RUDN university

5100

Программирование в среде R

Шевцов Василий Викторович,
директор ДИТ РУДН, shevtsov_vv@rudn.university

Регрессия

Определение

- Регрессионный анализ — статистический метод исследования влияния одной или нескольких независимых переменных X_1, X_2,..., X_n на зависимую переменную Y.
- Независимые переменные иначе называют регрессорами или предикторами, а зависимые переменные — критериальными.
- Терминология зависимых и независимых переменных отражает лишь математическую зависимость переменных, а не причинно-следственные отношения.

Разновидности регрессионного анализа

Тип регрессии	Для чего обычно используется
Простая линейная	Предсказание значений количественной зависимой переменной по значениям одной количественной независимой переменной
Полиномиальная	Предсказание значений количественной зависимой переменной по значениям количественной независимой переменной, когда взаимосвязь моделируется как полином n-ой степени
Множественная линейная	Предсказание значений количественной зависимой переменной по значениям двух и более количественных независимых переменных
Многомерная	Предсказание значений более чем одной зависимой переменной по значениям одной и более независимых переменных

Разновидности регрессионного анализа

Тип регрессии	Для чего обычно используется
Логистическая	Предсказание значений категориальной зависимой переменной по значениям одной и более независимых переменных
Пуассона	Предсказание значений зависимой счетной переменной по значениям одной или более независимых переменных
Пропорциональных рисков Кокса	Предсказание времени до наступления события (смерти, аварии, рецидива) по значениям одной или более независимых переменных
Временных рядов	Моделирование временных рядов с коррелированными ошибками

Разновидности регрессионного анализа

Тип регрессии	Для чего обычно используется
Нелинейная	Предсказание значений количественной зависимой переменной по значениям одной и более независимых переменных с использованием нелинейной модели
Непараметрическая	Предсказание значений количественной зависимой переменной по значениям одной и более независимых переменных с использованием полученной из данных и незаданной заранее модели
Устойчивая	Предсказание значений количественной зависимой переменной по значениям одной и более независимых переменных с использованием метода, устойчивого к выбросам

Метод наименьших квадратов (МНК, англ. Ordinary Least Squares, OLS)

- математический метод, применяемый для решения различных задач, основанный на минимизации суммы квадратов отклонений некоторых функций от искомых переменных.
- В случае МНК-регрессии значения количественной зависимой переменной предсказываются на основании взвешенной суммы значений независимых переменных, где веса переменных оцениваются, исходя из данных.

Метод наименьших квадратов

- **МНК-регрессия позволяет подгонять модели вида**

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \ldots + \widehat{\beta}_k X_{ki} \qquad i = 1 \ldots n,$$

n – это число наблюдений,
k – это число независимых переменных.

Метод наименьших квадратов

$\widehat{Y}_i$	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
$\widehat{\beta}_0$	Свободный член уравнения (предсказанное значение Y при нулевом значении всех независимых переменных)
$\widehat{\beta}_k$	Регрессионный коэффициент для k-ой независимой переменной (угол наклона для прямой, которая отражает изменение Y при изменении X на одну единицу измерения)

Метод наименьших квадратов

- цель – это выбрать такие параметры модели (свободный член и регрессионные коэффициенты), которые позволят минимизировать различия между реальными и предсказанными значениями зависимой переменной. То есть выбираются такие параметры модели, чтобы сумма квадратов остатков была минимальной:

$$\sum_{1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 = \sum_{1}^{n}\left(Y_i - \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \ldots + \widehat{\beta}_k X_{ki}\right)^2 = \sum_{1}^{n}\varepsilon^2$$

Метод наименьших квадратов

Для правильной интерпретации коэффициентов МНК-модели нужно, чтобы ваши данные удовлетворяли ряду требований:

- нормальность – значения зависимой переменной нормально распределены при фиксированных значениях независимых переменных;
- независимость – значения Yi независимы друг от друга;
- линейность – зависимая переменная линейно связана с независимыми;
- гомоскедастичность – дисперсия зависимой переменной постоянна при разных значениях независимых переменных (однородность дисперсии)

Подгонка регрессионных моделей при помощи команды lm()

- **lm(formula, data)**
- formula описывает вид модели, которую нужно подогнать,
- data – это таблица с данными, которые используются для создания модели. Полученный объект– это список, содержащий обширную информацию о подогнанной модели. Формула обычно записывается в таком виде: $Y \sim X_1 + X_2 + \ldots + X_k$
- ~ отделяет зависимую переменную слева от независимых переменных (разделенных знаками +) справа. Для различных изменений этой формулы можно использовать другие символы

Символы, которые используются в формулах R

Символ	Назначение
~	Отделяет зависимые переменные (слева) от независимых (справа). Например, предсказание значений y по значениям x, z и w будет закодировано так: y ~ x + z + w
+	Разделяет независимые переменные
:	Обозначает взаимодействие между независимыми переменными. Предсказание значений y по значениям x, z и взаимодействия между x и z будет закодировано как y ~ x + z + x:z
*	Краткое обозначение для всех возможных взаимодействий. Код y ~ x * z * w в полном виде означает y ~ x + z + w + x:z + x:w + z:w + x:z:w
^	Обозначает взаимодействия до определенного порядка. Код y ~ (x + z + w)^2 в полном виде будет записан как y ~ x + z + w + x:z + x:w + z:w

Символы, которые используются в формулах R

Символ	Назначение
.	Символ-заполнитель для всех переменных в таблице данных, кроме зависимой. Например, если таблица данных содержит переменные x, y, z и w, то код y ~ . будет означать y ~ x + z + w
-	Знак минуса удаляет переменную из уравнения. Например, y ~ (x + z + w)^2 – x:w соответствует y ~ x + z + w + x:z + z:w
-1	Подавляет свободный член уравнения. Например, формула y ~ x -1 позволяет подогнать такую регрессионную модель для предсказания значений y по x, чтобы ее график проходил через начало координат
I()	Элемент в скобках интерпретируется как арифметическое выражение. Например, y ~ x + (z + w)^2 означает y ~ x + z + w + z:w. Для сравнения y ~ x + I((z + w)^2) означает y ~ x + h, где h – это новая переменная, полученная при возведении в квадрат суммы z и w
function	В формулах можно использовать математические функции. Например, log(y) ~ x + z + w будет предсказывать значения log(y) по значениям x, z и w

Функции, полезные при подгонке линейных моделей

Функция	**Действие**
summary()	Показывает детальную информацию о подогнанной модели
coefficients()	Перечисляет параметры модели (свободный член и регрессионные коэффициенты)
confint()	Вычисляет доверительные интервалы для параметров модели (по умолчанию 95%)
fitted()	Выводит на экран предсказанные значения, согласно подогнанной модели
residuals()	Показывает остатки для подогнанной модели
anova()	Создает таблицу ANOVA (дисперсионного анализа) для подогнанной модели или таблицу ANOVA, сравнивающую две или более моделей

Функции, полезные при подгонке линейных моделей

Функция	Действие
vcov()	Выводит ковариационную матрицу для параметров модели
AIC()	Вычисляет информационный критерий Акаике (Akaike's Information Criterion)
plot()	Создает диагностические диаграммы для оценки адекватности модели
predict()	Использует подогнанную модель для предсказания зависимой переменной для нового набора данных

