Прикладная статистика в R

Лекция 6. Диагностика модели линейной регрессии

Елена Михайловна Парилина

д. ф.-м. н., проф.

2021

Основные предположения регрессионного анализа

Сформулируем основные предположения регрессионного анализа, которые относятся к случайным компонентам ε_i, $i = 1, \ldots, n$.

Первая группа предположений:

- Случайные величины ε_i, $i = 1, \ldots, n$ образуют так называемый слабый белый шум, т. е. последовательность центрированных ($E\varepsilon_i = 0$, $i = 1, \ldots, n$) и некоррелированных ($E(\varepsilon_l \varepsilon_u) = 0$ при $l \neq u$) случайных величин с одинаковыми дисперсиями σ^2 ($E\varepsilon_i^2 = \sigma^2$, $i = 1, \ldots n$).

Кроме первой группы предположений будем также рассматривать вторую группу, в которой сформулируем предположение о совместном распределении случайных величин ε_i, $i = 1, \ldots, n$.

Вторая группа предположений:

- Совместное распределение случайных величин ε_i, $i = 1, \ldots, n$ является нормальным распределением с нулевым вектором математических ожиданий и ковариационной матрицей $\sigma^2 E_n$, т. е. случайный вектор $\varepsilon^T = (\varepsilon_1, \ldots, \varepsilon_n) \sim N(0, \sigma^2 E_n)$, где E_n — единичная матрица порядка $n \times n$.

Массив данных "Denver Neighborhoods"

Массив содержит наблюдения вида

$$(X1, X2, X3, X4, X5, X6, X7),$$

где каждое наблюдение содержит характеристики района в Денвере:

- X1 = население (в тыс. человек),
- X2 = % изменения населения за последние несколько лет,
- X3 = % детей (младше 18) среди населения,
- X4 = % государственных школ, участвующих в программе бесплатных обедов,
- X5 = % изменения в заработной платы населения за последние несколько лет,
- X6 = криминальный показатель (на 1000 населения),
- X7 = % изменения в криминальном показателе за последние несколько лет.

Предполагается, что уровень криминала (X6) может зависеть от вышеперечисленных показателей. Задача: построить "наилучшую" модель линейной регрессии.

```
> m1<-lm(X6~X1+X2+X3+X4+X5+X7, neig)
> summary(m1)

Call:
lm(formula = X6 ~ X1 + X2 + X3 + X4 + X5 + X7, data = neig)

Residuals:
     Min       1Q   Median       3Q      Max
-117.010  -37.285   -2.292   28.046  241.633

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  67.5618    58.8035   1.149   0.2580
X1           -5.2679     3.1061  -1.696   0.0983 .
X2            1.2279     0.9558   1.285   0.2069
X3           -7.0990     1.6171  -4.390 9.11e-05 ***
X4            2.8437     0.5304   5.361 4.60e-06 ***
X5            4.5877     1.6586   2.766   0.0088 **
X7            0.5862     0.7550   0.776   0.4424
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 66.18 on 37 degrees of freedom
Multiple R-squared:  0.658,     Adjusted R-squared:  0.6025
F-statistic: 11.86 on 6 and 37 DF,  p-value: 2.187e-07
```

Графики в анализе модели линейной регрессии

- Plot: попарные графики рассеяния.
- "Residuals vs Fitted": Остатки должны быть "случайным образом разбросаны" относительно прямой $y = 0$. В нашем случае заметны выбросы, также наблюдается параболическая зависимость, а не линейная (красный график). Вы можете проверить предположение о равной дисперсии (гомоскедастичность). Хорошо, если вы увидите горизонтальную линию с одинаковым (случайным) разбросом точек вокруг нее.
- "Normal Q-Q" plot, или quantile-quantile plot: это графический инструмент, помогающий проверить, согласуется ли выборка остатков с теоретическим нормальным распределением. (ВНИМАНИЕ: это просто визуальная проверка, а не проверка гипотезы критерием, поэтому это субъективное мнение). График Q-Q — это диаграмма рассеяния, созданная путем сопоставления двух наборов квантилей друг другу (теоретических и эмпирических). Если оба набора квантилей принадлежат одному и тому же распределению, мы должны увидеть точки, образующие примерно прямую линию. В нашем случае точки наблюдения 32, 19, 7 не укладываются в нормальное распределение.

