ЛЕКЦИЯ 12. Методы отбора переменных (признаков) регрессионные модели

Прямое включение

Работа начинается с «нулевой модели», которая не содержит ни одной переменной. На первом шаге поочерёдно в «пустую модель» включаются по одной переменной и выбирается та, которая обеспечивает лучший результат. Затем в модель, содержащую единственную переменную поочерёдно добавляются оставшиеся переменные и выбирается та, которая обеспечивает наибольшее улучшение качества модели. Схематично метод представлен на рис. 1.

Рис. 1. Метод прямого включения

В роли критерия качества модели обычно используется F-критерий (критерий Фишера):

$$F = \frac{S_1 - S_2}{S_2} \times \frac{m - n_2}{n_1 - n_2}$$

где S_1 – сумма квадратов остатков для модели с исходным числом параметров n_1 (короткой модели), S_2 – сумма квадратов остатков для модели с увеличенным числом параметров n_2. Разности $n_1 - n_2$ и $m - n_2$ представляют собой числа степеней свободы F-распределения.

Несложно увидеть, что статистика F-теста практически представляет собой отношение двух масштабируемых с помощью чисел степеней свободы сумм квадратов, отражающих различные источники изменчивости исходных данных. Эти суммы квадратов построены таким образом, что статистика F-критерия имеет большие значения когда нулевая гипотеза не верна. Иными словами, если нулевая гипотеза критерия предполагает, что добавление новой переменной в модель не увеличивает значимо её точность, то чем выше F, тем больше вероятность что эта гипотеза неверна и переменная является значимой.

Если значение критерия больше заданного порога, то добавление переменной значимо увеличивает качество модели (соответствие модели исходным данным) и, следовательно, целесообразно. Процесс повторяется до тех пор пока не будет выполнено некоторое правило остановки, либо переменные не будут исчерпаны.

Иными словами, лучшим кандидатом на включение в модель будет та переменная, которая обеспечит наибольшее сокращение квадрата остатков регрессии или, что эквивалентно, наибольшее значение статистики F-критерия.

Альтернативными подходами могут быть следующие:

- наименьшее p-значение, обеспечиваемое «длинной моделью»;

Напомню p-значение – это наименьшее значение уровня значимости (т.е. вероятности отказа от справедливой гипотезы), для которого вычисленная проверочная статистика ведет к отказу от нулевой гипотезы. Обычно p-значение сравнивают с общепринятыми стандартными уровнями значимости 0,005 или 0,01. Например, если вычисленное по выборке значение проверочной статистики соответствует $p = 0{,}005$, это указывает на вероятность справедливости гипотезы 0,5%. Таким образом, чем p-значение меньше, тем лучше, поскольку при этом увеличивается «сила» отклонения нулевой гипотезы и увеличивается ожидаемая значимость результата.

- наибольшее увеличение коэффициента детерминации R-квадрат. В этом случае статистика F-критерия будет:

$$F = \frac{R^2/(k-1)}{(1-R^2)(m-k)} \tag{2}$$

где k – число параметров модели, m – количество наблюдений в наборе данных.

- наибольшее увеличение квадратичной ошибки регрессии (SSR — sum squares of regression) или объяснённой суммы квадратов (ESS — explained sum of squares). В этом случае вместо квадратов остатков $(y - \hat{y})^2$, которая описывает (объясняет) вариативность данных относительно линии регрессии,

минимизируемая в функционале (2), используется сумма квадратов отклонений оценок регрессии относительного среднего, т.е. $(\bar{y} - \hat{y})^2$

Иными словами, использование квадратичной суммы остатков показывает насколько хорошо модель соответствует данным, а использование квадратичной ошибки регрессии — насколько модель с заданным набором параметров объясняет изменчивость исходных данных лучше, чем простое среднее значение. Процедура, в которой используется сумма квадратов регрессии при оценке значимости переменной включаемой в модель (или исключаемой из неё), известна как частный F-тест.

