BIS623 Homework 5

Due on 10/31/2019 before the lecture

Joanna Chen

31 October, 2019

Problem 1

We collected a dataset for 25 customers satisfactory score (Y) at different time of the day (X) for a restaurant. The data was saved in `poly.rdata`. Please finish the following questions.

```
#Set the work directory and read in the data
setwd("~/Downloads/HW 5")
poly_data = read.csv("poly_data.csv", header=TRUE)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

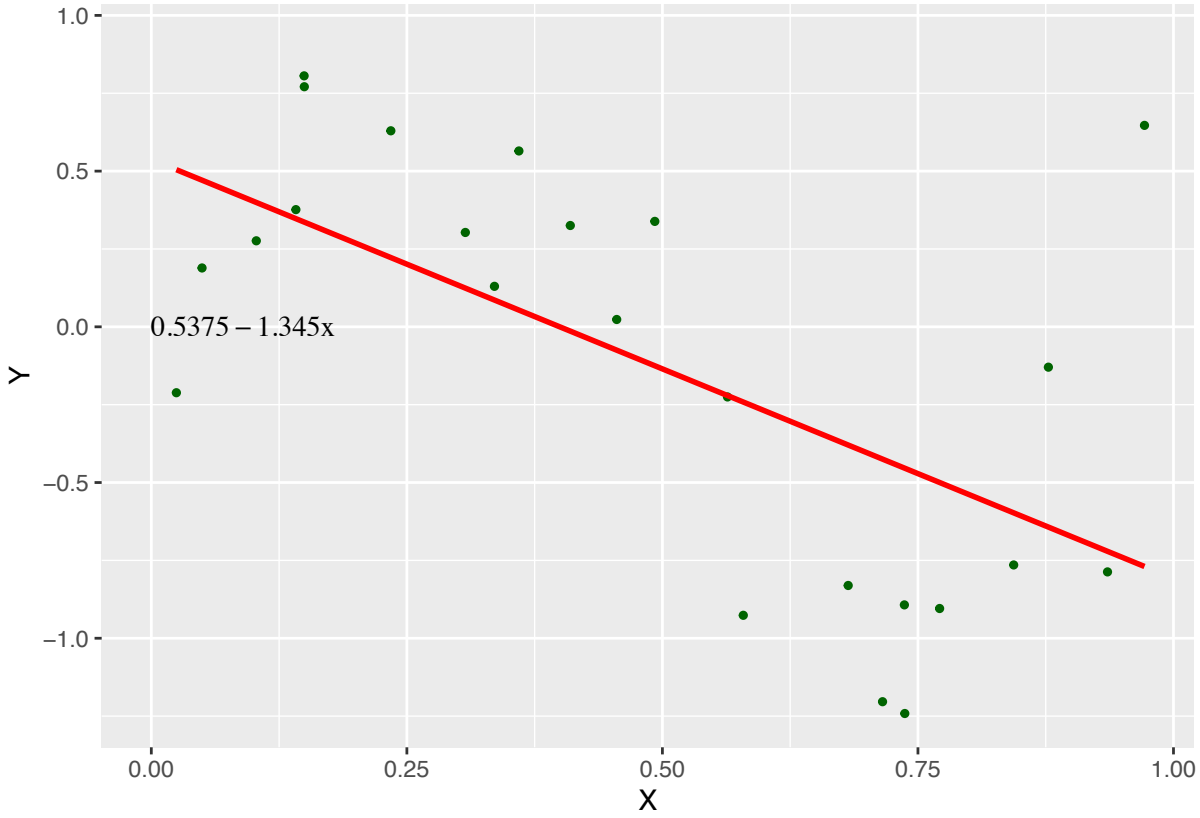
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

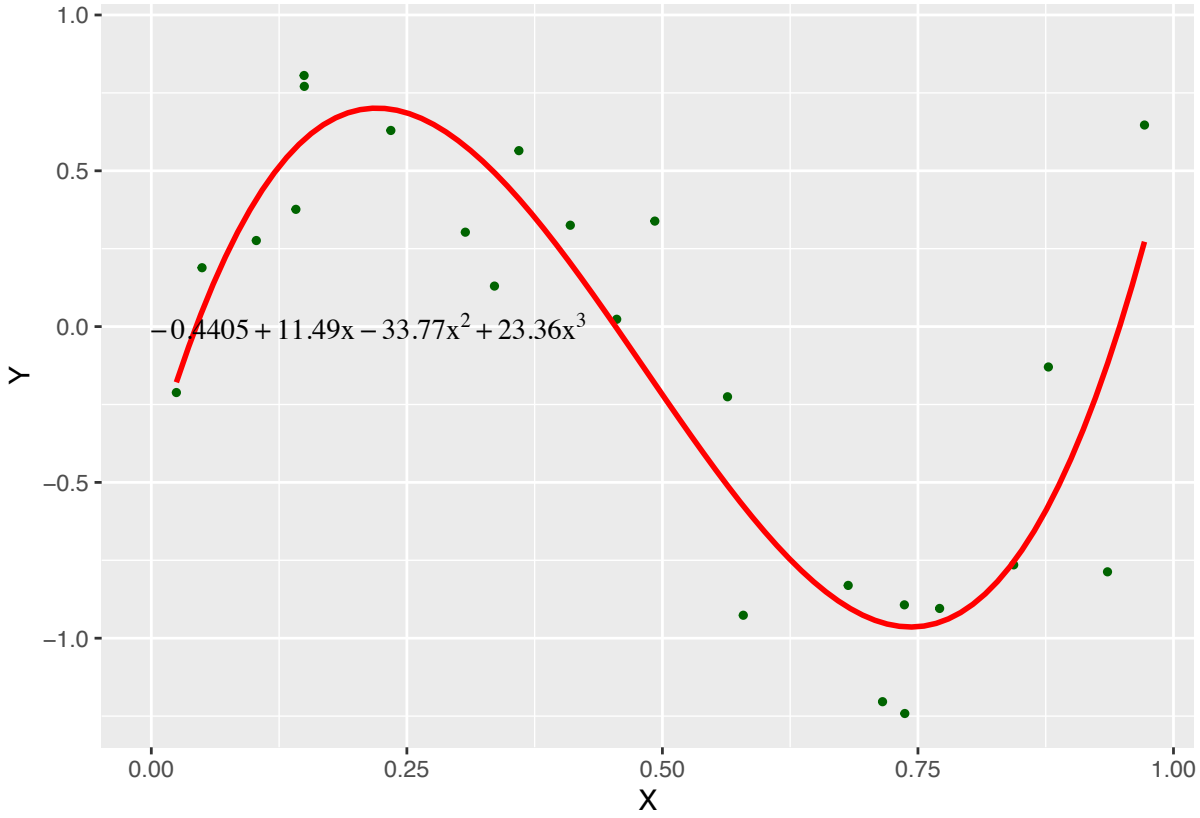
```
library(polynom)
```

1. Fit polynomial models from orders 1, 3, 5, 7, and 9 using R. Plot the fitted curve over original data for each of these models.

