Линейная регрессия

Постановка в одномерном случае

x - некий признак объекта (**независимая переменная**)

y - предсказываемая величина (**зависимая переменная**)

Предположим, что y = f(x) + eps (eps - шум, распределенный нормально)

Хотим найти такую функцию h(x) = bx + a, которая **лучше всего** аппроксимирует эту зависимость

Mean Squared Loss

Остаток, **residual**

$$r_i = y_i - \hat{y}_i = y_i - h(x) = y_i - bx - a$$

Хотим минимизировать функцию

$$MSE = \frac{1}{N}\sum r_i^2$$

Построение модели в R

```r
# install.packages('datarium')
data("marketing", package = "datarium")

model <- lm(sales ~ youtube, data = marketing)
model
```

Независимая переменная

Предсказываемая величина

Датасет

```
##
## Call:
## lm(formula = sales ~ youtube, data = marketing)
##
## Coefficients:
## (Intercept)      youtube
##     8.43911      0.04754
```

Коэффициент при переменной (b)

Свободный коэффициент (a)

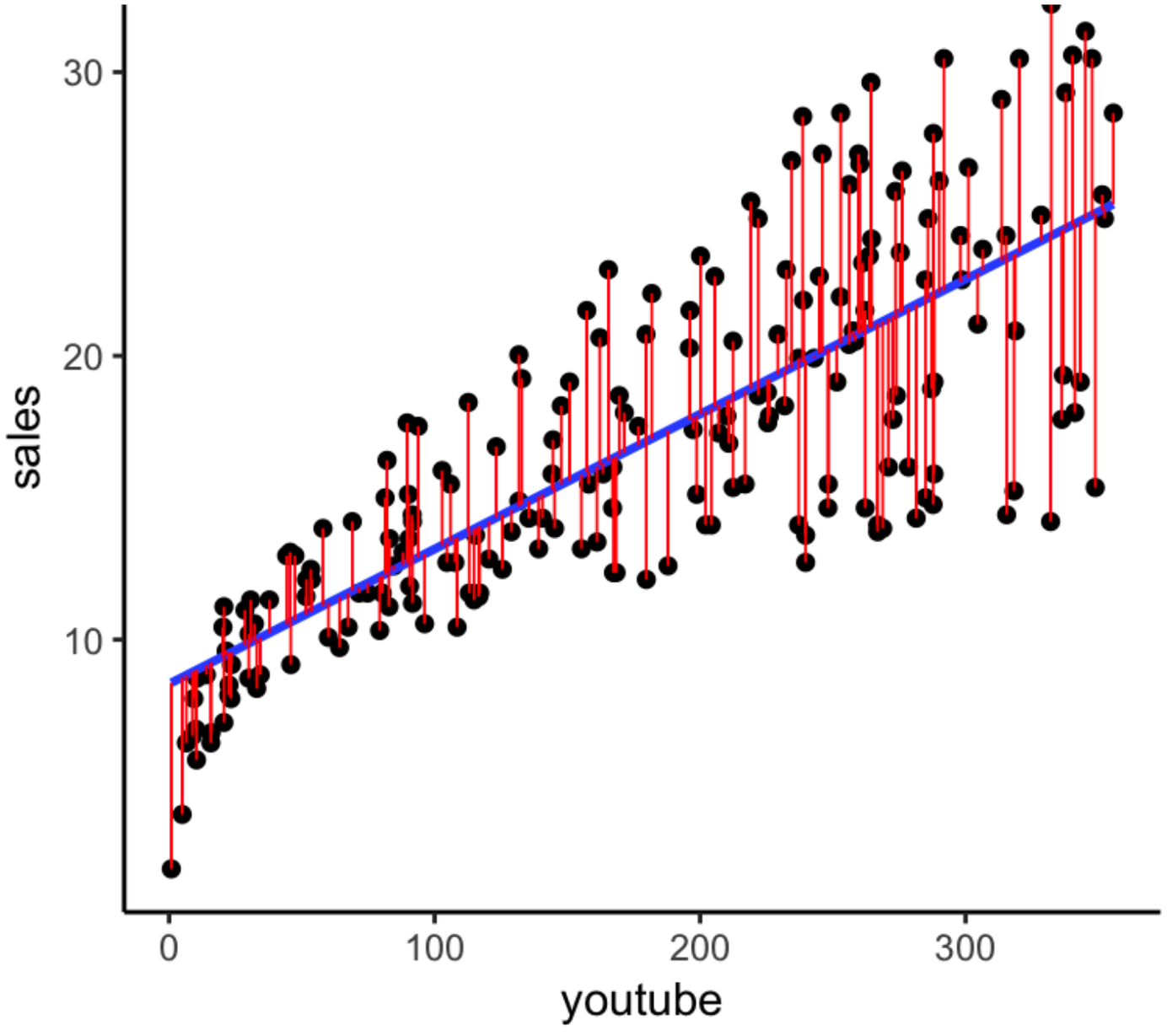
Данные о модели

```
library(broom)
model.diag.metrics <- augment(model)
head(model.diag.metrics)
```

```
## # A tibble: 6 x 9
##   sales youtube .fitted .se.fit .resid    .hat .sigma   .cooksd .std.resid
##   <dbl>   <dbl>   <dbl>   <dbl>  <dbl>   <dbl>  <dbl>     <dbl>      <dbl>
## 1 26.5    276.    21.6    0.385  4.96  0.00970   3.90 0.00794        1.27
## 2 12.5     53.4   11.0    0.431  1.50  0.0122    3.92 0.000920       0.387
## 3 11.2     20.6    9.42   0.502  1.74  0.0165    3.92 0.00169        0.449
## 4 22.2    182.    17.1    0.277  5.12  0.00501   3.90 0.00434        1.31
## 5 15.5    217.    18.8    0.297 -3.27  0.00578   3.91 0.00205       -0.839
## 6  8.64    10.4    8.94   0.525 -0.295 0.0180    3.92 0.0000534     -0.0762
```

Residuals

```
ggplot(model.diag.metrics, aes(youtube, sales)) +
  geom_point() +
  stat_smooth(method = lm, se = FALSE) +
  geom_segment(aes(xend = youtube, yend = .fitted), color = "red", size = 0.3)
```



Много переменных

xj - некий признак объекта (**независимая переменная**)

y - предсказываемая величина (**зависимая переменная**)

Предположим, что y = f(x) + eps (eps - шум, распределенный нормально)

Хотим найти такую функцию h(x) = bx1 + … + bxn + a, которая **лучше всего** аппроксимирует эту зависимость

Решается аналогично

Много переменных

```r
# install.packages('datarium')
data("marketing", package = "datarium")

model <- lm(sales ~ youtube + facebook + newspaper, data = marketing)
model
```

```
##
## Call:
## lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)
##
## Coefficients:
## (Intercept)      youtube     facebook    newspaper
##    3.526667     0.045765     0.188530    -0.001037
```

Что будет, если у нас есть 1000 наблюдений и 10000 переменных?

Что будет, если у нас есть 1000 наблюдений и 10000 переменных?

Всегда найдем все хорошо объясняющий набор переменных

Статистическая значимость переменных

```
summary(model)
```

```
##
## Call:
## lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5932  -1.0690   0.2902   1.4272   3.3951
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.526667   0.374290   9.422   <2e-16 ***
## youtube      0.045765   0.001395  32.809   <2e-16 ***
## facebook     0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.023 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Значимость коэффициента
(p-value гипотезы о том, что коэффициент равен

Статистическая значимость переменных

```
##
## Call:
## lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5932  -1.0690   0.2902   1.4272   3.3951
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.526667   0.374290   9.422   <2e-16 ***
## youtube      0.045765   0.001395  32.809   <2e-16 ***
## facebook     0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.023 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Значимость коэффициента (p-value гипотезы о том, что коэффициент равен 0)

Число степеней свободы теста

Статистическая значимость переменных

```
##
## Call:
## lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5932  -1.0690   0.2902   1.4272   3.3951
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.526667   0.374290   9.422   <2e-16 ***
## youtube      0.045765   0.001395  32.809   <2e-16 ***
## facebook     0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.023 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Значимость коэффициента (p-value гипотезы о том, что коэффициент равен 0)

Число степеней свободы теста

R-squared

$$SS_{tot} = \sum_i (y_i - \bar{y})^2$$

$$SS_{reg} = \sum_i (y_i - \hat{y}_i)^2 \equiv \sum_i r_i^2$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Коэффициент детерминации, в случае выполнения некоторых предположений, доля объясняемой **дисперсии**

R-squared

Какие проблемы вы видите?

R-squared

Какие проблемы вы видите?

$$SS_{reg} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i r_i^2$$

Чем лучше описываем наблюдения, тем лучше будет SSreg.

R-squared

Какие проблемы вы видите?

$$SS_{reg} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i r_i^2$$

Чем лучше описываем наблюдения, тем лучше будет SSreg.

Чем больше переменных, тем лучше описываем наблюдения

R-squared

Какие проблемы вы видите?

$$SS_{reg} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i r_i^2$$

Чем лучше описываем наблюдения, тем лучше будет SSreg.

Чем больше переменных, тем лучше описываем наблюдения

Чем больше переменных - тем лучше R-squared

Adjusted R-squared

$$R^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1}$$

n - число наблюдений, p - число независимых переменных

Статистическая значимость переменных

