**Липецкий государственный технический университет**

Факультет автоматизации и информатики

Кафедра автоматизированных систем управления

# ЛАБОРАТОРНАЯ РАБОТА №2

# СТАТИСТИЧЕСКИЕ МЕТОДЫ В ПРИКЛАДНЫХ ЗАДАЧАХ

# РЕГРЕССИОННЫЙ АНАЛИЗ

Студент _____ Станиславчук С.М.
подпись
Коновалов К. А.

Группа АС-21-1

Руководитель

Доцент _____ Рыжкова Д.В.
подпись

Липецк 2023 г.

Цель работы

На реальных данных (массив данных - независимая X и зависимая Y, 86 строк), несущих смысловую нагрузку, провести регрессионный анализ, а именно:

1. Проверить выборку данных на нормальность.
2. Вывести уравнение линейной регрессии $y=\alpha+\beta x$, оценки коэффициентов получить с помощью методов: МНК, метод Бартлетта-Кенуя.
3. Оценка статистической значимости выборочной регрессии.
4. Определение доверительных областей, включающих в истинную регрессию с заданной вероятностью.
5. Анализ регрессионных остатков.
6. Анализ наличия грубых отклонений от регрессии (выбросов).
7. Построение толерантных границ для регрессии.

Краткая теоретическая справка

Регрессионный анализ – статистический метод исследования влияния одной или нескольких независимых переменных X на зависимую переменную Y.

Линейный регрессионный анализ исходит из наличия зависимости $y=\alpha+\beta x$, где $\alpha$ и $\beta$ – неизвестные коэффициенты регрессии.

Ход выполнения лабораторной работы

Был сформирован массив из реальных данных. Y – % людей пенсионного возраста, а X – потребление алкогольных напитков на душу населения, литр. Данные представлены на рис. 1.

| | Города и субъ-екты РФ | Потребление алкогольных напитков на душу населения (15+ лет), литр, X | % людей пенсионного возраста, Y |
|---|---|---|---|
| 1 | | | |
| 2 | Алтайский край | 5,8 | 0,36 |
| 3 | Амурская область | 10,9 | 0,28 |
| 4 | Архангельская область | 12,7 | 0,39 |
| 5 | Астраханская область | 6,1 | 0,26 |
| 6 | Белгородская область | 5,3 | 0,33 |
| 7 | Брянская область | 7,1 | 0,35 |
| 8 | Владимирская область | 10,6 | 0,32 |
| 9 | Волгоградская область | 5 | 0,29 |
| 10 | Вологодская область | 13,7 | 0,32 |
| 11 | Воронежская область | 6,3 | 0,31 |
| 12 | рейская автономная облас | 11,8 | 0,29 |
| 13 | Забайкальский край | 7,4 | 0,27 |
| 14 | Ивановская область | 8,8 | 0,33 |
| 15 | Иркутская область | 8,7 | 0,31 |
| 16 | рдино-Балкарская Респуб | 1,2 | 0,22 |
| 17 | Калининградская область | 8,8 | 0,27 |
| 18 | Калужская область | 10,2 | 0,28 |
| 19 | Камчатский край | 13,7 | 0,30 |
| 20 | чаево-Черкесская Респуб | 1,9 | 0,26 |
| 21 | Кемеровская область | 7,7 | 0,31 |
| 22 | Кировская область | 10,6 | 0,36 |
| 23 | Костромская область | 10,5 | 0,35 |
| 24 | Краснодарский край | 6,1 | 0,28 |
| 25 | Красноярский край | 8 | 0,29 |
| 26 | Курганская область | 5,7 | 0,36 |
| 27 | Курская область | 5,9 | 0,35 |
| 28 | Ленинградская область | 12,4 | 0,24 |
| 29 | Липецкая область | 6,6 | 0,33 |
| 30 | Магаданская область | 16,9 | 0,30 |

Рисунок 1 – Исходные данные

1. Проверка выборки данных на нормальность.

