immunity to people who get inoculation. Kexing Covid-19 vaccine is intramuscular and belongs to inactive vaccine which means the virus is killed and have no ability to replicate in

human bodies(5). Compare to attenuated vaccine, the inactive vaccine has low pathogenesis and significantly reduces the risk due to the reason that attenuated vaccine may be able to weak replicate inside of body and may mutate to normal virus to cause sever disease(4). However, vaccine as a medicine that is inoculate to people who is health must satisfy an extremely high standards of safety and efficacy. In order to test vaccine for the two standards, health assessment after inoculation is significant, and also for analyzing the assessment, the help from statistics is indispensable. In order to obtain an intact and direct report of the health assessment of these volunteers, only tables of their individual information is insufficient, statistic analysis for their information is inevitable. Health assessment contains many aspects of their health condition related to Covid-19, such as the number of booster, the cholesterol level, the uric acid level and the BMI index and so on, to satisfy the two standards of the vaccine, safety and efficacy. Also, in one health assessment, people have to assess and respond to vaccine side effects which is the unexpected unpleasant medical occurrences that occasionally follow immunization. However, sometimes, the mild or serious events occurring after immunization may not have a direct relationship with the vaccine due to the fact that correlation is not causation. Thus, statistical analysis could provide an unbiased causal inference between these side effects and vaccine to assess if there is a cause-and-effect relationship.

From the report, one data set will be used to assess the causal relationship between the different side effects and the vaccine. In the methodology section, I selected the multiple linear regression model to test the causal influence. Also, the result section, there is the results for the actual causal relationship my dependent and independent variables. The discussion, challenge and the conclusion are summarized in the conclusion section.

4.Data Wrangling

I simulated the parameter $MAbconcentration(ug/ml)$ between 0 to 800ug/ml of the first model, representing the efficacy of Kexing vaccine due to the reason that a good vaccine is immunogenic and induce antibodies in volunteers' bodies. The $booster$ is 1 or 2, represents the number of times the volunteer inoculate the vaccine based on the information gave in the instruction of this vaccine. The variables of $Cholesterol(mmol/L)$, $UricAcid(umol/L)$ and BMI represents the basic healthy background of these volunteers and were selected from the previous healthy assessment of other vaccines which summarize the characteristics regarding to the healthy level of volunteer. Also the parameter of second model is $SideEffect$ represents the unexpected outcome after inoculation.

```
dataset = read.csv("STA304Final.csv")
set.seed(1003989053)
data= dataset[sample(nrow(dataset),150),]
#Select 150 samples from the data set
data1=na.omit(data)
#Removed missing values
```

5.Model

This project contains Model 1, testing the efficacy of the vaccine and efficacy represented by monoclonal antibody, and Model 2, testing the safety of the vaccine and the safety represented by side effect. Both of them contains the similar steps.

1.Use Covariance Correlation Matrix to test if there is a significant relationship between each variables to eliminate the possibility that there is correlation between each variable.

2.Use Multiple Linear Regression(MLR) model. In Model 1, the parameter is mAb concentration and the variables are the booster, age, cholesterol, uric acid, BMI and sex. In Model 2, the parameter is side effect and the variables are the mAb, the booster, age, cholesterol, uric acid, BMI and sex.

3.From R2 in the summary of MLR, check if this model is a strong and has linear relationship. And P-values of each variables in summary are able to check if the variables are significant.

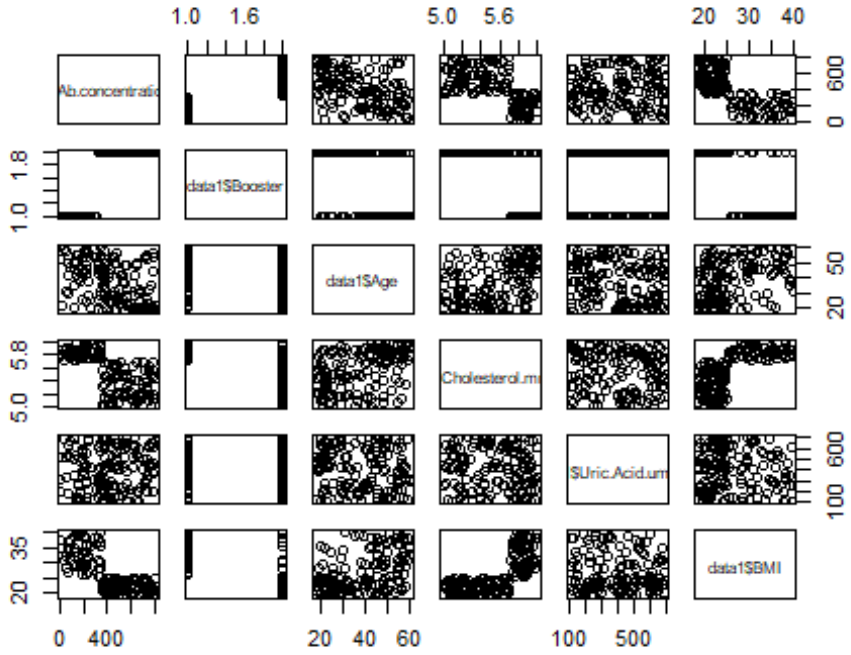
4.Use method Backward Elimination of AIC to find the most relative variables.

