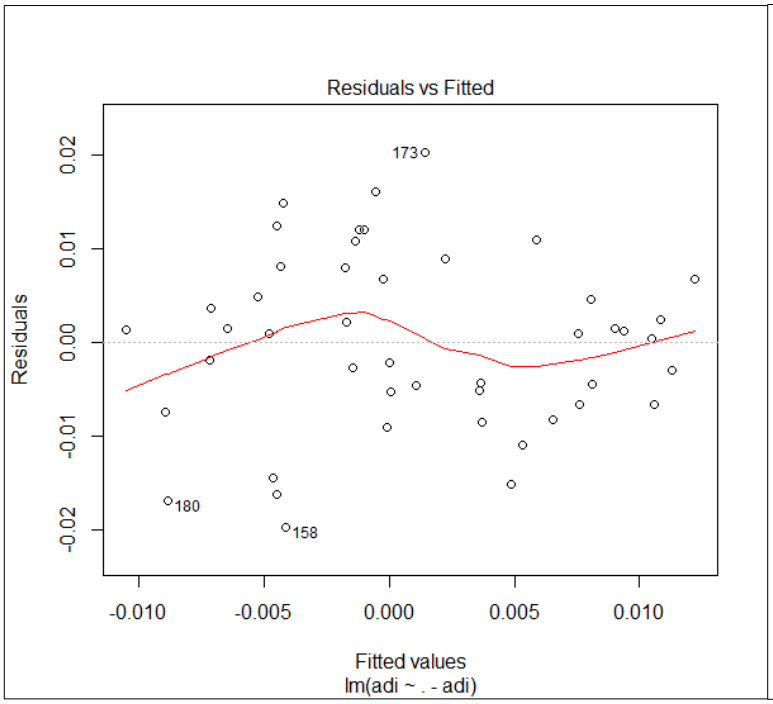
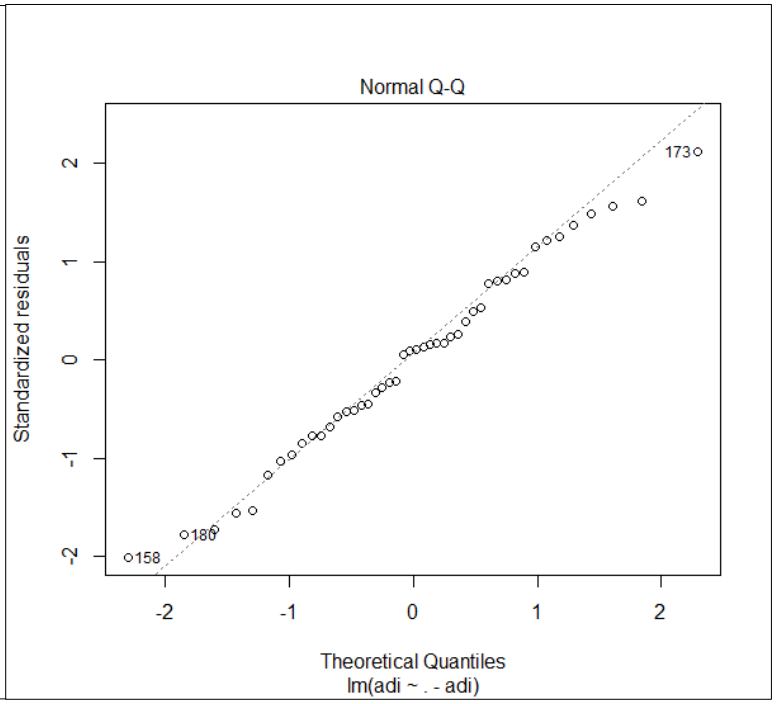
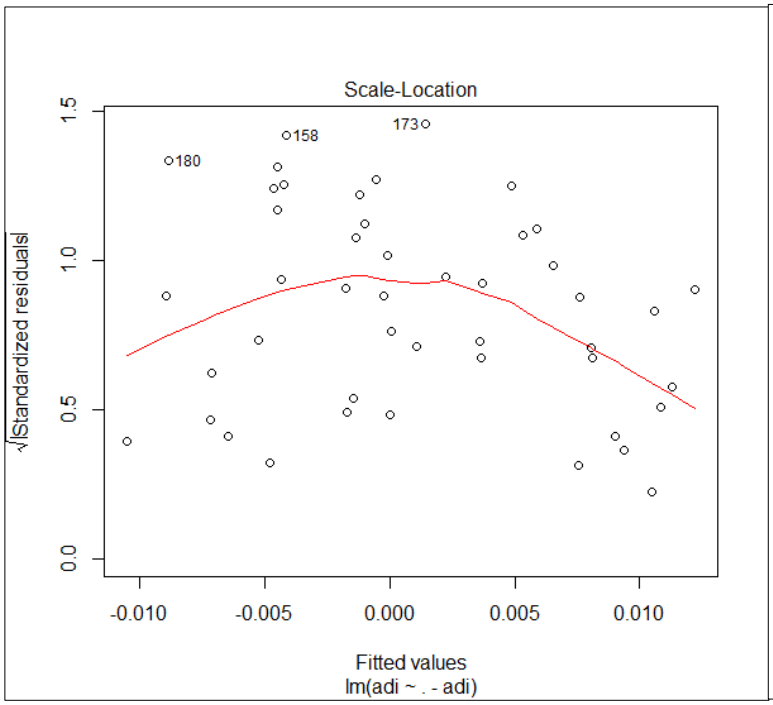
```
> filenames=list.files(path="C:/Users/Oksana/Desktop/statistics/data", ful
l.names=TRUE)
> y<-read.csv(file=filenames[1],header=F)[,c(1,6)]
> colnames(y)<-c("data", unlist(strsplit(filenames[1],"[_.]"))[2])
> for(i in 2:10) {x0<-read.csv(file=filenames[i],header=F)[,c(1,6)]; colna
mes(x0)<-c("data", unlist(strsplit(filenames[i],"[_.]"))[2]);y<-merge(y, x
0, by = "data")}
> n<-nrow(y)
> Data<-y[-nrow(y),-1]
> Data$adi<-y$adi[-1]
> nn<-nrow(Data)
```

