Методические материалы к семинару 4, 3 модуль, 2018

Проверка нормальности распределения остатков регрессии

По данным файла clothing1.dta, содержащем данные о продажах одежды в 400 голландских магазинах мужской одежды

1) Оцените коэффициенты уравнения регрессии

$sales = \beta_0 + \beta_1 hoursw + \beta_2 size + u$.

2) Постройте гистограмму распределения остатков регрессии и qqplot.
С помощью тестов Колмогорова-Смирнова, Харке-Бера, Шапиро-Уилка проверьте гипотезу о нормальности распределения случайных ошибок.
3) Рассчитайте Стьюдентизированные остатки регрессии, постройте их график и с его помощью определите возможные наблюдения, которые являются выбросами. Если такие наблюдения будут выявлены, то оцените регрессию
а) без наблюдений, являющихся выбросами,
б) R-регрессию Хуберта.
4) Сравните результаты оценки регрессий, оцененных с помощью МНК (с выбросами и без), с оценками Уайта для стандартных ошибок, R-регрессии Хуберта.

Методические рекомендации

1) Открыв файл clothing.dta в статистическом пакете STATA, оцените необходимую регрессию с помощью команды:

```
reg sales hoursw ssize
```

```
. reg sales hoursw ssize
```

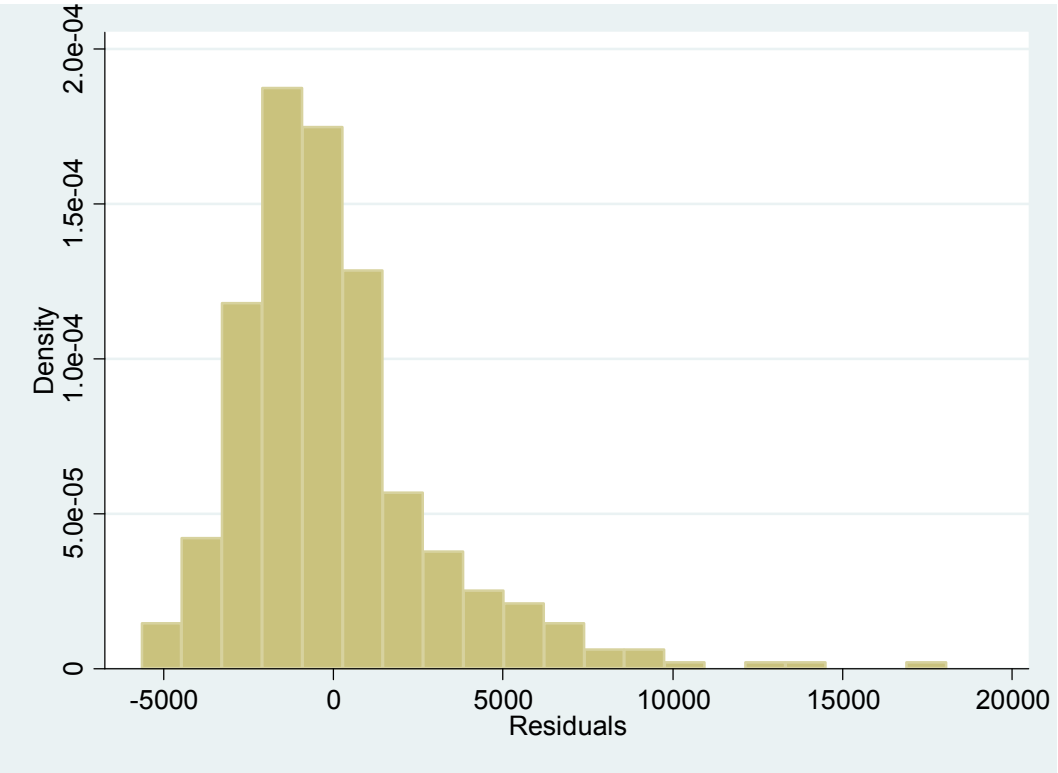
Source	SS	df	MS		
				Number of obs	= 400
				F(2, 397)	= 114.49
Model	2.0409e+09	2	1.0204e+09	Prob > F	= 0.0000
Residual	3.5382e+09	397	8912441.27	R-squared	= 0.3658
				Adj R-squared	= 0.3626
Total	5.5791e+09	399	13982691	Root MSE	= 2985.4

sales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hoursw	37.52842	2.83722	13.23	0.000	31.95056	43.10627
ssize	-22.14457	1.625067	-13.63	0.000	-25.33939	-18.94976
_cons	5133.59	321.6934	15.96	0.000	4501.155	5766.026

