Прикладная статистика в R

Лекция 5. Модель множественной линейной регрессии. Построение уравнения линейной регрессии в R

Елена Михайловна Парилина

д. ф.-м. н., проф.

2021

Модуль 3. Множественная линейная регрессия.

- Корреляционный анализ.
- Спецификация модели множественной линейной регрессии.
- Оценка параметров множественной линейной регрессии.
- Проверка значимости коэффициентов построенной модели.
- Проверка значимости построенного уравнения регрессии в целом.
- Основные предположения регрессионного анализа.
- Нарушение основных предположений регрессионного анализа.
- Интерпретация полученных результатов.
- План полного анализа регрессионной модели.
- Линейная регрессия в R.

Массив данных "Denver Neighborhoods"

Массив содержит наблюдения вида

$$(X1, X2, X3, X4, X5, X6, X7),$$

где каждое наблюдение содержит характеристики района в Денвере:

- X1 = население (в тыс. человек),
- X2 = % изменения населения за последние несколько лет,
- X3 = % детей (младше 18) среди населения,
- X4 = % государственных школ, участвующих в программе бесплатных обедов,
- X5 = % изменения в заработной платы населения за последние несколько лет,
- X6 = криминальный показатель (на 1000 населения),
- X7 = % изменения в криминальном показателе за последние несколько лет.

Предполагается, что уровень криминала (X6) может зависеть от вышеперечисленных показателей. Задача: построить "наилучшую" модель линейной регрессии.

Проверка значимости корреляции между величинами

Анализ зависимостей в R

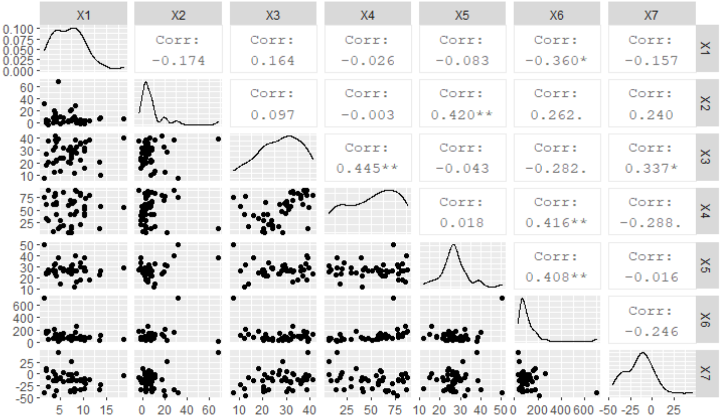
Анализ зависимостей (корреляция данных)

```
install.packages("GGally")
library("GGally")
ggpairs(neig)
ggcorr(neig, method = c("everything" ,"pearson"))
```

Пакеты для визуализации корреляции данных

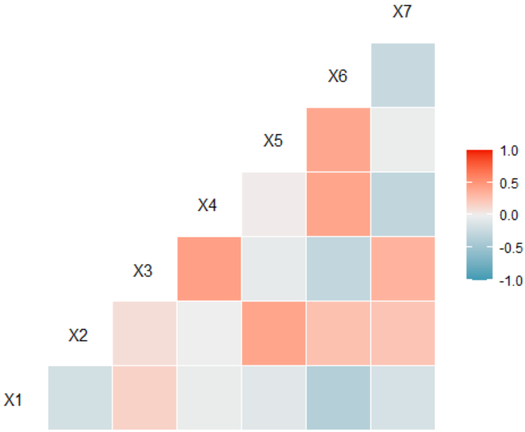
```
install.packages("PerformanceAnalytics")
install.packages("corrgram")
install.packages("ellipse")
```

Корреляционный анализ



Статистически значимые коэффициенты корреляции отмечены зведочками и точками. Чем больше звездочек, тем более "значим коэффициент".

