Домашнее задание 2
Часть 1
Дубровский Михаил, Пряничникова Алена
БЭК1811
2017 год

Введение

В представленной работе нами была рассмотрена зависимость показателя безработицы от изменения валового регионального продукта на основе данных Росстата. На основе регрессионного анализа в работе сделаны выводы о статистической взаимосвязи показателей ВРП и безработицы, а также о влиянии территориального деления региона на западный и восточный на зависимость.

Анализ dummy переменных

Прежде всего, мы разделили регионы на западные и восточные. Для этого была введена так называемая dummy переменную WEST, которая равна 1 для 58 западных регионов и 0 соответственно для 27 восточных. Мы воспользовались таблицей, данной нам в Приложении к домашнему заданию, для распределения 75 регионов, остальные прошли ручной отбор.

$$WEST_i = \begin{cases} 1, i - \text{западный регион} \\ 0, \text{иначе} \end{cases}$$

Следующим шагом было введение еще одной dummy переменной. Если предыдущая отвечала за изменение в свободном коэффициенте регрессии, теперь мы хотим ввести и дифференциацию для коэффициентов перед объясняющей переменной - GDPpercap. Данная переменная была названа ‘YWEST’ и задана с помощью статистического пакета STATA. Она была создана с помощью пакета STATA путем умножения ‘WEST’ на GDPpercap:

$$YWEST_i = \begin{cases} GDPpercap_i * WEST_i, i - \text{западный регион} \\ 0, \text{иначе} \end{cases}$$

Наконец, мы можем построить теоретическую модель регрессии:

$$Unempl_i = \beta_0 + \beta_1 * GDPpercap_i + \beta_2 * WEST_i + \beta_3 * YWEST_i + \varepsilon_i \quad (1)$$

Заметим, что в зависимости от значений dummy переменной, безработица региона может объясняться либо двумя переменными, либо четырьмя.

Перейдем к практической оценке регрессии. Используя STATA, нам были получены следующие результаты (Рис.1):

```
. reg Unemployment GDPpercap WEST YWEST
```

Source	SS	df	MS	Number of obs	=	85
				F(3, 81)	=	1.34
Model	50.8969072	3	16.9656357	Prob > F	=	0.2682
Residual	1028.06898	81	12.6922096	R-squared	=	0.0472
				Adj R-squared	=	0.0119
Total	1078.96588	84	12.8448319	Root MSE	=	3.5626

Unemployment	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDPpercap	-1.79e-06	1.01e-06	-1.77	0.081	-3.81e-06	2.26e-07
WEST	-1.620138	1.111022	-1.46	0.149	-3.830723	.5904468
YWEST	1.38e-06	1.30e-06	1.06	0.293	-1.21e-06	3.96e-06
_cons	8.034394	.9436033	8.51	0.000	6.156919	9.911868

Рис 1. Оцененные параметры регрессии (1).

Обсудим полученные результаты. Прежде всего, в глаза бросается значения $R^2 = 0.0472$. Он является очень низким, что говорит о довольно плохой способности модели в такой формации объяснить текущие зависимости. Данный факт подтверждает и F-статистика, $p_{value} = 0.2682$. Это говорит о том, что на любом разумном уровне значимости текущая модель не является адекватной.

Интересно заметить, что и введенные нами 2 dummy переменные также на любом адекватном уровне значимости не являются значимыми. Такой результат основан на полученных p_value. Значит, по идее, модель не сильно пострадает в качестве, если они будут выброшены.

Гипотезы о незначимости

Получив довольно плохие результаты, стоит задуматься о том, как улучшить текущую модель. И теперь можно задаться вопросом: имеет ли место единая зависимость для восточных и западных регионов? Перефразируем вопрос, а именно, нас интересует потенциальная значимость или незначимость введенных выше dummy переменных. Мы уже обсудили вариант выше, где отметили, что по отдельности они неважны для модели. Теперь подойдем к проблеме с другой стороны. Проверим одновременную значимость переменных через F-тест в начале, а потом перейдем к тесту Чоу.