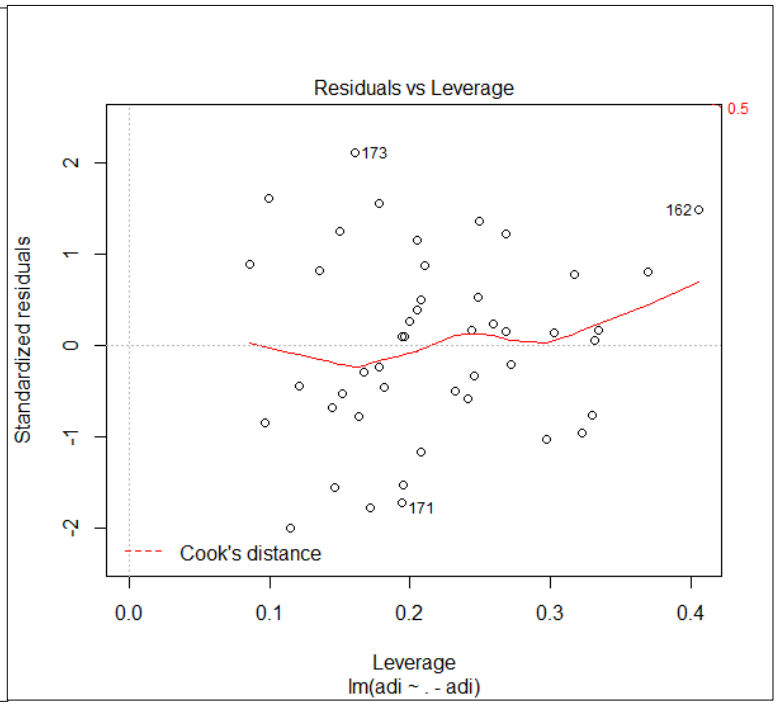
МОДЕЛЬ ЗА ОСТАННІМИ ДАНИМИ

```
> model1<-lm(adi~.-adi,data=Data[(nn-59):(nn-10),])
> summary(model1)

Call:
lm(formula = adi ~ . - adi, data = Data[(nn - 59):(nn - 10),
    ])

Residuals:
     Min       1Q   Median       3Q      Max
-1.01605 -0.33022 -0.00743  0.36222  0.85120

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.48799    6.96354   0.214  0.83188
adm          0.24580    0.27189   0.904  0.37138
adp          0.37269    0.21109   1.766  0.08511 .
adsk         0.24415    0.11022   2.215  0.03250 *
adt          0.10150    0.12554   0.809  0.42357
aee          0.08929    0.30115   0.297  0.76838
aep         -0.58991    0.27799  -2.122  0.04007 *
aes          1.35032    0.78282   1.725  0.09226 .
aet         -0.32722    0.10430  -3.137  0.00320 **
afl          0.43256    0.13205   3.276  0.00218 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5183 on 40 degrees of freedom
Multiple R-squared:  0.877,    Adjusted R-squared:  0.8494
F-statistic:  31.7 on 9 and 40 DF,  p-value: 1.796e-15
```

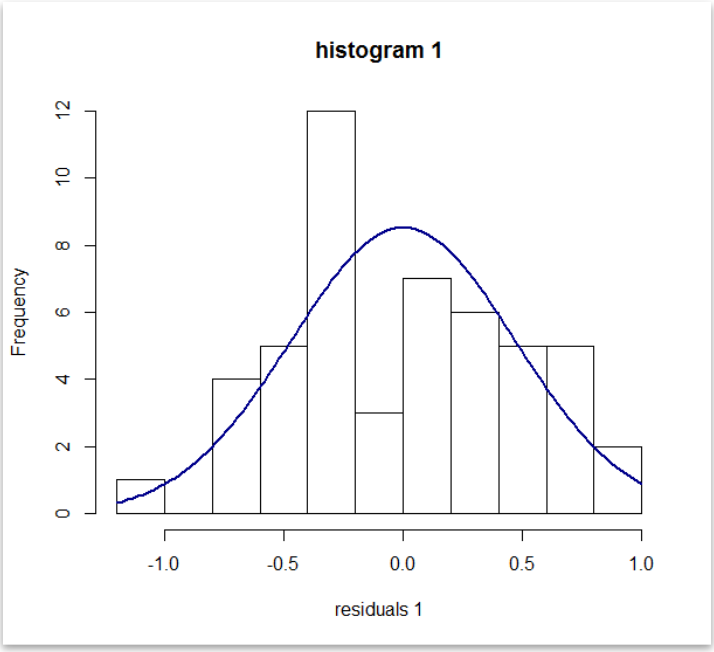








Графіки для моделі за останніми даними

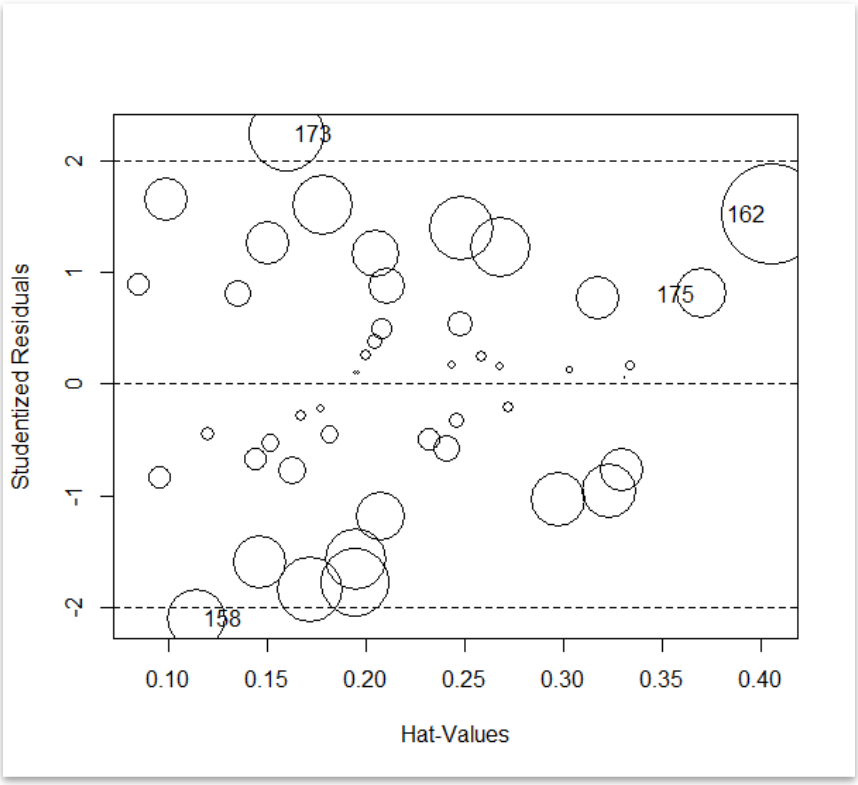


Проведемо тест Шапіро-Вілка для перевірки нормальності.

```
> shapiro.test(model1$residuals)

        Shapiro-Wilk normality test

data:  model1$residuals
W = 0.97464, p-value = 0.3537
```

p-value тесту суттєво більше досягнутого рівня значущості $\propto = 0.05$, отже, можна прийняти гіпотезу про нормальність залишків.

