Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и приближений

Анализ модели ценообразования на рынке недвижимости

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студента 4 курса 412 группы
направления 01.03.02 Прикладная математика и информатика

Механико-математический факультет

Зайцева Николая Николаевича

Научный руководитель
доцент, к.эк.н. ____________________ А.В. Харламов

Заведующий кафедрой
доцент, д.ф.-м.н. ____________________ С.П. Сидоров

Саратов 2016

Введение.

Одной из экономических систем, которая на сегодняшний день занимает важное место в жизни человека, является рынок жилой недвижимости. В последние несколько лет данный рынок получил бурное развитие, что связанно с развитием ипотечного кредитования, поиска объектов инвестирования и другими факторами. В рамках тенденций последних лет рост цен на недвижимость сделал ее очень привлекательным объектом инвестиции с целью получения коммерческой выгоды. Квартиры стали приобретаться не только для проживания, но и для вложения в них денежных средств с целью получения дохода.

Таким образом возникла необходимость в изучении и исследовании данного рынка. Существующие методы исследования являются либо упрощенными и недостаточно точными, либо трудоёмкими и сложными (экспертные в рамках индивидуальной оценки недвижимости). Таким образом, актуальным и востребованным становиться проведение исследований рынка недвижимости с помощью регрессионного анализа, суть которого заключается в определении основных тенденций ценообразования на рынке. В основе данного метода лежат факторные регрессионные модели, позволяющие оценить влияние на исследуемый объект различных ценообразующих факторов.

Данная методика имеет такие положительные моменты как:

- простота и быстрота расчета;
- способность выявлять и количественно описывать тенденции, которые невозможно оценить на глаз или экспертно на малом количестве наблюдений;
- отсутствие субъективизма со стороны конкретного эксперта ранка жилья.

Полученные результаты можно использовать как в общих аналитических целях, так и в прикладном направлении – для оценки жилой недвижимости.

При проведении теоретических и практических исследований на основе регрессионного анализа зачастую возникают определенные сложности. Большинство их них выражается в недостатках, вызванных как правило ошибками при формулировке модели.

Применение регрессионного анализа для исследования и оценки рынка жилья является актуальным, ввиду особенности изменения порядка определения налогооблагаемой базы при расчете налога на имущество.

Ценообразующим фактором является многообразие условий в которых формируется структура и уровень цены. Целью работы является построение и анализ модели ценообразования на рынках жилья.

К задачам данной работы можно отнести:

- сбор и первичный анализ данных;
- выбор и построение модели;
- выдвижение гипотез и проведение тестов для их проверки;
- анализ полученных данных.

Сама работа содержит в себе две части. В первой описывается классическая нормальная линейная модель множественной регрессии, метод наименьших квадратов, гипотезы и тесты, относящиеся к проверке этих гипотез, а также алгоритмы применяемых тестов. Во второй части эта теория применяется на конкретной задаче рынка недвижимости.

Основное содержание работы.

Рассмотрим модель множественной регрессии. Обозначим *i*-*е* наблюдение зависимой переменной y_i , а объясняющие переменные – $x_{i1}, x_{i2}, \ldots, x_{ip}$. Тогда модель можно представить в виде:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \ ,$$

где i=1,2,…,n; ε_i удовлетворяет предпосылкам регрессионного анализа:

1. В модели возмущение ε_i (или зависимая переменная y_i) есть величина случайная, а объясняющая переменная x_{ij} (где j=1,2,…,p) величина не случайная.

2. Математическое ожидание возмущения ε_i равно нулю:

$$M(\varepsilon_i) = 0 \, .$$

3. Дисперсия возмущения ε_i постоянна для любого i:

$$D(\varepsilon_i) = \sigma^2 \, .$$

Модель в которой зависимая переменная y_t, возмущения ε_i, и объясняющие переменные x_{i1}, x_{i2}, … , x_{ip} удовлетворяют приведенным предпосылкам регрессионного анализа и, кроме того, предпосылке о не вырожденности матрицы (независимости столбцов) значений объясняющих переменных, называется *классической нормальной линейной моделью множественной регрессии*.