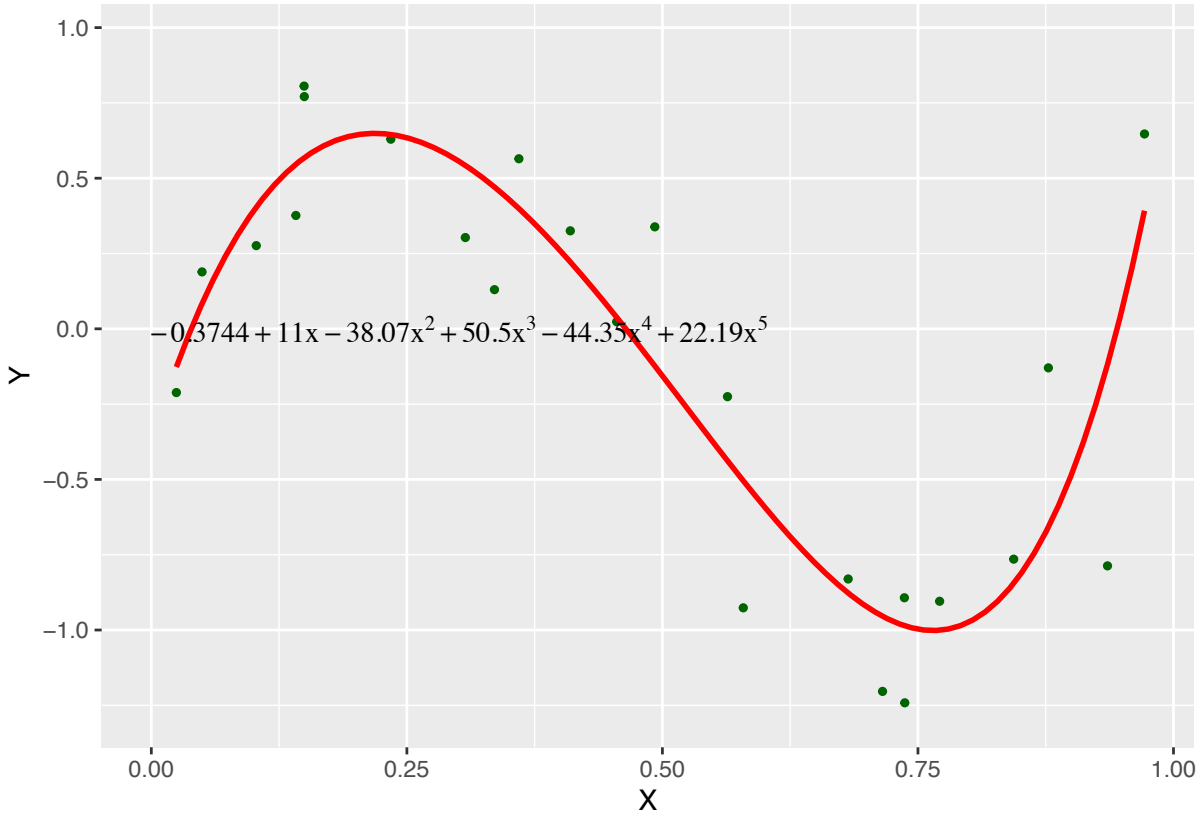
```
polyd1=lm(Y ~ poly(X, 1, raw=TRUE), data=poly_data)
polyd3=lm(Y ~ poly(X, 3, raw=TRUE), data=poly_data)
polyd5=lm(Y ~ poly(X, 5, raw=TRUE), data=poly_data)
polyd7=lm(Y ~ poly(X, 7, raw=TRUE), data=poly_data)
polyd9=lm(Y ~ poly(X, 9, raw=TRUE), data=poly_data)