Графики в анализе модели линейной регрессии

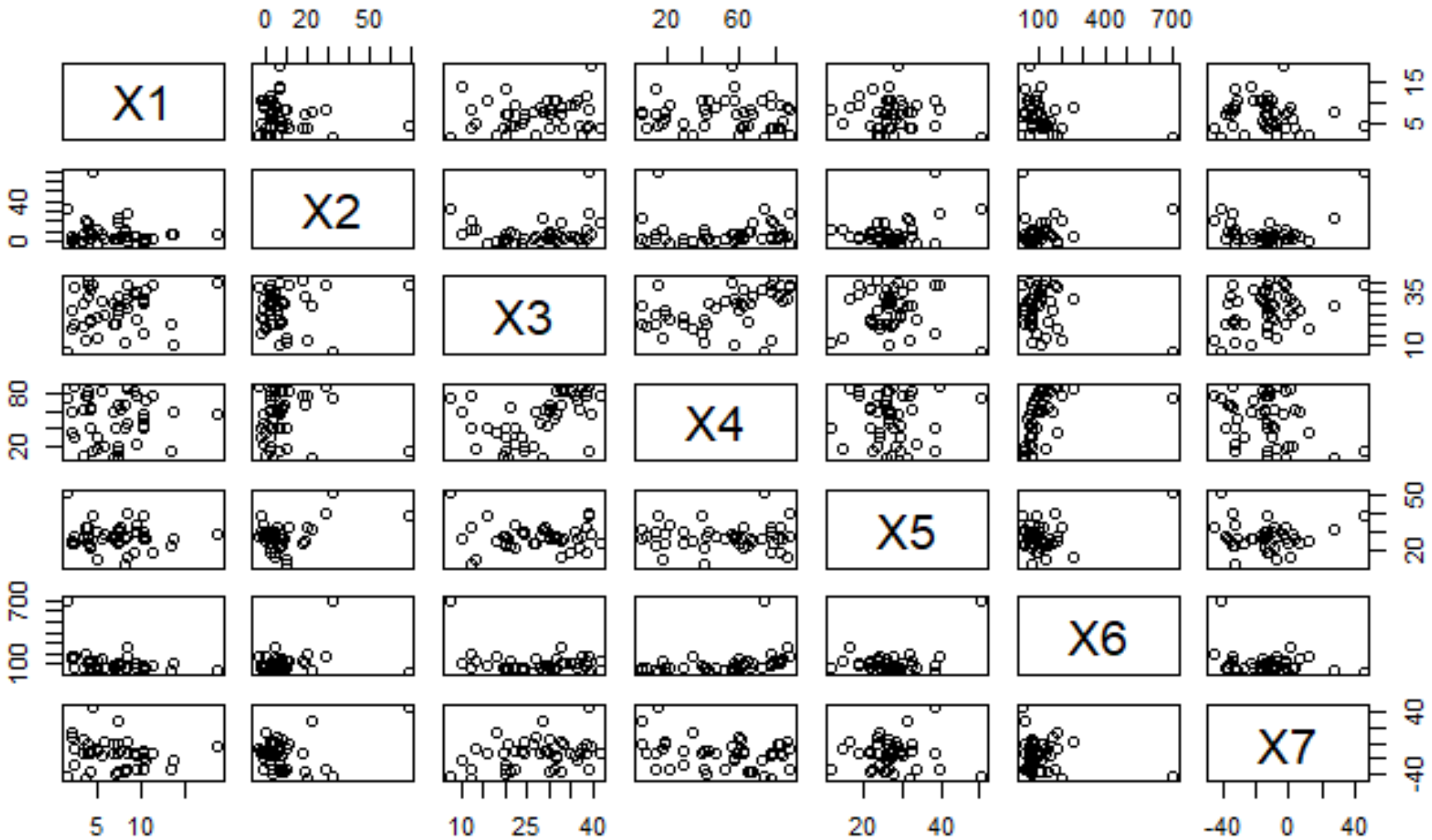
- "Residuals vs Leverage": Этот график помогает выявить влиятельные наблюдения, если таковые имеются. Не все выбросы существенно влияют на построенную модель линейной регрессии (что бы ни значили выбросы). Несмотря на то, что данные имеют экстремальные значения, они могут не оказывать существенного влияния на регрессию. Это означает, что результаты не будут сильно отличаться, если мы включим или исключим их из анализа. С другой стороны, некоторые наблюдения могут быть очень важными, даже если они находятся в разумном диапазоне значений. Они могут быть влиятельными при построении линии регрессии и могут изменить результаты, если мы исключим их из анализа.

Графики в анализе модели линейной регрессии

- "Residuals vs Leverage". Мы наблюдаем за отстающими значениями в правом верхнем или правом нижнем углу. Ищем случаи за пределами пунктирной линии расстояния Кука, т.е. когда наблюдение находится за пределами расстояния Кука (что означает, что они имеют большие значения расстояния по шкале Кука). Если мы исключим эти случаи, уравнение регрессии может существенно поменяться.

