Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Высшая школа прикладной математики и вычислительной физики

Многомерный статистический анализ

Отчет по лабораторной работе

Тема: Регрессионный анализ

Выполнил:
студент гр. 3630102/60401
Камалетдинова Ю. А.

Проверил:
к. ф-м. н., доцент
Павлова Л. В.

Санкт-Петербург
2020

Содержание

Постановка задачи

Цель лабораторной работы — восстановить зависимость между переменными с помощью модели линейной регрессии, рассчитать некоторые статистики, оценить полученную модель.

1. Ход работы

1.1. Построение модели и расчет характеристик

Пусть $X = \{x_0^m, \ldots, x_n^m\}$ — матрица независимых переменных размера (n, m), $Y = \{y_0, \ldots, y_n\}$ — вектор наблюдений. В данной задаче $n = 15, m = 3$ Рассматривается модель $Y = f(A, X) + \epsilon$, где ϵ — случайная величина, $A = \{a_0, ..., a_m\}^T$ — вектор параметров. Требуется найти оценки зависимой переменной $\hat{Y}$ и параметров $\hat{A}$ такие, что $\hat{Y} = \hat{A}X$

Найдем МНК-оценки параметров $a_0, \ldots, a_m$. Они получены по формуле 1. В таблице 1 представлены результаты минимизации.

$$\hat{A} = (X'X)^{-1}X'Y \tag{1}$$

$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
1.451	0.370	-0.003	5.711

Таблица 1: Коэффициенты линейной регрессии

Получим оценку дисперсии остатков $\hat{Y} - Y$ по формуле 2. Получено значение $\hat{s^2} \sim 3.743$.

$$\hat{s}^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n - m - 1} \tag{2}$$

Рассчитаем оценку матрицы ковариаций, которая определяется формулой 3. Результат представлен в таблице

$$\hat{cov}(\hat{A}) = \hat{s}^2(X^T X)^{-1} \tag{3}$$

	$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
$\hat{a}_0$	0.001	-0.005	0.008	0.332
$\hat{a}_1$	-0.005	0.082	-0.144	-5.358
$\hat{a}_2$	-0.008	-0.144	0.859	8.638
$\hat{a}_3$	0.332	-5.358	8.638	363.339

Таблица 2: Оценка матрицы ковариаций

Представим матрицу корреляций, элементы которой задаются формулой 4. Значения представлены в таблицу 3 .

$$cor(\hat{A})_{ij} = \frac{(X^T X)^{-1}_{ij}}{\sqrt{(X^T X)^{-1}_{ii}(X^T X)^{-1}_{jj}}} \tag{4}$$

	$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
$\hat{a}_0$	1	-0.608	-0.273	0.575
$\hat{a}_1$	-0.608	1	-0.544	-0.982
$\hat{a}_2$	-0.273	-0.544	1	0.489
$\hat{a}_3$	0.575	-0.982	0.489	1

Таблица 3: Корреляционная матрица

Построим графики $\hat{Y}$ и Y:

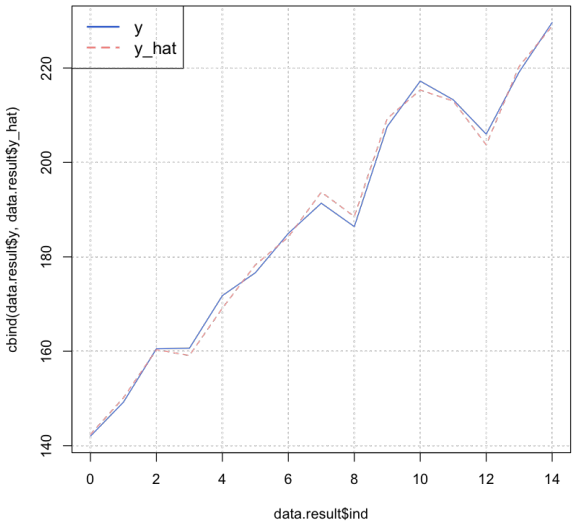

Рис. 1: Оцененные и реальные значения

Построим гистограмму остатков $\hat{Y} - Y$ для оценки адекватности регрессионной модели.

