Практическое занятие № 4.

«Применение Eviews при построении и анализе многофакторной модели регрессии. Выявление мультиколлинеарности и гетероскедастичности в модели. Проверка спецификации модели»

Пример 4. Имеются данные о вариации дохода кредитных организаций США за период 25 лет в зависимости от изменений годовой ставки по сберегательным депозитам и числа кредитных учреждений[2].

Введем следующие обозначения:

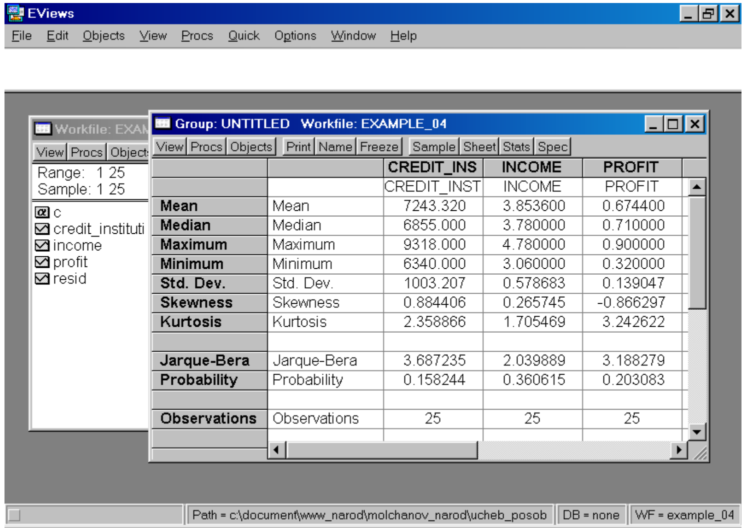
Y – прибыль кредитных организаций, %;

X_{1i} - чистый доход на 1$ депозита;

X_{2i} – число кредитных учреждений.

Год	X_{1i} (Income)	X_{2i} (Credit institutions)	Y (Profit)
1	3,92	7298	0,75
2	3,61	6855	0,71
3	3,32	6636	0,66
4	3,07	6506	0,61
5	3,06	6450	0,7
6	3,11	6402	0,72
7	3,21	6368	0,77
8	3,26	6340	0,74
9	3,42	6349	0,9
10	3,42	6352	0,82
11	3,45	6361	0,75
12	3,58	6369	0,77
13	3,66	6546	0,78
14	3,78	6672	0,84
15	3,82	6890	0,79
16	3,97	7115	0,7
17	4,07	7327	0,68
18	4,25	7546	0,72
19	4,41	7931	0,55
20	4,49	8097	0,63
21	4,7	8468	0,56
22	4,58	8717	0,41
23	4,69	8991	0,51
24	4,71	9179	0,47
25	4,78	9318	0,32

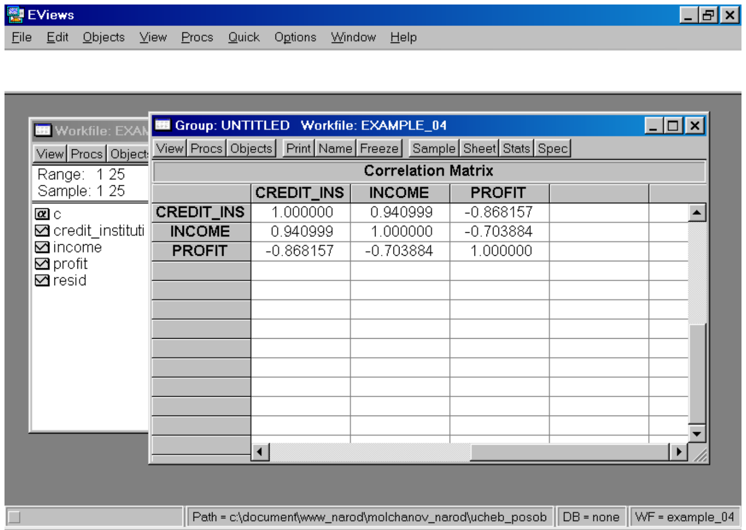
1. **Создать файл** с исходными данными в среде Excel (файл **example_04.xls**).
2. **Осуществить импорт** исходных данных в Eviews.
3. **Создать workfile.**
4. **Найти значения описательных статистик** по каждой переменной и объяснить их (рис. 58).