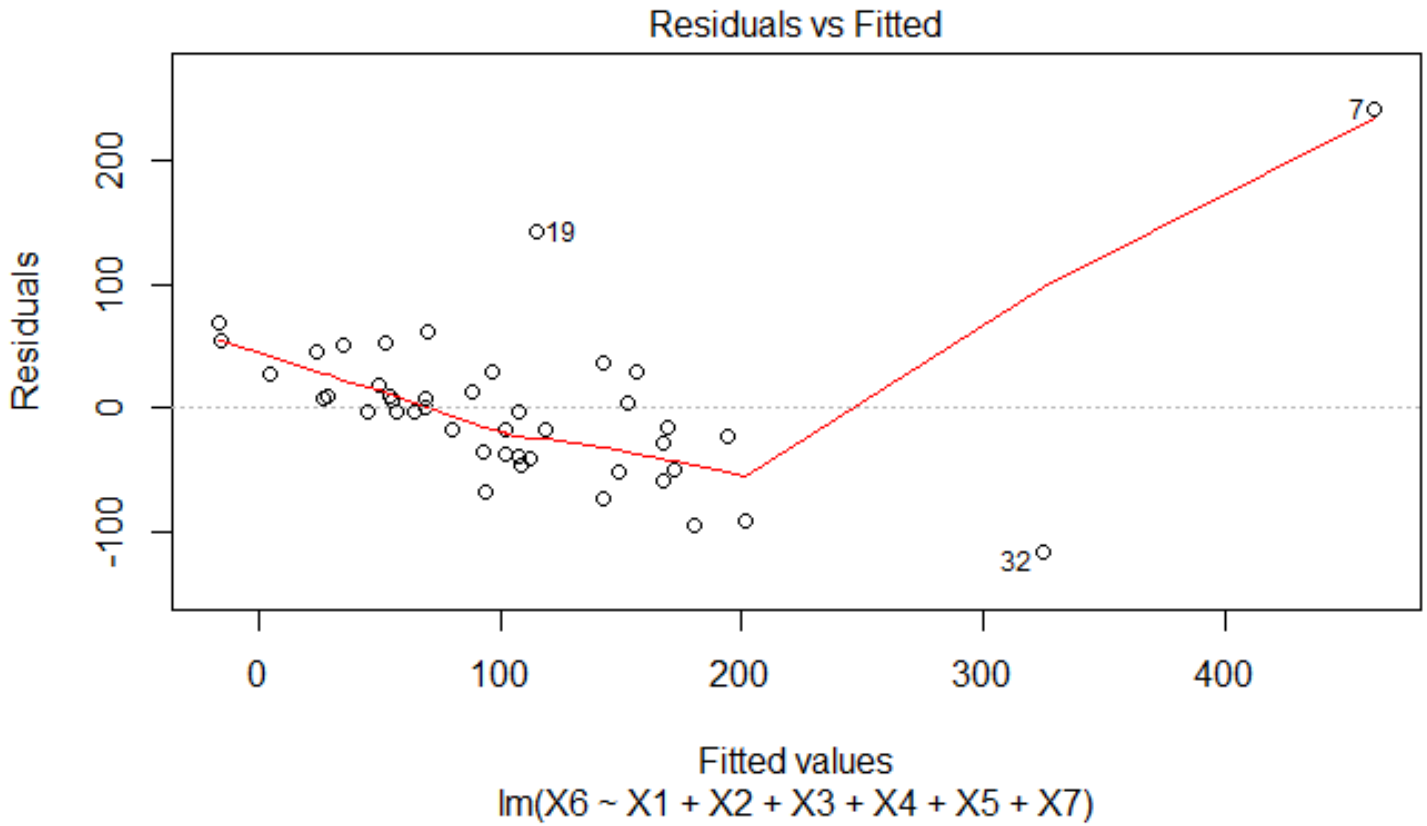
Основная модель (график: scatterplot)

`pairs(neig)`:



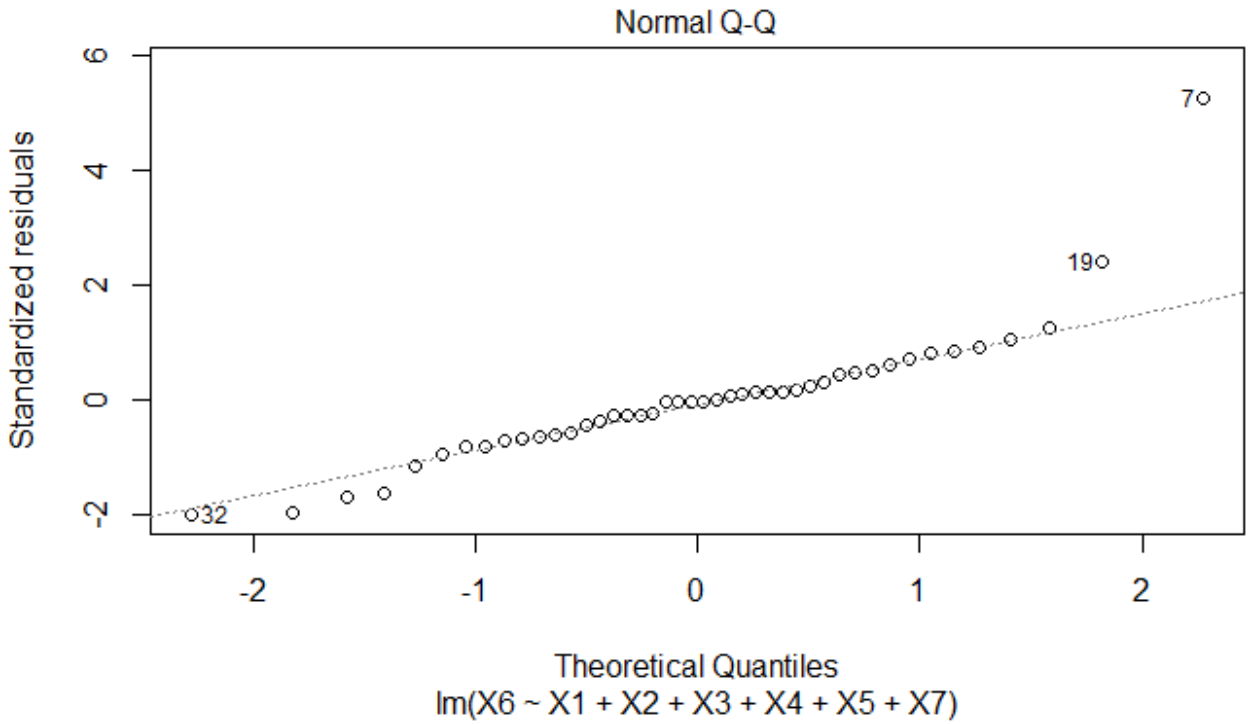
Основная модель (график: Residuals vs Fitted)

Функция `plot(m1)` выдает следующие графики:



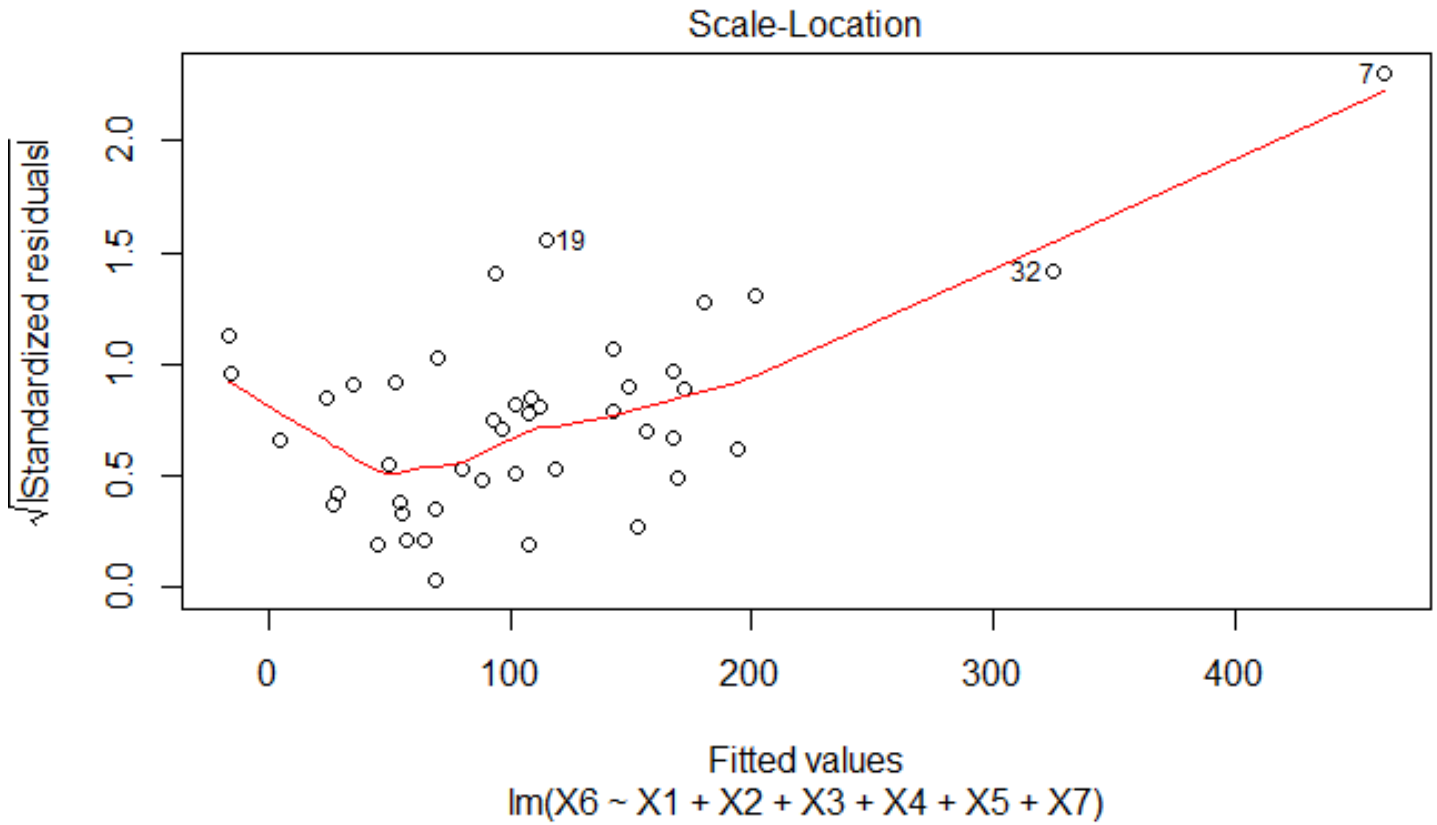
Анализируем распределение остатков по этому графику (см. предположения регрессионного анализа).

