Линейный регрессионный анализ

Проверка предположений и анализ результатов

Теорема Гаусса-Маркова

Рассматривается модель парной регрессии, в которой наблюдения Y связаны с X следующей зависимостью:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i.$$

На основе n наблюдений оценивается уравнение регрессии

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i.$$

Тогда, если данные обладают следующими свойствами:

1. Модель данных правильно специфицирована;
2. Все X_i детерминированы и не все равны между собой;
3. Ошибки не носят систематического характера, то есть $E(\varepsilon_i) = 0 \ \forall i$;
4. Дисперсия ошибок одинакова и равна некоторой σ^2;
5. Ошибки некоррелированы, то есть $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0 \ \forall i, j$;

— то оценки полученные методом наименьших квадратов оптимальны в классе линейных несмещённых оценок.

Пример: вес новорожденного

Файл `birthweight_reduced.csv` содержит данные о 42 новорожденных и их родителях. Нас интересует как зависит вес новорожденного (Birthweight, указан в фунтах, 1 брит. фунт = 453.59 грамма) от продолжительности беременности (Gestation, в неделях), роста и веса матери до беременности (mheight и mppwt в дюймах и фунтах соотвественно, 1 дюйм = 2.54 см) и от того, курит ли мать (smoker: 0 — не курит, 1 — курит)

Источник: http://www.statstutor.ac.uk/students/topics/r/statistical-analyses-using-r/

Чтение и очистка от пропусков

```
babyData = read.csv('week_09/data/birthweight_reduced.csv')
head(babyData, n=3)
```

```
##     id headcirumference length Birthweight Gestation smoker motherage
## 1 1313                12     17         5.8        33      0        24
## 2  431                12     19         4.2        33      1        20
## 3  808                13     19         6.4        34      0        26
##   mnocig mheight mppwt fage fedyrs fnocig fheight lowbwt mage35
## 1      0      58    99   26     16      0      66      1      0
## 2      7      63   109   20     10     35      71      1      0
## 3      0      65   140   25     12     25      69      0      0
##   LowBirthWeight
## 1            Low
## 2            Low
## 3         Normal
```

```
dim(babyData)
```

```
## [1] 42 17
```

```
babyData = na.omit(babyData); dim(babyData)
```

```
## [1] 42 17
```

smoker — это фактор

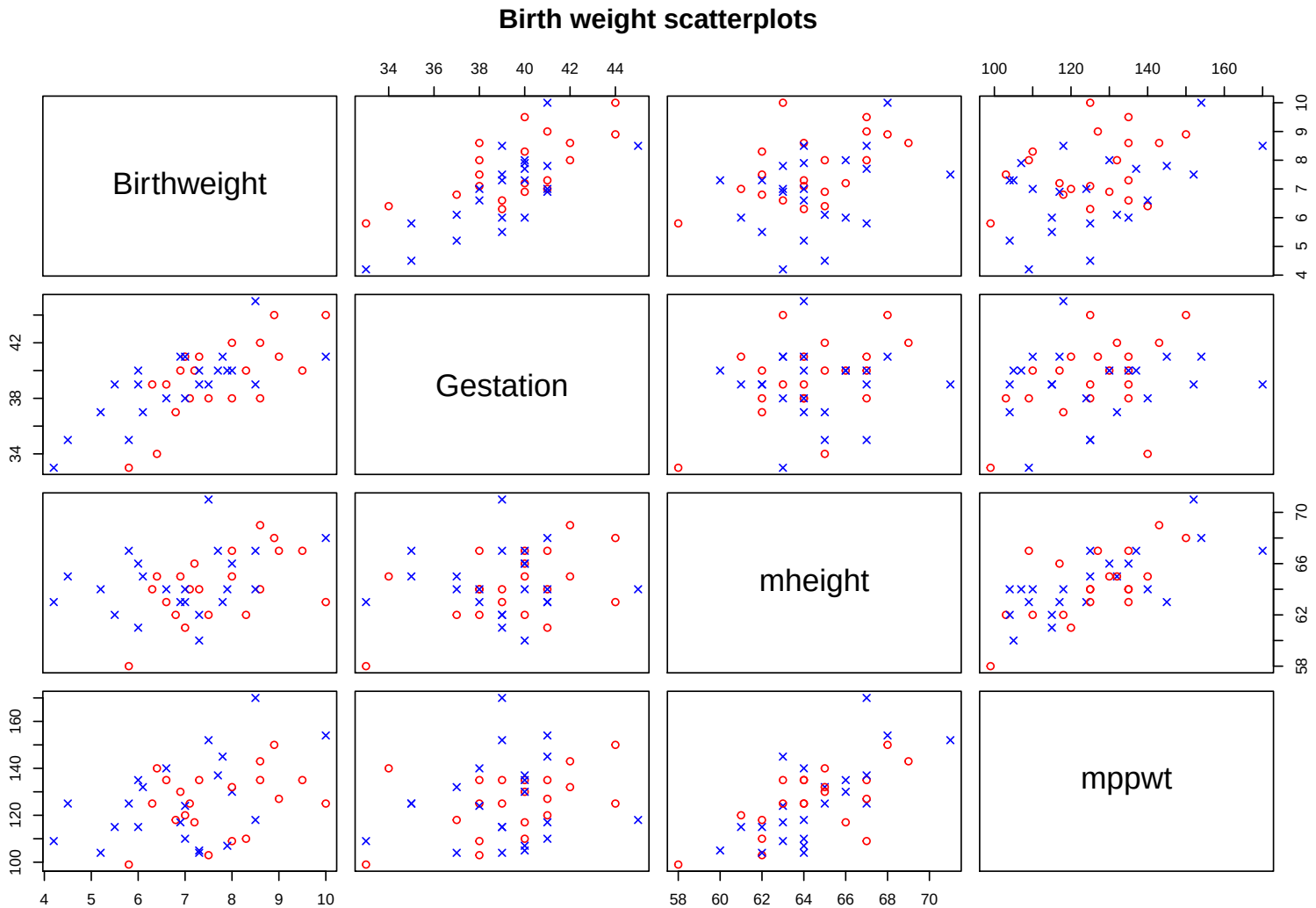
```
class(babyData$smoker)
```

```
## [1] "integer"
```

```
babyData$smoker<-factor(babyData$smoker,
                        labels=c('Non-smoker','Smoker'))
```

Проверка на мультиколлинеарность 1

```
attach(babyData)
pairs(~Birthweight+Gestation+mheight+mppwt,
      main='Birth weight scatterplots',
      col=c('red','blue')[smoker], pch=c(1,4)[smoker])
```



Проверка на мультиколлинеарность 2

Матрица корреляции

```
round(cor(cbind(Birthweight,Gestation,mppwt,mheight)),2)
```

```
##             Birthweight Gestation mppwt mheight
## Birthweight        1.00      0.71  0.39    0.37
## Gestation          0.71      1.00  0.25    0.23
## mppwt              0.39      0.25  1.00    0.67
## mheight            0.37      0.23  0.67    1.00
```

Проверка на мультиколлинеарность 3: tolerance

R_j^2 — коэффициент детерминации для регрессии j-го предиктора на остальные предикторы.

$$tolerance = 1 - R_j^2$$

Пороговые значения:

- *0.2 — допустимо*
- 0.1 — опасно!