2) Сохраните остатки регрессии с помощью команды

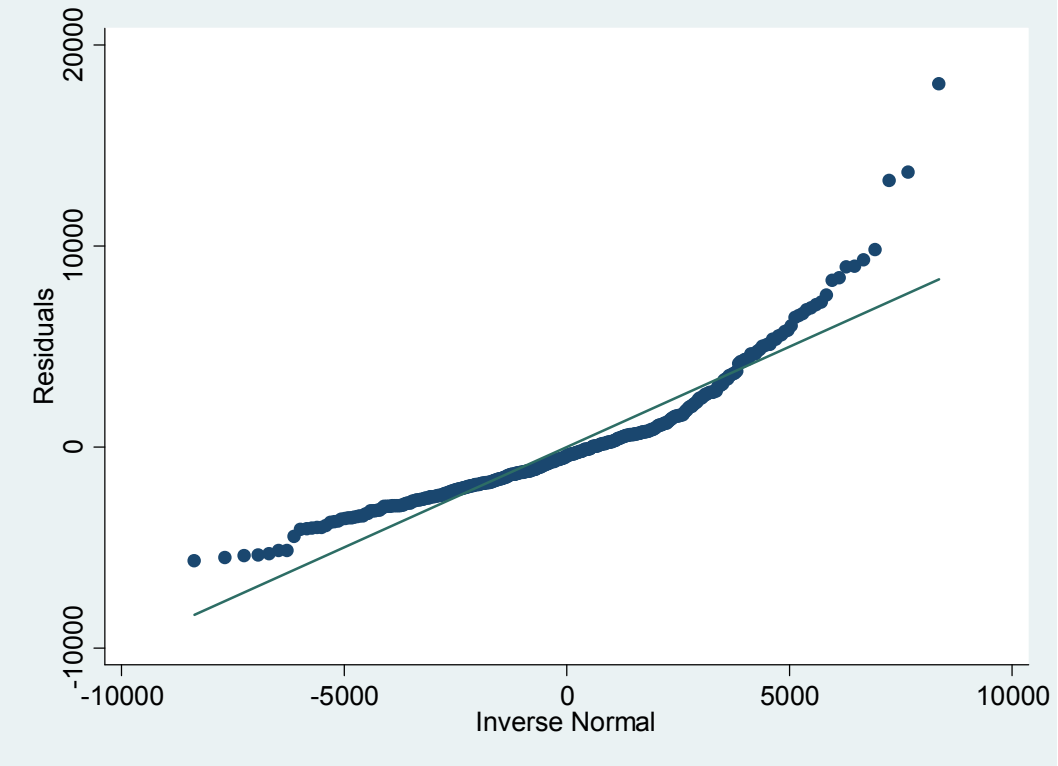
```
predict res, resid
```

3) Постройте их гистограмму с помощью команды `hist res`



Qq plot с помощью команды

```
qnorm res
```



4) Проведите тест Колмогорова-Смирнова с помощью команды

```
sum res
ksmirnov res = normal((res-r(mean))/r(sd))
```

```
. sum res
```

Variable	Obs	Mean	Std. Dev.	Min	Max
res	400	1.62e-07	2977.88	-5658.48	18076.32

```
. ksmirnov res = normal((res-1.62e-07)/2977.88)

One-sample Kolmogorov-Smirnov test against theoretical distribution
           normal((res-1.62e-07)/2977.88)
```

Smaller group	D	P-value	Corrected
res:	0.1314	0.000	
Cumulative:	-0.0764	0.009	
Combined K-S:	0.1314	0.000	0.000

```
Note: ties exist in dataset;
      there are 399 unique values out of 400 observations.
```

5) Проведите тест Харке-Бера с помощью команды `sktest res, noadjust`

```
. sktest res, noadjust
```

Skewness/Kurtosis tests for Normality

				joint	
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	chi2(2)	Prob>chi2
res	400	0.0000	0.0000	156.99	0.0000

