

ГЛАВА 4. ФИКТИВНЫЕ ПЕРЕМЕННЫЕ В РЕГРЕССИОННЫХ МОДЕЛЯХ

При построении уравнений регрессии может возникнуть необходимость включить в модель качественные признаки. Например, при моделировании зависимости между заработком и продолжительностью образования для женщин и мужчин, при учете этнических различий в модели потребления, при анализе влияния формы правления на эффективность помощи развивающимся странам и во многих других случаях.

Для учета качественного признака в модели используются так называемые *фиктивные* переменные (*d* или *dummy*), которым обычно присваиваются только два значения: 0 (признак отсутствует) и 1 (признак присутствует).

Если качественный признак имеет не два, а несколько значений, используют несколько фиктивных переменных. Количество фиктивных переменных должно быть на 1 меньше, чем количество возможных значений у качественного признака. Одно значение качественного признака (одна категория) принимается за эталон.

В качестве эталонной выбирают преобладающую или наиболее характерную категорию. Например, при моделировании зависимости между заработком и продолжительностью образования с учетом гендорного признака, в модель вводят переменную *Male*, равную единице, если респондент – мужчина и 0, если респондент – женщина.

Фиктивная переменная может быть включена в уравнение регрессии как слагаемое. Например, при моделировании зависимости издержек на обучения (переменная *COST*) от количества учеников (переменная *N*)

может быть учтено, что затраты на обучение в профессиональной школе в среднем выше, чем в обычной путем введения в уравнения регрессии фиктивной переменной.

$$COST = \beta_1 + \delta \cdot d + \beta_2 N + \varepsilon$$

$$d = \begin{cases} 1 - \text{профессиональная школа} \\ 0 - \text{обычная школа} \end{cases}$$

Коэффициент δ при фиктивной переменной показывает, насколько в среднем объем постоянных затрат в профессиональной школе выше, чем в обычной.

При помощи фиктивных переменных наклона можно построить кусочно-линейные модели, которые позволяют учесть структурные изменения в экономических процессах.

$$Y_i = \beta_0 + \beta_1 X_i + \delta \cdot d_i \cdot X_i + \varepsilon_i$$

$$d_i = \begin{cases} 0 & \text{до структурных изменений} \\ 1 & \text{после структурных изменений} \end{cases}$$

Альтернативой введения фиктивных переменных является разбиение исходной выборки на две подвыборки (до структурных изменений и после). Наличие структурных изменений проверяется тестом Чоу.

Практическое задание № 4. Использование фиктивных переменных в регрессионных моделях.

Цель работы: научиться работать с фиктивными переменными, изучить методы построения и оценки уравнений регрессии с использованием фиктивных переменных.

Условие задачи. По 493 наблюдениям изучается зависимость цены квартиры (тыс. долл. США) $price$ от

различных факторов. Имеются следующие данные о проданных квартирах: размер жилой площади (кв.м) livsp, площадь кухни (кв.м) kitsp, общая площадь (кв.м) totsp, расстояние до центра (км) dist, сколько времени добираться до метро (мин) metrdist, количество комнат room (зависимость от этих факторов была изучена в предыдущей работе). Для учета влияния факторов, отражающих качественные признаки, используется набор следующих фиктивных переменных

Признак	Значение переменной	Условие
наличие балкона	bal=1 bal=0	балкон есть в противном случае
наличие телефона	tel=1 tel=0	телефон есть в противном случае
категория дома	brick=1 brick=0	дом - кирпичный в противном случае
этаж	floor=1 floor=1	не первый/последний в противном случае
расположение по отношению к метро	walk=1 walk=0	пешком в противном случае

Исходные данные взяты из [6], приведены на странице учебника <http://econometrics.nes.ru/mkp/> в разделе **Материалы к примерам и задачам**, файл **flat98s.xls**, а также на странице курса «Эконометрика» на учебном портале economist.rudn.ru в файле **flat98.wf1**.

