Математические пакеты

Методы обработки данных средней сложности. Регрессия

Сучков Андрей Игоревич

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»

21 ноября 2020 г.

Дополнение к предыдущей лекции

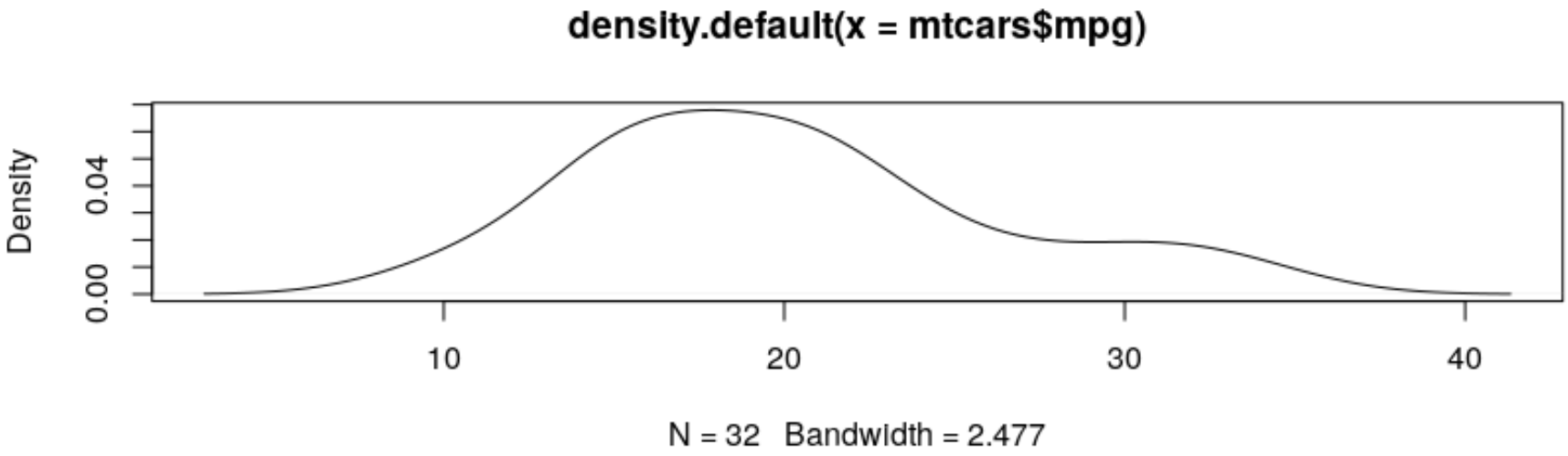
Диаграммы ядерной оценки функции плотности

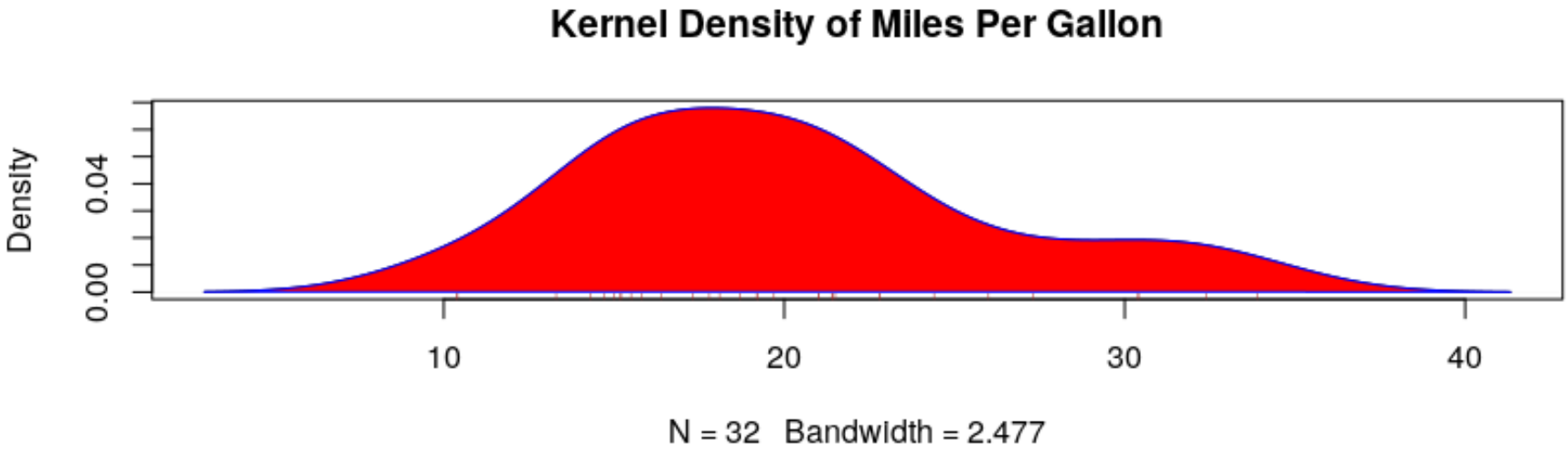
Листинг 1: Пример построения диаграммы ядерной оценки функции плотности

```
par(mfrow=c(2,1))
d <- density(mtcars$mpg)
plot(d)
d <- density(mtcars$mpg)
plot(d, main="Kernel Density of Miles Per Gallon")
polygon(d, col="red", border="blue")
rug(mtcars$mpg, col="brown")
```

Дополнение к предыдущей лекции

Результат построения диаграммы





Дополнение к предыдущей лекции

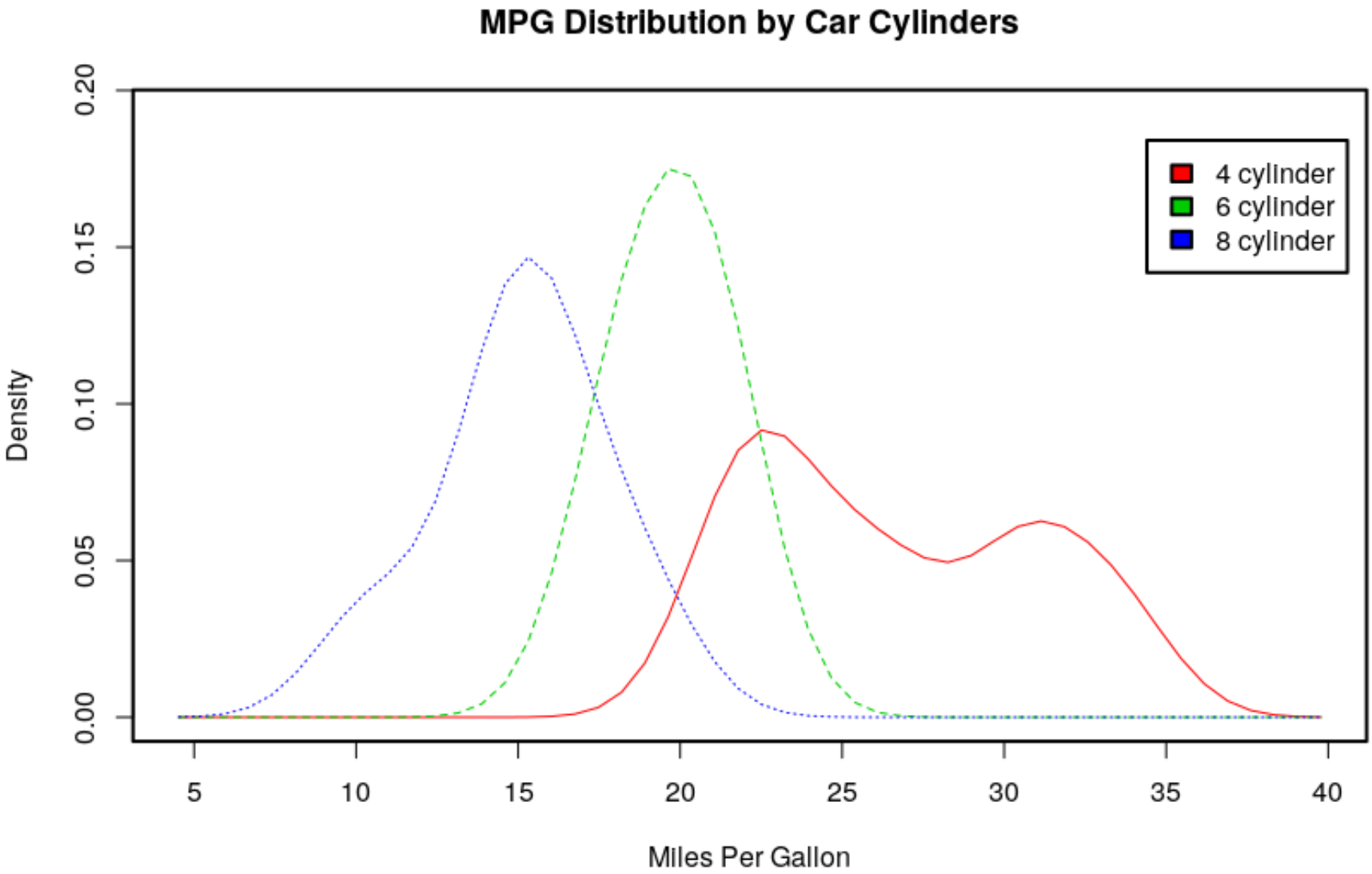
Диаграммы ядерной оценки функции плотности

