Домашнее задание 2
Часть 3
Пряничникова Алена, Дубровский Михаил
БЭК1811
2017 год

Введение

Закончим наше приключение со второй домашней работой. В данной части нам предстоит проверить правильность спецификации модели на наличие квадратичной или кубической зависимости от регрессоров. В конце будет проведена проверка на мультиколлинеарность и сделаны соответствующие выводы.

Тест Рамсея

Нами была оценена следующая линейная модель:

$$Unemployment = \beta_0 + \beta_1 * GDPpercap + \beta_2 * Urbanshare + \beta_3 * Higheduc$$

Затем мы провели тест Рамсея на правильную спецификацию модели:

H_0: модель правильно специфицирована

H_1: наличие ошибок спецификации модели

```
. reg Unemployment GDPpercap Urbanshare Higheduc
```

Source	SS	df	MS	Number of obs	=	85
				F(3, 81)	=	8.91
Model	267.746225	3	89.2487416	Prob > F	=	0.0000
Residual	811.219657	81	10.0150575	R-squared	=	0.2482
				Adj R-squared	=	0.2203
Total	1078.96588	84	12.8448319	Root MSE	=	3.1647

Unemployment	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDPpercap	-4.79e-08	5.81e-07	-0.08	0.935	-1.20e-06	1.11e-06
Urbanshare	-.1307659	.0273373	-4.78	0.000	-.1851586	-.0763732
Higheduc	-.0568294	.0712058	-0.80	0.427	-.1985066	.0848478
_cons	17.51559	2.821607	6.21	0.000	11.90148	23.12971

```
. ovtest

Ramsey RESET test using powers of the fitted values of Unemployment
       Ho:  model has no omitted variables
                  F(3, 78) =      3.18
                  Prob > F =      0.0287

. gen GDPpercap2 = (GDPpercap)^2

. gen Higheduc2 = (Higheduc)^2
```

Рис 1. Линейная регрессия и тест Рамсея

Для теста Рамсея $p_{value} = 0.0287$. Таким образом, на уровне значимости 5% мы отвергаем нулевую гипотезу и признаем, что в модели допущены ошибки спецификации и необходимо рассмотреть добавление квадратов или кубов переменных (можно степени и выше, но не нужно).

Вспоминая диаграммы рассеяния переменных из второй части ДЗ, заметим, что для переменных Urbanshare и Higheduc зависимость более-менее похожа на линейную. Соответственно, для них мы не будем вводить квадраты и кубы. А вот у GDPpercap зависимость явно нелинейная, и можно поэкспериментировать.

Для начала, добавим к уже имеющимся регрессорам квадрат GDPpercap. Теперь модель выглядит следующим образом:

$$Unemploymen = \beta_0 + \beta_1 * GDPpercap + \beta_2 * GDPpercap^2 + \beta_3 * Urbanshare + \beta_4 * Higheduc$$

Оценим созданную модель и проведем тест Рамсея.

```
. gen GDPpercap2 = (GDPpercap)^2

. reg Unemployment GDPpercap Urbanshare Higheduc GDPpercap2
```

Source	SS	df	MS	Number of obs	=	85
				F(4, 80)	=	7.36
Model	290.223191	4	72.5557977	Prob > F	=	0.0000
Residual	788.742692	80	9.85928364	R-squared	=	0.2690
				Adj R-squared	=	0.2324
Total	1078.96588	84	12.8448319	Root MSE	=	3.1399

Unemployment	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDPpercap	-3.50e-06	2.36e-06	-1.48	0.142	-8.18e-06	1.19e-06
Urbanshare	-.1082392	.0309563	-3.50	0.001	-.1698442	-.0466342
Higheduc	-.042229	.0713085	-0.59	0.555	-.1841375	.0996795
GDPpercap2	7.84e-13	5.19e-13	1.51	0.135	-2.49e-13	1.82e-12
_cons	16.68797	2.852733	5.85	0.000	11.01085	22.36509

```
. ovtest

Ramsey RESET test using powers of the fitted values of Unemployment
       Ho:  model has no omitted variables
                  F(3, 77) =      4.93
                  Prob > F =      0.0035
```

Рис 2. Модель с добавлением квадрата и тест Рамсея

И снова на уровне значимости 5% мы отвергаем гипотезу о правильной спецификации. Заметим, что в данном случае p_{value} еще меньше, а F-статистика больше, чем в прошлой модели.

