Санкт-Петербургский Политехнический университет
Петра Великого
Физико-механический институт
Высшая школа прикладной математики и вычислительной физики

Лабораторная работа

по дисциплине "Многомерный статистический анализ"
на тему "Линейная регрессионная модель"

Выполнил студент
гр.5030102/90401 Кунгурова Ф.А.
Руководитель:
Доцент, к.ф.-м.н. Павлова Л. В.

Санкт-Петербург
2023

Содержание

1 Постановка задачи

Требуется по заданным данным $\{y_i\}_{i=1}^n, \{x_{ij}\}_{\substack{i=\overline{1,n} \\ j=\overline{1,m}}}, m = 3$ построить регрессионную модель

$$y_t = \alpha_1 x_{t1} + \alpha_2 x_{t2} + \cdots + \alpha_{m+1} + \varepsilon_t, \ t = \overline{1,n} \tag{1}$$

Где ε_t - шум. Далее, требуется исследовать полученную модель - посчитать характеристики модели, проверить основные линейные гипотезы и построить доверительные интервалы для коэффициентов модели.

2 Обучение линейной регрессии

Модель предполагает, что между векторами регрессоров $\boldsymbol{x}_i$ и величинами y_i существует линейная зависимость, причем $\boldsymbol{x}_i$ - детерминированные величины, а ε_i - случайные шумы, т.ч.

- $M[\varepsilon_i] = 0, \ i = \overline{1,n}$
- $M[\varepsilon_i \varepsilon_j] = 0, \ i \neq j$, т.е. шумы некоррелированы
- $D[\varepsilon_i] = \sigma^2 < \infty, \ i = \overline{1,n}$

Если также предположить, что $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, то можно вывести выражения для доверительных интервалов для α_i, а также проверять различные гипотезы для коэффициентов регрессии. Также накладывается естественное ограничение на матрицу регрессоров $X = (\boldsymbol{x}_i^T)_{i=1}^n$: $rang(X) = m$.

Обучение модели заключается в вычислении оценок коэффициентов регрессии $a_i \equiv \hat{\alpha}_i$ по заданной выборке. Введя искусственный регрессор $x_{tm} \equiv 1$, можно записать модель в векторном виде - $y = X\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$, где к исходной матрице регрессоров добавлен столбец из единиц, а $\boldsymbol{\varepsilon} = (\varepsilon_i)_{i=1}^n, \boldsymbol{\alpha} = (\alpha)_{i=1}^n$, Известно, что МНК-оценку коэффициентов можно вычислить как $\boldsymbol{a} = (X^T X)^{-1} X^T \boldsymbol{y}$. Тогда предсказанные линейной регрессией значения откликов это $\hat{\boldsymbol{y}} = X\boldsymbol{a}$

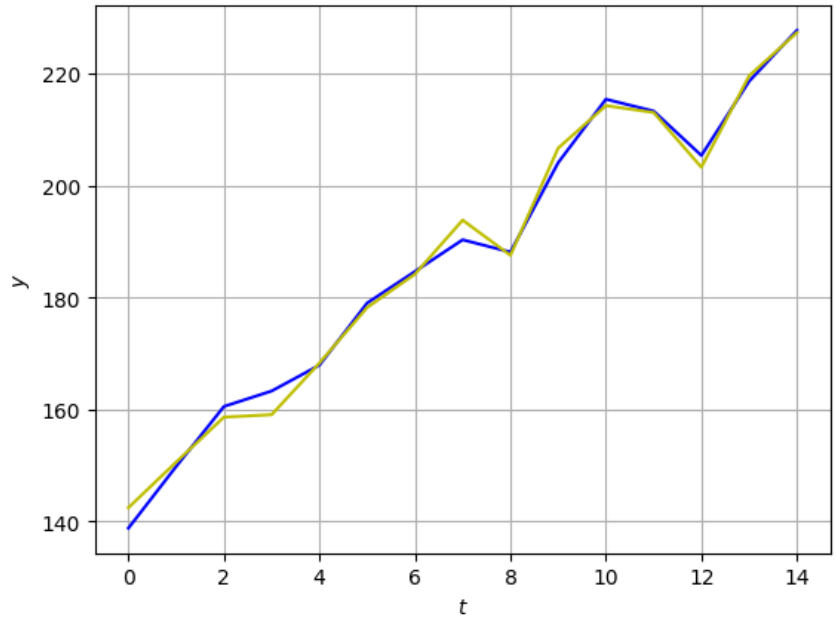


Рис. 1: Сравнение значений настоящих откликов и оцененных моделью на обучающих данных