Проверка на мультиколлинеарность 3: VIF

VIF (Variance Inflation Factors) = 1/tolerance

- от 1 до 5 — хорошо
- от 5 до 10 — внимание, опасно!
- 10+ — данный предиктор существенно зависит от других предикторов

Вычисление VIF в R

```
library(car)
vif(lm(Birthweight~Gestation+smoker+mheight+mppwt))
```

```
## Gestation    smoker   mheight     mppwt
##  1.087246  1.013175  1.838976  1.852606
```

```
vif(lm(Birthweight~Gestation+smoker+mppwt))
```

```
## Gestation    smoker     mppwt
##  1.077986  1.010494  1.068478
```

```
vif(itog1) # Для недвижимости в Альбукерке
```

```
    SQFT      AGE    FEATS       NE     CUST      COR      TAX
6.197432 1.692336 1.459148 1.374965 1.385906 1.108327 6.476598
```

Считаем VIF'ы в пакете usdm

```
library(usdm)
independents<-data.frame(cbind(Gestation,smoker,mppwt))
vif(independents)
```

```
  Variables      VIF
1 Gestation 1.077986
2    smoker 1.010494
3     mppwt 1.068478
```

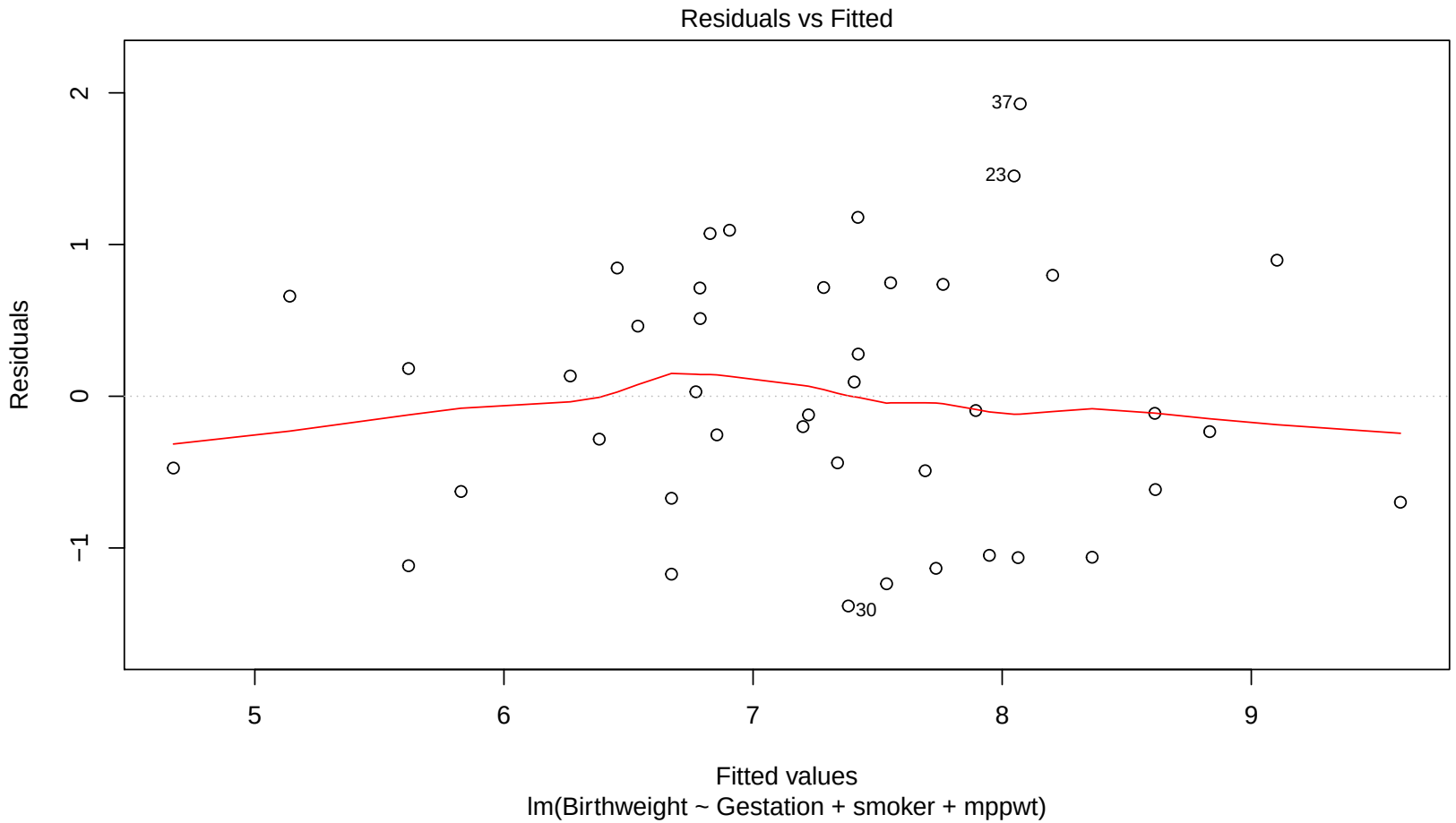
Строим регрессионную модель на основе независимых предикторов

```
reg1<-lm(Birthweight~Gestation+smoker+mppwt)
summary(reg1)
```

```
##
## Call:
## lm(formula = Birthweight ~ Gestation + smoker + mppwt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3824 -0.6246 -0.1033  0.7158  1.9276
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.164740   2.107344  -3.400   0.0016 **
## Gestation     0.313421   0.052887   5.926 7.19e-07 ***
## smokerSmoker -0.665279   0.267762  -2.485   0.0175 *
## mppwt         0.019819   0.008764   2.261   0.0295 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8622 on 38 degrees of freedom
## Multiple R-squared:  0.6104, Adjusted R-squared:  0.5796
## F-statistic: 19.84 on 3 and 38 DF,  p-value: 6.635e-08
```

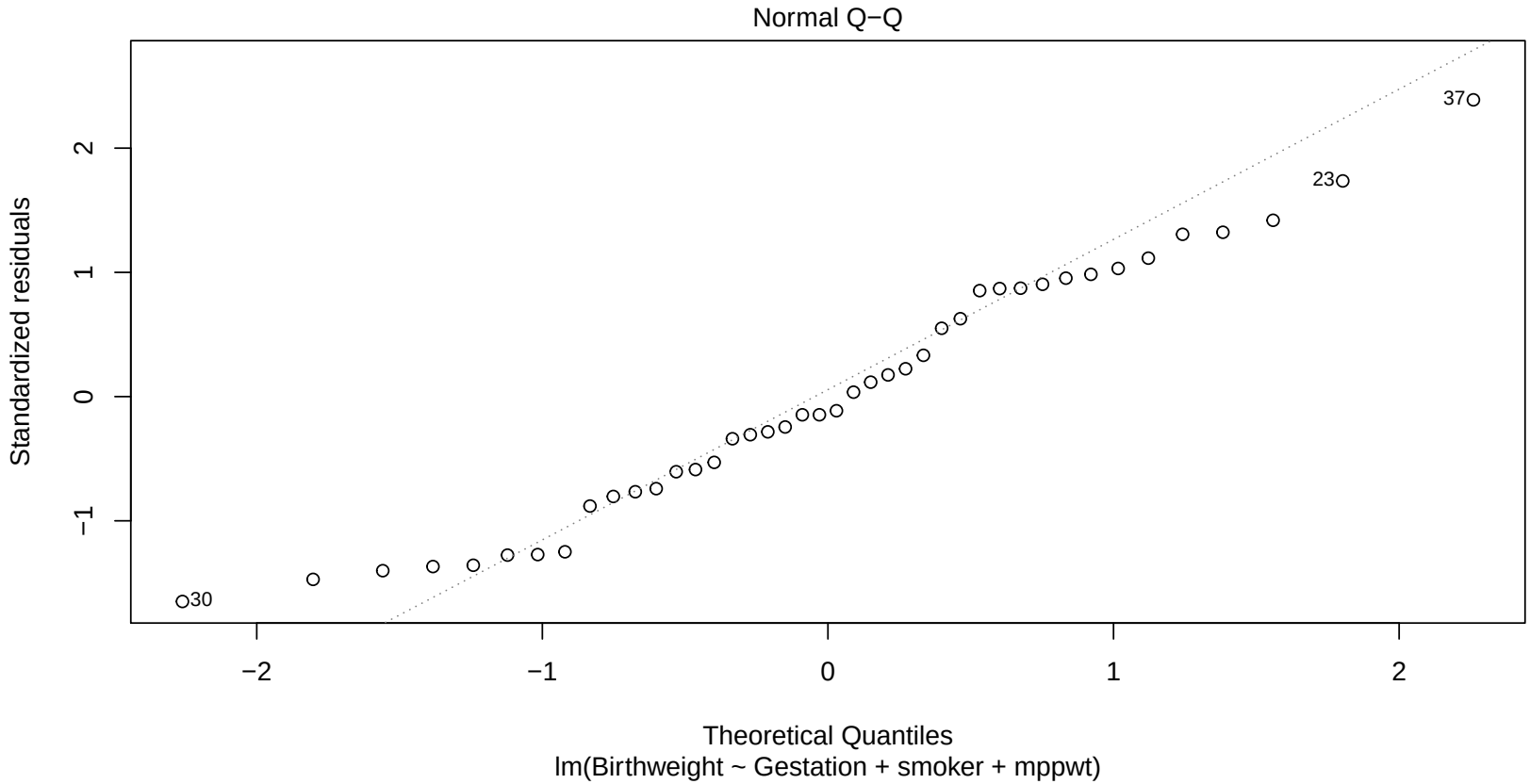
Проверка предположений: нормальность, гомоскедастичность и несмещенность оценок

```
plot(reg1, which = 1)
```