Листинг 2: Сравнение диаграмм ядерной оценки функции плотности

```
par(lwd=2)
library(sm)
attach(mtcars)
  cyl.f <- factor(cyl, levels = c(4,6,8),
                  labels = c("4 cylinder",
                             "6 cylinder",
                             "8 cylinder"))

  sm.density.compare(mpg, cyl,
                     xlab="Miles Per Gallon")
  title(main="MPG Distribution by Car Cylinders")

  colfill <- c(2:(1+length(levels(cyl.f))))
  legend(locator(1), levels(cyl.f), fill=colfill)
detach(mtcars)
```

Дополнение к предыдущей лекции

Результат построения диаграммы

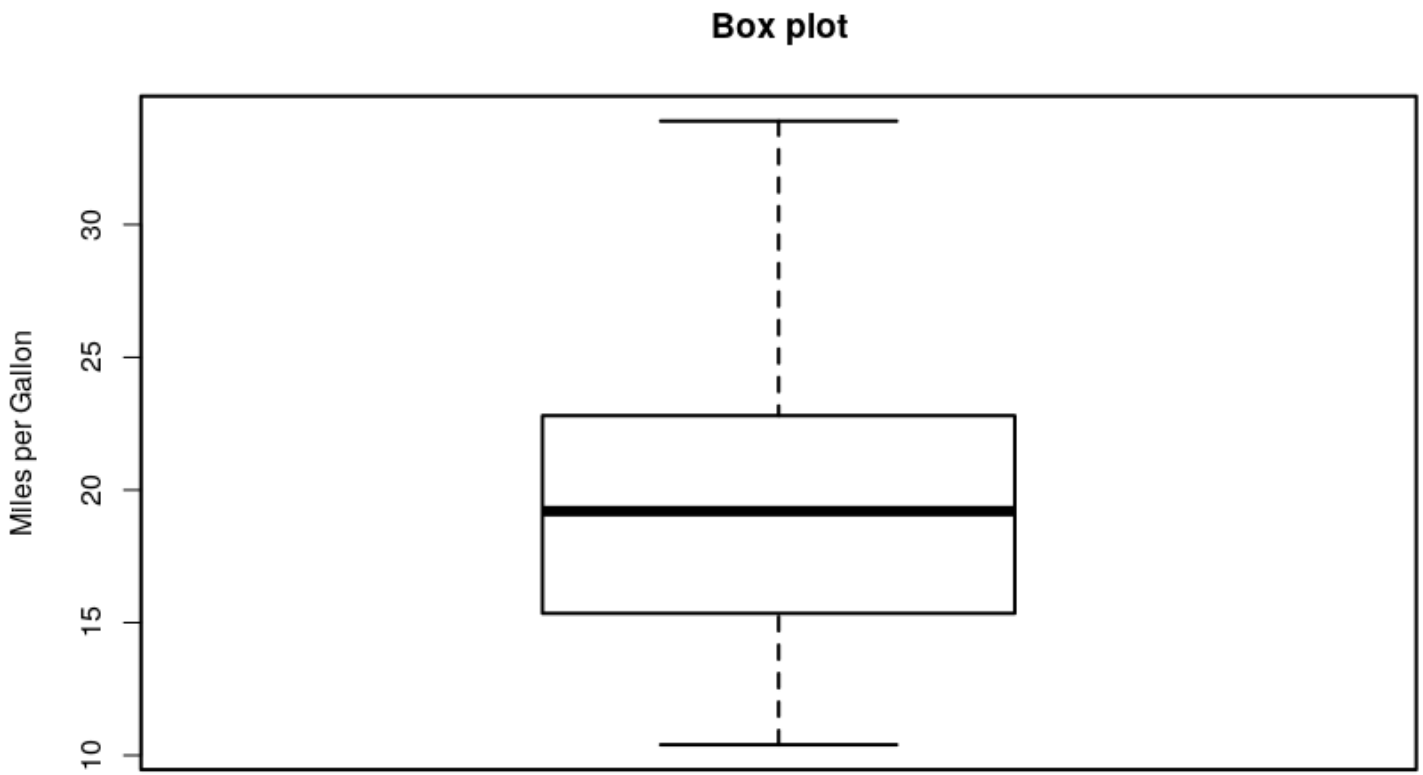


Дополнение к предыдущей лекции

Диаграммы размахов («ящик с усами»)

Построение «ящика с усами» в R

```
boxplot(mtcars$mpg, main="Box plot", ylab="Miles per
Gallon")
```



