Липецкий государственный технический университет

Кафедра прикладной математики

КОМПЬЮТЕРНЫЕ ТЕХНОЛОГИИ МАТЕМАТИЧЕСКИХ ИССЛЕДОВАНИЙ

Лекция 6

5-6. Корреляционный анализ. Регрессии

Составитель - Сысоев А.С., к.т.н.

Липецк – 2017

План

0. Корреляционные связи

Выявление корреляционных связей способствует решению широкого круга задач. В некоторых случаях требуется подтвердить не наличие, а отсутствие корреляционной связи.

Наличие корреляционной связи не всегда означает наличие причинно-следственной зависимости. Существуют три пути возникновения корреляционной связи:

→ **причинная зависимость** результативного признака (его вариации) от вариации факторного признака (плодородность почв - урожай);

→ **корреляционная связь между двумя следствиями общей причины** (количество пожарных - ущерб от пожара);

→ **взаимосвязь признаков, каждый из которых и причина, и следствие** (производительность труда - уровень оплаты труда).

Результаты корреляционного анализа необходимо проверять логикой, опираясь на теоретические и практические знания об исследуемых свойствах.

1. Корреляции

1.1. Типы корреляций

Основная задача *корреляционного анализа* - выявление связи между случайными переменными путем точечной и интервальной оценки различных (парных, множественных, частных) коэффициентов корреляции.

Коэффициенты корреляции используются для описания связей между количественными переменными. Знак коэффициента (+ или –) свидетельствует о направлении связи (положительная или отрицательная), а величина коэффициента показывает силу связи (варьирует от 0 – нет связи до 1 – абсолютно предсказуемая взаимосвязь).

Линейный коэффициент корреляции Пирсона (Pearson product moment correlation) отражает степень линейной связи между двумя количественными переменными.

Коэффициент ранговой корреляции Спирмена (Spearman's Rank Order correlation) – мера связи между двумя ранжированными переменными.

Тау Кэнделла (Kendall's Tau) – непараметрический показатель ранговой корреляции.

1. Корреляции
1.1. Типы корреляций

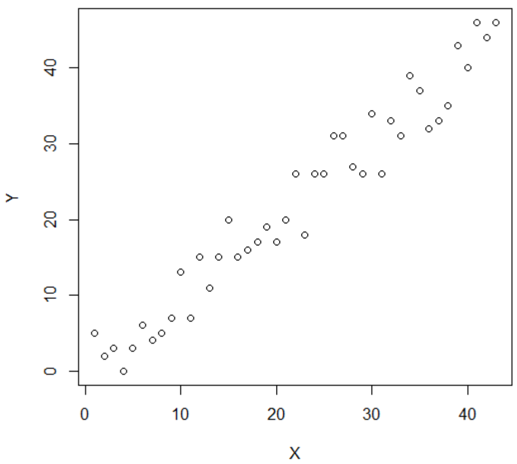
ЛИНЕЙНЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ ПИРСОНА

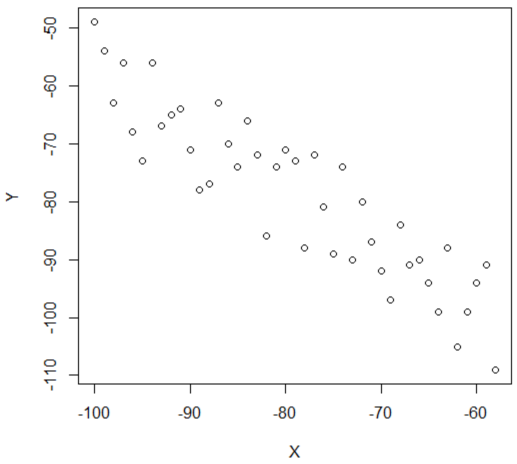
Пусть исходными данными является набор случайных векторов $(X,Y)=(x_i,y_i),\ i=1,...,n.$

Выборочным коэффициентом корреляции (выборочным линейным парным коэффициентом корреляции К. Пирсона) называют число

$$r=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}.$$

Если r = 1, то $y_i = ax_i + b$, причем $a > 0$. Если же r = -1, то $y_i = ax_i + b$, причем $a < 0$.