Виды рассматриваемых регрессий

простая линейная регрессия	в регрессионной модели есть одна зависимая и одна независимая переменная
полиномиальная регрессия	одна зависимая переменная, но в модель входят ее степени (например, X, X^2, X^3)
множественная регрессия	есть больше одной независимой переменной

Простая линейная регрессия

Простая линейная регрессия

```
> mtcars
                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C         17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Merc 450SE        16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Merc 450SL        17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC       15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
Fiat 128          32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
Honda Civic       30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
Toyota Corolla    33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
Toyota Corona     21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
Dodge Challenger  15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
AMC Javelin       15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
Camaro Z28        13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
Pontiac Firebird  19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
Fiat X1-9         27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
Porsche 914-2     26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
Lotus Europa      30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
Ford Pantera L    15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
Ferrari Dino      19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
Maserati Bora     15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
Volvo 142E        21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
> lm(mpg~hp,mtcars)

Call:
lm(formula = mpg ~ hp, data = mtcars)

Coefficients:
(Intercept)           hp
   30.09886     -0.06823
```

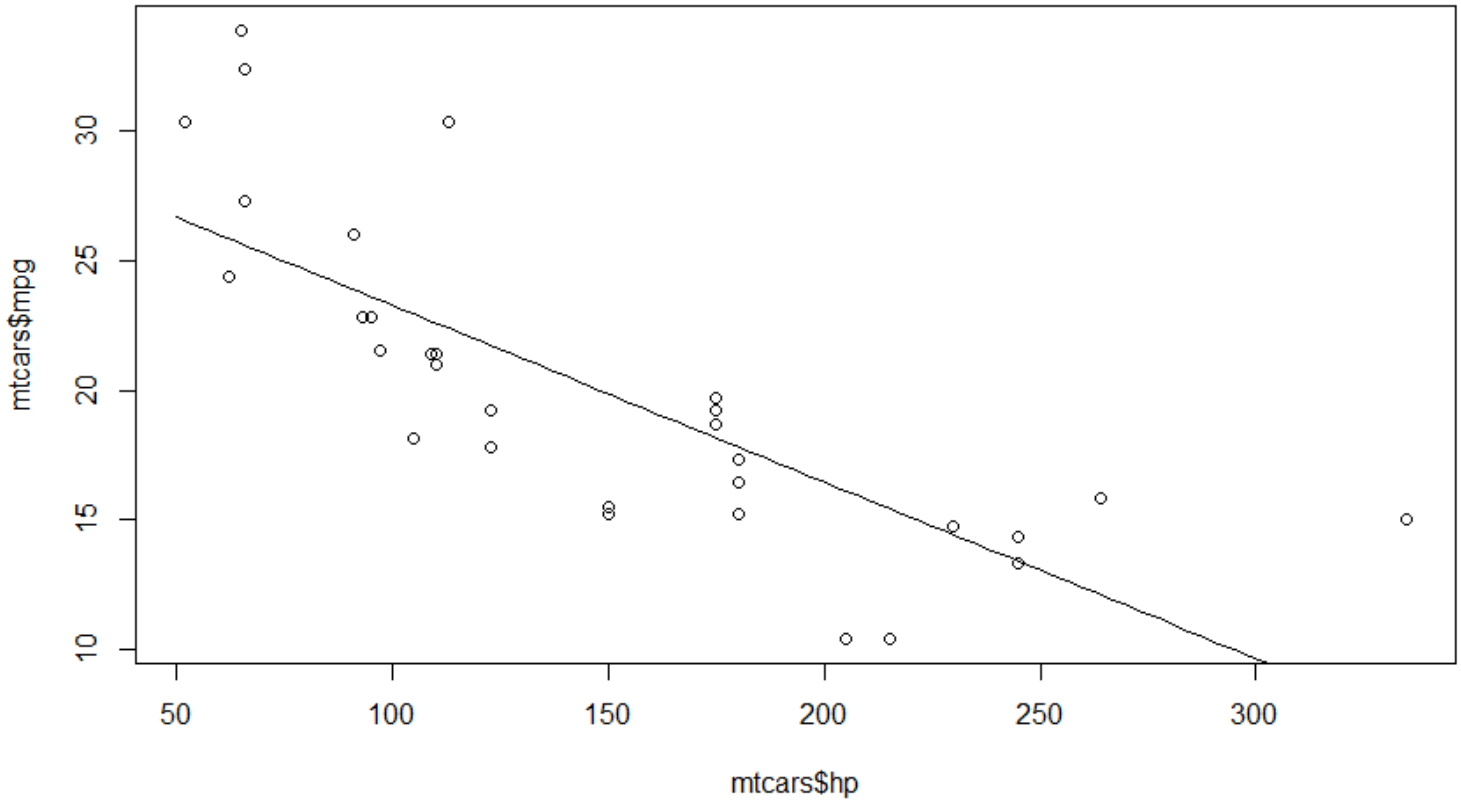
mpg=30.09886-0.06823*hp

Простая линейная регрессия

```
> lm1 <- lm(mpg~hp,mtcars)
> summary(lm1)

Call:
lm(formula = mpg ~ hp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7121 -2.1122 -0.8854  1.5819  8.2360

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
hp          -0.06823    0.01012  -6.742 1.79e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.863 on 30 degrees of freedom
Multiple R-squared:  0.6024,    Adjusted R-squared:  0.5892
F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

Простая линейная регрессия

```
plot(mtcars$hp,mtcars$mpg)
f1 <- function(x){30.09886-0.06823*x}
curve(f1,50,400,add=TRUE)
```



Простая линейная регрессия

```
> women
   height weight
1      58    115
2      59    117
3      60    120
4      61    123
5      62    126
6      63    129
7      64    132
8      65    135
9      66    139
10     67    142
11     68    146
12     69    150
13     70    154
14     71    159
15     72    164
```

Набор данных women, поставляемый с базовой версией программы, содержит данные о росте и весе 15 женщин в возрасте от 30 до 39 лет.

Рост – дюймы
Вес – фунты

Простая линейная регрессия

