Прогнозирование рыночной стоимости квартир на основе множественной модели регрессии

Ахмедзянова Татьяна Камильевна
Троицкий филиал Челябинского государственного университета

Собственное жильё занимает чрезвычайно важное место в нашей жизни, как в глобальном, так и повседневном уровне. Всем хочется иметь собственный уголок, который можно обставить по своему желанию, в котором можно укрыться от проблем внешнего мира и куда всегда будет приятно возвращаться. Поколения людей сменяют друг друга, но людям всегда нужно место для жительства, кроме того в настоящие время многие покупают недвижимость с целью вложения средств и получения дохода. Поэтому недвижимость всегда продавалась, продаётся и будет продаваться.

При покупке квартиры возникает вопрос, какую сумму правильно потратить, чтобы не оказаться в минусе. Для этого необходимо сравнение таких условий как: площадь, этаж, район, близости и отдаленность от остановок, магазинов и т.д. Спрогнозировать рыночную стоимость квартиры можно на основе модели множественной регрессии.

Множественная регрессия широко используется в решении проблем спроса, доходности акций, при изучении функции издержек производства, в макроэкономических расчетах и целом ряде других вопросов эконометрики. В настоящее время множественная регрессия – один из наиболее распространенных методов в эконометрике. Основная цель множественной регрессии – построить модель с большим числом факторов, определив при этом влияние каждого из них в отдельности, а также совокупное их воздействие на моделируемый показатель.

Целью работы является построение модели множественной регрессии зависимости стоимости квартир города Троицка от их характеристик в среде Eviews 5.0. Для достижения поставленной цели необходимо решить следующие задачи:

1. Подготовить и проанализировать данные эконометрического исследования;
2. Выявить существенные факторы влияющие на результативный признак.
3. Построить уравнения множественной регрессии и выбрать из них то, которое наилучшим образом отображает существующую зависимость между факторами и результативным признаком;
4. Проверить модель на мультиколлинеарность;
5. Определить адекватность и значимость модели;
6. По полученной модели построить прогноз.

В качестве исходных данных возьмём выборку из ста объявлений о продаже квартир в городе Троицке выложенных в свободном доступе на сайте domofond.ru.

В данном случае, для построения модели необходимо, чтобы в качестве объясняющих переменных выступали как численные величины, так и качественные переменные. Обычно в роли таких переменных выступают дихотомические или, фиктивные, которые могут принимать только два значения, 0 и 1 (есть признак или его нет). При этом необходимо иметь в виду, что количество таких переменных должно быть на единицу меньше, чем число уровней изучаемого признака. В качестве зависимой переменной У выступает цена квартиры, X1- размер жилплощади выраженный в квадратных метрах, X2- количество комнат в квартире, X3- этаж на котором продаётся квартира, X4 – качественная переменная, показывающая наличие или отсутствие ремонта в квартире, X5- расстояние от квартиры до центральной площади города, выражено в км., X6 и X7 качественные переменные отвечающие за материал из которого построен дом (кирпич; панельный или блочный), X9 и X10 качественный переменные, показывающие «возраст» дома (новый; «брежневка»)

Построив линейную модель множественной регрессии по исходным данным в программе Eviews 5, мы получили следующее уравнение:

Y=-578774,2+33383,69*X1-98788,88*X2-18076,54*X3+149135*X4-192,6506*X5+313420,1*X6+151482,2*X7+378857,1*X9-3555485,4*X10

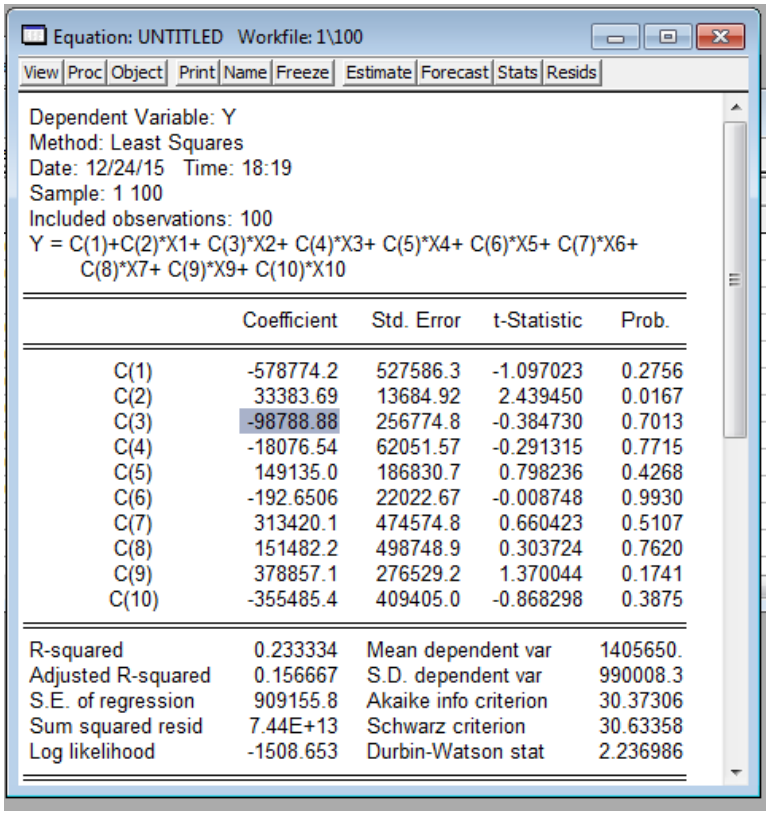
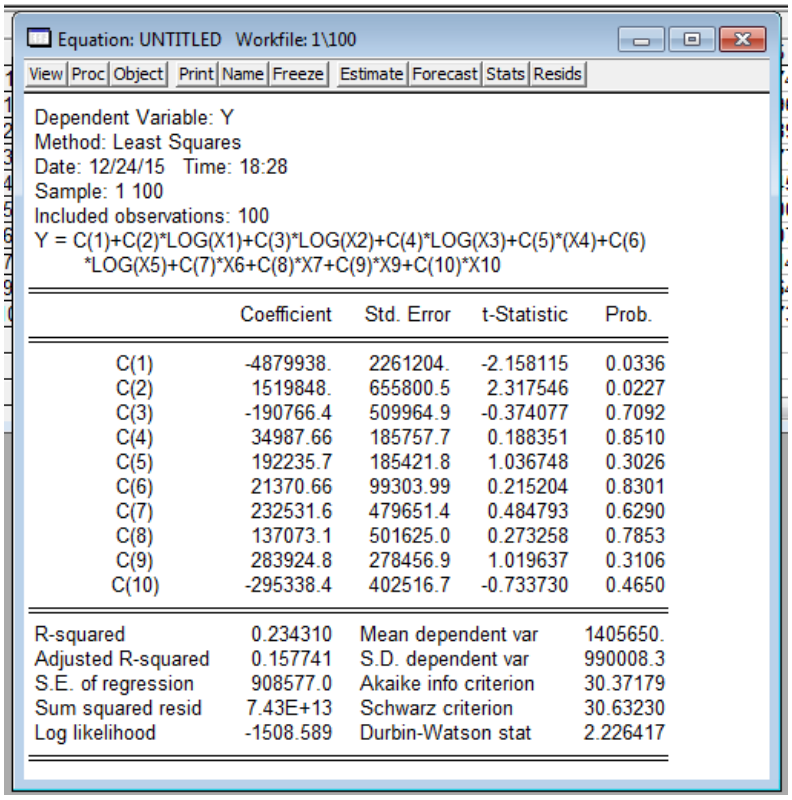


Рис.1. Расчётные данные линейного уравнения регрессии

Из рисунка 1, что значение коэффициента детерминации R^2=0.233334, поэтому полученное уравнение регрессии не значимое. Изменим вид модели на не линейный, прологарифмировав его. Общий вид модели:

$$Y = C(1)+C(2)*LOG(X1)+C(3)*LOG(X2)+C(4)*LOG(X3)+C(5)*(X4)+ C(6)*LOG(X5)+C(7)*X6+C(8)*X7+C(9)*X9+C(10)*X10$$



Equation: UNTITLED Workfile: 1\100

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y
Method: Least Squares
Date: 12/24/15 Time: 18:28
Sample: 1 100
Included observations: 100
Y = C(1)+C(2)*LOG(X1)+C(3)*LOG(X2)+C(4)*LOG(X3)+C(5)*(X4)+C(6) *LOG(X5)+C(7)*X6+C(8)*X7+C(9)*X9+C(10)*X10