Для проверки используем критерий «хи»-квадрат. Расчет по формуле:

$$\chi^2_{\text{экс}} = \sum_{i=1}^{M} \frac{\left(n_i - n_i^{\text{теор}}\right)^2}{n_i^{\text{теор}}}$$

$$n_i^{\text{теор}} = n[\Phi(x_i^{\text{кон}}) - \Phi(x_i^{\text{нач}})]$$

$$\Phi(x) = \int_{-\infty}^{z} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\bar{x})^2}{2\sigma^2}\right) dx$$

Проверка на нормальность осуществлялась для x и y. Для начала были определены: $x_{ср}$, $y_{ср}$, количество интервалов M, минимальные/максимальные значения переменных, а также шаг Δ/M, изображенные на рис. 2.

| | | |
|---|---|---|
| минимальное | 0,30 | 0,20 |
| максимальное | 17,60 | 0,48 |
| дисперсия | 13,14304388 | 0,002067312 |
| среднее | 8,38 | 0,30 |
| выборочное ско | 3,625333623 | 0,045467697 |
| ширина интервала | 2,471428571 | 0,039074331 |
| кол-во интервалов | 7 | |
| ХИ^2 крит | 64,74939583 | |
| ХИ^2 | 1527,576311 | 274234223,7 |
| тк ХИ^2>ХИ^2крит, то распределение нормальное | | |

Рисунок 2 – Поясняющие данные для X и Y

На рис. 3 представлены расчеты для X. $\chi^2_{\text{экс}}$ был посчитан для всех значений с помощью суммирования полученных результатов:

$\chi^2_{\text{экс}}$ = 1527,576311.

| E нач | E кон | n | n теор | n*n теор | (n-(n*n теор))^2 | ХИ^2 | ХИ^2 |
|---|---|---|---|---|---|---|---|
| -0,23 | -0,15 | 32 | 0,0286259 | 0,916027915 | 966,2133206 | 1054,785891 | 191340,4273 |
| -0,15 | -0,07 | 39 | 0,1663474 | 6,487550242 | 1057,059389 | 162,9366016 | |
| -0,07 | 0,01 | 13 | 0,3695879 | 4,804642956 | 67,16387707 | 13,97895279 | |
| 0,01 | 0,09 | 6 | 0,3160359 | 1,896215457 | 16,84104757 | 8,881399793 | |
| 0,09 | 0,17 | 2 | 0,1038674 | 0,207734739 | 3,212214766 | 15,46306015 | |
| 0,17 | 0,25 | 0 | 0,0130057 | 0 | 0 | | |
| 0,25 | 0,34 | 2 | 0,0006125 | 0,001224996 | 3,995101516 | 3261,317575 | |
| 0,34 | 0,42 | 2 | 0,0000107 | 2,14102E-05 | 3,99991436 | 186823,0638 | |

Рисунок 3 – Проверка на нормальность для Y

$\chi^2_{\text{экс}}$ для Y составил 274234223,7.

$\chi^2_{крит}$ подсчитывалось по встроенной формуле =ХИ2.ОБР(). Так как $\chi^2_{крит}$ зависит от α(0.05) и степеней свободы, то для X и Y этот параметр будет одинаков. $\chi^2_{крит}$ = 64,74939583.

Для X: 1527,576311 > 64,74939583 и для Y: 274234223,7 > 64,74939583

Вывод: выборка X и выборка Y подчиняются нормальному закону распределения.

2. Вывод уравнения линейной регрессии

Линейное парное уравнение регрессии имеет вид: $y=\alpha+\beta*x_i$, i=1, …,n , где n – объем совокупности (число наблюдений).

Оценки параметров линейной регрессии (a и b) могут быть найдены разными методами, наиболее распространенным является метод наименьших квадратов. Данный метод позволяет получить такие оценки параметров a и b, при которых сумма квадратов отклонений фактических значений результативного признака $y_i$ от расчетных (теоретических) значений $y_i^*$ (рассчитанных по уравнению регрессии) минимальна.