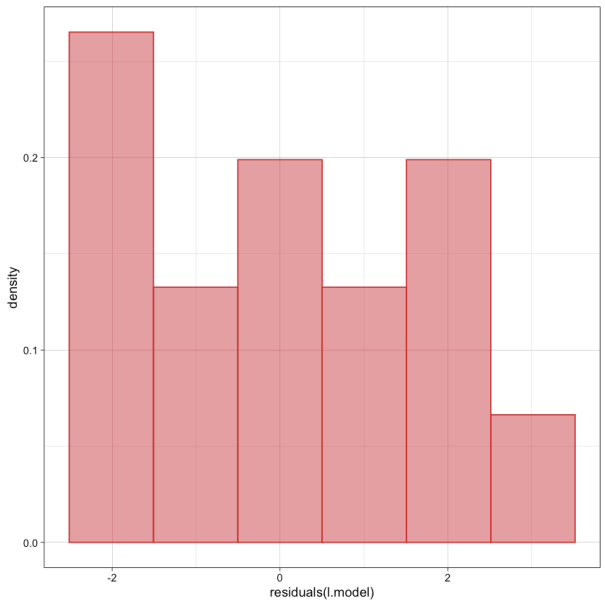

Рис. 2: Гистограмма регрессионных остатков

Вычислим оценки коэффициента детерминации и скорректированного коэффициента детерминации по формулам:

$$R^2 = \frac{\sum_i (\hat{y} - \overline{y})^2}{\sum_i (y - \overline{y})^2} \tag{5}$$

$$R^2_{adj} = \frac{R^2(n-1) - m + 1}{(n-m)}, \tag{6}$$

где n — число наблюдений, m — число параметров модели

Были получены следующие значения: $R^2 \sim 0.996$, $R^2_{adj} \sim 0.995$.

1.2. Проверка гипотез

Проверим гипотезу о равенстве нулю для близких к нулю коэффициентов регрессии, то есть $H_0 : \ a_i = 0, i = 1, 2$. Для этого рассчитаем статистику, имеющую распределение Стьюдента с $n - m = 11$ степенями свободы:

$$t = \frac{\hat{a}_i}{\hat{s}_{\hat{a}_i}} \tag{7}$$

В результате получены значения статистик, указанные в таблице 4. Квантили распределения Стьюдента уровней $1-\alpha/2$ при $\alpha = 0.05, 0.25$ — $t_{11,0.975} \sim 2.2$, $t_{11,0.875} \sim 1.214$.

	a_1	a_2
t	12.241	-0.011

Таблица 4: Значение распределения Стьюдента с 11 степенями свободы

Исходя из полученных результатов, гипотеза H_0 на уровнях значимости 0.05, 0.25 отвергается для коэффициента a_1, так как значение статистики находится за пределами $+-t_{11,0.975}, +-t_{11,0.875}$. Для коэффициента a_2 на уровнях значимости 0.05, 0.25 гипотеза не отвергается. Заключаем, что коэффициент a_2 не значим.

Перейдем к редуцированной модели без коэффициента a_2. Найдем оценки ее коэффициентов по формуле 1.

$\hat{a}_0^r$	$\hat{a}_1^r$	$\hat{a}_2^r$
1.255	0.370	5.706

Таблица 5: Коэффициенты редуцированной (r) линейной регрессии

Построим графики $\hat{Y}^r$ и Y^r:

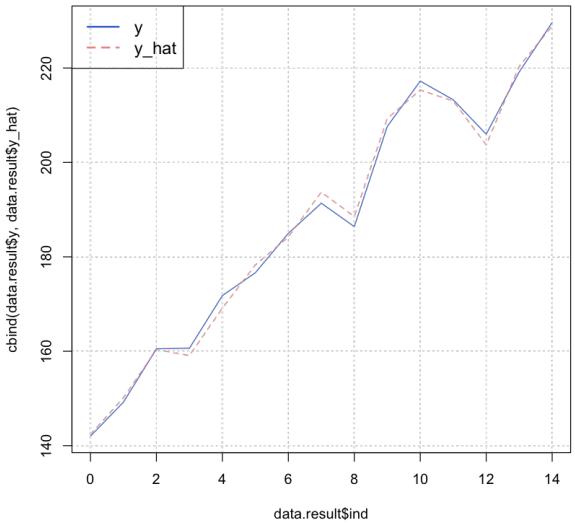

Рис. 3: Оцененные и реальные значения редуцированной (r) модели

Рассчитаем оценку дисперсии по формуле 2. Было получено значение $(\hat{s}^2)^r \sim 3.743 \sim \hat{s}^2$. Видим, что дисперсия практически не изменилась, значит, далее будем работать с редуцированной моделью. Теперь $m = 3$.