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	-4879938.	2261204.	-2.158115	0.0336
C(2)	1519848.	655800.5	2.317546	0.0227
C(3)	-190766.4	509964.9	-0.374077	0.7092
C(4)	34987.66	185757.7	0.188351	0.8510
C(5)	192235.7	185421.8	1.036748	0.3026
C(6)	21370.66	99303.99	0.215204	0.8301
C(7)	232531.6	479651.4	0.484793	0.6290
C(8)	137073.1	501625.0	0.273258	0.7853
C(9)	283924.8	278456.9	1.019637	0.3106
C(10)	-295338.4	402516.7	-0.733730	0.4650

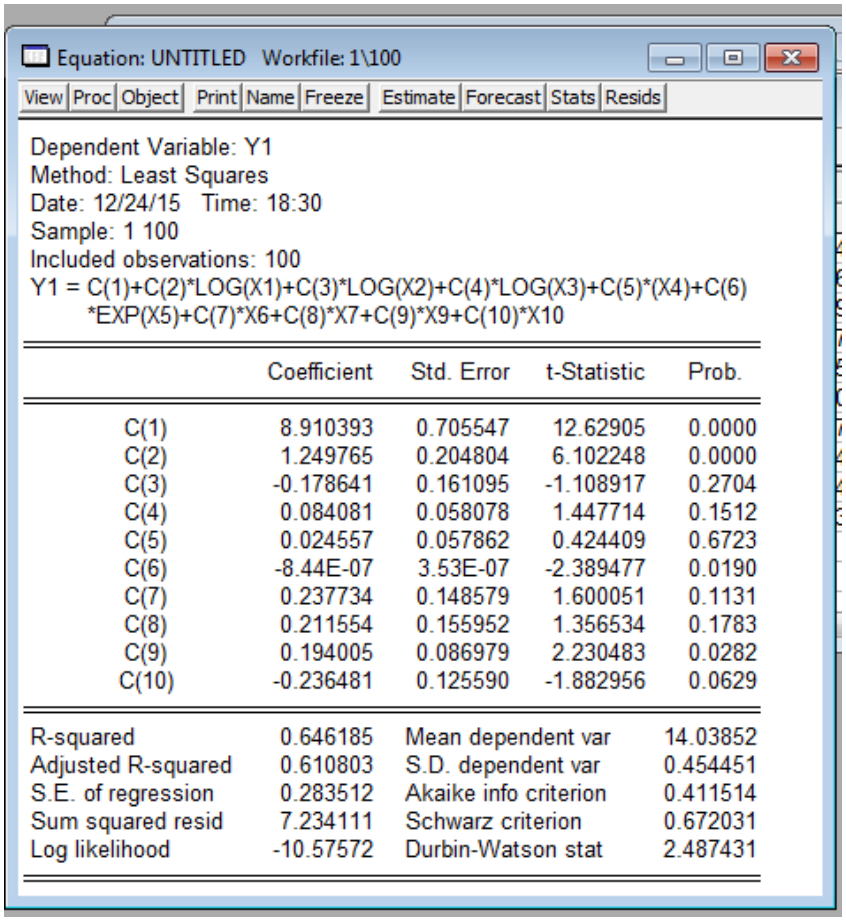
R-squared	0.234310	Mean dependent var	1405650.
Adjusted R-squared	0.157741	S.D. dependent var	990008.3
S.E. of regression	908577.0	Akaike info criterion	30.37179
Sum squared resid	7.43E+13	Schwarz criterion	30.63230
Log likelihood	-1508.589	Durbin-Watson stat	2.226417

Рис.2. Расчётные данные не линейного уравнения регрессии

Из рисунка 2 видно, что модель снова получилась, не значима, т.к. R^2=0.234310. Попробуем прологарифмировать зависимую переменную и вместо логарифмической функции у фактора X5 взять экспоненциальную. Обозначим теперь за Y1=LOG(Y).