Проверка предположений: нормальность

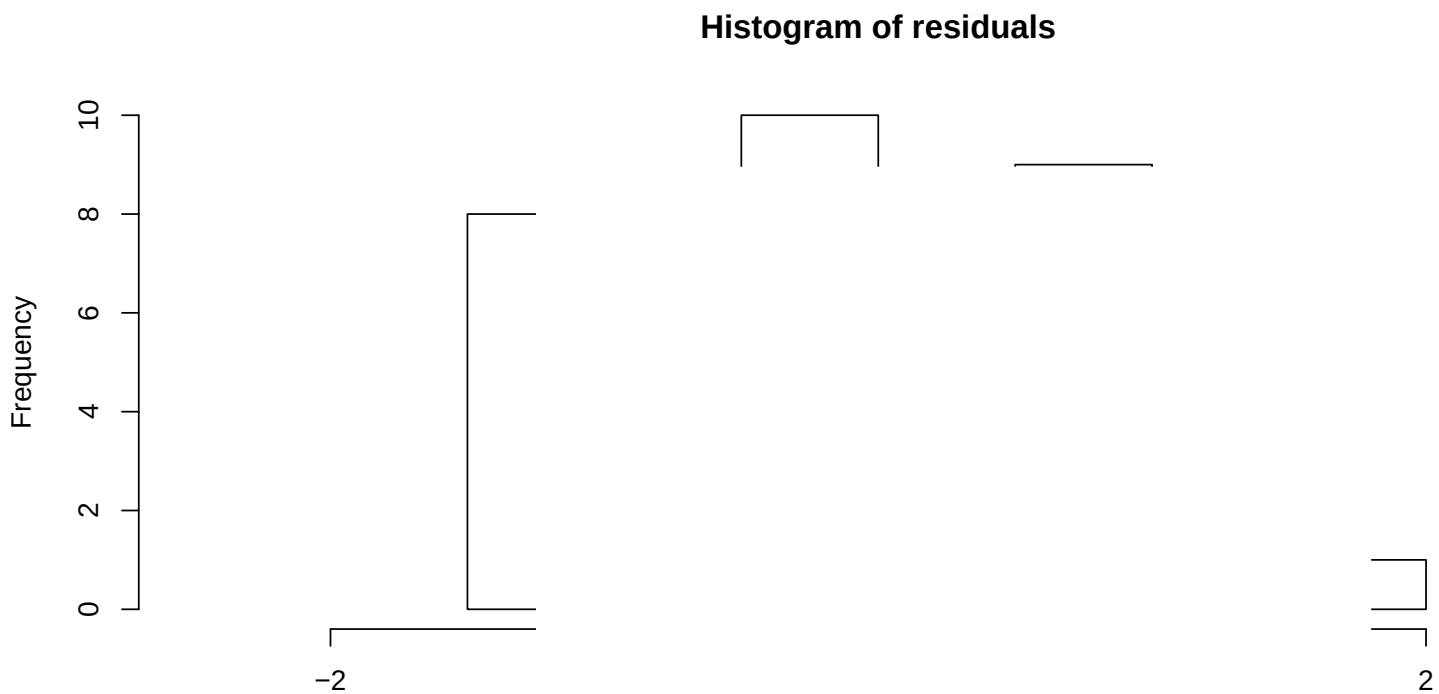
```
plot(reg1, which = 2)
```



О нормальных графиках КК (QQ plot): https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot

Гистограмма для проверки нормальности

```
hist(resid(reg1), xlim = range(c(-2.5,2.5)),
     main='Histogram of residuals',
     xlab='Standardised residuals',ylab='Frequency')
```



Мы ожидаем увидеть 95% стандартизованных остатков в диапазоне ± 1.96

Экстремальные значения

Если более 5% остатков выходит за пределы диапазона ± 1.96 или присутствуют экстремальные значения, выходящие из ± 3, то выполните регрессию с экстремальными значениями и без них, чтобы увидеть насколько сильно присутствие таких наблюдений изменяет коэффициенты модели.