```
> women_rus <- cbind(women$height * 2.54,women$weight * 0.454)
> colnames(women_rus) <- c("height","weight")
> women_rus <- as.data.frame(women_rus)
> women_rus
   height weight
1  147.32 52.210
2  149.86 53.118
3  152.40 54.480
4  154.94 55.842
5  157.48 57.204
6  160.02 58.566
7  162.56 59.928
8  165.10 61.290
9  167.64 63.106
10 170.18 64.468
11 172.72 66.284
12 175.26 68.100
13 177.80 69.916
14 180.34 72.186
15 182.88 74.456
```

Рост – сантиметры

Вес – килограммы

Простая линейная регрессия

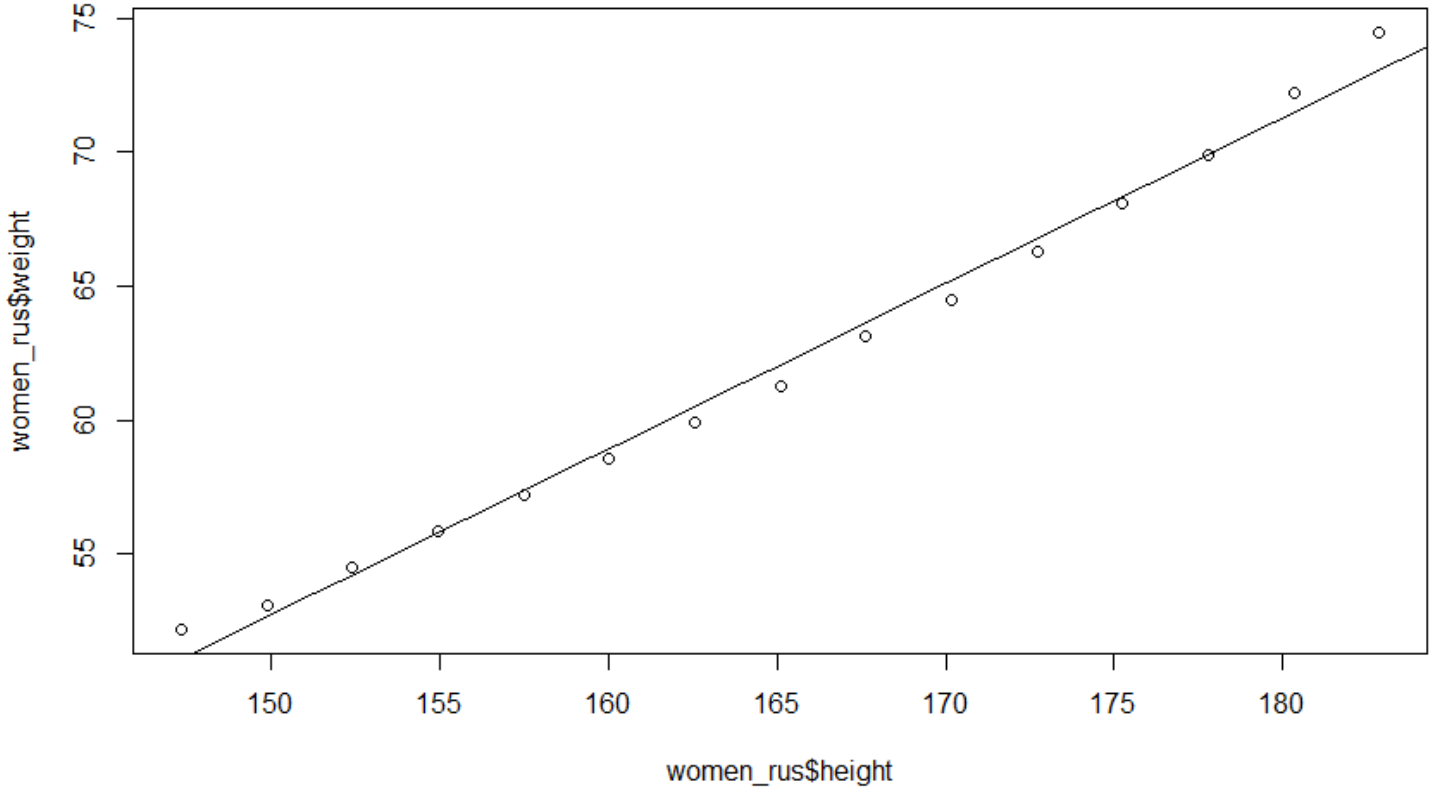
```
> lm2 <- lm(weight ~ height, women_rus)
> summary(lm2)

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

Residuals:
    Min      1Q  Median      3Q     Max
-0.7869 -0.5145 -0.1740  0.3367  1.4150

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -39.73257    2.69537  -14.74 1.71e-09 ***
height        0.61665    0.01629   37.85 1.09e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6924 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
```

RUDN university

weight=-39.73257+0.61665*height

Простая линейная регрессия

```
plot(women_rus$height,women_rus$weight)
f2 <- function(x){-39.73257+0.61665*x}
curve(f2,50,200,add=TRUE)
```



Полиномиальная регрессия

- точность предсказания можно улучшить, если использовать квадратичное регрессионное уравнение

$$\widehat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2$$

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

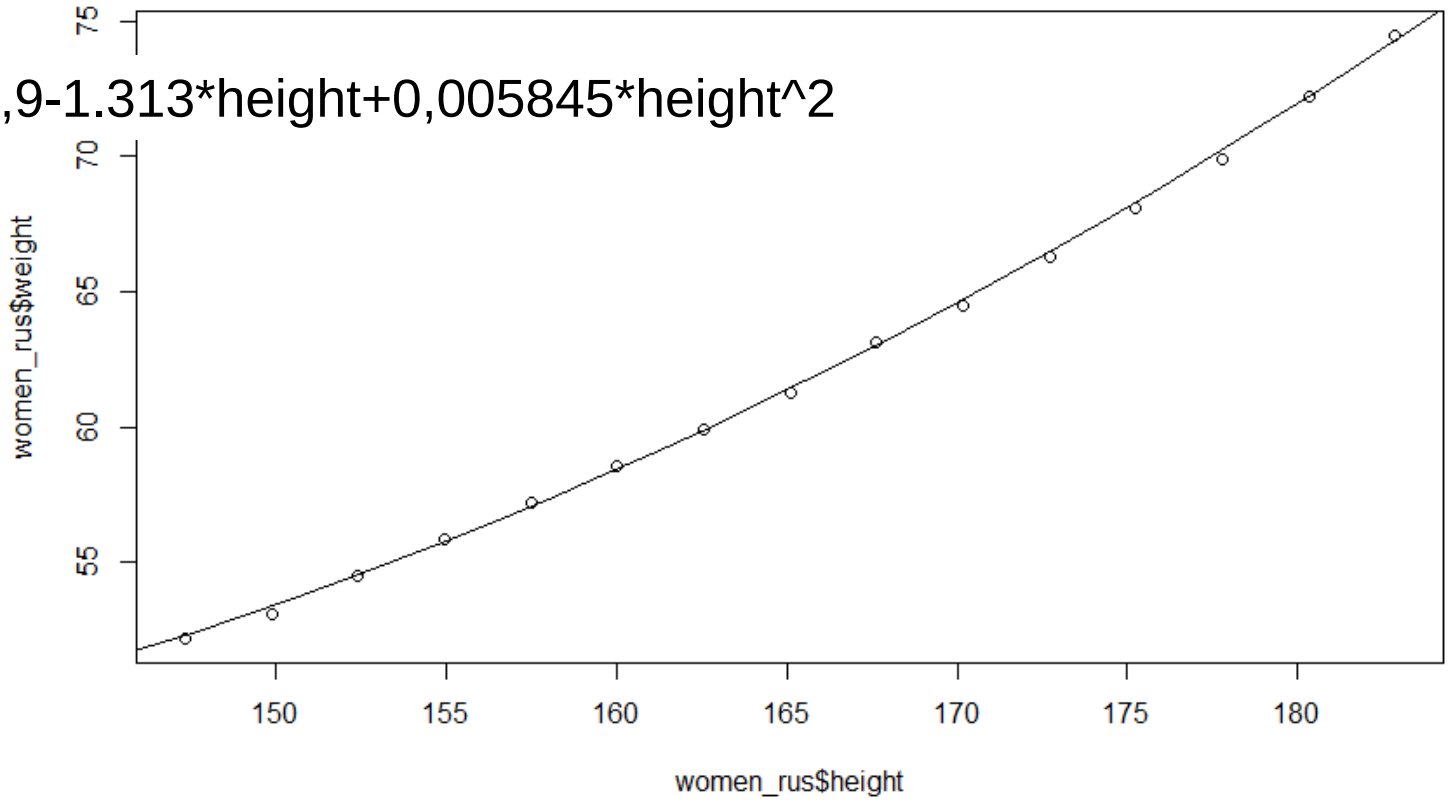
Полиномиальная регрессия