```
##
## Call:
## lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5932  -1.0690   0.2902   1.4272   3.3951
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.526667   0.374290   9.422   <2e-16 ***
## youtube      0.045765   0.001395  32.809   <2e-16 ***
## facebook     0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.023 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Значимость коэффициента (p-value гипотезы о том, что коэффициент равен 0)

Число степеней свободы теста

Допущения линейной регрессии

1) Предсказываемая переменная зависит от независимых линейно
2) Независимые переменные друг от друга не зависят
3) Residuals распределены нормально
4) Residuals имееют одинаковую дисперсию
5) Residuals независимы (= наблюдения независимы)

Допущения линейной регрессии

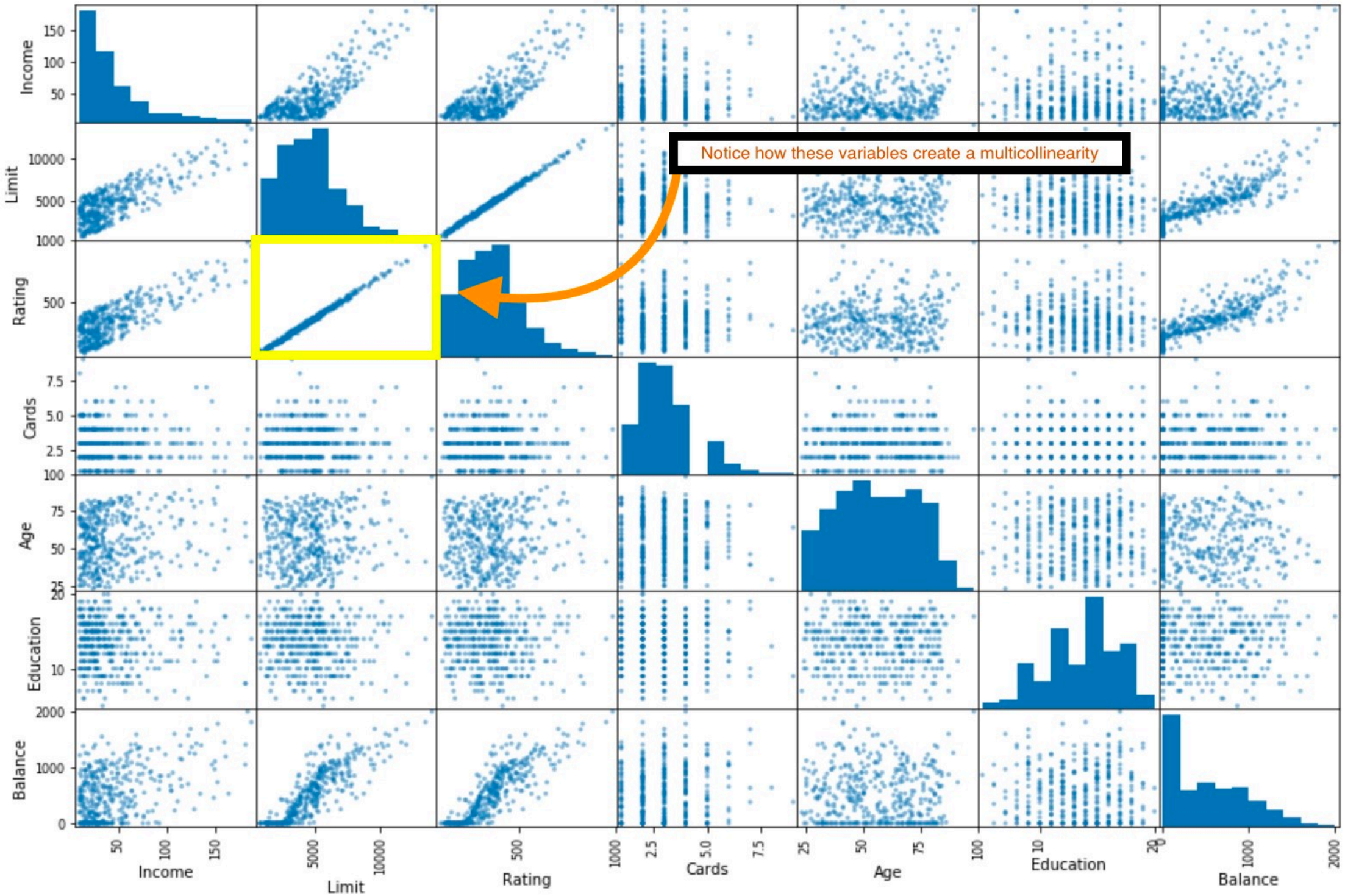
1) Предсказываемая переменная зависит от независимых линейно
2) Независимые переменные друг от друга не зависят
3) Residuals распределены нормально
4) Residuals имееют одинаковую дисперсию
5) Residuals независимы (= наблюдения независимы)

Независимые переменные друг от друга не зависят (нет мультиколлинеарности)

Мультиколлинеарность приводит к менее точно определенным коэффициентам, что может приводить к неприятным эффектам. Это более значимо, когда мы хотим интерпретировать модель. На предсказательную силу модели это (почти) не влияет