Проверим гипотезу об одновременном равенстве нулю всех коэффициентов, то есть, $H_0 : a_i^r = 0, \forall i = 0, \ldots, n$. Для этого посчитаем статистику теста:

$$F = \frac{R^2/(m-1))}{(1-R^2)(n-m)} \sim F(m-1, n-m) \tag{8}$$

Вычислим статистику для редуцированной модели. Потребуется рассчитать коэффициенты детерминации: $(R^2)^r \sim 0.996, (R^2_{adj})^r \sim 0.995$.

Критическое значения статистики $F_{0.95}(2, 12) \sim 3.885$. При расчетах получили значение $F \sim 1668.079$. Это значение гораздо больше критического, поэтому гипотеза о равенстве всех коэффициентов регресии нулю отвергается на уровне значимости 0.5.

Проверим гипотезу о равенстве коэффициентов двух регрессий, разбив имеющуюся выборку на 2 части. Для этого рассчитаем статистику:

$$F = \frac{((S - (S_1 + S_2))/m)}{(S_1 + S_2)/(n - 2m)} \sim F(m, n - 2m), \tag{9}$$

где $S,\ S_1,\ S_2$ — суммы квадратов остатков общей модели и моделей на двух подвыборках соответственно.

Значение распределения Фишера на уровне значимости 0.05 $F_{0.95}(3, 9) \sim 3.863$. При расчетах по формуле 9 было получено значение $F \sim 2.656 < F_{0.95}(3, 9)$, следовательно, гипотеза о равенстве коэффициентов двух регрессий не отвергается.

1.3. Доверительные интервалы

Построим индивидуальные доверительные интервалы для параметров линейной регрессии на уровне значимости $\alpha = 0.05$. Неравенство для доверительного интервала значений неизвестного коэффициента:

$$\hat{a}_i - t_{cr} \cdot S_{\hat{a}_i} < a_i < \hat{a}_i + t_{cr} \cdot S_{\hat{a}_i} \tag{10}$$

	a_0^r	a_1^r	a_2^r
$left$	0.320	4.099	-6.217
$right$	0.420	7.313	8.727

Таблица 6: Доверительные интервалы для коэффициентов редуцированной (r) регресии

Теперь построим совместные доверительные интервалы по принципу Тьюки

	a_0^r	a_1^r	a_2^r
$left$	0.308	3.700	-8.069
$right$	0.432	7.712	10.579

Таблица 7: Доверительные интервалы по методу Тьюки для коэффициентов редуцированной (r) регресии

Приведем доверительные интервалы, используя неравенство Чебышева:

$$P(|\hat{a}_i - a_i| \leq \tau s_i) \geq 1 - \frac{1}{\tau^2}, \tag{11}$$

где $\tau = \sqrt{20}$ для $\alpha = 0.05$

	a_0^r	a_1^r	a_2^r
$left$	0.263	2.229	-14.907
$right$	0.477	9.183	17.416

Таблица 8: Доверительные интервалы по Чебышеву для коэффициентов редуцированной (r) регресии

Заметим, что доверительные интервалы Чебышева гораздо шыре предыдущих.

1.4. Прогноз модели

Протестируем модель регрессии на случайном наблюдении, не участвовавшем в построении ее коэффициентов. Были получены следующие значения для наблюдения с индексом 14:

- $\hat{y}_{test}^r \sim 188.530, y_{test}^r \sim 186.405$
- $(\hat{s}_{test}^2)^r \sim 4.249$
- Индивидуальный доверительный интервал при $\alpha = 0.1$: $[184.828, 192.232]$

Заключение

Полученная модель линейной регрессии достаточно качественная. Это объясняется близостью коэффициента детерминации к единице, то есть, статистическая связь между зависимой Y и независимой X переменными функциональна, значения Y практически полностью определяются значениями факторов X.

В ходе работы были проверены гипотезы, удалось выяснить, что в полной модели присутствуют коэффициенты, гипотеза о равенстве нулю которых не отвергается. была получена редуцированная модель без фактора a_2. Исключение этого коэффициента не повлекло повышения дисперсии остатков, уменьшения коэффициентов детерминации, поэтому остальные расчеты приведены именно для редуцированной модели.

Наиболее узкими доверительными интервалами получились индивидуальные доверительные интервалы. Интервалы, полученные с помощью неравенства Чебышева, вышли наиболее широкими, что говорит о худшем результате. Этот метод оценки непараметрический, никаких предположений о распределении данных не требуется.

Список литературы