```
> lm3 <- lm(weight ~ height + I(height^2), women_rus)
> summary(lm3)

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

Residuals:
     Min       1Q   Median       3Q      Max
-0.23127 -0.13443 -0.00427  0.12991  0.27107

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.189e+02  1.144e+01  10.393 2.36e-07 ***
height      -1.313e+00  1.390e-01  -9.449 6.58e-07 ***
I(height^2)  5.845e-03  4.208e-04  13.891 9.32e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1744 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
```

Полиномиальная регрессия

```
plot(women_rus$height,women_rus$weight)
f3 <- function(x){1.189e+02-1.313e+00*x+5.845e-03*x^2}
curve(f3,50,200,add=TRUE)
```

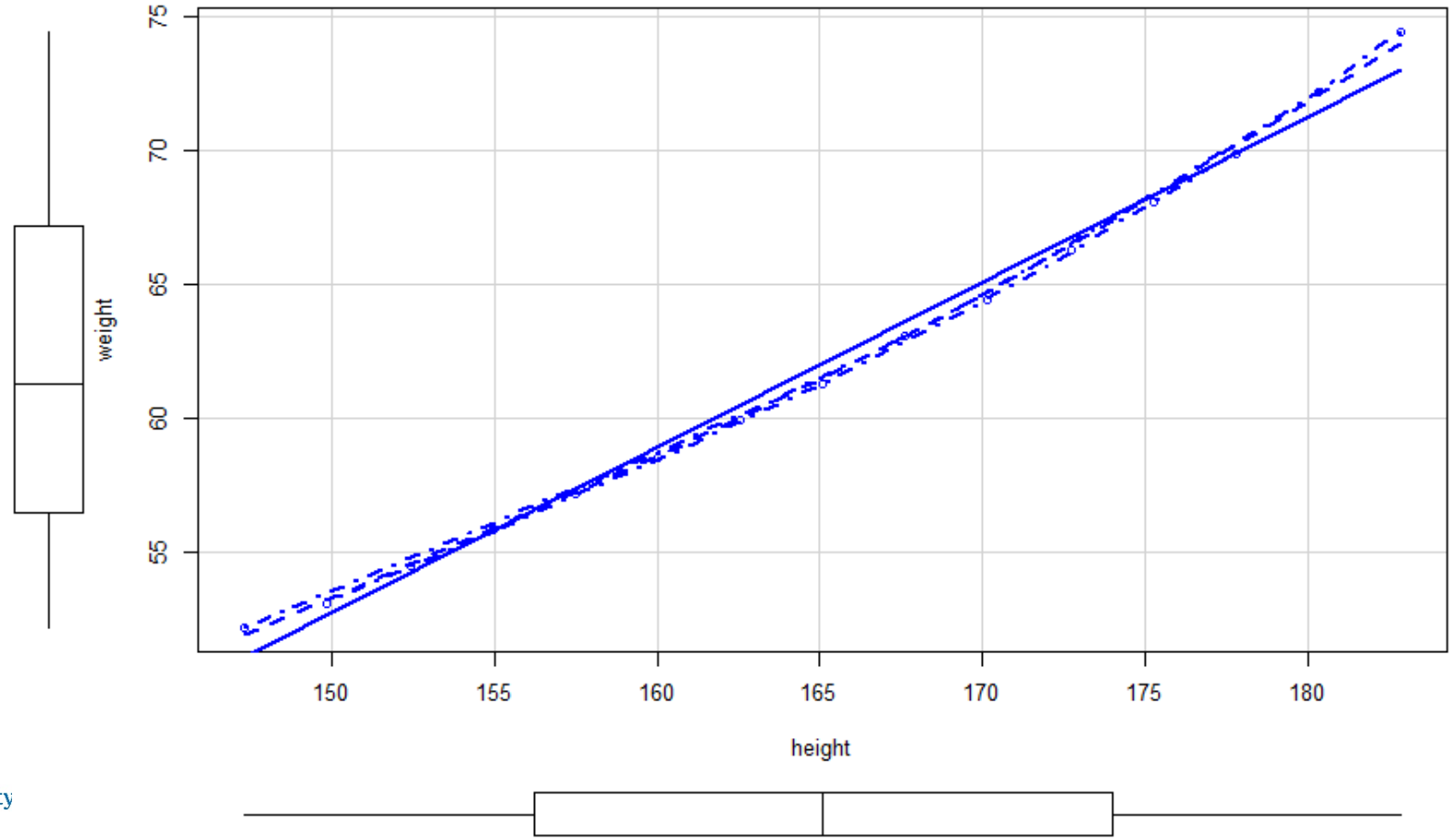
weight=118,9-1.313*height+0,005845*height^2



Функция scatterplot() из пакета car

диаграмма рассеяния для веса и роста, диаграмма размахов для обоих переменных на соответствующих полях диаграммы, регрессионная прямая и сглаженная кривая

```
install.packages("car")
library(car)
scatterplot(weight ~ height, data=women_rus)
```



Множественная линейная регрессия

- Если существует больше одной независимой переменной, простая линейная регрессия превращается во множественную линейную регрессию, а ход вычислений становится более сложным.
- С технической точки зрения, полиномиальная регрессия – это частный случай множественной регрессии.
- При квадратичной регрессии есть две независимые переменные (X и X2), а при кубической регрессии – три независимые переменные (X, X2, X3).

Множественная линейная регрессия

- набор данных state.x77
- исследование связи между уровнем преступности и другими характеристиками для каждого штата, включая численность населения, уровень неграмотности, средний доход и морозность (среднее число дней с отрицательной температурой).

```
> state.x77
```

	Population	Income	Illiteracy	Life Exp	Murder	HS Grad	Frost	Area
Alabama	3615	3624	2.1	69.05	15.1	41.3	20	50708
Alaska	365	6315	1.5	69.31	11.3	66.7	152	566432
Arizona	2212	4530	1.8	70.55	7.8	58.1	15	113417
Arkansas	2110	3378	1.9	70.66	10.1	39.9	65	51945
California	21198	5114	1.1	71.71	10.3	62.6	20	156361
Colorado	2541	4884	0.7	72.06	6.8	63.9	166	103766
Connecticut	3100	5348	1.1	72.48	3.1	56.0	139	4862
Delaware	579	4809	0.9	70.06	6.2	54.6	103	1982
Florida	8277	4815	1.3	70.66	10.7	52.6	11	54090
Georgia	4931	4091	2.0	68.54	13.9	40.6	60	58073
Hawaii	868	4963	1.9	73.60	6.2	61.9	0	6425
Idaho	813	4119	0.6	71.87	5.3	59.5	126	82677
Illinois	11197	5107	0.9	70.14	10.3	52.6	127	55748
Indiana	5313	4458	0.7	70.88	7.1	52.9	122	36097
Iowa	2861	4628	0.5	72.56	2.3	59.0	140	55941

Множественная линейная регрессия