Как выглядят проблемы

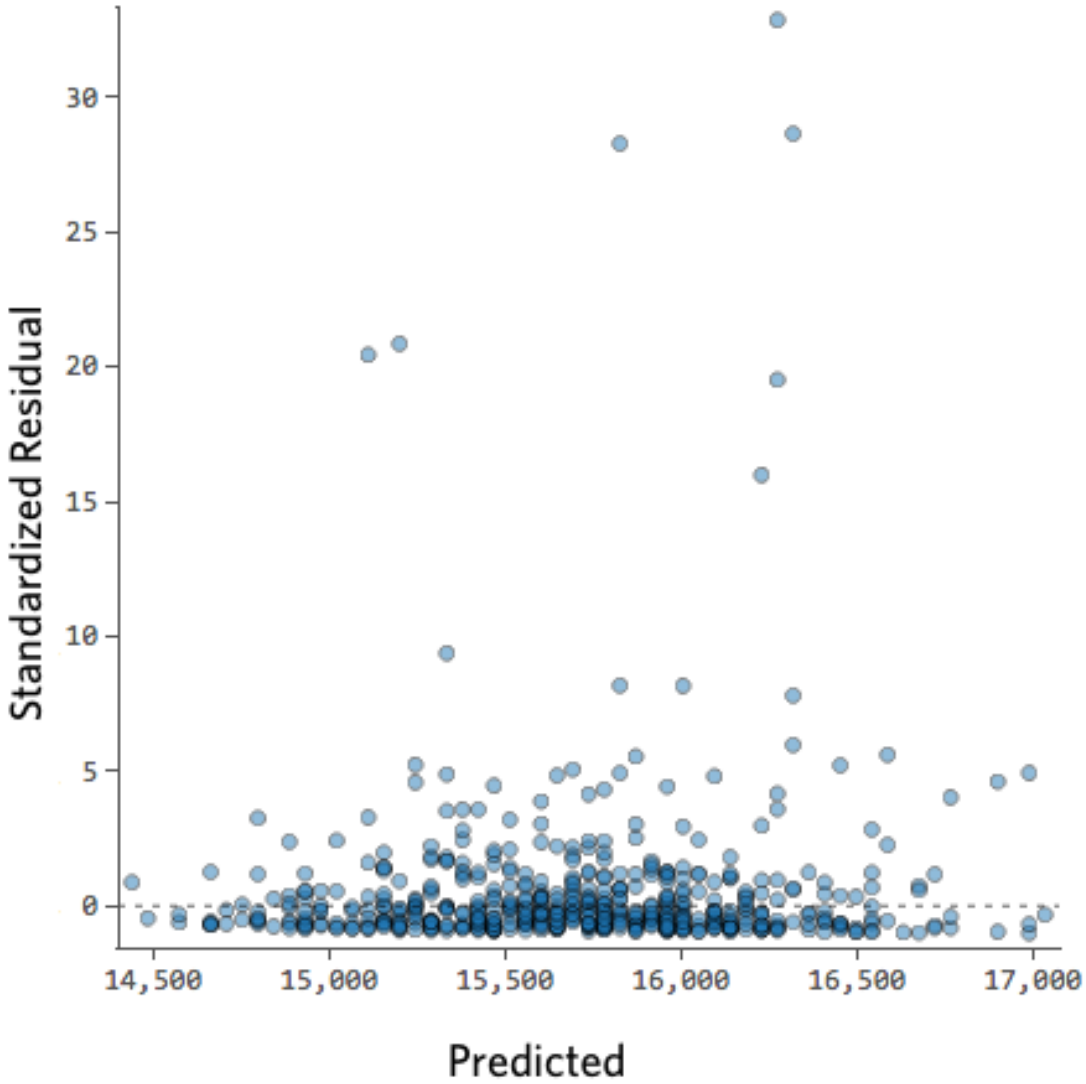


Figure 1:

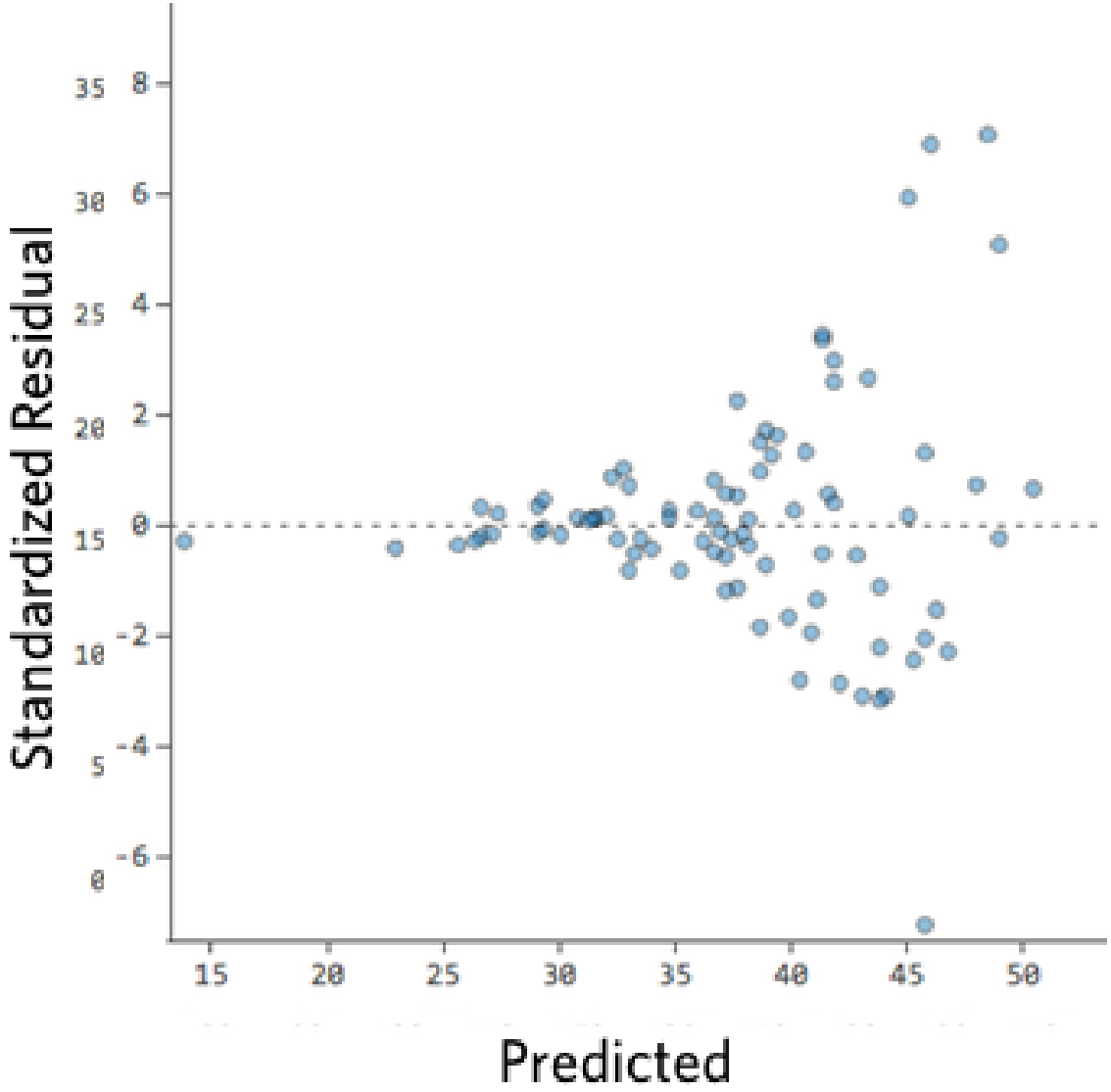

Standardized Residual
8
6
4
2
0
-2
-4
-6
15
20
25
30
35
40
45
50
Predicted

Figure 2:

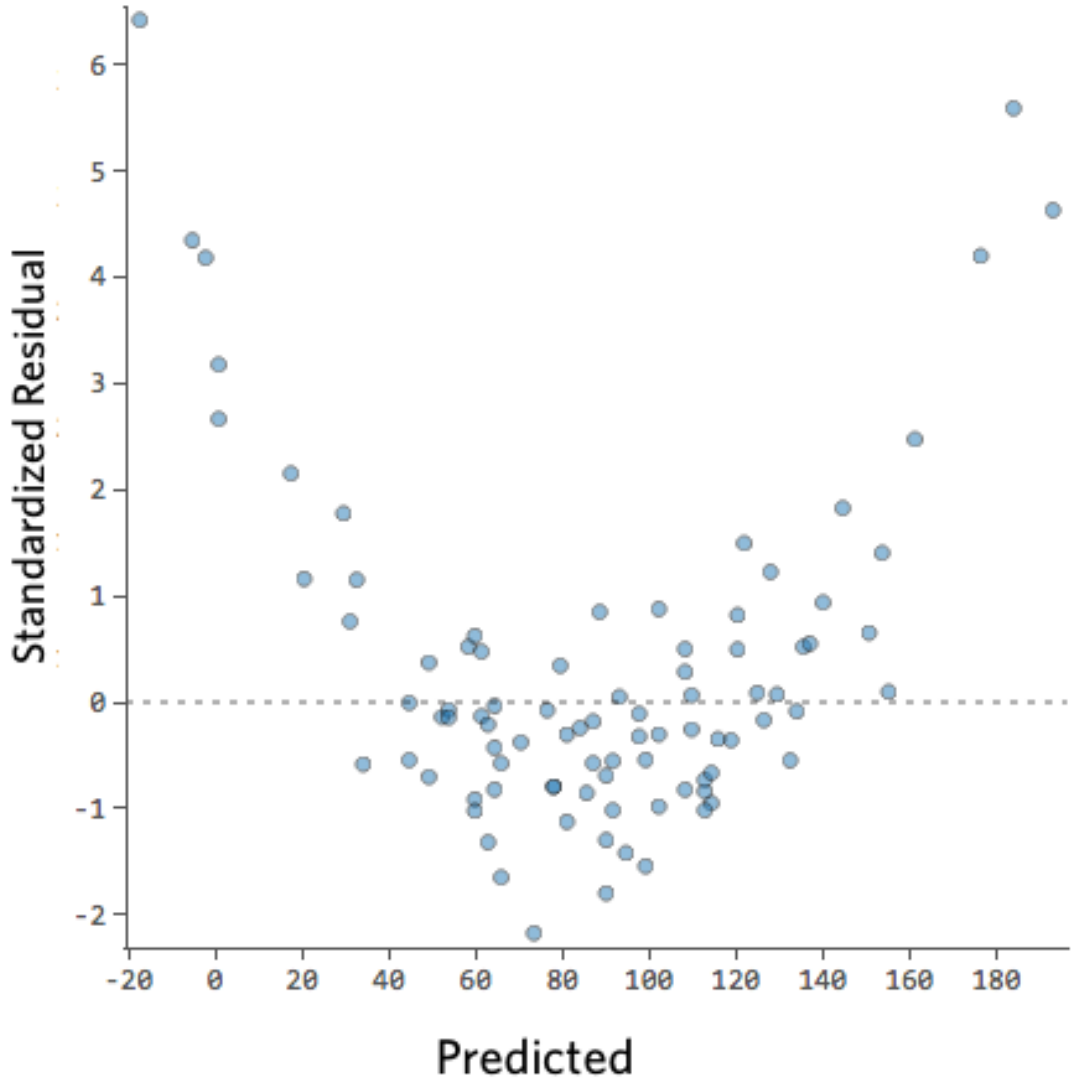

Standardized Residual
6
5
4
3
2
1
0
-1
-2
-20
0
20
40
60
80
100
120
140
160
180
Predicted

Figure 3:

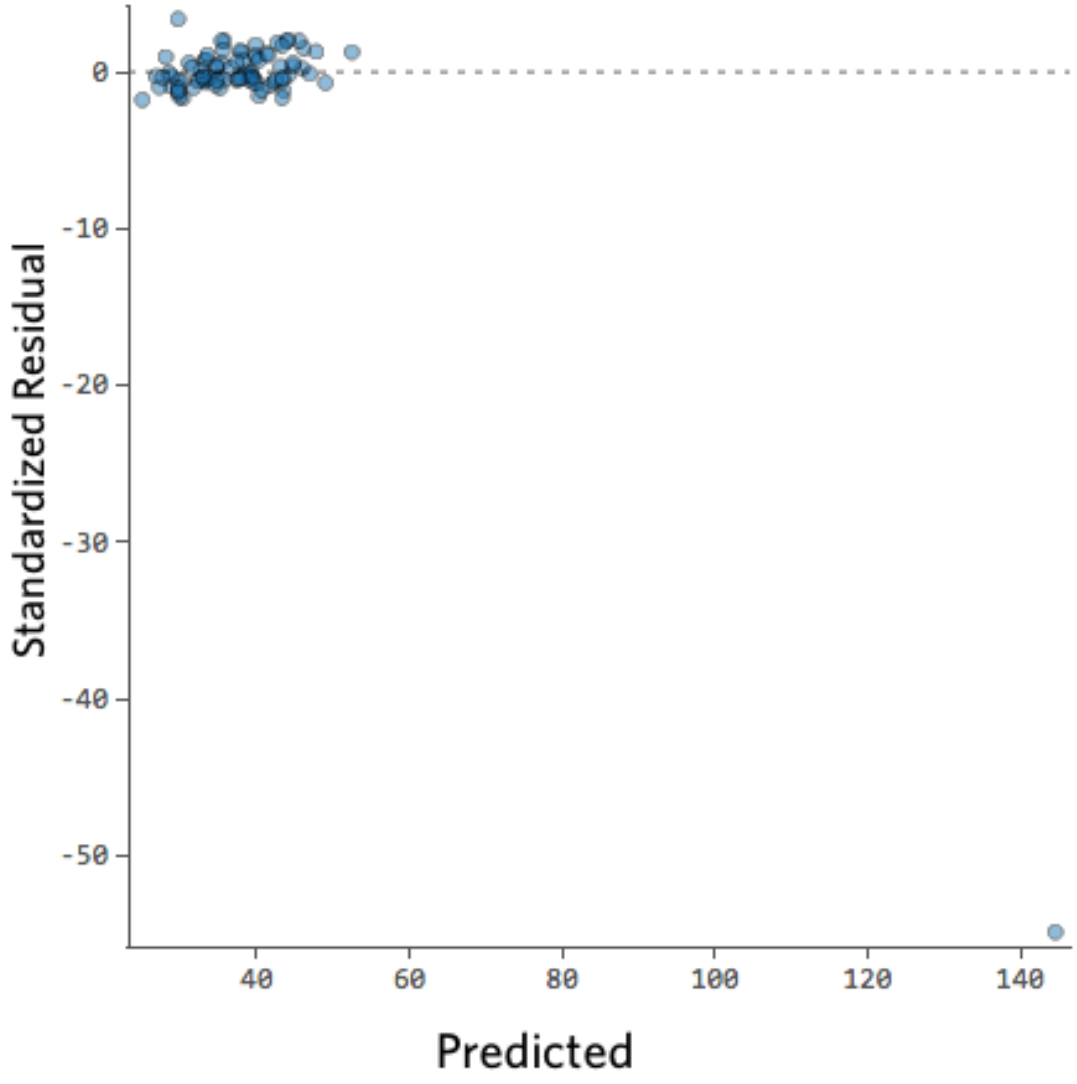

Standardized Residual
0
-10
-20
-30
-40
-50
40
60
80
100
120
140
Predicted

Figure 4:

Проверка на выбросы: расстояние Кука

Расстояние Кука (Cook's distance) для наблюдения показывает, насколько сильно изменится прогнозируемое значение $\hat{Y}$, если удалить это наблюдение из выборки. Иными словами, расстояние Кука характеризует влиятельность наблюдения.

Формула: https://en.wikipedia.org/wiki/Cook's_distance

Какие наблюдения считать влиятельными?

Те, для которых расстояние Кука D_i:

- > 1
- $> 4/N$, где N — число наблюдений (для параноиков).

Еще о влиятельности наблюдений: How to read Cook's distance plots?

Проверка на выбросы: рычаг

Рычаг (leverage) i-го наблюдения равен $h_{ii} = [H]_{ii}$.

Как понять, что такое рычаг в статистике:

https://stats.stackexchange.com/questions/58141/interpreting-plot-lm

Рассмотрим графики остатков для 4-х наборов данных:

1. все хорошо
2. есть точка с длинным рычагом, но малым остатком
3. есть точка с коротким рычагом и большим остатком
4. есть точка с длинным рычагом и большим остатком

```
set.seed(20)

x1 = rnorm(20, mean=20, sd=3)
y1 = 5 + .5*x1 + rnorm(20)

x2 = c(x1, 30);          y2 = c(y1, 20.8)
x3 = c(x1, 19.44);       y3 = c(y1, 20.8)
x4 = c(x1, 30);          y4 = c(y1, 10)
```

