Лекция 3
Линейная множественная регрессия

1. Линейная модель множественной регрессии.
2. Ранжирование факторов.
3. Оценка качества уравнения множественной регрессии.
4. Частные критерии.

1. Линейная модель множественной регрессии

Если любая парная регрессия статистически незначима, то следует искать зависимость объясняемой переменной либо от другого фактора, либо от нескольких факторов.

В последнем случае задача решается с помощью *множественного регрессионного анализа*.

Множественный регрессионный анализ является обобщением парного, однако здесь появляются новые проблемы, из которых следует выделить две.

Первая из них связана со спецификацией модели, которая теперь включает в себя *отбор факторов* и выбор *вида уравнения*.

При отборе факторов необходимо ответить на вопрос: какие факторы существенно влияют на , а какие – несущественно, и последние не следует включить в регрессию.

Вторая проблема связана с исследованием влияния конкретной независимой переменной на признак y, т.е. разграничения её воздействия от влияния других независимых переменных.

y

Будем далее считать, что факторы

$$x_1, x_2, ..., x_p$$

отобраны правильно, а в качестве уравнения связи с признаком y выбрана наиболее употребляемая и простая линейная модель множественной регрессии

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon$$

Для построения модели требуются исходные статистические данные в виде следующей многомерной выборки ($n > 6p$):

Номер измерения	Переменные				
	y	x_1	x_2	…	x_p
1	y_1	x_{11}	x_{21}	…	x_{p1}
2	y_2	x_{12}	x_{22}	…	x_{p2}
…	…	…	…	…	…
n	y_n	x_{1n}	x_{2n}	…	x_{pn}

Тогда наблюдаемые значения переменных должны удовлетворять уравнению

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + \varepsilon_i, i = \overline{1,n}\ , \qquad (1)$$

где y_i – значение признака y в i – м наблюдении, x_{ji} – значение j –го фактора в i – м наблюдении, ε_i – случайная составляющая в i – м наблюдении.

Оценкой уравнения (1) по выборке является *выборочное* уравнение регрессии

$$\tilde{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p \qquad (2)$$

В дальнейшем удобнее использовать матричные обозначения. Поэтому введем в рассмотрение следующие матрицы и векторы:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \boxtimes \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{21} & \boxtimes & x_{p1} \\ 1 & x_{12} & x_{22} & \boxtimes & x_{p2} \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 1 & x_{1n} & x_{2n} & \boxtimes & x_{pn} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \boxtimes \\ \varepsilon_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \boxtimes \\ \beta_p \end{pmatrix}, \quad b = \begin{pmatrix} b_0 \\ b_1 \\ \boxtimes \\ b_p \end{pmatrix}, \quad x = \begin{pmatrix} 1 \\ x_1 \\ \boxtimes \\ x_p \end{pmatrix}$$

Тогда уравнение регрессии (1) в матричной форме запишется

$$Y = X\beta + \varepsilon,$$

а выборочное уравнение (2) примет вид

$$\tilde{y} = b'x.$$

Отсюда нетрудно получить

$$Y = Xb + e,$$

где

$$e = Y - Xb.$$

Чтобы получить оценку вектора b методом наименьших квадратов, дополнительно к предпосылкам МНК для парной регрессии 1° -5° здесь требуется выполнение ещё одного условия:

6°. Столбцы матрицы X должны быть линейно-независимы, т.е. ранг матрицы X должен быть равен $p+1$ (числу столбцов).

При выполнении указанных предпосылок (1° - 6°) искомый вектор определяется из системы нормальных уравнений в матричной форме:

$$X'Xb = X'Y.$$

Решением этого уравнения является МНК - оценка

$$b = (X'X)^{-1} X'Y. \qquad (4)$$

2. Ранжирование факторов

Если бы все объясняющие переменные x_j в уравнении (2) измерялись в одних и тех же единицах, например, в кг, то непосредственно сопоставляя абсолютные значения коэффициентов регрессии $|b_j|$ можно было ранжировать факторы x_j по силе их воздействия на признак y.

Чем больше $|b_j|$, тем сильнее фактор x_j влияет на y.

