Task 2.2

Угнивенко Виталий

1 Постановка задачи

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:
- Записать задачу формально
- Провести отбор приизнаков линейной модели

Все выкладки должны быть сделаны аналитически, без использования компьютера.

2 Частный F-тест

Критерий призван оценить целесообразность ввода дополнительной независимой переменной в линейную модель множественной регрессии, уравнение которой, как известно, имеет вид:

$$y = w_0 + w_1 \cdot x_1 + \cdots + w_n \cdot x_n + \varepsilon$$

y - целевая переменная
$x_1, \ldots x_n$ - независимые переменные

ε - случайный шум
$w_0, \dots w_n$ - коффициенты модели

Введём такое понятие, как сумма квадратов регрессии

$$SSR = \sum_{i=1}^{n} (\hat{y_i} - \overline{y})^2$$

$\hat{y_i}$ - оценка, полученная на основе регрессионной модели
$\overline{y}$ - среднее по всем наблюдениям y
n - объем выборки

Данный показатель характеризует ту долю общей вариации результативного признака y, которую получилось объяснить при помощи регрессии.

Теперь предположим, что на основе переменных $x_1, \dots x_k$ была построена регрессионная модель, для которой доля изменчивости, объясненная линейной зависимостью, составила величину $SSR_{initial}$. Допустим, что мы хотим ввести в модель новый признак x_{extra}. Сумма квадратов регрессии SSR, построенной на независимых переменных $x_1, \dots x_k$ и x_{extra}, составляет SSR_{full}. Очевидно, что $SSR_{full} \geq SSR_{initial}$.

Рассчитаем, насколько увеличилась объясняющая способность модели в результате ввода новой переменной x_{extra}: $S_{extra} = SSR_{full} - SSR_{initial}$. Таким образом, S_{extra} можно назвать вкладом переменной x_{extra} в объяснение общей изменчивости результативного признака y. Очевидно, чем больше данное значение, тем весомей этот вклад.

Тогда возникает вопрос: Каким образом следует выбрать порог для SSR_{extra}, чтобы признать эту величину достаточно большой и, соответственно, принять решение о значимости признака x_{extra}?
С ответом на этот вопрос может помочь так называемый частный F-тест.

По сути, данный критерий призван проверить следующую гипотезу:

H_0 : вклад SSR_{extra}, вносимый x_{extra}, не достаточно велик, ввиду чего эту переменную не следует включать в модель против альтернативы

H_1 : вклад SSR_{extra} , вносимый x_{extra}, значительный, и потому эту переменную следует включить в модель.

Для того чтобы проверить эти гипотезы, следует перейти от рассматриваемого показателя SSR_{extra} к статистике следующего вида:

$$\gamma = \frac{SSR_{extra}}{MSE_{full}}$$

MSE_{full} представляет собой сумму квадратов ошибок SSE (модель построена по переменным $x_1, \dots x_k$ и x_{extra}), приходящуюся на одну степень свободы df_{sse}. Значение MSE_{full} может быть найдено по формуле:

$$MSE_{full} = \frac{SSE}{df_{sse}} = \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{n-k-2}$$

Доказано, что статистика γ при справедливости гипотезы H_0 распределена по закону Фишера.

Тогда проверка гипотезы H_0 будет сводиться к следующей последовательности действий:

1. Задаемся уровнем значимости α.
2. По специальным таблицам находим процентную точку K_α распределения Фишера со степенями свободы $d_1 = 1$ и $d_2 = n - k - 2$.Это значение будет являться граничным для статистики γ.
3. Сравниваем найденную процентную точку K_α со значением статистики γ. Если окажется, что $\gamma > K_\alpha$, то делается вывод о значимости признака x_{extra} и, соответственно, его следует включить в модель.
   Если же $\gamma \leq K_\alpha$, то принимается решение о неэффективности включения переменной x_{extra} в модель.

3 Метод прямого отбора

Данный алгоритм включает в себя следующие шаги:

1. Из списка всех возможных входных переменных выбирается та, которая имеет наибольшую корреляцию с y, после чего модель, содержащая лишь одну выбранную независимую переменную, проверяется на значимость при помощи частного F-критерия. Если значимость модели не подтверждается, то алгоритм на этом заканчивается за неимением существенных входных переменных. В противном случае эта переменная вводится в модель и осуществляется переход к следующему пункту алгоритма.Следует отметить, что в данном случае проверка на значимость всей модели в целом будет равносильна проверке на значимость выбранной независимой переменной, так как на данном этапе модель еще не содержит других входных переменных.

2. По всем оставшимся переменным рассчитывается значение статистики γ, которая представляет собой отношение прироста суммы квадратов регрессии, достигаемая за счет ввода в модель соответствующей дополнительной переменной (по сравнению с величиной $SSR_{initial}$, рассчитанной только лишь на основе ранее уже введенных переменных), к величине MSE_{full}.

3. Из всех переменных-претендентов на включение в модель выбирается та, которая имеет наибольшее значение критерия, рассчитанного в пункте 2.

