

Эконометрика, 2019-2020, 2 модуль
Семинары 2-3
11.11.19, 18.11.19

для
Группы Э_Б2017_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, `dumtmy`, 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 линейной 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

$$scalar \chi^2 = 0.5 * 540 * \text{abs}(\ln(rss4/rss3))$$

```
display \chi^2
```

и p-value для этой статистики:

$$display \chi^2_{tail}(1, \chi^2)$$

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

Дамми (фиктивные) переменные и тест Чоу

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

Задача 7.1. Оцененная зависимость почасовой оплаты труда индивида Y (измеряется в долларах в час) от результатов выпускного теста X (измеряется в баллах) и пола (D – фиктивная переменная, равная 1 для мужчин и 0 для женщин) имеет вид:

$$\hat{Y} = 2 + 3.7X + 2.4D.$$

Все коэффициенты являются значимыми при уровне значимости 1%.

При одинаковых результатах теста почасовая оплата мужчин выше почасовой оплаты женщин на

- 1) 0.024 \$ 2) 2.4 \$ 3) 0.024 % 4) 2.4%

Задача 7.2.

Оцененная зависимость почасовой оплаты труда американцев Y (измеряется в долларах) от стажа их работы X (измеряется в годах); пола, описываемого с помощью фиктивной переменной D_1 , равной 1 для мужчин и 0 для женщин; расовой принадлежности, описываемой с помощью фиктивной переменной D_2 , равной 1 для светлокожих и 0 для темнокожих американцев, имеет вид:

$$\hat{Y} = 4 + 0.8X + 0.04D_1 - 0.01D_2$$

Все коэффициенты являются значимыми при уровне значимости 1%.

Чему равна почасовая оплата труда темнокожих американцев при пятилетнем стаже работы?

Задача 7.3.

Зависимость расходов на продукты питания от располагаемого дохода X имеет вид:

$$\hat{Y} = 2 + 0.6X + 0.07D_1X,$$

где D_1 – фиктивная переменная, равная 1 для городских и 0 для сельских жителей.

а) Коэффициент наклона в линейной зависимости для сельских жителей равен

- 1) 0,67 2) 0,6 3) 0,53 4) 2

б) Если вместо D_1 использовать переменную D_2 , равную 0 для городских и 1 для сельских жителей, то зависимость примет вид:

- 1) $\hat{Y} = 2 + 0.67X - 0.07D_2X$
2) $\hat{Y} = 2 + 0.67X + 0.07D_2X$
3) $\hat{Y} = 2 + 0.6X - 0.07D_2X$
4) $\hat{Y} = 2.07 + 0.6X - 0.07D_2X.$

Оценена зависимость расходов потребителей на газ и электричество Y в США в 1977 – 1999 г.г. в постоянных ценах I квартала 1977г. от времени ($t = 1$ для 1977 г., $t = 2$ для

1978 г. и т.д.) с учетом сезонных факторов ($D_i = 1$, если наблюдение относится к i -му кварталу и 0 иначе, $i = 1, \dots, 4$):

$$\hat{Y} = 8 + 0.1t - 3D_2 - 2.6D_3 - 2D_4$$

Если в качестве выделенной категории будет выбран не первый квартал, а второй, то уравнение регрессии примет вид

$$1) \hat{Y} = 5 + 0.1t + 3D_1 + 0.4D_3 + D_4$$

$$2) \hat{Y} = 8 + 0.1t - 3D_1 - 2.6D_3 - 2D_4$$

$$3) \hat{Y} = 5 + 0.1t - 3D_1 - 2.6D_3 - 2D_4$$

$$4) \hat{Y} = 5 + 0.1t - 3D_2 - 0.4D_3 - D_4$$

Задача 7.5.

По данным для 570 индивидуумов оценили зависимость длительности обучения индивидуума S от способностей индивидуума, описываемых обобщенной переменной $ASVABC$ и пола индивидуума, описываемого с помощью фиктивной переменной $MALE$ (равной 1 для мужчин и 0 для женщин) с помощью двух регрессий:

$$\hat{S} = 6.12 + 0.147 \cdot ASVAB, \text{ } RSS_1 = 2099,9$$

(0.44) (0.0088)

$$\hat{S} = 6.72 + 0.137 \cdot ASVAB - 1.035 \cdot MALE + 0.0166 \cdot (MALE \cdot ASVABC), \text{ } RSS_2 = 2090,98$$

(0.73) (0.014) (0.933) (0.018)

Зависит ли длительность обучения от пола индивидуума и почему?