		CREDIT_INS	INCOME	PROFIT
		CREDIT_INST	INCOME	PROFIT
Mean	Mean	7243.320	3.853600	0.674400
Median	Median	6855.000	3.780000	0.710000
Maximum	Maximum	9318.000	4.780000	0.900000
Minimum	Minimum	6340.000	3.060000	0.320000
Std. Dev.	Std. Dev.	1003.207	0.578683	0.139047
Skewness	Skewness	0.884406	0.265745	-0.866297
Kurtosis	Kurtosis	2.358866	1.705469	3.242622
Jarque-Bera	Jarque-Bera	3.687235	2.039889	3.188279
Probability	Probability	0.158244	0.360615	0.203083
Observations	Observations	25	25	25

Рис. 58.

5. **Построить корреляционную матрицу для всех переменных, включенных в модель (рис. 59).**



	CREDIT_INS	INCOME	PROFIT
CREDIT_INS	1.000000	0.940999	-0.868157
INCOME	0.940999	1.000000	-0.703884
PROFIT	-0.868157	-0.703884	1.000000

Рис. 59.

6. **Построить регрессионное уравнение** МНК, в котором зависимая переменная – прибыль кредитных организаций, а независимые – чистый доход на 1$ депозита и число кредитных учреждений (рис. 60, 61).

[2] **Ниворожкина Л.И**. Текст лекций по начальному курсу эконометрики для аспирантов.

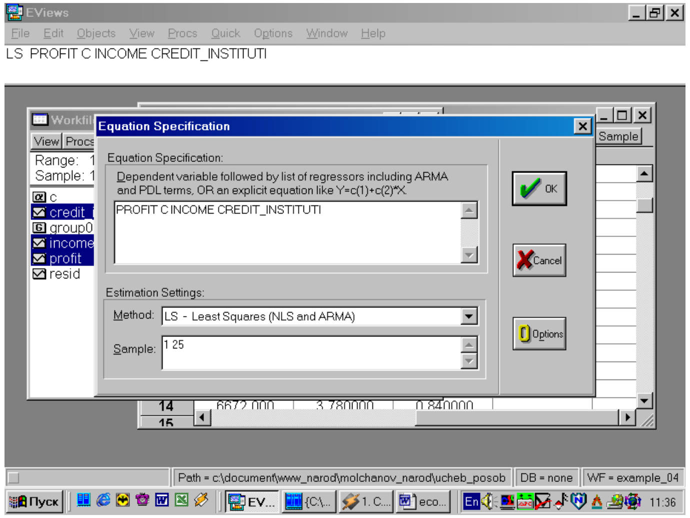


Рис. 60.

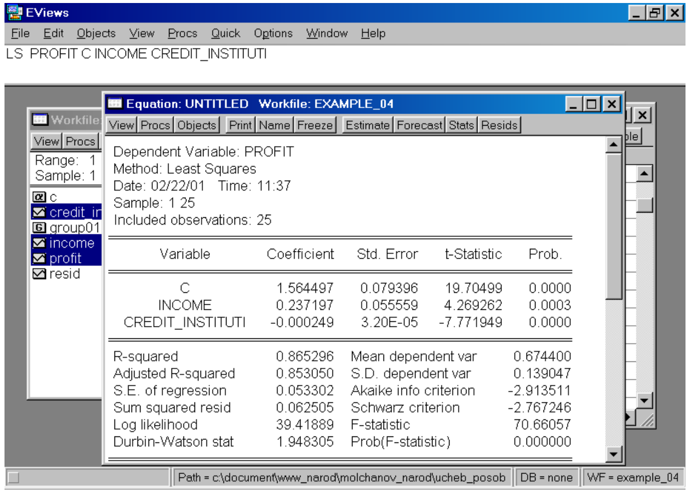


Рис. 61.

