Тема 5. ИСПОЛЬЗОВАНИЕ МНОГОФАКТОРНЫХ МОДЕЛЕЙ НА ОСНОВЕ ГЛАВНЫХ КОМПОНЕНТ ДЛЯ ПРОГНОЗИРОВАНИЯ СОСТОЯНИЯ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ СИСТЕМ

ПРОГНОЗИРОВАНИЕ НА ОСНОВЕ ГЛАВНЫХ КОМПОНЕНТ

Пространство состояний социально-экономической системы будет описываться в виде

$$\mathbf{x}^{o} = [x_1^{o} \quad x_2^{o} \quad \boxtimes \quad x_n^{o}] \quad (1)$$

Показатели социально-экономической системы, определяемые по такой модели, вычисляются по формуле

$$x_{ki}^{et} = \bar{x}_i + \sum_{h=1}^{p} v_{hi} z_{kh} \quad (2)$$

ПРОГНОЗИРОВАНИЕ НА ОСНОВЕ ГЛАВНЫХ КОМПОНЕНТ

Сценарное прогнозирование заключается в задание сценария в виде изменения показателей социально-экономической системы и вычисление по этим сценарным показателям значений главных компонент.

Для вычисления используется система уравнений вида

$$\begin{cases} x_{k1}^{sc} = \bar{x}_1 + \sum_{h=1}^{p} v_{hi} z_{kh} \\ x_{k2}^{sc} = \bar{x}_2 + \sum_{h=1}^{p} v_{hi} z_{kh} \\ \square \\ x_{kp}^{sc} = \bar{x}_p + \sum_{h=1}^{p} v_{hi} z_{kh} \end{cases} \qquad (3)$$

ПРОГНОЗИРОВАНИЕ НА ОСНОВЕ РЕГРЕССИОННЫХ МОДЕЛЕЙ

Построение регрессионной модели начинается с выдвижения гипотезы о том, что переменная зависит от набора эндогенных (независимых) переменных

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \square + \alpha_n x_n + \varepsilon \quad (1а)$$

или

$$y = \mathbf{X}\alpha + \varepsilon \quad (1б)$$

РЕГРЕССИОННАЯ МОДЕЛЬ

На практике вместо генеральной совокупности используется выборка данных,

$$\hat{y} = \mathbf{X}\alpha' + \varepsilon' \qquad (2)$$

α' - оценка коэффициентов регрессии

Минимизируется функционал (метод наименьших квадратов)

$$\Phi = (\mathbf{X}\ \alpha' - y)^T (\mathbf{X}\alpha' - y) \qquad (3)$$

В результате получаем

$$\mathbf{X}^T\ \mathbf{X}\alpha' - \mathbf{X}^T y = 0 \qquad (4)$$

РЕГРЕССИОННАЯ МОДЕЛЬ

С учетом формулы $\mathbf{A} = \frac{1}{m}\mathbf{X}^T\mathbf{X}$

уравнение (4) преобразуется к виду

$$m\mathbf{A}\alpha' = \mathbf{X}^T y \qquad (5)$$

Решая полученное уравнение, получается

$$\alpha' = \frac{1}{m}\mathbf{A}^{-1}\mathbf{X}^T y \qquad (6)$$

РЕГРЕССИОННАЯ МОДЕЛЬ

Матрица $\mathbf{A}$ может быть представлена

$$\mathbf{A} = \mathbf{V}_0 \Sigma \mathbf{V}_0^T \qquad (7)$$

Подставляя соотношение (7) в уравнение (5) получим

$$m\mathbf{V}_0 \Sigma \mathbf{V}_0^T \alpha' = \mathbf{X}^T y \qquad (8)$$

РЕГРЕССИОННАЯ МОДЕЛЬ

Далее умножаем справа на матрицу

$$m\mathbf{V}_0^T\mathbf{V}_0\Sigma\mathbf{V}_0^T\alpha' = \mathbf{V}_0^T\mathbf{X}^T y \qquad (9)$$