Листинг 3: Сравнение диаграмм размахов

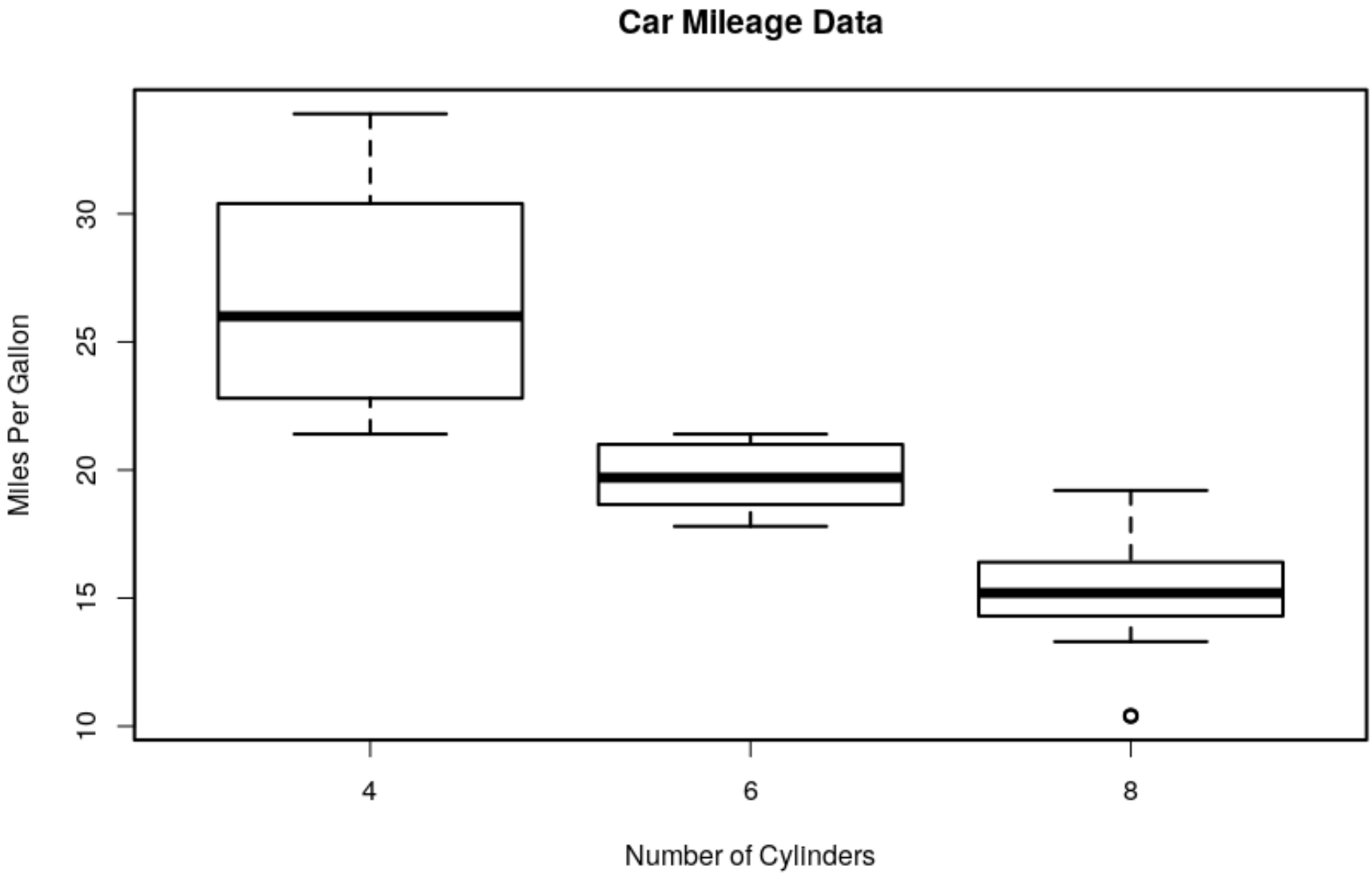
```
boxplot(mpg ~ cyl, data=mtcars,
        main="Car Mileage Data",
        xlab="Number of Cylinders",
        ylab="Miles Per Gallon")
```

Листинг 4: Построение диаграмм размахов с «насечками»

```
boxplot(mpg ~ cyl, data=mtcars,
        notch=TRUE,
        varwidth=TRUE,
        col="red",
        main="Car Mileage Data",
        xlab="Number of Cylinders",
        ylab="Miles Per Gallon")
```

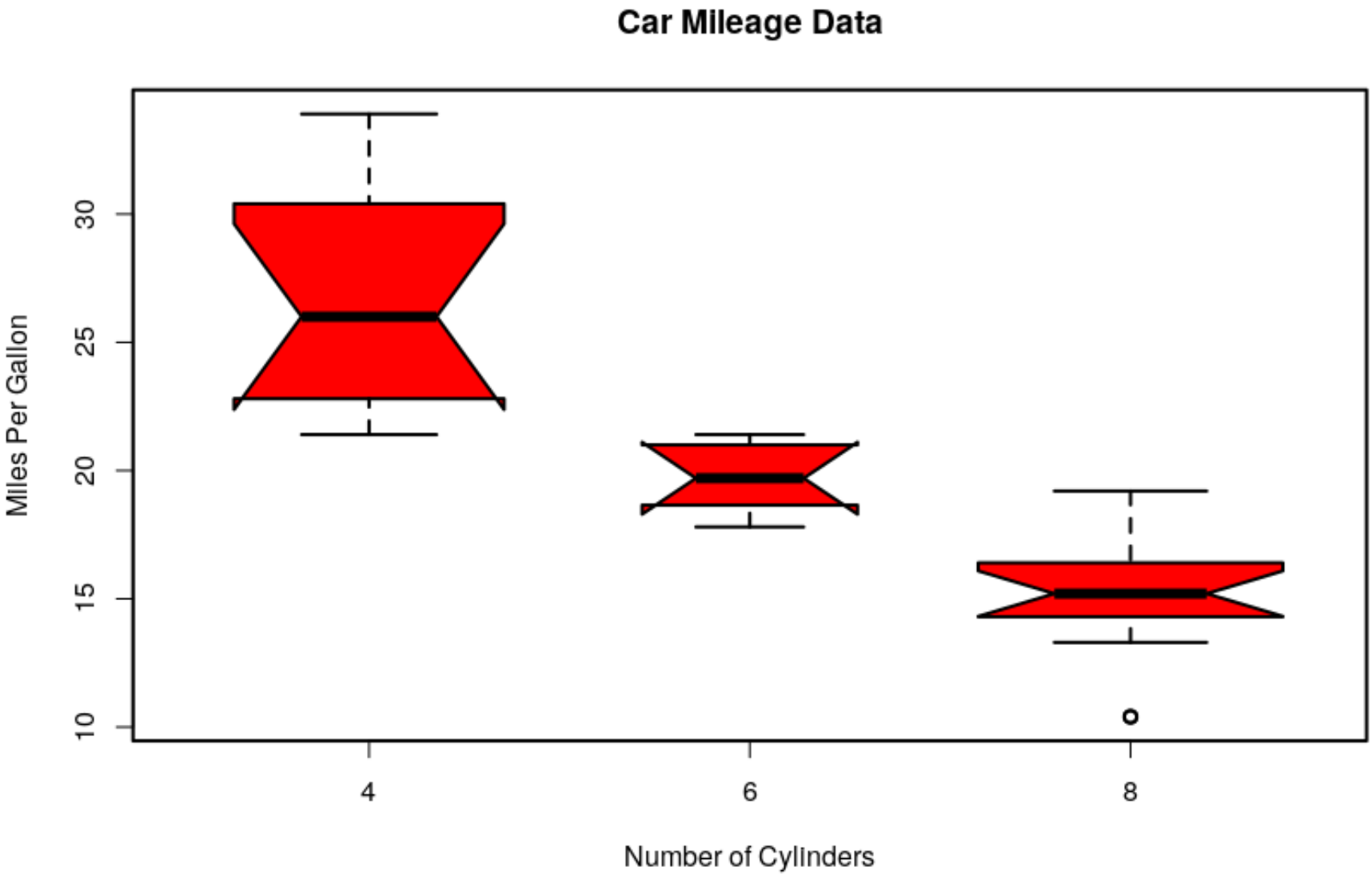
Дополнение к предыдущей лекции

Результат построения диаграммы



Дополнение к предыдущей лекции

Результат построения диаграммы с «насечками»

