МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ КОРРЕЛЯЦИЯ

ГЛАВА 5

Содержание

§ 22. Понятие множественной корреляции

Производственные взаимосвязи, как правило, определяются большим числом одновременно и совокупно действующих факторов. Например, овальность после чистового шлифования зависит от припуска на чистовое шлифование и от овальности после предварительного шлифования.

Себестоимость продукции зависит от стоимости материала, основной зарплаты рабочих, премиальных, расходов на содержание оборудования, отчислений на соцстрахование. В связи с этим возникает задача исследования зависимости между признаками $X_1, X_2, \ldots, X_n$, называемыми *факторными признаками* (а также регрессорами, предикторами) и признаком Y, называемым *результативным*. Эта задача решается методами так называемого *множественного корреляционного анализа*. Его характерной особенностью является то, что для измерения совместного влияния ряда показателей — факторов на величину анализируемого показателя строятся модели множественной корреляции.

§ 22. Понятие множественной корреляции

Построение многофакторной корреляционной модели начинается с установления формы связи, используя графический метод для пространства R^2 и метод перебора различных уравнений. От правильности выбора вида уравнения зависит, насколько построенная модель будет адекватна не только имеющимся экспериментальным данным, но и истинной зависимости между изучаемыми показателями. При прочих равных условиях предпочтение отдается модели, зависящей от меньшего числа параметров, так как для их оценки требуется меньшее количество эмпирических данных.

§ 22. Понятие множественной корреляции

После выбора формы многофакторной корреляционной модели проводят отбор факторных признаков и включение их в модель. Принято считать, что в уравнение множественной регрессии можно включать только независимые друг от друга факторные признаки X_j. Практически факторные признаки зависят либо слабо, либо сильно. Поэтому вопрос о включении факторных признаков в уравнение регрессии решают следующим образом.

Пусть, например, имеется три факторных признака X_1, X_2, X_3, влияющих на результативный признак Y и модель является линейной. Чтобы выяснить, какие факторные признаки включить в модель, находят коэффициенты парной корреляции r_{X1X2}, r_{X1X3}, r_{X2X3}.

§ 22. Понятие множественной корреляции

Если их значения меньше 0,8, то их можно включить в модель. Если же их значение больше 0,8, то следует какие-то из этих факторов исключить из модели.

Пусть, например, $r_{X1X3} = 0{,}85$. Ясно, что какой-то из признаков X_1 или X_3 надо исключить из модели. Для этого находят парные коэффициенты корреляции между каждым из факторов X_1 и X_3 и результативным признаком Y ,то есть вычисляют r_{YX1} и r_{YX3}. Затем сравнивают r_{YX1} и r_{YX3}. Пусть оказалось, что $r_{YX3} > r_{YX1}$. Это означает, что факторный признак X_3 сильнее связан с результативным признаком Y, чем признак X_1. Поэтому фактор X_3 следует включить в модель, а X_1 исключить из нее. Этот вывод подтверждаем путем вычисления коэффициентов частной корреляции $r_{YX3(X1)}$ и $r_{YX1(X3)}$.

§ 22. Понятие множественной корреляции

При исключении факторов из модели можно руководствоваться правилом. Если $|r_{ij}| > 1 - 3\sigma_{rij}$, где

$$\sigma_{r_{ij}} = \frac{1-r_{ij}^2}{\sqrt{n-1}},$$

то один из факторов, либо X_i , либо X_j следует исключить.

Рассмотрим случай построения многофакторной модели, когда результативный признак Y зависит от двух факторных признаков X_1 и X_2 . Если зависимость между ними носит линейный характер, то уравнение регрессии записывают в виде

$$\hat{Y}_{1,2} = a_0 + a_1 X_1 + a_2 X_2 .$$

§ 22. Понятие множественной корреляции

Коэффициенты уравнения регрессии a_0, a_1, a_2 находят по методу наименьших квадратов, решая систему нормальных уравнений:

$$\begin{cases} na_0 + [X_1]a_1 + [X_2]a_2 = [Y], \\ [X_1]a_0 + [X_1^2]a_1 + [X_1X_2]a_2 = [X_1Y], \\ [X_2]a_0 + [X_2X_1]a_1 + [X_2^2]a_2 = [X_2Y]. \end{cases}$$