После небольших преобразований получается соотношение

$$\begin{aligned}\alpha' &= {}^{1}\!/_{m}\,\mathbf{V}_0\Sigma^{-1}\mathbf{V}_0^T\mathbf{X}^T y \\ &= {}^{1}\!/_{m}\sum_{i=1}^{n}\lambda_i^{-1}\mathbf{v}_{0i}\mathbf{v}_{0i}^T\mathbf{X}^T y\end{aligned} \qquad (10)$$

РЕГРЕССИОННАЯ МОДЕЛЬ

Одним из путей повышения качества регрессионной модели является удаление членов, соответствующих очень маленьким , которое приводит вычислению оценки

$$\alpha'' = 1/m \sum_{i=1}^{p} \lambda_i^{-1} \mathbf{v}_{0i} \mathbf{v}_{0i}^T \mathbf{X}^T y \qquad (11)$$

РЕГРЕССИЯ ГЛАВНЫХ КОМПОНЕНТ

Регрессионное уравнение, использующее в качестве независимых переменных главные факторы, имеет вид

$$y = \mathbf{Z}\beta + \varepsilon \qquad (12)$$

Для получения оценки вектора также используется метод наименьших квадратов, в результате чего получаем уравнение

$$\mathbf{Z}^T \mathbf{Z}\beta' - \mathbf{Z}^T y = 0 \qquad (13)$$

РЕГРЕССИЯ ГЛАВНЫХ КОМПОНЕНТ

Главные факторы и исходные факторы связаны соотношением

$$\mathbf{Z} = \mathbf{V}_0^T \mathbf{X} \qquad (14)$$

С учетом соотношения (14) уравнение (13) преобразуется к виду

$$\mathbf{V}_0 \mathbf{X}^T \mathbf{X} \mathbf{V}_0^T \beta' = (\mathbf{X}\mathbf{V}_0)^T y \qquad (15)$$

или

$$m\mathbf{V}_0 \mathbf{A} \mathbf{V}_0^T \beta' = (\mathbf{X}\mathbf{V}_0)^T y$$

РЕГРЕССИЯ ГЛАВНЫХ КОМПОНЕНТ

В соответствии с соотношением (7) уравнение (15) преобразуется

$$m\mathbf{V}_0\mathbf{V}_0^T\boldsymbol{\Sigma}\mathbf{V}_0^T\mathbf{V}_0\beta' = (\mathbf{X}\mathbf{V}_0)^T y \qquad (16)$$

С учетом ортогональности собственных векторов $\mathbf{V}_0^T\mathbf{V}_0 = \mathbf{I}$

Уравнение (16) преобразуется

$$m\boldsymbol{\Sigma}\beta' = (\mathbf{X}\mathbf{V}_0)^T y \qquad (17)$$

РЕГРЕССИЯ ГЛАВНЫХ КОМПОНЕНТ

В итоге получаем оценку вектора

$$\beta' = 1/m \, \Sigma^{-1} \mathbf{V}_0^T \mathbf{X}^T y \qquad (18)$$

Для уменьшения колебаний оценки коэффициентов вводится смещение в оценку коэффициентов α'

$$\alpha'' = \alpha' - 1/m \sum_{i=m+1}^{p} \lambda_i^{-1} \mathbf{v}_{0i} \mathbf{v}_{0i}^T \mathbf{X}^T y \qquad (10)$$

Выбор числа главных компонент

Общая изменчивость процесса изменения признаков определяется как

$$\sigma = \sum_{i}^{n} \sigma_i$$

Наиболее простая стратегия выбора числа главных компонент представляет простое удаление главных компонент, вариации которых меньше некоторого граничного значения