Выбор критерия остановки отбора. Используется некоторый порог p-значения по достижении которого процесс отбора останавливается. Порог может быть выбран следующими способами:

- фиксированное значение;
- определяется с помощью информационного критерия Акаике (Akaike's information criterion – AIC). Критерий для выбора лучшей из нескольких статистических моделей, построенных на одном и том же наборе данных и использующих логарифмическую функцию правдоподобия. Предложен Хироцугу Акаикэ. Критерий является не статистическим, а информационным, поскольку основан на оценке потери информации при уменьшении числа параметров модели. Критерий позволяет найти компромисс между сложностью модели (числом параметров) и ее точностью. В общем случае AIC вычисляется по формуле:

$$AIC = 2k - 2\ln(L)$$

где k – число параметров модели, L – максимизированное начение функции правдоподобия модели. Лучшей признается та модель, для которой значение AIC минимально.

AIC тесно связан с байесовским информационным критерием (Bayesian information criterion – BIC), но, в отличие от него, содержит функцию штрафа, линейно зависящую от числа параметров.

Если модель использует метод наименьших квадратов, то критерий может быть вычислен следующим образом:

$$AIC = \ln\left(\frac{RSS_p}{n}\right) + 2\frac{p}{n} + 1 + \ln 2\pi$$

где RSS_p – сумма квадратов остатков модели, полученная при оценке коэффициентов модели методом наименьших квадратов, n – объем обучающей выборки.

Из выражения видно, что при фиксированном размере выборки рост критерия обусловлен в основном увеличением числа параметров модели, а не ее ошибкой. Т.е. за увеличение числа параметров модель «штрафуется» сильнее, чем за долю необъясненной дисперсии ошибки. Таким образом, задача заключается в том, чтобы выбрать модель с минимальным числом параметров, которые объясняют наибольшую долю дисперсии ошибки.

На практике это делается следующим образом. Берется «нулевая модель», которая содержит только свободный член, и для нее вычисляется значение критерия. Затем в нулевую модель поочередно добавляются параметры, и каждый раз AIC вычисляется вновь. Выбирается модель, для которой значение критерия окажется минимальным.

На малых выборках рекомендуется применять скорректированный критерий Акаике. Модифицированный критерий Акаике, применяемый для выборок малого размера, когда отношение числа содержащихся в выборке примеров к числу параметров модели меньше 40 (т.е. вводится поправка на ограниченный объем выборки). Значение критерия вычисляется следующим образом:

$$AIC_C = AIC + 2k\frac{k+1}{n-k-1}$$

где k – число параметров модели, n — объем обучающей выборки, AIC – информационный критерий Акаике.

Для моделей, построенных на выборочных данных, характерны смещенные оценки всех параметров и статистик. Поэтому для компенсации смещения используются специальные критерии и методы, к которым относится и модифицированный критерий Акаике. Несложно увидеть, что критерий «штрафует» модель на величину, пропорциональную отношению числа параметров модели к числу наблюдений выборки.

- определяется с помощью байесовского информационного критерия (Bayesian information criterion – BIC). Критерий выбора статистической модели из некоторого конечного набора. Предпочтение отдается модели с минимальным значением критерия.

Критерий основан на использовании функции правдоподобия и тесно связан с информационным критерием Акаике. Однако, поскольку при разработке подхода автор использовал и адаптировал идеи Байеса, за критерием закрепилось два варианта названия — Шварца и Байеса.

В основе подхода лежит тот факт, что при увеличении числа параметров модели значение функции правдоподобия растет, но при этом возможно наступление эффекта переобучения. Когда параметров модели оказывается

слишком много, доля каждого из них в объясняющей способности модели становится малой и они теряют свою значимость.

Поэтому задача выбора модели заключается в том, чтобы включить в нее минимум параметров, которые, тем не менее, вносили бы наибольший вклад в значение функции правдоподобия. Значение критерия вычисляется по формуле:

$$BIC = k \times \ln(n) - 2\ln(\hat{L})$$

где $\hat{L}$ – максимальное значение функции правдоподобия наблюдаемой выборки с известным числом параметров, k – число параметров модели, n – объем обучающей выборки.

BIC широко применяется для анализа временных рядов и решения задач линейной регрессии.

Фиксированное значение порога одинаково для всех переменных. Пороги, устанавливаемые с помощью AIC (критерия Акаике) и BIC (байесовского информационного критерия), могут быть индивидуальными для разных переменных.

Значение порога, определяемого AIC, зависит от числа степеней свободы переменной. Например, если переменная бинарная и имеет число степеней свободы, равное 1, то для включения в модель она должна иметь p-значение меньше 0,157.