Однако в общем случае переменные x_j имеют различные единицы измерения и такое ранжирование невозможно (ошибочно). В этом случае прибегают к нормированию коэффициентов регрессии – вычислению *стандартизованных коэффициентов регрессии* a_j по следующей формуле

$$a_j = b_j \frac{\sigma_{x_j}}{\sigma_y}, \quad j = \overline{1, p}, \qquad (5)$$

где σ_{x_j}, σ_y – средние квадратические отклонения переменных x_j, y соответственно.

Коэффициент a_j показывает, на сколько в среднем σ_y изменится переменная y, если соответствующий фактор x_j увеличится на одно σ_{x_j} при неизменном среднем уровне других факторов модели.

Имея значения a_j, можно построить уравнение множественной регрессии в *стандартизованном масштабе*

$$t_y = a_1 t_{x_1} + a_2 t_{x_2} + \ldots + a_p t_{x_p}, \qquad (6)$$

где $t_y = \dfrac{y - \bar{y}}{\sigma_y}$ $t_{x_j} = \dfrac{x_j - \bar{x}_j}{\sigma_{x_j}}$, $j = \overline{1, p}$ –

стандартизованные переменные, для которых средние значения равны нулю ($\bar{t}_y = \bar{t}_{x_j} = 0$), а средние квадратические отклонения равны единице ($\sigma_{t_y} = \sigma_{t_{x_j}} = 1$).

Чем больше модуль $|a_j|$, тем сильнее влияние фактора x_j на признак y, т.е. по значениям $|a_j|$ можно выполнить непосредственное ранжирование факторов по силе их воздействия на y.

От уравнения вида (6) можно перейти к уравнению регрессии *в натуральном масштабе* (2), используя формулы:

$$b_j = a_j \frac{\sigma_y}{\sigma_{x_j}}, \quad j = \overline{1, p},$$

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \ldots - b_p\bar{x}_p.$$

Для оценки влияния отдельных факторов на переменную также можно использовать средние коэффициенты эластичности

$$\overline{Э}_j = b_j \frac{\overline{x}_j}{\overline{y}}, \quad j = \overline{1, p}.$$

Ранжирование факторов по силе воздействия на можно выполнить также с помощью *частных коэффициентов корреляции*.

Коэффициент частной корреляции характеризует тесноту линейной связи между признаком y и фактором x_j при устранении (элиминировании) влияния других факторов, включенных в модель.

Различают коэффициенты частной корреляции 1, 2, …,($p-1$) – го порядков, если рассматривается регрессия с числом факторов, равным p.

Например, частными коэффициентами корреляции являются:

$r_{yx_1 \cdot x_2}$ –коэффициент первого порядка, учитывающий связь y и фактора x_1 при неизменном действии фактора x_2 ;

$r_{yx_1 \cdot x_2 x_3}$ –коэффициент 2-го порядка, учитывающий связь y и фактора x_1 при неизменном действии факторов x_2, x_3 ;

Отсюда коэффициент парной корреляции r_{yx_1} можно рассматривать как частный коэффициент 0-го порядка. Коэффициенты частной корреляции более высоких порядков определяются через коэффициенты низких порядков. Для случая, когда вычисляют два частных коэффициента первого порядка:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{(1 - r_{yx_2}^2)(1 - r_{x_1x_2}^2)}}, \quad r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{(1 - r_{yx_1}^2)(1 - r_{x_1x_2}^2)}}$$

При этом существует связь между частными коэффициентами корреляции и стандартизованными коэффициентами регрессии:

$$r_{yx_1 \cdot x_2} = a_1 \sqrt{\frac{1 - r^2_{x_1 x_2}}{1 - r^2_{yx_2}}}, \quad r_{yx_2 \cdot x_1} = a_2 \sqrt{\frac{1 - r^2_{x_1 x_2}}{1 - r^2_{yx_1}}}$$

3. Оценка качества уравнения множественной регрессии

По аналогии с парной регрессией можно определить долю результата , объясненной вариацией включенных в модель факторов в его общей дисперсии:

$$R^2 = \frac{Q_R}{Q} = 1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{u=1}^{n}(y_i - \bar{y})^2}.$$

Величину R^2 называют *коэффициентом множественной детерминации.* Он служит измерителем качества подбора уравнения. Его значения изменяются в пределах от 0 до 1, и чем ближе R^2 единице, тем больше уравнение регрессии объясняет поведение .

