Эконометрика, 2020-2021, 2 модуль
Семинар 5
23.11.20
для
Группы Э_Б2018_Э_3
Семинарист О.А.Демидова

Выбор функциональной формы модели

I. По данным файла dougherty.dta, используя тест Бокса-Кокса, с помощью статистического пакета STATA, оцените параметры модели

$$EARNINGS^{(\theta)} = \beta_0 + \beta_1 S^{(\lambda)} + \beta_2 ASVABC^{(\lambda)} + \varepsilon$$

1) Когда переменные левой и правой части преобразуются не одинаково:
Это можно сделать с помощью команды:

```
boxcox EARNINGS S ASVABC, model (theta)
```

2) Когда переменные в обеих частях модели преобразуются одинаково.
Это можно сделать с помощью команды:

```
boxcox EARNINGS S ASVABC, model (lambda),
```

3) Когда преобразуется только зависимая переменная
Это можно сделать с помощью команды

```
boxcox EARNINGS S ASVABC, model (lhsonly),
```

4) Когда преобразуются только независимые переменные
Это можно сделать с помощью команды

```
boxcox EARNINGS S ASVABC, model (rhsonly),
```

5) поэкспериментируйте с проведением теста Бокса-Кокса в модели с другим набором переменных (если Вы не хотите преобразовывать какие-то переменные, например, dummy, то это можно сделать с помощью команды, аналогичной следующей:

```
boxcox EARNINGS S ASVABC, notrans(MALE) model (theta)
boxcox EARNINGS S ASVABC, notrans(MALE) model (lambda)
```

II. По данным файла dougherty.dta выберите между линейной, полулогарифмической и линейной в логарифмах моделями с помощью теста
А) РЕ теста Дэвидсона, Уайта и МакКиннона
б) Бера и МакАлера.

а) РЕ теста Дэвидсона, Уайта и МакКиннона
Необходимые команды
Выбор между линейной и линейной в логарифмах моделью
Сначала оценим линейную модель с помощью команды:

```
reg EARNINGS S ASVABC,
```

Сохранить предсказанные значения зависимой переменной можно с помощью команды:

```
predict y_hat
```

Предварительно создав необходимые дополнительные переменные,
аналогично оценим линейную в логарифмах модель с помощью команды,

```
gen lnEARNINGS = ln(EARNINGS)
gen lnS = ln(S)
```

```
gen lnASVABC = ln(ASVABC)
reg lnEARNINGS lnS lnASVABC
```
и сохраним предсказанные значения зависимой переменной с помощью команды:
```
predict ln_y_hat.
```
Теперь переходим к шагу 2, оценим дополнительные модели.
Сначала создадим дополнительную разность для линейная модели:
```
gen lin_add= ln(y_hat)- ln_y_hat
```
и оценим эту модель:
```
reg EARNINGS S ASVABC lin_add
```
Создадим также дополнительную разность для линейной в логарифмах модели:
```
gen log_add=y_hat-exp(ln_y_hat)
```
оценим эту модель:
```
reg lnEARNINGS lnS lnASVABC log_add
```
Если

1) оба коэффициента при дополнительных разностях значимы или оба незначимы, то выбрать посредством теста Дэвидсона, Уайта и МакКиннона невозможно,
2) если незначим только коэффициент при дополнительной разности в линейной модели, то лучше линейная модель,
3) если незначим только коэффициент при дополнительной разности в линейной в логарифмах модели, то лучше линейная в логарифмах модель.