Основная модель (график: Normal Q-Q)



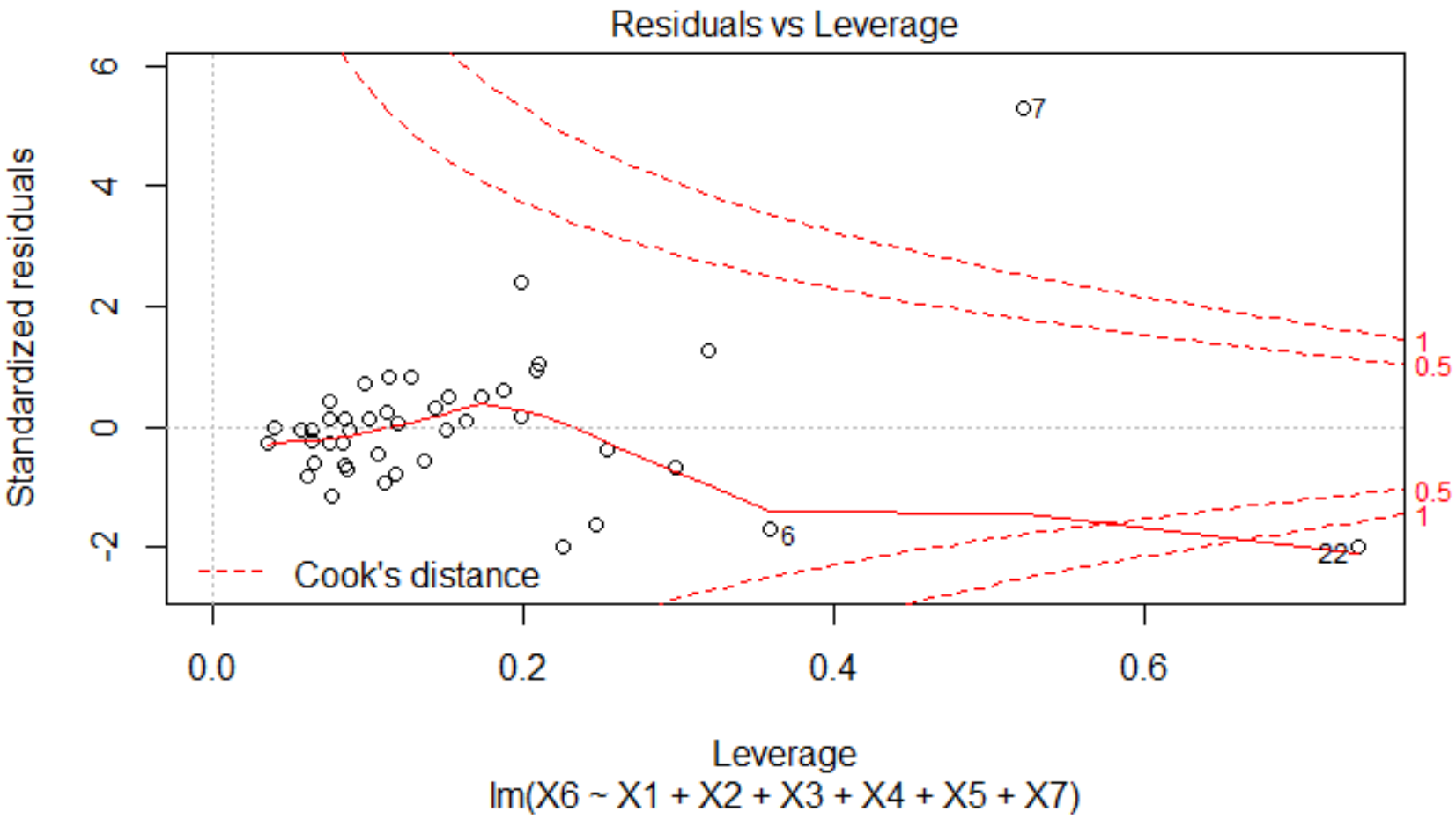
Нормальное ли распределение у остатков (см. предположения регрессионного анализа)?