Кроме коэффициента R^2 используют другой показатель качества – *коэффициент множественной корреляции*

$$R = \sqrt{R^2},$$

который представляет собой обобщение парного коэффициента корреляции r_{yx} и характеризует совместное (совокупное) влияние всех факторов на результат y. В отличие от r_{yx} коэффициент множественной корреляции R принимает значения от 0 до 1 и не может быть использован для интерпретации *направления* связи.

Коэффициент R^2 является неубывающей функцией числа объясняющих переменных. Если добавить в модель фактор, который совсем не влияет на y, то R^2 обязательно автоматически увеличится. Этот недостаток можно устранить, если определять показатель не через суммы квадратов, а через дисперсии на одну степень свободы. В результате получаем *скорректированный* (нормированный) *коэффициент множественной детерминации*:

$$\hat{R}^2 = 1 - \frac{\frac{\sum_{i=1}^{n} e_i^2}{(n-p-1)}}{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{(n-1)}} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

Доказано, что $\hat{R}^2$ увеличивается при добавлении нового фактора в модель тогда и только тогда, когда модуль t –статистики параметра по этой переменной больше единицы. Значение $\hat{R}^2$ может даже **уменьшится** при добавлении нового фактора.

Проверка статистического качества модели выполняется путем проверки совокупной значимости её коэффициентов, т.е. проверки гипотезы:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0.$$

На практике вместо указанной гипотезы проверяют тесно связанную с ней гипотезу о статистической значимости коэффициента детерминации R^2 :

$$H_0 : R^2 = 0$$

Для проверки данной гипотезы используется статистика:

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - p - 1}{p}, \qquad (7)$$

которая имеет распределение Фишера.

Найденное по формуле (7) значение F сравнивается с $F_{кр}$, которое находится по таблицам по заданному уровню значимости α и числу степеней свободы $k_1 = p$ и $k_2 = n - p - 1$. Если $F > F_{кр}$, то гипотеза H_0 отклоняется и это равносильно статистической значимости уравнения в целом.

Как и в случае парной регрессии выполняется статистическая значимость отдельных коэффициентов β_j уравнения на основе t – статистик:

$$t_{b_j} = \frac{b_j}{m_{b_j}}, \quad j = \overline{0, p}, \tag{8}$$

где m_{b_j} – стандартная ошибка параметра b_j, вычисляемая по формуле:

$$m_{b_j} = s\sqrt{\left[(X'X)^{-1}\right]_{j+1\,j+1}},$$

Здесь $\left[(X'X)^{-1}\right]_{j+1\,j+1}$ –диагональный элемент обратной матрицы $(X'X)^{-1}$, стоящий на пересечении $(j+1)$–й строки и $(j+1)$– го столбца; s^2 –несмещенная оценка дисперсии σ^2 возмущения ε, определяемая по формуле:

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p-1}.$$

Если $\left|t_{b_j}\right| > t_{кр}$, где $t_{кр}$ находится из таблиц по значению $\alpha / 2$ и числу степеней свободы $k = n - p - 1,$ то коэффициент β_j считается статистически значимым.

Приведенную строгую проверку значимости коэффициентов можно заменить простым сравнительным анализом ("грубое" правило):

- если $\left|t_{b_j}\right| \le 1$, то b_j статистически незначим;

- если $1 < \left|t_{b_j}\right| \le 2$, то b_j *относительно* значим, и для уточнения следует воспользоваться строгой методикой;
- если $2 < \left|t_{b_j}\right| \le 3$, то b_j статистически значим;

- если $\left|t_{b_j}\right| > 3$, то b_j считается сильно значимым и вероятность ошибки вывода не превосходит 0,001.

Так же как и в парной регрессии для статистически значимых коэффициентов модели можно построить интервальные оценки:

$$b_j - t_{кр} m_{b_j} \le \beta_j \le b_j + t_{кр} m_{b_j},$$

где $t_{кр}$ – прежнее значение критической точки распределения.