Отримані значення коефіцієнта детермінації R = 0.877 та p-value = 1.796e-15 свідчать на користь нашої моделі та залежності відгука від регресорів.



Спробуємо викинути точки, які можна вважати викидами згідно з діаграмою впливу та аналізом залишків.

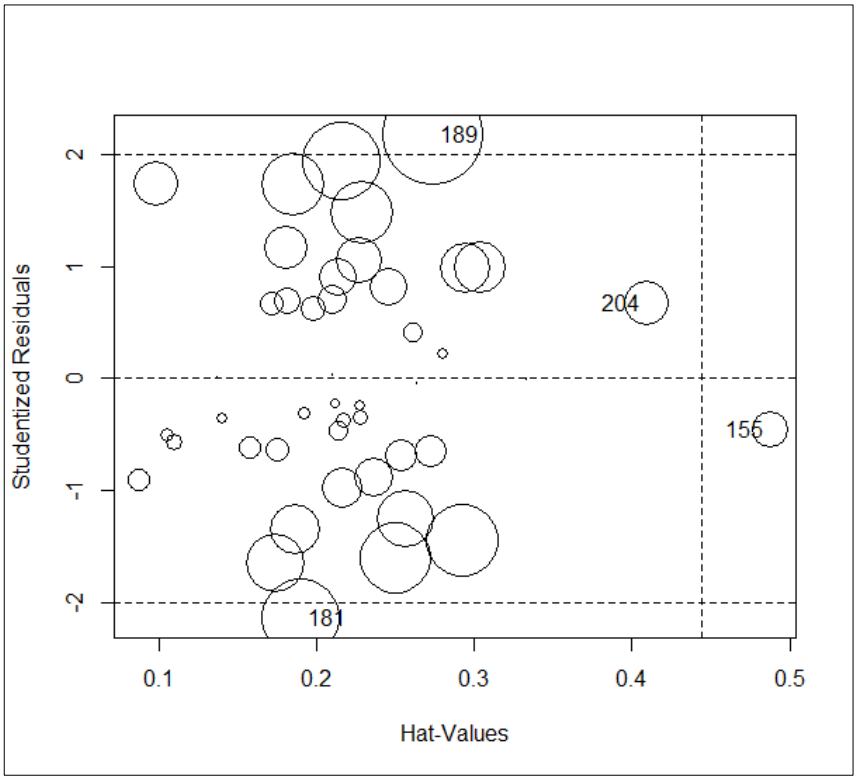
```
> influencePlot(model1)

StudRes         Hat         CookD
156  1.786875 0.3931625 0.19611393
159 -2.234981 0.1538650 0.08258538
179 -1.664345 0.2516503 0.08920200
193  1.792833 0.1144731 0.03937154
203  1.276478  0.3523905 0.08728852
```

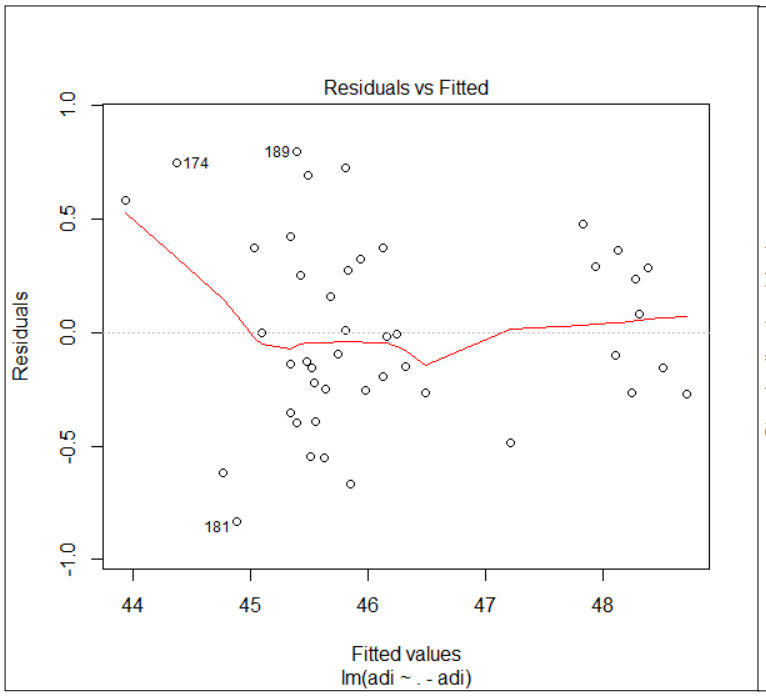
```
> Data2<-Data[(nn-59):(nn-10),]
> Data2<-Data2[-(c(156,159,179,193,203)-154),]
> model2<-lm(adi~.-adi,data=Data2)
> summary(model2)

Call:
lm(formula = adi ~ . - adi, data = Data2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.83285 -0.26221 -0.09274  0.28947  0.80009

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.08514    6.72603  -0.161 0.872757
adm          0.22974    0.25100   0.915 0.366302
adp          0.30790    0.19818   1.554 0.129270
adsk         0.19584    0.11508   1.702 0.097651 .
adt          0.13969    0.11183   1.249 0.219921
aee          0.01332    0.28869   0.046 0.963455
aep         -0.27963    0.30388  -0.920 0.363773
aes          0.55665    0.91243   0.610 0.545756
aet         -0.35405    0.09191  -3.852 0.000478 ***
afl          0.56740    0.14507   3.911 0.000403 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4541 on 35 degrees of freedom
Multiple R-squared:  0.9028, Adjusted R-squared:  0.8779
F-statistic: 36.14 on 9 and 35 DF,  p-value: 3.992e-15
```