Корреляционный анализ



Корреляционный анализ

Проверим значимость коэффициента корреляции:

```
> cor.test(neig$X5,neig$X6)

        Pearson's product-moment correlation

data:  neig$X5 and neig$X6
t = 2.8942, df = 42, p-value = 0.006005
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1261612 0.6285569
sample estimates:
      cor
0.4077686
```

В этом тесте нулевая гипотеза означает, незначимость коэффициента корреляции. Здесь получается, что нулевая гипотеза отклоняется при уровне значимости 0.05, так как *p*-value равно 0.006005.

Уравнение множественной линейной регрессии

y: зависимая переменная,
$x = (x_1, \ldots, x_k)^\mathsf{T}$: вектор независимых переменных, которые будут использованы для предсказания y.

Модель:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n,$$

где ε_i — ненаблюдаемая случайная величина.

Предположения модели:

- $E\varepsilon_i = 0, \quad i = 1, \ldots, n,$
- $E(\varepsilon_j \varepsilon_\ell) = 0, \quad j \neq \ell,$
- $E\varepsilon_i^2 = \sigma^2, \quad i = 1, \ldots n,$
- $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\mathsf{T} \sim N(0, \sigma^2 I_n).$

Множественная линейная регрессия

- Таким образом, переменные x_{i1}, x_{i2}, ..., x_{ik} представляют собой набор k независимых переменных, которые, как считается, влияют на y, и оценки коэффициентов β_1, β_2, ..., β_k — параметры, которые количественно определяют эффект каждого из этих объясняющие переменные на y. Каждый коэффициент теперь известен как частичный коэффициент регрессии (определяет влияние данной объясняющей переменной на объясненную переменную после сохранение постоянными или исключение влияния всех других объясняющих переменных).
- Например, β_2 измеряет влияние x_2 на y после устранения эффектов x_1, x_3, x_4, ..., x_k. Другими словами, каждый коэффициент измеряет среднее изменение зависимой переменной на единицу изменения в данной независимой переменной, сохраняя все остальные независимые переменные постоянными при их средних значениях.

Матричная форма записи

Запишем уравнение в матричном виде:

$$y = X\beta + \varepsilon$$

where

- y — вектор размерности $n \times 1$;
- X — матрица размерности $n \times (k+1)$;
- β вектор размерности $(k+1) \times 1$;
- ε — вектор размерности $n \times 1$.

Линейная регрессия

Запишем модель в матричной форме:

$$Y = X\beta + \varepsilon,$$

где $Y = (y_1, \dots, y_n)^\intercal$, $\beta = (\beta_0, \beta_1, \dots, \beta_k)^\intercal$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\intercal$,

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}.$$

Будем решать следующую задачу минимизации:

$$\min_{\beta_0, \beta_1, \dots, \beta_k} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_k x_{ik})^2.$$

MLE или МНК оценка: $\hat{\beta} = (X^T X)^{-1} X^T Y$.

Полученное решение: $\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$.

Свойства решения

Обозначим $\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY$,

$$S^2 = \frac{\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1} = \frac{(Y-\hat{Y})^\intercal(Y-\hat{Y})}{n-k-1}$$

1. $$\frac{\hat{\beta}_j - \beta_j}{S\sqrt{(X^TX)^{-1}_{j+1,j+1}}} \sim T_{n-k-1}, \quad j = 0, \ldots, k.$$

   Доверительный интервал:

$$\left(\hat{\beta}_j - t_{1-\frac{\alpha}{2},n-k-1}S\sqrt{(X^TX)^{-1}_{j+1,j+1}};\right.$$

$$\left.\hat{\beta}_j + t_{1-\frac{\alpha}{2},n-k-1}S\sqrt{(X^TX)^{-1}_{j+1,j+1}}\right),$$

Свойства

Обозначим

$$R^2 = \frac{(\hat{Y} - \bar{y}\mathbf{1})^T(\hat{Y} - \bar{y}\mathbf{1})}{(Y - \bar{y}\mathbf{1})^T(Y - \bar{y}\mathbf{1})} = 1 - \frac{(Y - \hat{Y})^\intercal(Y - \hat{Y})}{(Y - \bar{y}\mathbf{1})^T(Y - \bar{y}\mathbf{1})}.$$