ШКАЛА ЧЕДДОКА				
значение r				
0,1-0,3	0,3-0,5	0,5-0,7	0,7-0,9	0,9-0,99
Слабая связь	Умеренная связь	Заметная связь	Высокая связь	Весьма высокая связь

КОЭФФИЦИЕНТ РАНГОВОЙ КОРРЕЛЯЦИИ СПИРМЕНА

Для переменных, принадлежащих к порядковой шкале или для переменных, не подчиняющихся нормальному распределению, а также для переменных принадлежащих к интервальной шкале, вместо коэффициента Пирсона рассчитывается ранговая корреляция по Спирмену. Для этого отдельным значениям переменных присваиваются ранговые места, которые впоследствии обрабатываются с помощью соответствующих формул.

Практический расчет коэффициента ранговой корреляции Спирмена включает следующие этапы:

1) Сопоставать каждому из признаков их порядковый номер (ранг) по возрастанию (или убыванию).

2) Определить разности рангов каждой пары сопоставляемых значений.

3) Возвести в квадрат каждую разность и суммировать полученные результаты.

4) Вычислить коэффициент корреляции рангов по формуле $r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$,

где $\sum_{i=1}^{n} d_i^2$ - сумма квадратов разностей рангов, n - число парных наблюдений.

ТАУ КЭНДЕЛЛА

В этом методе одна переменная представляется в виде монотонной последовательности в порядке возрастания величин; другой переменной присваиваются соответствующие ранговые места. Количество инверсий (нарушений монотонности по сравнению с первым рядом) используется в формуле для корреляционных коэффициентов:

$$\tau = \frac{P(p) - P(q)}{n \cdot \frac{n-1}{2}},$$

где P(p) - число совпадений, P(q) - число инверсий, n - объем выборки.

Применение коэффициента Кендалла является предпочтительным, если в исходных данных встречаются выбросы.

Функция `cor(x, use= , method= )`

Опция	Описание
`x`	Матрица или таблица данных
`use=`	Упрощает работу с пропущенными данными. Может принимать следующие значения: `all.obs` (предполагается, что пропущенные значения отсутствуют; их наличие вызовет сообщение об ошибке), `everything` (любая корреляция, включающая строку с пропущенным значением, не будет вычисляться – обозначается как `missing`), `complete.obs` (учитываются только строки без пропущенных значений) и `pairwise.complete.obs` (учитываются все полные наблюдения для каждой пары переменных в отдельности)
`method=`	Определяет тип коэффициента корреляции. Возможные значения – `pearson`, `spearman` или `kendall`

Частная корреляция – это корреляция между двумя количественными переменными, зависящими, в свою очередь, от одной или более других количественных переменных.

`pcor(u, S)` (пакет `ggm`)

где *u* – это числовой вектор, в котором первые два числа – это номера переменных, для которых нужно вычислить коэффициент, а остальные числа – номера «влияющих» переменных (воздействие которых должно быть отделено), *S* – это ковариационная матрица для всех этих переменных.

1. Корреляции

1.3. Проверка статистической значимости коэффициента корреляции

Проверяя значимость коэффициента парной корреляции, устанавливают наличие или отсутствие корреляционной связи между исследуемыми явлениями. При отсутствии связи коэффициент корреляции генеральной совокупности равен нулю.

Процедура проверки начинается с формулировки нулевой и альтернативной гипотез:

H_0: различие между выборочным коэффициентом корреляции r и $\rho = 0$ незначимо;

H_1: различие между r и $\rho = 0$ значимо, и следовательно, между переменными y и x имеется существенная связь.

Вычисленная по результатам выборки статистика $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ сравнивается с критическим значением, определяемым по таблице распределения Стьюдента при заданном уровне значимости α и f = n-2 степенях свободы:

- $\rightarrow$ если $|t| > t_{f,\alpha}$, то нулевая гипотеза на уровне значимости α отвергается, т.е. связь между переменными значима;
- $\rightarrow$ если $|t| \le t_{f,\alpha}$, то нулевая гипотеза на уровне значимости α принимается. Отклонение значения r от $\rho = 0$ можно приписать случайной вариации. Данные выборки характеризуют рассматриваемую гипотезу как весьма возможную и правдоподобную, т.е. гипотеза об отсутствии связи не вызывает возражений.

1. Корреляции

1.4. Проверка статистической значимости коэффициента корреляции в R

Для проверки значимости отдельных корреляционных коэффициентов Пирсона, Спирмена и Кэнделла можно использовать функцию `cor.test()`

```
cor.test(x, y, alternative = , method = )
```

где *x* и *y* – это переменные, корреляция между которыми исследуется, опция `alternative` определяет тип теста (`"two.side"`, `"less"` или `"greater"`), опция `method` задает тип корреляции (`"pearson"`, `"kendall"` или `"spearman"`).