Задание

Будем основываться на результатах работы №3. В качестве исходной модели воспользуемся двойной логарифмической моделью (mod 3.3), результат оценки которого приведен на рис. 3.4.

1. Проверьте целесообразность включения фиктивных переменных в уравнение регрессии (mod 3.3), используя тест на пропущенные переменные. Сделайте выводы.
2. Оцените модель с полным набором значимых факторов, включая фиктивные переменные. Сравните качество оценки модели, учитывающей только количественные факторы, с моделью с фиктивными переменными по следующим критериям: скорректированный коэффициент детерминации, критерий Акаике, критерий Шварца.
3. Проинтерпретируйте коэффициенты регрессии при фиктивных переменных.
4. Сравните точность модели, учитывающей только количественные факторы, с точностью модели с включенными фиктивными переменными. Сделайте выводы.

Решение практического задания № 4 с использованием программы Eviews.

1. Чтобы проверить целесообразность включения в модель качественных переменных (наличие балкона, этажность квартиры, наличие телефона, возможность добраться до метро пешком) воспользуйтесь одним из вариантов, реализованных в Eviews, и проведите тест на пропущенные переменные. Для этого в окне уравнения выберите команду **View → Coefficient Tests → Omitted Variables – Likelihood Ratio...** и в открывшемся окне введите фиктивные переменные как показано на рис. 4.1.

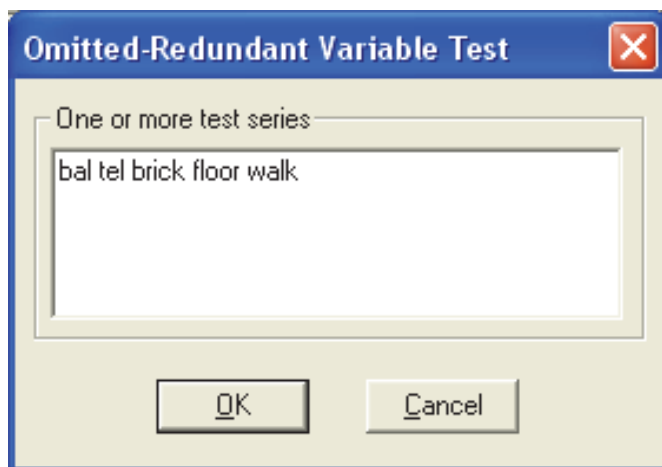


Рис. 4.1. Диалоговое окно для проведения теста на пропущенные переменные

В результате теста **Eviews** оценит уравнение, включив не только переменные из предыдущей модели, но и фиктивные переменные. Результат выполнения теста представлен на рис. 4.2.

Как видно, совокупное влияние введенных фиктивных переменных значимо ($F\text{-statistic} = 25.63$, $\text{Probability } 0.00$). Поэтому модель существенно улучшится после включения фиктивных переменных.

Обратите внимание, что, после введения в модель переменной **Walk** коэффициент при переменной ***log(metrdist)*** стал незначим ($\text{Prob.}=0,17$). При рассмотрении влияния этих двух переменных на качественном уровне, приходим к выводу, что переменная **Walk** более существенна и в окончательной модели можно опустить фактор ***metrdist***.

	A	B	C	D	E
1	Omitted Variables: BAL TEL BRICK FLOOR WALK				
2					
3	F-statistic	25.63067	Prob. F(5,483)	0.0000	
4	Log likelihood ratio	116.0183	Prob. Chi-Square(5)	0.0000	
5					
6					
7	Test Equation:				
8	Dependent Variable: LOG(PRICE)				
9	Method: Least Squares				
10	Date: 12/07/10 Time: 22:28				
11	Sample: 1 493				
12	Included observations: 493				
13					
14	Variable	Coefficient	Std. Error	t-Statistic	Prob.
15					
16	C	-0.480209	0.127521	-3.765734	0.0002
17	LOG(TOTSP)	0.958876	0.029444	32.56579	0.0000
18	LOG(KITSP)	0.284894	0.044413	6.414603	0.0000
19	LOG(METRDIST)	-0.021498	0.015756	-1.364447	0.1731
20	LOG(DIST)	-0.117557	0.013874	-8.473165	0.0000
21	BAL	0.084520	0.021969	3.847155	0.0001
22	TEL	0.151923	0.026883	5.651158	0.0000
23	BRICK	0.096838	0.019402	4.991189	0.0000
24	FLOOR	0.119121	0.022686	5.250932	0.0000
25	WALK	0.076490	0.022622	3.381165	0.0008
26					
27					