Уравнение примет следующий вид:

$$Y_i = \alpha_0 + \beta_1 INCOME + \beta_2 CREDIT\_INSTITUTI + u_i .$$

Подставим полученные оценки из итоговой формы вывода:

$$\hat{Y}_i = 1{,}5645 + 0{,}2372 INCOME - 0{,}00025 CREDIT\_INSTITUTI .$$

7. **Оценить статистическую значимость параметров полученного уравнения и всей модели в целом.**
8. **Проверить наличие мультиколлинеарности в модели. Сделать вывод.**

Мультиколлинеарность – это коррелированность двух или нескольких объясняющих переменных в уравнении регрессии. В результате высококоррелированные объясняющие переменные действуют в одном направлении и имеют недостаточно независимое колебание, чтобы дать возможность модели изолировать влияние каждой переменной. Проблема мультиколлинеарности возникает только в случае множественной регрессии. Мультиколлинеарность особенно часто имеет место при анализе макроэкономических данных (например, доходы, производство). Получаемые оценки оказываются нестабильными как в отношении статистической значимости, так и по величине и знаку (например, коэффициенты корреляции). Следовательно, они ненадежны. Значения коэффициентов R^2 могут быть высокими, но стандартные ошибки тоже высоки, и отсюда t- критерии малы, отражая недостаток значимости.

Для проверки появления мультиколлинеарности применяются два метода, доступные во всех статистических пакетах[3]:

- Вычисление матрицы коэффициентов корреляции для всех объясняющих переменных. Если коэффициенты корреляции между отдельными объясняющими переменными очень велики, то, следовательно, они коллинеарны. Однако, при этом не существует единого правила, в соответствии с которым есть некоторое пороговое значение коэффициента корреляции, после которого высокая корреляция может вызвать отрицательный эффект и повлиять на качество регрессии.
- Для измерения эффекта мультиколлинеарности используется показатель VIF – «фактор инфляции вариации»:
  - ✓ $VIF(X_h) = \dfrac{1}{1-R_h^2}$, где R_h^2 - значение коэффициента множественной корреляции, полученное для регрессора X_h как зависимой переменной и остальных переменных X_i. При этом степень мультиколлинеарности, представляемая в регрессии переменной X_h, когда переменные $X_1, X_2, ..., X_k$ включены в регрессию, есть функция множественной корреляции между X_h и другими переменными $X_1, X_2, ..., X_k$.
  - ✓ Если $VIF > 10$, то объясняющие переменные, коррелирующие между собой, считаются мультиколлинеарными.

Существует еще ряд способов, позволяющих обнаружить эффект мультиколлинеарности:

- Стандартная ошибка регрессионных коэффициентов близка к нулю.
- Мощность коэффициента регрессии отличается от ожидаемого значения.
- Знаки коэффициентов регрессии противоположны ожидаемым.
- Добавление или удаление наблюдений из модели сильно изменяют значения оценок.

[3] **Ниворожкина Л.И**. Текст лекций по начальному курсу эконометрики для аспирантов.

- Значение F-критерия существенно, а t-критерия – нет.

Для устранения мультиколлинеарности может быть принято несколько мер:

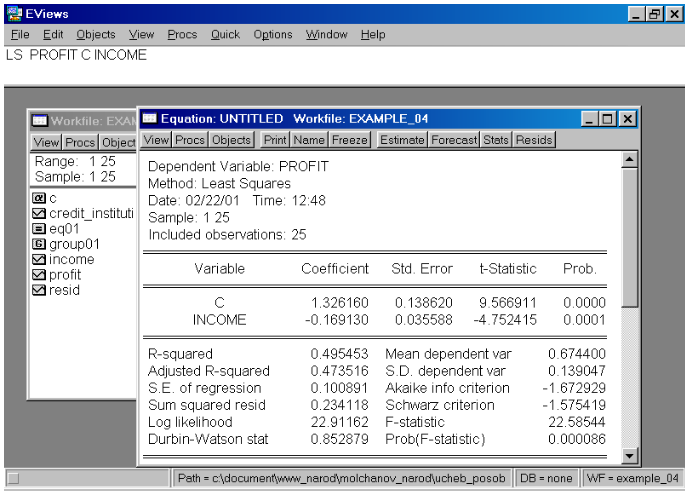
- Увеличивают объем выборки по принципу, что больше данных означает меньшие дисперсии оценок МНК. Проблема реализации этого варианта решения состоит в трудности нахождения дополнительных данных.
- Исключают те переменные, которые высококоррелированны с остальными. Проблема здесь заключается в том, что возможно переменные были включены на теоретической основе, и будет неправомочным их исключение только лишь для того, чтобы сделать статистические результаты «лучше».
- Объединяют данные кросс-секций и временных рядов. При этом методе берут коэффициент из, скажем, кросс-секционной регрессии и заменяют его на коэффициент из эквивалентных данных временного ряда.