Критерий BIC определяет порог в зависимости от размера выборки n. Например, для $n = 20$ переменной потребуется значение $p < 0{,}083$, чтобы войти в модель. Чем больше n, тем ниже будет порог.

BIC является более строгим критерием, чем AIC, и дает модели меньшего размера. Поэтому его рекомендуется использовать только при работе с большими выборками, когда число наблюдений превышает 100 на одну независимую переменную.

Обратное исключение (Backward elimination)

Алгоритм отбора начинает работу с модели, содержащей все переменные (такая модель называется «полной»). Затем начинает удалять наименее значимые переменные одну за другой до тех пор, пока не будет достигнуто предварительно заданное правило остановки, или пока в модели не останется ни одной переменной. Как и в случае прямого отбора требуется определить наименее значимую переменную на каждом шаге и правило остановки.

Очевидно, что первыми кандидатами на исключение являются переменные, которые наименее способствуют повышению качества модели. Аналогично методу прямого включения для оценки значимости изменения качества модели может быть использован критерий Фишера: лучшим кандидатом на исключение

будет та переменная, для которой значение критерия Фишера выше заданного порога.

Рис. 2. Метод обратного исключения

Наименее значимой является переменная:

- с которой связано наибольшее p-значение;
- исключение которой из модели вызывает наименьшее сокращение коэффициента детерминации R-квадрат;
- исключение которой из модели вызывает наименьшее увеличение RSS (суммы квадратов остатков) по сравнению с другими признаками.

Выбор правила остановки

Правило остановки выполняется, когда все оставшиеся переменные в модели имеют p-значение меньше некоторого заранее заданного порога. Когда модель достигнет этого состояния, алгоритм обратного исключения завершится.

Как и в случае прямого выбора, порог может быть:

- фиксированным значением (например: 0.05, 0.2 или 0.5);
- определяется AIC (критерия Акаике);
- определяется BIC (байесовского информационного критерия).

Прямой отбор предпочтительно использовать, когда количество рассматриваемых переменных велико. Это связано с тем, что он начинается с нулевой модели и продолжает добавлять переменные по одной, и поэтому, в отличие от обратного отбора, он не рассматривает полную и близкие к ней модели.

Обратный отбор предпочтительно использовать если нужно рассмотреть полную модель, когда одновременном учитываются все переменные. При обратном отборе кандидаты на исключение могут и не появиться и все переменные останутся в модели.

Преимущества пошагового отбора:

- простота реализации;
- улучшение интерпретируемости модели;
- снижение вычислительных затрат за счёт того, что рассматриваются не все переменные;
- объективность – автоматический выбор позволяет избежать субъективности экспертных оценок.

Особенно оказываются полезными методы пошагового отбора в случае разведочного анализа данных, когда априорные сведения о решаемой задаче отсутствуют.

Недостатки пошагового отбора:

- не рассматривает все возможные комбинации переменных, поэтому не гарантирует лучшего их набора;
- приводит к смещенным оценкам коэффициентов регрессии, доверительных интервалов, p-значений и коэффициента R-квадрат;
- формирует нестабильный набор переменных, особенно в случае, когда число переменных сравнимо с числом наблюдений. Это возможно когда разные наборы переменных одинаково воздействуют на выходную переменную и выражается в том, что каждый раз получается разный набор переменных. Чтобы избежать данного эффекта требуется, чтобы число наблюдений выборки на одну входную переменную было 50 и выше.
- не учитывает причинно-следственные связи между переменными.

Пошаговое включение/исключение (Stepwise)

В широком смысле «пошаговая» относится ко всем автоматическим методам отбора переменных с помощью их последовательного включения или исключения. В узком смысле «пошаговая» относится к технике, которая представляет собой комбинацию прямого и обратного отбора. Её часто называют также Двунаправленный отбор (Bidirectional forward/elimination). В дальнейшем

будем использовать «пошаговая» именно для обозначения этой техники (т.е. двунаправленный отбор или пошаговый отбор — синонимы).

В основе идеи пошаговой технологии лежит предположение, что признаки могут быть коррелированными. Это приводит к тому, что включение в модель новых переменных может вызывать снижение значимости ранее включенных переменных. И если это снижение значимости сильнее некоторого критического, то ранее включенные переменные следует удалить из модели.