Задача 7.6.

По квартальным данным 1960-1976 г.г. была оценена модель с тремя объясняющими факторами:

$$\hat{Y} = 1.03 + 0.1X_1 - 4.45X_2 + 0.26X_3, \text{ } ESS = 103.5, \text{ } RSS = 17.48.$$

При добавлении в модель трех сезонных dummy – переменных значение ESS увеличилось до 107.3.

Проверить гипотезу о наличии сезонности.

Задача 7.7.

По данным для 570 индивидуумов оценили зависимость почасовой заработной платы $EARN$ от длительности обучения S и от способностей индивидуума, описываемых обобщенной переменной $ASVABC$:

- по общей выборке

$$EARN = -9.96 + 0.93S + 0.21ASVABC \quad RSS_1 = 32189.36$$

(2.02) (0.16) (0.04)

- а также отдельно для мужчин

$$EARN = -7.23 + 1.01S + 0.35ASVABC \quad RSS_2 = 15223.7$$

(2.63) (0.27) (0.06)

- и женщин

$$EARN = -11.4 + 0.81S + 0.14ASVABC \quad RSS_3 = 10231.24$$

(3.24) (0.19) (0.03)

Можно ли считать, что эта зависимость одинакова для мужчин и женщин?

Упражнение 7.2.

В статистическом пакете Stata 12, по данным файла flats.dta, используя переменные

price_metr, livesp, kitsp, dist, metrdist, floors, walk, (где price_metr - стоимость квадратного метра однокомнатной квартиры, описание остальных переменных дано в приложении, определите, одинакова ли зависимость для двух групп квартир (для которых время пути от метро дано в минутах пешком и для которых время пути от метро дано в минутах езды на транспорте).

- 1) С помощью оценки регрессии вида (7.4),
- 2) С помощью теста Чоу.

Используйте 5% процентный уровень значимости.

Рекомендации.

- 1) Для оценки регрессии с варьирующимися коэффициентами наклона для двух групп переменных создадим новые переменные, которые являются перемножением (cross-terms) переменных livesp, kitsp, dist, metrdist, floors и переменной walk (объясните, для чего это нужно):
- 2)

```
. gen livesp_walk= livesp* walk  
. gen kitsp_walk= kitsp* walk  
. gen dist_walk= dist* walk  
. gen metrdist_walk= metrdist* walk  
. gen floors_walk= floors* walk
```

Оцените новую регрессию с включенными cross-terms переменными:

```
. reg price_meter livesp kitsp dist metrdist floors walk livesp_walk kitsp_walk  
dist_walk metrdist_walk floors_walk,
```

Проверим одновременную значимость всех коэффициентов переменных, содержащих walk, воспользовавшись командой test:

```
test livesp_walk= kitsp_walk= dist_walk= metrdist_walk= floors_walk= walk=0,
```

- 3) Проведем тест Чоу, оцениваем одну и ту же форму модели а) для всех квартир, б) для квартир, до которых время пути от метро дано в минутах пешим шагом, и в) для квартир, для которых время в пути дано в минутах езды на автомобиле.

Модель по данным для всех квартир можно оценить с помощью команды:

```
reg price_meter livesp kitsp dist metrdist floors,
```

Сохраним RSS с помощью команды scalar rssp=e(rss).

Модель для квартир, до которых время пути от метро дано в минутах пешим шагом можно оценить с помощью команды:

```
. reg price_meter livesp kitsp dist metrdist floors if walk==1,
```

Сохраним RSS с помощью команды scalar rss1=e(rss).

Модель для квартир, до которых время пути от метро дано в минутах езды можно оценить с помощью команды:

```
. reg price_meter livesp kitsp dist metrdist floors if walk==0,
```

Сохраним RSS с помощью команды `scalar rss2=e(rss)`.

Используя RSS из оцененных регрессий, рассчитаем тестовую F – статистику:

```
. scalar F=((rssp-rss1-rss2)/6)/((rss1+rss2)/(773-2*6))
```

```
. display F
```

```
20.060314
```

Для нахождения p-value для F-статистики используйте команду

```
di Ftail(6, 761, 20.060314)
```

Упражнение 7.3.