Формулы для расчета:

$$b = \frac{\overline{X \cdot Y} - \bar{X} \cdot \bar{Y}}{\sigma_x^2}; a = \bar{Y} - b \cdot \bar{X}$$

| | | | |
|---|---|---|---|
| 104 | по матрицам | 10 | 712,10 |
| 105 | | 712,10 | 7082,88 |
| 106 | | | |
| 107 | | -0,016235548 | 0,001632292 |
| 108 | | 0,001632292 | -0,000022922 |
| 109 | | | |
| 110 | b | -0,060389091 | |
| 111 | k | 0,037061947 | |
| 112 | МНК | | |
| 113 | b | 0,003549872 | |
| 114 | a | 0,27 | |

Рисунок 4 – МНК

Пары наблюдений ($y_i$, $x_i$) упорядочиваются по x и разбиваются на 3 примерно равные группы, так чтобы первая и третья группа были обязательно разного объема.

В каждой группы находятся суммы $\sum_{i=1}^{n} y_i$ и $\sum_{i=1}^{n} x_i$.

Тогда коэффициенты регрессии находятся с помощью соотношений:

$$\tilde{b} = \frac{Y_3 - Y_1}{X_3 - X_1}, \quad \tilde{a} = \bar{y} - \tilde{b}\bar{x}.$$

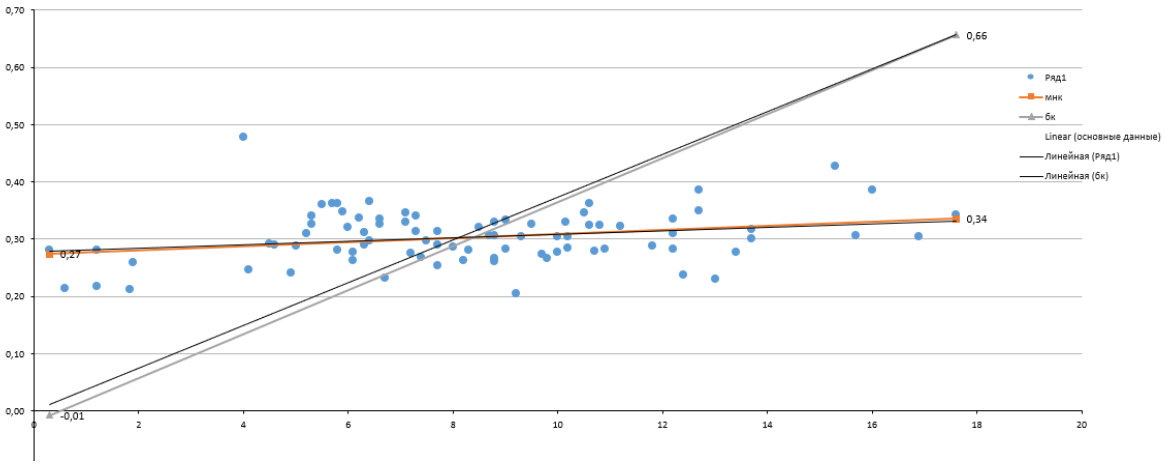


Рисунок 5 – График

Был построен график (рис. 5), на котором отображен разброс начальных данных, и прямые, полученные с помощью МНК и метода Бартлетта-Кенуя.

Как видно на рис. 5, линия тренда для X,Y совпала с прямой МНК, а значит значения были получены верно.

3. Оценка статистической значимости выборочной регрессии

Статистические выводы относительно коэффициента β регрессии y=α+βx получаются с помощью статистики:

$$S_\beta = \frac{S}{S_x\sqrt{n-1}},\ S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - a - bx_i)^2,\ S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Значение коэффициента β регрессии является значимым, если выполняется $|b| > t_{\frac{1+\alpha}{2}} S_\beta$.