Основная модель (график: Scale-Location)



Является ли распределение остатков нормальным с равным дисперсиями (см. предположения регрессионного анализа)?

Основная модель (график: Residuals vs Leverage)



Выявляем наиболее «влиятельные» наблюдения.

Расстояние Кука

Пусть

$$H = x_i^T (X^T X)^{-1} x_i,$$

— матрица проекции, а i-ый элемент главной диагонали есть

$$h_{ii} \equiv x_i^{\mathsf{T}} (X^{\mathsf{T}} X)^{-1} x_i$$

также имеет название leverage i-го наблюдения.
Остатки или ошибки есть вектор

$$\varepsilon = y - \widehat{y} = (I - H)\, y,$$

где ε_i — i-ый остаток.
Расстояние Кука D_i наблюдения i — это сумма квадратов изменений в предсказаниях моделей при удалении i-го наблюдения:

$$D_i = \frac{\sum_{j=1}^{n} \left(\widehat{y}_j - \widehat{y}_{j(i)} \right)^2}{ks^2},$$

где $\widehat{y}_{j(i)}$ — предсказанное значение y_j, когда в модели удалено наблюдение i.

Расстояние Кука

Значение

$$s^2 = \frac{\varepsilon^\top \varepsilon}{n-k}$$

есть усредненная ошибка.
Расстояние Кука может быть вычислено так:

$$D_i = \frac{\varepsilon_i^2}{ks^2}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right].$$

С помощью Bonferroni Outlier Test мы можем протестировать наблюдения на линейный сдвиг (нулевая гипотеза о том, что сдвига нет):

```
> outlierTest(m1,n.max=5)
  rstudent unadjusted p-value Bonferonni p
7 10.48506         1.7263e-12   7.5956e-11
```

Для наблюдения 7 отклоняется гипотеза о нулевом сдвиге.

Гомоскедастичность

Гомоскедастичность означает, что все остатки ε_i имеют одинаковую дисперсию.
H_0: дисперсии остатков одинаковы (гомоскедастичность).
H_1: дисперсии остатков различны (гетероскедастичность).

Breusch-Pagan test:

Функция ncvTest и bptest в R

```
library(car)
ncvTest(m1)
```

или

```
library(lmtest)
bptest(m1)
```

Гомоскедастичность

Результаты тестов показывают наличие проблемы гетероскедастичности.

```
> library(car)
> ncvTest(m1)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 68.60054, Df = 1, p = < 2.22e-16
```

```
> bptest(m1)

        studentized Breusch-Pagan test

data:  m1
BP = 25.696, df = 6, p-value = 0.0002536
```

Отсутствие vs. наличие автокорреляции

Автокорреляция означает, что остатки удовлетворяют уравнению:

$$\varepsilon_i = \rho\varepsilon_{i-1} + \nu_i,$$

где $\{\nu_i\}$ — независимые нормально распределенные величины, $(0, \sigma_\nu^2)$, и $|\rho| < 1$ — коэффициент автокорреляции.
H_0: $\rho = 0$ (отсутствие автокорреляции),
H_1: $\rho \neq 0$ (наличие автокорреляции).

Функции в R

```
library(car)
durbinWatsonTest(m1)
```

или

```
library(lmtest)
dwtest(m1)
```

Тест на автокорреляцию (Дарбина-Уотсона)

Гипотеза об отсутствии автокорреляции (нулевая гипотеза) принимается на основе выводов теста Дарбина-Уотсона:

```
> library(car)
> durbinWatsonTest(m1)
 lag Autocorrelation D-W Statistic p-value
   1      -0.1980801      2.378141   0.246
 Alternative hypothesis: rho != 0
> library(lmtest)
> dwtest(m1)

        Durbin-Watson test

data:  m1
DW = 2.3781, p-value = 0.8741
alternative hypothesis: true autocorrelation is greater than 0

> dwtest(m1,alternative = "two.sided")

        Durbin-Watson test

data:  m1
DW = 2.3781, p-value = 0.2517
alternative hypothesis: true autocorrelation is not 0
```