Наконец, оценим модель с добавлением и квадрата, и куба регрессора GDPpercap и проведем тест Рамсея.

```
. reg Unemployment GDPpercap Urbanshare Higheduc GDPpercap2 GDPpercap3
```

Source	SS	df	MS	Number of obs	=	85
				F(5, 79)	=	6.02
Model	297.524371	5	59.5048742	Prob > F	=	0.0001
Residual	781.441512	79	9.8916647	R-squared	=	0.2757
				Adj R-squared	=	0.2299
Total	1078.96588	84	12.8448319	Root MSE	=	3.1451

Unemployment	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDPpercap	-6.41e-06	4.13e-06	-1.55	0.125	-.0000146	1.81e-06
Urbanshare	-.1012052	.0320698	-3.16	0.002	-.1650385	-.0373718
Higheduc	-.0559502	.0731893	-0.76	0.447	-.20163	.0897295
GDPpercap2	2.90e-12	2.52e-12	1.15	0.253	-2.11e-12	7.91e-12
GDPpercap3	-3.39e-19	3.95e-19	-0.86	0.393	-1.12e-18	4.47e-19
_cons	17.40564	2.977014	5.85	0.000	11.48004	23.33124

```
. ovtest

Ramsey RESET test using powers of the fitted values of Unemployment
       Ho:  model has no omitted variables
                  F(3, 76) =      6.07
                  Prob > F =      0.0009
```

Рис 3. Модель с добавлением квадрата и куба; тест Рамсея

И снова после проведения теста Рамсея нулевая гипотеза отвергается на уровне значимости 5%, поэтому и эта модель не может быть признана правильно специфицированной.

На основании трех тестов продолжим наше исследование с линейной моделью без добавления квадратов и кубов, так как, несмотря на то что при добавлении многочленов R^2 немного увеличился, тест Рамсея с каждым разом показывал результаты хуже и хуже.

Мультиколлинеарность

Перейдем к исследованию проблемы мультиколлинеарности. Необходимость в обработке данных может возникнуть, если среди всего множества признаков существует подмножество признаков, корреляция которых примерно равна 1. Наличие такого подмножества отрицательно влияет на способность модели прогнозировать целевую переменную: небольшое изменение во входных данных может привести к очень сильному изменению коэффициентов перед регрессорами.

Одним из способов выявления наличия мультиколлинеарности является проверка так называемых VIF'ов данных – Variance Inflation Factor. Необходимо рассчитать VIF для каждого регрессора следующим образом:

1) Предположим, есть следующая линейная модель:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \varepsilon$$

2) Построим следующую линейную модель для j-го регрессора:

$$X_j = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_{j-1} X_{j-1} + \alpha_{j+1} X_{j+1} + \ldots + \alpha_n X_n + \varepsilon$$

3) Оценим R^2 данной модели и обозначим его за R_j^2

4) $VIF = \frac{1}{1-R_j^2}$

Будем исходить из следующих предположений: VIF > 6 свидетельствует о наличии мультиколлинеарности - линейной зависимости между объясняющими переменными, а VIF < 6 - о ее отсутствии.

Создадим простую линейную регрессию и проверим в Стате VIF'ы для каждого регрессора:

```
. reg Unemployment GDPpercap Urbanshare Higheduc
```

Source	SS	df	MS	Number of obs	=	85
				F(3, 81)	=	8.91
Model	267.746225	3	89.2487416	Prob > F	=	0.0000
Residual	811.219657	81	10.0150575	R-squared	=	0.2482
				Adj R-squared	=	0.2203
Total	1078.96588	84	12.8448319	Root MSE	=	3.1647

Unemployment	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDPpercap	-4.79e-08	5.81e-07	-0.08	0.935	-1.20e-06	1.11e-06
Urbanshare	-.1307659	.0273373	-4.78	0.000	-.1851586	-.0763732
Higheduc	-.0568294	.0712058	-0.80	0.427	-.1985066	.0848478
_cons	17.51559	2.821607	6.21	0.000	11.90148	23.12971

```
. vif
```

Variable	VIF	1/VIF
GDPpercap	1.09	0.915482
Urbanshare	1.09	0.921338
Higheduc	1.03	0.971315
Mean VIF	1.07	

Рис 4. VIF для линейной модели

Заметим, что как среднее значение, так и значение VIF'а каждой переменной не превосходит 6. Таким образом, мы можем спокойно сделать вывод об отсутствии мультиколлинеарности в нашей модели.