Используя первый метод, выдвинем гипотезу H_0, что dummy переменные WEST и YWEST незначимы:

$$\begin{cases} H_0: WEST = YWEST = 0 \\ H_1: WEST^2 + YWEST^2 > 0 \end{cases}$$

Проверим данную гипотезу о незначимости dummy переменных на 5% уровне доверия. Для этого используем F-тест в STATA. Полученное значение p-value намного превышает заданный уровень значимости, следовательно, не отвергаем H_0, то есть признаем dummy переменные незначимыми (Рис. 2). Вновь пришли к выводу, что модель не слишком сильно пострадала бы в случае их отсутствия.

```
. test WEST YWEST

 ( 1)  WEST = 0
 ( 2)  YWEST = 0

       F(  2,    81) =    1.07
            Prob > F =    0.3469
```

Рис 2. F-test dummy переменных

Тест Чоу

Теперь же попробуем подтвердить наше предположение тестом Чоу. Для этого разделим общую выборку на 2 части: там, где WEST = 1, и где WEST = 0. После этого построим регрессии и рассчитаем RSS новых моделей. Зная значения RSS всех трех моделей, мы можем рассчитать F- статистику и сравнить ее с критическим значением $F(2, 81)$. Распишем имеющиеся модели. Модель (2) основана на выборке, где dummy переменная WEST = 0, и представляет из себя подвыборку восточных регионов:

$$Unemployment_{east_i} = \beta_0 + \beta_1 * GDPpercap_{east_i} + \varepsilon_i \quad (2)$$

Модель (3) представляет из себя такую же регрессию, только составленную на выборке из западных регионов (т.е. WEST = 1):

$$Unemployment_{west_i} = \beta_0 + \beta_1 * GDPpercap_{west_i} + \varepsilon_i \quad (3)$$

Наконец, модель (4) является общей, и в ней присутствуют как западные регионы, так и восточные:

$$Unemployment_j = \beta_0 + \beta_1 * GDPpercap_j + \varepsilon_j \quad (4)$$

Протестируем имеющиеся теоретические регрессии на данных и подробно изучим результаты.

```
. reg Unemployment GDPpercap
```

Source	SS	df	MS	Number of obs	=	85
				F(1, 83)	=	1.86
Model	23.665161	1	23.665161	Prob > F	=	0.1762
Residual	1055.30072	83	12.7144665	R-squared	=	0.0219
				Adj R-squared	=	0.0101
Total	1078.96588	84	12.8448319	Root MSE	=	3.5657

Unemployment	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDPpercap	-8.55e-07	6.27e-07	-1.36	0.176	-2.10e-06	3.92e-07
_cons	6.869102	.4979621	13.79	0.000	5.878676	7.859529

Рис 3. Модель (4) – общая. $RSS_R = 1055.30072$

```
. reg Unemployment GDPpercap if WEST == 1
```

Source	SS	df	MS	Number of obs	=	58
				F(1, 56)	=	0.23
Model	3.34073313	1	3.34073313	Prob > F	=	0.6322
Residual	807.567025	56	14.4208397	R-squared	=	0.0041
				Adj R-squared	=	-0.0137
Total	810.907759	57	14.2264519	Root MSE	=	3.7975

Unemployment	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDPpercap	-4.16e-07	8.65e-07	-0.48	0.632	-2.15e-06	1.32e-06
_cons	6.414256	.625165	10.26	0.000	5.1619	7.666612

Рис 4. Модель (3) – западные регионы. $RSS_1 = 807.567025$

```
. reg Unemployment GDPpercap if WEST == 0
```

Source	SS	df	MS	Number of obs	=	27
				F(1, 25)	=	4.49
Model	39.624717	1	39.624717	Prob > F	=	0.0441
Residual	220.50195	25	8.82007798	R-squared	=	0.1523
				Adj R-squared	=	0.1184
Total	260.126667	26	10.0048718	Root MSE	=	2.9699

Jnemployment	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDPpercap	-1.79e-06	8.46e-07	-2.12	0.044	-3.53e-06	-5.08e-08
_cons	8.034394	.7866056	10.21	0.000	6.414349	9.654438

Рис 5. Модель (2) – восточные регионы. $RSS_2 = 220.50195$

Получив качественные метрики по моделям, можем воспользоваться F-статистикой, которая используется для теста Чоу:

$$F = \frac{(RSS_R - [RSS_1 + RSS_2])/(k+1)}{(RSS_1 + RSS_2)/(n-2(k+1))} \sim F\big(k+1, n-2(k+1)\big),$$

где k – количество коэффициентов-“slopes”, n – количество наблюдений. Посчитаем для нашей выборки: $F_{test} = \frac{(1055.30072 - [807.567025 + 220.50195])/2}{(807.567025 + 220.50195)/(85-2(1+1))} = \frac{13.61}{12.69} = 1.07 \sim F(2, 81)$.
Сравним полученную статистику с $F_{crit}^{5\%}(2, 81) = 3.11$. Видим, что $F_{test} < F_{crit}$ $\Rightarrow$ гипотеза о необходимости рассмотрения двух разных моделей отбрасывается и подтверждается гипотеза о равенстве модели. Тест Чоу подтвердил всё сказанное выше: при такой предобработке данных и спецификации модели нет разницы между тем, западный или нет регион.

