Лекция 3. Методы корреляционно-регрессионного анализа фондового рынка.

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

α - ошибка или значение помехи, также называемая остатком.

β - коэффициент регрессии, отражает наклон линии, вдоль которой рассеяны данные наблюдений. Он может быть истолкован как показатель, характеризующий процентное изменение переменной Y, которое вызвано изменением значения X на единицу.

ε - - ошибка или значение помехи, также называемая остатком.

Определение параметров уравнения регрессии с помощью метода наименьших квадратов

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2 \to \min$$

$$\beta = \frac{\operatorname{cov}(X, Y)}{\sigma_X^2} = \frac{\sum [(X_i - \overline{X})(Y_i - \overline{Y})]}{\sum (X_i - \overline{X})}$$

$$\alpha = \overline{Y} - \beta \overline{X}$$

При использовании МНК к ошибкам предъявляются следующие требования, называемые условиями Гаусса - Маркова:

1) величина ε_i является случайной переменной;
2) математическое ожидание ε_i равно нулю: $M(\varepsilon_i) = 0$;
3) дисперсия постоянна: $D(\varepsilon_i) = \sigma^2$ для всех i;
4) значения независимы между собой. Откуда вытекает, в частности, что

$$\operatorname{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0 \ldots при \ldots i \neq j \\ \sigma^2 \ldots при \ldots i = j \end{cases}$$

5) величины ε_i статистически независимы от значений x_i.

Критерии значимости коэффициентов и в уравнении регрессии.

$$t = \frac{\beta}{S_\beta}$$

Коэффициент детерминации R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \varepsilon_i^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}$$

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m}$$

При оценке значимости коэффициента линейной регрессии можно использовать следующее грубое правило. Если стандартная ошибка коэффициента больше его модуля, т.е. $t < 1$, то он не может быть признан хорошим (значимым). Если стандартная ошибка меньше модуля коэффициента, но больше его половины, т.е. $1 < t < 2$, то сделанная оценка может рассматриваться как более или менее значимая. Доверительная вероятность здесь примерно от 0,7 до 0,95. Значение t от 2 до 3 свидетельствуете весьма значимой связи (доверительная вероятность от 0,95 до 0,99), и $t > 3$ есть практически стопроцентное свидетельство ее наличия. Конечно, в каждом случае играет роль число наблюдений; чем их больше, тем надежнее при прочих равных условиях выводы о наличии связи и тем меньше верхняя граница доверительного интервала для данных числа

Коэффициент детерминации характеризует долю вариации (разброса) зависимой переменной, объясненной с помощью данного уравнения. В качестве меры разброса зависимой переменной обычно используется ее дисперсия, а остаточная вариация может быть измерена как дисперсия отклонений вокруг линии регрессии. Если числитель и знаменатель вычитаемой из единицы дроби разделить на число наблюдений n, то получим, соответственно, выборочные оценки остаточной дисперсии и дисперсии зависимой переменной Y. Отношение остаточной и общей дисперсий представляет собой долю необъясненной дисперсии. Если же эту долю вычесть из единицы, то получим долю дисперсии зависимой переменной, объясненной с помощью регрессии.

Иногда при расчете коэффициента детерминации для получения несмещенных оценок дисперсии в числителе и знаменателе вычитаемой из единицы дроби делается поправка на число степеней свободы; тогда

$$R^2 = 1 - \left[\frac{\sum_{i=1}^{n} \varepsilon_i^2}{n - m - 1} \right] / \left[\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 1} \right]$$

Для определения статистической значимости коэффициента детерминации проверяется нулевая гипотеза для F-статистики, рассчитываемой по формуле:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}$$

Величина F, если предположить, что выполнены предпосылки относительно отклонений , имеет распределение Фишера с (m; n-m-1) степенями свободы, где m - число объясняющих переменных, n - число наблюдений.

Гетероскедастичность.

Если остатки имеют постоянную дисперсию, они называются гомоскедастичными, но если они непостоянны, то гетероскедастичными.

Гетероскедастичность приводит к тому, что коэффициенты регрессии больше не представляют собой лучшие оценки или не являются оценками с минимальной дисперсией, следовательно, они больше не являются наиболее эффективными коэф фициентами.