Доверительный интервал можно построить и для индивидуальных прогнозных значений зависимой переменной y .

Зафиксируем значения прогнозных объясняющих переменных

$$x_{10}, x_{20}, \square \quad , x_{p0}$$

и по вектору-столбцу

$$x_0 = (1 \quad x_{10} \quad x_{20} \square \quad x_{p0})'$$

найдем прогнозное значение зависимой переменной y :

$$\tilde{y}_0 = b'x_0 = b_0 + b_1 x_{10} + b_2 x_{20} + \square \quad + b_p x_{p0}$$

Тогда доверительный интервал для индивидуального прогнозного значения y_0 в точке x_0 примет вид

$$\tilde{y}_0 - t_{кр} m_{\tilde{y}_0} \le y_0 \le \tilde{y}_0 + t_{кр} m_{\tilde{y}_0},$$

где стандартная ошибка $\tilde{y}_0$ вычисляется по формуле:

$$m_{\tilde{y}_0} = s\sqrt{1 + x_0'(X'X)^{-1} x_0}$$

4. Частные F _ критерии

Не каждый фактор, дополнительно включаемый в модель, может существенно увеличить долю объясненной вариации зависимой переменной. Ввиду корреляции между факторами значимость одного и того же фактора может быть различной в зависимости от последовательности включения его в модель.

Мерой оценки значимости улучшения качества модели, в которой были включены факторы $x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots x_p$, после включения в неё дополнительно фактора x_j, служит частный F – критерий:

$$F_{x_j} = \frac{R^2 - R_j^2}{1 - R^2}(n - p - 1),$$

где R_j^2 – коэффициент множественной детерминации для модели без фактора x_j, R^2 – тот же коэффициент с включенным в модель фактором x_j.

Если в модели $p = 2$, то используются два частных критерия:

$$F_{x_1} = \frac{R^2 - r^2_{yx_2}}{1 - R^2}(n-3), \qquad (9)$$

$$F_{x_2} = \frac{R^2 - r^2_{yx_2}}{1 - R^2}(n-3).$$

Фактические значения частных критериев , найденных по формуле (9), сравнивается с $F_{кр}$, определяемое по таблицам распределения Фишера по заданному уровню значимости α и числам степеней свободы $k_1 = 1$ и $k_2 = n - 3$.

Если, например, $F_{x_1} > F_{кр}$, то включение фактора x_1 в модель, после того как в уравнение уже включен фактор x_2 , статистически оправдано и параметр b_1 при факторе x_1 статистически значим.

В противном случае дополнительное включение в модель фактора x_1 не увеличивает существенно долю объясненной вариации y и, следовательно, включение фактора x_1 в модель нецелеобразно.

По аналогичной схеме проверяется целесообразность включения (или исключения) не одного, а группы факторов.

Пусть по n наблюдениям построено уравнение регрессии с p факторами и коэффициент множественной детерминации равен R_1^2 . Дополнительно в модель включают ещё k факторов и коэффициент детерминации при этом составит величину R_2^2 ($R_2^2 \geq R_1^2$). Тогда проверяется гипотеза $H_0 : R_1^2 = R_2^2$ с помощью статистики

$$F = \frac{R_2^2 - R_1^2}{1 - R_2^2} \cdot \frac{n - p - 1}{k}.$$

Если $F > F_{кр}(\alpha; k_1 = k; k_2 = n - p - 1)$, то гипотеза H_0 отклоняется и одновременное включение k факторов в модель обоснованно.

Если из модели одновременно исключаются k факторов, то используют статистику

$$F = \frac{R_1^2 - R_2^2}{1 - R_1^2} \cdot \frac{n - p - 1}{k},$$

где $R_1^2, R_2^2 \ (R_1^2 \geq R_2^2)$ – коэффициенты детерминации с p и $(p - k)$ факторами соответственно.

Если при этом $F > F_{кр}(\alpha; k_1 = k; k_2 = n - p - 1)$, то гипотеза H_0 отклоняется и одновременное исключение k факторов из модели некорректно, так как R_1^2 существенно превышает R_2^2.