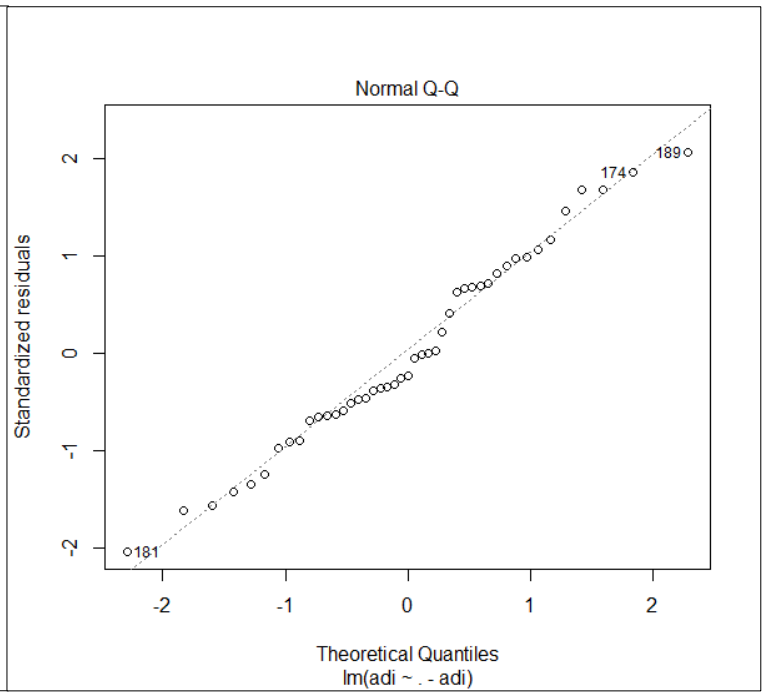
```
> influencePlot(model2)
       StudRes       Hat      CookD
155 -0.4561932 0.4880438 0.02029843
181 -2.1397682 0.1900785 0.09748662
189  2.1750829 0.2737517 0.16115092
204  0.6736042 0.4094220 0.03195473
```

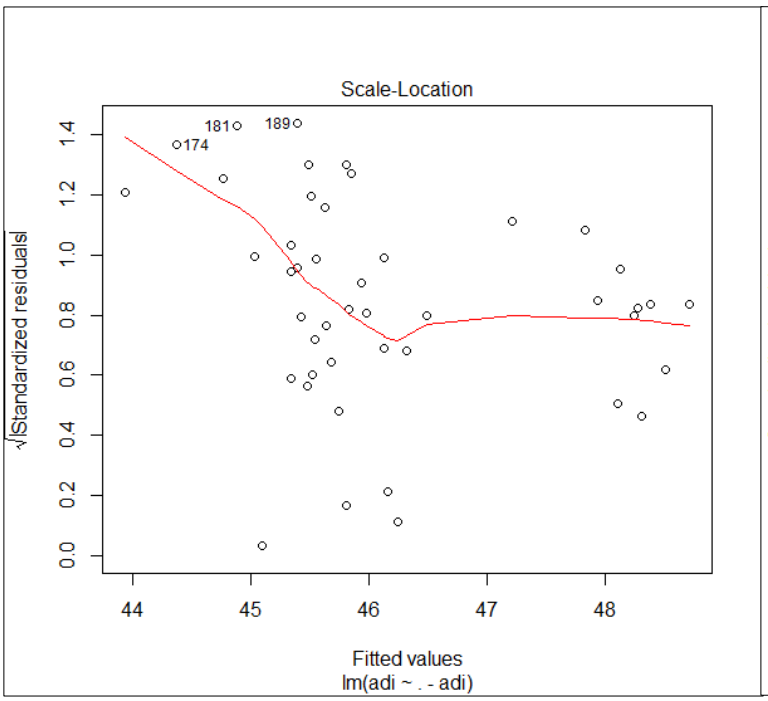
```
> shapiro.test(model2$residuals)

Shapiro-Wilk normality test

data:  model2$residuals
W = 0.97682, p-value = 0.4968
```

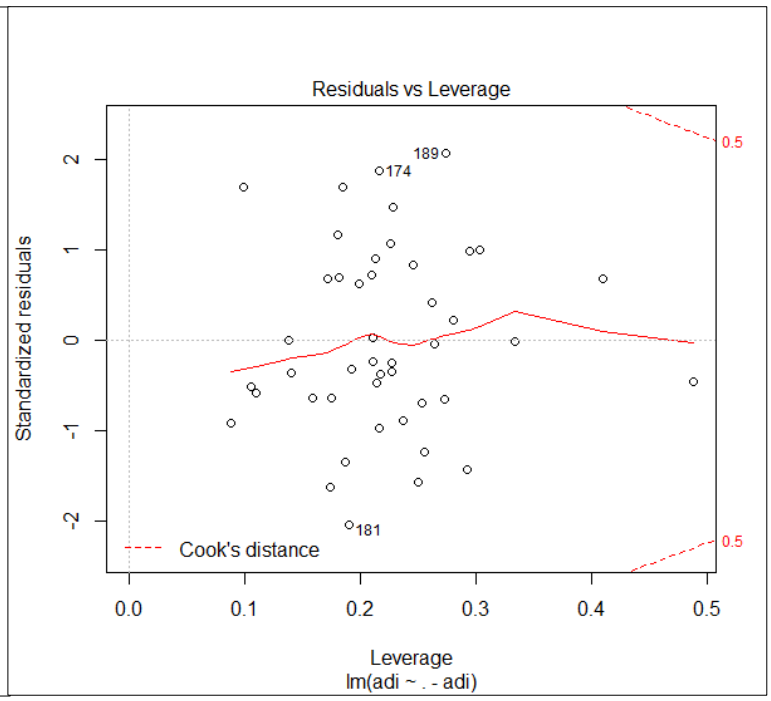
(На жаль, для моделей за останнім и даними не вдалося підібрати варіант з хорошими характеристика ми,при якому використовувались би регресори з великою статистичною значущістю.)