4. Проводится проверка на значимость выбранной в пункте 3 независимой переменной. Если ее значимость подтверждается, то она включается в модель, и осуществляется переход к пункту 2 (но уже с новой независимой переменной в составе модели). В противном случае алгоритм останавливается.

4 Решение

Найдём корреляцию Пирсона каждого признака с целевой переменной по следующей формуле:

$$\text{corr}(x_j, y) = \frac{\sum_{i=1}^{30}(x_{ji} - \overline{x_j})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{30}(x_{ji} - \overline{x_j})^2 \cdot \sum_{i=1}^{30}(y_i - \overline{y})^2}}$$

j - номер признака, а i - номер экземпляра в выборке

feature	$\text{corr}(x_j, y)$
x_1	-0.244527
x_2	0.085782
x_3	0.785655
x_4	0.185746
x_5	0.010491
x_6	-0.354543
x_7	0.019549
x_8	0.110317
x_9	-0.134460
x_{10}	0.096125

Признак x_3 имеет наибольшую корреляцию, потому рассмотрим модель $\hat{y} = b + w \cdot x_3$. Коэффициенты найдём с помощью МНК: $b = 1.78$ и $w = 3.16$

Посчитаем всё необходимое:

x_{3i}	y_i	$\hat{y_i}$	$\overline{y}$	$(y_i - \hat{y_i})^2$	$(y_i - \overline{y})^2$
-1.2	0.0	-2.02	2.7	4.07	22.25
0.5	1.0	3.36	2.7	5.59	0.44
0.8	3.0	4.31	2.7	1.73	2.61
0.5	3.0	3.36	2.7	0.13	0.44
0.4	1.0	3.05	2.7	4.2	0.12
0.9	2.0	4.63	2.7	6.92	3.73
1.5	10.0	6.53	2.7	12.04	14.67
1.1	4.0	5.26	2.7	1.6	6.58
-0.1	1.0	1.47	2.7	0.22	1.52
2.0	5.0	8.11	2.7	9.69	29.31
-1.2	0.0	-2.02	2.7	4.07	22.25
1.8	14.0	7.48	2.7	42.5	22.85
-0.2	1.0	1.15	2.7	0.02	2.41
-0.7	1.0	-0.43	2.7	2.06	9.82
0.5	3.0	3.36	2.7	0.13	0.44
-0.6	0.0	-0.12	2.7	0.01	7.94
1.5	10.0	6.53	2.7	12.04	14.67
0.2	1.0	2.42	2.7	2.0	0.08
0.7	3.0	4.0	2.7	1.0	1.68
0.3	1.0	2.73	2.7	3.0	0.0
0.1	1.0	2.1	2.7	1.21	0.36
-0.3	1.0	0.83	2.7	0.03	3.49
-0.1	1.0	1.47	2.7	0.22	1.52
0.5	4.0	3.36	2.7	0.4	0.44
-0.4	1.0	0.52	2.7	0.23	4.77
-0.6	1.0	-0.12	2.7	1.25	7.94
0.1	1.0	2.1	2.7	1.21	0.36
0.1	2.0	2.1	2.7	0.01	0.36
0.9	4.0	4.63	2.7	0.4	3.73
-0.3	1.0	0.83	2.7	0.03	3.49

Таким образом

$$\mathrm{MSE}_{\mathrm{full}} = 4.21$$

$$\mathrm{SSE}_{extra} = 190.27$$

$$\gamma = 45.14$$

Табличное же значение при $d_1 = 1$ и $d_2 = 28 \rightarrow K_\alpha = 4,20$, что меньше, чем γ. Соответсвенно, взяв в модель x_3 мы ошибёмся с вероятностью не более 5%.

Теперь для всех остальных признаков считаем γ аналогичным образом:

feature	SSE_{extra}	MSE_{full}	γ
x_1	3.55	22.19	6.25
x_2	4.31	1.52	0.35
x_4	4.23	3.75	0.89
x_5	4.31	1.54	0.36
x_6	4.27	2.67	0.63
x_7	4.34	0.95	0.22
x_8	4.36	0.38	0.09
x_9	4.37	0.09	0.02
x_{10}	4.36	0.26	0.06

Самое большое значение γ у признака x_1. Табличное же значение при $d_1 = 1$ и $d_2 = 27 \rightarrow K_\alpha = 4,21$, что меньше, чем $\gamma = 6.25$ у x_1. Соответсвенно, взяв в модель x_1 мы ошибёмся с вероятностью не более 5%.

Повторим итерацию вычислений:

feature	SSE_{extra}	MSE_{full}	γ
x2	3.49	1.48	0.42
x4	3.34	5.58	1.67
x5	3.54	0.25	0.07
x6	3.43	3.09	0.9
x7	3.51	1.18	0.34
x8	3.52	0.77	0.22
x9	3.55	0.08	0.02
x10	3.55	-0.03	-0.01

При этом табличное значение при $d_1 = 1$ и $d_2 = 26 \rightarrow K_\alpha = 4,23$, что больше, чем любая γ. Соответсвенно остальные признаки в модель включать не стоит.

Необходимо включить в модель признаки x_1 и x_3.