Проверкой на гетероскедастичность служит тест Голдфелда-Кванта. Он требует, чтобы остатки были разделены на две группы из n наблюдений, одна группа с низкими, а другая - с высокими значениями. Обычно срединная одна шестая часть наблюдений удаляется после ранжирования в возрастающем порядке, чтобы улучшить разграничение между двумя группами.

Гетероскедастичность

Критерий Голдфелда-Кванта - это отношение суммы квадратов отклонений (СКО) высоких остатков к СКО низких остатков:

$$ГК = \frac{СКО_в}{СКО_н}$$

Этот критерий имеет F-распределение с (n-d)/2-k степенями свободы.

Чтобы решить проблему гетероскедастичности, нужно исследовать взаимосвязь между значениями ошибки и переменными и трансформировать регрессионную модель так, чтобы она отра жала эту взаимосвязь.

Автокорреляция.

$$\varepsilon_t = \beta \varepsilon_{t-1} + u_t$$

$$DW = \frac{\sum_{t=2}^{n} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n} \varepsilon_t^2}$$

Автокорреляция

Автокорреляция, также известная как сериальная корреляция, имеет место, когда остатки не являются независимыми друг от друга, потому что текущие значения Y находятся под влиянием прошлых значений. Зависимость между остатками описывается с помощью авторегрессионной зависимости. Эмпирическое правило гласит, что если критерий Дарбина-Уотсона равен двум, то не существует положительной автокорреляции, если он равен нулю, то имеет место совершенная по ложительная автокорреляция, а если он равен четырем, то имеет место совершенная отрицательная автокорреляция. *Если статистика DW находится в интервале* ***от 1.3 до 2.7*** *мы можем считать, что статистическая значимая автокорреляция остатков отсутствует.*

Мультиколлинеарность

Если некоторые или все независимые переменные в множественной регрессии являются высоко коррелированными, то регрессионной модели трудно разграничить их отдельные объясняющие воздействия на Y. В результате высококоррелированные независимые переменные действуют в одном направлении и имеют недостаточно независимое колебание, чтобы дать возможность модели изолировать влияние каждой переменной. Не существует точного граничного значения уровня корреляции переменных, при котором возникает проблема мультиколлинеарности. Это явление особенно часто имеет место при анализе фондовых переменных, таких, как доходность и объемы продаж, когда инфляция, например, может повлиять на оба временных ряда.

Для уменьшения мультиколлинеарности может быть принято несколько мер:

- Увеличивают объем выборки по принципу, что больше данных означает меньшие дисперсии оценок МНК. Проблема реализации этого варианта решения состоит в трудности на хождения дополнительных данных.
- Исключают те переменные, которые высококоррелированны с остальными. Проблема здесь заключается в том, что возможно переменные были включены на теоретической основе, и будет неправомочным их исключение только лишь для того, чтобы сделать статистические результаты "лучше".

Фиктивные переменные

Иногда необходимо включение в регрессионную модель одной или более качественных переменных, например, степени качества управления инвестиционным портфелем. Альтернативно может понадобиться сделать качественное различие между наблюдениями одних и тех же данных. Например, если проверяется взаимосвязь между размером компании и ежемесячными доходами по акциям, может быть желательным включение качественной переменной, представ ляющей месяц январь, по причине хорошо известного "январского эффекта" во временных рядах доходов по ценным бумагам.

Нелинейная регрессия.

$$\ln(Y) = \ln(\alpha) + \beta \ln(x)$$

Интервал прогнозирования:

$$\hat{Y} \pm t_{99} \cdot s \sqrt{1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{\sum (X_i - \overline{X})^2}}$$

Выявление наличия корреляционной связи между парой показателей и оценка ее тесноты.

линейный (парный) коэффициент корреляции:

$$r_{yx} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Для качественной оценки коэффициента корреляции применяются различные шкалы, наиболее часто - шкала Чеддока. В зависимости от значения коэффициента корреляции связь может иметь одну из оценок:

0.1 - 0.3 - слабая;

0.3 - 0.5 - заметная;

0.5 - 0.7 - умеренная;

0.7 - 0.9 - высокая;

0.9 - 1.0 - весьма высокая.