Проделанные манипуляции позволяют предположить, что мультиколлинеарность может присутствовать (оценки любой регрессии будут страдать от нее в определенной степени, если только все независимые переменные не окажутся абсолютно некоррелированными), однако в данном примере это не влияет на результаты оценки регрессии. Следовательно, выделять «лишние» переменные не стоит, так как это отражается на содержательном смысле модели.

9. Проверить спецификацию модели. Объяснить полученные результаты.

Подробно теоретические вопросы, связанные с проблемами спецификации эконометрических моделей, были рассмотрены в лекционном курсе.

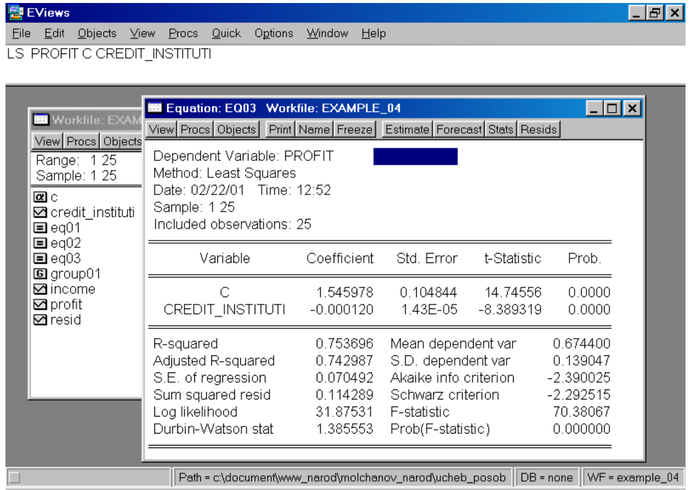
В нашем случае мы ограничимся тем, что попробуем исключить поочередно независимые переменные. Первой исключаем переменную **CREDIT_INSTITUTI** (рис. 62). Коэффициент при переменной **INCOME** изменил знак на противоположный.



Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.326160	0.138620	9.566911	0.0000
INCOME	-0.169130	0.035588	-4.752415	0.0001

R-squared	0.495453	Mean dependent var	0.674400
Adjusted R-squared	0.473516	S.D. dependent var	0.139047
S.E. of regression	0.100891	Akaike info criterion	-1.672929
Sum squared resid	0.234118	Schwarz criterion	-1.575419
Log likelihood	22.91162	F-statistic	22.58544
Durbin-Watson stat	0.852879	Prob(F-statistic)	0.000086

Рис. 62.



Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.545978	0.104844	14.74556	0.0000
CREDIT_INSTITUTI	-0.000120	1.43E-05	-8.389319	0.0000

R-squared	0.753696	Mean dependent var	0.674400
Adjusted R-squared	0.742987	S.D. dependent var	0.139047
S.E. of regression	0.070492	Akaike info criterion	-2.390025
Sum squared resid	0.114289	Schwarz criterion	-2.292515
Log likelihood	31.87531	F-statistic	70.38067
Durbin-Watson stat	1.385553	Prob(F-statistic)	0.000000

Рис. 63.

В случае исключения из первоначальной модели переменной **INCOME**, знак регрессионного коэффициента при переменой **CREDIT_INSTITUTI** остался без изменения (рис. 63). Представляется разумным разделять эффект двух независимых переменных на зависимую переменную в модели с совместным их влиянием в регрессионном уравнении. Данный пример иллюстрирует важность использования множественной регрессии вместо парной в случае, когда изучаемое явление существенно детерминирует несколько независимых переменных.

10. Проверить наличие гетероскедастичности в модели. Объяснить полученные результаты.

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 коэффициентами.

Воздействие гетероскедастичности на оценку интервала прогнозирования и проверку гипотезы заключается в том, что хотя коэффициенты не смещены, дисперсии и, следовательно, стандартные ошибки этих коэффициентов будут смещены. Если смещение отрицательно, то оценочные стандартные ошибки будут меньше, чем они должны быть, а критерий проверки будет больше, чем в реальности. Таким образом, мы можем сделать вывод, что коэффициент значим, когда он таковым не является. И наоборот, если смещение положительно, то оценочные ошибки будут больше, чем они должны быть, а критерии проверки – меньше. Значит, мы можем принять нулевую гипотезу, в то время как она должна быть отвергнута.

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. Отсюда число остатков в каждой группе составляет $(n-c)/2$, где c представляет одну шестую часть наблюдений.