5.Use Residual Plot Analysis, from the 4 plots, testing if the assumptions of MLR is satisfy.

Model 1:What factors will significantly affect the concentration of antibodies level?

I. Exploratory Data Analysis

```
pairs(data1$MAb.concentration.ug.ml. ~
data1$Booster+data1$Age+data1$Cholesterol.mmol.L.+data1$Uric.Acid.umol.
L.+data1$BMI)
```




```
numericxy=cbind(data1$MAb.concentration.ug.ml.,data1$Booster,data1$Age,
data1$Cholesterol.mmol.L.,data1$Uric.Acid.umol.L.,data1$BMI)
#COrrelation matrix

round(cor(numericxy), 4)

##         [,1]    [,2]    [,3]    [,4]    [,5]    [,6]
## [1,]  1.0000  0.7954 -0.4374 -0.6687  0.0748 -0.7323
## [2,]  0.7954  1.0000 -0.4039 -0.6687  0.0467 -0.7607
## [3,] -0.4374 -0.4039  1.0000  0.4147 -0.1287  0.4173
## [4,] -0.6687 -0.6687  0.4147  1.0000 -0.0528  0.7025
## [5,]  0.0748  0.0467 -0.1287 -0.0528  1.0000 -0.0762
## [6,] -0.7323 -0.7607  0.4173  0.7025 -0.0762  1.0000
```

II. Methods and Model

```
full_model1=lm(data1$MAb.concentration.ug.ml. ~
data1$Booster+data1$Age+data1$Cholesterol.mmol.L.+data1$Uric.Acid.umol.
L.+data1$BMI+data1$Sex)

summary(full_model1)

##
## Call:
## lm(formula = data1$MAb.concentration.ug.ml. ~ data1$Booster +
##     data1$Age + data1$Cholesterol.mmol.L. + data1$Uric.Acid.umol.L.
+
```

```
##     data1$BMI + data1$Sex)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -215.82  -93.64  -21.45   83.61  318.37
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                903.56340  287.26153   3.145  0.00202 **
## data1$Booster              229.22674   35.13017   6.525  1.1e-09 ***
## data1$Age                   -1.57406    0.89109  -1.766  0.07945 .
## data1$Cholesterol.mmol.L. -116.19918   49.87748  -2.330  0.02122 *
## data1$Uric.Acid.umol.L.      0.02578    0.05570   0.463  0.64414
## data1$BMI                   -7.28905    2.81428  -2.590  0.01059 *
## data1$SexMale               24.31551   21.41155   1.136  0.25801
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122.7 on 143 degrees of freedom
## Multiple R-squared:  0.6931, Adjusted R-squared:  0.6802
## F-statistic: 53.82 on 6 and 143 DF,  p-value: < 2.2e-16
```

```
step(full_model1,direction="backward")
```

```
## Start:  AIC=1449.64
## data1$MAb.concentration.ug.ml. ~ data1$Booster + data1$Age +
##     data1$Cholesterol.mmol.L. + data1$Uric.Acid.umol.L. + data1$BMI
+
##     data1$Sex
##
##                              Df Sum of Sq     RSS    AIC
## - data1$Uric.Acid.umol.L.     1      3224 2154486 1447.9
## - data1$Sex                   1     19401 2170664 1449.0
## <none>                                    2151262 1449.6
## - data1$Age                   1     46942 2198204 1450.9
## - data1$Cholesterol.mmol.L.   1     81650 2232912 1453.2
## - data1$BMI                   1    100917 2252179 1454.5
## - data1$Booster               1    640512 2791775 1486.7
##
## Step:  AIC=1447.86
## data1$MAb.concentration.ug.ml. ~ data1$Booster + data1$Age +
##     data1$Cholesterol.mmol.L. + data1$BMI + data1$Sex
##
##                              Df Sum of Sq     RSS    AIC
## - data1$Sex                   1     18119 2172605 1447.1
## <none>                                    2154486 1447.9
## - data1$Age                   1     49626 2204112 1449.3
## - data1$Cholesterol.mmol.L.   1     80914 2235401 1451.4
## - data1$BMI                   1    102913 2257399 1452.9
## - data1$Booster               1    639012 2793498 1484.8
```

```
##
## Step:  AIC=1447.12
## data1$MAb.concentration.ug.ml. ~ data1$Booster + data1$Age +
##     data1$Cholesterol.mmol.L. + data1$BMI
##
##                              Df Sum of Sq     RSS    AIC
## <none>                                    2172605 1447.1
## - data1$Age                   1     39916 2212521 1447.8
## - data1$Cholesterol.mmol.L.   1     73031 2245637 1450.1
## - data1$BMI                   1    107949 2280554 1452.4
## - data1$Booster               1    669035 2841640 1485.4

##
## Call:
## lm(formula = data1$MAb.concentration.ug.ml. ~ data1$Booster +
##     data1$Age + data1$Cholesterol.mmol.L. + data1$BMI)
##
## Coefficients:
##               (Intercept)              data1$Booster
##                   877.580                    232.948
##                 data1$Age  data1$Cholesterol.mmol.L.
##                    -1.416                   -109.068
##                 data1$BMI
##                    -7.519
```