Введем обозначения: Y - матрица-столбец, или вектор, значений зависимой переменной размера n; X - матрица значений объясняющих переменных, или матрица плана размера n×(p+1); β – матрица-столбец, или вектор, параметров размера (p+1); ε - матрица-столбец, или вектор, возмущений размера n.

Тогда в матричной форме модель примет вид:

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

Оценкой этой модели по выборке является уравнение

$$\hat{Y} = Xb + e \ .$$

Для оценки вектора неизвестных параметров β применяется метод наименьших квадратов. Так как произведение транспонированной матрицы на саму матрицу есть:

$$e'e = \sum_{i=1}^{n} e_i^2 \ ,$$

условие минимизации остаточной суммы квадратов запишется в виде:

$$S = \sum_{i=1}^{n} (\hat{y}_{x_i} - y_i)^2 = e'e = \sum_{i=1}^{n} e_i^2 = (Y - Xb)'(Y - Xb) \to min \, .$$

На основании необходимого условия экстремума функции нескольких переменных $S(b_1, b_2, \ldots, b_n)$, представляющей необходимо приравнять нулю частные производные по этим переменным или в матричной форме – вектор частных производных

$$\frac{\partial S}{\partial b} = \left(\frac{\partial S}{\partial b_0} \frac{\partial S}{\partial b_1} \ldots \frac{\partial S}{\partial b_p} \right).$$

Для вектора частных производных получается система нормальных уравнений в матричной форме для определения вектора b:

$X'Xb = X'Y$

Для решения матричного уравнения относительно вектора оценок параметров b необходимо ввести предпосылку:

Векторы значений объясняющих переменных, или столбцы матрицы плана X, должны быть линейно независимыми, т.е. ранг матрицы X – максимальный ($r(X)=p+1$).

Теорема Гаусса-Маркова:

1. $Y_t = a + bX_t + \varepsilon_t, \quad t = 1,\ldots,n;$
2. $X_t - \textit{детерминированная} \quad \textit{величина};$
3. $E\varepsilon_t = 0, \quad E\left(\varepsilon_t^2\right) = V\left(\varepsilon_t\right) = \sigma^2;$
4. $E\left(\varepsilon_t \varepsilon_s\right) = 0, \quad \textit{при} \quad t \neq s.$

Оценки, полученные по методу наименьших квадратов, имеют наименьшую дисперсию в классе всех линейных несмещенных оценок.

Рассмотрим оценку значимости коэффициентов регрессии b_j и построение доверительного интервала для параметров регрессионной модели β_j (j=1,2,…,p).

Значимость коэффициента регрессии b_j можно проверить, если учесть, что статистика $(b_j - \beta_j)/s_{b_j}$ имеет t-распределение Стьюдента с k=n-p-1 степенями свободы. Поэтому b_j значимо отличается от нуля на уровне значимости α, если $|t| = \frac{|b_j|}{s_{b_j}} > t_{1-\alpha;n-p-1}$, где $t_{1-\alpha;n-p-1}$ – табличное значение t-критерия Стьюдента, определенное на уровне значимости α при числе степеней свободы k=n-p-1.

В общей постановке гипотеза о равенстве параметра β_j заданному числу β_{j0}, отвергается если

$$|t| = \frac{|b_j - \beta_{j0}|}{s_{b_j}} > t_{1-\alpha;n-p-1} .$$

Поэтому доверительный интервал для параметра β_j есть

$$b_j - t_{1-\alpha;n-p-1} s_{b_j} \le \beta_j \le b_j + t_{1-\alpha;n-p-1} s_{b_j} .$$

В модели множественной регрессии общая вариация Q – сумма квадратов отклонений зависимой переменной от средней может быть разложена на две составляющие:

$$Q = Q_R + Q_e .$$

Уравнение множественной регрессии значимо, если

$$F = \frac{Q_R(n-p-1)}{Q_e p} > F_{\alpha;p;n-p-1} ,$$

где $F_{\alpha;p;n-p-1}$- табличное значение F-критерия Фишера-Снедекора.