Мультиколлинеарность

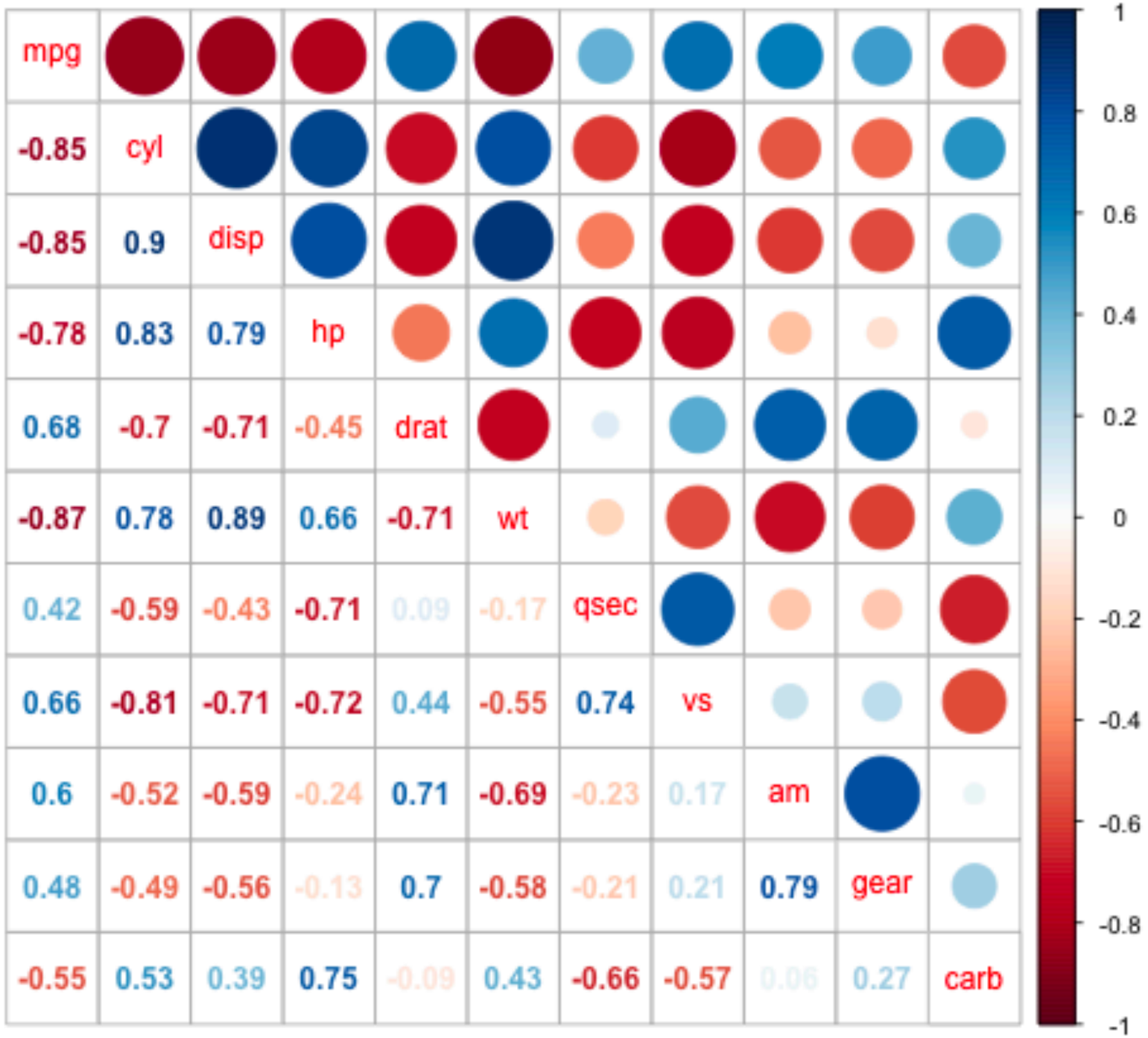
Построить зависимости между переменными



Мультиколлинеарность

Посчитать корреляцию между переменными

https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html

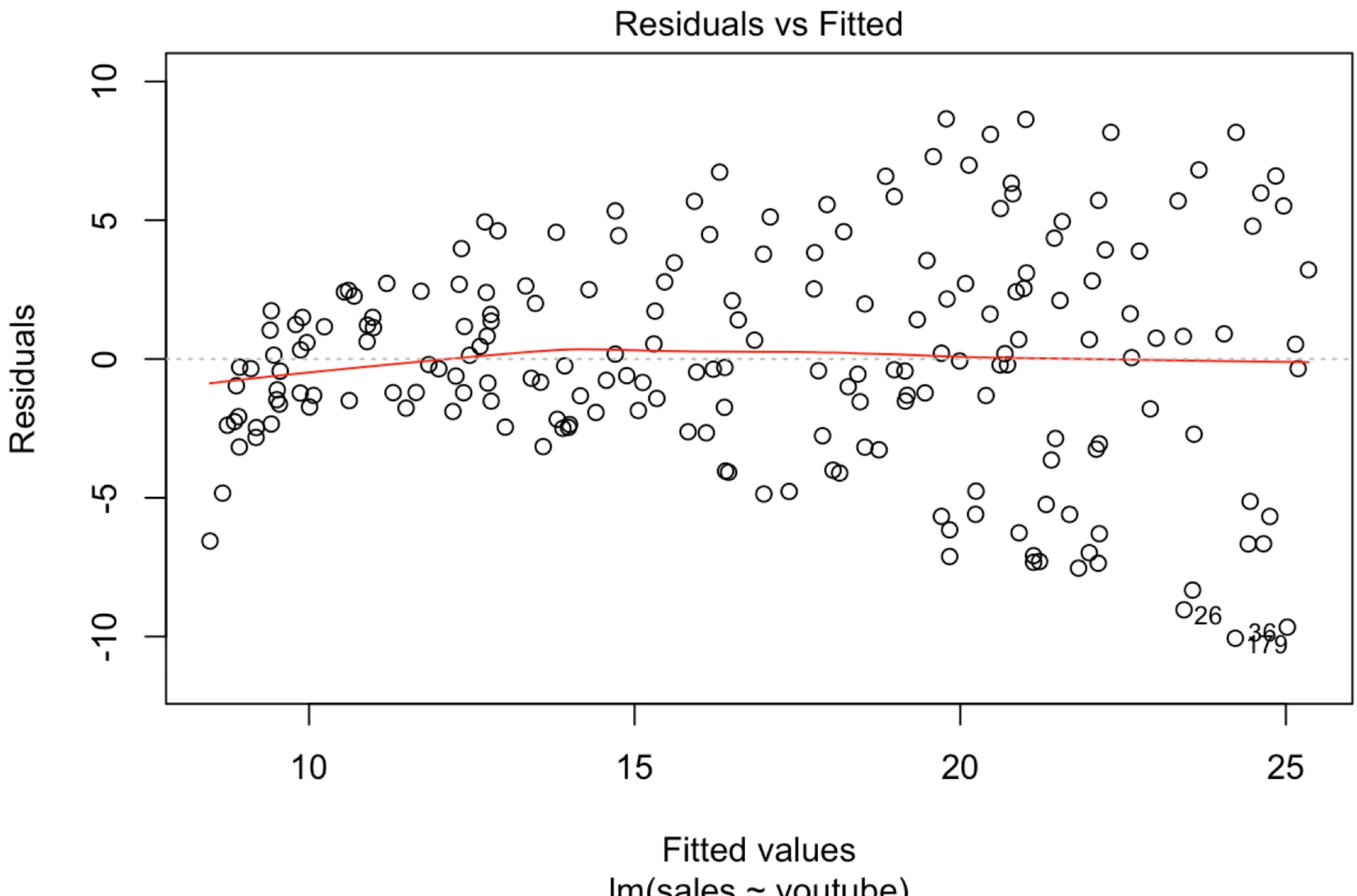


Допущения линейной регрессии

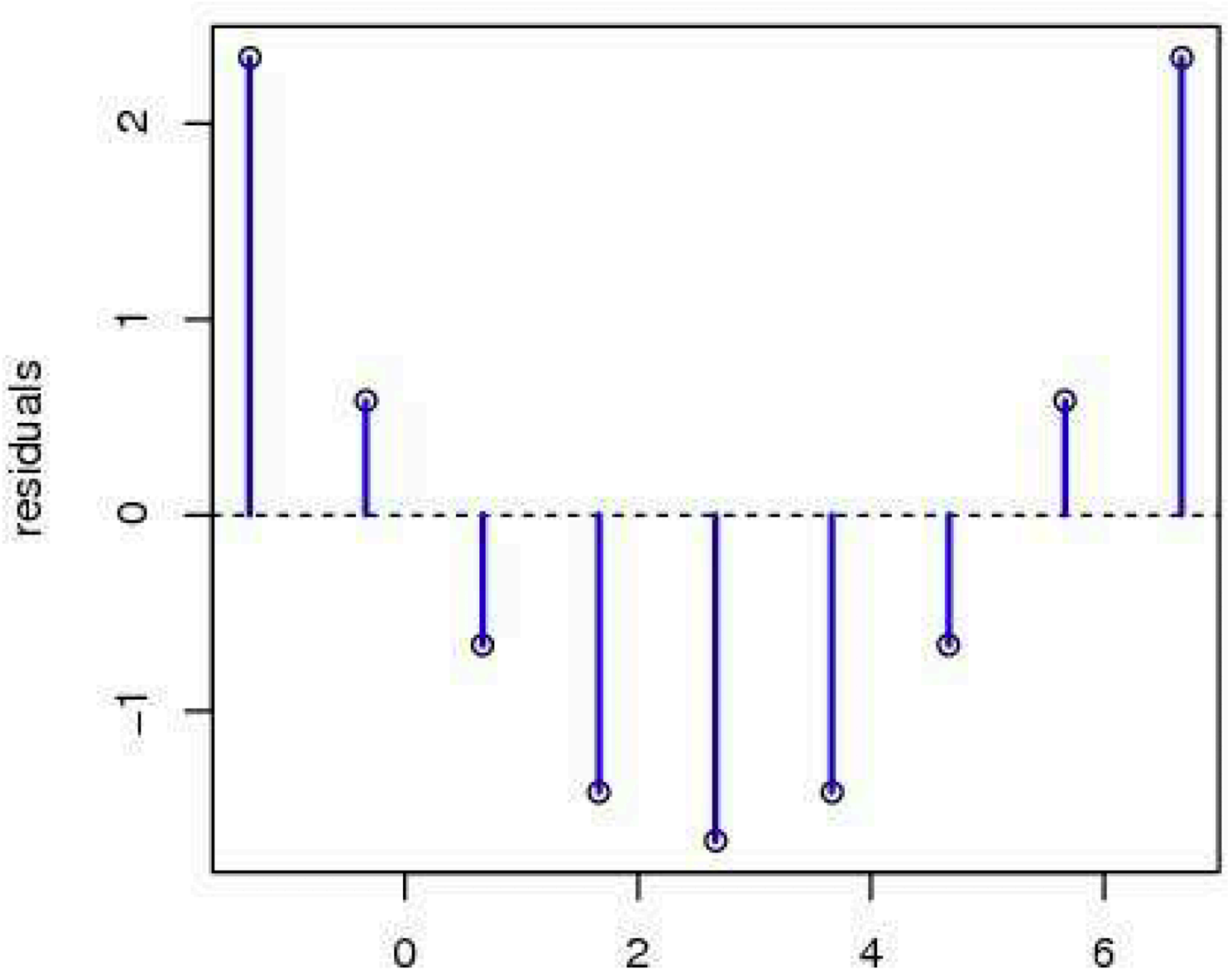
1) **Предсказываемая переменная зависит от независимых линейно**
2) **Независимые переменные друг от друга не зависят**
3) **Residuals распределены нормально**
4) **Residuals имееют одинаковую дисперсию**
5) **Residuals независимы (= наблюдения независимы)**

Проверка на линейность

```
model <- lm(sales ~ youtube, data = marketing)
plot(model, 1)
```