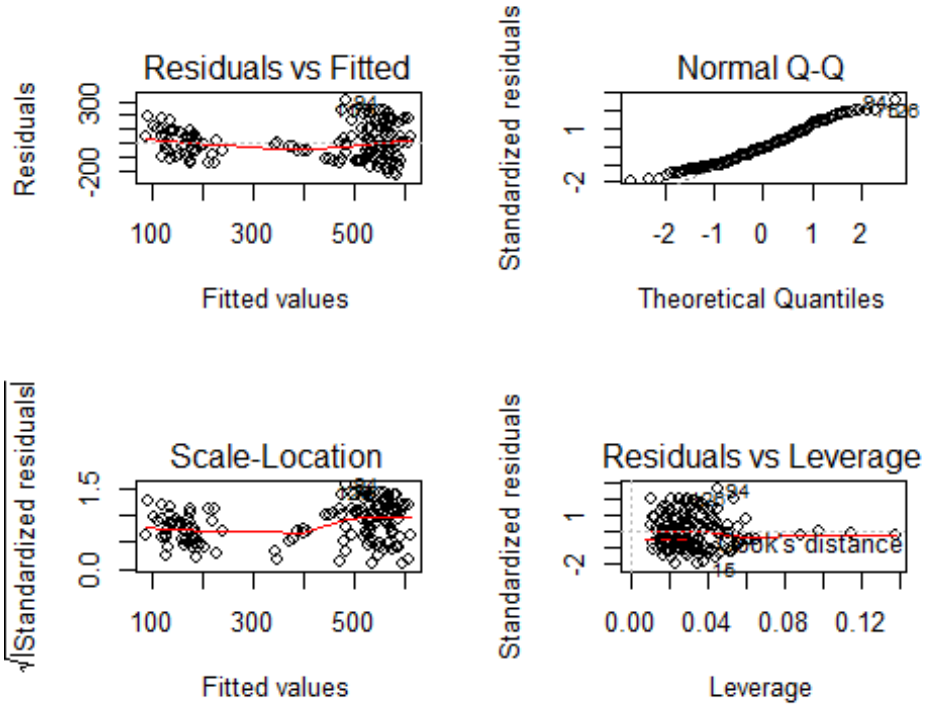
```
#Backward elimination with AIC - to see if we can get a better model

reduced_model1 = lm(formula = data1$MAb.concentration.ug.ml. ~
data1$Booster + data1$Age + data1$Cholesterol.mmol.L. + data1$BMI)


summary(reduced_model1)
```

```
##
## Call:
## lm(formula = data1$MAb.concentration.ug.ml. ~ data1$Booster +
##     data1$Age + data1$Cholesterol.mmol.L. + data1$BMI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -229.51  -98.85  -14.92   86.53  302.19
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  877.5799   283.6067   3.094  0.00237 **
## data1$Booster                232.9479    34.8611   6.682 4.71e-10 ***
## data1$Age                     -1.4159     0.8675  -1.632  0.10481
## data1$Cholesterol.mmol.L. -109.0682    49.4026  -2.208  0.02883 *
## data1$BMI                    -7.5189     2.8012  -2.684  0.00812 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122.4 on 145 degrees of freedom
## Multiple R-squared:   0.69,  Adjusted R-squared:  0.6815
## F-statistic:  80.7 on 4 and 145 DF,  p-value: < 2.2e-16
```

```
#Adjusted R-Square increases after we eliminate some explanatory
variables
```

```
par(mfrow=c(2,2))
plot(reduced_model1)
```




```
#Residual plot ananlysis

cook=cooks.distance(reduced_model1)
round(sort(cook,decreasing = TRUE)[1:10],4)
```