2. $F = \frac{R^2}{1 - R^2} \cdot \frac{n - k - 1}{k} \sim \mathcal{F}_{k,n-k-1}.$

Построение линейной регрессии в R

Функция lm в R

Функция lm

```
lm(formula, data, weights, na.action, method = "qr" , subset,
model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok =
TRUE, contrasts = NULL, offset, ...)
```

Для вывода всей информации по построенному уравнению необходимо использовать следующую функцию:

Функция summary

```
summary(object)
```

Для просмотра графиков попарной зависимости можно использовать следующую функцию:

Функция pairs

```
pairs(formula, data = NULL, ..., subset, na.action =
stats::na.pass)
```

Arguments (lm)

formula: объект класса "formula": описание устанавливаемой модели. Детали спецификации модели приведены в ‘Details’.

subset: необязательный вектор, определяющий подмножество наблюдений, которые будут использоваться в процессе подбора.

weights: необязательный вектор весов, который будет использоваться в процессе подгонки (МНК). Может равняться NULL или числовому вектору. Если не NULL, то МНК использует эти веса (т.е. минимизируется $sum(w \cdot e^2)$); в противном случае используются единичные веса.

Найдем оценки МНК уравнения линейной регрессии:

```
> lm(X6~X1+X2+X3+X4+X5+X7, neig)

Call:
lm(formula = X6 ~ X1 + X2 + X3 + X4 + X5 + X7, data = neig)

Coefficients:
(Intercept)           X1           X2           X3           X4
    67.5618      -5.2679       1.2279      -7.0990       2.8437
         X5           X7
     4.5877       0.5862
```

```
> m1<-lm(X6~X1+X2+X3+X4+X5+X7, neig)
> summary(m1)

Call:
lm(formula = X6 ~ X1 + X2 + X3 + X4 + X5 + X7, data = neig)

Residuals:
     Min       1Q   Median       3Q      Max 
-117.010  -37.285   -2.292   28.046  241.633 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  67.5618    58.8035   1.149   0.2580    
X1           -5.2679     3.1061  -1.696   0.0983 .  
X2            1.2279     0.9558   1.285   0.2069    
X3           -7.0990     1.6171  -4.390 9.11e-05 ***
X4            2.8437     0.5304   5.361 4.60e-06 ***
X5            4.5877     1.6586   2.766   0.0088 ** 
X7            0.5862     0.7550   0.776   0.4424    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 66.18 on 37 degrees of freedom
Multiple R-squared:  0.658,	Adjusted R-squared:  0.6025 
F-statistic: 11.86 on 6 and 37 DF,  p-value: 2.187e-07
```

Интерпретация результатов

- Квантили уровней 0.00, 0.25, 0.5, 0.75, 1 остатков.
- Оценки коэффициентов β_0, ..., β_k, их стандартные ошибки $\hat{\beta}_1/t_{\beta_1}$, ..., $\hat{\beta}_k/t_{\beta_k}$, значения статистик t_{β_1}, ..., t_{β_k} для проверки гипотез типа $H_0 : \beta_i = 0,\ i = 1, \ldots, k$. Эта гипотеза говорит о незначимости коэффициента β_i.
- В столбце $Pr(>|t|)$ записываются p-values для этих гипотез. Если $p - value > \alpha$ (по умолчанию, $\alpha = 0.05$), тогда $H_0 : \beta_i = 0$ принимается, и коэффициент принимается незначимым. В противном случае, нулевая гипотеза отвергается и коэффициент признается значимым. В примере, значимыми признаются $X3$, $X4$, $X5$. Значение `Residual standard error` — это статистика S.
- `Multiple R-squared` — значение R^2. статистика F для проверки гипотезы $H_0 : \beta_1 = \ldots = \beta_k = 0$ равна 11.86. Значение p-value равно 2.187e-07, что меньше 0.05. Следовательно, нулевая гипотеза отвергается и мы можем утверждать, что построенное уравнение регрессии значимо в целом.