Общий вид модели:

Y1= C(1)+C(2)*LOG(X1)+C(4)*EXP(X3)+C(5)*EXP(X4)+C(6)*EXP(X5) +C(8)*X7+C(9)*X9+C(10)*X10



Equation: UNTITLED Workfile: 1\100

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y1
Method: Least Squares
Date: 12/24/15 Time: 18:30
Sample: 1 100
Included observations: 100
Y1 = C(1)+C(2)*LOG(X1)+C(3)*LOG(X2)+C(4)*LOG(X3)+C(5)*(X4)+C(6)
*EXP(X5)+C(7)*X6+C(8)*X7+C(9)*X9+C(10)*X10

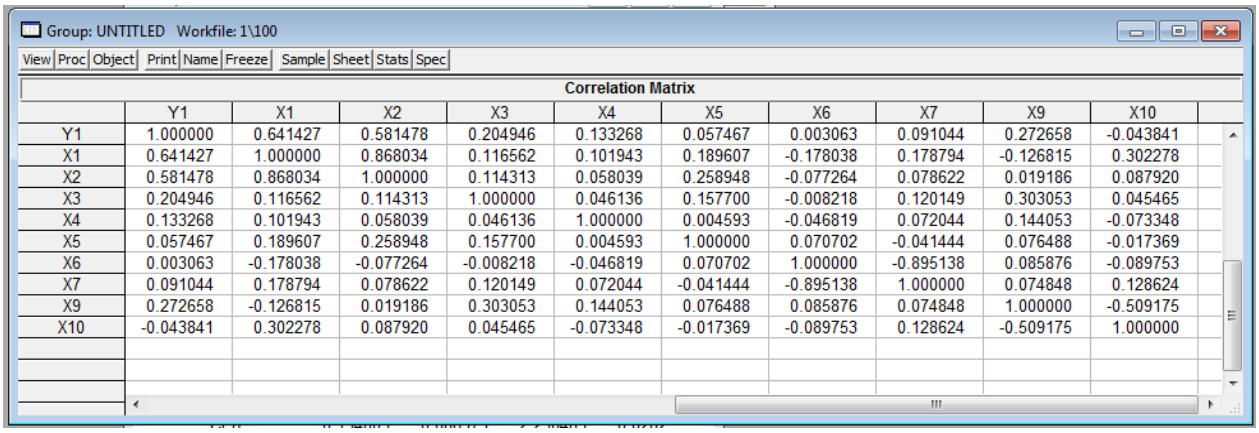
	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	8.910393	0.705547	12.62905	0.0000
C(2)	1.249765	0.204804	6.102248	0.0000
C(3)	-0.178641	0.161095	-1.108917	0.2704
C(4)	0.084081	0.058078	1.447714	0.1512
C(5)	0.024557	0.057862	0.424409	0.6723
C(6)	-8.44E-07	3.53E-07	-2.389477	0.0190
C(7)	0.237734	0.148579	1.600051	0.1131
C(8)	0.211554	0.155952	1.356534	0.1783
C(9)	0.194005	0.086979	2.230483	0.0282
C(10)	-0.236481	0.125590	-1.882956	0.0629

R-squared	0.646185	Mean dependent var	14.03852
Adjusted R-squared	0.610803	S.D. dependent var	0.454451
S.E. of regression	0.283512	Akaike info criterion	0.411514
Sum squared resid	7.234111	Schwarz criterion	0.672031
Log likelihood	-10.57572	Durbin-Watson stat	2.487431

Рис.3. Расчётные данные не линейного уравнения регрессии и логарифмированного Y

Рисунок 3 показывает, что выбранная нами модель значима, R^2=0.646185.

Теперь проверим факторы на мультиколлениарность и исключим не значимые коэффициенты.