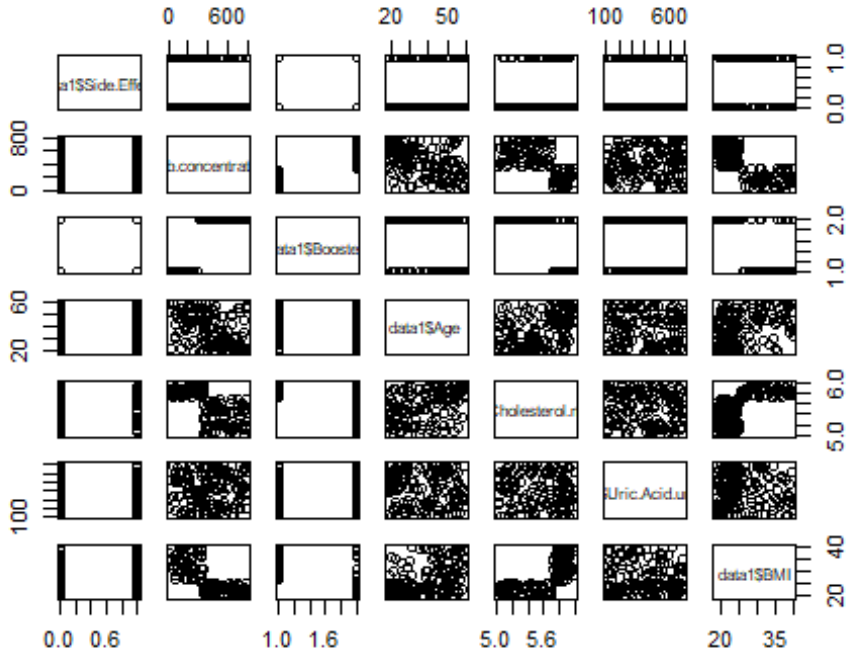
```
##     94    126     15     10     93     43    143     22    133
74
## 0.0598 0.0447 0.0258 0.0256 0.0251 0.0244 0.0237 0.0234 0.0227
0.0213
```

```
#Apparently, the largest cook distance is only 0.1820, so it is a good
sign that there is no influential point in the model.
```

Model 2:What factors will significantly cause the side effects?

I. Exploratory Data Analysis

```
pairs(data1$Side.Effect ~
data1$MAb.concentration.ug.ml.+data1$Booster+data1$Age+data1$Cholestero
l.mmol.L.+data1$Uric.Acid.umol.L.+data1$BMI)
```




```
numericxy=cbind(data1$Side.Effect,
data1$MAb.concentration.ug.ml.,data1$Booster,data1$Age,data1$Cholestero
l.mmol.L.,data1$Uric.Acid.umol.L.,data1$BMI)

round(cor(numericxy), 4)
```

```
##         [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]
## [1,]  1.0000 -0.2729 -0.2900  0.2385  0.3449 -0.0849  0.1756
## [2,] -0.2729  1.0000  0.7954 -0.4374 -0.6687  0.0748 -0.7323
## [3,] -0.2900  0.7954  1.0000 -0.4039 -0.6687  0.0467 -0.7607
## [4,]  0.2385 -0.4374 -0.4039  1.0000  0.4147 -0.1287  0.4173
## [5,]  0.3449 -0.6687 -0.6687  0.4147  1.0000 -0.0528  0.7025
## [6,] -0.0849  0.0748  0.0467 -0.1287 -0.0528  1.0000 -0.0762
## [7,]  0.1756 -0.7323 -0.7607  0.4173  0.7025 -0.0762  1.0000
```