Посчитав для наших данных $\boldsymbol{a} = (0.3486, 0.404, 4.5925, -23.016)$, можно увидеть, что линейная регрессия хорошо описывает данные - на обучающей выборке отклонения $y_i - \hat{y}_i$ малы.

3 Характеристики коэффициентов регрессии

Зная оценку вектора коэффициентов $\boldsymbol{a}$, можно вычислить

- Оценку дисперсии шумов $s^2 \equiv \hat{\sigma}^2 = \frac{\|\boldsymbol{e}\|_2^2}{n-m}$
- Оценку матрицы ковариаций $\widehat{cov}(\boldsymbol{a}) = s^2(X^T X)^{-1}$
- Стандартную ошибку оценки i-го коэффицента $s_i = s(\boldsymbol{a})_i = \sqrt{\widehat{cov}(\boldsymbol{a})_{ii}}$
- Матрицу корреляций оценок коэффициентов $cor(\boldsymbol{a})_{ij} = \frac{\widehat{cov}(\boldsymbol{a})_{ii}}{s(a_i)s(a_j)}$

Где $\boldsymbol{e} = (e_i)_{i=1}^n = (y_i - \hat{y}_i)_{i=1}^m$ - вектор остатков.

В нашем случае, получаются следующие значения

- $s^2 = 5.7112$

- $\widehat{cov}(\boldsymbol{a})$ =
$$\begin{pmatrix} 0.0014 & -0.008 & -0.0117 & 0.5058 \\ -0.008 & 0.125 & -0.2203 & -8.1732 \\ -0.0117 & -0.2203 & 1.1301 & 13.1777 \\ 0.5058 & -8.1732 & 13.1777 & 554.2771 \end{pmatrix}$$

- $s(\boldsymbol{a})$ = (0.03730.35361.144623.5431)

- $cor(\boldsymbol{a})$ =
$$\begin{pmatrix} 1 & -0.6075 & -0.2731 & 0.5752 \\ -0.6076 & 1 & -0.5443 & -0.9819 \\ -0.2731 & -0.5443 & 1 & 0.4891 \\ 0.5752 & -0.9819 & 0.4891 & 1 \end{pmatrix}$$

Судя по корреляционной матрице, 2-ой и 4-ый регрессоры сильно коррелируют - $cor(\boldsymbol{a})_{24}$ почти равняется -1.