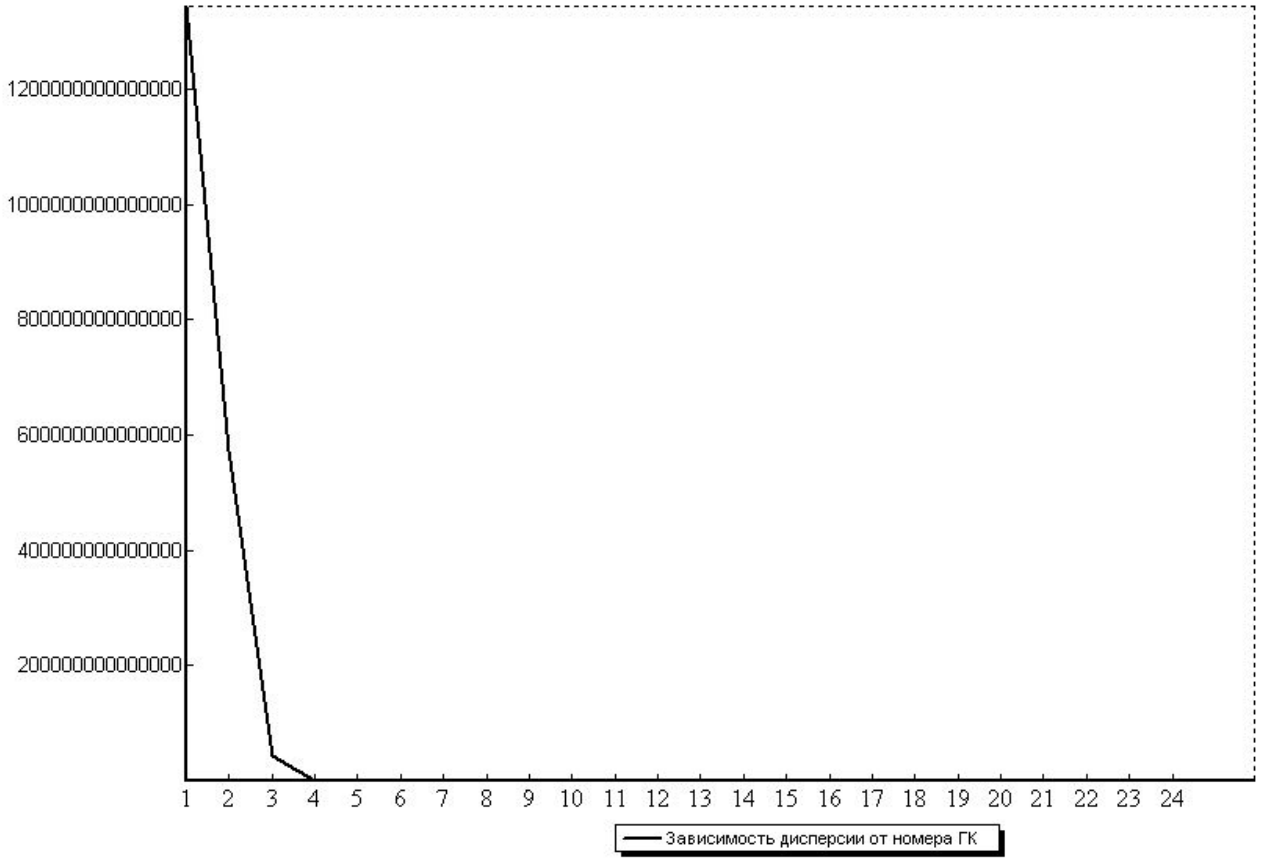
$$\lambda_i < \lambda_c$$

Критерий Кайзера. В соответствие с этим критерием отбираются только факторы, с собственными значениями, большими дисперсий отдельных факторов.

Выбор числа главных компонент

Критерий каменистой осыпи.

Критерий «каменистой осыпи» базируется на графическом представлении собственных значений.