my.eq1 <- as.character(signif(as.polynomial(coef(polyd1)), 4))
my.eq3 <- as.character(signif(as.polynomial(coef(polyd3)), 4))
my.eq5 <- as.character(signif(as.polynomial(coef(polyd5)), 4))
my.eq7 <- as.character(signif(as.polynomial(coef(polyd7)), 4))
my.eq9 <- as.character(signif(as.polynomial(coef(polyd9)), 4))

p = ggplot(data = poly_data, aes(x = X, y = Y)) +
    geom_point(color = "darkgreen", size = 1)  #Plotting the scatter point

p + stat_smooth(method="lm", se=TRUE, fill=NA,formula=y ~ poly(x, 1, raw=TRUE),colour="red") +
    annotate(geom = "text", x = 0, y = 0, label = my.eq1, family = "serif", hjust = 0, parse = TRUE, si
```



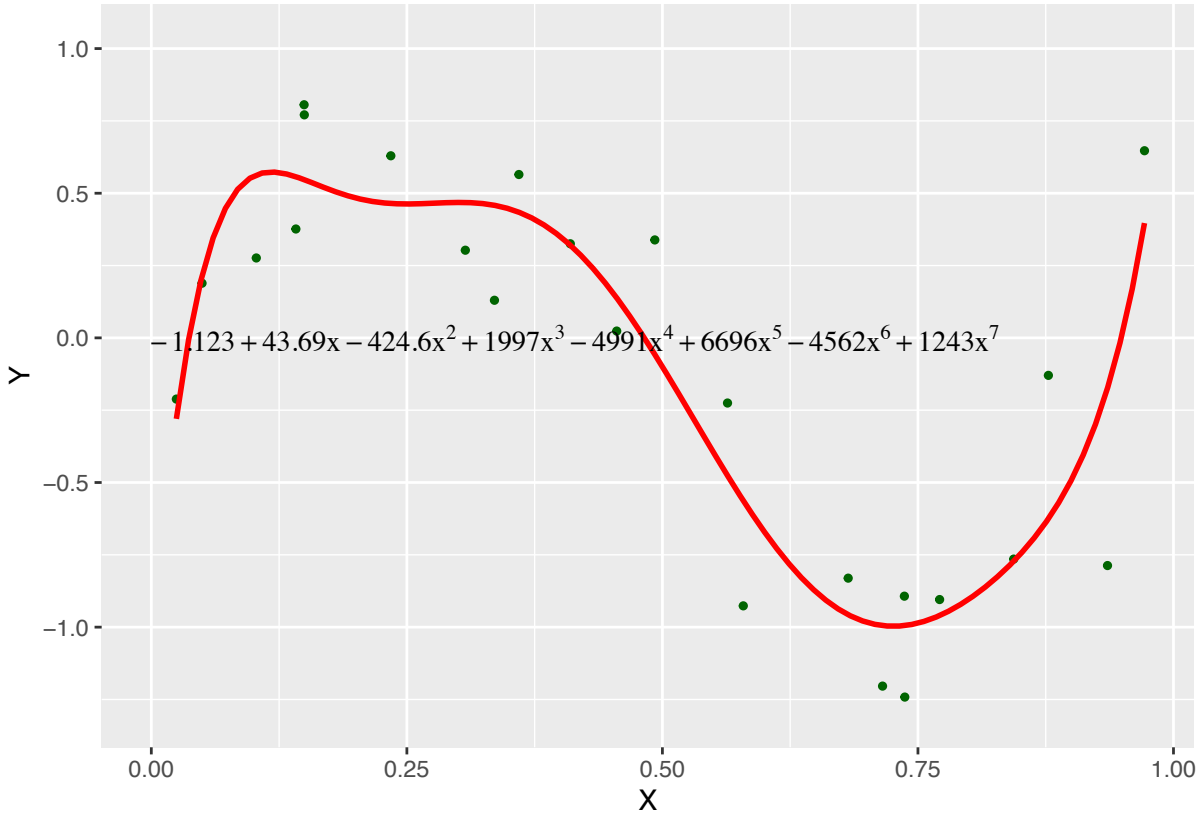