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

$$(n-c)/2.$$

Этот критерий имеет $t-$распределение с $(n-c)/(2-k)$ степенями свободы.

Чтобы решить проблему гетероскедастичности, нужно исследовать взаимосвязь между значениями ошибки и переменными и трансформировать регрессионную модель так, чтобы она отражала эту взаимосвязь. Это может быть достигнуто посредством регрессии значений ошибок по различным формам функций переменной, которая приводит к гетероскедастичности, например,

$$e_i = \alpha + \beta \bullet X_i^H,$$

где X_i - независимая переменная (или какая-либо функция независимой переменной), которая предположительно является причиной гетероскедастичности, а H отражает степень взаимосвязи между ошибками и данной переменной, например, X^2 или $X^{1/n}$ и т. д.

Следовательно, дисперсия коэффициентов запишется:

$$E(\sigma_i^2) = \sigma^2 X_i^H.$$

Отсюда если $H=1$, мы трансформируем регрессионную модель к виду:

$$\frac{Y_i}{\sqrt{X_i}} = \frac{\alpha}{\sqrt{X_i}} + \beta_i \frac{e_i}{\sqrt{X_i}}.$$

Если $H=2$, т.е. дисперсия увеличивается в пропорции к квадрату рассматриваемой переменной X, трансформация приобретает вид:

$$\frac{Y_i}{X_i} = \frac{\alpha}{X_i} + \beta_i \frac{e_i}{X_i}.$$

Используя Eviews, можно провести проверку и устранение гетероскедастичности следующим образом:

- Запустить стандартную регрессию.
- Вычислить остатки.
- Запустить регрессию с использованием квадрата остатков как зависимой переменной и оценить зависимую переменную $\hat{y}$ как независимую переменную (тест White).

- Оценить nR^2, где n – объем выборки, R^2 – коэффициент детерминации.
- Использовать статистику χ^2 с одной степенью свободы (в EVIEWS – используется F – статистика) для проверки существенности отличия nR^2 от нуля.
- Основным способом устранения гетероскедастичности является применение взвешенного метода наименьших квадратов.

Выбираем тест White (см. рис. 64).

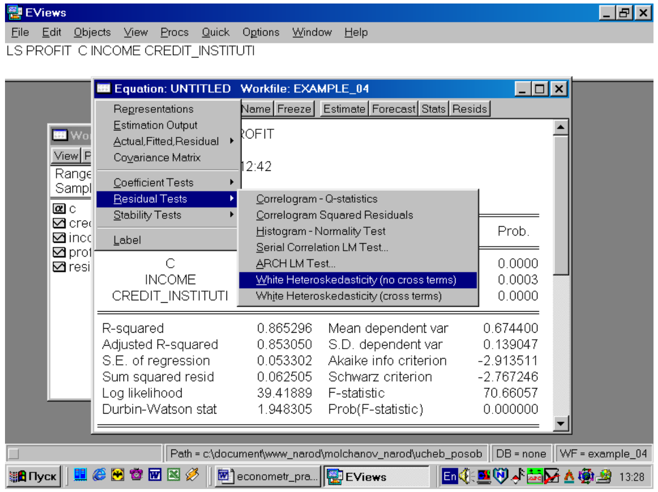


Рис. 64.

Итог формы вывода представлен на рис. 65.

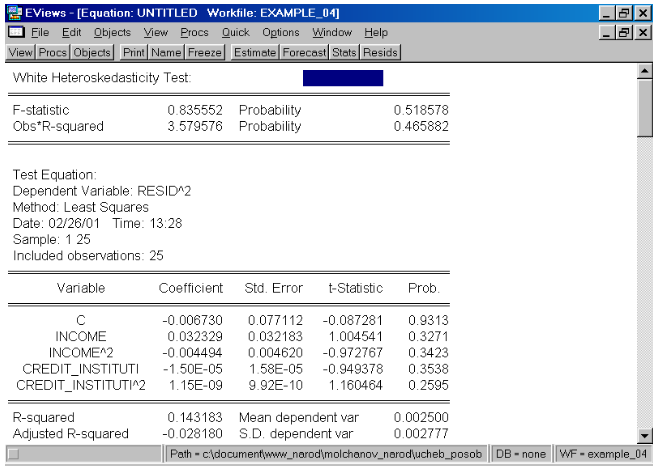


Рис. 65.