- → `alternative = "less"` для проверки гипотезы о том, что в генеральной совокупности коэффициент корреляции меньше нуля;
- → `alternative = "greater"` для проверки того, что в генеральной совокупности коэффициент корреляции больше нуля;
- → `alternative = "two.side"` - проверяется гипотеза о том, что коэффициент корреляции в генеральной совокупности не равен нулю (по умолчанию).

```
        Pearson's product-moment correlation

data:  r$Q and r$L
t = 18.2825, df = 25, p-value = 6.661e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9228305 0.9839296
sample estimates:
     cor
0.964578
```

p-value - фактически это вероятность ошибки при отклонении нулевой гипотезы (ошибки первого рода).

Проверка гипотез с помощью p-значения является альтернативой классической процедуре проверки через критическое значение распределения.

2. Регрессии

2.1. Парная линейная регрессия

Для характеристики формы связи при изучении корреляционной зависимости пользуются уравнением регрессии. Задача ставится таким образом: по данной выборке объема n найти уравнение регрессии и оценить допускаемую при этом ошибку.

Уравнение прямой на плоскости $Y = \beta_0 + \beta_1 X$. Для определения линии регрессии необходимо непременно статистически оценить коэффициент регрессии β_1 и постоянное число β_0. Для этого должны быть удовлетворены два следующих условия:

1. Линия регрессии должна проходить через точку с координатами (X; Y) средних значений $\bar{X}$ и $\bar{Y}$.

2. Сумма квадратов отклонений от линии регрессии вдоль оси Oy должна быть наименьшей:

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \to \min.$$

Оценки коэффициентов регрессии:

$$\beta_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \bar{Y}\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2 - \bar{X}\sum_{i=1}^{n} X_i}, \qquad \beta_0 = \bar{Y} - \beta_1 \bar{X}.$$

ЗНАЧИМОСТЬ УРАВНЕНИЯ ПАРНОЙ РЕГРЕССИИ

Для проверки значимости уравнения регрессии используют F-критерий (критерий Фишера).

Для этого определяют общую дисперсию σ_Y^2 и остаточную $\sigma_{ОСТ}^2$:

$$\sigma_Y^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}\right], \qquad \sigma_{ОСТ}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

и определяют их отношение $F_0 = \dfrac{\sigma_Y^2}{\sigma_{ОСТ}^2}$.

Если $F_0 > F_{n-1,\, n-2,\, \alpha}$, то уравнение статистически значимо описывает результаты экспериментов.

ЗНАЧИМОСТЬ ОЦЕНОК ПАРАМЕТРОВ УРАВНЕНИЯ ПАРНОЙ РЕГРЕССИИ

Для проверки значимости параметров уравнения регрессии используют t-критерий (критерий Стьюдента).

Выдвигают следующие гипотезы:

H_0: $\beta_k = b_k$, т.е. нет существенного различия между оценкой параметра регрессии, полученной по результатам выборки, и истинным значением параметра;

H_1: $\beta_k \neq b_k$, т.е. имеется значимая разница между оценкой параметра регрессии и соответствующим параметром генеральной совокупности.

Вычисляют значения t-статистики для каждого параметра модели:

$$t_{\beta_k} = \frac{\beta_k}{m_{\beta_k}} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 (n-2)}{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}.$$

Число степеней свободы статистики f = n - m - 1, где m - количество объясняющих переменных, включенных в регрессию. Значение t сравнивают с критическим табличным значением: если $t > t_{f,\alpha}$, то β_k значимо отличается от b_k т.е. нельзя предположить, что выборка отобрана из генеральной совокупности с параметром регрессии b_k.

ЗНАЧИМОСТЬ ОЦЕНОК ПАРАМЕТРОВ УРАВНЕНИЯ ПАРНОЙ РЕГРЕССИИ

Проверка значимости свободного члена выполняется аналогично.

H_0: $\beta_0 = 0$; H_1: $\beta_0 \neq 0$.

Однако значимость этого параметра имеет второстепенное значение, т.к. чаще всего его значение лишено содержательного смысла.

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ

$$\underbrace{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}_{TSS} = \underbrace{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2}_{RSS} + \underbrace{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}_{ESS}$$

TSS – вся дисперсия: сумма квадратов отклонений от среднего.