Для коэффициента α используется статистика: $S_\alpha = S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2}}$.

| | | | | | | |
|---|---|---|---|---|---|---|
| 129 | | | | | | |
| 130 | S^2 | 0,001985805 | | S | 0,044562371 | |
| 131 | S^2_X | 13,14304388 | | S_X | 3,625333623 | |
| 132 | S_betta | 0,001333248 | | | | |
| 133 | S_alpha | 0,012159277 | | | | |
| 134 | \|B\| | 0,003549872 | 0,000693129 | => | Значения A и B являются значимыми | |
| 135 | \|A\| | 0,272937761 | 0,006321367 | | | |
| 136 | | | | | | |
| 137 | | | | | | |
| 138 | t((1+alpha)/2) | 0,519880152 | | | | |
| 139 | t*S_b | 0,000693129 | | | | |
| 140 | t*S_a | 0,006321367 | | | | |

Рисунок 6 – Оценка значимости

Исходя из расчетов (рис. 6), можно сделать вывод о том, что коэффициенты a и b являются значимыми (t*S_b < |B| и t*S_a < |A|)

4. Определение доверительных интервалов

Двусторонний доверительный интервал для β имеет вид:

$$b - t_{\frac{1+\alpha}{2}} S_\beta \le \beta \le b + t_{\frac{1+\alpha}{2}} S_\beta.$$

Двусторонний доверительный интервал для α имеет вид:

$$a - t_{\frac{1+\alpha}{2}} S_\alpha \le \alpha \le a + t_{\frac{1+\alpha}{2}} S_\alpha.$$

| | | | |
|---|---|---|---|
| 2 | Доверительный интервал | | |
| 3 | 0,002856743 | beta | 0,004243001 |
| 4 | 0,26661639460 | alpha | 0,279259128 |
| 5 | | | |

Рисунок 7 – Определение доверительных интервалов для значений a и b.

5. Анализ регрессионных остатков

Значительную информацию об адекватности уравнения регрессии дает анализ ее остатков. Остатки вычисляются по формуле $e_i = y_i - \hat{y}_i$, где $\hat{y}_i = a + bx_i$.

Если уравнение регрессии хорошо описывает исходные данные, то остатки должны быть распределены по нормальному закону. Поэтому исследование остатков должно включать в себя проверку гипотезы о нормальности распределения.

Независимость остатков может быть проверена с помощью статистики Дарбина-Уотсона:

$$D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}.$$

| Ei^2 | Ei-Ei-1 | (Ei-Ei-1)^2 |
|---|---|---|
| 0,004830089 | -0,10 | 0,009640433 |
| 0,000822936 | 0,10 | 0,009239432 |
| 0,004547492 | -0,10 | 0,009738191 |
| 0,000976386 | 0,07 | 0,004233188 |
| 0,001143506 | 0,01 | 0,000185359 |
| 0,002249645 | -0,03 | 0,001175912 |
| 0,000172629 | -0,02 | 0,000260307 |
| 8,97114E-06 | 0,00 | 2,48002E-06 |
| 2,08848E-05 | 0,02 | 0,000409381 |
| 0,000245335 | -0,04 | 0,001825822 |
| 0,000732594 | 0,00 | 1,49876E-05 |
| 0,000957151 | 0,06 | 0,003132258 |
| 0,000626437 | -0,02 | 0,000541603 |
| 3,08481E-06 | -0,06 | 0,003905128 |
| 0,003688699 | 0,02 | 0,000533804 |
| 0,001416052 | 0,01 | 0,000172683 |
| 0,00059974 | 0,00 | 1,90846E-05 |
| 0,000404854 | 0,00 | 8,30964E-08 |
| 0,000416538 | 0,03 | 0,001151482 |
| 0,000182905 | 0,04 | 0,001453996 |
| 0,002668294 | -0,02 | 0,000244521 |
| 0,001297323 | -0,05 | 0,002804606 |
| 0,000286969 | 0,00 | 3,37744E-06 |
| 0,000228082 | 0,08 | 0,007039751 |
| 0,004733557 | -0,01 | 0,000211412 |
| 0,00294424 | -0,13 | 0,017899727 |
| 0,006324863 | 0,11 | 0,011786559 |
| 0,000843142 | 0,00 | 1,12011E-06 |
| 0,000905725 | -0,09 | 0,008971787 |
| 0,004176294 | -0,02 | 0,000616086 |
| 0,008000467 | 0,12 | 0,014764697 |
| 0,001028148 | -0,01 | 0,00019975 |
| 0,000321537 | 0,00 | 1,33126E-08 |
| 0,000317412 | 0,00 | 3,62782E-06 |
| 0,000388908 | -0,03 | 0,000674093 |
| 3,89695E-05 | 0,02 | 0,000592431 |
| 0,000327514 | 0,01 | 5,71228E-05 |
| 0,000658195 | 0,04 | 0,00190048 |
| 0,004795537 | -0,03 | 0,000902487 |
| 0,001537299 | -0,04 | 0,001725222 |
| 5,41675E-06 | -0,03 | 0,000920417 |
| 0,001067053 | 0,03 | 0,000968931 |
| 2,36579E-06 | -0,05 | 0,002283628 |
| 0,002432998 | 0,03 | 0,000771291 |
| 0,000464545 | 0,00 | 1,54652E-06 |
| 0,000519698 | -0,02 | 0,000439389 |
| 0,001914807 | -0,02 | 0,000582536 |

