Лабораторная работа №2
(теоретическая часть)

Садиев Абдурахмон, 674 группа

17.05.2020

Задача 2.2

Условие:

Рассмотрим задачу предсказание числа заболевших некоторой болезнью от некоторых экологических анализов. Гарантируется, что предсказание описывается линейной моделью.

Так как проведение анализов не является бесплатным, то стоит вопрос о том какие из анализов являются лишними (на уровне значимости $\alpha = 0.05$) для предсказания линейной модели.

Требуется:

1. Записать задачу формально;
2. Провести отбор приизнаков линейной модели.

Все выкладки должны быть сделаны аналитически, без использования компьютера.

Решение

Данные

Данные предоставленны на странице курса: целевая переменная соответствует столбцу $y \in \mathbb{R}^n$, матрица признаков $X^* \in \mathbb{R}^{n \times k}$, где $n = 30$, а $k = 10$. Будем обозначать $x_1, x_2, x_3, x_4, x_5, x_7, x_7, x_8, x_9, x_{10}$ признаками (независимые переменные).

Модель линейной регрессии

Как известно, в модели линейной регрессии предполагается, что

$$y = X^* \beta + \varepsilon,$$

где $\beta \in \mathbb{R}^{k+1}$ - параметры модели, а ε есть случайная ошибка модели, причем $\mathbb{E}(\varepsilon) = 0.X \in \mathbb{R}^{n \times (k+1}$- есть матрица X^* с добавленном столбцом из единичек (для константного параметра мо) Тогда задача решается с помощью метода наименьших квадратов:

$$\|X\beta - y\|_2^2 \to \min_{\beta}$$

. Праметры вычисляются по формуле $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$, а оценка целевой переменной нашей моделью- $\hat{y} = X \left(X^T X\right)^{-1} X^T y$.

Отбор признаков модели

Для отбора признаков воспользуемся критерием Фишера (частным F-тест). Данный критерий помогает определить стоит ли добавлять переменную в нашу модель, то есть мы будем добавлять переменные (усложнять модель).

Для применения критерия Фишера введем необходимые понятия:

1. (Explained Sum of Squares): $ESS = \sum_{i=1}^{n} (\hat{y}_i - \overline{y}_i)^2$, $\overline{y}$ есть среднее значение y. ESS измеряет степень вариации смоделированных значений и сравнивается с общей суммой квадратов, которая измеряет степень вариации наблюдаемых данных, и с остаточной суммой квадратов, которая измеряет степень вариации погрешностей моделирования. Физический смысл: доля изменчивости, объясняемая линейной зависимостью.

2. Выше мы говорили, что будем добавлять признаки. Тогда возникает вопрос о пересчете показателя. Пусть $x_1, x_2, ..., x_{k'}$ - переменные, на которых построена модель линейной регресси. для этой модели посчитана Explained Sum of Squares: ESS.Пусть x_{new} - переменная, которую мы хотим добавить. Обозначим ESS_{new} как Explained Sum of Squares для модели, построеннаяной на переменых $x_1, x_2, ..., x_{k'}, x_{new}$. Тогда величина отвечающая за вклад переменной x_{new} в объясняющую способность модели есть $Con_{new} = ESS_{new} - ESS$. Важное свойство: чем больше Con_{new}, тем значительнее вклад пременной. Теперь стоит задуматься, как мы опеределим порог для принятия решения о достаточности велечины Con_{new}. Иначе говоря, значим признак x_{new} или нет. В этом поможет критерий Фишера.

3. Residual Sum of Squares: $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Наш критерий проверяет следующую гипотизу H_0 против альтернативы $H1$:

1. H_0: Con_{new} не достаточно велик. (переменная не значима)
2. H_1: Con_{new} достаточно велик.(переменная значима)

Если быть более формальным, то

1. H_0: $\beta_{new} = 0$
2. H_1: $\beta_{new} \neq 0$

Для этого определим статистику:

$$T = \frac{Con_{new}(n - k' - 2)}{RSS_{new}},$$

где RSS_{new} - Residual Sum of Squares для модели, построеннаяной на переменых $x_1, x_2, ..., x_{k'}, x_{new}$. Для данной статистики найдено нулевое распределение: $T \sim \mathcal{F}(1, n - k' - 2)$ - распределение Фишера со степенями свободы 1, $n - k' - 2$.