О построенной модели

- `coef(model)` выдает коэффициенты построенного уравнения.
- `fitted(model)` дает оценки $\hat{y}$.
- `summary(model)` дает "summary" модели.
- `confint(model,"variable")` выдает доверительный интервал для указанной переменной.
- `anova(model)` предоставляет результаты анализа ANOVA.
- `resid(model)` выдает остатки по каждому наблюдению.

***	p-value в интервале от 0 до 0.001
**	p-value в интервале от 0.001 до 0.01
*	p-value в интервале от 0.01 до 0.05
.	p-value в интервале от 0.05 до 0.1
(пусто)	p-value в интервале от 0.1 до 1.0

```
> coef(m1)
(Intercept)          x1          x2          x3 
 67.5618471  -5.2678778   1.2278735  -7.0989590 
         x4          x5          x7 
  2.8436595   4.5877261   0.5862316 
> fitted(m1)
         1          2          3          4          5 
101.740884 194.287666 169.539344  26.719347  69.128420 
         6          7          8          9         10 
201.474124 462.467156 142.492204 102.206063  69.740061 
        11         12         13         14         15 
143.012224 167.763761 167.162139 172.481450  52.326685 
        16         17         18         19         20 
 54.833597  49.558103 148.905956 114.653005   4.308262 
        21         22         23         24         25 
 96.594265  94.085811 111.872473 -16.075254  56.757503 
        26         27         28         29         30 
119.117692 156.790284  64.067670  28.154302 -16.519456 
        31         32         33         34         35 
 80.133015 324.709778  44.723533 107.461200 107.322153 
        36         37         38         39         40 
152.813801  92.905214 108.534999 180.271642  69.499926 
        41         42         43         44 
 54.242814  23.550403  88.530418  35.255366 
```

```
> confint(m1,c(1,2,3,4,5,6,7))
                  2.5 %     97.5 %
(Intercept) -51.5854350 186.709129
x1          -11.5613635   1.025608
x2           -0.7088486   3.164596
x3          -10.3754152  -3.822503
x4            1.7689196   3.918399
x5            1.2271073   7.948345
x7           -0.9435052   2.115968
> anova(m1)
Analysis of Variance Table

Response: x6
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1  61486   61486 14.0394 0.0006095 ***
x2         1  19407   19407  4.4313 0.0421361 *
x3         1  30561   30561  6.9780 0.0120193 *
x4         1 165490  165490 37.7870  3.97e-07 ***
x5         1  32178   32178  7.3474 0.0101210 *
x7         1   2641    2641  0.6029 0.4423983
Residuals 37 162043    4380
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> resid(m1)
          1           2           3           4 
-16.84088351 -21.68766639 -15.33934365   8.48065331 
          5           6           7           8 
  0.07157952 -90.47412422 241.63284360 -72.59220357 
          9          10          11          12 
-36.80606261  62.35993924  36.88777641 -27.86376051 
         13          14          15          16 
-58.46213949 -49.28144953  52.37331541   6.66640295 
         17          18          19          20 
 18.64189747 -52.00595551 143.34699523  27.69173787 
         21          22          23          24 
 30.40573463 -66.98581143 -41.17247264  54.37525373 
         25          26          27          28 
 -2.75750253 -17.61769163  29.10971566  -2.86767017 
         29          30          31          32 
 10.44569797  69.11945636 -17.53301463 -117.00977755 
         33          34          35          36 
 -2.32353348  -2.26120038 -38.72215288   4.48619926 
         37          38          39          40 
-34.40521418 -45.43499902 -93.87164158   8.00007393 
         41          42          43          44 
  9.25718564  45.34959655  14.26958233  51.34463402 
```

Важные наблюдения в выборке

Все наблюдения влияют на регрессионную модель, пусть и незначительно. Когда аналитик говорит, что наблюдение имеет большое значение, это означает, что его удаление значительно изменит подобранную модель регрессии. Мы хотим идентифицировать эти наблюдения, потому что они могут быть выбросами, искажающими нашу модель; мы обязаны исследовать их детально.