Иными словами, при пошаговом методе чередуются шаги прямого и обратного отбора, добавляя и удаляя переменные, которые соответствуют критериям для включения или исключения, до тех пор, пока не будет достигнут стабильный набор переменных.

Останов алгоритма производится при достижении порога, заданного критерием Маллоуза:

$$C_p = \frac{SSE_k}{MSE} - n + 2k + 2$$

где $SSE_k = \sum_{i=1}^{m}(y_i - \hat{y}_i)^2$ – (SSE — sum square error, сумма квадратов ошибки) модели, содержащей k переменных, $MSE = \frac{1}{n}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$ – средняя сумма квадратов ошибок регрессии для полной модели.

Очевидно, что критерий штрафует модели с большей сложностью (числом переменных). Действительно, чем больше переменных в модели, тем меньше её ошибка и, соответственно, значение в числителе. Поэтому для модели с большим числом переменных значение критерия будет меньше. Минимизация критерия позволяет найти подмножество наиболее значимых переменных. К недостаткам метода, можно отнести то, что важные переменные могут никогда не включаться в модель, а второстепенные будут включены.

Ridge (Гребневая регрессия)

В матричном виде формула для квадрата остатков регрессии может быть записана в виде:

$$S = (y - bX)^{\mathrm{T}}(y - bX)$$

Дифференцируя эту функцию по вектору параметров b и приравняв производные к нулю, получим систему уравнений в матричной форме:

$$(X^{\mathrm{T}}X)b = X^{\mathrm{T}}y$$

Решение этой системы уравнений и дает общую формулу оценок метода наименьших квадратов для модели линейной регрессии. Плохая обусловленность матрицы $\sum X^{\mathrm{T}}X$ приводит к неустойчивости решения уравнения линейной регрессии. Причиной плохой обусловленности матрицы является корреляция между независимыми переменными.

Неустойчивость решения проявляется в том, что даже небольшие изменения в исходных данных приводит к значительным изменениям параметров регрессионной модели. В результате, на практически одних и тех же данных могут быть построены существенно отличающиеся модели.

Чтобы повысить устойчивость решения применяется специальный математический метод, называемые регуляризацией (а именно, регуляризация по А.Н. Тихонову или L2 - регуляризация). В основе идеи регуляризации лежит применение так называемого регуляризующего функционала с помощью которого на решение накладываются ограничения. При этом улучшается обусловленность матрицы $\sum X^{\mathrm{T}} X$.

В случае линейной регрессии это означает, что параметры модели оцениваются не с помощью минимизации функционала (1), а с помощью минимизации функционала, в который введён специальный элемент — параметр регуляризации, который обычно обозначается α:

$$\mathbf{b}^* = \underset{\mathbf{b}}{argmin} \left(\sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} b_i x_{ij} \right)^2 + \alpha \left\| \mathbf{b} \right\|^2 \right).$$

Увеличение параметра α приводит к уменьшению нормы вектора параметров модели. Проинтерпретируем метод гребневой регрессии графически (рис. 3).

