Фамилия и имя: Вареник Наталия

Прикладной статистический анализ данных

Лабораторная работа 2

Задача 2.2 Имеются признаки $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{10}$ и целевая переменная $\boldsymbol{y}$. Требуется построить линейную модель и отобрать наиболее значимые признаки на уровне значимости $\alpha = 0.05$.

Здесь отбор признаков будет осуществляться не выбрасыванием незначимых признаков из модели, а добавлением значимых признаков по одному начиная с признака, имеющего наибольшую корреляцию с целевой переменной.

Для оценки значимости признака будем пользоваться частным F-тестом, идея которого состоит в подсчете суммы квадратов регрессии $SSR = \sum\limits_{i=1}^{n}(\hat{y}_i - \overline{\boldsymbol{y}})^2$, которая описывает объясняющую способность модели. Тогда пусть мы хотим добавить в модель новую переменную x_{add}. Для этого введем переменную SSR_{add}, которая будет описывать вклад добавленной переменной в объясняющую способность модели. Ее можно вычислить как разность объясняющей способности модели после введения новой переменной и до введения. Тогда принятие решения о значимости признака будет состоять в выборе порога на вклад нового признака SSR_{add} после которого мы будем признавать признак значимым и добавлять его в модель. Теперь можно сформулировать гипотезу, проверяемую в частном F-тесте.

H_0 : вклад SSR_{add} меньше порога, признак добавлять в модель не нужно

H_1 : вклад SSR_{add} больше порога, признак нужно добавить в модель

Для проверки данной гипотезы используется статистика $T = \frac{SSR_{add}}{MSE}$, где $MSE = \frac{\sum\limits_{i=1}^{n}(y_i - \hat{y_i})^2}{n-k-2}$ — это сумма квадратов ошибок модели с новым признаком, приходящая на одну степень свободы (n - размер выборки, k - число признаков модели, не учитывая новый, который мы хотим добавить). Данная статистика описывает отношение прироста объясняющей способности модели в результате добавления нового признака к ошибке модели с учетом добавленного признака. В условии истинности нулевой гипотезы статистика T имеет распределение Фишера $\mathcal{F}(1, n-k-2)$. В качестве критической области возьмем правый хвост распределения, соотвествущий $1-\alpha$-квантили распределения Фишера. Тогда при попадании статистики T в критическую область гипотеза H_0 отклоняется.

Тогда отбор признаков будем проводить следующим образом. Сначала выберем признак, который имеет наибольшую корреляцию с целевой пременной. Проверим его на значимость, если значимость признака подтверждается, то он добавляется в модель, в ином случае отбор признаков прекращается ввиду малоинформативности признаков. В случае, если признак добавился в модель смотрим на оставшиеся признаки. Среди них выбирается новый признак, который мы хотим добавить как тот, который имеет наибольшее значение статистики T и он проверяется на значимость. Если значимость подтверждается, то он добавляется в модель, в ином случае отбор признаков прекращается.

Попробуем отобрать признаки для выборки из задания.

Посчитаем коэффициент корреляции Пирсона каждого признака $\boldsymbol{x}_i$ с целевой переменной.

$$r_{\boldsymbol{x}_i,\boldsymbol{y}} = \frac{\sum\limits_{j=1}^{30}(x_{ij} - \hat{\boldsymbol{x}}_i)(y_j - \hat{\boldsymbol{y}})}{\sqrt{\sum\limits_{j=1}^{30}(x_{ij} - \hat{\boldsymbol{x}}_i)^2}\sqrt{\sum\limits_{j=1}^{30}(y_j - \hat{\boldsymbol{y}})^2}}$$

Признак $\boldsymbol{x}_3$ имеет наибольшую корреляцию, выбираем его в качестве первого. Проверяем его на значимость. Имеем модель вида $\boldsymbol{y} = c_0 + c_3\boldsymbol{x}_3$. Коэффициенты $c_0 = 1.78$, $c_3 = 3.17$ находим с помощью МНК. Найдем значение статистики T. Посчитав все получаем $SSR_{add} = 190.29$,

	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_6$	$\boldsymbol{x}_7$	$\boldsymbol{x}_8$	$\boldsymbol{x}_9$	$\boldsymbol{x}_{10}$
$r_{\boldsymbol{x}_i,\boldsymbol{y}}$	-0.24	0.09	0.79	0.19	0.01	-0.35	0.02	0.11	-0.13	0.10

Таблица 1: Коэффициент корреляции Пирсона между признаками и целевой переменной

$MSE = 4.21, T = 45.16$. Т.к. $\mathcal{F}_{1-\alpha}(1, 28) = 4.2$ получаем, что гипотеза H_0 отклоняется и признак x_3 нужно включить в модель. Смотрим на оставшиеся признаки. Считаем для них значение статистики T.

	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_6$	$\boldsymbol{x}_7$	$\boldsymbol{x}_8$	$\boldsymbol{x}_9$	$\boldsymbol{x}_{10}$
T	6.24	0.35	0.88	0.35	0.62	0.21	0.08	0.01	0.05

Таблица 2: Значение статистики T на второй итерации для всех признаков кроме $\boldsymbol{x}_3$

Видим, что у $\boldsymbol{x}_1$ значение статистики самое большое, выбираем теперь его. Проверим его значимость. Для этого находим $\mathcal{F}_{1-\alpha}(1, 27) = 4.21$, т.к. $T = 6.24 > 4.21$ делаем вывод, что признак $\boldsymbol{x}_1$ значим, добавляем его в модель. Для полученной модели и для оставшихся признаков повторяем аналогичные шаги. Получаем следующие значения статистики

	$\boldsymbol{x}_2$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_6$	$\boldsymbol{x}_7$	$\boldsymbol{x}_8$	$\boldsymbol{x}_9$	$\boldsymbol{x}_{10}$
T	0.41	1.61	0.06	0.87	0.32	0.22	0.02	0.002

Таблица 3: Значение статистики T на третьей итерации для всех признаков кроме $\boldsymbol{x}_1, \boldsymbol{x}_3$

Но так как $\mathcal{F}_{1-\alpha}(1, 26) = 4.26$, то выбирая любой признак мы не сможем опровергнуть его незначимость. Следовательно, в модель будут входить признаки $\boldsymbol{x}_1$ и $\boldsymbol{x}_3$ с коэффициентами $c_0 = 1.73$, $c_1 = -0.74$, $c_3 = 3.20$