Коэффициенты a_0, a_1, a_2 можно находить по формулам

$$a_1 = \frac{r_{X_1Y} - r_{X_2Y} \cdot r_{X_1X_2}}{1 - r_{X_1X_2}^2} \cdot \frac{S_Y}{S_{X_1}},$$

$$a_2 = \frac{r_{X_2Y} - r_{X_1Y} \cdot r_{X_1X_2}}{1 - r_{X_1X_2}^2} \cdot \frac{S_Y}{S_{X_2}},$$

$$a_0 = \overline{Y} - a_1\overline{X}_1 - a_2\overline{X}_2.$$

§ 22. Понятие множественной корреляции

Здесь r_{X1Y}, r_{X2Y}, r_{X1X2} — коэффициенты парной корреляции между признаками X_1 и Y, X_2 и Y, X_1 и X_2;

S_{X1}, S_{X2}, S_Y — средние квадратические отклонения;

$\bar{X}_1$, $\bar{X}_2$, Y — средние признаков X_1, X_2 , Y.

Если уравнение линейной регрессии имеет вид

$$\hat{Y}_{1,2,\ldots,k} = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_k X_k ,$$

то коэффициенты a_0, a_1, a_2, . . ., a_k находят, решая систему нормальных уравнений:

$$\begin{cases} na_0 + [X_1]a_1 + \ldots + [X_k]a_k = [Y], \\ [X_1]a_0 + [X_1^2]a_1 + \ldots + [X_1X_k]a_k = [X_1Y], \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ [X_k]a_0 + [X_kX_1]a_1 + \ldots + [X_k^2]a_k = [X_kY]. \end{cases}$$

§ 22. Понятие множественной корреляции

Для того чтобы проще было запомнить систему, запишем ее в матричной форме, предварительно введя следующие условные обозначения (только для этого конкретного случая):

$$[X_0]=n\,,\ [X_r]=\sum_{i=1}^{n}x_{ri}\,,\ [Y]=\sum_{i=1}^{n}y_i\,,\ [X_rY]=\sum_{i=1}^{n}x_{ri}y_i\,,\ r,s=1\ ..\ k,$$

$$[X_0][X_0]=[X_0^2]=n\,,\qquad [X_0][X_s]=[X_s],\qquad [X_r][X_s]=[X_rX_s],$$

$$[X_0][Y]=[Y],\qquad [X_r][Y]=[X_rY].$$

§ 22. Понятие множественной корреляции

Рассмотрим вектор

$$X = \begin{pmatrix} [X_0] \\ [X_1] \\ \vdots \\ [X_k] \end{pmatrix} = ([X_0]\ [X_1]\ \cdots\ [X_k])^{\mathsf{T}}$$

Тогда перепишем систему в сокращенной матричной форме:

$$XX^{\mathsf{T}}A = XY\,,$$

§ 22. Понятие множественной корреляции

или, в развернутой форме,

$$\begin{pmatrix}[X_0]\\[X_1]\\\vdots\\[X_k]\end{pmatrix}([X_0]\ [X_1]\ \cdots\ [X_k])\begin{pmatrix}a_0\\a_1\\\vdots\\a_k\end{pmatrix}=\begin{pmatrix}[X_0]\\[X_1]\\\vdots\\[X_k]\end{pmatrix}[Y],$$

или:

$$\begin{pmatrix}[X_0X_0] & [X_0X_1] & \cdots & [X_0X_k]\\[X_1X_0] & [X_1X_1] & \cdots & [X_1X_k]\\\cdots & \cdots & \cdots & \cdots\\[X_kX_0] & [X_kX_1] & \cdots & [X_kX_k]\end{pmatrix}\begin{pmatrix}a_0\\a_1\\\vdots\\a_k\end{pmatrix}=\begin{pmatrix}[X_0Y]\\[X_1Y]\\\vdots\\[X_kY]\end{pmatrix},$$

§ 22. Понятие множественной корреляции

или наконец

$$\begin{pmatrix} n & [X_1] & \cdots & [X_k] \\ [X_1] & [X_1^2] & \cdots & [X_1X_k] \\ \cdots & \cdots & \cdots & \cdots \\ [X_k] & [X_kX_1] & \cdots & [X_k^2] \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{pmatrix} = \begin{pmatrix} [Y] \\ [X_1Y] \\ \vdots \\ [X_kY] \end{pmatrix}$$

Для оценки надежности коэффициентов a_0, a_1, ..., a_k находят средние квадратические ошибки этих коэффициентов, то есть S_{ar}, r = 0..k.