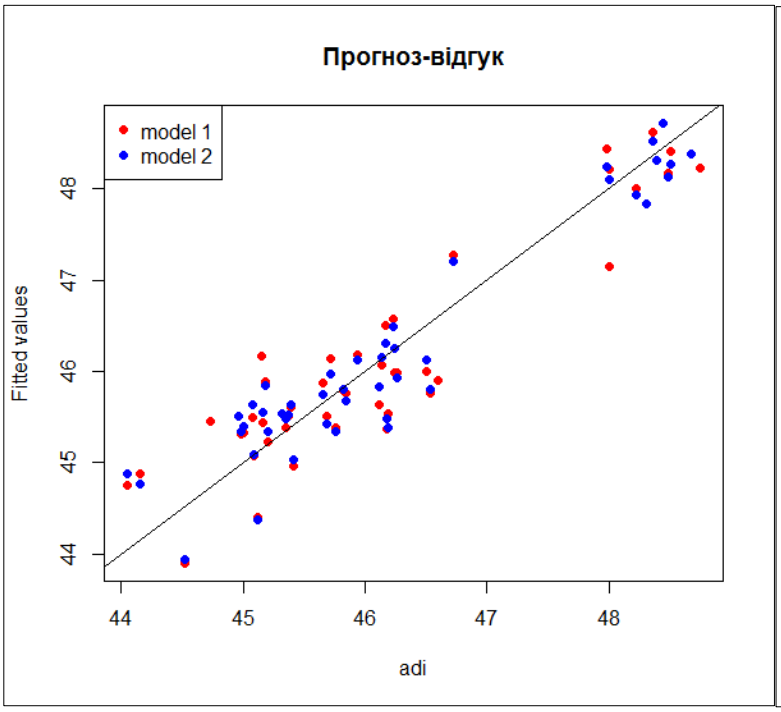


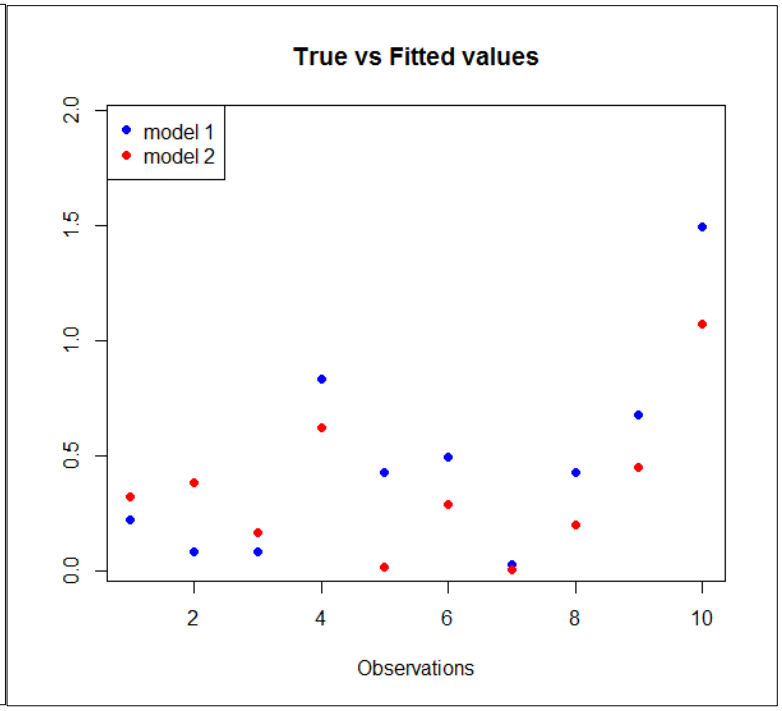




Графіки для моделі за останніми даними з викинутими спостереженнями зі значним впливом.

```
> plot(Data$adi[(n1-59):(n1-10)], model1$fitted.values, pch = 19, col = "red",
main = "Прогноз-відгук", xlab = "adi", ylab = "Fitted values")
> points(Data2$adi, model2$fitted.values, col = "blue")
> points(Data2$adi, model2$fitted.values, pch = 19, col = "blue")
> legend("topleft", c("model 1", "model 2"), pch = 19, col = c("red", "blue"))
> abline(0, 1)
> xadi1<-predict(model1, Data[(n1-9):n1,])
> xadi2<-predict(model2, Data[(n1-9):n1,])
> points(1:10, abs(xadi- xadi1), pch = 19, col = "blue")
> points(1:10, abs(xadi- xadi2), pch = 19, col = "red")
> legend("topleft", c("model 1", "model 2"), pch = 19, col = c("blue", "red"))
```

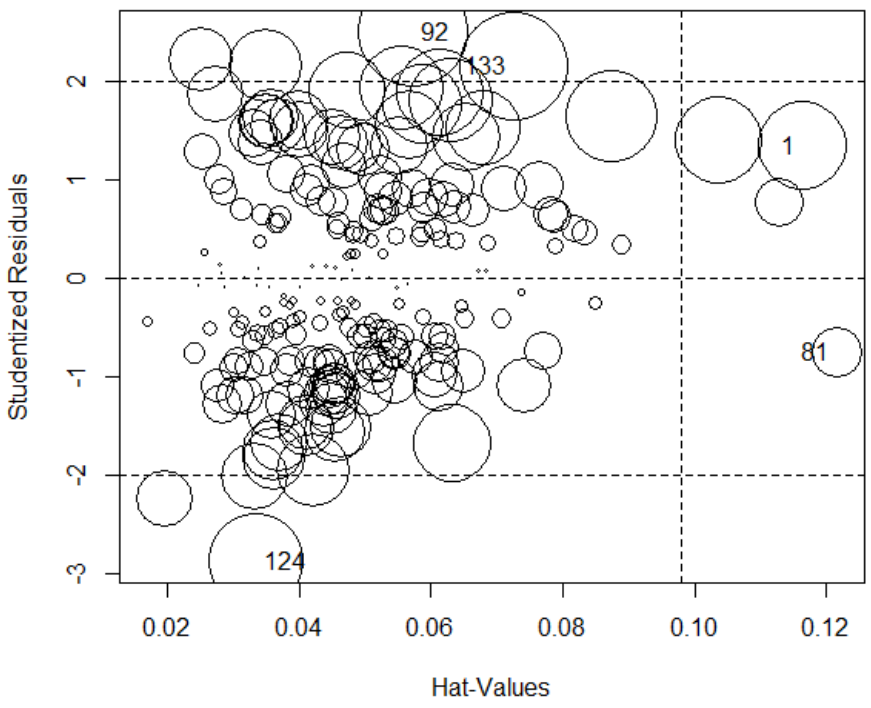




При повторному моделюванні досить сильно змінилась оцінка вільного коефіцієнта, але значної різниці в прогнозах моделей немає. Коефіцієнт детермінації та p-value також покращились при повторному моделюванні: 0.877 та 1.796e-15 для першої моделі, 0.9028 та 3.992e-15 - для другої відповідно. Обидві моделі задовольняють умову нормальності залишків.