Рисунок 8 – Анализ регрессионных остатков

Отдельно посчитаны $e_i$, $e_i^2$, $(e_i - e_{i-1})^2$.

| | | |
|---|---|---|
| DW= | 1,754383212 | |
| p | 0,122808394 | ->0, => null |

Рисунок 9.

На рис. 9 видно, что выполняется условие DW -> 2, p->0, это говорит о том, что автокорреляция отсутствует.

| Z score(останки) <= | X останков | Дисперсия останков | Стандартное отклонение выборки останков | |
|---|---|---|---|---|
| 1,5403 | 0,0010 | 0,001975246 | 0,044443741 | |
| -0,6689 | | | | |
| 1,4938 | A^2 | p | | |
| -0,7266 | 0,4508 | 0,2749 | >=0,05 , | => нет достаточных доказательств, чтобы утверждать, что данные выборки значительно отклоняются от нормальности, поэтому мы не отвергаем нормальность. |
| 0,7374 | | | | |
| 1,0437 | | | | |
| 0,2721 | | | | |
| -0,0909 | | | | |
| -0,1263 | | | | |
| 0,3289 | | | | |
| -0,6325 | | | | |
| -0,7196 | | | | |
| 0,5397 | | | | |
| 0,0160 | | | | |

Рисунок 10. Проверка гипотезы нормальности – Андерсон-Дарлинг

Для проверки на нормальность так же использовался критерий «хи»-квадрат, который показал, что выборка подчиняется нормальному закону распределения.

6. Анализ наличия грубых отклонений от регрессии (выбросов)

Способ - межквартильный диапазон.

Межквартильный размах (IQR) — это разница между 75-м процентилем (Q3) и 25-м процентилем (Q1) в наборе данных. Он измеряет разброс средних 50% значений.

| Останки | | |
|---|---|---|
| 0,0695 | 0 | 0,0543 |
| -0,0287 | 0 | |
| 0,0674 | 0 | |
| -0,0312 | 0 | |
| 0,0338 | 0 | |
| 0,0474 | 0 | |
| 0,0131 | 0 | |
| -0,0030 | 0 | |
| -0,0046 | 0 | |
| 0,0157 | 0 | |
| -0,0271 | 0 | |
| -0,0309 | 0 | |
| 0,0250 | 0 | |
| 0,0018 | 0 | |
| -0,0607 | 0 | |
| -0,0376 | 0 | |
| -0,0245 | 0 | |
| -0,0201 | 0 | |
| -0,0204 | 0 | |
| 0,0135 | 0 | |
| 0,0517 | 0 | |
| 0,0360 | 0 | |
| 0,0169 | 0 | |

Рисунок 12. Анализ наличия выбросов (IQR)

Мы можем определить наблюдение как выброс, если оно в 1,5 раза превышает межквартильный размах, превышающий третий квартиль (Q3), или в 1,5 раза превышает межквартильный размах меньше, чем первый квартиль (Q1)

В нашем случае выбросов среди остатков не обнаружено.