```
> st1 <- as.data.frame(state.x77[,c("Murder","Population","Illiteracy","Income","Frost")])
> st1
```

	Murder	Population	Illiteracy	Income	Frost
Alabama	15.1	3615	2.1	3624	20
Alaska	11.3	365	1.5	6315	152
Arizona	7.8	2212	1.8	4530	15
Arkansas	10.1	2110	1.9	3378	65
California	10.3	21198	1.1	5114	20
Colorado	6.8	2541	0.7	4884	166
Connecticut	3.1	3100	1.1	5348	139
Delaware	6.2	579	0.9	4809	103
Florida	10.7	8277	1.3	4815	11
Georgia	13.9	4931	2.0	4091	60
Hawaii	6.2	868	1.9	4963	0
Idaho	5.3	813	0.6	4119	126
Illinois	10.3	11197	0.9	5107	127
Indiana	7.1	5313	0.7	4458	122

Множественная линейная регрессия

- Важный первый шаг в множественной регрессии – исследование парных взаимосвязей между переменными. Двухмерные корреляции вычисляются при помощи функции cor(), а диаграммы рассеяния создаются при помощи функции scatterplotMatrix() из пакета car()

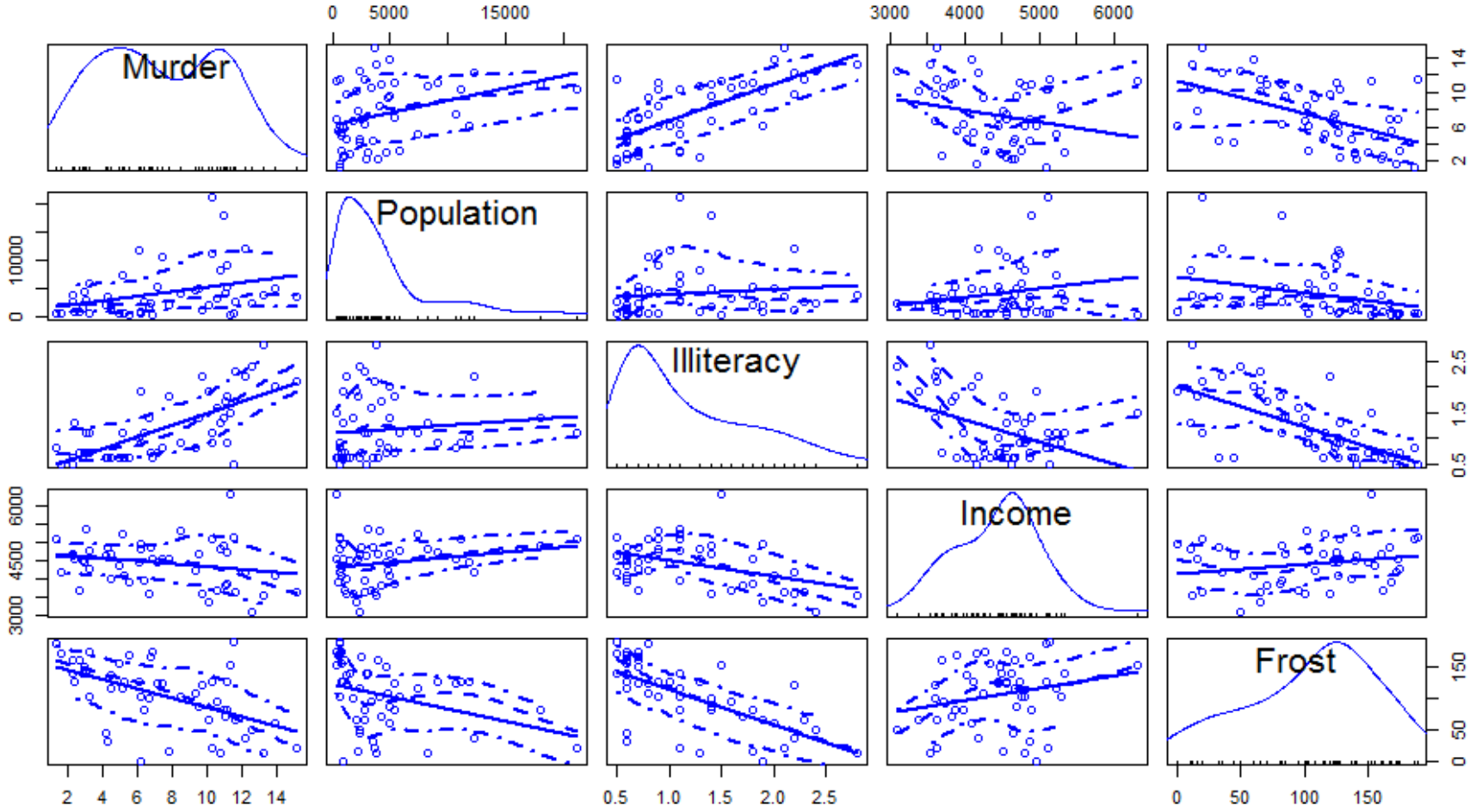
```
> cor(st1)
             Murder Population Illiteracy     Income       Frost
Murder     1.0000000  0.3436428  0.7029752 -0.2300776 -0.5388834
Population 0.3436428  1.0000000  0.1076224  0.2082276 -0.3321525
Illiteracy 0.7029752  0.1076224  1.0000000 -0.4370752 -0.6719470
Income     -0.2300776 0.2082276 -0.4370752  1.0000000  0.2262822
Frost      -0.5388834 -0.3321525 -0.6719470 0.2262822  1.0000000
```

scatterplotMatrix() ≡ scatterplot.matrix()

Множественная линейная регрессия

scatterplotMatrix(st1, spread=FALSE, lty.smooth=2)

По умолчанию функция создает диаграммы рассеяния для всех пар переменных с наложенными сглаженной (loess) кривой и регрессионной прямой. На главной диагонали представлены диаграммы плотности.



Множественная линейная регрессия

```
> lm1 <- lm(Murder ~ Population + Illiteracy + Income + Frost, data=st1)
> summary(lm1)

Call:
lm(formula = Murder ~ Population + Illiteracy + Income + Frost,
    data = st1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7960 -1.6495 -0.0811  1.4815  7.6210

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.235e+00  3.866e+00   0.319   0.7510
Population  2.237e-04  9.052e-05   2.471   0.0173 *
Illiteracy  4.143e+00  8.744e-01   4.738 2.19e-05 ***
Income      6.442e-05  6.837e-04   0.094   0.9253
Frost       5.813e-04  1.005e-02   0.058   0.9541
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.535 on 45 degrees of freedom
Multiple R-squared:  0.567,     Adjusted R-squared:  0.5285
F-statistic: 14.73 on 4 and 45 DF,  p-value: 9.133e-08
```

Множественная линейная регрессия со взаимодействиями

- Исследования взаимодействий между независимыми переменными (mtcars).
- Влияние веса автомобиля и мощности двигателя на расход топлива. Можно подобрать регрессионную модель, включающую обе независимые переменные, а также взаимодействие между ними

```
> lm1 <- lm(mpg ~ hp + wt + hp:wt, data=mtcars)
> summary(lm1)

Call:
lm(formula = mpg ~ hp + wt + hp:wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0632 -1.6491 -0.7362  1.4211  4.5513