Оценка качества регрессионной модели

Сумма квадратов, объясняемая регрессией (СКР) – это сумма возведённых в квадрат разностей между прогнозируемыми величинами зависимой переменной и средней величиной наблюдаемых значений зависимой переменной

$$СКР = \sum (\hat{y} - \bar{y})^2$$

Общая сумма квадратов отклонений (ОСК) – это сумма возведённых в квадрат разностей между наблюдаемой величиной зависимой переменной и средней наблюдаемых величин зависимо переменной

$$ОСК = \sum (y_i - \bar{y})^2$$

Оценка качества регрессионной модели

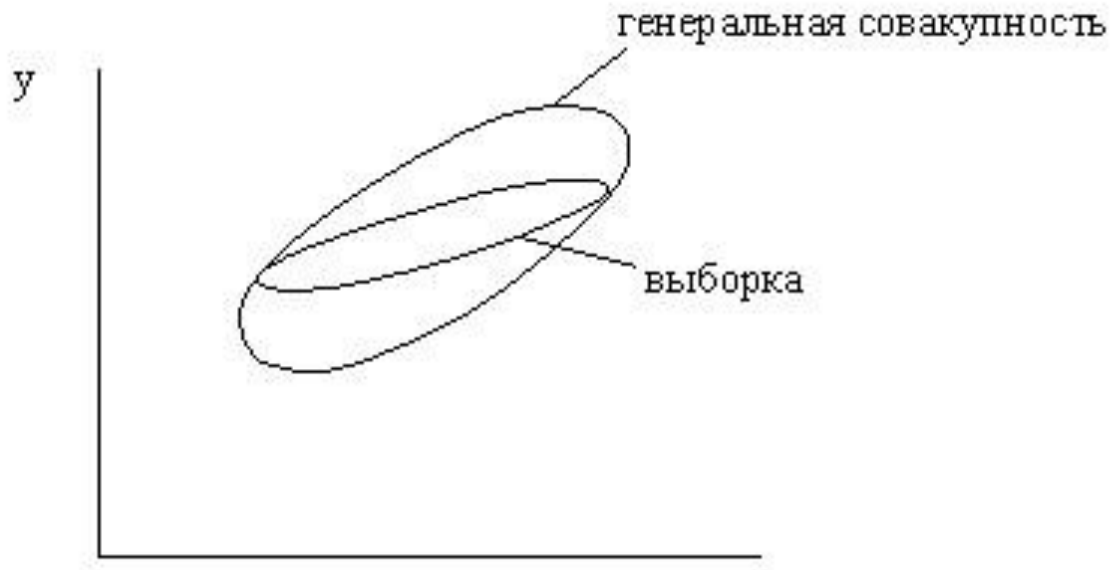
Результат деления СКР на ОСК называется коэффициентом детерминации

$$R^2 = \frac{СКР}{ОСК}$$

Например, если коэффициент детерминации равен 0.4, то регрессионная модель может объяснить 40% дисперсии критериального показателя, остальные же 60% определяются факторами, которые отсутствуют в модели.

ПРОВЕРКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ

Точки из генеральной совокупности попадают в выборку случайным образом, по этому в соответствии с теорией вероятности среди прочих случаев возможен вариант, когда выборка из “широкой” генеральной совокупности окажется “узкой”



ПРОВЕРКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ

В случае «узкой» выборки:

а) уравнение регрессии, построенное по выборке, может значительно отличаться от уравнения регрессии для генеральной совокупности, что приведет к ошибкам прогноза;

б) коэффициент детерминации и другие характеристики точности окажутся неоправданно высокими и будут вводить в заблуждение о прогнозных качествах уравнения.

ПРОВЕРКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ

Один из наиболее часто используемых вариантов проверки заключается в следующем. Для полученного уравнения регрессии определяется F -статистика – характеристика точности уравнения регрессии, представляющая собой отношение той части дисперсии зависимой переменной которая объяснена уравнением регрессии к необъясненной (остаточной) части дисперсии.

$$F = \frac{\left(\sum (\hat{y}_i - \bar{y})^2\right) / m)}{\left(\sum (y_i - \hat{y}_i)^2 / (n - m - 1)\right)}$$

ПРОВЕРКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ

Для осуществления статистической проверки значимости уравнения регрессии формулируется нулевая гипотеза об отсутствии связи между переменными (все коэффициенты при переменных равны нулю) и выбирается уровень значимости α

Уровень значимости – это допустимая вероятность совершить ошибку первого рода – отвергнуть в результате проверки верную нулевую гипотезу.

Чем выше уровень значимости (чем меньше α), тем выше уровень надежности теста, равный $1-\alpha$, т.е. тем больше шанс избежать ошибки признания по выборке наличия связи у генеральной совокупности на самом деле несвязанных между собой переменных.

ПРОВЕРКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ

Для выбранного уровня значимости по распределению Фишера определяется табличное значение $F_{m,\alpha}$.

$F_{m,\alpha}$ сравнивается с фактическим значением критерия для регрессионного уравнения F_{Φ}

Если выполняется условие

$$F_{\Phi} > F_{m,\alpha}$$

то ошибочное обнаружение связи будет происходить с вероятностью меньшей чем уровень значимости.

В соответствии с правилом “очень редких событий не бывает”, приходим к выводу, что установленная по выборке связь между переменными имеется и в генеральной совокупности.

ПРОВЕРКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ

Если же оказывается

$$F_{ф} < F_{m,\alpha}$$

то уравнение регрессии статистически не значимо.

Иными словами существует реальная вероятность того, что по выборке установлена не существующая в реальности связь между переменными.

ПРОВЕРКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ

После того как выполнена проверка статистической значимости регрессионного уравнения в целом полезно, особенно для многомерных зависимостей осуществить проверку на статистическую значимость полученных коэффициентов регрессии.

Полученные фактические значения критерия Стьюдента сравниваются с табличными значениями , полученными из распределения Стьюдента. Если оказывается, что

$$t_{\phi} > t_{m}$$

то соответствующий коэффициент статистически значим, в противном случае нет.

Критические точки распределения Стьюдента

	Уровень значимости (двусторонняя критическая область)					
	0.10	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
1	6.31	12.7	31.82	63.7	318.3	637.0
2	2.92	4.30	6.97	9.92	22.33	31.6
3	2.35	3.18	4.54	5.84	10.22	12.9
4	2.13	2.78	3.75	4.60	7.17	8.61
40	1.68	2.02	2.42	2.70	3.31	3.55
60	1.67	2.00	2.39	2.66	3.23	3.46
120	1.66	1.98	2.36	2.62	3.17	3.37
∞	1.64	1.96	2.33	2.58	3.09	3.29

Оценка ошибок прогноза

Ошибки прогноза

$$e_i = y_i - \hat{y}_i$$

Среднее абсолютное отклонение (Mean Absolute Derivation, **MAD**) измеряет точность прогноза, усредняя величины ошибок прогноза

$$e_{MAD} = \frac{1}{m} \sum_{i=1}^{m} \left| y_i - \hat{y}_i \right|$$

Оценка ошибок прогноза

Среднеквадратическая ошибка (Mean Squared Error, **MSE**)

$$e_{MSE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

Средняя абсолютная ошибка в процентах (Mean Absolute Percentage Error, **MAPE**)

$$\varepsilon_{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}$$

Оценка ошибок прогноза

Стандартная ошибка оценки

$$e_{SSE} = \sqrt{\frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{m - n - 1}}$$

Относительная среднеквадратическая ошибка

$$e_{MSEN} = \sqrt{\frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m} y_i^2}}$$

Оценка ошибок прогноза

Стандартная ошибка оценки

$$e_{SSE} = \sqrt{\frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{m - n - 1}}$$

Относительная среднеквадратическая ошибка

$$e_{MSEN} = \sqrt{\frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m} y_i^2}}$$