Выбор между линейной и полулогарифмической моделями
Сначала оценим линейную модель с помощью команды:
```
reg EARNINGS S ASVABC
```
Сохранить предсказанные значения зависимой переменной можно с помощью команды:
```
predict y_hat
```
Аналогично оценим полулогарифмическую модель с помощью команды:
```
reg lnEARNINGS S ASVABC
```
и сохраним предсказанные значения зависимой переменной с помощью команды:
```
predict semiln_y_hat
```
Теперь переходим к шагу 2, оценим дополнительные модели.
Сначала создадим дополнительную разность для линейной модели:
```
gen lin_adds= ln(y_hat)- semiln_y_hat
```
и оценим эту модель:
```
reg EARNINGS S ASVABC lin_adds
```
Создадим также дополнительную разность для полулогарифмической модели:
```
gen semilog_add=y_hat-exp(ln_y_hat)
```
оценим эту модель:
```
reg lnEARNINGS S ASVABC semilog_add
```
Если

1) оба коэффициента при дополнительных разностях значимы или оба незначимы, то выбрать посредством теста Дэвидсона, Уайта и МакКиннона невозможно,
2) если незначим только коэффициент при дополнительной разности в линейной модели, то лучше линейная модель,
3) если незначим только коэффициент при дополнительной разности в полулогарифмической модели, то лучше полулогарифмическая модель.

б) Тест Бера и МакАлера
Необходимые команды
Выбор между линейной и полулогарифмической моделями
Сначала оценим полулогарифмическую модель с помощью команды:
```
reg lnEARNINGS S ASVABC
```
и сохраним предсказанные значения зависимой переменной с помощью команды:
```
predict semiln_y_hat
```

Аналогично оценим линейную модель с помощью команды:

```
reg EARNINGS S ASVABC
```

Сохранить предсказанные значения зависимой переменной можно с помощью команды:

```
predict y_hat
```

Теперь переходим к шагу 2, оценим дополнительные модели.

```
1) gen y1= exp(semiln_y_hat)
reg y1 S ASVABC
```

Сохранить остатки регрессии можно с помощью команды

```
predict res1, resid
2) gen y2= ln(y_hat)
reg y2 S ASVABC
```

Сохранить остатки регрессии можно с помощью команды

```
predict res2, resid
```

Теперь переходим к шагу 3, оценив еще 2 дополнительные модели.

```
reg lnEARNINGS S ASVABC res1
reg EARNINGS S ASVABC res2
```

Если

1) оба коэффициента при res1 и res2 незначимы или оба значимы, то выбрать посредством теста Бера и МакАлера невозможно,
2) если незначим только коэффициент при res1, то лучше полулогарифмическая модель,
3) если незначим только коэффициент при res2, то лучше линейная модель.

III. По данным файла dougherty.dta выберите между линейной и полулогарифмической моделями с помощью теста Зарембки.

Необходимые команды

Выбор между линейной и полулогарифмической моделями

```
reg EARNINGS S ASVABC
gen lnEARNINGS = ln(EARNINGS)
reg lnEARNINGS S ASVABC

means EARNINGS
```

Variable	Type	Obs	Mean	[95% Conf.	Interval]
EARNINGS	Arithmetic	540	19.71924	18.48493	20.95355
	Geometric	540	16.3442	15.54379	17.18584
	Harmonic	540	13.77391	13.05586	14.57555

```
gen EARNINGSstar= EARNINGS/16.3442
gen lnEARNINGSstar = ln(EARNINGSstar)
reg EARNINGSstar S ASVABC
сохраните RSS с помощью команды scalar rss3=e(rss)
reg lnEARNINGSstar S ASVABC
сохраните RSS с помощью команды scalar rss4=e(rss)
```

Используя RSS из оцененных регрессий, следует рассчитать тестовую F – статистику

*scalar xi2 = 0.5*540*abs(ln(rss4/rss3))*

display xi2

и p-value для этой статистики:

display chi2tail(1, xi2)

Если рассчитанное p-value не превышает выбранного уровня значимости, то основная гипотеза отвергается, есть разница между исходными линейной и полулогарифмическими моделями.

1. Выбор включаемых в модель факторов. Тест Рамсея

Материалы из учебника О.Демидовой и Д.Малахова «Эконометрика. Учебник и практикум»

Задача 9.1. (К.Доугерти, Введение в эконометрику, изд.3, стр. 216, задача № 6.7).