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 49.80842    3.60516  13.816 5.01e-14 ***
hp          -0.12010    0.02470  -4.863 4.04e-05 ***
wt          -8.21662    1.26971  -6.471 5.20e-07 ***
hp:wt        0.02785    0.00742   3.753 0.000811 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.153 on 28 degrees of freedom
Multiple R-squared:  0.8848,    Adjusted R-squared:  0.8724
F-statistic: 71.66 on 3 and 28 DF,  p-value: 2.981e-13
```

mpg = 49.80842 – 0.12010*hp – 8.21662*wt + 0.03*hp*wt

:	Обозначает взаимодействие между независимыми переменными. Предсказание значений y по значениям x, z и взаимодействия между x и z будет закодировано как y ~ x + z + x:z

Множественная линейная регрессия со взаимодействиями

- Столбец Pr(>|t|) hp:wt 0.000811
- Взаимодействие между мощностью двигателя и весом машины значимо. Значимое взаимодействие между двумя независимыми переменными свидетельствует о том, что на взаимосвязь между одной независимой переменной и зависимой влияют значения другой независимой переменной.
- В данном случае характер зависимости между расходом топлива и мощностью двигателя не одинаков для автомобилей разного веса.

Интерпретация взаимодействия hp и wt

```
> summary(mtcars$wt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.513   2.581   3.325   3.217   3.610   5.424
> sd(mtcars$wt)
[1] 0.9784574
```

mpg = 49.80842 – 0.12010*hp – 8.21662*wt + 0.03*hp*wt

среднее значение wt - 3.217
значения на одно стандартное отклонение меньше - 2.238543
больше - 4.195457

Для wt=2.238543 уравнение принимает вид:
mpg = 49.80842 – 0.12010*hp – 8.21662*2.238543 + 0.03*hp*2.238543
mpg = 31.41516 – 0.1872563*hp

Для wt=3.217 оно становится таким:
mpg = 49.80842 – 0.12010*hp – 8.21662*3.217 + 0.03*hp*3.217
mpg = 23.37555 – 0.02359*hp

Для wt=4.195457 выражение приобретает вид:
mpg = 49.80842 – 0.12010*hp – 8.21662*4.195457 + 0.03*hp*4.195457
mpg = 15.33594 + 0.00576371*hp

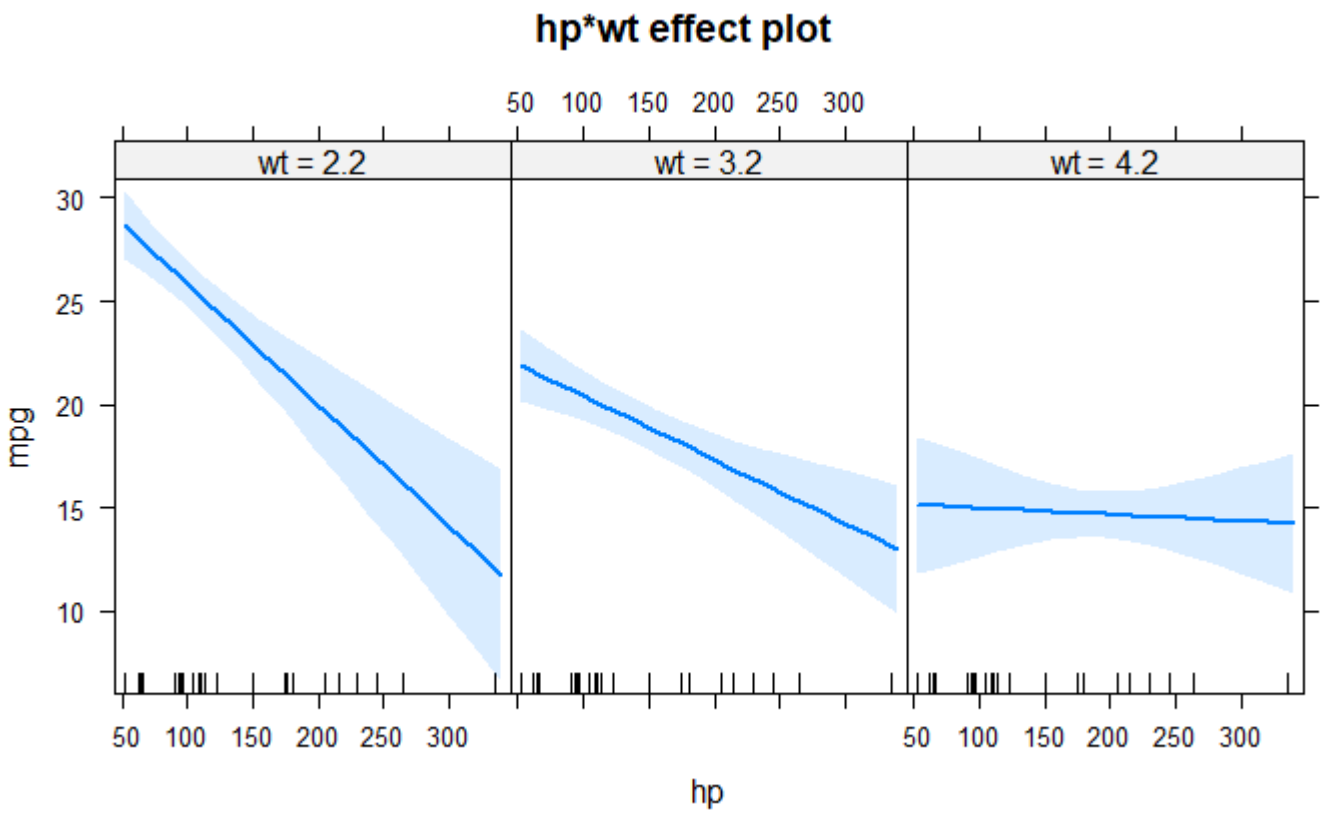
с увеличением веса ожидаемое изменение mpg на единицу изменения hp уменьшается

Функция effect() из effect

- **plot(effect(term, mod, xlevels), multiline=TRUE)**
- term – это член модели, который нужно отобразить на диаграмме,
- mod – подогнанная модель, выдаваемая функцией lm(),
- xlevels – это список переменных, значения которых будут фиксированы, и самих этих значений.
- multiline=TRUE позволяет наложить на диаграмму линии.

Функция effect() из effect