Поскольку $Con_{new} = ESS_{new} - ESS \geq 0$, то у нас будет правостороняя альтернатива. Правило принятия решения будет выглядеть таким образом:

$$\boxed{\text{нулевая гипотеза } \mathrm{H}_0 \text{ отклоняется} \Leftrightarrow Con_{new} > \mathcal{F}(1, n - k' - 2)_{(1-\alpha)}},$$

где $\mathcal{F}(1, n - k' - 2)_{(1-\alpha)}$ - $1 - \alpha$ квантиль распределения Фишера.

Таким образом, мы имеем

1. нулевая гипотеза: H_0: $\beta_{new} = 0$
2. альтернатива: H_1: $\beta_{new} \neq 0$
3. статистика: $T = \frac{Con_{new}(n-k'-2)}{RSS_{new}}$
4. нулевое распределение: $T \sim \mathcal{F}(1, n - k' - 2)$

Алгоритм отбора признаков модели

1. Выберем признак, имеющий наибольшую корреляцию с целевой переменной. (Наибольшее значение коэффициента Пирсона)

$$r_{x_i,y} = \frac{\sum_{j=1}^{n}((x_i)_j - \overline{x_i})(y_j - \overline{y})}{\sqrt{\sum_{j=1}^{n}((x_i)_j - \overline{x_i})^2 \sum_{j=1}^{n}(y_j - \overline{y})^2}}, n = 30$$

2. Проверяем на значимость выбранный признак при помощи критерия. Согласно ему, если статистические данные не противоречат гипотезе, алгоритм закончен (нет значимых признаков). В противном случае добавляем это признак в модель.

3. Далее считаем для оставшихся переменных статистику T и выбираем переменную с максимальным значением статистики.

4. Проверяем на значимоть выбранный признак при помощи критерия. Согласно ему, если статистические данные не противоречат гипотезе, алгоритм закончен (больше нет значимых признаков). В противном случае добавляем это признак в модель. И возвращаемся к пункту 3.

Переменная	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
$r_{x_i y}$	-0.245	0.086	0.786	0.186	0.010	-0.355	0.020	0.110	-0.134	0.096

Таблица 1: Коэффициент Пирсона между x_i и y.

Вычисление

Согласно приведенному алгоритму найдем значения коэффициента Пирсона. Как мы видим, максимальное значение соответствует переменной x_3. Тогда определим модель линейной регрессии, как $y = \beta_0 + \beta_3 x_3$. Согласно пункту о модели линейной регрессии вычислим параметры модели с помощью метода наименьших квадратов : $\beta_0 = 1.78, \beta_3 = 3.16$. Теперь рассчитаем значение статистики для переменной x_3: $Con(x_3) = 190.27, T = 45.14$. Согласно таблицам $\mathcal{F}(1, 28)_{(1-\alpha)} = 4.20$, тогда $\mathcal{F}(1, 28)_{(1-\alpha)} < T$, то есть признак x_3 значим.

Теперь для оставшихся признаков нужно рассчитать значение статистики T. Как мы видим макси-

Переменная	x_1	x_2	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
Статистика T	6.25	0.35	0.89	0.36	0.63	0.22	0.09	0.02	0.06

Таблица 2: Значение статистики T для оставшихся переменных.

мальное значение статистики сооответствует переменной x_1. Теперь проверим его значимость: находим $\mathcal{F}(1, 27)_{(1-\alpha)} = 4.21$. Сравнивая $\mathcal{F}(1, 27)_{(1-\alpha)} < T$, получаем, что признак x_1 значим, следовательно, добавляем его в модель. То есть мы усложняем нашу модель: $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3$, где $\beta_0 = 1.73, \beta_1 = -0.74, \beta_3 = 3.20$.

Снова считаем значение статистики T для нового набора оставшихся переменных.

Переменная	x_2	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
Статистика T	0.42	1.67	0.07	0.90	0.34	0.22	0.02	-0.01

Таблица 3: Значение статистики T для оставшихся переменных.

Мы видем, что максимальное значение соответствует x_4. Теперь применяем критерий для проверки значимости переменной x_4. Квантить: $\mathcal{F}(1, 26)_{(1-\alpha)} = 4.23$, сравниваем: $\mathcal{F}(1, 26)_{(1-\alpha)} > T = 1.67$. Статистические данные гипотизе не противоречат. Таким образом, нетолько x_2, но и остальные из оставшихся не значимы.

Вывод

Используя частный F-тест, мы выяснили, что только x_1 и x_3 признаки значимы. Это хорошо с экономической точки зрения, поскольку мы можем умешьнить количество проводимых анализов.