§ 22. Понятие множественной корреляции

Пусть $C = XX^T$ — матрица коэффициентов при переменных a_r системы . Эта матрица является симметрической, то есть $C^T = C$, но, — существенно для дальнейшего, — она не является вырожденной: det $C = k + 1$. Тогда средние квадратические ошибки коэффициентов a_i вычисляют по формуле:

$$S_{a_r} = \hat{S}_\varepsilon \sqrt{c'_{rr}} ,$$

$$\hat{S}_\varepsilon = \sqrt{\sum (Y - \hat{Y}_X)^2} , \; c'_{rr}$$

— диагональный элемент матрицы C^{-1}, соответствующий факторному признаку X_r.

§ 22. Понятие множественной корреляции

П р и м е р. По опытным данным найдена матрица

$$(XX^{\mathrm{T}})^{-1} = \begin{pmatrix} 70{,}31 & -1{,}65 & 0{,}56 & 0{,}21 \\ -1{,}65 & 0{,}08 & -0{,}03 & -0{,}02 \\ 0{,}56 & -0{,}03 & 0{,}02 & 0{,}01 \\ 0{,}21 & -0{,}02 & 0{,}01 & 0{,}003 \end{pmatrix},$$

величина $\widehat{S}_{\varepsilon} = 3{,}08$ и уравнение регрессии

$$\hat{Y}_X = 59{,}5 - 0{,}81X_1 - 0{,}29X_2 + 0{,}13X_3.$$

Требуется вычислить средние квадратические ошибки коэффициентов уравнения регрессии.

§ 22. Понятие множественной корреляции

Для решения воспользуемся вышеуказанной формулой. Учитывая условие задачи, находим:

$$c'_{00} = 70{,}31, \quad c'_{11} = 0{,}08, \quad c'_{22} = 0{,}02, \quad c'_{33} = 0{,}003.$$

Тогда

$$S_{a_0} = 3{,}08\sqrt{70{,}31} = 25{,}826, \qquad S_{a_1} = 3{,}08\sqrt{0{,}08} = 0{,}871,$$

$$S_{a_2} = 3{,}08\sqrt{0{,}02} = 0{,}436, \qquad S_{a_3} = 3{,}08\sqrt{0{,}03} = 0{,}533.$$

§ 22. Понятие множественной корреляции

Записываем уравнение регрессии и под коэффициентами их полученные ошибки:

$$\hat{Y}_X = 59{,}5 \quad -0{,}81X_1 - 0{,}29X_2 + 0{,}13X_3,$$

$$(25{,}826),\ (0{,}871),\quad (0{,}436),\quad (0{,}533).$$

§ 23. Измерение тесноты множественной линейной корреляционной связи

За меру тесноты линейной связи между факторными и результативным признаками в совокупности принимают множественный или совокупный коэффициент корреляции R , который вычисляют по формуле:

$$R_{1.2.\cdots.k} = \sqrt{1 - \frac{\hat{S}^2_{1.2.\cdots.k}}{\hat{S}^2_Y}},$$

где

$$\hat{S}^2_{1.2.\cdots.k} = \frac{1}{n-k-1}\sum (Y_i - \hat{Y}_{1.2.\cdots.k})^2$$ — остаточная дисперсия;

$$\overline{S}^2_Y = \frac{1}{n-1}\sum (Y_i - \overline{Y})^2$$ — общая дисперсия результативного признака.

§ 23. Измерение тесноты множественной линейной корреляционной связи

Множественный коэффициент корреляции можно рассчитать, используя парные коэффициенты корреляции. Так, например, для линейной множественной корреляции между Y , X_1, X_2 коэффициент R вычисляют по формуле:

$$R_{1.2.3} = R_{Y.X_1.X_2} = \sqrt{\frac{r_{YX_1}^2 + r_{YX_2} - 2r_{X_1X_2} \cdot r_{YX_2} r_{YX_1}}{1 - r_{X_1X_2}^2}}.$$