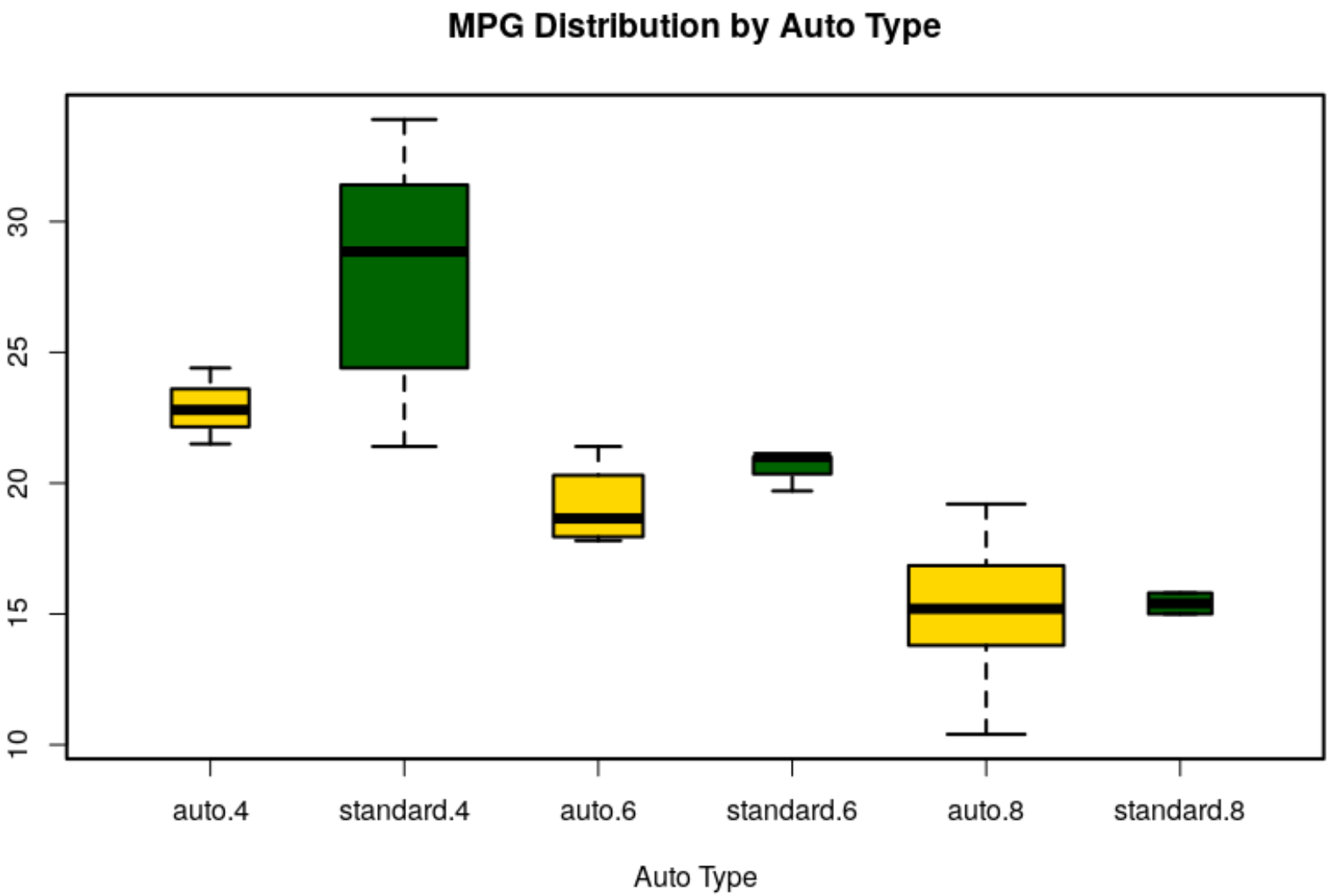


Листинг 5: Диаграмма размахов для всех сочетаний значений двух факторов

```
mtcars$cyl.f <- factor(mtcars$cyl,
                       levels=c(4,6,8),
                       labels=c("4","6","8"))

mtcars$am.f <- factor(mtcars$am,
                      levels=c(0,1),
                      labels=c("auto", "standard"))

boxplot(mpg ~ am.f *cyl.f, data=mtcars,
        varwidth=TRUE,
        col=c("gold","darkgreen"),
        main="MPG Distribution by Auto Type",
        xlab="Auto Type")
```

Дополнение к предыдущей лекции

Диаграмма размахов, отражающая расход топлива для всех комбинаций числа цилиндров и типа коробки передач

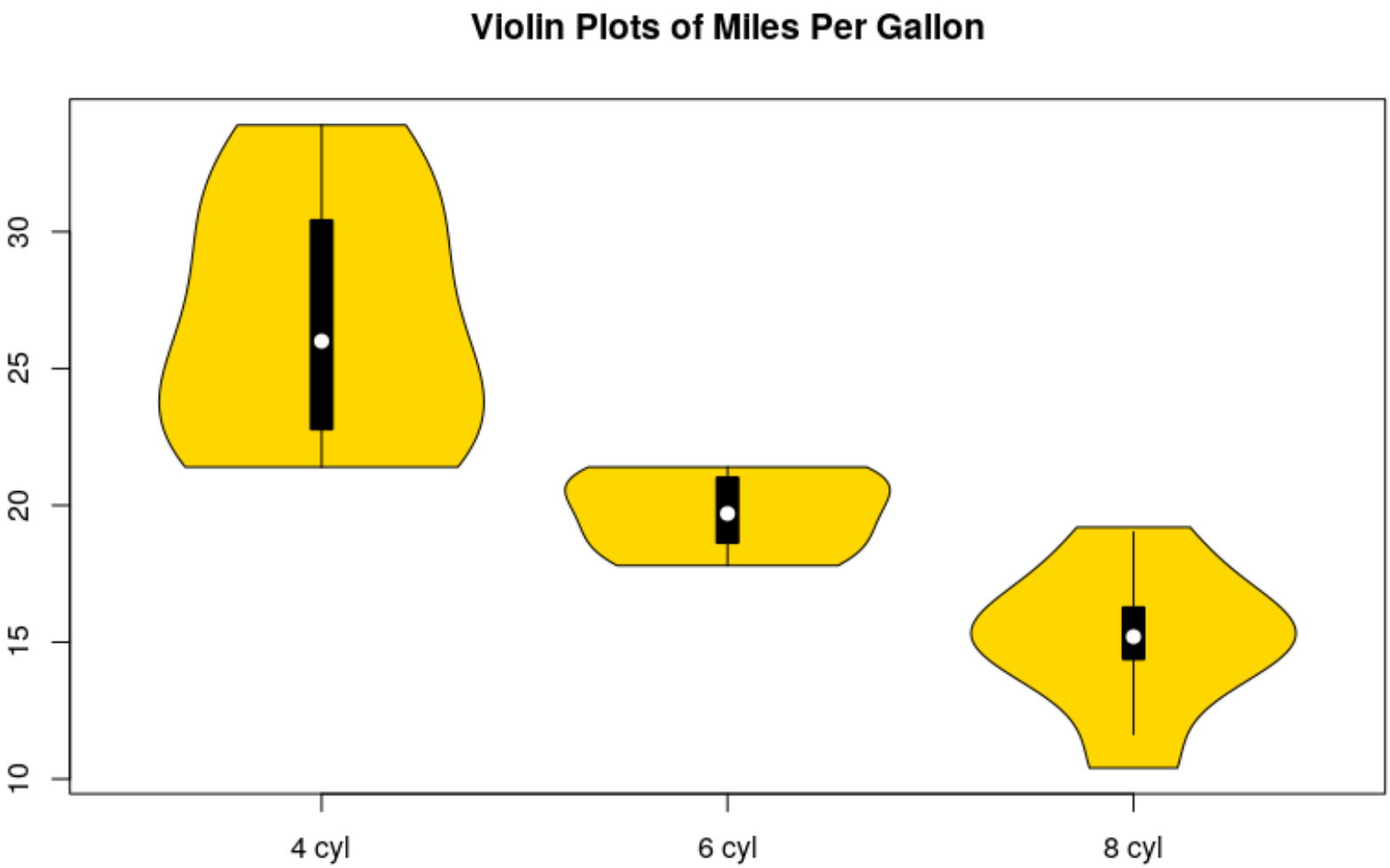


Листинг 6: Построение скрипичных диаграмм

```
library(vioplot)
x1 <- mtcars$mpg[mtcars$cyl==4]
x2 <- mtcars$mpg[mtcars$cyl==6]
x3 <- mtcars$mpg[mtcars$cyl==8]
vioplot(x1, x2, x3,
        names=c("4 cyl", "6 cyl", "8 cyl"),
        col="gold")
title("Violin Plots of Miles Per Gallon")
```

Дополнение к предыдущей лекции

Скрипичные диаграммы, отражающие расход топлива у автомобилей с разным числом цилиндров