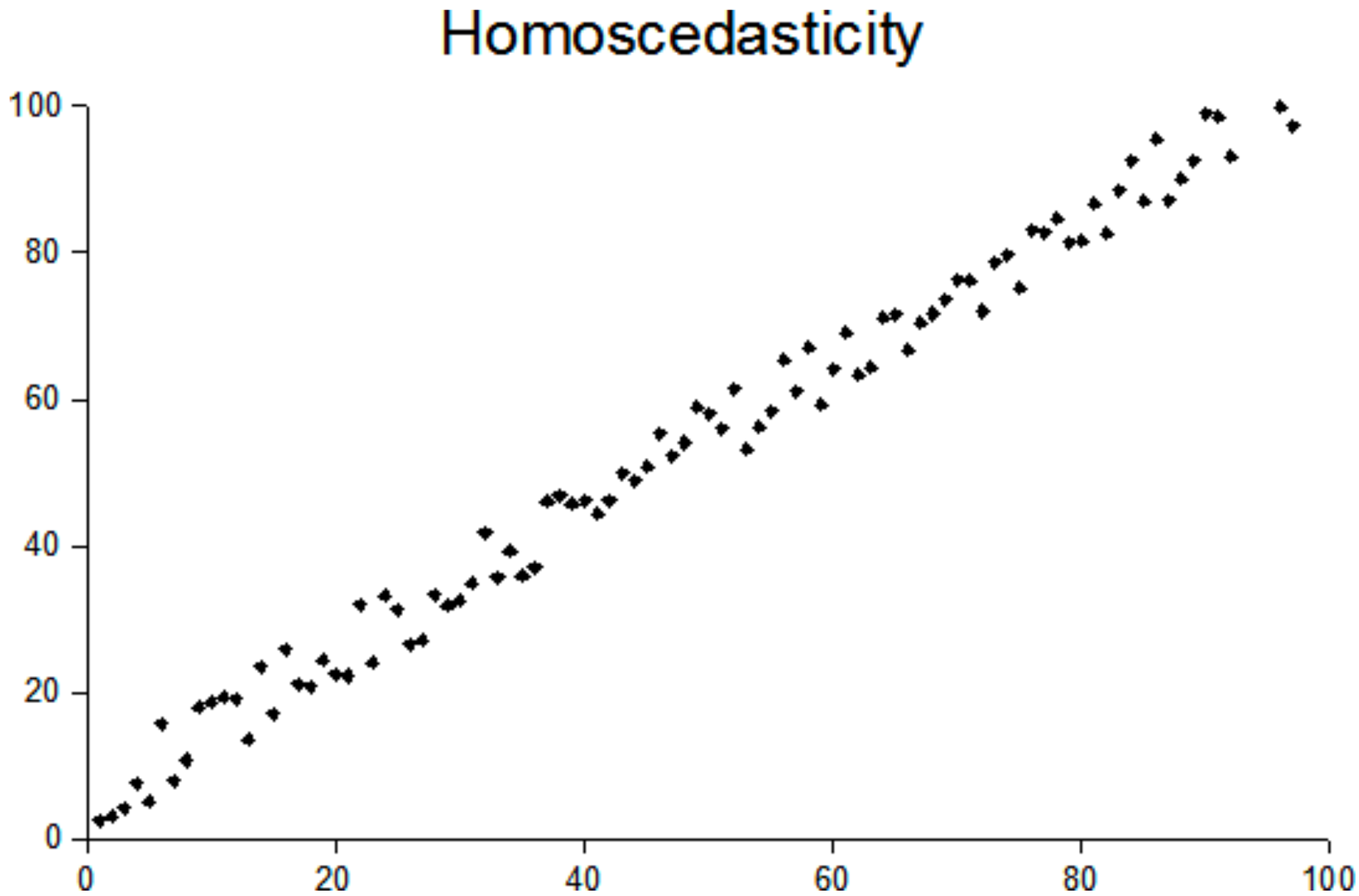
Пример нелинейности



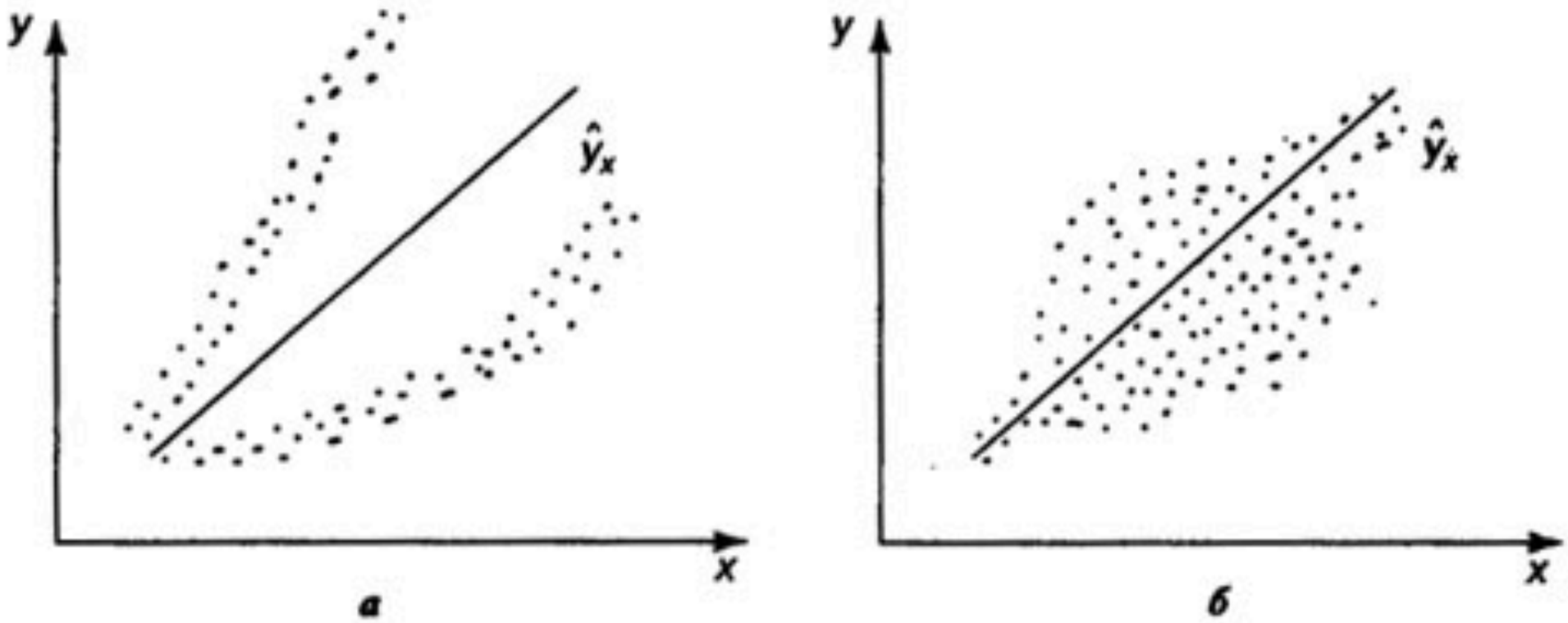
Допущения линейной регрессии

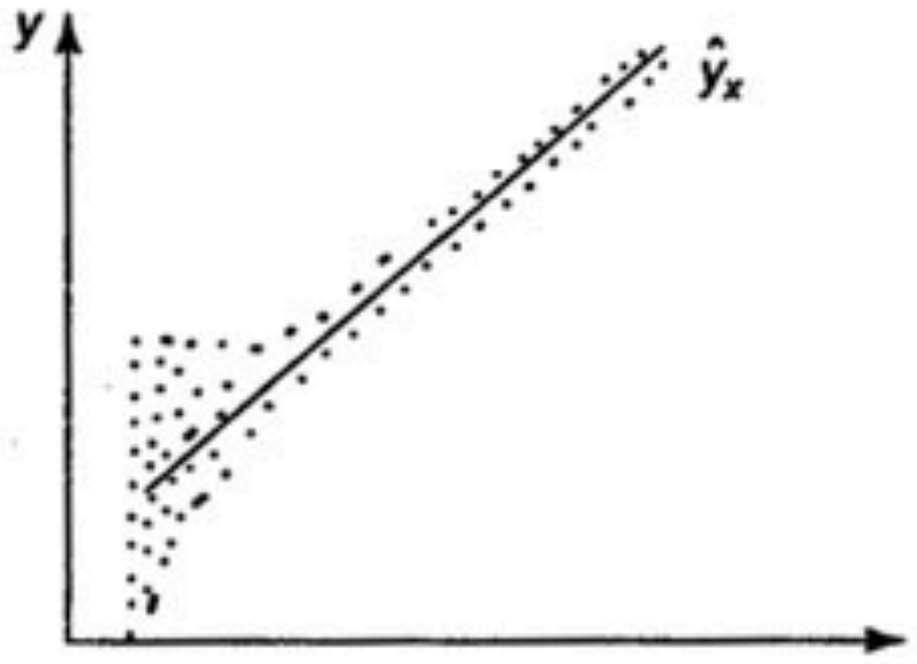
1) Предсказываемая переменная зависит от независимых линейно
2) Независимые переменные друг от друга не зависят
3) Residuals распределены нормально
4) Residuals имееют одинаковую дисперсию
5) Residuals независимы (= наблюдения независимы)

Гомоскедастичность (homogenity of variance)