Path = i:\документы - sveta\статистика\финансы DB = finmarke WF = flat98

Рис. 4.2. Результат теста на пропущенные переменные

2. Оцените модель с полным набором значимых факторов

$$\log(\text{price}) = \beta_0 + \beta_1 \log(\text{totsp}) + \beta_2 \log(\text{kitsp}) + \beta_3 \log(\text{dist}) + \\ + \delta_1 \text{bal} + \delta_2 \text{tel} + \delta_3 \text{brick} + \delta_4 \text{floor} + \delta_5 \text{walk}$$

(mod 4.1)

Для этого в окне оценки уравнения введите набор переменных как показано на рис. 4.3.

**LOG(PRICE) C LOG(TOTSP) LOG(KITSP) LOG(DIST) BAL TEL
BRICK FLOOR WALK**

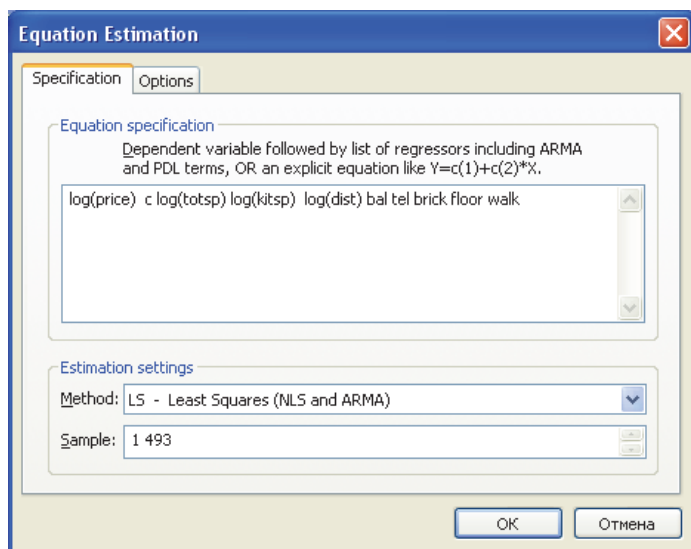


Рис. 4.3. Окно ввода параметров для оценки уравнения

Результат оценки уравнения представлен на рис. 4.4. Сохраните уравнение для последующей работы под именем EQ04.

Полученное уравнение статистически значимо и имеет высокий коэффициент детерминации ($F\text{-statistic}=303$, $\text{Prob}(F\text{-statistic})=0$, $R^2 = 0.83$).

Чтобы сравнить качество оценки двух уравнений с разным количеством факторов (в данном случае уравнение (mod 3.3) имеет 4 фактора, а уравнение (mod 4.1) – 8 факторов), более корректно использовать скорректированный коэффициент детерминации R^2_{adj} , который компенсирует автоматическое увеличение R^2 за счет увеличения количества факторов.

При оценке уравнения (3.4) $R^2_{\text{adj}} = 0,79$ (см. рис. 3.4), а при оценке уравнения (4.1) $R^2_{\text{adj}} = 0,83$ (см. рис. 4.4), что

говорит о существенном улучшении качества оценки после введения фиктивных переменных.