##II.Methods and Model

```
full_model2=lm(data1$Side.Effect~ data1$MAb.concentration.ug.ml.+
data1$Booster+data1$Age+data1$Cholesterol.mmol.L.+data1$Uric.Acid.umol.
```

```
L.+data1$BMI+data1$Sex)

summary(full_model2)

##
## Call:
## lm(formula = data1$Side.Effect ~ data1$MAb.concentration.ug.ml. +
##     data1$Booster + data1$Age + data1$Cholesterol.mmol.L. +
data1$Uric.Acid.umol.L. +
##     data1$BMI + data1$Sex)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8060 -0.3534 -0.1500  0.4185  0.9605
##
## Coefficients:
##                                  Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                    -1.6426606  1.1044958  -1.487
0.13917
## data1$MAb.concentration.ug.ml. -0.0001280  0.0003110  -0.412
0.68117
## data1$Booster                  -0.2212263  0.1488105  -1.487
0.13933
## data1$Age                       0.0041832  0.0033494   1.249
0.21374
## data1$Cholesterol.mmol.L.       0.5459623  0.1889528   2.889
0.00447 **
## data1$Uric.Acid.umol.L.        -0.0001665  0.0002073  -0.803
0.42308
## data1$BMI                      -0.0253443  0.0107074  -2.367
0.01928 *
## data1$SexMale                   0.0027030  0.0799756   0.034
0.97309
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4561 on 142 degrees of freedom
## Multiple R-squared:  0.1697, Adjusted R-squared:  0.1288
## F-statistic: 4.146 on 7 and 142 DF,  p-value: 0.000353

step(full_model2,direction="backward")

## Start:  AIC=-227.75
## data1$Side.Effect ~ data1$MAb.concentration.ug.ml. + data1$Booster +
##     data1$Age + data1$Cholesterol.mmol.L. + data1$Uric.Acid.umol.L.
+
##     data1$BMI + data1$Sex
##
##                                  Df Sum of Sq    RSS     AIC
```