RSS – объясненная часть всей дисперсии (обусловленная регрессией).

ESS – остаточная сумма.

Коэффициентом детерминации, или долей объясненной дисперсии называется

$$R^2 = 1 - \frac{ESS}{TSS} = \frac{RSS}{TSS} \quad \text{(в силу определения } 0 \le R^2 \le 1\text{).}$$

Скорректированный коэффициент детерминации $R^2_{adj} = 1 - (1 - R^2)\dfrac{n-1}{n-k-1}$

2. Регрессии
2.2. МНК-регрессии

МНК-регрессия позволяет подгонять модели вида

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

где *n* – это число наблюдений, а *k* – это число независимых переменных.

$\widehat{Y}_i$	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)

Необходимо выбрать такие параметры модели, чтобы:

$$\sum_1^n \left(Y_i - \widehat{Y}_i\right)^2 = \sum_1^n \left(Y_i - \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \ldots + \widehat{\beta}_k X_{ki}\right)^2 = \sum_1^n \varepsilon^2$$

Требования к данным:

нормальность – значения зависимой переменной нормально распределены при фиксированных значениях независимых переменных;

независимость – значения Y_i независимы друг от друга;

линейность – зависимая переменная линейно связана с независимыми;

гомоскедастичность – дисперсия зависимой переменной постоянна при разных значениях независимых переменных.

В R основная функция для подгонки регрессионных моделей – это `lm()`.

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

где `formula` описывает вид модели, которую нужно подогнать, а `data` – это таблица с данными, которые используются для создания модели. Полученный объект (`myfit`) – это список, содержащий обширную информацию о подогнанной модели.

Формула обычно записывается в таком виде:

`Y ~ X1 + X2 + … + Xk`

где ~ отделяет зависимую переменную слева от независимых переменных (разделенных знаками +) справа.

Символ	Обозначение
~	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`

Символ	Обозначение
`^`	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`
`.`	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`

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
`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

Задача. Набор содержит данные о выпуске и капитале металлургической промышленности.

```
> fit <-lm(r$Q ~ r$K, data=r)
> summary(fit)

Call:
lm(formula = r$Q ~ r$K, data = r)

Residuals:
     Min       1Q   Median       3Q      Max 
-1282.26  -216.22     2.42   109.78  1183.90 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 391.96909  131.43334   2.982   0.0063 ** 
r$K           0.71491    0.03234  22.104   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 506.6 on 25 degrees of freedom
Multiple R-squared:  0.9513,    Adjusted R-squared:  0.9494 
F-statistic: 488.6 on 1 and 25 DF,  p-value: < 2.2e-16
```

Из полученных результатов следует, что уравнение для предсказания веса по росту имеет следующий вид:

$$Q = 391.97 + 0.71 \cdot K$$

Точность предсказания предыдущей задачи можно улучшить, если использовать квадратичное регрессионное уравнение (то есть включить в него X^2).

```
fit <-lm(r$Q ~ r$K + I(r$K^2) , data=r)
```

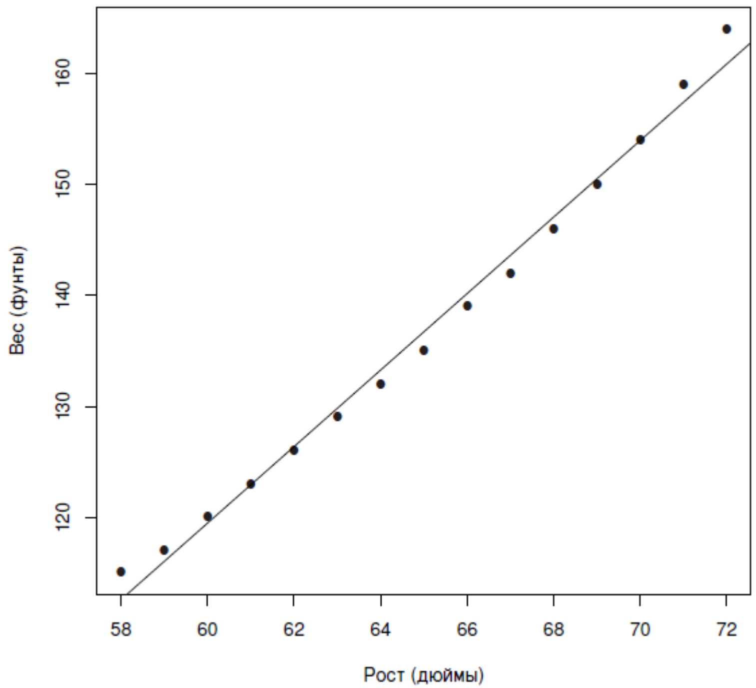
```
> fit <-lm(r$Q ~ r$K + I(r$K^2) , data=r)
> summary(fit)

Call:
lm(formula = r$Q ~ r$K + I(r$K^2), data = r)

Residuals:
     Min       1Q   Median       3Q      Max
-1401.12  -240.66    18.71   168.11  1116.46

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.940e+02  1.830e+02   1.607    0.121
r$K          7.862e-01  9.738e-02   8.073 2.68e-08 ***
I(r$K^2)    -5.823e-06  7.500e-06  -0.777    0.445
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 510.7 on 24 degrees of freedom
Multiple R-squared:  0.9525,    Adjusted R-squared:  0.9486
F-statistic: 240.7 on 2 and 24 DF,  p-value: < 2.2e-16
```