Как следует из приведенной распечатки, вероятность ошибки первого рода равна 51,86%. Следовательно, нулевую гипотезу (об отсутствии гетероскедастичности) нельзя отклонить.

Для случая, когда гетероскедастичность присутствует, проблему гетероскедастичности можно решать следующим образом:

Выбираем в пунктах меню текущего окна опцию **Proc/Specify/Estimate…** (рис. 66). Появляется окно оценки регрессии, где необходимо нажать клавишу **Options** и в появившимся окне отметить **Heteroskedasticity** (рис. 67).

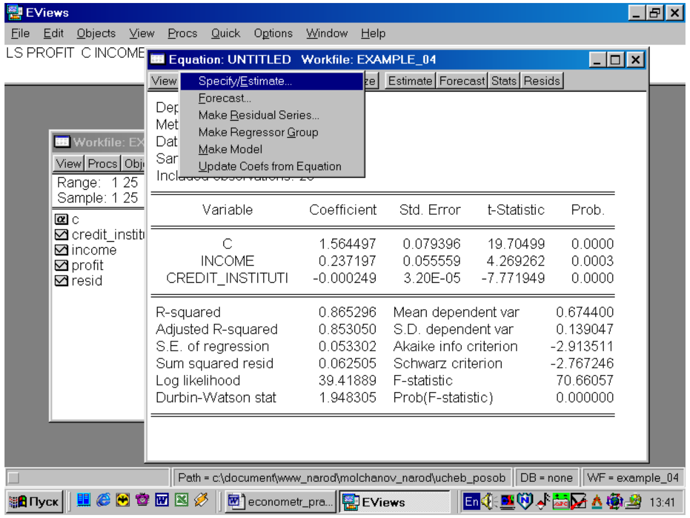


Рис. 66.

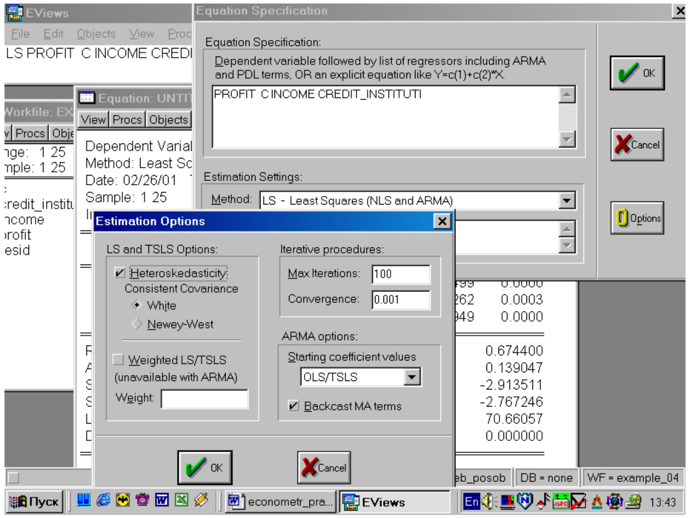


Рис. 67.

Появилось новое, переоцененное уравнение (рис. 68). Полученное уравнение можно вновь проверить по тесту White.

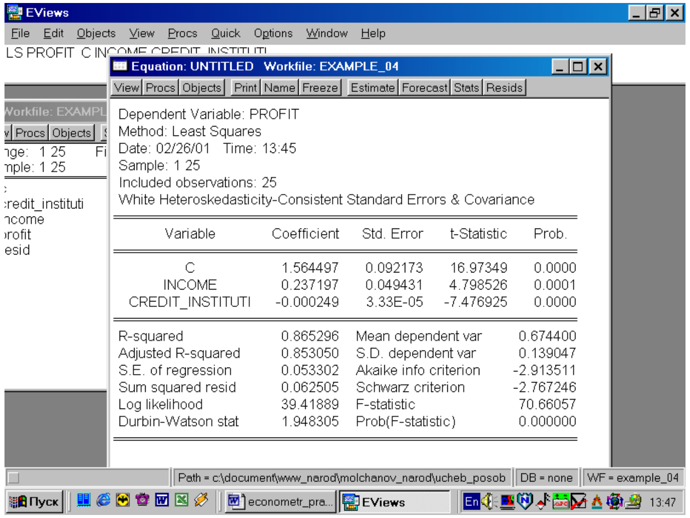


Рис. 68.

11. Оформить отчет.