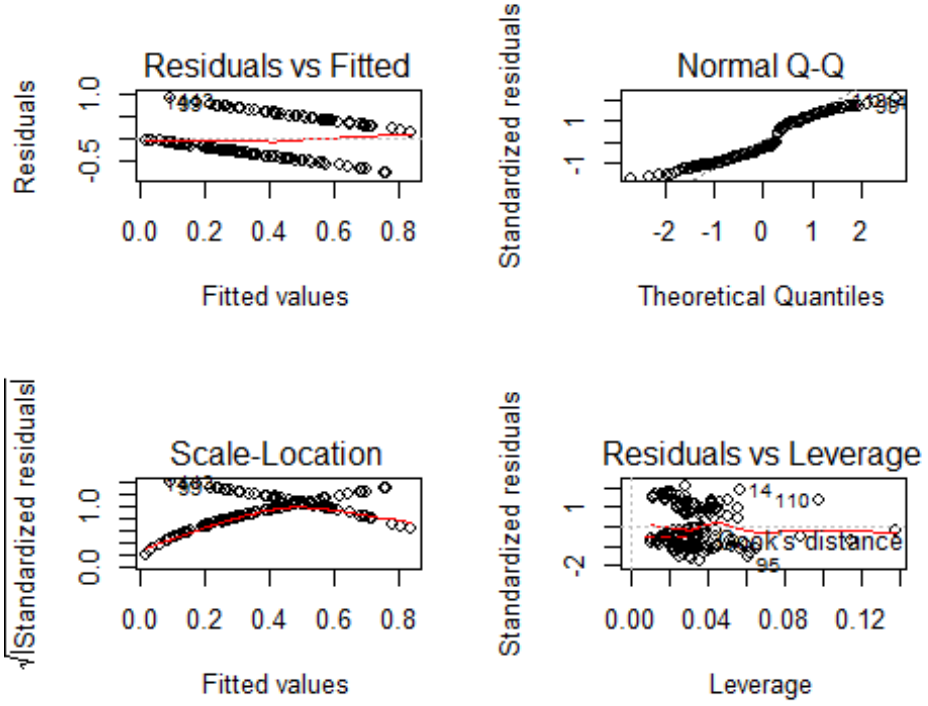
```
## - data1$Sex                         1   0.00024 29.537 -229.75
## - data1$MAb.concentration.ug.ml.    1   0.03526 29.572 -229.57
## - data1$Uric.Acid.umol.L.           1   0.13426 29.671 -229.07
## - data1$Age                         1   0.32446 29.861 -228.11
## <none>                                          29.537 -227.75
## - data1$Booster                     1   0.45971 29.997 -227.43
## - data1$BMI                         1   1.16539 30.702 -223.94
## - data1$Cholesterol.mmol.L.         1   1.73658 31.273 -221.18
##
## Step:  AIC=-229.75
## data1$Side.Effect ~ data1$MAb.concentration.ug.ml. + data1$Booster +
##     data1$Age + data1$Cholesterol.mmol.L. + data1$Uric.Acid.umol.L.
+
##     data1$BMI
##
##                                   Df Sum of Sq    RSS     AIC
## - data1$MAb.concentration.ug.ml.    1   0.03503 29.572 -231.57
## - data1$Uric.Acid.umol.L.           1   0.13660 29.674 -231.06
## - data1$Age                         1   0.34193 29.879 -230.02
## <none>                                          29.537 -229.75
## - data1$Booster                     1   0.45968 29.997 -229.43
## - data1$BMI                         1   1.16818 30.705 -225.93
## - data1$Cholesterol.mmol.L.         1   1.77553 31.313 -222.99
##
## Step:  AIC=-231.57
## data1$Side.Effect ~ data1$Booster + data1$Age +
data1$Cholesterol.mmol.L. +
##     data1$Uric.Acid.umol.L. + data1$BMI
##
##                                 Df Sum of Sq    RSS     AIC
## - data1$Uric.Acid.umol.L.        1   0.14089 29.713 -232.86
## - data1$Age                      1   0.37746 29.950 -231.67
## <none>                                       29.572 -231.57
## - data1$Booster                  1   0.77388 30.346 -229.69
## - data1$BMI                      1   1.13543 30.708 -227.92
## - data1$Cholesterol.mmol.L.      1   1.92990 31.502 -224.09
##
## Step:  AIC=-232.86
## data1$Side.Effect ~ data1$Booster + data1$Age +
data1$Cholesterol.mmol.L. +
##     data1$BMI
##
##                                 Df Sum of Sq    RSS     AIC
## <none>                                       29.713 -232.86
## - data1$Age                      1   0.43601 30.149 -232.67
## - data1$Booster                  1   0.75558 30.469 -231.09
## - data1$BMI                      1   1.10313 30.816 -229.39
## - data1$Cholesterol.mmol.L.      1   1.91665 31.630 -225.48
```

```
##
## Call:
## lm(formula = data1$Side.Effect ~ data1$Booster + data1$Age +
##     data1$Cholesterol.mmol.L. + data1$BMI)
##
## Coefficients:
##                (Intercept)              data1$Booster
##                   -1.84226                   -0.24756
##                  data1$Age  data1$Cholesterol.mmol.L.
##                    0.00468                    0.55875
##                  data1$BMI
##                   -0.02404
```

```
#Backward elimination with AIC - to see if we can get a better model

reduced_model2 = lm(data1$Side.Effect ~ data1$Booster + data1$Age +
data1$Cholesterol.mmol.L. + data1$BMI)

summary(reduced_model2)
```

```
##
## Call:
## lm(formula = data1$Side.Effect ~ data1$Booster + data1$Age +
##     data1$Cholesterol.mmol.L. + data1$BMI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7608 -0.3581 -0.1376  0.4177  0.9111
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -1.842263   1.048817  -1.757  0.08111 .
## data1$Booster             -0.247557   0.128921  -1.920  0.05679 .
## data1$Age                  0.004680   0.003208   1.459  0.14682
## data1$Cholesterol.mmol.L.  0.558746   0.182698   3.058  0.00265 **
## data1$BMI                 -0.024036   0.010359  -2.320  0.02173 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4527 on 145 degrees of freedom
## Multiple R-squared:  0.1647, Adjusted R-squared:  0.1417
## F-statistic:  7.15 on 4 and 145 DF,  p-value: 2.781e-05
```

```
#Adjusted R-Square increases after we eliminate some explanatory
variables

par(mfrow=c(2,2))
plot(reduced_model2)
```




```
#Residual plot analysis

cook=cooks.distance(reduced_model2)
round(sort(cook,decreasing = TRUE)[1:10],4)
```

```
##     14    110     95    113    128     51     56    129     12
136
## 0.0431 0.0398 0.0311 0.0240 0.0227 0.0217 0.0208 0.0205 0.0189
0.0187
```