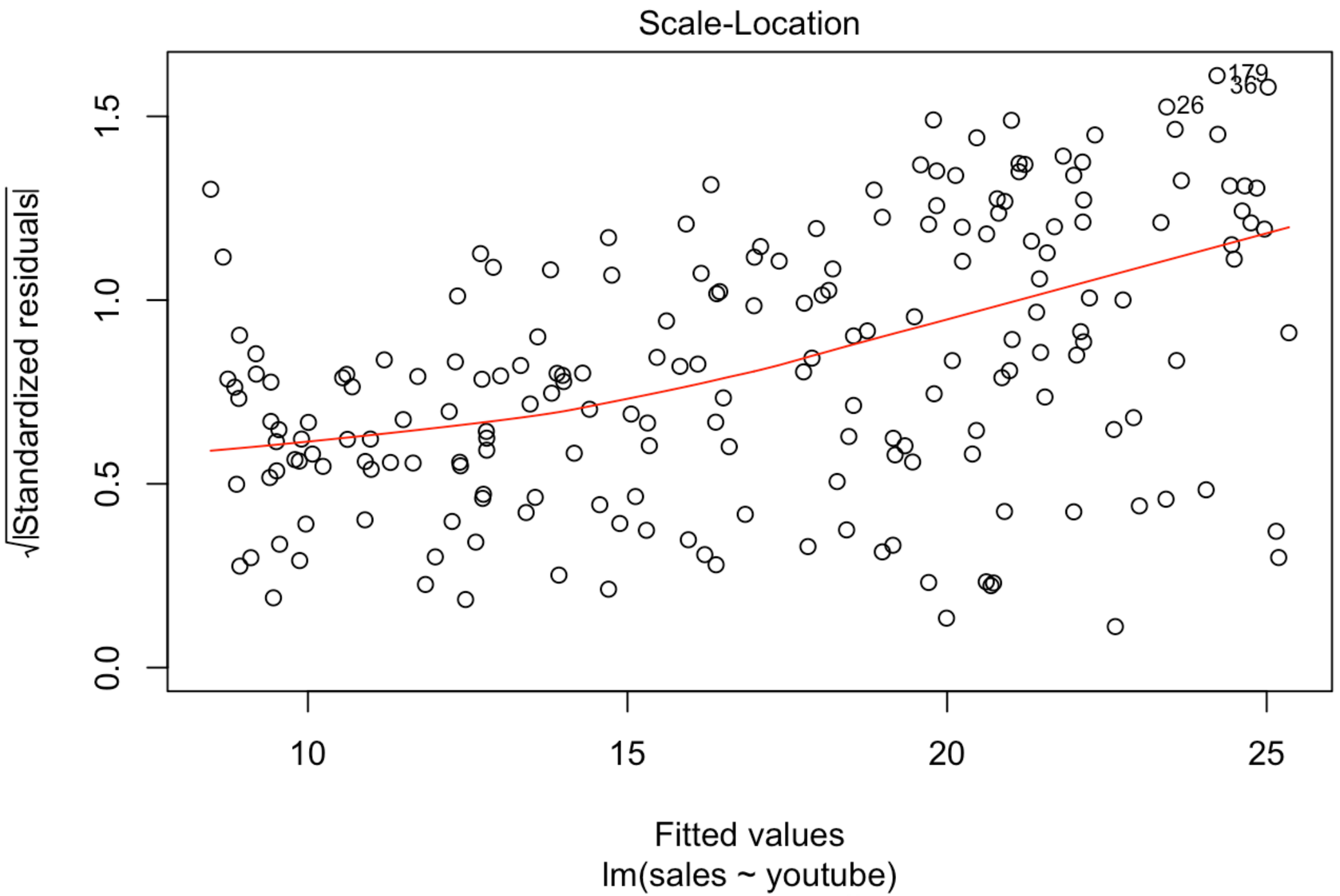
Гетероскедастичность (heterogenity of variance)





Homogeneity of variance (гомоскедастичность)

```
plot(model, 3)
```

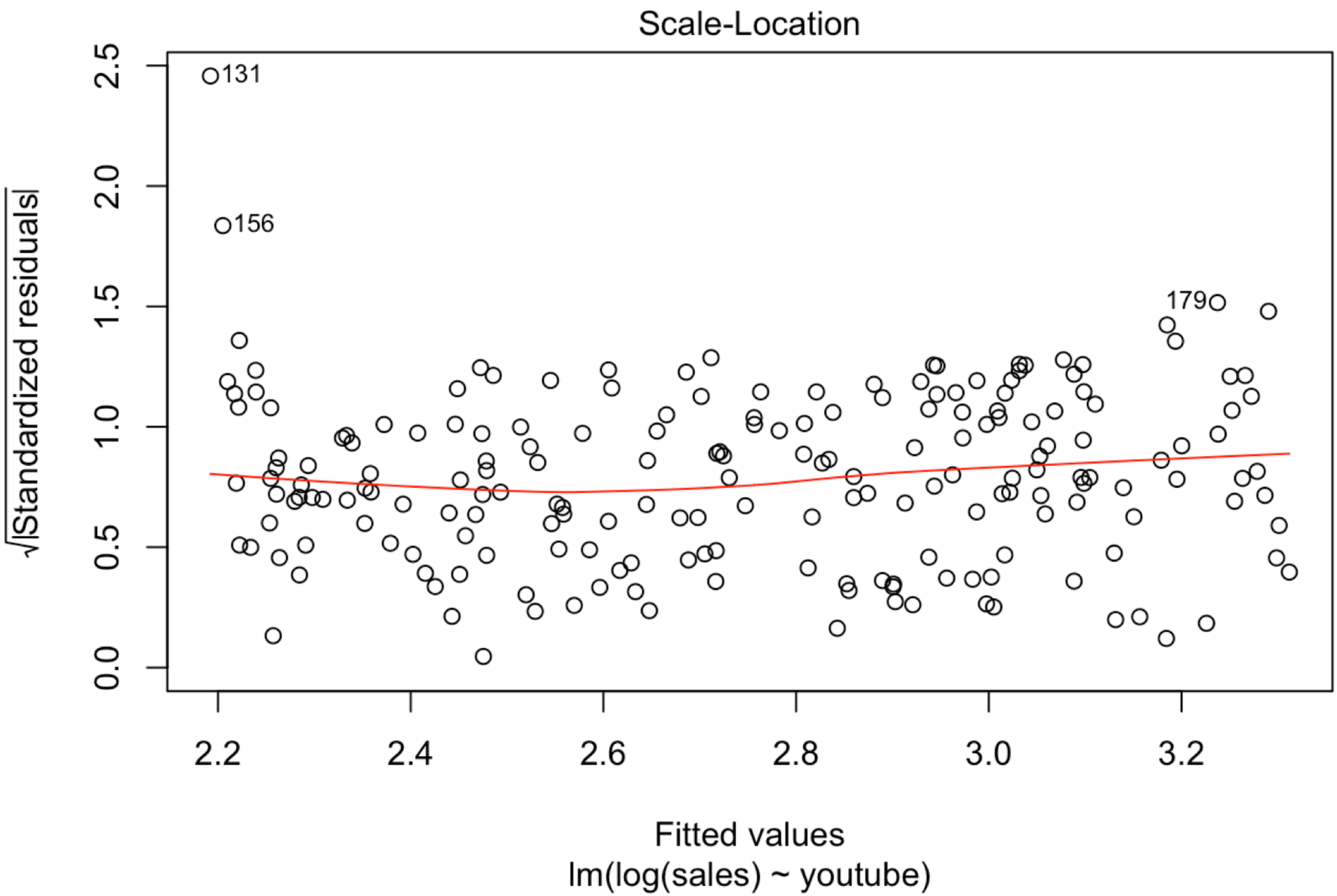


Не наш случай

Homogeneity of variance (гомоскедастичность)

Иногда помогают преобразования (например,логарифмирование)