```
install.packages("effects")
library(effects)
lm1 <- lm(mpg ~ hp + wt + hp:wt, data=mtcars)
plot(effect("hp:wt", lm1, xlevels=list(wt=c(2.2,3.2,4.2))))
```

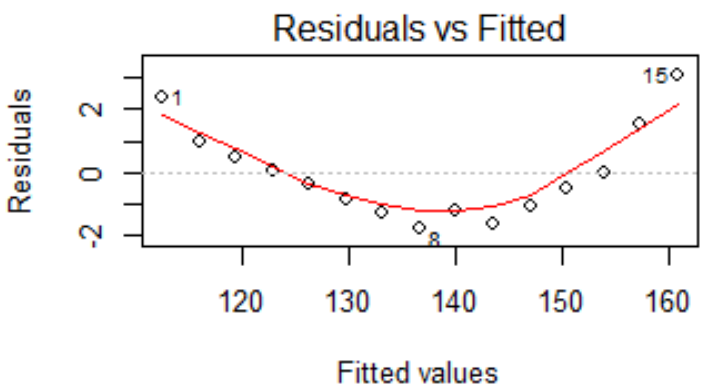


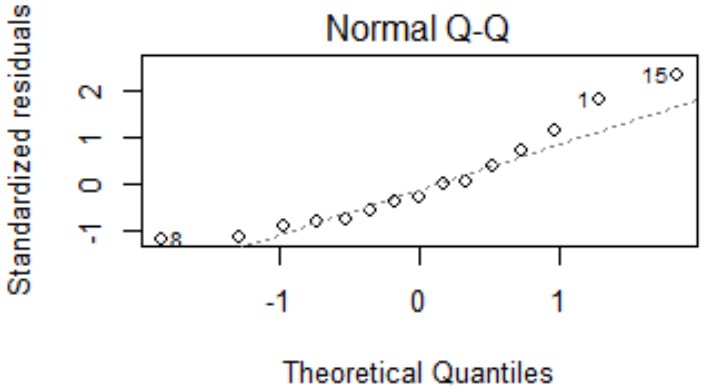
С увеличением веса автомобиля взаимодействие между мощностью мотора и расходом топлива ослабевает.
Для wt=4.2 линия почти горизонтальная, что говорит о незначительном изменении значений mpg при увеличении hp.

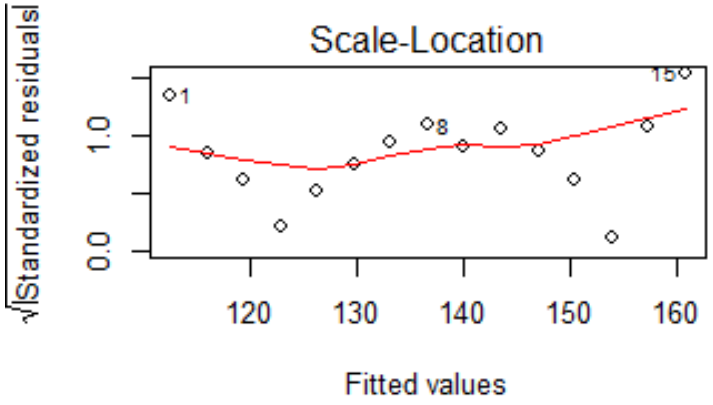
Диагностика регрессионных моделей

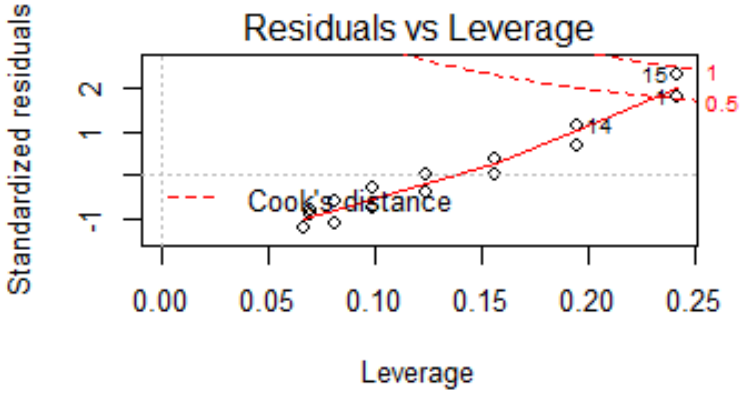
Стандартный подход

```
lm1 <- lm(weight ~ height, data=women)
par(mfrow=c(2,2))
plot(lm1)
```









Допущения МНК-регрессии

- **Нормальность**. Если значения зависимой переменной нормально распределены при постоянных значениях независимых переменных, тогда остатки должны быть нормально распределены со средним значением 0. Графическая проверка данных на нормальность (Normal Q-Q plot) это построение графика распределения вероятностей, сопоставляющего стандартизованные остатки и значения, которые ожидаются при нормальном распределении. Если допущение о нормальном распределении выполняется, то точки на этой диаграмме должны ложиться на прямую с углом наклона в 45°. Поскольку здесь это не наблюдается, это допущение не выполняется.
- **Независимость**. Из этих диаграмм нельзя сказать, насколько значения прогнозируемой переменной независимы. Для этого нужно понимать, как были собраны данные. Нет никаких априорных оснований полагать, что вес одной женщины зависит от веса другой женщины.

Допущения МНК-регрессии

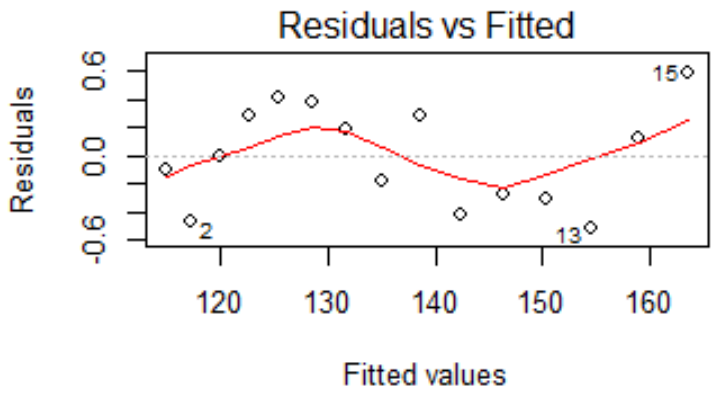
- **Линейность**. Если зависимая переменная линейно связана с независимой, то связь между остатками и предсказанными (то есть подогнанными) значениями отсутствует. Другими словами, модель должна отражать всю закономерную изменчивость в данных, учитывая все, кроме белого шума. На диаграмме зависимости остатков от предсказанных значений (Residuals vs Fitted) вы ясно видите нелинейную зависимость, что позволяет задуматься о добавлении квадратного члена в уравнение регрессии.
- **Гомоскедастичность** (однородность дисперсии). Если допущение о постоянной изменчивости выполняется, то точки на диаграмме (Scale-Location) должны располагаться в форме полосы вокруг горизонтальной линии.

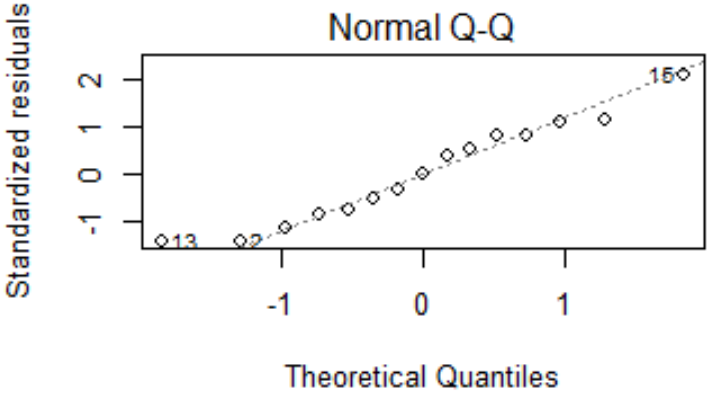
Допущения МНК-регрессии

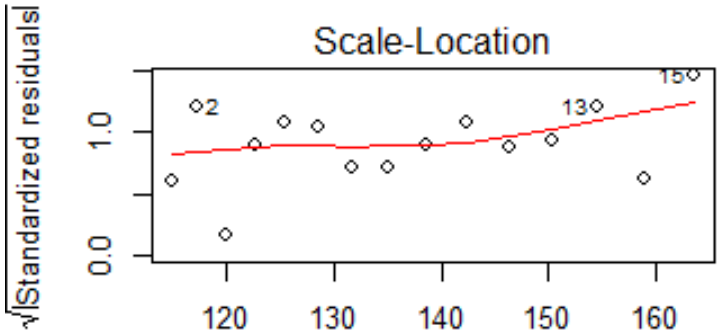
- Диаграмма зависимости остатков от «показателя напряженности» (Residuals vs Leverage) содержит информацию о наблюдениях, на которые следует обратить внимание.
- Диаграмма выявляет выбросы, точки высокой напряженности и влиятельные наблюдения.
  - Выброс – это значение, которое плохо предсказывается подобранной моделью (то есть имеет большой положительный или отрицательный остаток).
  - Значение с высоким значением напряженности описывается необычной комбинацией независимых переменных. Таким образом, это выброс в пространстве независимых переменных.
  - Значения зависимой переменной не используются при вычислении напряженности.
  - Влиятельное наблюдение – это значение, которое вносит непропорциональный вклад в расчет параметров модели. Влиятельные наблюдения выявляются при помощи статистики, называемой расстоянием Кука (Cook's distance, Cook's D).

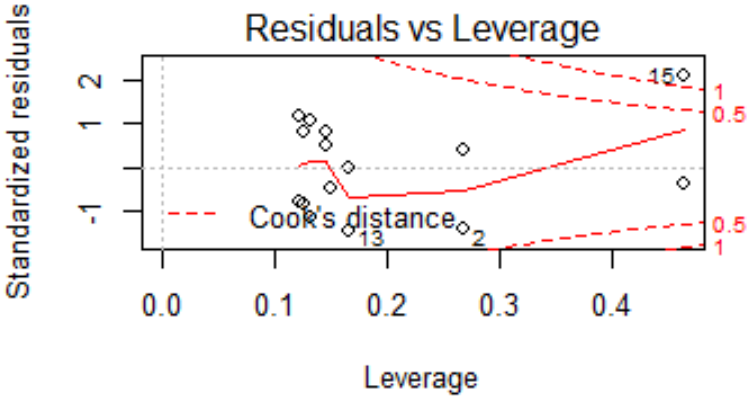
Диагностические диаграммы для квадратичной регрессии