Исследователь считает, что уровень активности в теневой экономике Y зависит либо положительно от налогового бремени X, либо отрицательно от уровня государственных расходов на предотвращение теневой экономической деятельности Z. Переменная Y может также зависеть от обеих переменных X и Z. Получены международные данные двух перекрестных выборок по Y, X и Z (в млн долл. США): для группы из 30 индустриально развитых и для группы из 30 развивающихся стран. Исследователь оценивает регрессионные зависимости: 1) log Y от log Z; 2) log Y только от log X; 3) log Y только от log Z одновременно для каждой выборки, получая следующие результаты (в скобках приведены стандартные отклонения):

	Индустриально развитые страны			Развивающиеся страны		
	(1)	(2)	(3)	(4)	(5)	(6)
log X	0.699	0.201	-	0.806	0.727	-
s.e.	(0.154)	(0.112)		(0.137)	(0.090)	
log Z	-0.646	-	-0.053	-0.091	-	0.427
s.e.	(0.162)		(0.124)	(0.117)		(0.116)
Константа	-1.137	-1.065	1.230	-1.122	-1.024	2.824
s.e.	(0.863)	(1.069)	(0.896)	(0.873)	(0.858)	(0.835)
R^2	0.44	0.10	0.01	0.71	0.70	0.33

Переменная X положительно коррелированна с Z в обеих выборках. Выполнив соответствующие статистические тесты, напишите краткий обзор, дав рекомендации исследователю относительно интерпретации полученных результатов. Выбрать, какая из моделей лучше

А) Для индустриально развитых стран, Б) Для развивающихся стран. Объяснить изменения в оценках коэффициентов и их стандартных отклонений в других моделях.

Задача 9.2.

1) По 150 наблюдениям оценили зависимость почасовой заработной платы от пола (переменная MALE равно 1 для мужчин и 0 для женщин), длительности обучения S и возраста AGE.

$$\hat{Y} = \underset{(3.09)}{3.6} + \underset{(1.21)}{3.5}\,MALE + \underset{(0.53)}{3.24}\,S + \underset{(0.057)}{0.44}\,AGE,\ RSS = 7632$$

Используя результаты двух вспомогательных регрессий, приведенных ниже, проведите RESET – тест и ответьте, правильная ли спецификация модели выбрана.

$$\hat{\hat{Y}} = \underset{(4.09)}{12.37} - \underset{(1.7)}{0.29}\,MALE - \underset{(1.3)}{0.49}\,S - \underset{(0.17)}{0.08}\,AGE + \underset{(0.002)}{0.0064}\,\hat{Y}^2,\ RSS = 7154$$

$$\hat{\hat{Y}} = \underset{(4.42)}{-18.1} + \underset{(2.44)}{9.2}\,MALE + \underset{(2.05)}{7.93}\,S + \underset{(0.28)}{1.1}\,AGE - \underset{(0.004)}{0.012}\,\hat{Y}^2 + \underset{(3.45\cdot10^{-11})}{1.75\cdot10^{-10}}\,\hat{Y}^3,\ RSS = 6069$$

Задание 9.5.

По данным для 23 демократических стран оценили зависимость индекса Джини (меры неравенства , 0 – полное равенство, по мере роста этого показателя степень неравенства увеличивается) от ВНР на душу населения с учетом ППС (паритета покупательной способности) и провели тест Рамсея. Результаты оценивания указаны в таблице. Прокомментируйте результаты теста Рамсея.

```
reg gini gdp if democ==1

      Source |       SS       df       MS              Number of obs =      23
-------------+------------------------------           F(  1,    21) =   13.05
       Model |  506.853501     1  506.853501           Prob > F      =  0.0016
    Residual |  815.572523    21  38.8367868           R-squared     =  0.3833
-------------+------------------------------           Adj R-squared =  0.3539
       Total |  1322.42602    22  60.1102738           Root MSE      =  6.2319

------------------------------------------------------------------------------
        gini |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gdp |  -.0006307   .0001746    -3.61   0.002    -.0009937   -.0002676
       _cons |   44.30983   3.572733    12.40   0.000     36.87993    51.73974
------------------------------------------------------------------------------

 ovtest

Ramsey RESET test using powers of the fitted values of gini
       Ho:  model has no omitted variables
                  F(3, 18) =      5.16
                  Prob > F =      0.0095
```

Упражнение 9.5.