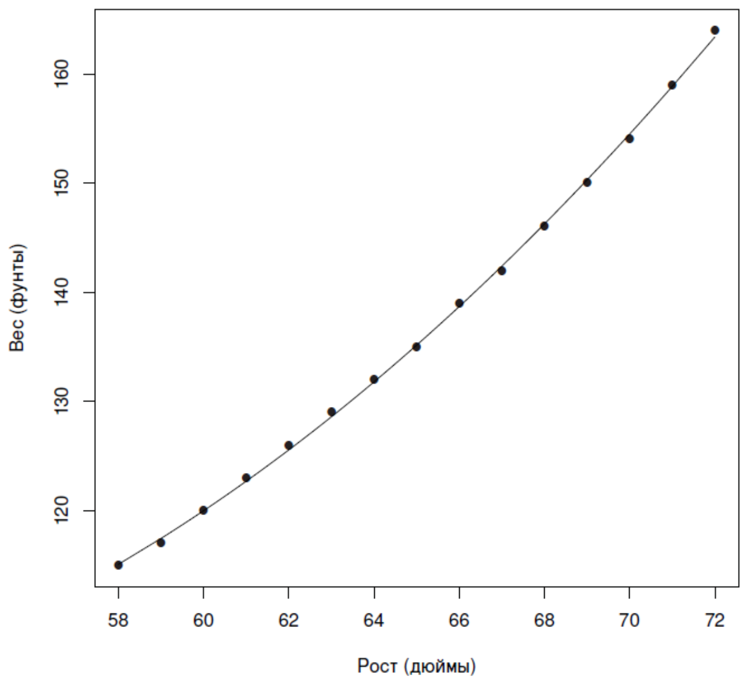
По результатам этого нового анализа регрессионное уравнение приобретает вид

$$\widehat{\mathrm{Weight}} = 261.88 + 7.35 \times Height + 0.083 \times Height^2$$

В общем случае полиномиальная функция *n*-ой степени соответствует кривой с *n – 1* изгибами. Для подгонки модели кубической функции нужно использовать выражение

```
fit3 <- lm(weight ~ height + I(height^2) +I(height^3), data=women)
```





2. Регрессии

2.6. Множественная линейная регрессия

Множественная линейная регрессия имеет вид

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

где *n* – это число наблюдений, а *k* – это число независимых переменных.

```
fit_mn <- lm (CCM_ASTC_Thickness_final ~ CCM_MCO_Water_flow_NL +
CCM_MCO_Water_flow_NR + CCM_MCO_Water_flow_WF + CCM_MCO_Water_flow_WL,
data = dannye)
```

```
Call:
lm(formula = CCM_ASTC_Thickness_final ~ CCM_MCO_Water_flow_NL +
    CCM_MCO_Water_flow_NR + CCM_MCO_Water_flow_WF + CCM_MCO_Water_flow_WL,
    data = dannye)

Residuals:
     Min       1Q   Median       3Q      Max
-0.02457 -0.00788 -0.00197  0.00091  1.09474

Coefficients:
                       Estimate Std. Error  t value Pr(>|t|)
(Intercept)           2.478e+02  1.520e-01 1630.358  < 2e-16 ***
CCM_MCO_Water_flow_NL  3.595e-04  2.270e-04    1.584  0.11378
CCM_MCO_Water_flow_NR -1.203e-03  4.956e-04   -2.427  0.01548 *
CCM_MCO_Water_flow_WF  8.323e-04  8.279e-05   10.053  < 2e-16 ***
CCM_MCO_Water_flow_WL  2.585e-04  8.919e-05    2.899  0.00387 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04626 on 669 degrees of freedom
Multiple R-squared:  0.9886,    Adjusted R-squared:  0.9885
F-statistic: 1.452e+04 on 4 and 669 DF,  p-value: < 2.2e-16
```