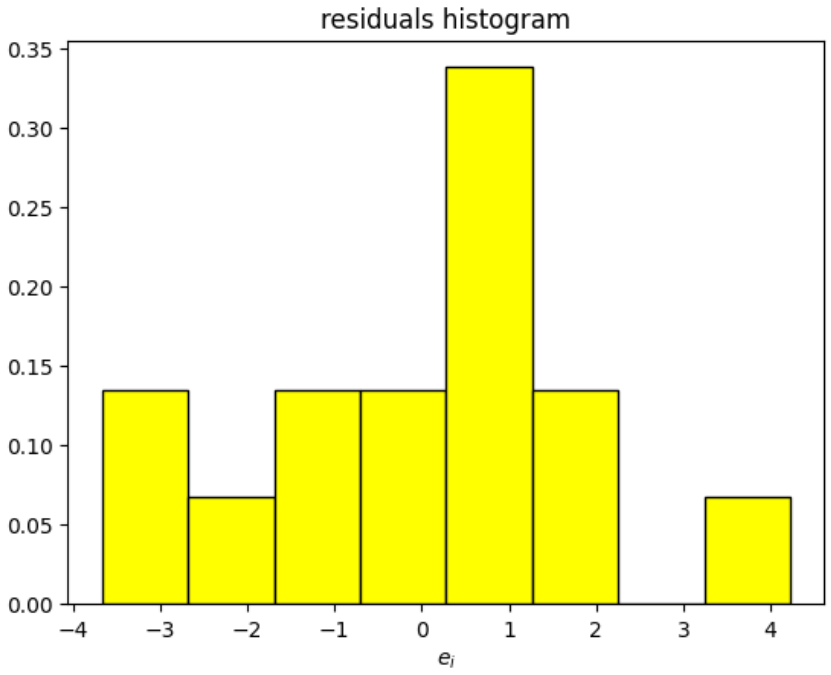


Рис. 2: Гистограмма невязок

4 Доверительные интервалы для коэффициентов регрессии

4.1 Индивидуальные доверительные интервалы

Если дополнительно предположить, что шумы в нашей модели нормально распределены, то можно вывести выражения для доверительных интервалов для α_i - статистика $t = \frac{|a_i - \alpha_i|}{s_i}$ в таком случае имеет распределение Стьюдента с $n-m$ степенями свободы. В таком случае доверительные интервалы с доверительной вероятностью $1 - \alpha$ это:

$$D_i = [\alpha_i^{lb}, \alpha_i^{ub}] = [a_i - s_i t_{1-\alpha/2,n-m},\ a_i + s_i t_{1-\alpha/2,n-m}] \tag{2}$$

Где $t_{1-\alpha/2,n-m}-$ квантиль распределения Стьюдента с $n-m$ степенями свободы уровня $1-\alpha/2$. В нашем случае получаются следующие интервалы для стандартной доверительной вероятности 0.95:

1. $D_1 = [0.2664, 0.4308], a_1 = 0.3486$
2. $D_2 = [-0.3742, 1.1822], a_2 = 0.4039$
3. $D_3 = [2.0733, 7.1117], a_3 = 4.5925$
4. $D_1 = [-74.8340, 28.8020], a_4 = -23.0160$

Таким образом, все оценки для коэффициентов попали в доверительные области.

4.2 Совместная доверительная область

Принцип Тьюки позволяет строить совместной доверительную область D для всех коэффициентов регрессии в виде m-мерного прямоугольника на основе индивидуальных доверительных интервалов: для построения совместного интервала D с доверительной вероятностью $1-\alpha$ достаточно построить индивидуальные доверительные интервалы D_i для α_i с доверительной вероятностью $1 - \alpha/m$ – тогда $D = \bigotimes_{i=1}^{m} D_i$.

В нашем случае все сводится к построения индивидуальных доверительных областей с доверительной вероятностью $1 - 0.05/4 = 0.9875$:

1. $D_1 = [0.2372, 0.4547], a_1 = 0.3486$
2. $D_2 = [-0.6463, 1.4508], a_2 = 0.4039$
3. $D_3 = [1.1859, 8.0043], a_3 = 4.5925$
4. $D_1 = [-93.2317, 47.1047], a_4 = -23.0160$

То есть, вектор $\boldsymbol{a}$ попал в совместную доверительную области с доверительным уровнем 0.95.

5 Линейные гипотезы для коэффициентов регрессии

Опять же, использую нормальность шумов, строятся определенны статистики, которые позволяет проверять гипотезы для коэффициентов регрессии. В частности, можно построить линейные гипотезы:

- Гипотезу $H_0 = \{\alpha_1 = \cdots = \alpha_m = 0\}$ об адекватности модели среднего $y_t = \alpha_{m+1} + \varepsilon_t$
- Гипотезу $H_0 = \{\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2\}$ об идентичности двух регрессий - при разбиении исходной выборки на две подвыборки $\{\boldsymbol{y}_1, X_1\}, \{\boldsymbol{y}_2, X_2\}$ объемом n_1 и n_2 соответственно $(n_1 + n_2 = n)$

5.1 Гипотеза об адекватности модели среднего

Гипотеза использует тот факт, что следующая статистика t принадлежит распределению Фишера:

$$t = \frac{R^2}{1 - R^2} \cdot \frac{n - m}{m - 1} \sim F(m - 1, n - 1) \tag{3}$$