В файле rlms14.dta приведены данные о заработной плате, возрасте, образовании, профессиональной принадлежности россиян из базы данных RLMS, 14 раунд опроса, проводившийся в 2005 г. (описание переменных дано в Приложении 1).

1) Оцените зависимость заработной платы от возраста, продолжительности рабочей недели, длительность работы на последнем месте. Выберите в качестве зависимой

переменной income или wage, а в качестве независимых – возраст (переменную необходимо предварительно рассчитать по формуле age = 2005 – birth_year), длительность работы на последнем месте (переменную необходимо предварительно рассчитать по формуле tenure=2005 – beginning), и т.д.
2) Адекватна ли оцененная регрессия? Все ли оценки коэффициентов имеют ожидаемый знак? Правильно ли выбрана спецификация модели, исходя из полученных результатов?
3) Проведите тест Рамсея. Если гипотеза H_0 о правильной спецификации модели будет отвергнута, переходите к следующему пункту.
4) Оцените регрессию снова, добавив в уравнение регрессии квадрат одной из независимых переменных, например, возраста. Если новая оцененная регрессия является адекватной, дайте экономическую интерпретацию полученным результатам.
5) Поэкспериментируйте с включением в модель других переменных.

Методические рекомендации

1) Создайте новые переменные age и tenure с помощью команд:

```
gen age = 2005 - birth_year,
gen tenure=2005 - beginning
```

2) Оцените коэффициенты уравнения регрессии $income = \beta_0 + \beta_1 age + \beta_2 tenure + u$, набрав в командном окне

```
reg income age tenure
```

3) Проведите тест Рамсея с помощью команды

```
ovtest
```

6. Оценено уравнение регрессии (в скобках указаны значения t – статистик)

$$\hat{Y} = 1 + \underset{(0.5)}{2.1} X_1 + \underset{(1.5)}{5.7} X_2 - \underset{(-1.7)}{7.5} X_3 + \underset{(2.1)}{3} X_4 - \underset{(-3.5)}{6.2} X_5$$

При удалении каких переменных качество подгонки регрессии может увеличиться?

Борзых Д.А., Демешев Б.Б., Эконометрика в задачах и упражнениях, Издание 2, URSS, 2017

9.3 По 30 наблюдениям при помощи метода наименьших квадратов оценена модель $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$, для которой $RSS = 150$. При помощи вспомогательной регрессии $\hat{\hat{y}} = \hat{\gamma}_1 + \hat{\gamma}_2 x + \hat{\gamma}_3 z + \hat{\gamma}_4 \hat{y}^2 + \hat{\gamma}_5 \hat{y}^3$, для которой $RSS = 120$, выполните тест Рамсея на уровне значимости 5%.

9.4 По 35 наблюдениям при помощи метода наименьших квадратов оценена модель $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$, для которой $R^2 = 0.7$. При помощи вспомогательной регрессии $\hat{\hat{y}} = \hat{\gamma}_1 + \hat{\gamma}_2 x + \hat{\gamma}_3 z + \hat{\gamma}_4 \hat{y}^2 + \hat{\gamma}_5 \hat{y}^3$, для которой $R^2 = 0.8$, выполните тест Рамсея на уровне значимости 5%.

9.5 Используя 80 наблюдений, исследователь оценил две конкурирующие модели: $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$, в которой $RSS_1 = 36875$ и $\widehat{\ln y} = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$, в которой $RSS_2 = 122$.

Выполнив преобразование $y_i^* = y_i / \sqrt[n]{\prod y_i}$, исследователь также оценил две вспомогательные регрессии: $\hat{y}^* = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$, в которой $RSS_1^* = 239$ и $\widehat{\ln y^*} = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$, в которой $RSS_2^* = 121$.

Завершите тест Бокса-Кокса на уровне значимости 5%.