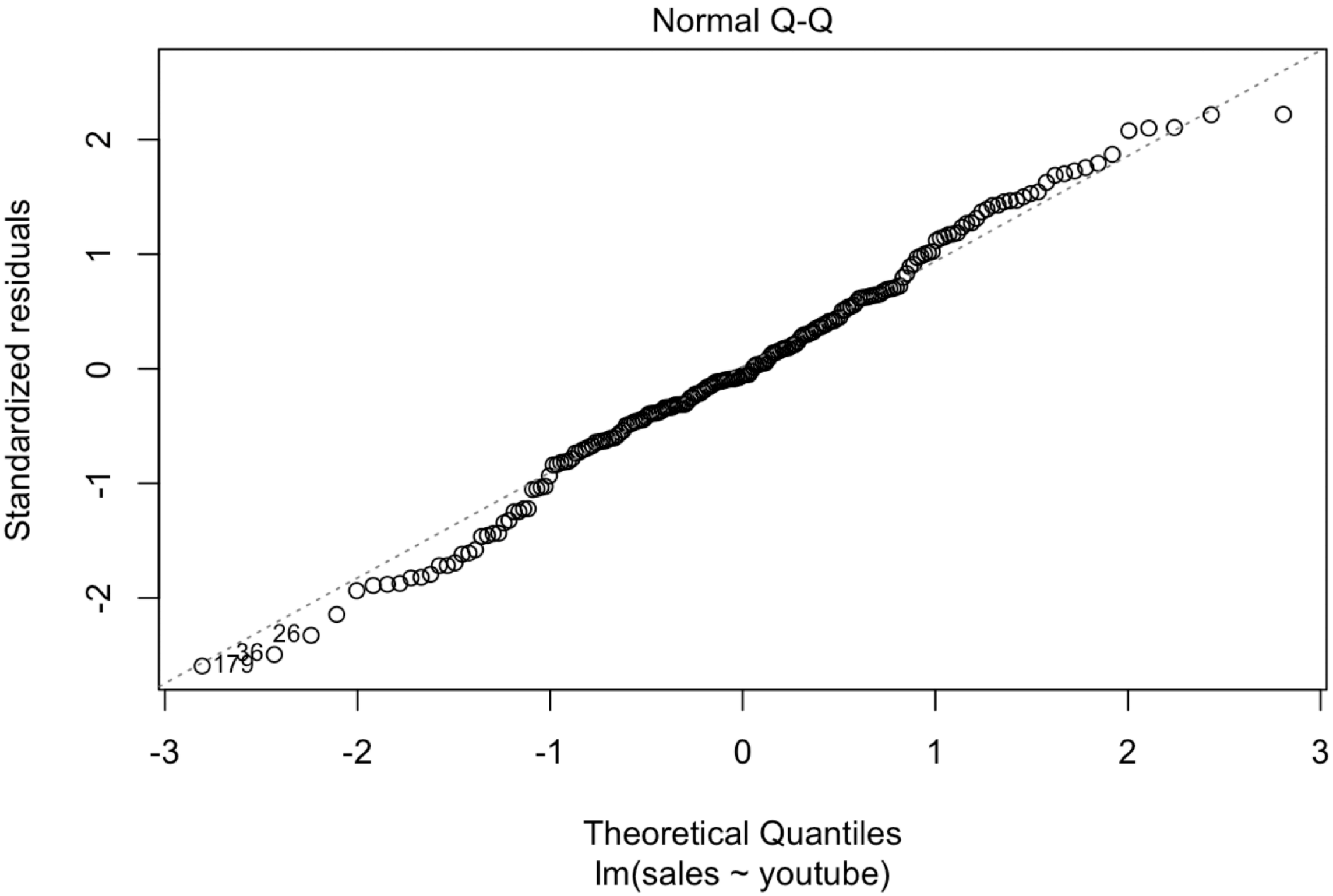
```
model2 <- lm(log(sales) ~ youtube, data = marketing)
plot(model2, 3)
```



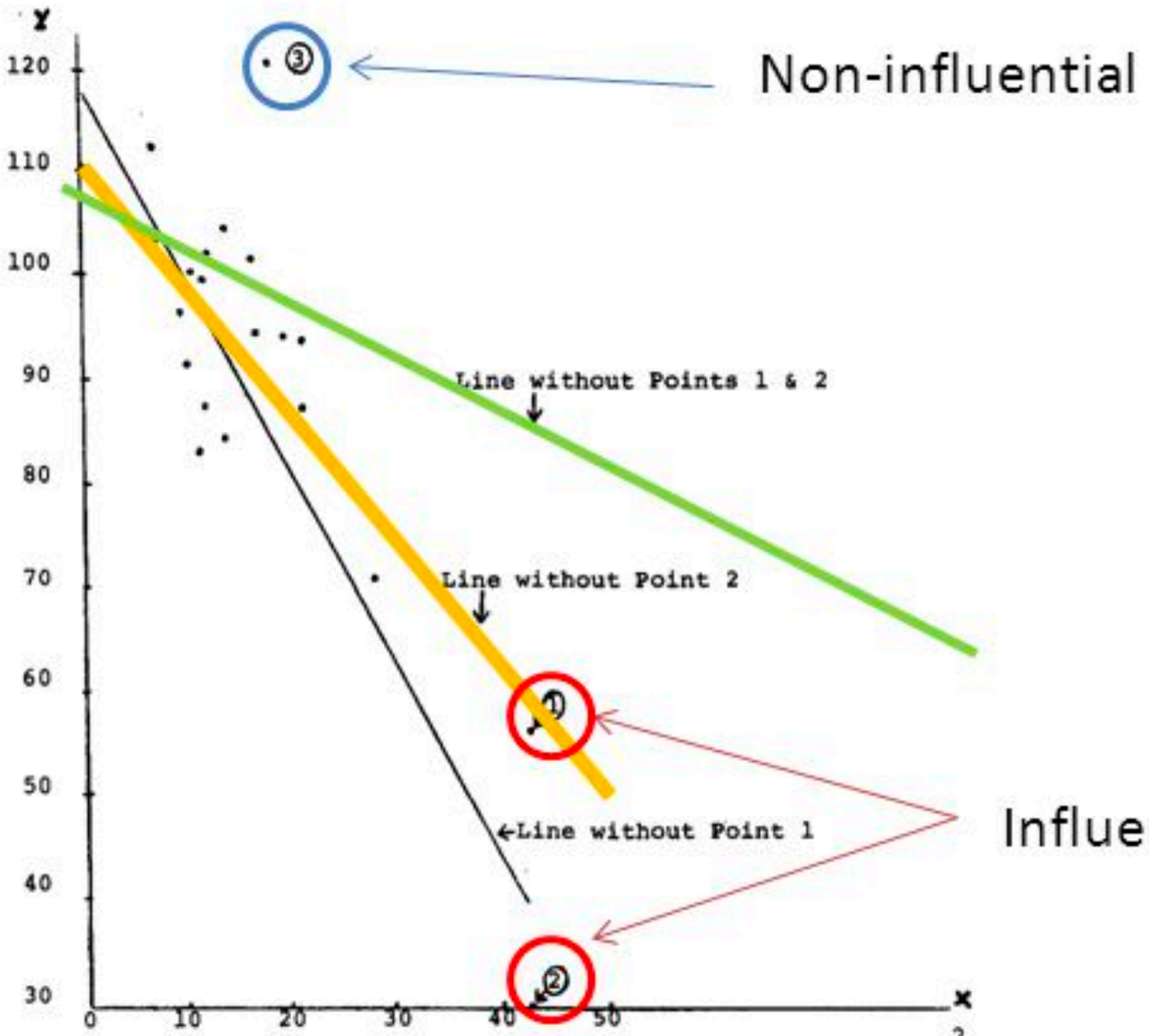
Допущения линейной регрессии

1) Предсказываемая переменная зависит от независимых линейно
2) Независимые переменные друг от друга не зависят
3) Residuals распределены нормально
4) Residuals имееют одинаковую дисперсию
5) Residuals независимы (= наблюдения независимы)