Коэффициент детерминации R^2 одна из наиболее эффективных оценок регрессионной модели, мера качества уравнения регрессии. Определяется по формулам:

$$R^2 = \frac{Q_R}{Q} = \frac{b'X'Y - n\bar{y}^2}{Y'Y - n\bar{y}^2},$$

$$R^2 = 1 - \frac{Q_e}{Q}.$$

R² *характеризует* долю вариации зависимой переменной, обусловленной регрессией или изменчивостью объясняющих переменных.

Перейдем к поиску доверительного интервала и проверке гипотез.

Проверим гипотезу $H_0 : \beta_i = \beta_{i0}$, с статистическими результатами: вектор оценок $\hat{\beta}_{OLS}$ имеет нормальное распределение со средним β и матрицей ковариаций $V\left(\beta_{OLS,i}\right) = \sigma^2 \left(X'X\right)^{-1}$, т.е. $\beta_{OLS,i} - \beta_i \sim N\left(0, \sigma^2_{\beta_i}\right)$, где $\sigma^2_{\beta_i} = \sigma^2 q^{ii}; q^{ii}$ - *i-й* диагональный элемент матрицы. В качестве оценки дисперсии берется $s^2_{\beta_i} = \sigma_{\beta_i}{}^2 = \sigma^2 q^{ii} = s^2 q^{ii}$. Случайная величина $\left(n-k\right)\frac{s^2}{\sigma^2}$ распределена по закону $\chi^2\left(n-k\right)$. Оценки β_{OLS} и s^2 независимы. Получаем статистику

$$t = \frac{\beta_{OLS,i} - \beta_i}{s_{\beta_i}} = \frac{\left(\beta_{OLS,i} - \beta_i\right) / \sigma_{\beta_i}}{s_{\beta_i} / \sigma_{\beta_i}} \sim t\left(n-k\right),$$

из которой интервал

$$\left[\beta_{OLS,i} - t_c \, s_{\beta_i} ; \beta_{OLS,i} + t_c \, s_{\beta_i}\right]$$

является $100(1-\alpha)\%-$*ным доверительным интервалом* для истинного значения коэффициента β_i , где $t_c = t_{\alpha/2}(n-k) - 100(\alpha/2)\%-$*ная* точка распределения Стьюдента.

Пусть $H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0$. Предположим, что в число регрессоров включена константа: $y_t = \beta_1 + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + \varepsilon_t$. Нулевая гипотеза состоит в том, что коэффициенты при всех регрессорах равны нулю.

Рассмотрим статистику

$$F=\frac{R^2}{1-R^2}\frac{n-k}{k-1}=\frac{RSS}{ESS}\frac{n-k}{k-1}=\frac{\sum\left(\hat{y}_t-\bar{y}\right)^2/(k-1)}{\sum e_t^2/(n-k)}=\frac{\frac{\hat{y}_*'\hat{y}_*}{\sigma^2}\frac{1}{n-k}}{\frac{e'e}{\sigma^2}\frac{1}{n-k}}.$$

Из Теоремы Гаусса-Маркова и оценок дисперсии возмущений, $\hat{\beta}_{OLS}$ и e независимы, поэтому статистика F имеет распределение Фишера

$$F=\frac{R^2}{1-R^2}\frac{n-k}{k-1}=\frac{RSS}{ESS}\frac{n-k}{k-1}=\frac{\left(\hat{y}_t-\bar{y}\right)^2/(k-1)}{e_t^2/(n-k)}=\frac{\hat{y}_*'\hat{y}_*\Big/(\mathrm{k}-1)}{e'e\Big/(n-k)}\sim F(k-1,n-k),$$

и её можно использовать для проверки гипотезы $H_0:\beta_2=\beta_3=...=\beta_k=0$. Гипотеза отвергается на 5%-ном уровне значимости, если $F>F_c$, где $F_c=F_{0.05}(k-1,n-k)$ - 5%-ная точка распределения Фишера.