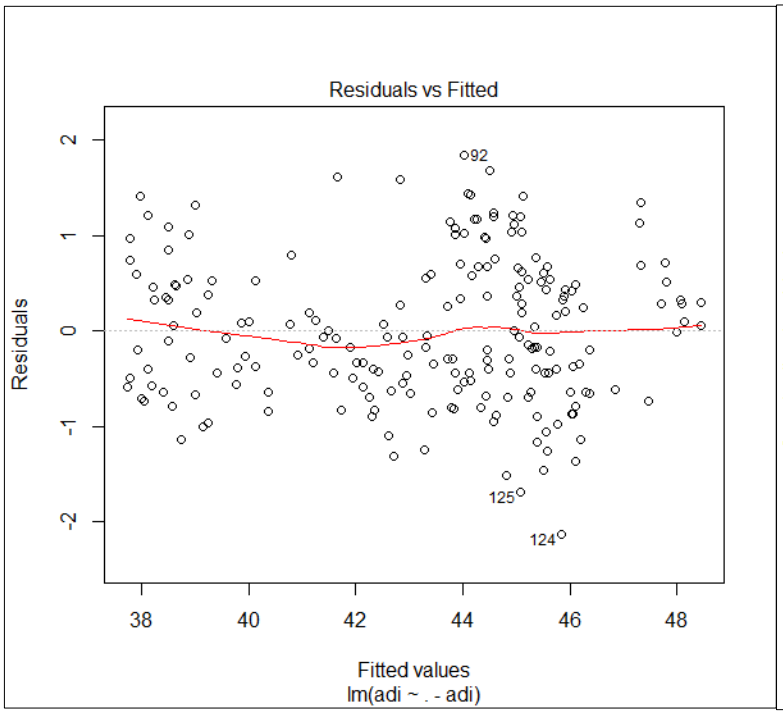
Модель за повними даними

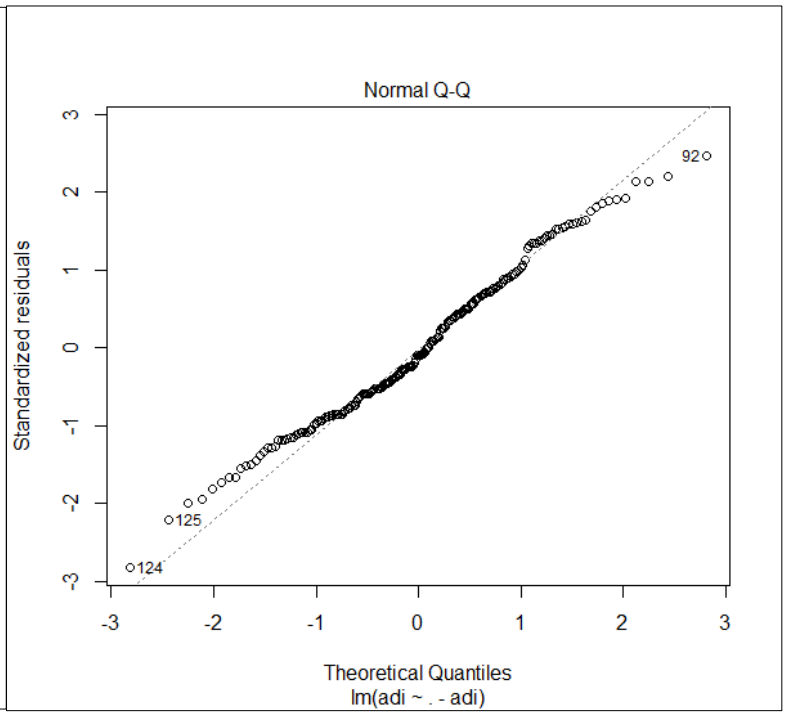
```
> modelf1<-lm(adi~.-adi,data=Data[1:(nn-10),])
> summary(modelf1)

Call:
lm(formula = adi ~ . - adi, data = Data[1:(nn - 10), ])

Residuals:
     Min       1Q   Median       3Q      Max
-2.13067 -0.57694 -0.07082  0.53188  1.84405

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.211094   3.137307   1.342 0.181079
adm          0.485400   0.079297   6.121 5.05e-09 ***
adp          0.140635   0.072439   1.941 0.053656 .
adsk         0.263215   0.052012   5.061 9.66e-07 ***
adt          0.220021   0.038995   5.642 5.88e-08 ***
aee         -0.070103   0.108199  -0.648 0.517810
aep          0.008799   0.110597   0.080 0.936667
aes         -1.109784   0.288715  -3.844 0.000164 ***
aet          0.100493   0.034223   2.936 0.003722 **
afl          0.112633   0.043789   2.572 0.010854 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7695 on 194 degrees of freedom
Multiple R-squared:  0.9352,   Adjusted R-squared:  0.9322
F-statistic: 311.1 on 9 and 194 DF,  p-value: < 2.2e-16
```



Аналіз впливу:

```
> influencePlot(modelf1)
       StudRes         Hat        CookD
1    1.3488570 0.11652931 0.023897054
81  -0.7426461 0.12168017 0.007658364
92   2.5014300 0.05731804 0.037041676
124 -2.8685576 0.03355757 0.027545690
133  2.1633190 0.07241340 0.035854589
```






```
> shapiro.test(modelf1$residuals)

        Shapiro-Wilk normality test

data:  modelf1$residuals
W = 0.98788, p-value = 0.08025
```

Бачимо, що отримане p-value тесту Шапіро-Вілка перевищує значення $\propto = 0.05$, що свідчить на користь гіпотези про нормальність розподілу залишків регресії.
Вилучимо спостереження, що можуть мати великий вплив на модель.

```
> Data2<-Data[1:(nn-10),]
> Data3<-Data2[-c(1,81,92,124,133),]
> modelf2<-lm(adi~.-adi,data=Data3)
```