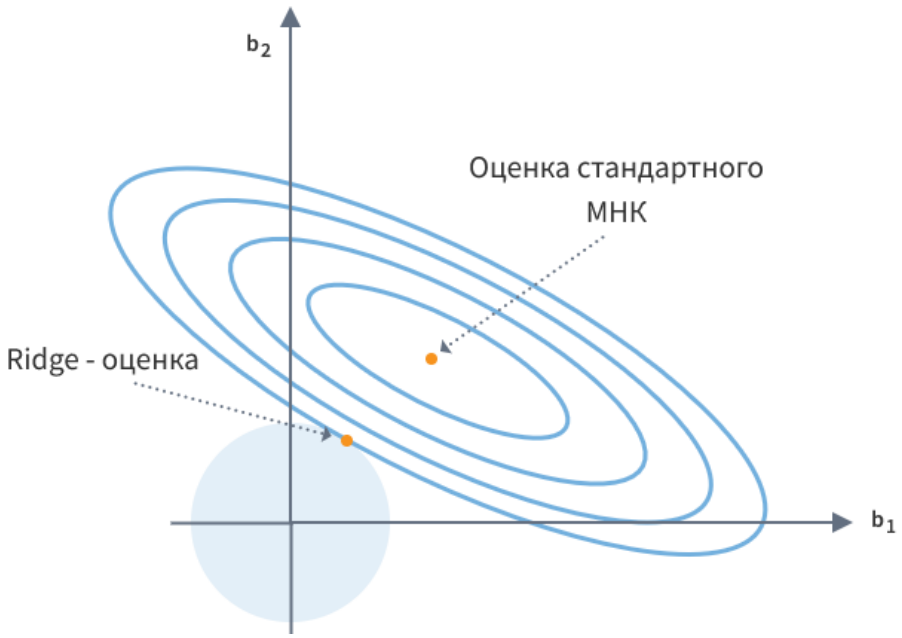

Рис. 3. Гребневая регрессия

На рисунке показано пространство параметров модели. Критерий S – квадратичная функция относительно параметров b, поэтому кривая $S = const$ является эллипсоидом. Регуляризирующий параметр, отличный от нуля, задает

сферу в этом пространстве. Точка касания эллипсоида и сферы является решением нормального уравнения при фиксированном α. При этом касание эллипсоида в нулевой точке исключено и обнуления параметров модели не происходит. Метод улучшает устойчивость параметров регрессионной модели, но не приводит к обращению в ноль ни одного из них.

Следует отметить, что в результате корректировки оценок параметров модели при использовании гребневой регрессии они никогда не принимают нулевых значений, поэтому гребневая регрессия не является методом отбора переменных. С её помощью производится корректировка оценок параметров регрессионной модели с целью повышения её устойчивости, снижающейся из-за корреляции признаков набора данных.

Регрессия LASSO

Ещё одним методом оценивания параметров модели линейной регрессии с использованием регуляризации является метод LASSO (Least absolute shrinkage and selection operator – оператор наименьшего абсолютного сокращения и выбора). В отличие от гребневой регрессии оценки параметров, которые даёт регрессия LASSO, могут принимать нулевые значения. Таким образом, данный метод можно рассматривать и как регуляризацию с целью повышения точности, и как процедуру отбора переменных.

Метод LASSO использует ограничение на сумму абсолютных значений параметров модели. Рассматривается сумма модулей параметров модели:

$$T(b) = \sum_{j=1}^{n} |b_j|$$

Параметры регрессии выбираются из условия минимизации критерия (1)

$$S = \sum_{j=1}^{n} \left(y_i - f(b_s x_i)\right)^2 \quad (1)$$

при ограничении $T(b) \leq t$, где t – параметр регуляризации.

При больших t решение совпадает с решением, полученным методом наименьших квадратов. Чем меньше t, тем больше коэффициентов регрессии принимают нулевое значение.

Графическая интерпретация метода LASSO представлена на рис. 4.

Эллипсоид, как и в случае гребневой регрессии (рис. 4) образован точками, в которых сумма квадратов остатков регрессии, минимизируемая в процессе решения, постоянна. Параметр t, отличный от нуля, задает многомерный октаэдр. Точка касания эллипсоида и октаэдра является решением стандартного уравнения регрессии при фиксированном t. При касании эллипсоида и ребра октаэдра происходит обнуление коэффициента.