6) Проведите тест Шапиро-Уилка с помощью команды `swilk res`

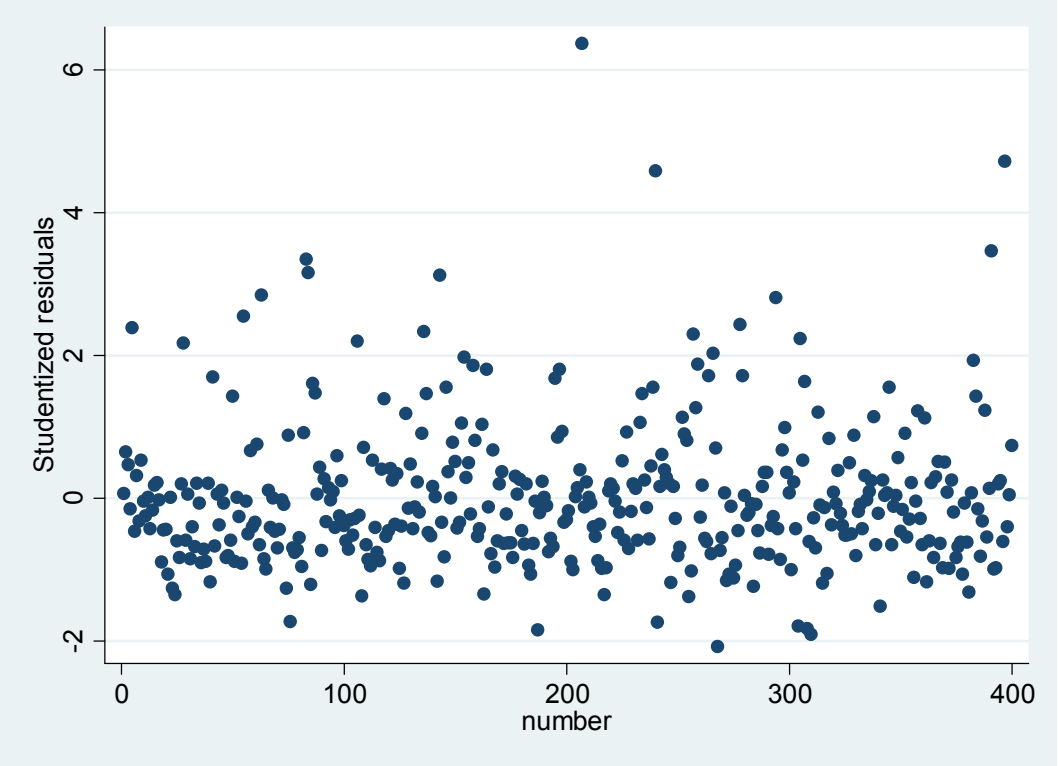
```
. swilk res
```

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
res	400	0.88695	31.121	8.180	0.00000

7) Постройте график стьюдентизированных остатков регрессии с помощью команды

```
predict residst, rstudent
twoway(scatter residst number, sort)
list number if abs(residst) > 2
```




```
. list number if abs(residst) > 2
```

	number
5.	5
28.	28
55.	55
63.	63
83.	83
84.	84
106.	106
136.	136
143.	143
207.	207
240.	240
257.	257
266.	266
268.	268
278.	278
294.	294
305.	305
391.	391
397.	397

Оценить регрессию без наблюдений - выбросов можно с помощью команды

```
reg sales hoursw ssize if abs(residst) < 2
```

```
. reg sales hoursw ssize if abs(residst) < 2
```

Source	SS	df	MS		
Model	1.2479e+09	2	623945884	Number of obs =	381
Residual	1.8479e+09	378	4888502.86	F(2, 378) =	127.64
				Prob > F =	0.0000
Total	3.0957e+09	380	8146699.61	R-squared =	0.4031
				Adj R-squared =	0.3999
				Root MSE =	2211

sales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hoursw	34.84753	2.4932	13.98	0.000	29.94525	39.74981
ssize	-21.73711	1.465495	-14.83	0.000	-24.61865	-18.85556
_cons	4985.046	258.1984	19.31	0.000	4477.361	5492.732

8) Оценить R-регрессию Хуберта можно с помощью команды

```
rreg sales hoursw ssize
```

```
. rreg sales hoursw ssize

   Huber iteration 1:  maximum difference in weights = .83822297
   Huber iteration 2:  maximum difference in weights = .18323368
   Huber iteration 3:  maximum difference in weights = .08958103
   Huber iteration 4:  maximum difference in weights = .04554895
Biweight iteration 5:  maximum difference in weights = .29064309
Biweight iteration 6:  maximum difference in weights = .12010312
Biweight iteration 7:  maximum difference in weights = .03363972
Biweight iteration 8:  maximum difference in weights = .01166615
Biweight iteration 9:  maximum difference in weights = .0041783

Robust regression                                     Number of obs =     399
                                                      F(  2,   396) =  130.57
                                                      Prob > F      =  0.0000
```

sales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hoursw	31.44067	2.178818	14.43	0.000	27.15717	35.72416
ssize	-19.50244	1.386781	-14.06	0.000	-22.22882	-16.77607
_cons	4899.219	253.5963	19.32	0.000	4400.656	5397.782

9) Сохранить результаты оценки регрессии можно с помощью команды

```
est store reg1 (и т.д.)
```

Для сравнения результатов удобно сформировать общую таблицу с помощью команды

```
est tab reg1 (и т.д.), star (0.1 0.05 0.01) b(%7.3f)
```

```
. qui reg sales hoursw ssize

. est store reg1

. qui reg sales hoursw ssize if abs(residst) < 2

. est store reg2

. qui reg sales hoursw ssize, robust

. est store reg3

. qui rreg sales hoursw ssize

. est store reg4

. est tab reg1 reg2 reg3 reg4, star (0.1 0.05 0.01) b(%7.3f)
```

Variable	reg1	reg2	reg3	reg4
hoursw	37.528***	34.848***	37.528***	31.441***
ssize	-22.145***	-21.737***	-22.145***	-19.502***
_cons	5133.590***	4985.046***	5133.590***	4899.219***

legend: * p<.1; ** p<.05; *** p<.01