Group: UNTITLED Workfile: 1\100

View | Proc | Object | Print | Name | Freeze | Sample | Sheet | Stats | Spec

Correlation Matrix

	Y1	X1	X2	X3	X4	X5	X6	X7	X9	X10
Y1	1.000000	0.641427	0.581478	0.204946	0.133268	0.057467	0.003063	0.091044	0.272658	-0.043841
X1	0.641427	1.000000	0.868034	0.116562	0.101943	0.189607	-0.178038	0.178794	-0.126815	0.302278
X2	0.581478	0.868034	1.000000	0.114313	0.058039	0.258948	-0.077264	0.078622	0.019186	0.087920
X3	0.204946	0.116562	0.114313	1.000000	0.046136	0.157700	-0.008218	0.120149	0.303053	0.045465
X4	0.133268	0.101943	0.058039	0.046136	1.000000	0.004593	-0.046819	0.072044	0.144053	-0.073348
X5	0.057467	0.189607	0.258948	0.157700	0.004593	1.000000	0.070702	-0.041444	0.076488	-0.017369
X6	0.003063	-0.178038	-0.077264	-0.008218	-0.046819	0.070702	1.000000	-0.895138	0.085876	-0.089753
X7	0.091044	0.178794	0.078622	0.120149	0.072044	-0.041444	-0.895138	1.000000	0.074848	0.128624
X9	0.272658	-0.126815	0.019186	0.303053	0.144053	0.076488	0.085876	0.074848	1.000000	-0.509175
X10	-0.043841	0.302278	0.087920	0.045465	-0.073348	-0.017369	-0.089753	0.128624	-0.509175	1.000000

Рис.4. Матрица корреляции не линейного уравнения

Из рисунка 4 видно, что можно исключить переменную X2. Это связано с высоким уровнем коррелируемости между X1 и X2. Но коррелируемость между X1 и У всё же выше, чем между X2 и У.

В итоге получаем модель вида:

Equation: UNTITLED22 Workfile: 1\100

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y1
Method: Least Squares
Date: 12/24/15 Time: 18:35
Sample: 1 100
Included observations: 100
Y1 = C(1)+C(2)*LOG(X1)+C(4)*LOG(X3)+C(5)*(X4)+C(6)*EXP(X5)+C(7)
*X6+C(8)*X7+C(9)*X9+C(10)*X10

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	9.574933	0.372859	25.67976	0.0000
C(2)	1.046504	0.091476	11.44019	0.0000
C(4)	0.085847	0.058129	1.476816	0.1432
C(5)	0.026187	0.057916	0.452148	0.6522
C(6)	-9.10E-07	3.49E-07	-2.609261	0.0106
C(7)	0.230885	0.148638	1.553337	0.1238
C(8)	0.211260	0.156148	1.352943	0.1794
C(9)	0.193810	0.087089	2.225440	0.0285
C(10)	-0.204555	0.122400	-1.671201	0.0981

R-squared	0.641351	Mean dependent var	14.03852
Adjusted R-squared	0.609821	S.D. dependent var	0.454451
S.E. of regression	0.283870	Akaike info criterion	0.405085
Sum squared resid	7.332953	Schwarz criterion	0.639551
Log likelihood	-11.25426	Durbin-Watson stat	2.481126

Рис.5. Расчётные данные не линейной модели с исключением фактора X2

Действую аналогично исключаем из модели фактор X6 – кирпичные дома. Получаем корреляционную матрицу:

Group: UNTITLED Workfile: 1\100

View | Proc | Object | Print | Name | Freeze | Sample | Sheet | Stats | Spec

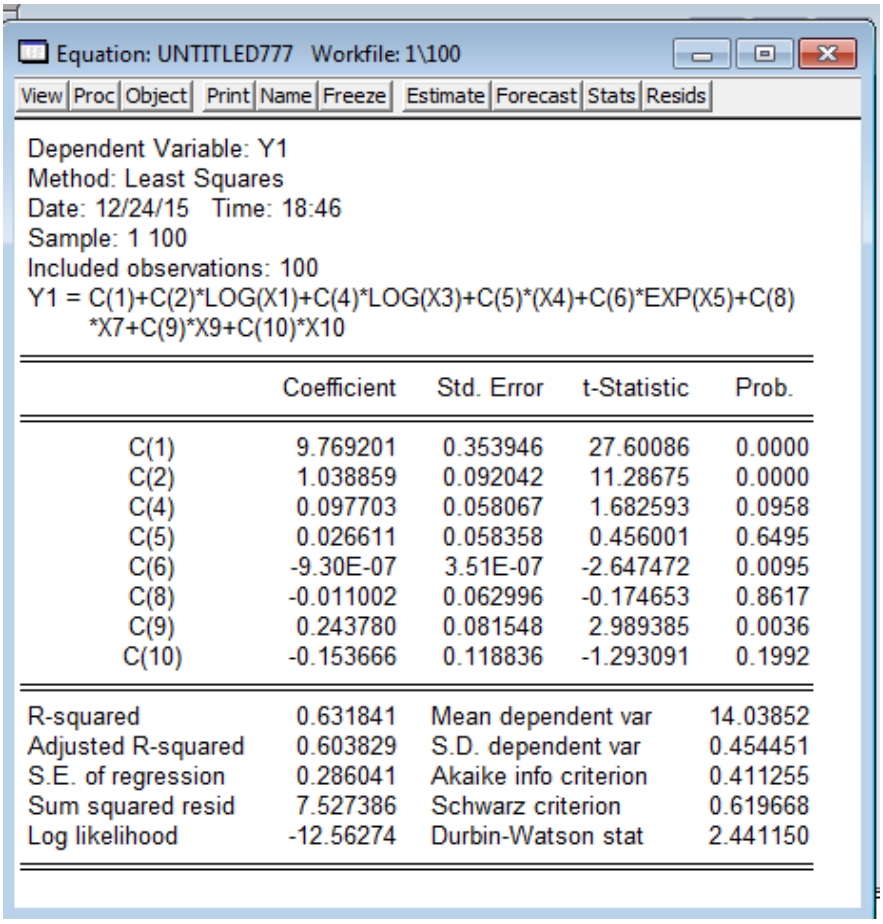
Correlation Matrix

	Y1	X1	X3	X4	X5	X7	X9	X10
Y1	1.000000	0.641427	0.204946	0.133268	0.057467	0.091044	0.272658	-0.043841
X1	0.641427	1.000000	0.116562	0.101943	0.189607	0.178794	-0.126815	0.302278
X3	0.204946	0.116562	1.000000	0.046136	0.157700	0.120149	0.303053	0.045465
X4	0.133268	0.101943	0.046136	1.000000	0.004593	0.072044	0.144053	-0.073348
X5	0.057467	0.189607	0.157700	0.004593	1.000000	-0.041444	0.076488	-0.017369
X7	0.091044	0.178794	0.120149	0.072044	-0.041444	1.000000	0.074848	0.128624
X9	0.272658	-0.126815	0.303053	0.144053	0.076488	0.074848	1.000000	-0.509175
X10	-0.043841	0.302278	0.045465	-0.073348	-0.017369	0.128624	-0.509175	1.000000

Рис.6. Матрица корреляции не линейного уравнения

Общий вид модели: Y1 = C(1)+C(2)*LOG(X1)+C(4)*LOG(X3)+C(5)*(X4)+ C(6)*EXP(X5)+C(8)*X7+C(9)*X9+C(10)*X10

Получаем расчётные данные для нового уравнения не линейной регрессии:



Equation: UNTITLED777 Workfile: 1\100

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y1
Method: Least Squares
Date: 12/24/15 Time: 18:46
Sample: 1 100
Included observations: 100
Y1 = C(1)+C(2)*LOG(X1)+C(4)*LOG(X3)+C(5)*(X4)+C(6)*EXP(X5)+C(8)*X7+C(9)*X9+C(10)*X10

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	9.769201	0.353946	27.60086	0.0000
C(2)	1.038859	0.092042	11.28675	0.0000
C(4)	0.097703	0.058067	1.682593	0.0958
C(5)	0.026611	0.058358	0.456001	0.6495
C(6)	-9.30E-07	3.51E-07	-2.647472	0.0095
C(8)	-0.011002	0.062996	-0.174653	0.8617
C(9)	0.243780	0.081548	2.989385	0.0036
C(10)	-0.153666	0.118836	-1.293091	0.1992

R-squared	0.631841	Mean dependent var	14.03852
Adjusted R-squared	0.603829	S.D. dependent var	0.454451
S.E. of regression	0.286041	Akaike info criterion	0.411255
Sum squared resid	7.527386	Schwarz criterion	0.619668
Log likelihood	-12.56274	Durbin-Watson stat	2.441150

Рис.7. Расчётные данные итогового уравнения

Таким образом, модель является значимой R^2=0.631841 и значимой.