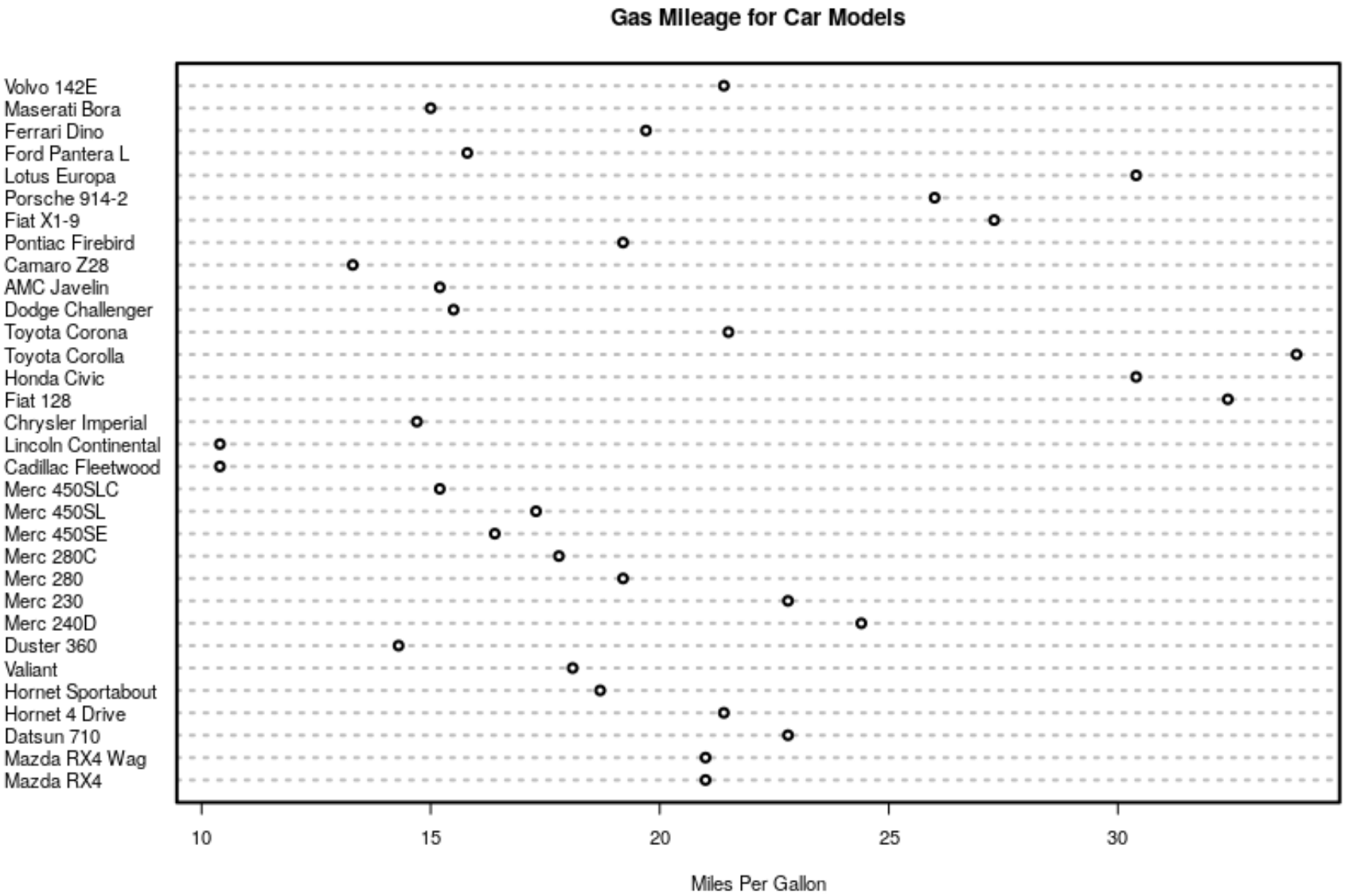


Листинг 7: Построение точечных диаграмм

```
with(mtcars, {
  dotchart(mpg, labels=row.names(mtcars),
           cex=.7,
           main="Gas Mileage for Car Models",
           xlab="Miles Per Gallon")})
```

Дополнение к предыдущей лекции

Точечная диаграмма для расхода топлива у автомобилей разных марок



Дополнение к предыдущей лекции

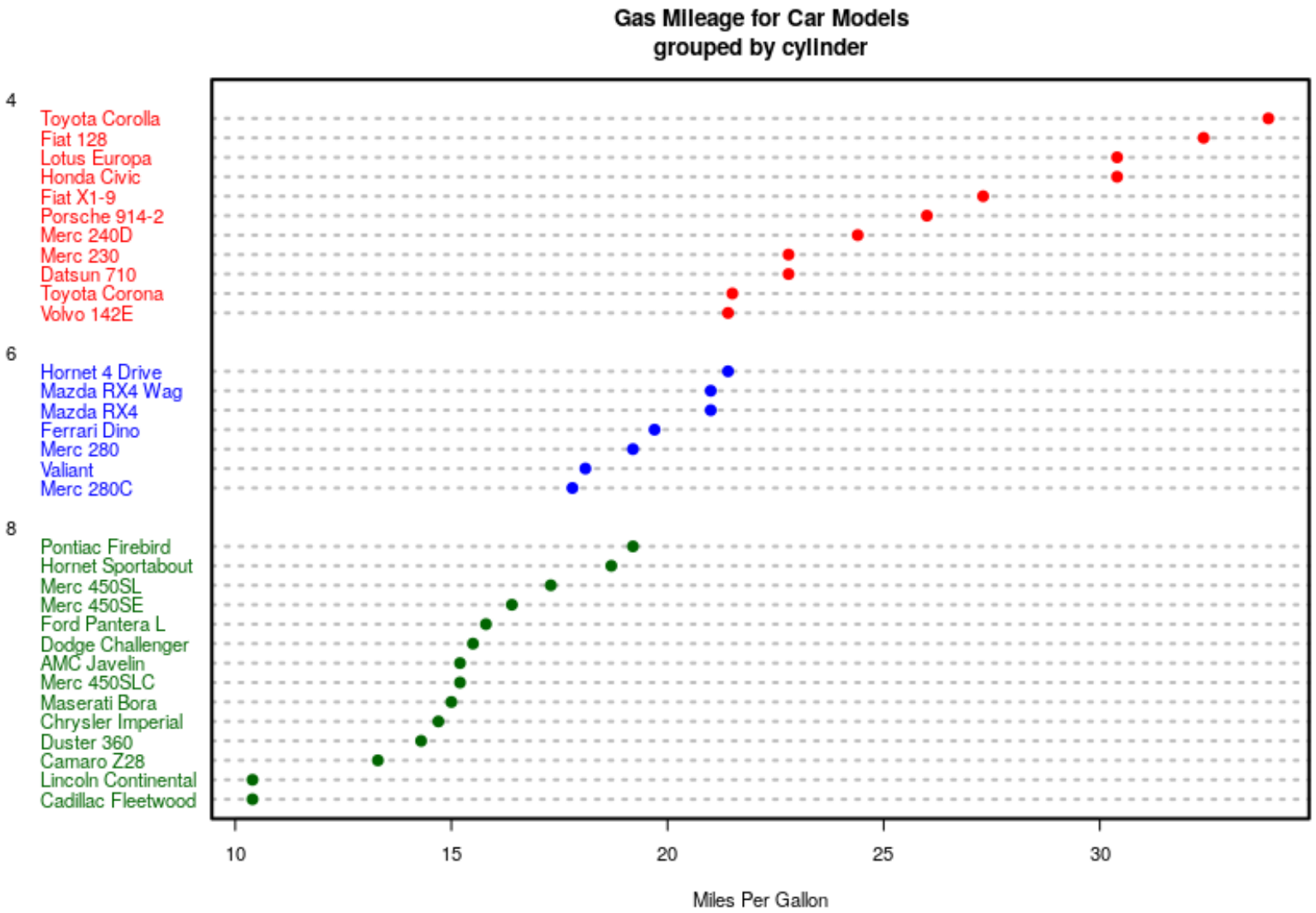
Точечная диаграмма с отсортированными, сгруппированными и раскрашенными значениями