6.Results

Model 1 is to test if the variables boosters, age, cholesterol, uric acid, BMI and sex have linear relationship to the parameter, the mAb concentration. First, summarize the multiple linear regression model, the p-value for booster is 1.1*10-9 which indicates the times of booster the volunteer inoculated significantly predict the mAb concentration which represents the efficacy of the Kexing vaccine. And the intercept shows that if the time of booster is increased by 1, the concentration of monoclonal antibody is increased by 229.23ug/ml. Also, the p-values for cholesterol and BMI are 0.02122 and 0.01059 respectively, which represents the two variables also have significant linear relationship to the efficacy of the vaccine, and if the cholesterol level and BMI index are increased by 1 unit, the concentration of mAb is decreased by 116.20ug/ml and 7.29ug/ml. However, the p-values for the age, the uric acid level and the sex are not significant which is higher than 0.05 which indicates these three variables give no significant prediction to the efficacy of vaccine. Thus, I used the method of backward elimination of AIC to eliminate the sex and uric acid that

have the highest p-value. The adjusted R-square increased from 0.6802 to 0.6815 which indicates the model becomes stronger to predict a linear relationship. Also, from the residual plot analysis, the Residuals vs Fitted plot is to evaluate the linear relationship between parameter and variables. From the plot, most data points are between 100 to 200 and 500 to 600 that are not random distributed and has a clear pattern. So, it is not a strong linear relationship. The second plot, the Normal Q-Q plot shows if residuals are normally distributed. It's good if residuals are lined well on the straight dashed line, almost all the points are on the dashed line in the plot. The Scalre-Location plot shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance. It's good if there is a horizontal line with equally (randomly) spread points. However, in the plot these points are not equally distributed and have clear pattern. The Residuals vs Leverage plot is to find influential cases if any and with further Cook's distance analysis, there is no influential points in these dataset.

Model 2 is to test if the variables mAb concentration, boosters, age, cholesterol, uric acid, BMI and sex have linear relationship to the side effect of the vaccine. First, from the summary table of MLR, the p-value for Cholesterol is 0.0047 which indicates the cholesterol level in volunteer significantly predict the side effect which represents the safety of the vaccine. And the intercept shows that if the cholesterol is increased by 1mmol/L, the probability of having side effect increases 54%. Also, the p-values for the BMI index is 0.01928, which represents the BMI also have significant linear relationship to the safety of the vaccine, and if the BMI index are increased by 1 unit, the probability of having side effect increases 2.5%. However, the p-values for the other variables include the mAb concentration are not significant which is higher than 0.05 which indicates these three variables give no significant prediction to the safety of vaccine. Thus, through the method of backward elimination of AIC, I generated to a reduced model only contains the variables of booster, age, cholesterol and the BMI index. The adjusted R-square increased from 0.1288 to 0.1417 which indicates the model becomes stronger to predict a linear relationship. Also, from the residual plot analysis, from the Residuals vs Fitted plot, most data points are not random distributed and has a clear pattern. So, it is not a strong linear relationship. The the Normal Q-Q plot shows all the points are on the dashed line in the plot. For the Scalre-Location plot these points are not equally distributed and have clear pattern. The Residuals vs Leverage plot is to find influential cases if any and with further Cook's distance analysis, there is no influential points in these dataset.

7.Discussion

I.Summary

In this project of Subsequent Health Assessment of Volunteers after Kexing Covid-19 Vaccine inoculation, I uses two Multiple Linear Regression models to test the efficacy and safety of the Kexing vaccine. To quantify the efficacy, I used the concentration of antibodies to represents how strong the immunogenicity is in volunteers after inoculation. Also to represent the safety of this vaccine, I choose to

use whether the volunteers have side effect. In this project, six variables, boosters, age, cholesterol, uric acid, BMI and sex are used to predict the concentration of antibodies. And seven variables mAb concentration, boosters, age, cholesterol, uric acid, BMI and sex are used to predict the side effect of the vaccine.