Выбор регрессоров называется *спецификацией модели*. Поэтому возникает естественный вопрос соотношения между МНК-оценками параметров в истинной и выбранной моделях. Рассмотрим две ситуации: в оцениваемой модели отсутствует часть независимых переменных, имеющихся в истинной модели (исключение существенных переменных); в оцениваемой модели присутствуют независимые переменные, которых нет в истинной модели (включение несущественных переменных). Рассмотрим два основных случая.

Случай 1. *Исключены существенные переменные.*

Процесс, порождающий данные: $y=X\beta+Z\gamma+\varepsilon$.

Модель: $y=X\beta+\varepsilon$.

Случай 2. *Включены несущественные переменные.*

Процесс, порождающий данные: $y=X\beta+\varepsilon$.

Модель: $y=X\beta+Z\gamma+\varepsilon$.

где $y-n\times1$ вектор наблюдений зависимой переменной; $X-n\times k$ и $Z-n\times l$ матрицы; $\beta-k\times1$ и $\gamma-l\times1$ векторы коэффициентов.

Рассмотрим постановку задачи, когда нам неизвестен процесс, порождающий данные. Таким образом, мы сравниваем две модели:

I. $y = X\beta + Z\gamma + \varepsilon$ (модель без ограничений),

II. $y = X\beta + \varepsilon$ (модель с ограничениями),

будем использовать индексы *u* и *r* для моделей без ограничений и с ограничениями соответственно.

Для выбора одной из этих моделей существует два способа.

Способ 1

$$R^2 = 1 - \frac{e'e}{y'Ay}; \quad A = I_n - \frac{1}{n}ii'; \quad R^2_{adj} = 1 - \frac{e'e/(n-k)}{y'Ay/(n-1)}.$$

Для модели с ограничением и без ограничения получаем соответственно

$$R^2_{u,adj} = 1 - \frac{e'_u e_u/(n-k-1)}{y'Ay/(n-1)}; \quad R^2_{r,adj} = 1 - \frac{e'_r e_r/(n-k)}{y'Ay/(n-1)};$$

$$R^2_{r,adj} - R^2_{u,adj} = \frac{e'_u e_u/(n-k-1) - e'_r e_r/(n-k)}{y'Ay/(n-1)}.$$

Гипотезу $H_0 : \gamma = 0$ проверяем с помощью *F*-статистики (или *t*-статистики):

$$F = \frac{(e'_r e_r - e'_u e_u)/1}{e'_u e_u/(n-k-1)} \sim F(1, n-k-1) \sim t^2(n-k-1).$$

Следовательно,

$$R^2_{r,adj} - R^2_{u,adj} = \frac{e'_u e_u/(n-k-1) - e'_r e_r/(n-k)}{y'Ay/(n-1)} \cdot \frac{1-t^2}{n-k}.$$

Таким образом, если $|t| > 1$, то $R^2_{r,adj} < R^2_{u,adj}$, и наоборот. Если верить, что скорректированный коэффициент детерминации R^2_{adj} является правильным критерием, то мы должны выбрать модель без ограничения тогда, когда $|t| > 1$.

Способ 3 (основан на наименьшем среднеквадратичном отклонении MSE).

Сравним модели I и II по критерию $MSE\left(\hat{\beta}\right)' = E\left(\left(\hat{\beta}-\beta\right)'\left(\hat{\beta}-\beta\right)\right)$.