Листинг 8: Построение точечных диаграмм

```
x <- mtcars[order(mtcars$mpg),]
x$cyl <- factor(x$cyl)
x$color[x$cyl==4] <- "red"
x$color[x$cyl==6] <- "blue"
x$color[x$cyl==8] <- "darkgreen"
dotchart(x$mpg,
         labels = row.names(x),
         cex=.7,
         groups = x$cyl,
         gcolor = "black",
         color = x$color,
         pch=19,
         main = "Gas Mileage for Car Models\ngrouped by
             cylinder",
         xlab = "Miles Per Gallon")
```

Дополнение к предыдущей лекции

Точечная диаграмма для расхода топлива у автомобилей, сгруппированных по числу цилиндров



Вид модели МНК-регрессии

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki} \quad i = 1..n$$

- $\hat{Y}_i$ – предсказанное значение зависимой переменной для i-го наблюдения (а именно оценка среднего значения распределения Y по набору независимых переменных)
- X_{ki} – значение k-ой независимой переменной для i-го наблюдения
- $\hat{\beta}_0$ – свободный член уравнения (предсказанное значение Y при нулевом значении всех независимых переменных)
- $\hat{\beta}_k$ – регрессионный коэффициент для k-ой независимой переменной (угол наклона для прямой, которая отражает изменение Y при изменении X на одну единицу измерения)

Цель

Выбрать такие параметры модели (свободный член и регрессионные коэффициенты), которые позволят минимизировать различия между реальными и предсказанными значениями зависимой переменной. То есть мы выбираем такие параметры модели, чтобы сумма квадратов остатков была минимальной:

$$\sum_{1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \cdots - \hat{\beta}_k X_{ki})^2 = \sum_{1}^{n}\varepsilon^2$$

Требования:

- нормальность
- независимость
- линейность
- гомоскедастичность

Формат функции `lm()`

```
myfit <- lm(formula, data)
```

Формула обычно записывается в таком виде:

```
Y ~ X1 + X2 + ... + Xk
```

Символы, которые часто используются в формулах R:

- `~` – Отделяет зависимые переменные (слева) от независимых (справа)
- `+` – Разделяет независимые переменные
- `:` – Обозначает взаимодействие между независимыми переменными
- `*` – Краткое обозначение для всех возможных взаимодействий
- `^` – Обозначает взаимодействия до определенного порядка

Регрессионный анализ

Подгонка регрессионных моделей при помощи команды `lm()`

Символы, которые часто используются в формулах R:

- `.` – Символ-заполнитель для всех переменных в таблице данных, кроме зависимой
- `-` – Знак минуса удаляет переменную из уравнения
- `-1` – Подавляет свободный член уравнения
- `I()` – Элемент в скобках интерпретируется как арифметическое выражение
- `function` – В формулах можно использовать математические функции

Листинг 9: Простая линейная регрессия

```
fit <- lm(weight ~ height, data=women)
summary(fit)

women$weight

fitted(fit)

residuals(fit)

plot(women$height, women$weight,
     xlab="Height (in inches)",
     ylab="Weight (in pounds)")
abline(fit)
```

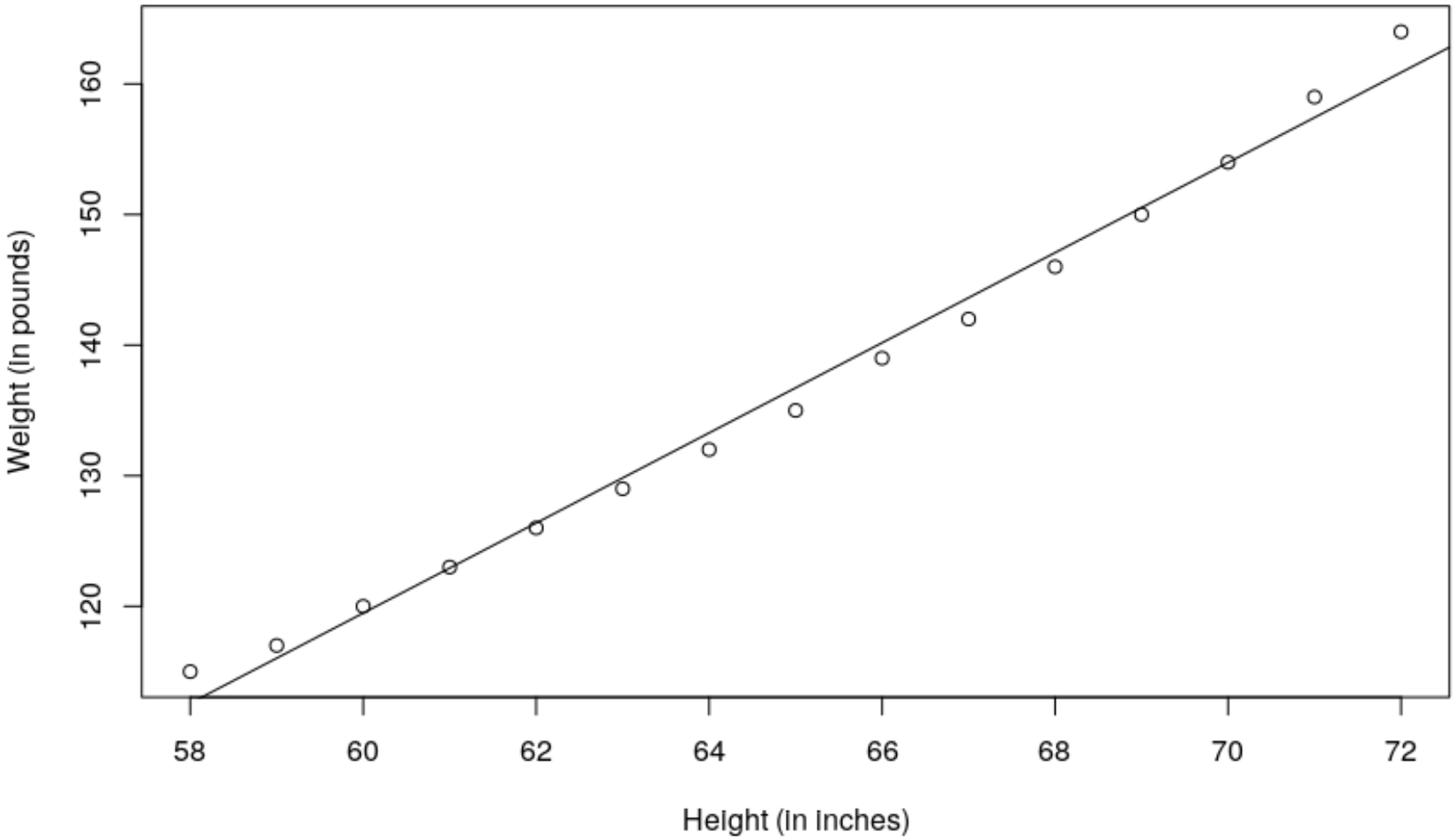
Регрессионный анализ

Результат выполнения команды `summary(fit)`

```
Call:
lm(formula = weight ~ height, data = women)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7333 -1.1333 -0.3833  0.7417  3.1167

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
height        3.45000    0.09114   37.85 1.09e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.525 on 13 degrees of freedom
Multiple R-squared:  0.991,     Adjusted R-squared:  0.9903
F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

Регрессионный анализ

Диаграмма рассеяния с регрессионной прямой для значений веса, предсказанных по значениям роста



Листинг 10: Полиномиальная регрессия