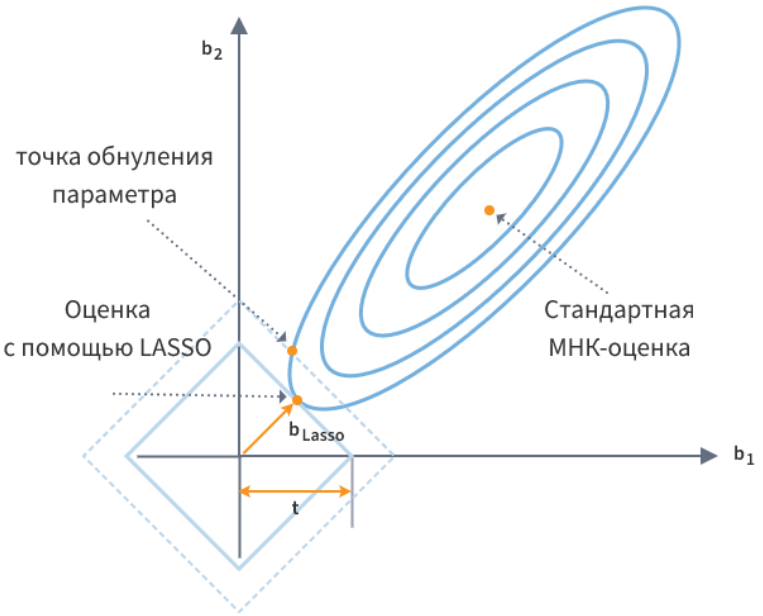

Рис. 4. Регрессия LASSO

Таким образом, LASSO (регрессия Лассо) — также как и Ridge, применяется для регуляризации (защиты от переобучения) обучаемой модели. Подразумевает введение "штрафов" (вычисляются как сумма модулей коэффициентов переменных умноженные на коэффициент регуляризации) для уменьшения значений коэффициентов регрессии. Регуляризация Lasso (L1) позволяет снизить размерность и упростить регрессионную модель, за счёт зануления коэффициентов некоторых признаков.

Регрессия «Эластичная сеть»

В рассмотренных выше методах регуляризации регрессионной модели (гребневая и LASSO) используется единственный регуляризатор. Метод «Эластичная сеть» комбинирует обе эти техники, что позволяет преодолеть присущие им недостатки (рис. 5).

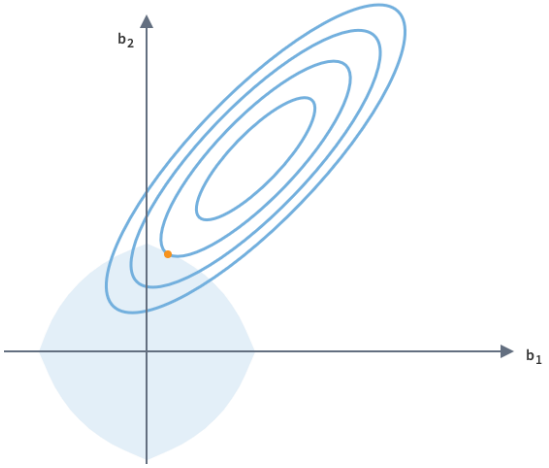

Рис. 5. Регрессия «Эластичная сеть»

Например, если в наборе данных присутствует большое число признаков и малое число наблюдений, то метод LASSO может включать в модель «лишние» переменные. И, наоборот, если переменные модели сильно коррелированы, то LASSO выбирает только одну переменную.

Метод эластичной сети использует два регуляризующих члена:

$$\mathbf{b}^* = \underset{\mathbf{b}}{argmin} \left(\sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} b_i x_{ij} \right)^2 + \alpha_1 \|\mathbf{b}\|^2 + \alpha_2 \|\mathbf{b}\| \right).$$

Квадратичный член делает целевую функцию более выпуклой и имеющей ярко выраженный минимум. Метод эластичной сети содержит два шага. Сначала фиксируется значение первого регуляризующего члена, т.е. ищутся оценки коэффициентов для гребневой регрессии. А затем производится их сокращение с помощью LASSO.

Пример использования методов отбора

Рассмотрим пример работы методов отбора переменных на реальных данных о заёмщике банка. Используемые признаки представлены таблице 1.

Таблица 1. Описание набора данных о заёмщике

Признак	Обозначение	Тип
Количество просрочек	y	Зависимая переменная
Стаж на последнем месте работы	x_1	Независимая переменная
Срок кредита	x_2	Независимая переменная
Сумма кредита	x_3	Независимая переменная

Данные о 10 заёмщиках по описанным в таблице 1 признакам, представлены в таблице 2.

Таблица 2. Исходные данные для модели линейной регрессии

ID заёмщика	Кол-во просрочек (y)	Стаж, лет (x_1)	Срок кредита, мес.(x_2)	Сумма кредита, руб (x_3)
1	0	7.5	12	170 000
2	0	4.5	12	120 000
3	0	6.5	12	85 000
4	1	2.5	12	160 000
5	1	3.5	24	105 000
6	0	6.5	12	90 000
7	3	2.0	24	80 000
8	2	3.5	24	395 000
9	2	6.0	36	150 000
10	4	2.0	60	70 000

Метод прямого отбора

Начинаем с пустой модели. Первым признаком, который будет выбран в качестве переменной регрессионной модели, будет тот, который сильнее коррелирован с независимой переменной. Рассмотрим таблицу 3, в которой для каждой независимой переменной представлен коэффициент корреляции в зависимой.