```
p + stat_smooth(method="lm", se=TRUE, fill=NA,formula=y ~ poly(x, 3, raw=TRUE),colour="red") +
    annotate(geom = "text", x = 0, y = 0, label = my.eq3, family = "serif", hjust = 0, parse = TRUE, si
```



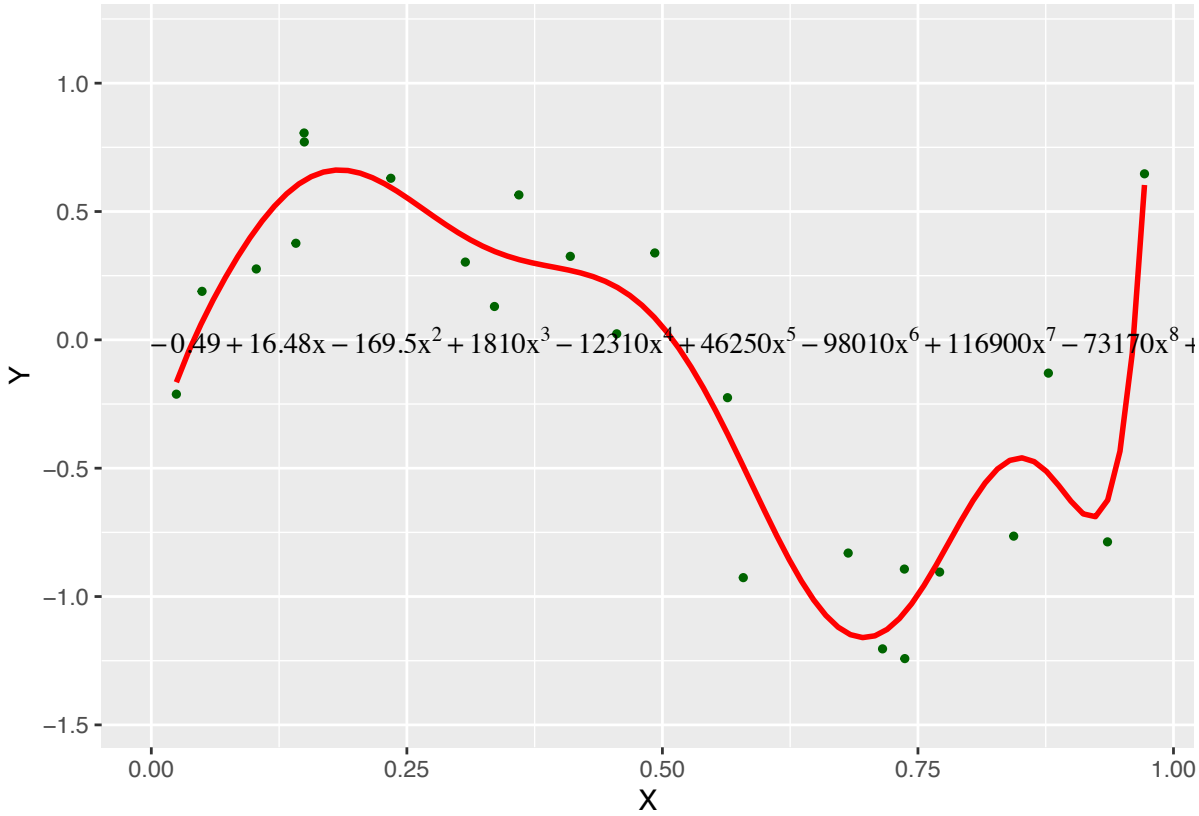

```
p + stat_smooth(method="lm", se=TRUE, fill=NA,formula=y ~ poly(x, 5, raw=TRUE),colour="red") +
    annotate(geom = "text", x = 0, y = 0, label = my.eq5, family = "serif", hjust = 0, parse = TRUE, si
```




```
p + stat_smooth(method="lm", se=TRUE, fill=NA,formula=y ~ poly(x, 7, raw=TRUE),colour="red") +
    annotate(geom = "text", x = 0, y = 0, label = my.eq7, family = "serif", hjust = 0, parse = TRUE, si
```




```
p + stat_smooth(method="lm", se=TRUE, fill=NA,formula=y ~ poly(x, 9, raw=TRUE),colour="red") +
    annotate(geom = "text", x = 0, y = 0, label = my.eq9, family = "serif", hjust = 0, parse = TRUE, si
```



2. Which polynomial order will you choose eventually? State your reasons.
   I will choose the third order polynomial. We don't want to the model overfits the data.
3. What is the "Adjusted R-squared" for the final model you choose?
   From the following summary statistics, we know that the adjusted R-squared is 0.7935.