Используя статистический пакет Stata, по данным файла `nlsw88.dta` (эта база данных встроена в статистический пакет Stata, сделать ее активной можно выбрав `File->Example Datasets...-> Example Datasets Installed in Stata`, описание можно найти, нажав на `describe` (также описание переменных дано в Приложении 1),

1) Оцените модель $wage_i = \beta_0 + \beta_1 \cdot hours_i + \beta_2 \cdot ttl_exp_i + \beta_3 \cdot tenure_i + \beta_4 \cdot union_i + u_i, i = 1, \dots, n$.

Проинтерпретируйте значение оценки коэффициента перед переменной `union`.

2) Проанализируйте, нужно ли оценивать модель

$$wage_i = \beta_0 + \beta_1 \cdot hours_i + \beta_2 \cdot ttl_exp_i + \beta_3 \cdot tenure_i + u_i$$

отдельно для тех женщин, которые состоят в союзе и для тех, которые не состоят.

Решение.

1) Заметим, что в базе данных значение `union` переменной `union` соответствует 1, значение `nonunion` соответствует 0, пропуски в значениях переменных обозначаются как “.” (соответствующие наблюдения выкидываются).

Оценим искомую модель с помощью команды:

```
reg wage hours ttl_exp tenure union,
```

получим:

Source	SS	df	MS	Number of obs =	1867
-----+-----				F(4, 1862) =	84.07
Model	4961.07427	4	1240.26857	Prob > F =	0.0000
Residual	27470.4872	1862	14.7532155	R-squared =	0.1530
-----+-----				Adj R-squared =	0.1512
Total	32431.5615	1866	17.380258	Root MSE =	3.841

wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----					
hours	.0191089	.0092112	2.07	0.038	.0010436 .0371743
ttl_exp	.2785683	.0243063	11.46	0.000	.2308978 .3262389
tenure	.0468629	.0196121	2.39	0.017	.0083988 .0853269

union	1.190311	.2085296	5.71	0.000	.7813344	1.599287
_cons	2.691105	.392896	6.85	0.000	1.920542	3.461667

Из результатов оценки модели, можно заметить, что если респондент состоит в профсоюзе, то при прочих равных его почасовая зарплата выше на 1.190311 долл. (коэффициент при переменной union значим на любом адекватном уровне значимости).

2) Теперь проанализируем, нужно ли оценивать вышеуказанную модель отдельно для каждой подвыборки.

Для этого создадим переменные, которые являются перемножением регрессоров и дамми-переменной:

```
. gen hours_union=hours*union
(369 missing values generated)

. gen ttl_exp_union=ttl_exp*union
(368 missing values generated)

. gen tenure_union=tenure*union
(378 missing values generated)
```

Оценим регрессию, включив в нее вновь созданные переменные:

```
reg wage hours ttl_exp tenure union hours_union ttl_exp_union tenure_union,
получим:
```

Source	SS	df	MS	Number of obs =	1867
-----+-----				F(7, 1859) =	50.31
Model	5165.36882	7	737.909831	Prob > F =	0.0000
Residual	27266.1926	1859	14.6671289	R-squared =	0.1593
-----+-----				Adj R-squared =	0.1561
Total	32431.5615	1866	17.380258	Root MSE =	3.8298

wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----					
hours	.0356629	.0103232	3.45	0.001	.0154165 .0559093
ttl_exp	.282476	.0273754	10.32	0.000	.2287863 .3361656
tenure	.0449328	.0228423	1.97	0.049	.0001335 .0897322
union	4.483446	.9768301	4.59	0.000	2.567647 6.399245
hours_union	-.080646	.0226316	-3.56	0.000	-.125032 -.0362599
ttl_exp_union	-.0235813	.0588922	-0.40	0.689	-.139083 .0919204
tenure_union	.0148574	.0446927	0.33	0.740	-.0727958 .1025106
_cons	2.036556	.4369703	4.66	0.000	1.179552 2.89356

Проверим совместную значимость переменных, содержащих union с помощью команды:

```
( 1) union - hours_union = 0
( 2) union - ttl_exp_union = 0
( 3) union - tenure_union = 0
( 4) union = 0
```