Dependent Variable: LOG(PRICE)				
Method: Least Squares				
Sample: 1 493				
Included observations: 493				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.530746	0.122132	-4.345675	0.0000
LOG(TOTSP)	0.959855	0.029462	32.57971	0.0000
LOG(KITSP)	0.284913	0.044453	6.409325	0.0000
LOG(DIST)	-0.119264	0.013830	-8.623643	0.0000
BAL	0.086850	0.021922	3.961685	0.0001
TEL	0.152808	0.026900	5.680712	0.0000
BRICK	0.099050	0.019351	5.118574	0.0000
FLOOR	0.119604	0.022703	5.268168	0.0000
WALK	0.081983	0.022281	3.679537	0.0003
R-squared	0.833762	Mean dependent var		4.209515
Adjusted R-squared	0.831014	S.D. dependent var		0.487255
Akaike info criterion				
S.E. of regression	0.200300	criterion		-0.3599
Sum squared resid	19.41821	Schwarz criterion		-0.2832
Log likelihood	97.71773	Durbin-Watson stat		1.652900
F-statistic	303.4353			
Prob(F-statistic)	0.000000			

Рис. 4.4. Результат оценки уравнения (mod 4.1)

Качество оценки моделей можно также сравнивать по критериям Акаике и Шварца. Из двух моделей с разным количеством факторов предпочтение отдается той, у которой значения этих критериев меньше. Как видно из рис. 3.4 и 4.4 оба эти критерия меньше в уравнении (mod 4.1) с фиктивными переменными.

Критерий	Уравнение (mod 3.3)	Уравнение (mod 4.1)
Akaike info criterion	-0.144653	-0.3599
Schwarz criterion	-0.102051	-0.2832

Таким образом, из двух моделей по всем рассмотренным критериям предпочтительной является модель (mod 4.1).

3. Интерпретация коэффициентов при фиктивных переменных:

- Коэффициент при переменной **Bal** показывает, что при наличии балкона при прочих равных условиях (одинаковой общей площади, площади кухни и т.д.) цена квартиры вырастет в среднем на 8,7% (коэффициент при факторе умножается на 100 и выражается в процентах).
- Коэффициент при переменной **tel** показывает, что при наличии телефона при прочих равных условиях цена квартиры вырастет в среднем на 15%.
- Коэффициент при переменной **brick** показывает, что цена квартиры в кирпичном доме выше при прочих равных условиях в среднем на 9,9%.
- Коэффициент при переменной **floor** показывает, что цена квартиры на любом этаже, кроме первого или последнего, выше в среднем на 12% по сравнению с такой же квартирой на первом или последнем этаже.
- Коэффициент при переменной **walk** показывает, что цена квартиры при прочих равных условиях возрастает в среднем на 8,2% при возможности дойти до метро пешком.

4. Одной из характеристик **точности оценки** модели является средняя относительная ошибка аппроксимации (чем значение этого показателя меньше, тем модель лучше), которую можно получить, используя меню уравнения с помощью команды **Forecast**.

Для уравнения mod 3.3, откройте EQ02 (оценка уравнения mod 3.3), щелкните подменю **Forecast** и в открывшемся окне введите в поле **Forecast name** имя **Pricef_3_3**, в поле **Forecast sample** задайте диапазон 1 493 и нажмите ОК.

Для уравнения mod 4.1, откройте EQ04 (оценка уравнения 4.1), щелкните подменю **Forecast** и в открывшемся окне введите в поле **Forecast name** имя **Pricef_4_1**, в поле **Forecast sample** задайте диапазон 1 493 и нажмите ОК. Для дальнейшей работы результаты графиков и данных сохраните под предлагаемыми программой именами. Сравните значения средних относительных ошибок аппроксимации (**Mean Abs. Percent Error**) для двух моделей.

Forecast PRICEF_3_3		Forecast PRICEF_4_1	
Actual: PRICE		Actual: PRICE	
Forecast sample: 1 493		Forecast sample: 1 493	
Included observations: 493		Included observations: 493	
Root Mean Squared Error	24.8091	Root Mean Squared Error	22.71818
Mean Absolute Error	14.8481	Mean Absolute Error	13.32530
Mean Abs. Percent Error	17.1702	Mean Abs. Percent Error	15.31403
Theil Inequality Coefficient	0.14631	Theil Inequality Coefficient	0.133586
Bias Proportion	0.00720	Bias Proportion	0.006217
Variance Proportion	0.09745	Variance Proportion	0.101474
Covariance Proportion	0.89534	Covariance Proportion	0.892308

Как видно из представленных выше результатов средняя относительная ошибка аппроксимации для модели 3.3 равна 17,17%, а для модели 4.1 равна 15,3%. В таком случае по данному критерию модель 4.1 предпочтительнее.