```
summary(polyd3)
```

```
##
## Call:
## lm(formula = Y ~ poly(X, 3, raw = TRUE), data = poly_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66789 -0.21219  0.03661  0.15844  0.52158
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -0.4405     0.2388  -1.844     0.08 .
## poly(X, 3, raw = TRUE)1   11.4878     2.1566   5.327 3.26e-05 ***
## poly(X, 3, raw = TRUE)2 -33.7699     5.0532  -6.683 1.67e-06 ***
## poly(X, 3, raw = TRUE)3  23.3646     3.3456   6.984 8.89e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3014 on 20 degrees of freedom
## Multiple R-squared:  0.8205, Adjusted R-squared:  0.7935
## F-statistic: 30.47 on 3 and 20 DF,  p-value: 1.177e-07
```

4. Fit another polynomial model under the polynomial order you choose but using orthogonal polynomials. Does LSE change compared with those in your original model? Does Adjusted R-squared change compared that in your original model?
   Comparing LSE (the Estimate Std column), it changes. In the orthogonal polynomial model, Adjusted R-squared = 0.7935. It doesn't change.

```
polyd3_orthogonal=lm(Y ~ poly(X, 3, raw=FALSE), data=poly_data)
summary(polyd3_orthogonal)
```

```
##
## Call:
## lm(formula = Y ~ poly(X, 3, raw = FALSE), data = poly_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66789 -0.21219  0.03661  0.15844  0.52158
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -0.11410    0.06152  -1.855   0.0784 .
## poly(X, 3, raw = FALSE)1 -1.93456    0.30139  -6.419 2.92e-06 ***
## poly(X, 3, raw = FALSE)2  0.36055    0.30139   1.196   0.2456
## poly(X, 3, raw = FALSE)3  2.10478    0.30139   6.984 8.89e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3014 on 20 degrees of freedom
## Multiple R-squared:  0.8205, Adjusted R-squared:  0.7935
## F-statistic: 30.47 on 3 and 20 DF,  p-value: 1.177e-07
```

```
#Mean square error
#mean(polyd3_orthogonal$residuals^2)
#mean(polyd3$residuals^2)
```

Problem 2

We want to fit a model with age, race (four categories–white, black, asian and other) and their interaction for the `income` example in the class. Please finish the following question.

1. Please first create a reference cell coding for "race", and then write down the regression model formulation.

$$x_2 = \begin{cases} 1 & \text{if white} \\ 0 & \text{otherwise} \end{cases}$$
$$x_3 = \begin{cases} 1 & \text{if black} \\ 0 & \text{otherwise} \end{cases}$$
$$x_4 = \begin{cases} 1 & \text{if asian} \\ 0 & \text{otherwise} \end{cases}$$

   Let x_1 be the age. Then

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4$$

For the case race is White, $E(Y) = \beta_0 + \beta_2 + \beta_1 x_1 + \beta_5 x_1$;

For the case race is Black, $E(Y) = \beta_0 + \beta_3 + \beta_1 x_1 + \beta_6 x_1$;

For the case race is Asian, $E(Y) = \beta_0 + \beta_4 + \beta_1 x_1 + \beta_7 x_1$.

For the case race is others, $E(Y) = \beta_0 + \beta_1 x_1$;

2. Please state the null and alternative hypothesis for "Coincident". If you are going to use Wald test, what is the R matrix?

$$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$

$$H_a : \text{At least one of them is non-zero.}$$

$$\begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}_{6\times 8}$$

3. Please state the null and alternative hypothesis for "Parallel". If you are going to use Wald test, what is the R matrix?

$$H_0 : \beta_5 = \beta_6 = \beta_7 = 0$$

$$H_a : \text{At least one of them is non-zero.}$$

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}_{3\times 8}$$

4. Please state the null and alternative hypothesis for "Equal intercept". If you are going to use Wald test, what is the R matrix?

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_a : \text{At least one of them is non-zero.}$$

$$\begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}_{3\times 8}$$