Функция `influence.measures` сообщает несколько значений: DFBETAS, DFFITS, коэффициент ковариации, расстояние Кука. Если какая-либо из этих мер указывает на то, что наблюдение является важным, функция отмечает это наблюдение звездочкой (*) справа в таблице вывода.

Функция в R

```
influence.measures(m1)
```

Использование модели линейной регрессии для предсказаний

predict

```
predict(object, newdata, se.fit = FALSE, scale = NULL, df = Inf,
interval = c("none" , "confidence" , "prediction"), level = 0.95,
type = c("response" , "terms"), terms = NULL, na.action =
na.pass, pred.var = res.var/weights, weights = 1, ...)
```

Выбор наилучшей модели

Коэффициент AIC для выбора «лучшей» модели

Информационный критерий Акаике (AIC) — критерий, применяющийся исключительно для выбора из нескольких статистических моделей (используется для сравнения моделей). Разработан в 1971 как "an information criterion" Хироцугу Акаике и предложен им в статье 1974 года.

Предпосылкой к созданию критерия послужила задача оценки качества предсказаний модели на тестовой выборке при известном качестве на обучающей выборке при условии, что модель мы «настраивали» по методу максимума правдоподобия.

Коэффициент AIC вычисляется по формуле:

$$AIC = 2k - 2\ln(L),$$

где k — число параметров в статистической модели, L — максимизированное значение функции правдоподобия модели.

Пошаговый выбор регрессионной модели

Функция step в R

```
step(object, scope, scale = 0, direction =
c("both "backward "forward"), trace = 1, keep = NULL, steps =
1000, k = 2, ...)
```

Больше информации по этой функции:
https://www.rdocumentation.org/packages/stats/versions/3.5.2/topics/step

- object — модель типа `lm`
- scope — определяет диапазон моделей, исследуемых при пошаговом поиске. Это должна быть либо одна формула, либо список, содержащий компоненты верхнего и нижнего аргументов, обе формулы.
- scale — параметр, используемый в вычислении коэффициента AIC.

```
> step(m1)
Start:  AIC=375.3
x6 ~ x1 + x2 + x3 + x4 + x5 + x7

       Df Sum of Sq    RSS    AIC
- x7    1      2641 164684 374.01
- x2    1      7227 169270 375.22
<none>              162043 375.30
- x1    1     12597 174640 376.60
- x5    1     33508 195551 381.57
- x3    1     84406 246449 391.75
- x4    1    125875 287918 398.59

Step:  AIC=374.01
x6 ~ x1 + x2 + x3 + x4 + x5

       Df Sum of Sq    RSS    AIC
<none>              164684 374.01
- x2    1      9351 174035 374.44
- x1    1     18060 182744 376.59
- x5    1     32178 196862 379.87
- x3    1    102079 266762 393.24
- x4    1    154310 318994 401.10

Call:
lm(formula = x6 ~ x1 + x2 + x3 + x4 + x5)

Coefficients:
(Intercept)           x1           x2           x3           x4
     58.553       -6.005        1.371       -6.374        2.614
         x5
      4.480
```

stepAIC [MASS package]

Функция выбирает лучшую модель по коэффициенту AIC. У функции есть опция "direction" : “both” (для пошаговой регрессии, как прямой, так и обратный выбор); “backward” (для выбора назад) и “forward” (для выбора вперед). Он возвращает лучшую финальную модель.

stepAIC()

```
library(MASS)
full.model <- lm(y ~., data = swiss) # строим базовую модель
# Пошаговый выбор модели регрессии
step.model <- stepAIC(full.model, direction = "both" , trace =
FALSE)
summary(step.model)
```