```
fit2 <- lm(weight ~ height + I(height^2), data=women)
summary(fit2)

plot(women$height,women$weight,
     xlab="Height (in inches)",
     ylab="Weight (in lbs)")
lines(women$height, fitted(fit2))
```

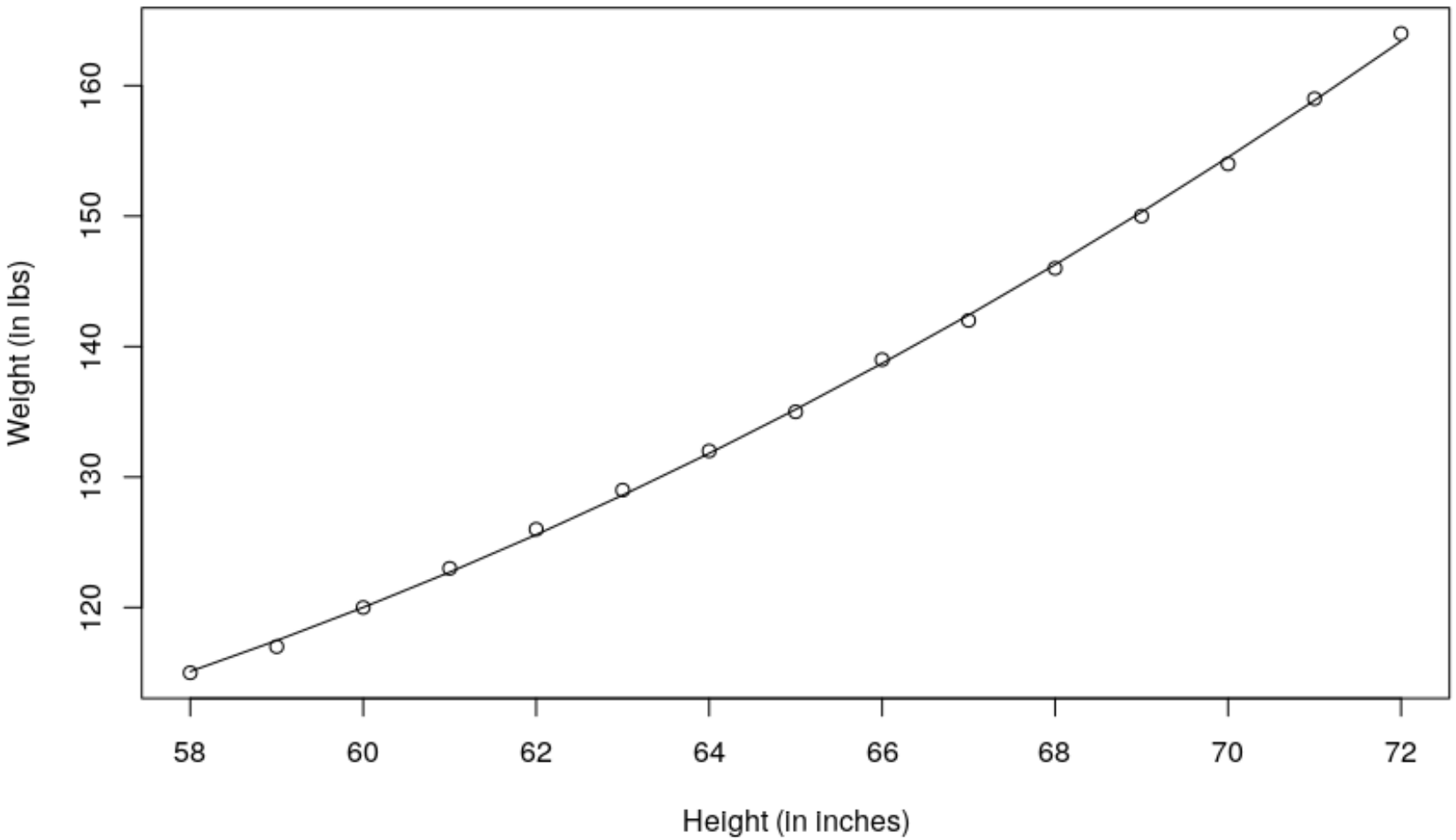
Регрессионный анализ

Результат выполнения команды `summary(fit2)`

```
Call:
lm(formula = weight ~ height + I(height^2), data = women)

Residuals:
     Min       1Q   Median       3Q      Max
-0.50941 -0.29611 -0.00941  0.28615  0.59706

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 261.87818   25.19677  10.393 2.36e-07 ***
height       -7.34832    0.77769  -9.449 6.58e-07 ***
I(height^2)   0.08306    0.00598  13.891 9.32e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3841 on 12 degrees of freedom
Multiple R-squared:  0.9995,    Adjusted R-squared:  0.9994
F-statistic: 1.139e+04 on 2 and 12 DF,  p-value: < 2.2e-16
```

Регрессионный анализ

Квадратичная регрессия для предсказаний значений веса по значениям роста

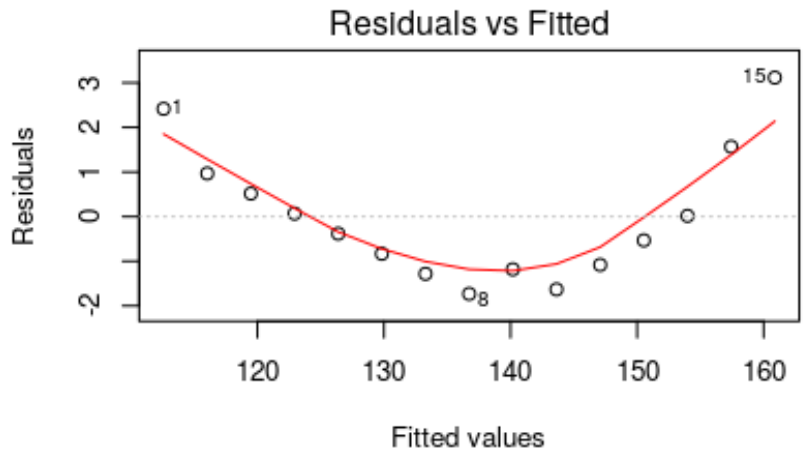


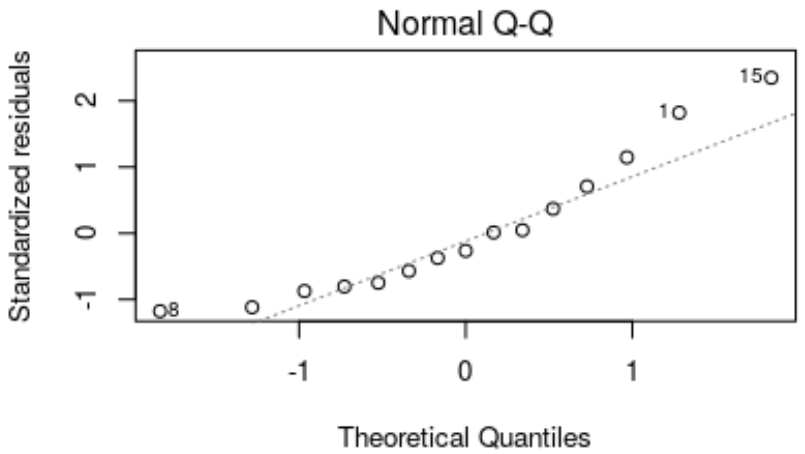
Листинг 11: Полиномиальная регрессия

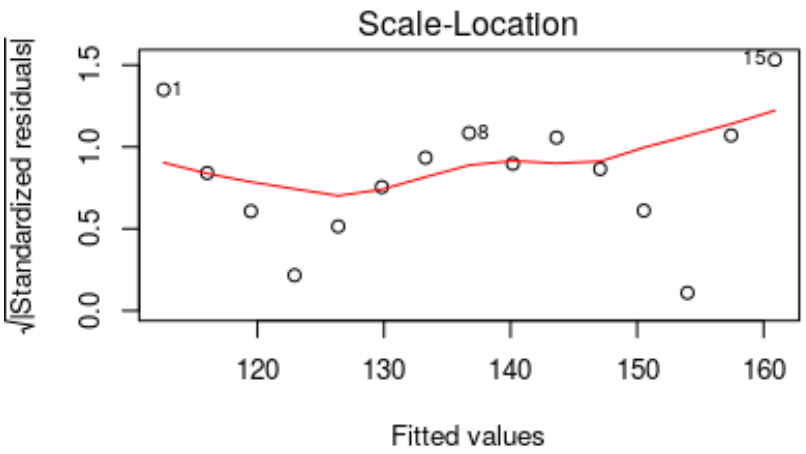
```
fit <- lm(weight ~ height, data=women)
par(mfrow=c(2,2))
plot(fit)

