Flipped Assignment 8

Group 5

2/21/2022

Input Data

```
setwd('G:/OneDrive - Texas Tech University/IE 5344 Statistical Data Analysis/Flipped Assignment 8')
data <- read.csv('data-table-B9.csv', header = TRUE)
colnames(data) <- c('x1', 'x2', 'x3', 'x4', 'y')
```

Part a.

Fitting the Model

```
fit1 <- lm(y ~ x1 + x2 + x3 + x4 + x1:x2 + x1:x3 + x1:x4 + x2:x3 + x2:x4 + x3:x4, data)
summary(fit1)
```

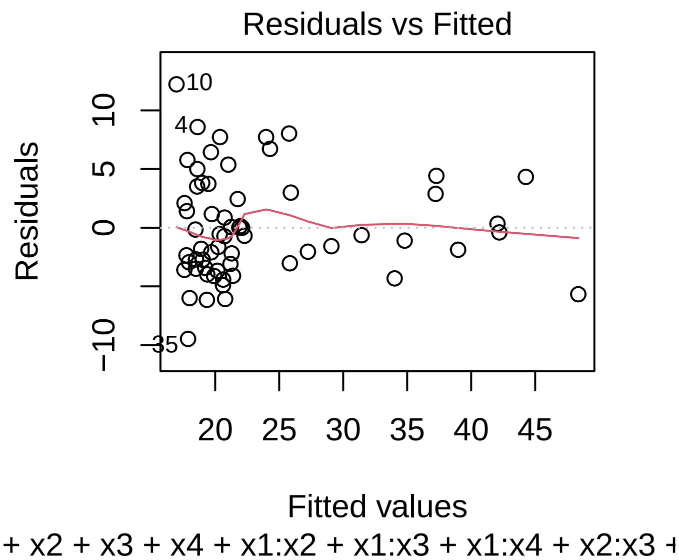
```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x1:x2 + x1:x3 + x1:x4 +
##     x2:x3 + x2:x4 + x3:x4, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4804 -3.0766 -0.6635  2.9625 12.2221
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.88376   23.17863   0.685  0.49616
## x1            0.18696    0.78447   0.238  0.81255
## x2            0.37921    0.06332   5.989 1.89e-07 ***
## x3          -11.99940   67.31148  -0.178  0.85919
## x4           -8.86442   35.62553  -0.249  0.80446
## x1:x2         0.01155    0.00869   1.329  0.18955
## x1:x3              NA         NA      NA       NA
## x1:x4        -1.11525    1.14847  -0.971  0.33592
## x2:x3              NA         NA      NA       NA
## x2:x4        -0.38547    0.11962  -3.222  0.00218 **
## x3:x4        72.85976  103.15353   0.706  0.48308
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.683 on 53 degrees of freedom
## Multiple R-squared:  0.7496, Adjusted R-squared:  0.7118
## F-statistic: 19.83 on 8 and 53 DF,  p-value: 1.947e-13
```

Here, two interactions, x_1x_3 and x_2x_3 are dropped because of multicollinearity. So,

$\hat{y} = 15.88376 + 0.18696x_1 + 0.37921x_2 - 11.99940x_3 - 8.86442x_4 + 0.01155x_1x_2 - 1.11525x_1x_4 - 0.38547x_2x_4 + 72.85976x_3x_4$. R^2 is 0.7496 and the adjusted one is 0.7118.

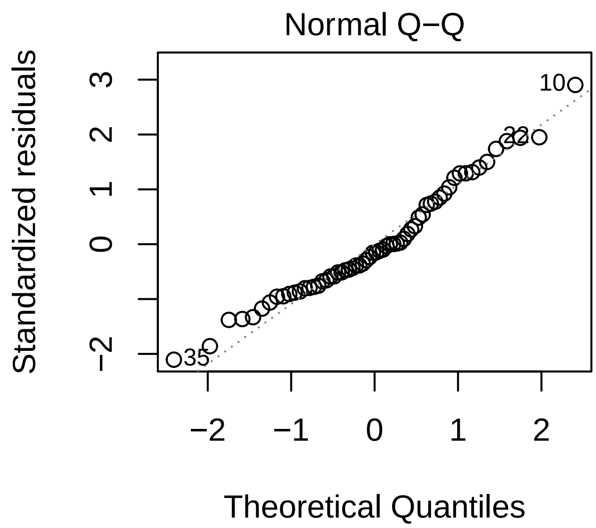
Checking for Model Adequacy

```
plot(fit1,1)
```



This figure ensures the constant variance for we cannot see any patterns.