Информационный критерий Акаике (AIC) позволяет вам проверить, насколько хорошо построенная модель соответствует набору данных, не «перегружая» ее. Предполагается, что модель с **более низким показателем AIC** обеспечит лучший баланс между ее способностью соответствовать набору данных и способностью избегать переобучения набором данных.

```
> library(MASS)
> step.model<-stepAIC(m1, direction = "both", trace = FALSE)
> summary(step.model)

Call:
lm(formula = x6 ~ x1 + x2 + x3 + x4 + x5)

Residuals:
     Min       1Q   Median       3Q      Max 
-122.529  -41.573   -3.694   24.642  240.264 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  58.5527    57.3456   1.021  0.31369    
x1           -6.0054     2.9418  -2.041  0.04820 *  
x2            1.3706     0.9331   1.469  0.15008    
x3           -6.3740     1.3133  -4.853 2.10e-05 ***
x4            2.6141     0.4381   5.967 6.32e-07 ***
x5            4.4801     1.6441   2.725  0.00967 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 65.83 on 38 degrees of freedom
Multiple R-squared:  0.6524,	Adjusted R-squared:  0.6067 
F-statistic: 14.27 on 5 and 38 DF,  p-value: 7.16e-08
```

Выбор модели с использованием AIC

1. Базовая модель имеет коэффициент AIC=375.3.
2. Посмотрим на последний столбец в таблице. Он показывает, каким станет AIC, если мы удалим ту или иную переменную. В нашем случае, удаление $X7$ приведет к значению AIC 374.01.
3. На следующем этапе удаляется $X7$. Полученное значение AIC является наименьшим. Если удалить еще какую-то переменную, то значение AIC только увеличится.
4. Полученная на втором этапе модель — наилучшая с точки зрения минимизации AIC.
5. На следующем этапе следует проверить новую модель на значимость коэффициентов и регрессии в целом.

План построения регрессионной модели в R (часть 1)

1. Провести корреляционный анализ имеющихся данных (используем функции `ggpairs`, `ggcorr`).
2. Построить базовую модель линейной регрессии (функция `lm`).
3. Вывести результаты анализа базовой модели (функции `lm` и `summary`).
4. Записать уравнение линейной регрессии (функция `coef`).
5. Проверить значимость каждого отдельного коэффициента с помощью T-test. P-values для данного критерия выводятся в Таблице `summary` в столбце $Pr(>|t|)$.
6. Проверить значимость построенного уравнения регрессии с помощью F-test. Значение статистики теста и соответствующее p-value выводятся в результате работы функции `summary`.
7. Построить график рассеяния и уравнения регрессии `pairs`.
8. Построить доверительные интервалы для коэффициентов регрессии с помощью функции `confint`.

9. Используя функцию `anova`, проверить значимость парных линейных регрессий с помощью F-теста.
10. В случае подозрения на наличие выбросов, проверить так называемые важные наблюдения, которые значительно влияют на построение модели (функция `influence.measures(m1)`)
11. Используя функцию `step` или `stepAIC`, постараться улучшить модель.
12. В случае получения в предыдущем пункте модели, отличной от базовой, повторить пп. 3–9 для новой модели.

Итоги

Что мы узнали на Лекции 5?

- Как проверить значимость корреляции переменных.
- Что такое уравнение линейной регрессии.
- Как построить модель линейной регрессии в R.
- Как проверить значимость коэффициентов построенной модели, а также значимость уравнения в целом.
- Как выбрать лучшую модель и сделать это автоматически в R.

Что мы узнаем на Лекции 6?

- Какие предположения при построении уравнения линейной регрессии нужно проверять.
- Какие тесты существуют для диагностики модели и как их использовать в R.
- Как составить комплексный отчет о построенной модели линейной регрессии.

.

Спасибо за внимание и до встречи на Лекции 6!