Все хорошо

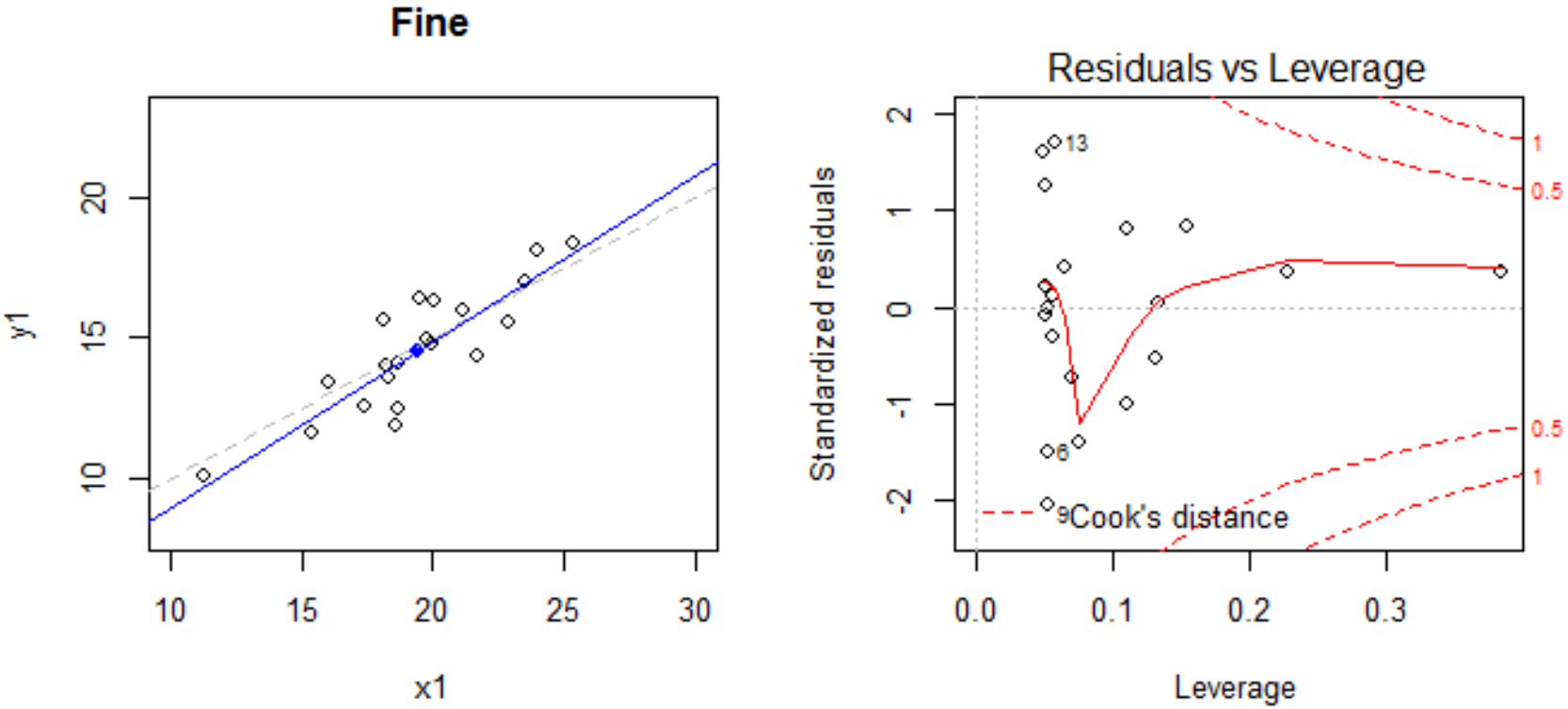


Figure 5:

Точка с длинным рычагом и малым остатком

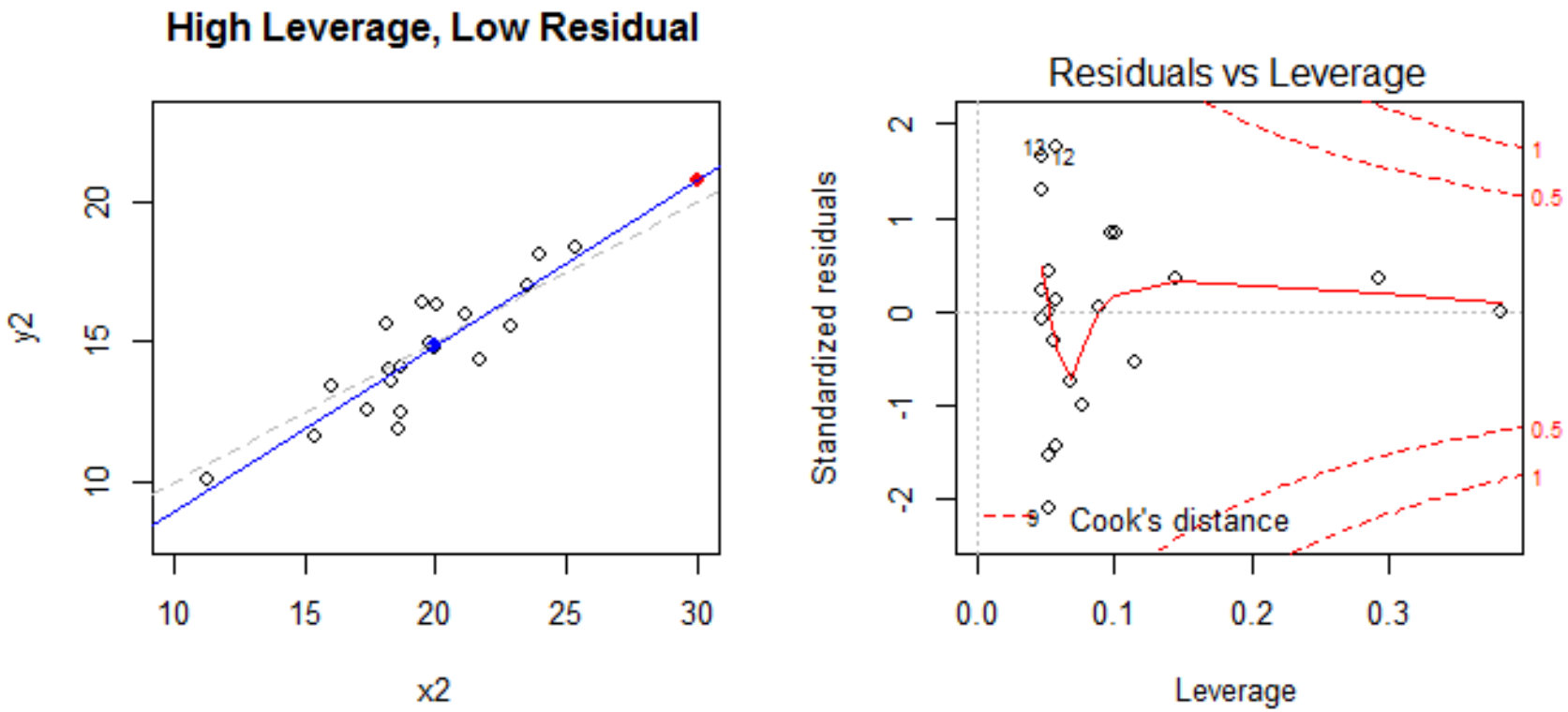


Figure 6:

Точка с коротким рычагом и большим остатком

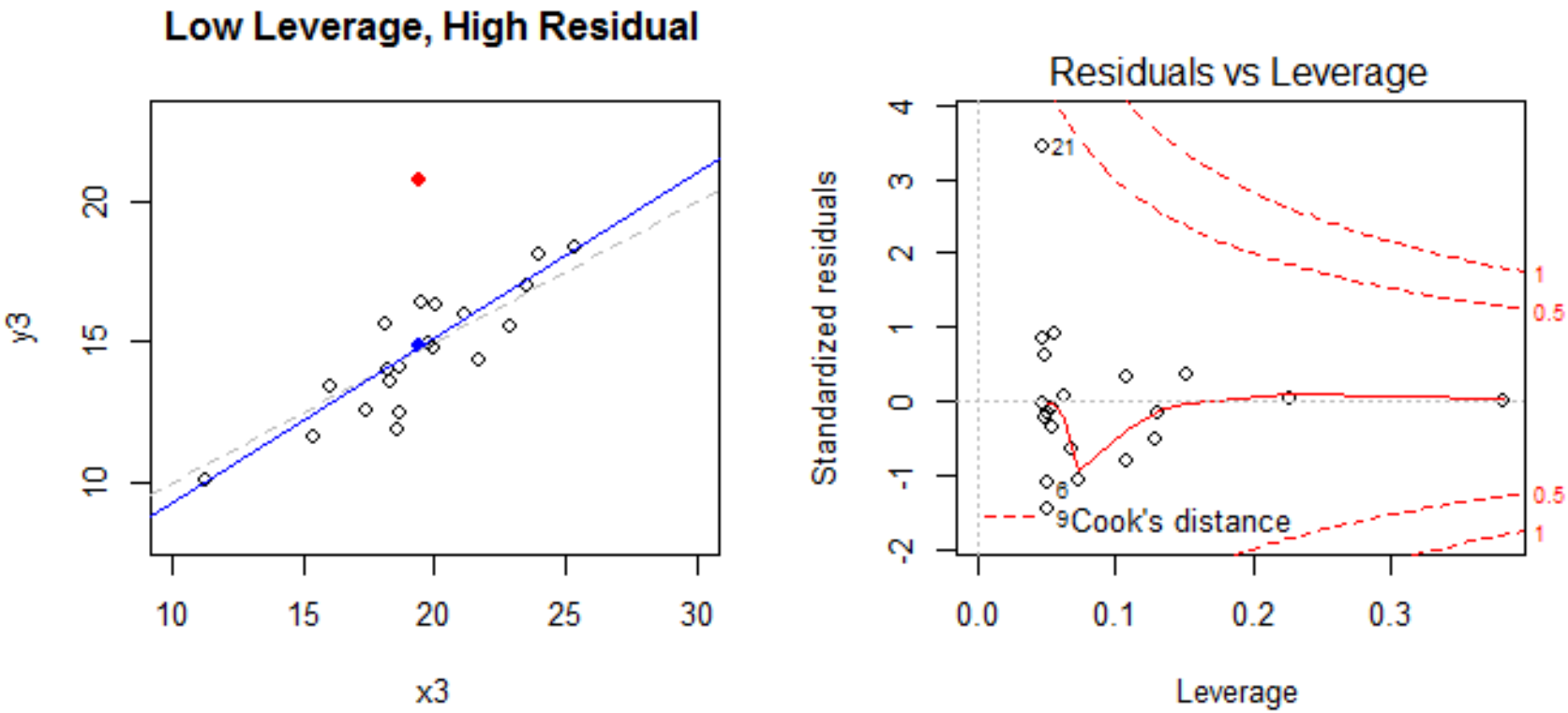


Figure 7:

Точка с длинным рычагом и большим остатком

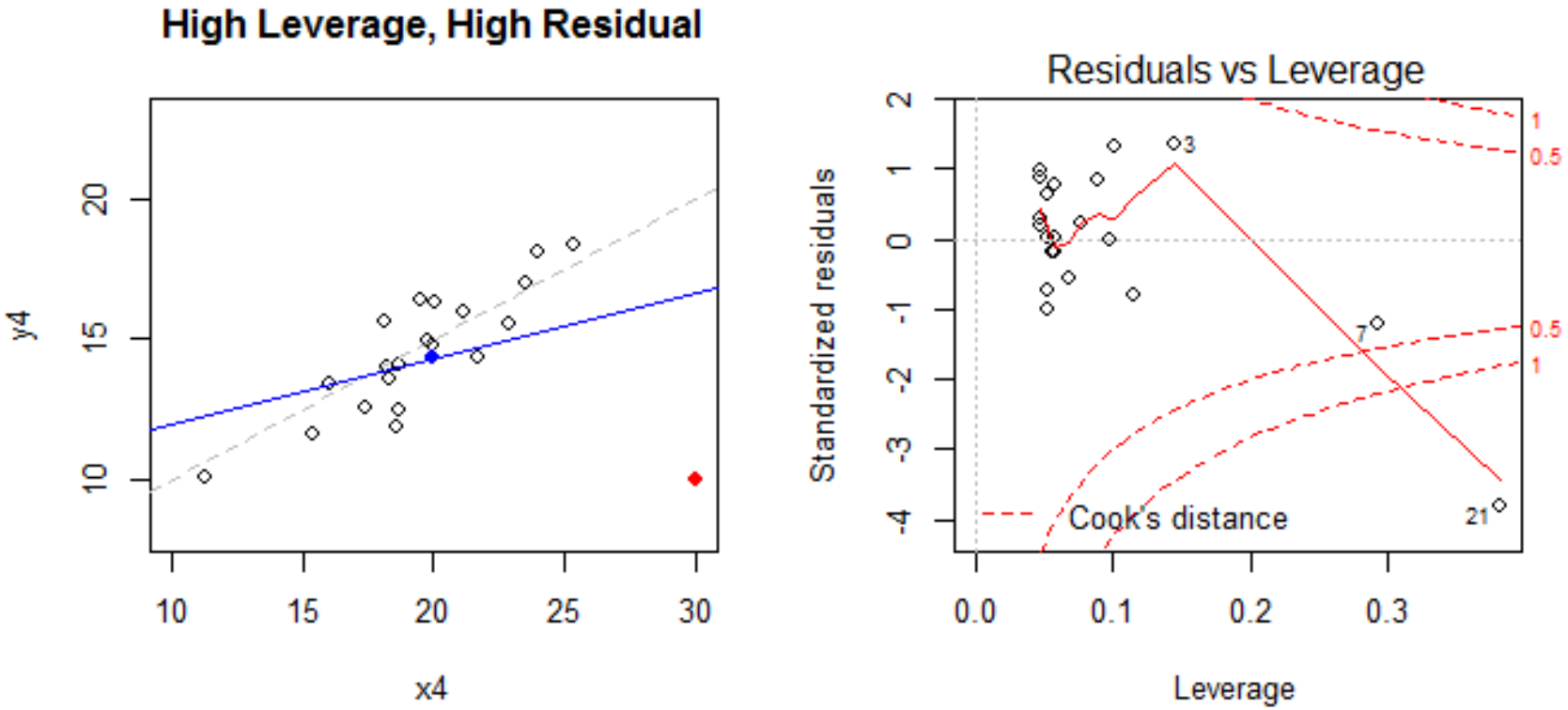
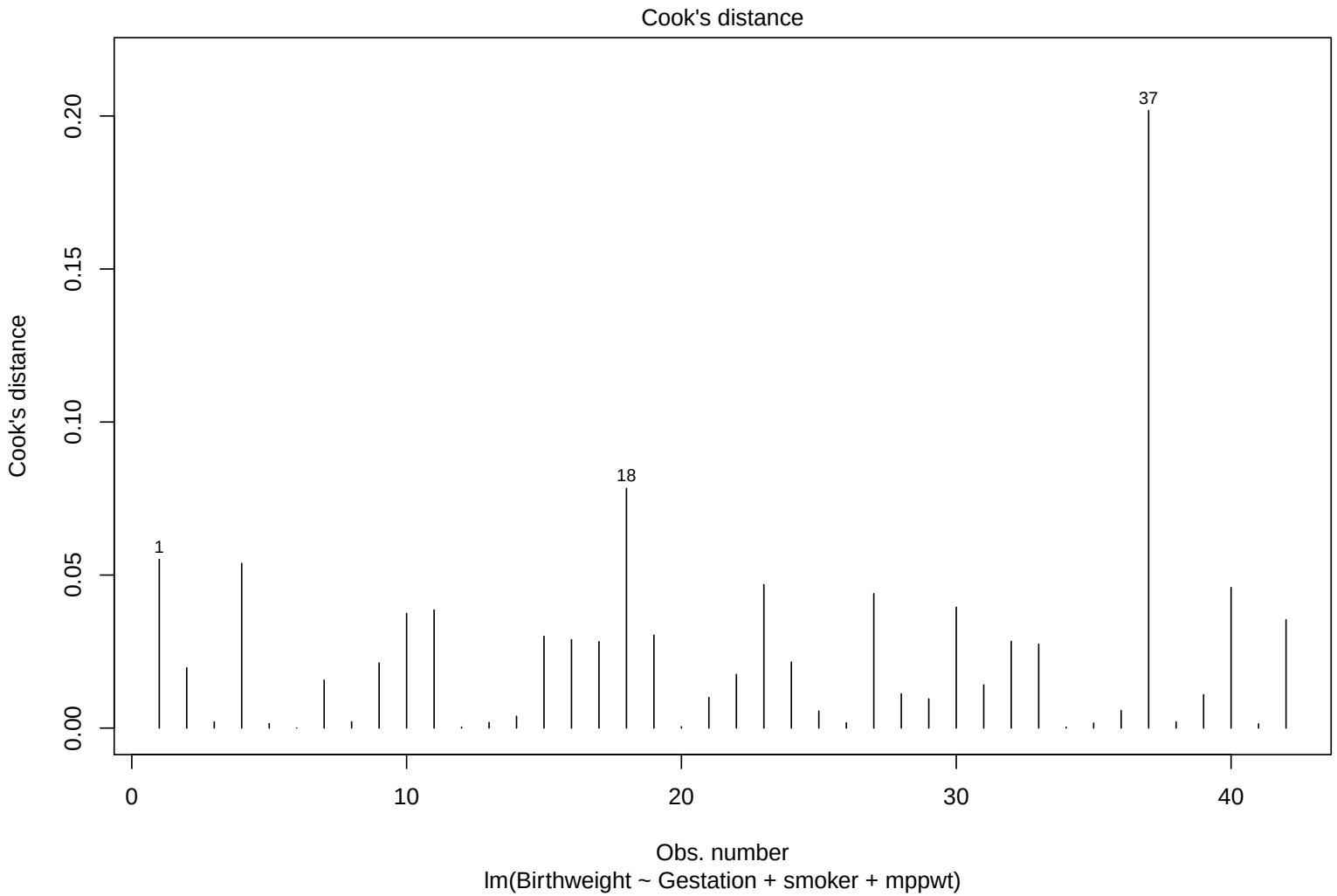