II.Conclusions

From Model 1 in the result section, there are three variables, the booster, the cholesterol and the BMI index are able to significantly predict the concentration of antibodies. For the booster, it has the smallest p-value, and if the time of booster is increased by 1, the concentration of monoclonal antibody is increased by 229.23ug/ml. This result indicate that the booster times significantly increase the efficacy of the Kexing vaccine. From the instrument of this vaccine, the best time between the first and second inoculation time is 14 days which generates to the highest efficacy(1). Thus, if this vaccine pasts the clinical trial and enter to the market in future, the two times inoculation are indispensable and the time between the two shots should be regulated. The reason to have two boosters is because our immune system that the adaptive immunity exerts memory to the pathogen which infected our body. Compare to the primary infection, the secondary infection of the same pathogen gives rise to a higher affinity to the pathogen of the antibodies and a stronger and quicker response to this infected pathogen(2, 7). Thus, if we have the second shot of the vaccine, we can obtain a better immunogenic effect. For the cholesterol and BMI index, basically, these two indexes represent the obesity of these volunteers. If the indexes are high for the two variables(i.e. obesity), the efficacy of the vaccine is reduced. The reason to explain the relationship between obesity and vaccine is that the doses of vaccine is related to the weight of inoculator. Theoretically, inoculator with a heavier weight should inoculate more doses of antigens in the vaccine(6). Thus, if people with obesity receive the same doses, then the efficacy compare to normal people would reduce. Also, obesity may trigger the disturbance of metabolism, especially for the disturbance in immune system. This can lead to decreased function of immune cells responding to the antigens in the vaccine and cause the reduction of vaccine efficacy(8).

For the Model 2 result, the variables that have the linear relationship to the side effect of Kexing vaccine are the cholesterol and BMI. This result reveals that people with obesity tend to have side effect after vaccine inoculation. The most interesting point in the result is that the concentration of monoclonal antibody has no significant linear relationship to the side effect which represents the safety of this vaccine. And due to the fact that the concentration of antibody represents the caused immunogenicity by this vaccine, if it dose not directly predict the safety. Then, this may conclude that the side effects occurred in these volunteers have no cause-and-effect relationship. Because no matter the concentration of antibody is, there is no significant relationship to induce a side effect. These side effect may generate by coincidence rather than the vaccine.

III.Weakness & Next Steps

In Model 1: The adjusted R Square in the first model is 0.68, which is relatively strong but apparently still needs to be improved. To improve the adjusted R Square, I can attempt to find out more relevant explanatory variables that can strongly explain the variations in the response variable. Also, the clear pattern that indicated in Residual vs. Fitted and the Scale-Location plots demonstrate that the assumption of linear relationship and equal variance are violated. In other words, the multiple linear regression could be unworkable because of these violations in important assumptions. To figure out this issue, I can use weighted least square method or make some transformations to the explanatory variables until the assumption of linear relationship and equal variance can be satisfied. Finally, the correlation between cholesterol and BMI is notably high (up to 0.7025). In order to avoid a multicollinearity problem that leads to an inaccurate model, I should delete one of them or replace one of them by another unrelated explanatory variable.

In Model 2: Compared to the first model, the second multiple linear regression model is far less credible since the adjusted R Square in the second model is only 0.13. In other words, the explanatory variables I chose are failed to explain the side effects of vaccines. Therefore, the best way to improve my prediction is changing those insignificant variables which have a p-value much higher than significance level (0.05). Alternatively, we can keep the remaining explanatory variables in the reduced model that is generated from the backward elimination with AIC, and then add more new variables into the data set. From the residual plots, it is clear to see that basically all the assumptions of linear regression model are violated. Therefore, I will try to make some transformations such as logarithms to the existing explanatory variables to see if those violations can be adjusted.

8.Reference

1.Le, T. Thanh, et al. “The COVID-19 vaccine development landscape.” Nat Rev Drug Discov 19.5 (2020): 305-306.

2.Tillett, Richard L., et al. “Genomic evidence for reinfection with SARS-CoV-2: a case study.” The Lancet Infectious Diseases (2020).

3.Grijalva, Carlos G., et al. “Transmission of SARS-COV-2 infections in households—Tennessee and Wisconsin, April–September 2020.” Morbidity and Mortality Weekly Report 69.44 (2020): 1631.

4.Siegrist, Claire-Anne. “Vaccine immunology.” Vaccines 5.1 (2008): 17-36.

5.La Montagne, John R., et al. “Summary of clinical trials of inactivated influenza vaccine—1978.” Reviews of infectious diseases 5.4 (1983): 723-736.

6.Rhorer, Janelle, et al. “Efficacy of live attenuated influenza vaccine in children: a meta-analysis of nine randomized clinical trials.” Vaccine 27.7 (2009): 1101-1110.

7.Pfefferbaum, Betty, and Carol S. North. “Mental health and the Covid-19 pandemic.” New England Journal of Medicine (2020).

8.Watkins, John. “Preventing a covid-19 pandemic.” (2020).