Где $R^2 = \frac{\|\boldsymbol{e}\|_2^2}{\|\boldsymbol{y} - \bar{y} \cdot \mathbf{1}\|_2^2}$ - коэффициент детерминации регресии, который показывает, насколько лучше данная регрессия модели среднего. Значения R^2, близкие к 1, показывают превосходство регрессии над моделью среднего. Несмещенный коэффициент детерминации вычисляется по формуле $R^2_{unbiased} = 1 - (1 - R^2)\frac{n-1}{n-m}$

Так как распределение Фишера задано на положительной полуоси, то тест односторонний, p-value вычисляется как:

$$p = 1 - F_{F(m-1,n-m)}(t) \tag{4}$$

В нашем случае $R^2 = 0.9938, R^2_{unbiased} = 0.9921, p = 8.1027 \cdot 10^{-12}$. Следовательно, модель среднего неадекватно описывает данные и должна быть отброшена, как и нулевая гипотеза.

5.2 Гипотеза об идентичности двух регрессий

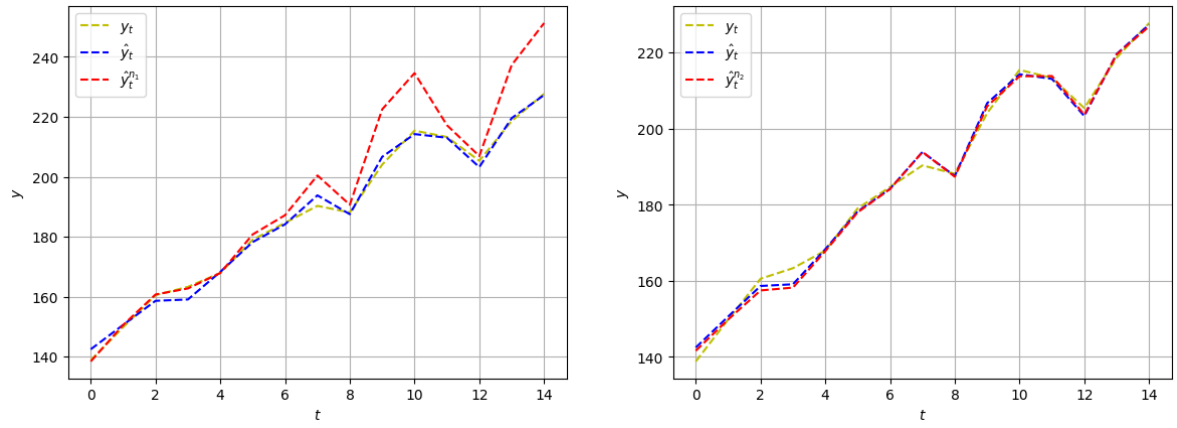
Пусть выборка разбита на две подвыборки объема n_1 и n_2, т.ч. $n_i \geq m$. Затем на каждой из подвыборок обучается линейная регрессия с итоговыми коэффициентами $\boldsymbol{a_1}, \boldsymbol{a_2}$ и ставится гипотеза $H_0 = \{\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2\}$ об идентичности двух регрессий. В условиях H_0 статистика t:

$$t = \frac{(Q_R - Q_1 - Q_2)/m}{s^2} \sim F(m, n - m) \tag{5}$$

Где $Q_R = \|y - X\boldsymbol{a}_R\|_2^2, s^2 = (Q_1 + Q_2)/(n_1 + n_2 - 2m), Q_i = \|y - X\boldsymbol{a}_i\|_2^2$, а $\boldsymbol{a}_R$ - коэффициенты регрессии, обученной на всей исходной выборке. Тогда p-value есть

$$p = 1 - F_{F(m,n-m)}(t) \tag{6}$$

В нашем случае, при $n_1 = 5, n_2 = n - n_1 = 10$ получается p-value равное 0.1432, т.е. гипотеза H_0 принимается.