По всем рассмотренным в работе критериям модель (mod 4.1) является более адекватной, введение фиктивных

переменных, отражающих качественные признаки, существенно улучшило качество исходной модели.

Задание 4 для самостоятельной работы

При составлении задания использованы материалы из [5].

По результатам опроса 540 жителей США в возрасте от 37 до 44 лет, проведенного в 2002г., собраны данные по следующим факторам

Обозначение	Описание	Единица измерения
EARNINGS	Текущая почасовая заработная плата	долл США
S	Число завершенных лет обучения	год
MALE	=1 для респондентов -мужчин	безразмерная
FEMALE	=1 для респондентов -женщин	безразмерная
ETHBLACK	=1 для респондентов -афроамериканцев	безразмерная
ETHHISP	=1 для респондентов -латиноамериканцев	безразмерная
ETHWHITE	=1 для респондентов остальных этнических групп	безразмерная
ASVABC	общий результат выполнения тестов на познавательные способности	балл

Требуется:

1. Построить модель, отражающую зависимость почасовой заработной платы от образования и умственных способностей (модель 1).

$$\ln(EARNINGS) = \beta_0 + \beta_1 S + \beta_2 ASVABC + \varepsilon \quad (1)$$

2. Оценить полученное уравнение, проверить его качество, дать интерпретацию коэффициентам уравнения.
3. Проверить, есть ли различие в уровне заработной платы для мужчин и женщин, введя фиктивную переменную *Male* (модель 2). Дать интерпретацию полученным результатам.

$$\ln(EARNINGS) = \beta_0 + \beta_1 S + \beta_2 ASVAVC + \delta \cdot Male + \varepsilon \quad (2)$$

4. Проверить, влияет ли на уровень заработной платы этническое происхождение. Построить модель с фиктивными переменными *ETHBLACK* и *ETHHISP*, выбрав в качестве эталонной переменной *ETHWHITE* (модель 3). Проинтерпретировать коэффициенты при фиктивных переменных.

$$\ln(EARNINGS) = \beta_0 + \beta_1 S + \beta_2 ASVAVC + \delta \cdot Male + \gamma_1 ETHBLACK + \gamma_2 ETHHISP + \varepsilon \quad (3)$$

5. Проверить, лучше ли модель 3 объясняет поведение эндогенной переменной, чем модель 2 с помощью соответствующего F-теста и скорректированного коэффициента детерминации. Сделать вывод.

Варианты заданий

Исходные данные взяты из [5], приведены на странице учебника в разделе **Data sets in Eviews format** <http://econ.lse.ac.uk/courses/ec220/G/iedata/eecs/>, а также на странице курса «Эконометрика» на учебном портале economist.rudn.ru. Для выполнения самостоятельной работы скачайте набор данных **Data set#** (#=2-22 в зависимости от номера варианта) с указанного ресурса или с портала [Economist](http://economist.rudn.ru) (файлы **eaef02.wf1- eaef22.wf1**)

Контрольные вопросы

1. Что такое фиктивные переменные? Какие значения они могут принимать?
2. При каких условиях строится уравнение множественной регрессии с фиктивными переменными?
3. Влияние каких факторов учитывают фиктивные переменные: качественных или количественных?
4. Как называют нулевое значение фиктивной переменной?
5. Как интерпретируются коэффициенты при фиктивных переменных?
6. Как проверить, улучшилось ли качество модели после введения фиктивных переменных?
7. На сколько процентов в среднем заработная плата у мужчин выше, чем у женщин для вашего набора данных?