Задача. Рассмотрим данные об автомобилях из таблицы `mtcars`. Допустим, нас интересует влияние веса автомобиля и мощности двигателя на расход топлива. Можно подобрать регрессионную модель, включающую обе независимые переменные, а также взаимодействие между ними

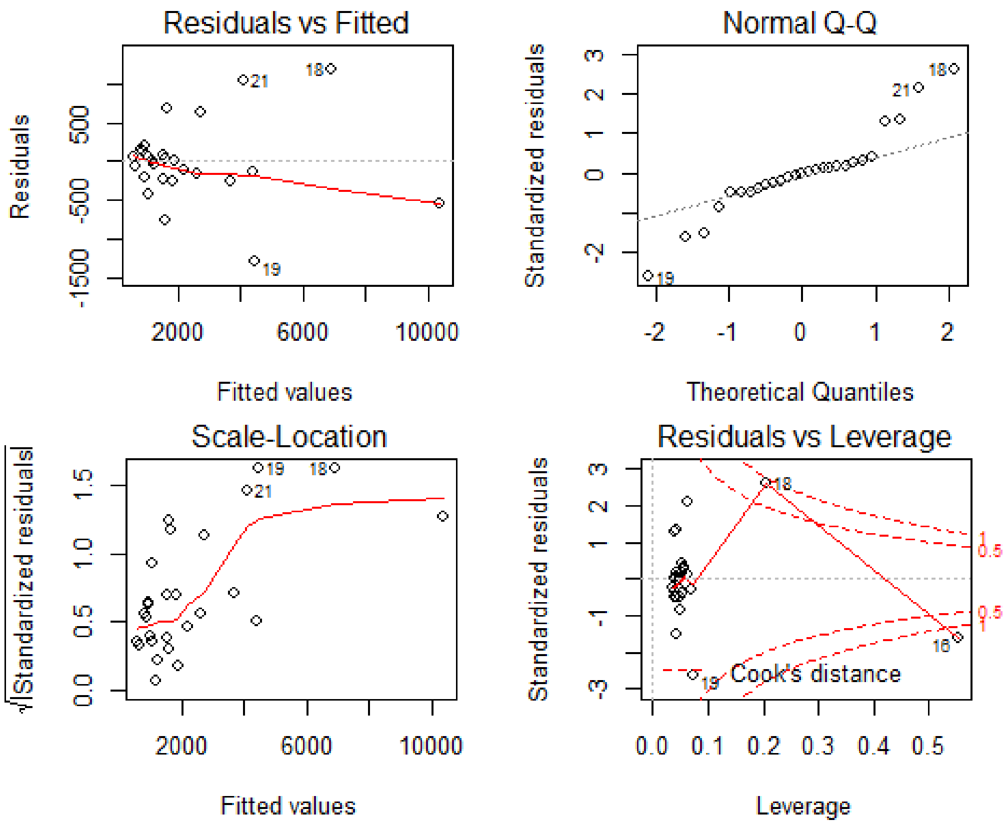
```
> fit <- lm(mpg ~ hp + wt + hp:wt, data=mtcars)
> summary(fit)

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

Residuals:
   Min     1Q Median     3Q    Max
-3.063 -1.649 -0.736  1.421  4.551

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 49.80842    3.60516   13.82  5.0e-14 ***
hp          -0.12010    0.02470   -4.86  4.0e-05 ***
wt          -8.21662    1.26971   -6.47  5.2e-07 ***
hp:wt        0.02785    0.00742    3.75  0.00081 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Модель для предсказания значений

$$\widehat{mpg} = 49.81 - 0.12 \times hp - 8.22 \times wt + 0.03 \times hp \times wt.$$

Наиболее распространенный подход – применить функцию `plot()` к объекту, представляющему собой результат действия функции `lm()`.