Оценка значимости коэффициента корреляции при малых объемах выборки выполняется с использованием t-критерия Стьюдента. При этом фактическое (наблюдаемое) значение этого критерия определяется по формуле

$$t_{набл} = \sqrt{\frac{r_{y,x}^2}{1 - r_{y,x}^2}(n-2)}$$

Линейный коэффициент корреляции изменяется в пределах от -1 до +1. Его положительные значения свидетельствуют о прямой связи между переменными, отрицательные - об обратной. Близость коэффициента корреляции к нулю свидетельствует о слабой связи между переменными и о нецелесообразности ее моделирования. Следует отметить, что величина коэффициента корреляции не является доказательством того, что между исследуемыми признаками существует причинно-следственная связь, а представляет собой оценку степени взаимной согласованности в изменениях признаков. Для того чтобы установить причинно-следственную зависимость, необходим анализ качественной природы явлений.

Вычисленное по этой формуле значение $t_{набл}$ сравнивается с критическим значением t-критерия, которое берется из таблицы значений t-критерия Стьюдента с учетом заданного уровня значимости и числа степеней свободы (n - 2).

Если $t_{набл} > t_{таб}$, то полученное значение коэффициента корреляции признается значимым (т.е. **нулевая гипотеза,** утверждающая равенство нулю коэффициента корреляции, **отвергается).** И таким образом делается вывод, что между исследуемыми переменными есть тесная статистическая взаимосвязь.

Матрица коэффициентов парной корреляции

$$R = \begin{pmatrix} 1 & r_{y,x_1} & r_{y,x_2} & \cdots & r_{y,x_m} \\ r_{y,x_1} & 1 & r_{x_1x_2} & \cdots & r_{x_1x_m} \\ r_{y,x_2} & r_{x_1x_2} & 1 & \cdots & r_{x_2x_m} \\ r_{y,x_m} & r_{x_1x_m} & r_{x_2x_m} & \cdots & 1 \end{pmatrix}$$

Анализ матрицы коэффициентов парной корреляции используют при построении моделей множественной регрессии. Одной корреляционной матрицей нельзя полностью описать зависимости между величинами. В связи с этим **в многомерном корреляционном анализе рассматривается две задачи:**

1. Определение тесноты связи одной случайной величины с совокупностью остальных величин, включенных в анализ.

2. Определение тесноты связи между двумя величинами при фиксировании или исключении влияния остальных величин.

Эти задачи решаются соответственно с помощью коэффициентов множественной и частной корреляции.

Множественный коэффициент корреляции

$$R_{j,1,2,\ldots,j-1,j+1,\ldots,m} = \sqrt{1 - \frac{|R|}{R_{jj}}}$$

Решение первой задачи (определение тесноты связи одной случайной величины с совокупностью остальных величин, включенных в анализ) осуществляется с помощью ***выборочного коэффициента множественной корреляции*** по формуле, где |R| - определитель корреляционной матрицы R;
- алгебраическое дополнение

элемента R_{ij} той же матрицы R.

Коэффициенты множественной корреляции и детерминации являются величинами положительными, принимающими значения в интервале от 0 до 1. При приближении коэффициента R^2 к единице можно сделать вывод о тесноте взаимосвязи случайных величин, но не о ее направлении.

Частный коэффициент корреляции

Если рассматриваемые случайные величины коррелируют друг с другом, то на величине коэффициента парной корреляции частично сказывается влияние других величин. В связи с этим возникает необходимость исследования частной корреляции между величинами при исключении влияния других случайных вели чин (одной или нескольких).

Частный коэффициент корреляции определяется по формуле:

Частный коэффициент корреляции

$$r_{jk(1,2,\ldots,m)} = -\frac{R_{jk}}{\sqrt{R_{jj}R_{kk}}}$$

$$r_{1,2,(3)} = \frac{r_{1,2} - r_{1,3}r_{2,3}}{\sqrt{(1-r_{1,3}^2)(1-r_{2,3}^2)}}$$

Частный коэффициент корреляции, так же как и парный коэффициент корреляции, изменяется от -1 до +1.

Выражение при условии m = 3 будет иметь вид

$$r_{1,2,(3)} = \frac{r_{1,2} - r_{1,3} r_{2,3}}{\sqrt{(1 - r_{1,3}^2)(1 - r_{2,3}^2)}}$$

Коэффициент $r_{1,2,(3)}$ называется коэффициентом корреляции между x1, и x2 при фиксированном x_3. Он симметричен относительно первичных индексов 1, 2. Его вторичный индекс 3 относится к фиксированной переменной.