На жаль, при такій модифікації даних тест Шапіро-Вілка вказує на те, що залишки не мають нормального розподілу.
(такий наслідок, здається, спричиняє в основному 124-те спостереження, що має найбільший вплив: саме при його вилученні порушується нормальність залишків):

```
> shapiro.test(modelf2$residuals)

Shapiro-Wilk normality test

data:  modelf2$residuals
W = 0.98448, p-value = 0.02733
```

```
> summary(modelf2)

Call:
lm(formula = adi ~ . - adi, data = Data3)

Residuals:
     Min       1Q   Median       3Q      Max
-1.71337 -0.52914 -0.07536  0.51022  1.67200

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.11346    3.11509   0.357 0.721162
adm          0.50736    0.07709   6.581 4.47e-10 ***
adp          0.12603    0.07016   1.796 0.074070 .
adsk         0.26974    0.05164   5.223 4.62e-07 ***
adt          0.21448    0.03847   5.575 8.46e-08 ***
aee         -0.05778    0.10381  -0.557 0.578446
aep          0.08477    0.10945   0.774 0.439628
aes         -1.30765    0.28902  -4.524 1.07e-05 ***
aet          0.08429    0.03392   2.485 0.013827 *
afl          0.16361    0.04362   3.751 0.000234 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7357 on 189 degrees of freedom
Multiple R-squared:  0.9413, Adjusted R-squared:  0.9386
F-statistic:   337 on 9 and 189 DF,  p-value: < 2.2e-16
```

ПОВНА МОДЕЛЬ ЗА ЧАСТИНОЮ РЕГРЕСОРІВ

Розглянемо модель за повними даними, в якій відкинемо регресори, що не мають високої статистичної значущості в початковій моделі за повними даними.

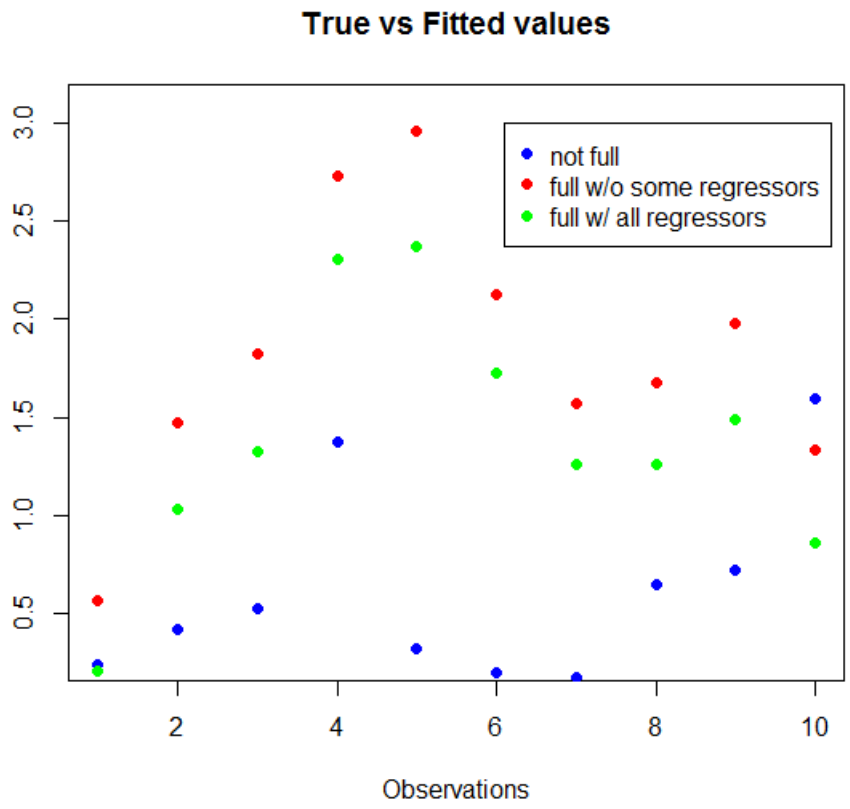
```
> model4<-lm(adi~adm+adsk+adt+aes+aet, data=Data2)
> summary(model4)
Call:
lm(formula = adi ~ adm + adsk + adt + aes + aet, data = Data2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.22163 -0.58927  0.00498  0.61815  1.78161

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.90592    0.98459  11.077  < 2e-16 ***
adm          0.52224    0.05521   9.458  < 2e-16 ***
adsk         0.23033    0.05233   4.401 1.76e-05 ***
adt          0.23378    0.03417   6.843 9.50e-11 ***
aes         -1.18833    0.13346  -8.904 3.47e-16 ***
aet          0.22665    0.01981  11.440  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8064 on 198 degrees of freedom
Multiple R-squared:  0.9274,       Adjusted R-squared:  0.9256
F-statistic: 505.7 on 5 and 198 DF,  p-value: < 2.2e-16
```

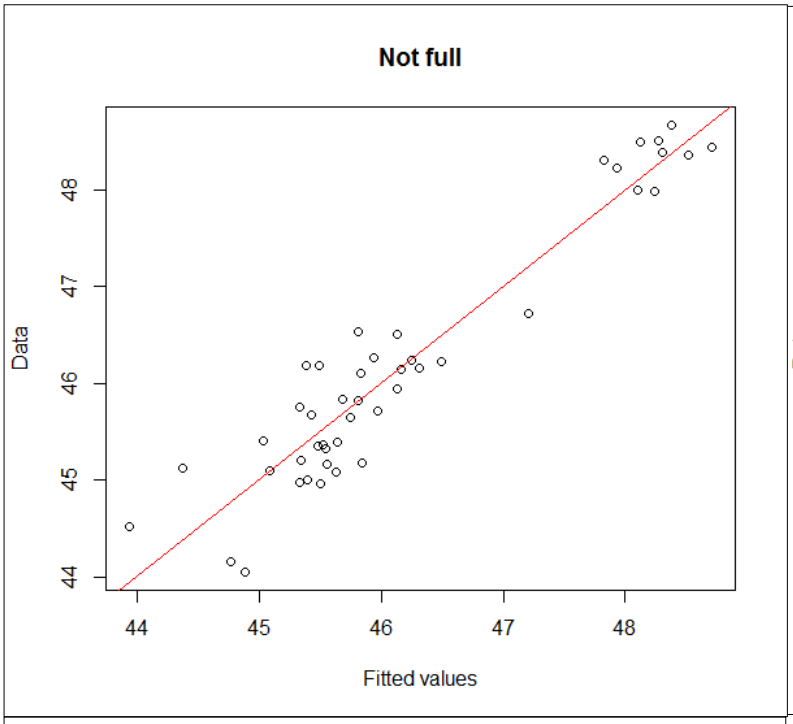
Ця модель має спостереження з можливим великим впливом (1,81,124,125,197), але можна переконатись, що при виключенні відповідних спостережень з моделі, суттєвих змін коефіцієнти регресії не зазнають:

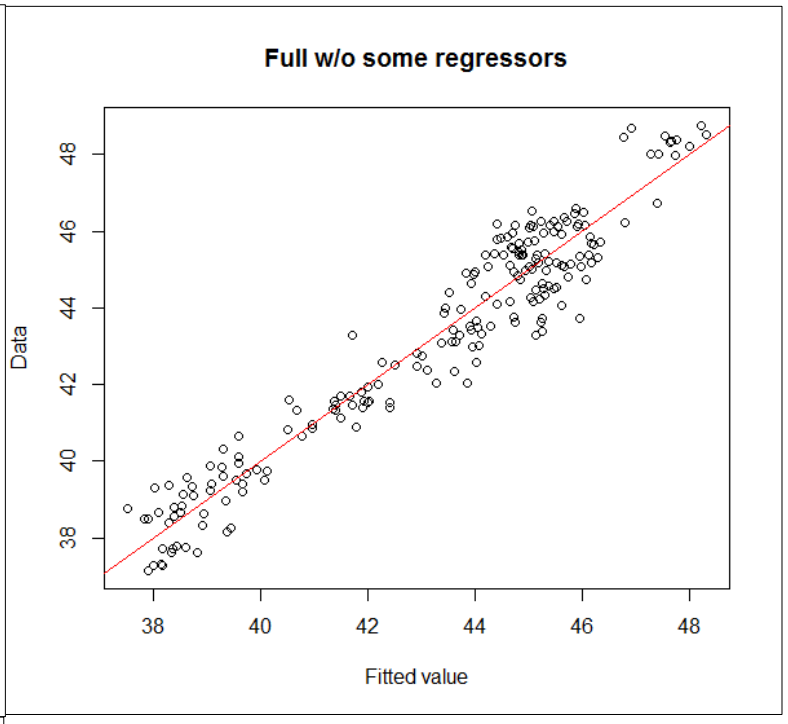
```
Call:
lm(formula = adi ~ adm + adsk + adt + aes + aet, data = dat)
     Residuals:
     Min       1Q   Median       3Q      Max
-1.85509 -0.56229 -0.01031  0.62387  1.79517
     Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.46865    0.97642  10.721  < 2e-16 ***
adm          0.51985    0.05366   9.688  < 2e-16 ***
adsk         0.23863    0.05233   4.560 9.07e-06 ***
adt          0.24030    0.03435   6.996 4.22e-11 ***
aes         -1.20431    0.12980  -9.279  < 2e-16 ***
aet          0.22902    0.01952  11.733  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7724 on 193 degrees of freedom
Multiple R-squared:  0.9332,       Adjusted R-squared:  0.9314
F-statistic:   539 on 5 and 193 DF,  p-value: < 2.2e-16

> shapiro.test(model42$residuals)
Shapiro-Wilk normality test
data:  model42$residuals
W = 0.98867, p-value = 0.1117
```

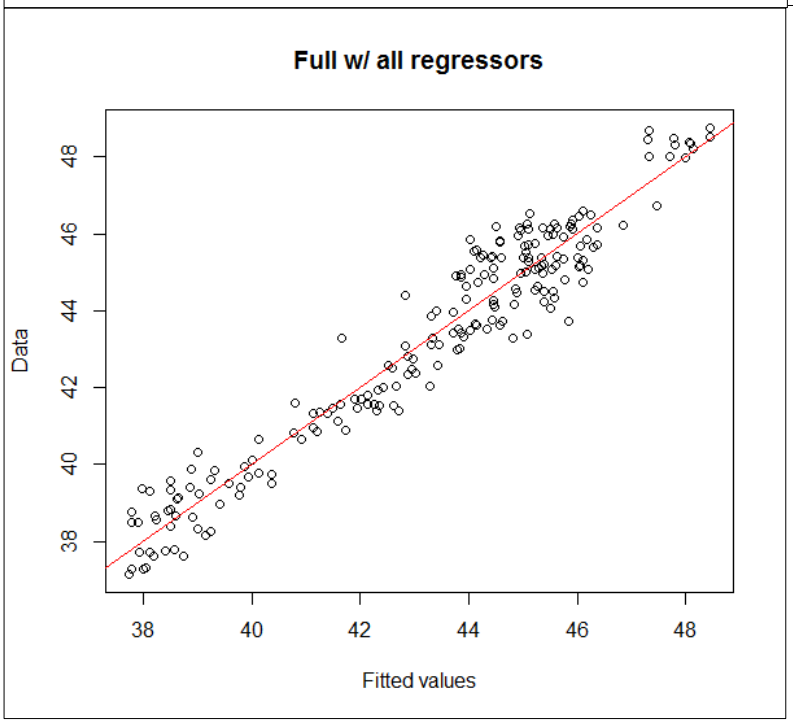
Графік абсолютних відхилень прогнозів моделей від справжніх значень даних adi

```
> plot(1:10, 1.3*abs(xadi - xadif1),type = "n", xlab = "Observations", yla
b = "",main = "True vs Fitted values")
> points(1:10, abs(xadi - xadiM2), pch = 19, col = "blue")
> points(1:10, abs(xadi - adi4), pch = 19, col = "red")
> points(1:10, abs(xadi - xadif1), pch = 19, col = "green")
> legend(6.1, 3.0, c("not full", "full w/o some regressors","full w/ all r
egressors"), pch = 19, col = c("blue", "red", "green"))
```









1 Модель за останніми даними
(без елементів з великим впливом):
R = 0.9028, p-value = 3.992e-15,
статистичну значущість мають лише 2 регресори

2 Модель за повними даними:
R = 0.9352, p-value = 2.2e-16,
статистичної значущості не мають 4 регресори

3 Модель за повними даними:
(без регресорів з низькою значущістю):
R = 0.9274, p-value = 2.2e-16,
статистично значущість мають всі регресори

В моделі за останніми даними наявно багато коефіцієнтів, що не мають високого рівня значущості. В моделі за повними даними з всіма регресорами також, але коефіцієнт детермінації вище, а p-value - менше; також, якщо викинути єдине спостереження, то оцінки коефіцієнтів МНК досить сильно зміняться, а залишки втратять нормальний розподіл.
В моделі за повними даними, в котрій не враховувались деякі регресори, усі коефіцієнти мають високий рівень значущості, хоча коефіцієнт детермінації трохи менший, ніж при моделі за повними даними. Загалом найкращий прогноз в термінах абсолютного відхилення прогнозу за допомогою моделі та справжнх значень досліджуваної змінної дає модель за останніми даними.