```
fit1 <-lm(r$Q ~ r$K, data=r)
par(mfrow=c(2,2))
plot(fit1)
```



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 – справа сверху) – 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°.

Независимость. Из этих диаграмм нельзя сказать, насколько значения прогнозируемой переменной независимы. Для этого нужно понимать, как были собраны данные.

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 (сверху слева) ясно видно нелинейную зависимость.

Гомоскедастичность. Если допущение о постоянной изменчивости выполняется, то точки на нижней правой диаграмме должны располагаться в форме полосы вокруг горизонтальной линии.

Диаграмма зависимости остатков от «показателя напряженности» (слева внизу) содержит информацию о наблюдениях, на которые, возможно, следует обратить внимание (например, выбросы).

Необычные наблюдения требуют отдельного изучения: либо потому, что они каким-то образом отличаются от прочих, либо потому, что они значительно влияют на общие результаты.

ВЫБРОСЫ

Характеризуются большими положительными или отрицательными остатками $\hat{Y}_i - Y_i$. Положительные остатки свидетельствуют о том, что модель *недооценивает* зависимую переменную, отрицательные остатки – признак *переоценки*.

В пакете `car` также реализован статистический тест на выбросы. Функция `outlierTest()` вычисляет *значение вероятности статистической ошибки первого рода для наибольшего остатка Стьюдента*

```
> library(car)
> outlierTest(fit1)

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
19 -3.027448           0.0058134      0.15696
```

Если тест не дал значимого результата, то в данных нет выбросов. А если результат теста значимый, нужно удалить выброс и провести тест заново, чтобы узнать, есть ли другие выбросы.

УДАЛЕНИЕ НАБЛЮДЕНИЙ

Удаление выбросов часто может улучшить соответствие набора данных требованию нормальности. Влиятельные наблюдения также часто удаляют, поскольку они слишком сильно влияют на результаты. После удаления наибольшего выброса или влиятельного наблюдения модель подбирается заново. Если после этого все равно остаются выбросы или влиятельные наблюдения, процесс повторяется, пока не будет достигнуто допустимое соответствие модели данным.

ПРЕОБРАЗОВАНИЕ ПЕРЕМЕННЫХ

Когда модели не отвечают требованию нормальности, линейности или гомоскедастичности, трансформация одной или более переменных может улучшить или исправить ситуацию. Преобразования обычно заключаются в замене переменной Y на переменную Y^{λ}.

	−2	−1	−0.5	0	0.5	1	2
Преобразование	$1/Y^2$	$1/Y$	$1/\sqrt{Y}$	$\log(Y)$	$\sqrt{Y}$	отсутствует	Y^2

Если модель не соответствует требованиям нормальности, обычно пытаются преобразовать зависимую переменную. При помощи функции `powerTransform()` из пакета `car` можно оценить по *методу максимального правдоподобия* величину λ, возведение в которую, скорее всего, нормализует переменную X^{λ}.

ДОБАВЛЕНИЕ ИЛИ УДАЛЕНИЕ ПЕРЕМЕННЫХ

Изменение числа переменных, входящих в модель, будет влиять на степень ее соответствия данным. Иногда добавление важной переменной может исправить многие проблемы. Удаление причиняющих беспокойство переменных может привести к аналогичному эффекту.

ПРИМЕНЕНИЕ ДРУГОГО ПОДХОДА

- При наличии выбросов и/или влиятельных наблюдений можно использовать устойчивую регрессионную модель, а не МНК-регрессию.
- Если не выполняется требование нормальности, можно подобрать нелинейную регрессионную модель.
- В случае отклонения от независимости ошибок можно применить модели, которые учитывают структуру остатков, – такие как модели временных рядов или многоуровневые регрессионные модели.
- Если требования, лежащие в основе МНК-регрессии, не выполняются, вы можете обратиться к обобщенным линейным моделям.

СРАВНЕНИЕ МОДЕЛЕЙ

Информационный критерий Акаике (Akaike Information Criterion, AIC):

$$AIC = 2k + n(\ln(RSS)),$$

где k - число оцениваемых параметров модели,

n - размер выборки,

RSS - объясненная часть всей дисперсии (обусловленная регрессией).

При расчете этого критерия учитывается статистическое соответствие модели данным и число необходимых для достижения этого соответствия параметров. Предпочтение нужно отдавать моделям с меньшими значениями AIC, указывающими на хорошее соответствие данным при использовании меньшего числа параметров. Этот критерий вычисляется при помощи функции `AIC()`