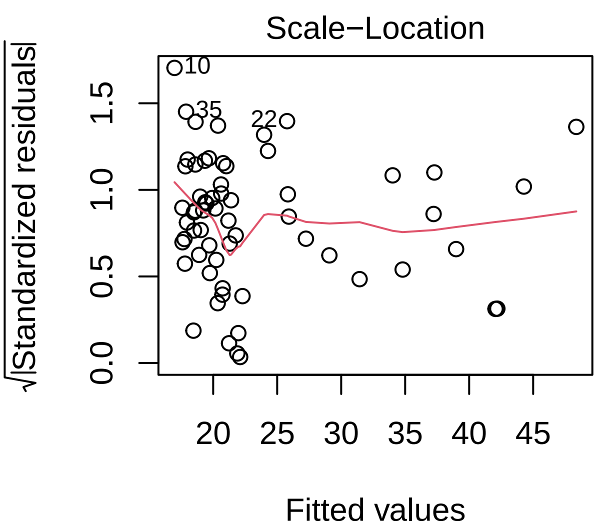
```
plot(fit1,2)
```



+ x2 + x3 + x4 + x1:x2 + x1:x3 + x1:x4 + x2:x3 +

This figure ensures the normality for the line is almost straight.

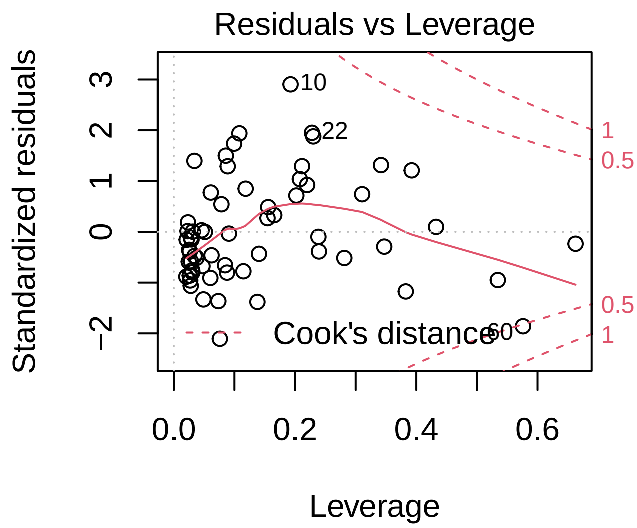
```
plot(fit1,3)
```



+ x2 + x3 + x4 + x1:x2 + x1:x3 + x1:x4 + x2:x3 +

We don't think there is any outliers in this figure.

```
plot(fit1,5)
```



+ x2 + x3 + x4 + x1:x2 + x1:x3 + x1:x4 + x2:x3 +

Also, we don't think there is any outliers in this figure.

Thus, we think this model has model adequacy.

Test of Significance of the Full Regression Model

```
X <- add_column(data, x0 = rep(1,nrow(data)), .before = 'x1')
X <- cbind(X, data$x1*data$x2, data$x1*data$x4, data$x2*data$x4, data$x3*data$x4)
X <- X[,-c(6)]
Y <- data$y
SST <- t(Y)%*%Y - (sum(Y)^2)/nrow(data)
beta <- na.omit(fit1$coefficients)
SSE <- t(Y)%*%Y - t(as.vector(beta))%*%t(X)%*%Y
SSR <- SST - SSE
MSR <- SSR/(length(beta) - 1)
MSE <- SSE/(nrow(data) - length(beta))
f <- MSR/MSE
pvalue <- 1 - pf(f,length(beta) - 1, nrow(data) - length(beta))
pvalue
```

```
##               [,1]
## [1,] 1.947331e-13
```

```
f
```

```
##          [,1]
## [1,] 19.8342
```

Reject H_0 because $p - value < 0.001$. So, we conclude that at least one of these regressors contributes significantly to the model, which implies that the pressure drop in a screen plate bubble column is related to at least of these factors. (This can be seen directly from the summary table of this model.)

Part b.

Reducecd Model

```
fit1 <- lm(y~x1+x2+x3+x4+x1:x2+x1:x4+x2:x4+x3:x4,data)
fit2 <- lm(y~x1+x2+x3+x4,data)
```

Test of Significance of the interactions

```
anova(fit2,fit1)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x3 + x4
## Model 2: y ~ x1 + x2 + x3 + x4 + x1:x2 + x1:x4 + x2:x4 + x3:x4
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     57 1432.8
## 2     53 1162.4  4    270.37 3.0819 0.02352 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
X2 <- add_column(data, x0 = rep(1,nrow(data)), .before = 'x1')
X2 <- X2[,-c(6)]
beta2 <- as.vector(fit2$coefficients)
SSRreduced <- t(beta2)%*%t(X2)%*%Y - (sum(Y)^2)/nrow(data)
fpartial <- ((SSR-SSRreduced)/(length(beta)-length(beta2)))/MSE
fpartial
```

```
##          [,1]
## [1,] 3.081881
```

```
pvaluepartial <- 1 - pf(fpartial,length(beta) - length(beta2), nrow(data) - length(beta))
pvaluepartial
```

```
##            [,1]
## [1,] 0.02352117
```

Reject H_0 because $p - value < 0.05$. We conclude that at least one of these interactions contributes significantly to the model.