Нормальное распределение остатков

Функции в R

```
library(olsrr)
ols_test_normality(m1)
```

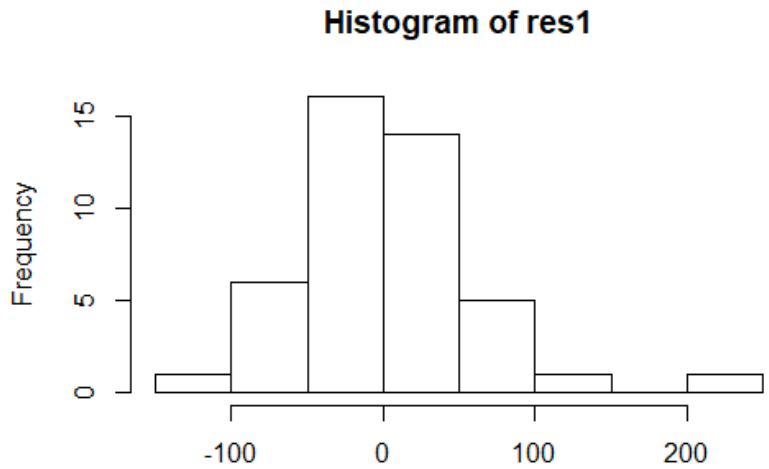
или

```
res1=residuals(m1,type="response") # Определяем вектор остатков,
shapiro.test(res1) # используем тест Шапиро-Уилка на нормальность.
```

```
> ols_test_normality(m1)
-----------------------------------------------
       Test             Statistic       pvalue
-----------------------------------------------
Shapiro-Wilk              0.9017         0.0012
Kolmogorov-Smirnov        0.1143         0.5744
Cramer-von Mises          3.6862         0.0000
Anderson-Darling          0.889          0.0211
-----------------------------------------------
> res1=residuals(m1,type="response")
> shapiro.test(res1)

        Shapiro-Wilk normality test

data:  res1
W = 0.90167, p-value = 0.001225
```

Нормальное распределение остатков

```
> library(nortest)
Warning message:
пакет 'nortest' был собран под R версии 3.5.2
> lillie.test(res1)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  res1
D = 0.11426, p-value = 0.159

> pearson.test(res1)

        Pearson chi-square normality test

data:  res1
P = 6, p-value = 0.5397
```



Проблема мультиколлинеарности

Мультиколлинеарность означает линейную зависимость одной независимой переменной от линейной комбинации других независимых переменных. Нам нужно рассчитать коэффициенты *VIF* для всех независимых переменных:

$$VIF(\beta_k) = \frac{1}{1 - R_k^2},$$

где R_k^2 — коэффициент детерминации (multiple R-squared), вычисленный для модели линейной регрессии, когда X_k является зависимой переменной, а все другие предикторы являются независимыми переменными. Считается, что проблема мультиколлинеарности не возникает, если все $VIF(\beta_k)$ меньше 10. Иногда используется граница 5.

Функции в R

```
library(olsrr)
ols_vif_tol(m1)
```

или

```
library(car)
vif(m1)
```

Проблема мультиколлинеарности

Проблема мультиколлинеарности не наблюдается в модели $m1$, поскольку все коэффициенты VIF меньше 10 (и даже 5).

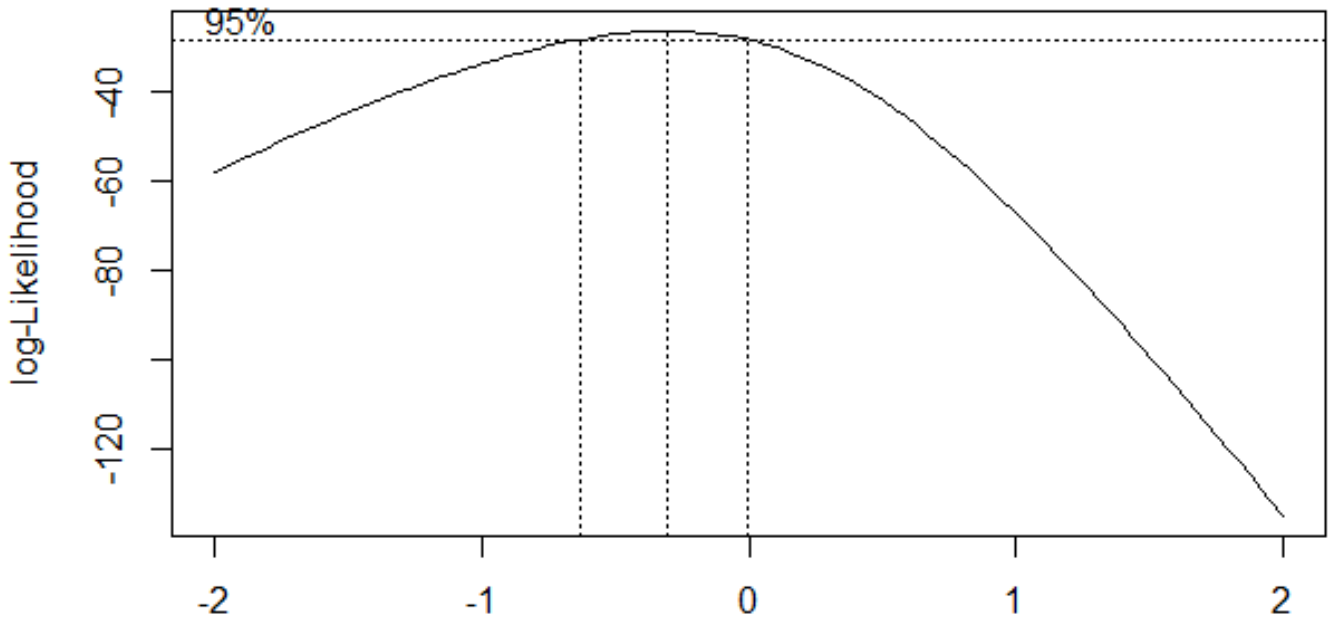
```
> library(olsrr)
> ols_vif_tol(m1)
  Variables Tolerance      VIF
1        X1 0.8354701 1.196931
2        X2 0.7483817 1.336217
3        X3 0.4961859 2.015374
4        X4 0.5402921 1.850851
5        X5 0.8068418 1.239400
6        X7 0.5525069 1.809932
> library(car)
> vif(m1)
      X1       X2       X3       X4       X5       X7
1.196931 1.336217 2.015374 1.850851 1.239400 1.809932
```