§ 23. Измерение тесноты множественной линейной корреляционной связи

Множественный коэффициент корреляции можно получить на основе вычисления определителей, составленных из парных коэффициентов корреляции:

$$\Delta^* = \begin{vmatrix} 1 & r_{12} & r_{13} & \dots & r_{1n} \\ r_{21} & 1 & r_{23} & \dots & r_{2n} \\ r_{31} & r_{32} & 1 & \dots & r_{3n} \\ . & . & . & . & . \\ r_{n1} & r_{n2} & r_{n3} & \dots & 1 \end{vmatrix}, \Delta = \begin{vmatrix} 1 & r_{23} & r_{24} & \dots & r_{2n} \\ r_{32} & 1 & r_{34} & \dots & r_{3n} \\ r_{42} & r_{43} & 1 & \dots & r_{4n} \\ . & . & . & . & . \\ r_{n2} & r_{n3} & r_{n4} & \dots & 1 \end{vmatrix}, R = \sqrt{\frac{\Delta^*}{\Delta}}.$$

§ 23. Измерение тесноты множественной линейной корреляционной связи

Множественный коэффициент корреляции R обладает следующими свойствами:

1. $R \in [0, 1]$.
2. Если $R = 0$, то линейная корреляционная связь между признаками X_j и Y отсутствует, но другая зависимость (функциональная или нелинейная корреляционная) между ними может существовать.
3. Если $R = 1$, то между факторами X_j и Y существует функциональная линейная зависимость.

§ 23. Измерение тесноты множественной линейной корреляционной связи

Величину множественного коэффициента корреляции корректируют, так как при малом числе наблюдений значение R получается завышенным. Корректировку осуществляют по формуле:

$$\hat{R}_{1.2.\cdots.k} = \sqrt{1-(1-R^2)\frac{n-1}{n-k}},$$

где $\hat{R}$ — скорректированное значение R, n — число наблюдений, k — число факторных признаков. Корректировка R не производится при условии, если $\frac{n-k}{k} \geq 20$. Для коэффициента множественной корреляции определяют среднеквадратическую ошибку по формуле:

$$S_R = \frac{1}{\sqrt{n-1}}.$$

§ 23. Измерение тесноты множественной линейной корреляционной связи

Если выполняется неравенство $\frac{R}{S_R} > 3$, то с вероятностью $p = 0{,}99$ можно считать R значимым.

Наряду с определением показателя, отражающего тесноту связи результативного признака Y с факторными, вместе взятыми, определяют степень влияния каждого фактора в отдельности на изменение результативного фактора с помощью коэффициентов частной корреляции.

§ 23. Измерение тесноты множественной линейной корреляционной связи

Если уравнение множественной линейной регрессии между факторами X_1, X_2 и Y имеет вид $\hat{Y}_{1,2} = a_0 + a_1X_1 + a_2X_2$, то коэффициенты частной корреляции рассчитывают по формулам:

$$r_{YX_1(X_2)} = \frac{r_{YX_1} - r_{X_1X_2} \cdot r_{YX_2}}{\sqrt{(1-r^2_{X_1X_2})(1-r^2_{YX_2})}},$$

$$r_{YX_2(X_1)} = \frac{r_{YX_2} - r_{X_1X_2} \cdot r_{YX_1}}{\sqrt{(1-r^2_{X_1X_2})(1-r^2_{YX_1})}}.$$

§ 23. Измерение тесноты множественной линейной корреляционной связи

Если линейная регрессия имеет вид

$$\hat{Y}_{1.2.3} = a_0 + a_1 X_1 + a_2 X_2 + a_3 X_3$$

то частные коэффициенты корреляции находят по формуле:

$$r_{ij.kh} = \frac{r_{ij.k} - r_{ih.k} \cdot r_{jh.k}}{\sqrt{(1-r_{ih.k}^2)\cdot(1-r_{jh.k}^2)}},$$

где $i \neq j \neq k \neq h = 1,2,3,4$. Для общего случая.

$$r_{1.2\ldots p} = \frac{r_{12.34\ldots(p-1)} - r_{1p.34\ldots(p-1)} \cdot r_{2p.34\ldots(p-1)}}{\sqrt{(1-r_{1p.34\ldots(p-1)}^2)(1-r_{2p.34\ldots(p-1)}^2)}}.$$

§ 23. Измерение тесноты множественной линейной корреляционной связи

Если в корреляционную модель включено *k* факторных признаков, воздействующих на результативный признак *Y*, то коэффициент частной корреляции, например, для первого фактора, можно определить по формуле:

$$r_{Y1.2\ldots k} = \sqrt{1 - \frac{\hat{S}^2_{Y.1.2\ldots k}}{\hat{S}^2_{Y.2.3\ldots k}}},$$

где

$$\hat{S}^2_{Y.1.2\ldots k} = \frac{1}{n-1}\sum (y_i - \hat{y}_{1.2\ldots k})^2$$

— средний квадрат отклонений фактических значений признака Y от значений, вычисленных по формуле с учетом всех факторных признаков;

$$\hat{S}^2_{Y.2.3\ldots k} = \frac{1}{n-1}\sum (y_i - \hat{y}_{2.3\ldots k})^2$$

— средний квадрат отклонений фактических значений признака Y от значений, вычисленных по формуле, включающей все факторы кроме первого.