Таблица 3. Корреляция между независимой переменной и зависимыми

Независимая переменная	Коэффициент корреляции
x_1	-0,721
x_2	0,871
x_3	0,018

Из таблицы 3 несложно увидеть, что наибольшая линейная зависимость наблюдается между независимой переменной и переменной x_2, т.е. между количеством просрочек и сроком кредита. При этом корреляция положительная, т.е. с ростом срока кредита число просрочек растёт. Поэтому первой переменной, которая будет включена в модель будет именно x_2.

Рассчитаем регрессионные оценки для модели, содержащей единственную переменную.

Сумма квадратов остатков для модели с единственной независимой переменной x_2 будет $S(x_2) = 4{,}38$. Если добавить в модель переменную x_1, то $S(x_2, x_1) = 2{,}07$. Если добавить в модель переменную x_3, то $S(x_2, x_3) = 4{,}13$.

Рассчитаем значения критерия Фишера для модели, включающей переменные x_2 и x_1.

$$S(x_2, x_1) = \frac{4{,}38 \times 2{,}07}{2{,}07} \times \frac{10-2}{2-1} = 1{,}1 \times 8 = 8{,}8$$

Зададимся уровнем значимой вероятности $p = 0.05$. Это означает, что вероятность ошибочного отклонения гипотезы о значимости новой переменной не превышает 5%.

Далее нам необходимо воспользоваться таблицами критических значений распределения Фишера. Фрагмент такой таблицы представлен в таблице 4.

Таблица 4. Критические значения распределения Фишера

$k_1 k_2$	1	2	3	4	5	6
1	161.45	199.50	215.72	224.57	230.17	233.97
2	18.51	19.00	19.16	19.25	19.30	19.33
3	10.13	9.55	9.28	9.12	9.01	8.94
4	7.71	6.94	6.59	6.39	6.26	6.16
5	6.61	5.79	5.41	5.19	5.05	4.95

6	5.99	5.14	4.76	4.53	4.39	4.28
7	5.59	4.74	4.35	4.12	3.97	3.87
8	5.32	4.46	4.07	3.84	3.69	3.58
9	5.12	4.26	3.86	3.63	3.48	3.37
10	4.96	4.10	3.71	3.48	3.33	3.22

Строки и столбцы таблицы образованы числами степеней свободы, которых у распределения Фишера два (в таблице они обозначены как k_1 и k_2). При этом $k_2 = n - 2 = 8$, где n – число наблюдений в наборе данных, на котором строится модель, а $k_1 = m - 1 = 1$, где m – число свободных (независимых) переменных модели после добавления новой переменной.

На пересечении столбца для $k_1 = 1$ и строки для $k_2 = 8$ находим в таблице 4 критическое значение 5,32 (выделено красным цветом). Если рассчитанное значение критерия выше критического, то гипотезу о том, что новая переменная не увеличивает значимо точность модели должно быть отвергнуто. При этом вероятность обратного не превышает 5%. Поскольку рассчитанное значения F-критерия превышает критическое, можно сделать вывод о значимом улучшении качества модели при добавлении в неё переменной x_1.

Проведём аналогичную проверку для переменной x_3:

$$S(x_2, x_3) = \frac{4{,}38 - 4{,}13}{4{,}13} \times \frac{10 - 2}{2 - 1} = 0{,}06 \times 8 = 0{,}48$$

Данное значение меньше критического значения F-распределения, что позволяет отклонить гипотезу о значимости улучшения модели при добавлении в неё переменной x_3. Таким образом, мы получили, что из двух переменных-кандидатов на включение в модель x_1 и x_3, только первая из них обеспечивает значимое улучшение качества модели и может быть включена в модель.

Метод обратного исключения

Начинаем с полной модели, которая содержит все признаки, доступные в наборе данных. Требуется произвести проверку, которая позволит определить нельзя ли исключить из модели какие-то переменные без значимого ухудшения её качества. Для этого найдём переменную, с которой связано минимальным значением F-критерия, найденного при условии, что остальные переменные включены в модель.