Так как p-value соответствующей тестовой статистики равно 0, то оценки всех коэффициентов совместно отличны от нуля и необходимо оценивать модели для двух подвыборок отдельно, что мы и сделаем ниже.

```
. reg wage hours tenure ttl exp if union==1
```

Source	SS	df	MS	Number of obs = 460		
-----+-----				F(3, 456) = 19.63		
Model	914.385516	3	304.795172	Prob > F = 0.0000		
Residual	7079.67078	456	15.5255938	R-squared = 0.1144		
-----+-----				Adj R-squared = 0.1086		
Total	7994.05629	459	17.4162447	Root MSE = 3.9403		

wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----+-----						
hours	-.0449831	.020721	-2.17	0.030	-.0857037	-.0042625
tenure	.0597902	.0395226	1.51	0.131	-.0178788	.1374593
t1l_exp	.2588947	.0536471	4.83	0.000	.1534685	.3643209
cons	6.520002	.8988477	7.25	0.000	4.753605	8.286399

```
. reg wage hours ttl exp tenure if union==0
```

9

tenure	.0449328	.022624	1.99	0.047	.0005522	.0893134
_cons	2.036556	.432794	4.71	0.000	1.187563	2.885549

Стоит сделать важное замечание. При проведении теста Чоу необходимо сначала определить оптимальную модель для всей выборки, а затем уже проводить сам тест.

Исходя из простого сопоставления результатов оценки этих моделей, можно заключить, что результаты сильно отличаются для двух подвыборок, что свидетельствует о корректности проведенного теста.

Задание 7.4.

В приведенных ниже таблицах содержатся результаты оценивания функции спроса на молоко (в таблице 1 по всем наблюдениям, в таблице 2 – по наблюдениям для сельской местности, в таблице 3 – для городской местности).

Переменные:

buymilk – стоимость молока, купленного семьей за последние 7 дней (в руб.),

income – доход семьи за месяц,

pr_milk - цена 1 л молока (в руб.),

status – тип населенного пункта (1 – областной центр, 2 – город, 3 – поселок городского типа, 4 – село),

Таблица 1.

reg buymilk_c inc pr_milk

Source	SS	df	MS	Number of obs = 2127		
Model	7855703.78	2	3927851.89	F(2, 785)	= 943.58	
Residual	8841601.29	2124	4162.71247	Prob > F	= 0.0000	
Total	16697305.1	2126	7853.85939	R-squared	= 0.4705	
				Adj R-squared	= 0.4700	
				Root MSE	= 64.519	

buymilk_c	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
inc	.0002428	.0000762	3.19	0.001	.0000934	.0003922
pr_milk	.8768133	.02023	43.34	0.000	.837140	.9164859
_cons	32.96319	1.746953	18.87	0.000	29.53727	36.38911

Таблица 2.

reg buymilk_c inc pr_milk if status==4

Source	SS	df	MS	Number of obs = 348		
Model	3184511.16	2	1592255.58	F(2, 785)	= 319.33	
Residual	1720236.56	345	4986.19293	Prob > F	= 0.0000	
Total	4904747.72	347	14134.7197	R-squared	= 0.6493	
				Adj R-squared	= 0.6472	
				Root MSE	= 70.613	

buymilk_c	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
inc	.0002418	.0002566	0.94	0.347	-.0002629	.0007465
pr_milk	.9387025	.0371628	25.26	0.000	.8656084	1.011797
_cons	32.57962	4.539265	7.18	0.000	23.65151	41.50774

Таблица 3.

```
reg buymilk_c inc pr_milk if status==1 status==2 status==3
```

Source	SS	df	MS	Number of obs = 1779
Model	4688916.24	2	2344458.12	F(2, 785) = 586.49
Residual	7099423.62	1776	3997.42321	Prob > F = 0.0000
Total	11788339.9	1778	6630.11241	R-squared = 0.3978
				Adj R-squared = 0.3971
				Root MSE = 63.225

buymilk_c	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
inc	.0002451	.0000793	3.09	0.002	.0000896	.0004006
pr_milk	.8425161	.0246925	34.12	0.000	.7940866	.8909456
_cons	33.35554	1.894483	17.61	0.000	29.63989	37.07119

Можно ли считать зависимость спроса на молоко от его цены и дохода единой для городской и сельской местности? Ответ обоснуйте подходящим тестом.