Figure 8:

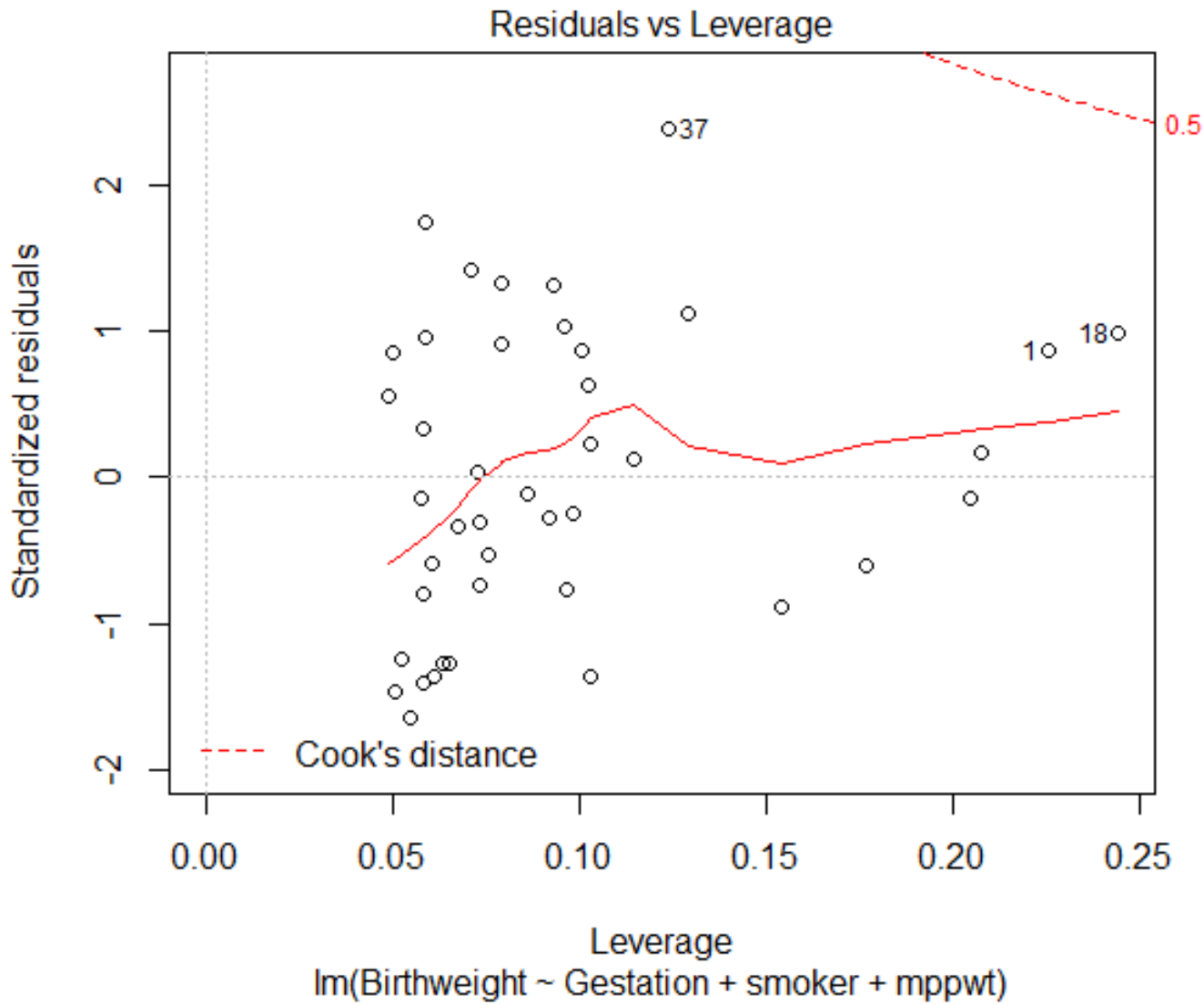
Возвращаемся к исследованию веса новорожденных...

```
plot(reg1, which = 4)
```



4-й график из серии plot()

```
plot(reg1, which = 5)
```



Автокорреляция остатков

Одним из предположений относительно регрессии является независимость наблюдений, которая выражается в отсутствии корреляции между остатками. Однако, если наблюдается процесс развивающийся во времени, то вполне возможно, что последовательные на[illegible] с другом.

Тест Дурбина-Ватсон[illegible] т отсутствие автокорреляции. Ес[illegible] последовательные на[illegible] то статистика Дурбина-[illegible] делах от 1.5 до 2.5. p-значение о[illegible] 5.

Тест Дурбина-Ватсона

```
library(car)
dwt(reg1)
```

```
 lag Autocorrelation D-W Statistic p-value
   1       -0.0774652      2.122239   0.782
 Alternative hypothesis: rho != 0
```

Предположение об отсутствии автокорреляции обычно нарушается для временных рядов.

Выводы на основе регрессионной модели

Множественная линейная регрессия проводилась в целях изучения взаимосвязи между весом новорожденного и 1) продолжительностью беременности (в неделях), 2) весом матери до беременности и 3) курением матери в течение беременности.

Показано наличие значимой взаимосвязи между продолжительностью беременности (гестационным возрастом) и весом новорожденного ($p < 0.001$), курением матери и весом новорожденного ($p = 0.017$), а также весом матери до беременности и весом новорожденного ($p = 0.03$).

Отмечено увеличение веса новорожденного на 0.313 фунта за каждую дополнительную неделю гестационного возраста.

На каждый дополнительный фунт веса матери, вес новорожденного увеличивается на 0.02 фунта.

Новорожденные у курящих матерей весят на 0.665 фунта меньше, чем у некурящих.

Скорректированный коэффициент детерминации (adjusted R-squared) равен 0.58, следовательно 58% изменений веса новорожденных может быть объяснено регрессионной моделью, включающей гестационный возраст, вес и курение матери.

Дополнительные материалы

- Interpreting residual plots to improve your regression
- Statistical Analyses Using R resources — интересный вводный материал по регрессионному анализу и вообще по анализу данных (лекции, скрипты, наборы данных)
- Quick-R: Regression Diagnostics