Таким образом, на основе анализа регрессии мы выяснили, что зависимость безработицы от ВРП не зависит от типа регионов: восточный или западный.

Анализ данных, выбросов и улучшение модели

Как уже было много раз сказано, представленные выше модели рассчитывались на неочищенных данных. Учитывая, что обучается Ridge регрессия, а она довольно чувствительна к выбросам, теоретически можно было ожидать такой результат. Посмотрим график рассеяния, где по Ox находится $GDPpercap$, а по $Oy - Unemployment$. На нём также видна построенная регрессия из модели (4) (Рис 6.):

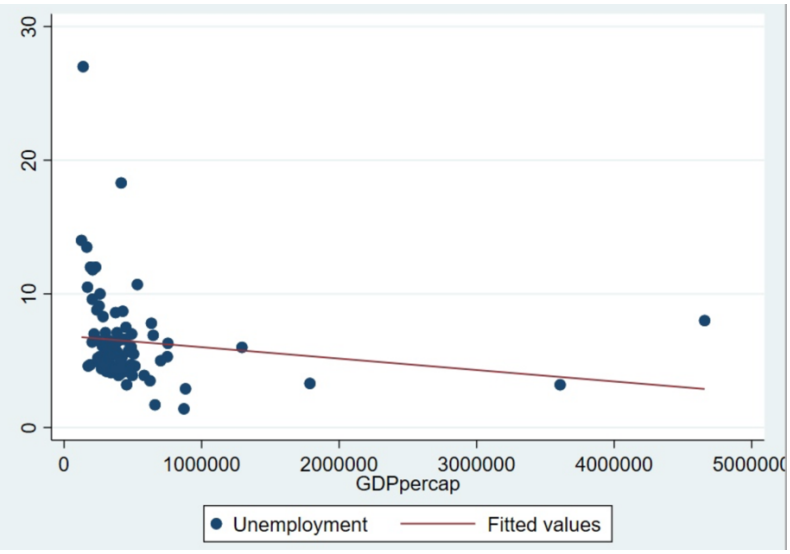

Рис. 6 – Scatter Plot и регрессия из (4).

На графике можно легко выделить выбросы в данных. Есть два пути, как с этим бороться: можно либо вообще выкинуть их из модели, однако можем потерять общность и способность выдавать более качественные прогнозы на выбросах; либо же сделать специальную метку (dummy переменную), что этот сэмпл является выбросом. Мы пойдем по второму пути и введём 1 dummy переменную: XWEST

$$XWEST_i = \begin{cases} 1, \text{если } GDP_{percap} > 20\,000\,000 \\ 0, \text{иначе} \end{cases}$$

Подчеркнём момент, почему мы отказались от введения dummy переменных для вертикальных выбросов. В самом деле, это не имеет смысла: суть регрессии заключается в построении модели прогнозирования и предсказания на входящих данных. Задача – подобрать оптимальные веса. И если мы введем в «обучающую» выборку dummy переменную, которая основана на значениях таргета модели – безработицы – при том, что она не известна на входе, получается, что модель просто запоминает выборку. Для действительно адекватной модели это бессмысленно, поэтому такая идея была отброшена.

Вернемся к новой модели с dummy переменной для выбросов по Ox. Обучим ее и посмотрим на результат: видим, что R^2 стал выше, однако, F_{stat} стала хуже. Видим, что небольшое улучшение, пусть и небольшое (Рис. 7)

```
. reg Unemployment GDPpercap XWEST
```

Source	SS	df	MS	Number of obs	=	85
				F(2, 82)	=	1.31
Model	33.4410294	2	16.7205147	Prob > F	=	0.2750
Residual	1045.52485	82	12.7503031	R-squared	=	0.0310
				Adj R-squared	=	0.0074
Total	1078.96588	84	12.8448319	Root MSE	=	3.5708

Unemployment	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDPpercap	-1.89e-06	1.34e-06	-1.41	0.162	-4.56e-06	7.75e-07
XWEST	3.926465	4.484193	0.88	0.384	-4.994023	12.84695
_cons	7.249766	.6615581	10.96	0.000	5.933716	8.565816

Рис 7. Модель с введением переменной XWEST.

Также стоит заметить, что безработица слабоотрицательно зависит от ВРП, поскольку коэффициент GDPpercap является незначимым. То есть при измении ВРП на единицу существенного изменения безработицы не произойдет. Значение коэффициента XWEST, отвечающего за выбросы, равно 3.926, следовательно, в регионах-выбросах в среднем значение на 4 % выше значения безработицы в других регионах.