```
lm2 <- lm(weight ~ height + I(height^2), data=women)
par(mfrow=c(2,2))
plot(lm2)
```

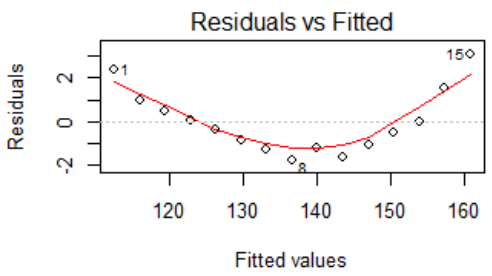


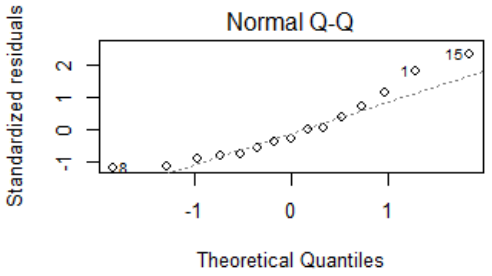


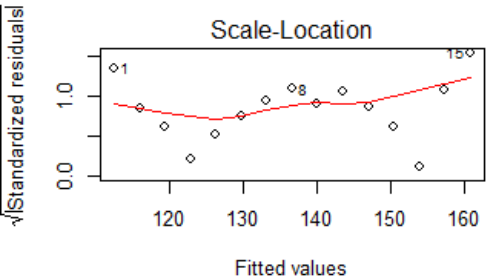


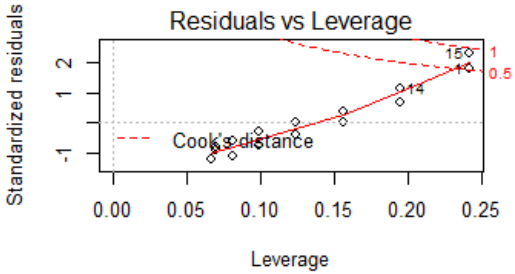


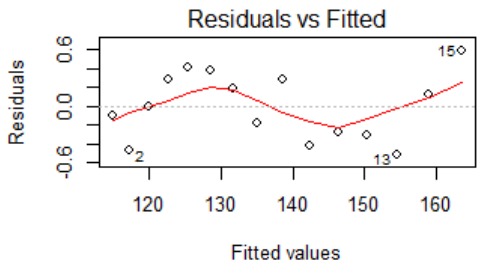
RUDN university

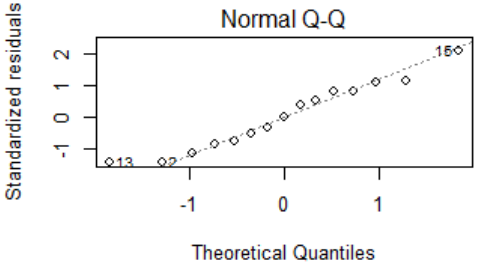


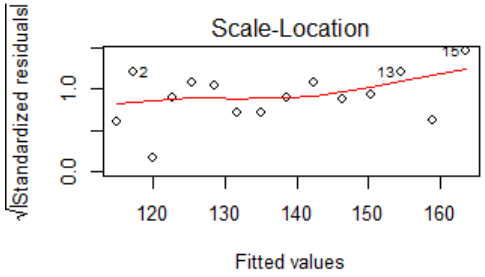


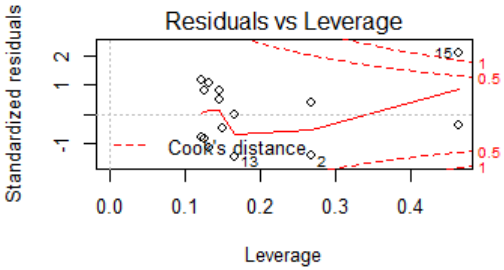








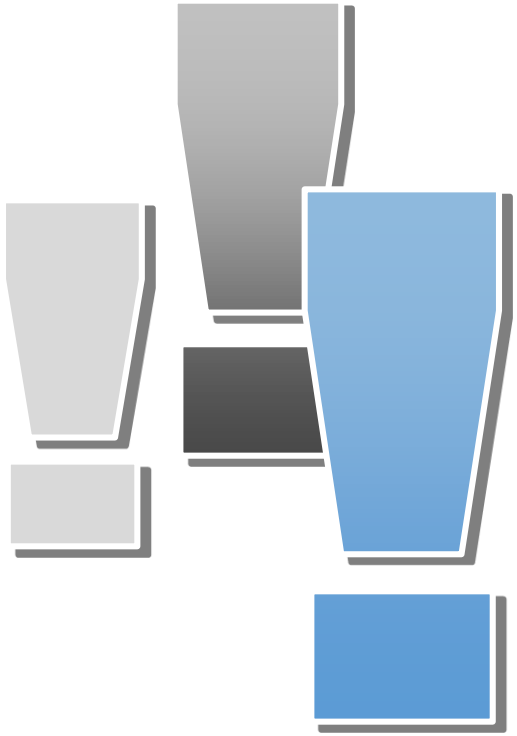




Полиномиальная регрессия подходит лучше, поскольку учитывает требования:

- линейности (Residuals vs Fitted)
- нормального распределения остатков (Normal Q-Q plot)
- гомоскедастичности (Scale-Location).

Спасибо за внимание!



Шевцов Василий Викторович

shevtsov_vv@rudn.university
+7(903)144-53-57