Нормальность остатков



Example: Influential Points

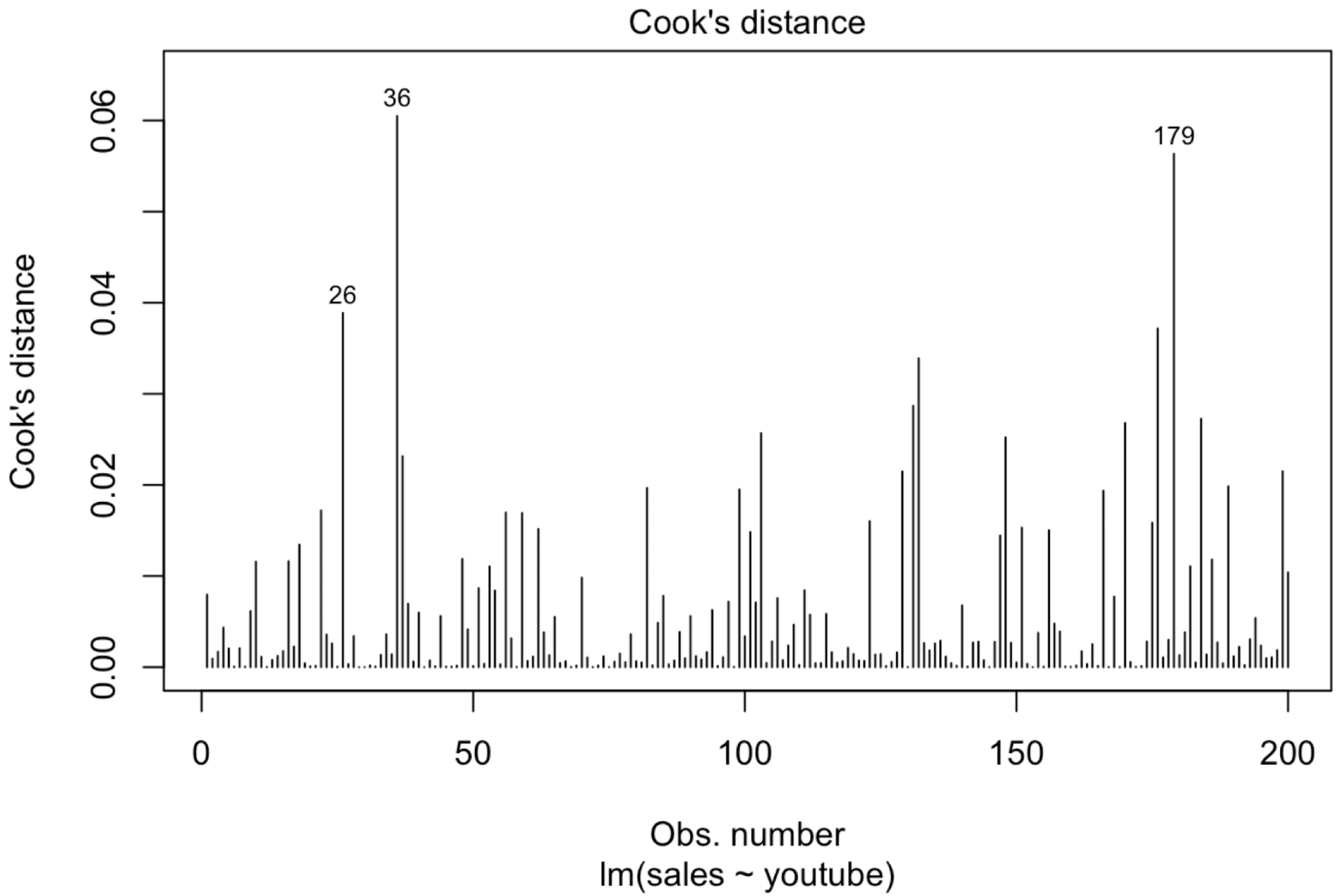


Line without	Regression coeffs.	r_i	h_{ii}	CD_i	D_i^2
Point 2	111.51, -1.18	-.753	.42	.508	10.604
Point 1	118.64, -1.76	-2.00	.42	3.583	10.604
Points 1 & 2	107.92, -.89				

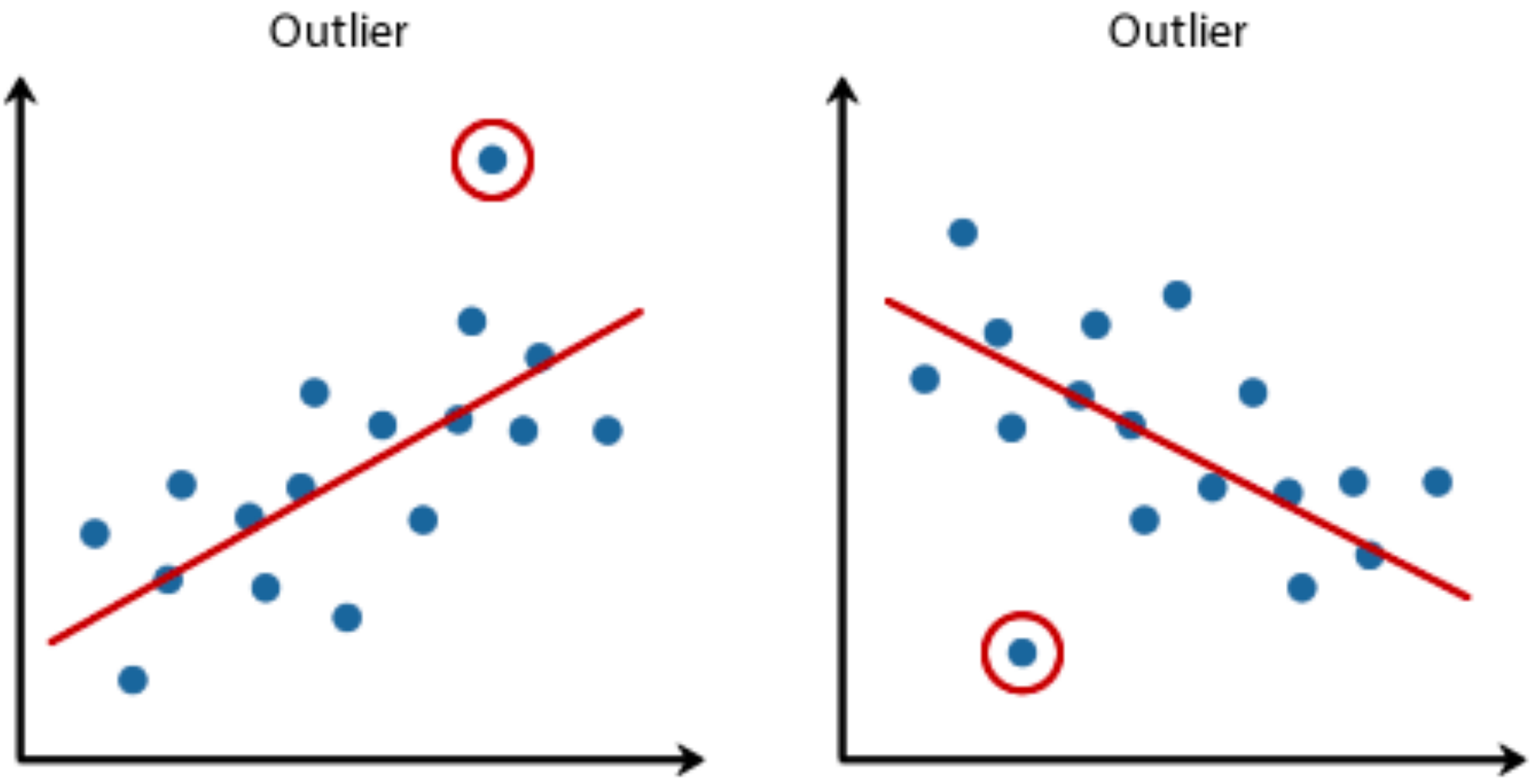
Figure 1. Regression lines and diagnostics for Mickey, Dunn, & Clark (1967) data. (A variation (42.30) on data point 1 has been added.)

