Data 03: bộ dữ liệu gồm 39 quan trắc được thực hiện trên vài đoạn đường cao tốc ở tiểu bang Minnesota vùng Trung Tây của Hoa Kỳ, gồm các biến sau :
– X1 : chiều dài đoạn đường (dặm) ;
– X2 : lượng giao thông trung bình hàng ngày (nghìn xe) ;
– X3 : tỷ lệ % xe tải trên tổng số ;
– X4 : tốc độ giới hạn cho phép (dặm/giờ) ;
– X5 : chiều rộng làn đường (bước chân) ;
– X6 : chiều rộng làn đường khẩn cấp (bước chân) ;
– X7 : số lần thay đổi làn đường trống (trên dặm đường của một đoạn đường cao tốc) ;
– X8 : số lần thay đổi làn đường được báo (trên dặm đường) ;
– X9 : số cửa vào đoạn đường cao tốc ;
– X10 : tổng số làn đường (trên hai chiều của đường cao tốc) ;
– X11 : 1 nếu là tuyến đường liên thông xa lộ và cao tốc , 0 nếu ngược lại ;
– X12 : 1 nếu là tuyến đường lớn của cao tốc , 0 nếu ngược lại ;
– X13 : 1 nếu là tuyến đường cao tốc chính, 0 nếu ngược lại.

Sử dụng phương pháp stepwise và tiêu chuẩn AIC/BIC:

Với AIC:

```
###stepAIC : stepwise regression proceeded
mod <- lm(y_i ~ ., data = accident)# Full model
modAIC <- MASS::stepAIC(mod, k = 2, direction = "backward", trace =
FALSE) # With AIC  k = 2
summary(modAIC)
```

Kết quả cho mô hình chọn bởi tiêu chuẩn AIC:

```
Call:
lm(formula = y_i ~ x_i.1 + x_i.4 + x_i.8 + x_i.9 + x_i.12, data =
accident)

Residuals:
     Min       1Q   Median       3Q      Max
-1.93493 -0.79927  0.00224  0.71824  2.50054

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.10349    2.61780   3.860   0.0005 ***
x_i.1       -0.06871    0.02375  -2.894   0.0067 **
x_i.4       -0.11023    0.04175  -2.640   0.0126 *
x_i.8        0.79395    0.37254   2.131   0.0406 *
x_i.9        0.06545    0.03051   2.145   0.0394 *
x_i.12      -0.72254    0.41278  -1.750   0.0893 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.086 on 33 degrees of freedom
Multiple R-squared:  0.7404,     Adjusted R-squared:  0.7011
F-statistic: 18.82 on 5 and 33 DF,  p-value: 8.002e-09
```

Vậy với tiêu chuẩn AIC ta chọn được mô hình dự đoán tỷ lệ tai nạn gồm 5 biến là X1, X4, X8, X9, X12 với mức ý nghĩa 10%. Còn với mức ý nghĩa 5% thì mô hình không còn biến X12.
– X1 : chiều dài đoạn đường (dặm) ;
– X4 : tốc độ giới hạn cho phép (dặm/giờ) ;
– X8 : số làn đường thay đổi (báo hiệu) trên đoạn đường cao tốc ;
– X9 : số cửa vào đoạn đường cao tốc ;
– X12 : 1 nếu là tuyến đường lớn của cao tốc , 0 nếu ngược lại ;

Với BIC:

```
modBIC <- MASS::stepAIC(mod, k = log(nrow(accident)), direction =
"backward", trace = FALSE)
summary(modBIC)
```

Kết quả cho mô hình chọn bởi tiêu chuẩn BIC:

```
Call:
lm(formula = y_i ~ x_i.1 + x_i.4 + x_i.9, data = accident)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0344 -0.7593  0.1639  0.8400  2.3836

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.61318    2.63889   3.643 0.000865 ***
x_i.1       -0.07352    0.02410  -3.050 0.004342 **
x_i.4       -0.10871    0.04306  -2.525 0.016268 *
x_i.9        0.10112    0.02738   3.693 0.000752 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.136 on 35 degrees of freedom
Multiple R-squared:  0.6986,     Adjusted R-squared:  0.6728
F-statistic: 27.05 on 3 and 35 DF,  p-value: 3.121e-09
```

Tiêu chuẩn BIC ta chọn được mô hình dự đoán tỷ lệ tai nạn gồm ít biến hơn, gồm 3 biến là X1, X4, X9 với mức ý nghĩa 5%.
– X1 : chiều dài đoạn đường (dặm) ;
– X4 : tốc độ giới hạn cho phép (dặm/giờ) ;
– X9 : số cửa vào đoạn đường cao tốc ;
Mô hình:

$$\hat{Y} = 9.61318 - 0.07352 * X1 - 0.10871 * X4 + 0.10112 * X9$$

Ta so sánh hai mô hình được chọn bởi tiêu chuẩn AIC và BIC bằng kiểm định F từng phần như bên dưới, liệu ta có thể bỏ đi biến X8 và X12 hay không?

```
> anova(modBIC, modAIC)
Analysis of Variance Table
```

```
Model 1: y_i ~ x_i.1 + x_i.4 + x_i.9
Model 2: y_i ~ x_i.1 + x_i.4 + x_i.8 + x_i.9 + x_i.12
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     35 45.17
2     33 38.91  2    6.2601 2.6546 0.0853 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Kết quả từ bảng ANOVA cho thấy ta không thể bác bỏ giả thuyết H0: X8 = X12 = 0 với mức ý nghĩa 5%, do đó ta tạm thời chấp nhận mô hình modBIC có ba biến là X1, X4 và X9. Vậy tỷ lệ tai nạn có thể được dự báo bởi:

– X1 : chiều dài đoạn đường (dặm) ;

– X4 : tốc độ giới hạn cho phép (dặm/giờ) ;

– X9 : số cửa vào đoạn đường cao tốc ;

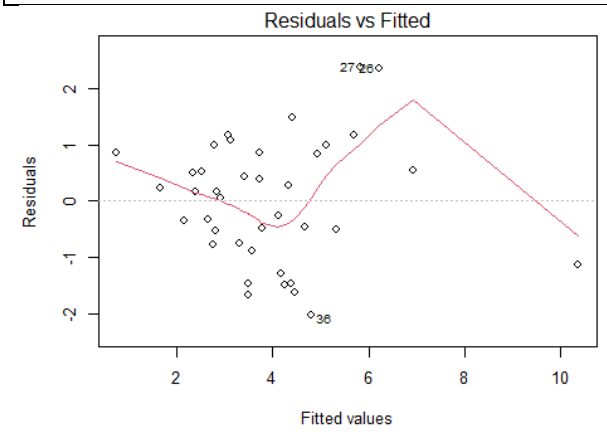
$$\hat{Y} = 9.61318 - 0.07352 * X1 - 0.10871 * X4 + 0.10112 * X9$$

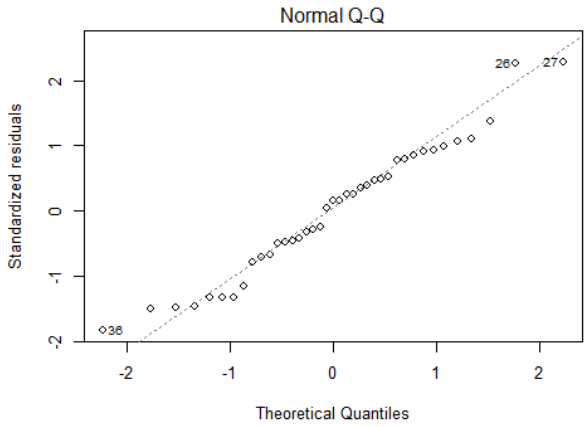
Ta tiến hành kiểm tra phân phối của sai số cho mô hình được chọn bên trên:

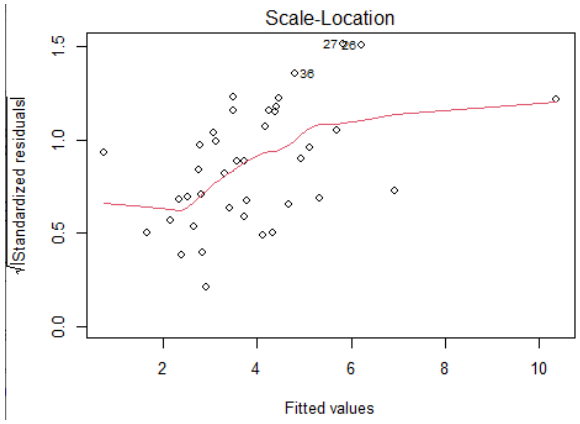
```
shapiro.test(residuals(modBIC))
Shapiro-Wilk normality test

data:  residuals(modBIC)
W = 0.97525, p-value = 0.5346
```

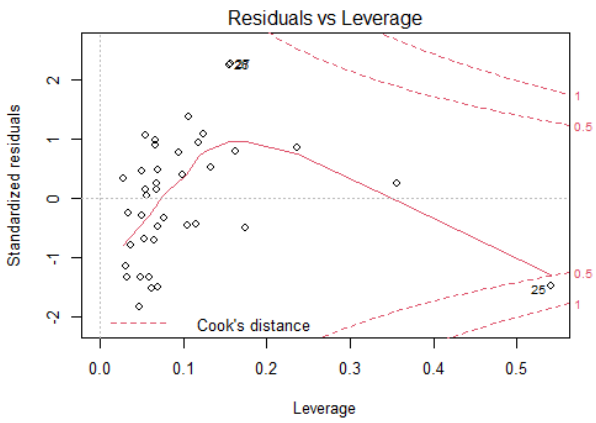
Trị số P lớn (0.5346) nên ta không thể bác bỏ giá thuyết H0 là sai số tuân theo phân phối chuẩn, vậy có thể “chấp nhận” rằng sai số có phân phối chuẩn.

```
op <- par(mfrow=c(2,2))
plot(modBIC)
```






Scale-Location
√|Standardized residuals|
Fitted values
27
26
36


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
25