7. Построение толерантных границ для регрессии

Для того, чтобы найти толерантный интервал, для которого можно утверждать с вероятностью α, что внутри него лежит β часть значений, необходимо найти доверительные интервалы для каждого значения икса. Искомая толерантная область будет геометрическим местом точек толерантных интервалов для отдельных $x_i$.

Двусторонний толерантный интервал имеет вид:

$$\hat{y}(x) - k(n^*(x), \alpha, \beta) S(x) \le \hat{y}(x) \le \hat{y}(x) + k(n^*(x), \alpha, \beta) S(x)$$, где

$$n^*(x) = \frac{n \sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 - n(x - \bar{x})^2},$$

$$S^2(x) = S_{\hat{y}}^2 \left( \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right), \qquad S_{\hat{y}}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (\hat{y}_i - a - b x_i)^2.$$

Показанные на рис. 11 и найденные по приведенным выше формулам толерантные границы для регрессии были построены

| | | |
|---|---|---|
| Толерантный интервал регрессии | | |
| | Уровень доверия | 0,95 |
| | Доля "населения" на интервале | 0,99 |
| | | |
| | Размер выборки | 85 |
| | Среднее выборочное | 0,0011 |
| | Выборочное стандартное отклонение | 0,044443741 |
| | Z критическое (0,95) | 1,959963985 |
| | X^2 крит | 56,81298069 |
| | | |
| Интервал | Нижняя граница: | -0,1054844 |
| | Верхняя граница: | 0,10759643 |

Рисунок 9 – Построение толерантных границ

Выводы

В ходе лабораторной работы был проведен регрессионный анализ и получены следующие результаты:

1. Выборки данных X и Y подчиняются нормальному закону распределения.

($\chi^2_{экс}$ (X)) 1527,576311 > 64,74939583 ($\chi^2_{крит}$)

($\chi^2_{экс}$ (Y)) 274234223,7 > 64,74939583 ($\chi^2_{крит}$)

=> мы можем применять множество разных методов, рассчитанных на нормальность распределения данных.

2. Получено уравнение линейной регрессии.

МНК: y = 0,27 + 0,003549872x;

Метод Бартлетта-Кенуя: y= -0,02 - 0,038424653x

Исходя из полученного уравнения, видим некоторую положительную зависимость, между X и Y => употребление алкоголя довольно-таки распространено именно среди пенсионеров.

3. Проведена проверка статистической значимости выборочной регрессии. Коэффициенты a и b признаны значимыми, т.к. $t*S_b < |B|$ и $t*S_a < |A|$

=> между X и Y есть функциональная связь

4. Определены доверительные интервалы для коэффициентов α и β.

$0,002856743 < \beta < 0,004243001$

$0,26661639460 < \alpha < 0,279259128$

=> с вероятностью 95% доверительный интервал будет содержать наши значения коэффициентов.

*коэффициенты α и β в нашем случае попали в этот интервал.

5. Проведен анализ регрессионных остатков. Принята гипотеза о отсутствии автокорреляции между остатками (т.е. значения предыдущих остатков никак не влияют на значения последующих), а также проверка на нормальность показала, что выборка подчиняется нормальному закону распределения.

6. Проведен анализ наличия грубых отклонений от регрессии. Выявлено, что выбросов нет => данные подобраны корректно, т.к. находятся близко друг к другу.

7. Построение толерантных границ выполнено успешно: все значения линии регрессии попадают в толерантные интервалы.