(a) Модель, обученная на первых 5 элементах

(b) Модель, обученная на последних 10 элементах

Рис. 3: Сравнение исходной регрессии и регрессии, обученной на некотором подмножестве выборки

5.3 Гипотеза о незначимости регрессора

Проверка гипотезы $H_0 = \{\alpha_i = 0\}$ сводится к тому факту, что следующая статистика t_i принадлежит распределению Стьюдента:

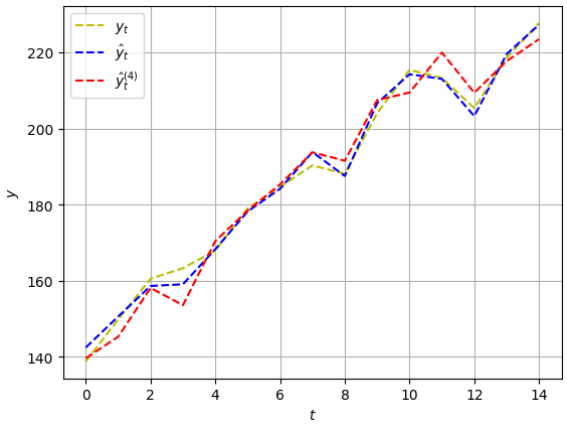
$$t_i = \frac{|a_i|}{s_i} \sim S(n-m) \tag{7}$$

Значит, если альтернативная гипотеза - $H_1 = \{\alpha_i \neq 0\}$, то тест - двусторонний. Учитывая, что распределение Стьюдента симметрично относительно нуля, p-value есть:

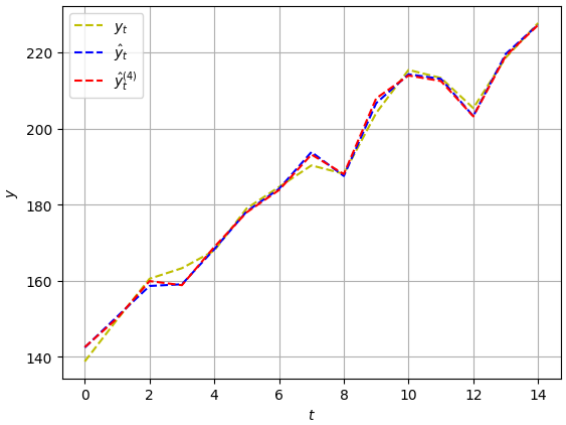
$$p_i = \mathbb{P}\{|t_i| > T\} = 2(1 - F_{S(n-m)}(t_i)),\ T \sim S(n-m) \tag{8}$$

В нашем случае $\boldsymbol{p} = (7.3125 \cdot 10^{-7}, 0.1387, 0.0010, 0.1746)$. Первый и третий регрессоры имеют достаточно малое значение p-value - их определенно нельзя отбросить без потери точности модели.

Второй и четвертый регрессоры имеют высокое значение p-value и могут быть проигнорированы при построении регрессии. Стоит заметить, что четвертый регрессор фиктивен и соответствует константному слагаемому в модели, что еще раз подтверждает неадекватность модели среднего для наших данных(учитывая тот факт, что данные, очевидно, не похожи на белый шум и имеют отличимый тренд).



(a) Модель без второго регрессора



(b) Модель без четвертого регрессора

Рис. 4: Сравнение исходной модели и модели, не учитывающей незначительный признак

На графиках видно, что обе регрессии $\hat{y}_t^{(2)}, \hat{y}_t^{(4)}$ не намного потеряли точность при исключении малозначимого второго и четвертого регрессора в первом и втором случае соответственно.

6 Оценка на тестовой выборке

Обучаем выборку на первых 14-ти элементах, а для 15-го строим прогноз. Чтобы проверить модель на тестовых данных, исходная выборка была разбита на обучающую и тестовую выборку, объем тестовой выборки равен 1. В качестве тестовой выборки был выбран первый элемент $\{y_1, \boldsymbol{x}_1\}$. В результате невязка $y_1 - \hat{y}_1$ равнялась -2.0624, а относительная невязка $\frac{y_1 - \hat{y}_1}{y_1} = -0.01$, т.е. прогноз оказался достаточно точным.

7 Вывод

Анализируя результаты, можно сделать выводы:

- Заданная выборка хорошо описывается линейной регрессией
- Модель среднего не подходит для данной выборки