```
fit_mn_1 <- lm (CCM_ASTC_Thickness_final ~ CCM_MCO_Water_flow_NL + CCM_MCO_Water_flow_NR + CCM_MCO_Water_flow_WF + CCM_MCO_Water_flow_WL, data = dannye)
```

```
fit_mn_2 <- lm (CCM_ASTC_Thickness_final ~ CCM_MCO_Water_flow_NL + CCM_MCO_Water_flow_NR + CCM_MCO_Water_flow_WF, data = dannye)
```

```
AIC(fit_mn_1, fit_mn_2)
```

```
         df       AIC
fit_mn_1  6 -2223.229
fit_mn_2  5 -2216.816
```

ПОШАГОВАЯ РЕГРЕССИЯ

При пошаговом выборе переменные добавляются в модель или удаляются из нее по одной, пока не будет достигнуто заданное значение критерия для остановки процесса.

При методе пошагового включения (forward stepwise) переменные по одной добавляются в модель, пока добавление новых переменных не перестанет ее улучшать.

При пошаговом исключении (backward stepwise) начинают с модели, включающей все независимые переменные, а потом удаляют их по одной до тех пор, пока модель не начнет ухудшаться.

При комбинированном методе (stepwise stepwise) совмещены оба подхода. Переменные добавляются по одной, однако на каждом шаге происходит переоценка модели, и те переменные, которые не вносят значительного вклада, удаляются.

Результат применения метода пошаговой регрессии зависит от критериев включения или удаления переменных. При помощи функции `stepAIC()` из пакета `MASS` можно провести все три типа пошаговой регрессии с использованием точного критерия `AIC`.

2. Регрессии

2.10. Сравнение моделей и выбор лучшей

```
lm(formula = CCM_ASTC_Thickness_final ~ BOF_TAP_Temperature_at_end +
    CCM_ASTC_Pos_soft_reduc_end + CCM_ASTC_Pos_soft_reduct_start +
    CCM_ASTC_Soft_reduction + CCM_ASTC_Soft_reduction_rate +
    CCM_LD_Gate_sealing_gas_flow + CCM_MCO_Water_d_temp_NL +
    CCM_MCO_Water_d_temp_NR + CCM_MCO_Water_d_temp_WF + CCM_MCO_Water_d_temp_WL +
    CCM_MCO_Water_flow_NL + CCM_MCO_Water_flow_NR + CCM_MCO_Water_flow_WL,
    data = dannye)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25795 -0.00455  0.00088  0.00626  0.82751

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     2.508e+02  5.430e-01 461.840  < 2e-16 ***
BOF_TAP_Temperature_at_end     -1.149e-03  3.340e-04  -3.440 0.000617 ***
CCM_ASTC_Pos_soft_reduc_end     1.769e-02  2.905e-03   6.091 1.90e-09 ***
CCM_ASTC_Pos_soft_reduct_start -1.789e-02  2.351e-03  -7.611 9.45e-14 ***
CCM_ASTC_Soft_reduction        -2.651e-02  9.141e-03  -2.900 0.003855 **
CCM_ASTC_Soft_reduction_rate    6.601e-02  2.359e-02   2.798 0.005295 **
CCM_LD_Gate_sealing_gas_flow    1.393e-03  4.751e-04   2.931 0.003495 **
CCM_MCO_Water_d_temp_NL        -3.490e-02  5.967e-03  -5.849 7.77e-09 ***
CCM_MCO_Water_d_temp_NR        -1.346e-02  5.655e-03  -2.381 0.017559 *
CCM_MCO_Water_d_temp_WF        -4.958e-02  3.147e-02  -1.576 0.115609
CCM_MCO_Water_d_temp_WL         6.360e-02  2.988e-02   2.128 0.033688 *
CCM_MCO_Water_flow_NL          -1.308e-03  5.300e-04  -2.468 0.013822 *
CCM_MCO_Water_flow_NR          -1.316e-03  4.765e-04  -2.762 0.005904 **
CCM_MCO_Water_flow_WL           1.113e-03  3.090e-05  36.012  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03812 on 660 degrees of freedom
Multiple R-squared:  0.9924,	Adjusted R-squared:  0.9922
F-statistic:  6606 on 13 and 660 DF,  p-value: < 2.2e-16
```

СПАСИБО ЗА ВНИМАНИЕ!