Преобразование зависимой переменной

Возможно, степенное преобразование зависимой переменной y приведет к улучшению модели. Воспользуемся процедурой Box–Cox:

Функции в R

```
library(MASS)
bc <- boxcox(m1)
```



Преобразование зависимой переменной

Функция boxcox отображает значения λ в зависимости от логарифмической функции правдоподобия полученной модели, как показано на рисунке. Мы хотим максимизировать эту логарифмическую функцию правдоподобия. На рисунке изображено наилучшее значение, а также границы доверительного интервала. В этом случае, похоже, что наилучшее значение составляет около -0.3 с доверительным интервалом около (–0.6, 0.0).

Найдем максимальное значение λ, которое дают наибольшее значение логарифмической функции правдоподобия. Мы используем функцию `which.max`:

Функции в R

```
which.max(bc$y)
lambda <- bc$x[which.max(bc$y)]
lambda
```

Преобразование зависимой переменной

```
> library(MASS)
> boxcox(m1)
> bc <- boxcox(m1)
> which.max(bc$y)
[1] 43
> lambda <- bc$x[which.max(bc$y)]
> lambda
[1] -0.3030303
```

Делаем преобразование переменной $X6$, возводя ее в степень λ, после этого строим новую модель регрессии и проводим ее анализ:

Функции в R

```
 z <- neig$X6 ^ lambda
neig2<-neig[,-6]
neig2$X6 <- z
m2 <- lm(X6~., data=neig2)
summary(m2)
```

Анализ новой модели

```
> summary(m2)

Call:
lm(formula = X6 ~ ., data = neig2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.050647 -0.019910  0.001391  0.014067  0.040362

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.2522098  0.0209897  12.016 2.45e-14 ***
X1           0.0031793  0.0011087   2.868  0.00679 **
X2           0.0003006  0.0003412   0.881  0.38405
X3           0.0028826  0.0005772   4.994 1.44e-05 ***
X4          -0.0017128  0.0001893  -9.047 6.54e-11 ***
X5          -0.0003797  0.0005920  -0.641  0.52522
X7          -0.0003993  0.0002695  -1.482  0.14689
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02362 on 37 degrees of freedom
Multiple R-squared:  0.7744,    Adjusted R-squared:  0.7378
F-statistic: 21.17 on 6 and 37 DF,  p-value: 1.345e-10
```

План диагностики регрессионной модели в R (часть 2)

1. Построить графики: scatterplot, "Residuals vs Fitted" , "Normal Q-Q" , "Residuals vs Leverage" с помощью функции `plot(m1)` и дать интерпретации.
2. Проверить модель на наличие выбросов с помощью функции `outlierTest(m1)`.
3. Проверить модель на гетероскедастичность с помощью функции `ncvTest(m1)` или `bptest(m1)`.
4. Проверить остатки модели на автокорреляцию с помощью функции `durbinWatsonTest(m1)` или `dwtest(m1)`.
5. Проверить остатки модели на нормальность распределения с помощью функции `ols_test_normality(m1)`.
6. Проверить модель на мультиколлинеарность данных с помощью функции `vif(m1)`.
7. Попробовать применить трансформацию Box-Cox зависимой переменной.
8. В случае получения новой модели в предыдущем пункте проанализировать новую модель.

Литература

1. **Brooks, Chris**: Introductory econometrics for finance. Cambridge, 2002 or newer.
2. **Tsay, Ruey S.**: Analysis of Financial Time Series. Wiley, 2002 or newer.
3. **Levine, David M., Stephan, David F., Szabat, Kathryn A**: Statistics for Managers Using Microsoft Excel. Pearson; 8 edition, 2016 or newer.

Итоги

Что мы узнали на Лекции 6?

- Какие предположения делаются при построении уравнения линейной регрессии.
- Как проверить, выполняется ли каждое предположение с использованием R.
- Как составить отчет с целью диагностики модели линейной регрессии.

Что мы узнаем на Лекции 7?

- На лекции 7 мы разберем пример построения наилучшей регрессионной модели для одного датасета, составим отчет по результатам работы в R.

Спасибо за внимание и до встречи на Лекции 7!