§ 23. Измерение тесноты множественной линейной корреляционной связи

Коэффициенты частной корреляции изменяются от 0 до 1 и обладают всеми свойствами парного коэффициента корреляции. Коэффициенту частной корреляции приписывается тот же знак, который имеет в уравнении множественной линейной регрессии коэффициент регрессии a_i при соответствующем факторном признаке X_i.

§ 24. Проверка адекватности модели множественной линейной корреляции

Адекватность модели означает не только количественное, но, прежде всего, качественное соответствие описания объекта. Для проверки соответствия полученного уравнения множественной линейной регрессии опытным данным используют коэффициент множественной регрессии R.

Если $R = 0$, то модель полностью не адекватна. Если $R = 1$, то модель в общем и в целом воспроизводит свойства моделируемого объекта.

Количественным показателем адекватности модели служит коэффициент детерминации R^2, который показывает долю дисперсии, объясняемой данной моделью в общей дисперсии.

§ 24. Проверка адекватности модели множественной линейной корреляции

Адекватность построенной модели может быть проверена по критерию Фишера — Снедекора. Для этого вычисляют статистику F_H по формуле:

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

где n — объем выборки; p — число факторных признаков, включенных в модель.

§ 24. Проверка адекватности модели множественной линейной корреляции

Затем находят при заданном уровне значимости α и числах степеней свободы $k_1 = p$, $k_2 = n - p - 1$ по таблице критических точек распределения Фишера F_T . Если $F_H > F_T$, то уравнение регрессии согласуется с опытными данными, если же $F_H < F_T$, то уравнение регрессии не согласуется с данными эксперимента. Адекватность модели множественной корреляции можно определять по средней ошибке аппроксимации

$$\varepsilon = \frac{1}{p}\left[\frac{\left|Y - \hat{Y}_{1,2,\cdots,p}\right|}{Y}\right] \cdot 100\%.$$

§ 25. Экономическая интерпретация уравнения регрессии

Заключительным этапом, завершающим построение корреляционной модели, является интерпретация полученного уравнения регрессии, т.е. перевод его с языка статистики и математики на язык экономики. Интерпретация начинается с выяснения, как каждый факторный признак, входящий в модель, влияет на величину результативного признака.

§ 25. Экономическая интерпретация уравнения регрессии

Чем больше величина коэффициента *ai* регрессии, тем сильнее фактор X_j влияет на результативный признак *Y*. Знаки коэффициентов регрессии a_j говорят о характере влияния на результативный признак.

Если коэффициент a_j имеет знак (+), то с увеличением данного фактора X_jрезультативный фактор *Y* возрастает. Если коэффициент a_j имеет знак (–), то с увеличением данного фактора X_j результативный признак уменьшается. Интерпретация знаков зависит от экономической сущности результативного признака.

§ 25. Экономическая интерпретация уравнения регрессии

Если величина результативного признака должна изменяться в сторону увеличения (объем реализованной продукции, фондоотдача, производительность труда и т.д.), то плюсовые знаки коэффициентов a_j свидетельствуют о положительном влиянии соответствующих факторов.

Если величина результативного признака изменяется в сторону снижения (себестоимость продукции, материалоемкость, простои оборудования и т.д.), то в этом случае положительное влияние на результативный признак будут оказывать факторы, коэффициенты которых отрицательны.

§ 25. Экономическая интерпретация уравнения регрессии

Рис. 12. Алгоритм построения многофакторной модели: объяснение — управление — предсказание.

§ 25. Экономическая интерпретация уравнения регрессии

Если экономический анализ подсказывает, что факторный признак должен влиять положительно, а коэффициент при нем имеет знак (–), то необходимо проверить расчеты. Так получается за счет допущенных ошибок при решении и в силу наличия взаимосвязей между факторными признаками, включенными в модель, влияющих в совокупности на результативный признак. При построении регрессионной модели можно рекомендовать алгоритм выполнения операций, представленный на рис. 12.

Спасибо за внимание!