Обозначим $M = I - X\left(X'X\right)^{-1}X'$ и также введем обозначения

$$q = \frac{\sigma}{\sqrt{z'Mz}}\left(X'X\right)^{-1}X'z; \quad \theta = \frac{\gamma}{\sigma/\sqrt{z'Mz}}.$$

Предположим, что вектор ошибок имеет стандартное многомерное нормальное распределение, тогда следует, что

$$\hat{\beta}_r \sim \mathrm{N}\left(\beta+\theta q, \sigma^2\left(X'X\right)^{-1}\right); \quad \hat{\beta}_u \sim \mathrm{N}\left(\beta, \sigma^2\left(X'X\right)^{-1}+qq'\right).$$

Отсюда по свойствам нормального распределения получаем

$$MSE\left(\hat{\beta}_r\right) - MSE\left(\hat{\beta}_u\right) = \left(\theta^2-1\right)q'q.$$

Условие $|\theta|>1$ важно, но тут θ - «теоретическое *t*-отношение», а не полученное из наблюдений.

Необходимо знать «что лучше $\hat{\beta}_r$ или $\hat{\beta}_u$» и правильно будет задать вопрос «верна ли гипотеза $H_0 : |\theta| > 1$?», так как это условие различает, какое из чисел больше: $MSE\left(\hat{\beta}_r\right)$ или $MSE\left(\hat{\beta}_u\right)$.

Применение обобщенного метода наименьших квадратов при наличии гетероскедастичности сводится к минимизации суммы взвешенных квадратов отклонений. использование доступного обобщенного метода наименьших квадратов в общем случае требует оценивания n параметров по n наблюдениям, что не позволяет получать состоятельные оценки. В некоторых ситуациях (ошибка пропорциональна одной из независимых переменных, дисперсии ошибок принимают два значения) можно применять доступный обобщенный метод наименьших квадратов и получать состоятельные оценки коэффициентов регрессии.

Заключение.

В данной работе рассматривается такой социально-экономический вопрос как ценообразующие факторы рынка недвижимости. Данный рынок имеет постоянный интерес со стороны общества. Вопрос сравнительных преимуществ жилья приходилось, приходится и придется решать многим, тем кто меняет свои жилищные условия и тем, кто инвестирует свои денежные средства.

Интерес анализа данного рода задач объясним тем, что к ним можно подойти с математической точки зрения, применяя как в данной работе регрессионный анализ и соответствующие тесты. Такой подход в данном вопросе помогает систематизировать информацию и упростить анализ и решение такого рода задач, путем автоматизации по средствам вычислительной техники и различного рода программного обеспечения.

Применяемый метод наименьших квадратов показывает свою адекватность при должном подборе оцениваемой модели. В ходе работы выяснилось, что логарифмическая форма модели и даваемые ей оценки наиболее эффективны. В случае использования линейной формы модели возникает необходимость борьбы с гетероскедастичностью в выборке, что является не тривиальной задачей.

Проделанные тесты на гетероскедастичность отвечают на вопрос о неоднородности наблюдений. По этому параметру логарифмическая форма модели наиболее приемлема для использования в исследованиях.

Про тест Чоу можно сказать то, что данная процедура проверяет стабильность параметров регрессионной модели. По факту тест проверяет неоднородность выборки в контексте регрессионной модели. Проведенный в работе тест указал на возможность сравнения двух рынков недвижимости одной общей моделью. Но стоит заметить, что такое возможно для теоретических исследований и может не работать на практике.

Анализируя модели и проделанные тесты ценообразующими факторами на рынке недвижимости для 1-комнатных квартир, можно указать такие параметры как: площадь квартиры, площадь кухни и жилой комнаты для вторичного рынка недвижимости, основной материал из которого построен дом, расстояние до центра города. Стоит отметить, что такие параметры являются достаточно общими и могут добавляться различными параметрами при дополнительном и более детальном рассмотрении рынка.