Переменная	F
x_1 (Стаж)	7,92
x_2 (Срок кредита)	24,36
x_3 (Сумма)	0,47

Из таблицы видно, что первым кандидатом на исключение является переменная x_3. По таблице F-распределения определим, что для $k_1 = 10 - 3 = 7$ и $k_2 = 3 - 2 = 1$ критическое значение $F_{\text{кр}} = 5{,}59$ (выделено зеленым цветом). Значение F-критерия для переменной x_3 меньше критического, что подтверждает предположение о низкой значимости переменной и приводит к выводу о целесообразности её исключения из модели.

Проведём аналогичные действия для оставшихся переменных, учитывая, что $F_{\text{кр}} = 5{,}32$

Переменная	F
x_1 (Стаж)	8,87
x_2 (Срок кредита)	27,05

Таким образом, значение F-критерия для остальных переменных превышает критическое, что позволяет сделать вывод о нецелесообразности их исключения из модели. При этом связанное с переменной x_2 значение F-критерия значительно превышает значение для x_1. Это говорит о том, что значимость переменной x_2 с точки зрения повышения точности модели, существенно выше, чем x_1, что делает её исключение наименее целесообразным.

Таким образом, порядок кандидатов на исключение следующий: x_3, x_1, x_2, что согласуется с результатами метода прямого отбора, полученными выше.

Пошаговый отбор

1. Осуществляем прямой ход процедуры пошагового отбора, т.е. первый шаг прямого отбора. Как показано выше, его результатом является включение в модель переменной x_2. Поскольку на данном шаге регрессионная модель не содержит других переменных, обратный ход процедуры пошагового отбора не выполняется.

2. Рассматриваем следующую переменную-кандидата на включение в модель. Это будет переменная x_1, значимость которой была показана при рассмотрении метода прямого включения. После включения в модель новой переменной, переменная включенная ранее может потерять свою значимость и её использование в модели теряет смысл. Выяснить, потеряла ли переменная x_2 значимость «на фоне» x_1 и предстоит на фазе обратного хода алгоритма отбора.

3. Для проверки целесообразности оставления переменной x_2 на обратном ходе, нужно оценить значимость увеличения суммы квадратов остатков регрессии при её исключении. Для этого определим соответствующее значение F-критерия.

Переменная	S
$x_2 \setminus x_1$	2,07
x_1	8,68

$$F = \frac{8{,}68 - 2{,}07}{8{,}68} \times \frac{10 - 2}{2 - 1} = 6{,}1$$

Данное значение превышает соответствующее критическое значение F-распределения $F_{\text{кр}} = 5{,}32$, поэтому можно считать что исключение переменной x_2 на обратном ходе алгоритма значимо ухудшает точность модели и, следовательно, нецелесообразно.

Поскольку переменная x_3 в прямом включении не смогла показать значимость, достаточную для включения в модель, то использовать её в процедуре пошагового отбора также не целесообразно и поэтому она завершает свою работу.

И так были рассмотрены методы отбора переменных, наиболее широко применяемые в статистических моделях линейной регрессии. Однако эта проблема актуальна и для других задач и типов моделей анализа данных — кластеризации, классификации, прогнозирования и т.д. Во всех случаях включение в модель избыточных и незначимых переменных приводит к возрастанию сложности модели без адекватного увеличения её качества (а иногда приводит и к его ухудшению).

Поэтому отбор переменных можно рассматривать как часть более общей задачи снижения размерности (dimensionality reduction) пространства признаков. Она позволяет не только отобрать наиболее значимые переменные, но и обойти «проклятие размерности». Помимо отбора признаков, задача снижения размерности включает проецирование признаков, где производится попытка выразить информацию, содержащуюся в наборе исходных признаков с помощью меньшего числа новых переменных, представляющих линейные комбинации исходных. Сюда входят: факторный анализ и метод главных компонент, линейный и обобщённый дискриминантный анализ, корреляционный анализ и др.

Таким образом, технологии отбора переменных для аналитических моделей не ограничиваются описанными в данной лекции, и могут применяться не только в рамках статистики, но и в машинном обучении. Важно лишь правильно выбрать метод, адекватный решаемой задаче, корректно его применить и проинтерпретировать результаты.