Influential points

```
plot(model, 4)
```

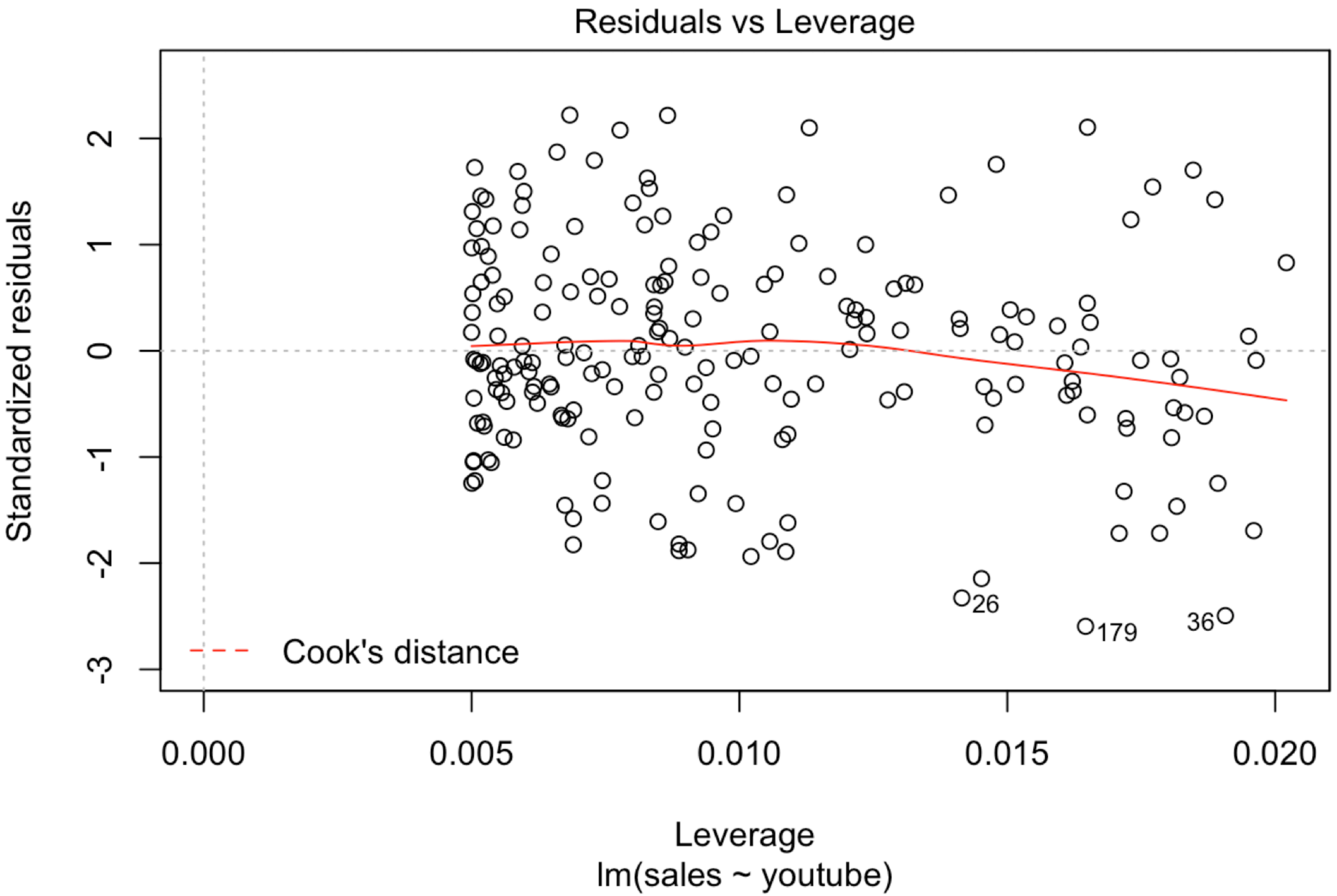


Outliers

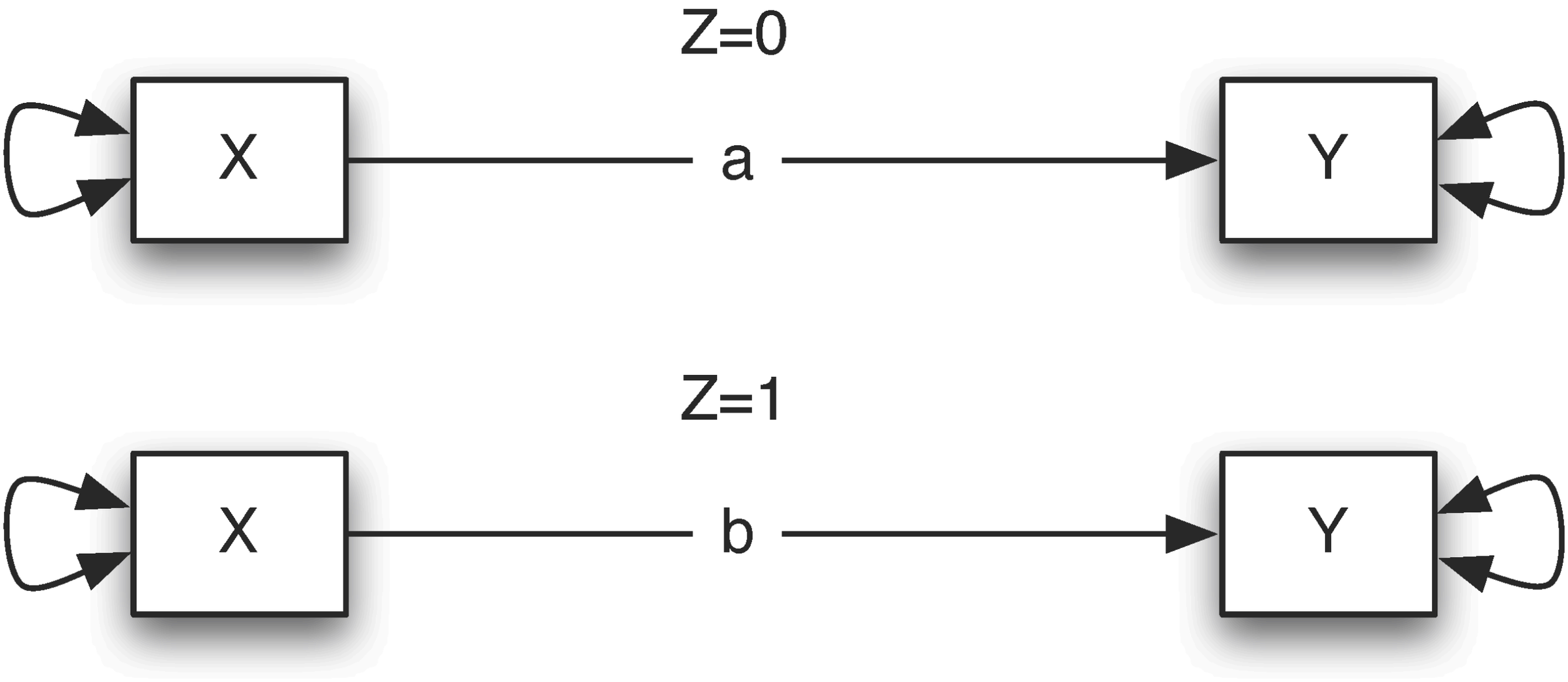


```
plot(model, 5)
```

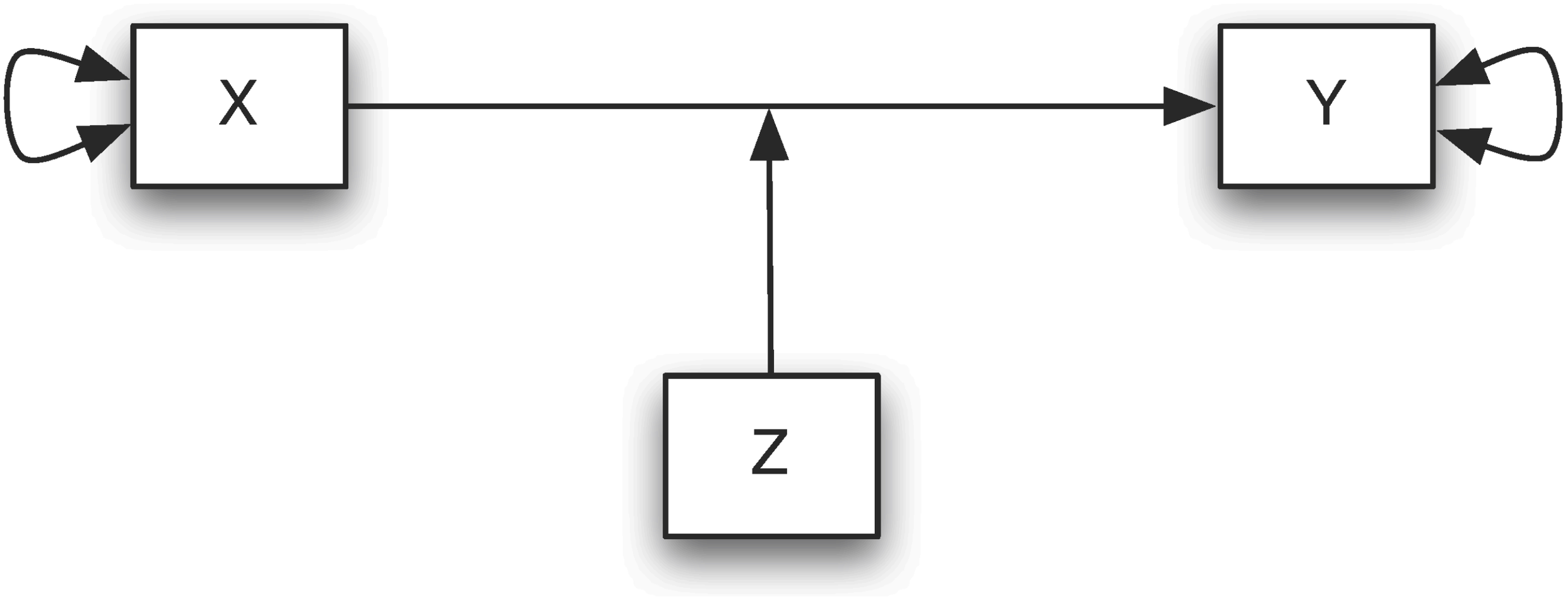
Если отклонение больше 3, то имеем основание подозревать, что точка - outlier



Moderation effect



Moderation effect



Как выглядит линейная модель

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot Z + \beta_3 \cdot X \cdot Z + \epsilon$$

Вклад X **Вклад Z** **moderation effect Z**

Как выглядит линейная модель

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot Z + \beta_3 \cdot X \cdot Z + \epsilon$$

Вклад X **Вклад Z** **moderation effect Z**

Если Z либо 0, либо 1, то как выглядит?

Как выглядит линейная модель

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot Z + \beta_3 \cdot X \cdot Z + \epsilon$$

Вклад X **Вклад Z** **moderation effect Z**

Если Z либо 0, либо 1, то как выглядит?

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon, \;\; Z = 0$$

$$Y = \beta_0 + \beta_2 + (\beta_1 + \beta_3) \cdot X + \epsilon, \;\; Z = 1$$

```r
library(psych)
```

```r
example <- lm(bdi ~ stateanx*epiNeur, data=epi.bfi)
example
```

```
##
## Call:
## lm(formula = bdi ~ stateanx * epiNeur, data = epi.bfi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0493  -2.2513  -0.4707   2.1135  11.9949
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.06367    2.18559   0.029   0.9768
## stateanx          0.03750    0.06062   0.619   0.5368
## epiNeur          -0.14765    0.18869  -0.782   0.4347
## stateanx:epiNeur  0.01528    0.00466   3.279   0.0012 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.12 on 227 degrees of freedom
## Multiple R-squared:  0.4978, Adjusted R-squared:  0.4912
## F-statistic: 75.02 on 3 and 227 DF,  p-value: < 2.2e-16
```