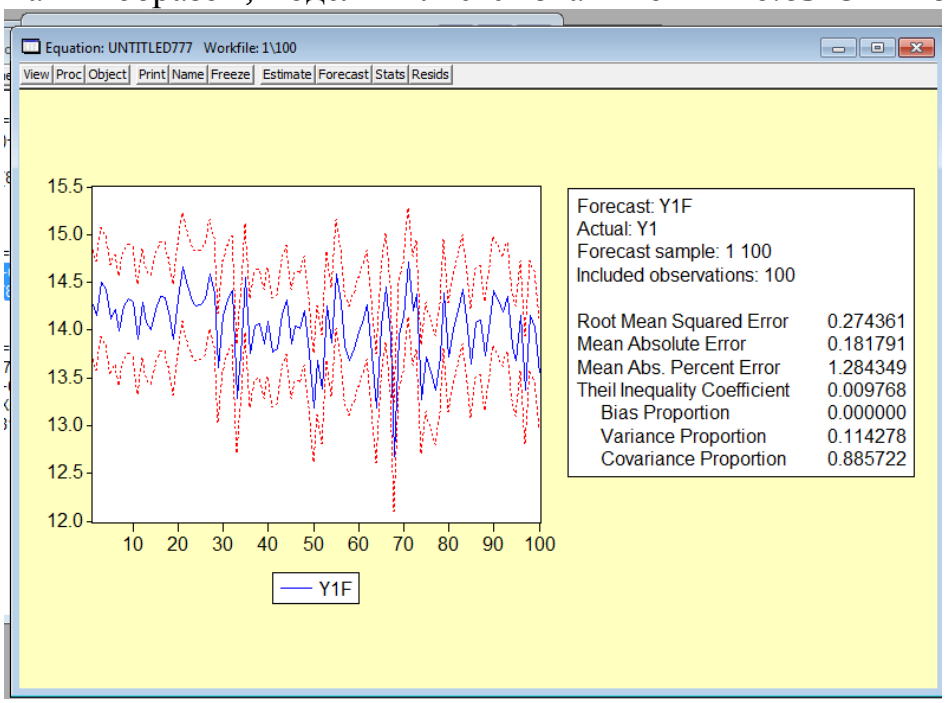


Рис.8. Ошибка аппроксимации

Средняя ошибка аппроксимации равна 1,284349<5%. Можно сделать вывод, что модель точная.

Подставив коэффициенты получаем не линейную модель множественной регрессии в виде:

$$Y1 = 9{,}7692 + 1{,}0389*LOG(X1) + 0{,}0977*LOG(X3) + 0{,}0266*(X4) - 0{,}0001*EXP(X5) - 0{,}011*X7 + 0{,}2438*X9 - 0{,}1537*X10$$

Построим точеный прогноз при условиях: X1=47,5(общая площадь); X3=2(этаж); X4=1(наличие ремонта); X5=1(расстояние от квартиры до центра города); X7=1(материал дома панельный); X9=0(не новый); X10=1(«брежневка»). Получаем:

Y1=9.762+1.0389*ln(47.5)+0.0977*ln(2)+0.0266-0.000001*exp(1)-0.011-0.1537.

Y1=13,7025, откуда получаем, что У=exp(y)= 891700

Таким образом, можно сделать вывод, что цена на квартиру зависит от жилищной площади, этажа, наличия ремонта, расстояния до центра, материала из которого сделан дом и «возраста» дома. Полученная модель соответствует реальным данным и позволяет прогнозировать стоимость квартир с различными исходными данными. Для этого необходимо всего лишь выбрать параметры интересуемой квартиры (площадь, месторасположение, «возраст дома» и т.д.) и подставить их в уравнение. В результате расчета мы получим стоимость квартиры. Однако использовать эту модель можно с определенной осторожностью, так как она не может дать 100 процентный результат, а лишь отражает среднюю стоимость квартир со схожими параметрами.