fit2 <- lm(weight ~ height + I(height^2), data=women)
par(mfrow=c(2,2))
plot(fit2)

states <- as.data.frame(state.x77[,c("Murder", "Population",
                                     "Illiteracy", "Income",
                                     "Frost")])
fit3<-lm(Murder ~ Population + Illiteracy + Income + Frost,
         data=states)
par(mfrow=c(2,2))
plot(fit3)
```

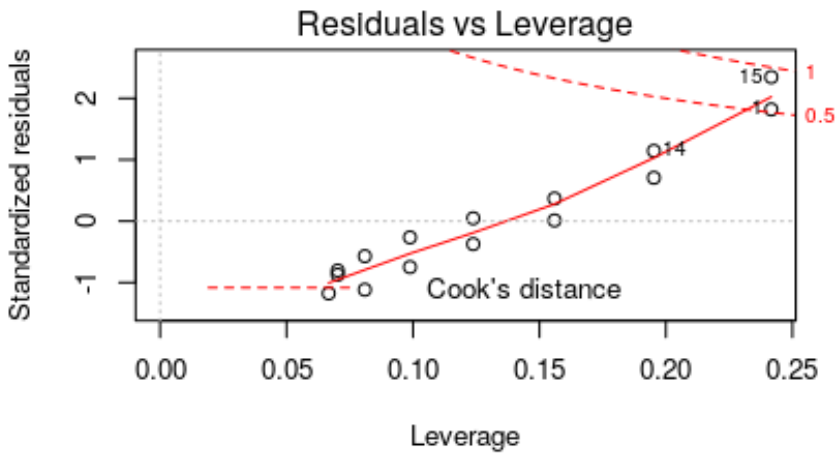
Диагностика регрессионных моделей

Диагностические диаграммы для регрессии веса по росту



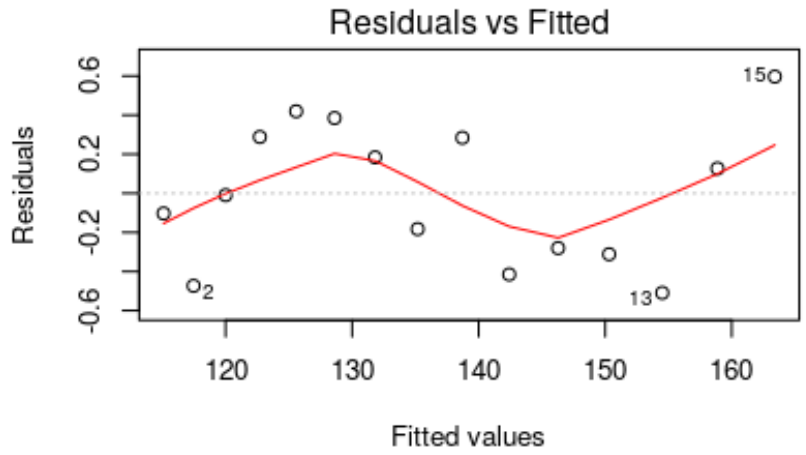


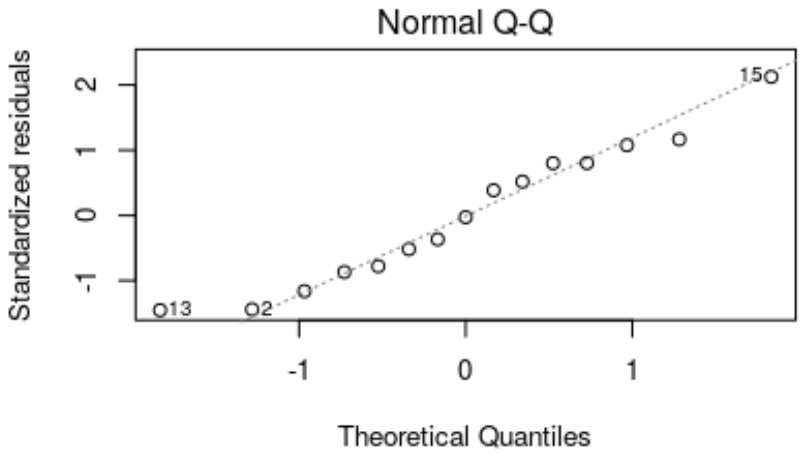


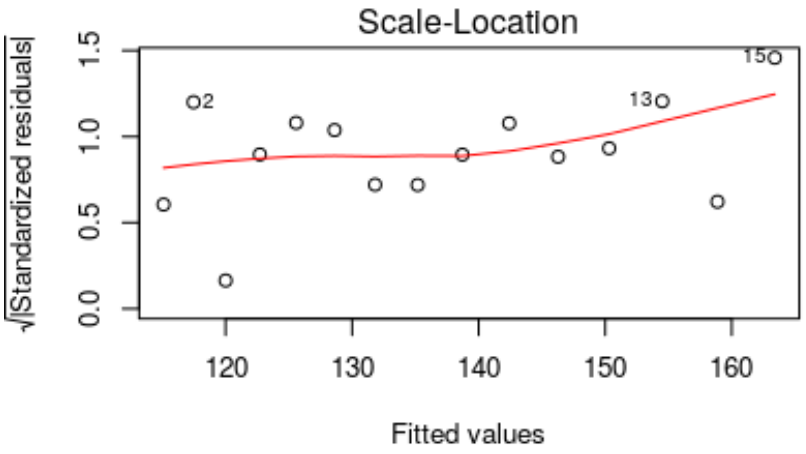


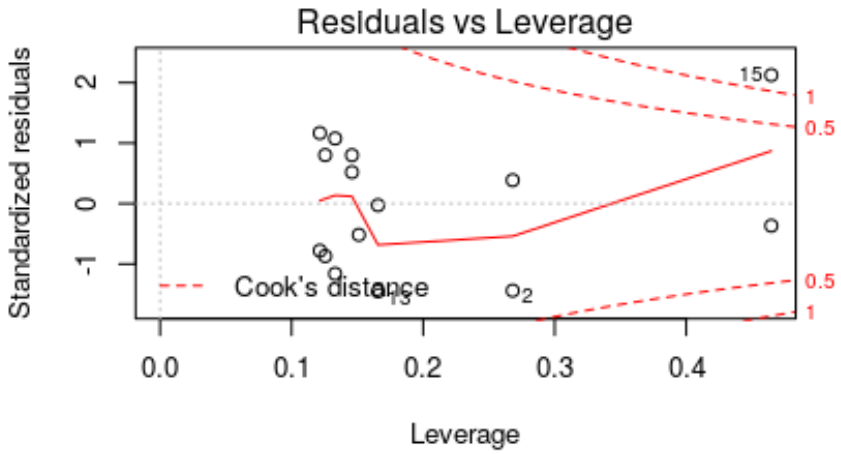
Диагностика регрессионных моделей

Диагностические диаграммы для регрессии веса по росту и по росту, возведенному в квадрат









Диагностика регрессионных моделей

Диагностические диаграммы для регрессии уровня преступности по характеристикам штатов