Part c. Finding the Best Model

Test x_3x_4

```
fit3 <- lm(y~x1+x2+x3+x4+x1:x2+x1:x4+x2:x4,data)
anova(fit3, fit1)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x3 + x4 + x1:x2 + x1:x4 + x2:x4
## Model 2: y ~ x1 + x2 + x3 + x4 + x1:x2 + x1:x4 + x2:x4 + x3:x4
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     54 1173.4
## 2     53 1162.4  1    10.942 0.4989 0.4831
```

Don't reject H_0 because $p - value > 0.05$. So we drop x_3x_4.

Test x_2x_4

```
fit4 <- lm(y~x1+x2+x3+x4+x1:x2+x1:x4,data)
anova(fit4, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x3 + x4 + x1:x2 + x1:x4
## Model 2: y ~ x1 + x2 + x3 + x4 + x1:x2 + x1:x4 + x2:x4
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     55 1401.8
## 2     54 1173.4  1    228.39 10.511 0.002036 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Reject H_0 because $p - value < 0.05$. So we keep x_2x_4.

Test x_1x_4

```
fit5 <- lm(y~x1+x2+x3+x4+x1:x2+x2:x4,data)
anova(fit5, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x3 + x4 + x1:x2 + x2:x4
## Model 2: y ~ x1 + x2 + x3 + x4 + x1:x2 + x1:x4 + x2:x4
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     55 1193.2
## 2     54 1173.4  1    19.837 0.9129 0.3436
```

Don't reject H_0 because $p - value > 0.05$. So we drop x_1x_4.

Test x_1x_2

```
fit6 <- lm(y~x1+x2+x3+x4+x2:x4,data)
anova(fit6, fit5)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x3 + x4 + x2:x4
## Model 2: y ~ x1 + x2 + x3 + x4 + x1:x2 + x2:x4
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     56 1225.5
## 2     55 1193.2  1    32.307 1.4892 0.2275
```

Don't reject H_0 because $p - value > 0.05$. So we drop x_1x_2.

Test x_1

```
fit7 <- lm(y~x2+x3+x4+x2:x4,data)
anova(fit7, fit6)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x2 + x3 + x4 + x2:x4
## Model 2: y ~ x1 + x2 + x3 + x4 + x2:x4
```

```
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     57 1236.8
## 2     56 1225.5  1    11.262 0.5146 0.4761
```

Don't reject H_0 because $p - value > 0.05$. So we drop x_1.

Test x_3

```
fit8 <- lm(y~x2+x4+x2:x4,data)
anova(fit8, fit7)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x2 + x4 + x2:x4
## Model 2: y ~ x2 + x3 + x4 + x2:x4
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     58 1480.4
## 2     57 1236.8  1     243.6 11.227 0.001435 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Reject H_0 because $p - value > 0.05$. So we keep x_3.

The Best Fitting

```
fitbest <- lm(y~x2+x3+x4+x2:x4,data)
summary(fitbest)
```

```
##
## Call:
## lm(formula = y ~ x2 + x3 + x4 + x2:x4, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.959 -3.358 -1.131  3.040 11.646
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.52261    4.03964   0.377  0.70763
## x2           0.38056    0.06084   6.255 5.47e-08 ***
## x3          34.51062   10.29961   3.351  0.00144 **
## x4           9.52471    2.96093   3.217  0.00214 **
## x2:x4       -0.30472    0.09056  -3.365  0.00137 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.658 on 57 degrees of freedom
## Multiple R-squared:  0.7336, Adjusted R-squared:  0.7149
## F-statistic: 39.24 on 4 and 57 DF,  p-value: 9.297e-16
```

The best fitting is $\hat{y} = 1.52261 + 0.38056x_2 + 34.51062x_3 + 9.52471x_4 - 0.30472x_2x_4$. The R^2 is 0.7336 and the adjusted one is 0.7149. We don't think there exists a significant difference between the best fitting and the original one.

Part d. Confidence Interval

```r
newx1 <- c(5.0, 10.0)
newx2 <- c(10.0, 3.0)
newx3 <- c(0.5, 0.25)
newx4 <- c(0.75, 0.85)
CI <- predict(fitbest, data.frame(x1 = newx1, x2 = newx2,
                                  x3 = newx3, x4 = newx4), interval='confidence')
CI
```

```
##        fit      lwr      upr
## 1 27.44168 24.00618 30.87717
## 2 18.61092 15.76975 21.45208
```

So, CI for the first point is $(24.00618, 30.87717)$ and that for the second point is $(15.76975, 21.45208)$.

Part e. Prediction Interval

```r
PI <- predict(fitbest, data.frame(x1 = newx1, x2 = newx2,
                                  x3 = newx3, x4 = newx4), interval='prediction')
PI
```

```
##        fit       lwr      upr
## 1 27.44168 17.501496 37.38186
## 2 18.61092  8.860183 28.36165
```

So, PI for the first point is $(17.501496, 37.38186)$ and that for the second point is $(8.860183, 28.36165)$.