РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИННОВАЦИОННЫХТЕХНОЛОГИЙ И ПРЕДПРИНИМАТЕЛЬСТВА
ПЕНЗЕНСКИЙ ФИЛИАЛ

ОТЧЕТ

о выполнении лабораторной работы №2

ДисциплинаЭконометрика

Тема: «Множественная линейная регрессия»

Выполнил: студент гр. 08э2 Макуев Тимур

Проверил: преподаватель И.Ю. Денисова

2011

Цели работы:

- рассмотреть множественную линейную регрессию и ее характеристики;
- закрепить навыки решением типовой задачи на основе использованияIBMSPSSStatistics.

Ход работы:

Условие задачи (43)

Для изучения проблемы рассмотрите следующие показатели и их значения по территориям Центрального федерального округа за 2001 г.:

y_1– численность безработных, тыс. чел.;

x_1– годовой фонд заработной платы занятых в экономике региона, млрд руб.;

x_2 – численность мигрантов за год, тыс. чел.;

x_3 – численность безработных в расчете на одну заявленную вакансию, чел.;

x_4 – число малых предприятий в регионе, тыс.

Задание:

Установить зависимость числа совершенных преступлений в регионе от социально-экономических факторов, оказывающих наибольшее воздействие на данный процесс. Выполните расчет прогнозного значения результата, предполагая, что прогнозные значения факторов составят 102,9% от их среднего уровня.

Необходимо:

1. Построить линейное уравнение множественной регрессии с полным перечнем заданных показателей и оценить его;

2. Провести исключение неинформативных переменных и получить модель только с информативными переменными для уровня значимости $\alpha = 10\%$;

3. Построить матрицу парных коэффициентов корреляции. Установить, какие факторы мультиколлинеарны.Рассчитать множественный коэффициент корреляции;

4. Дать оценку полученного уравнения на основе коэффициента детерминации и общего *F*-критерия Фишера.

5. Выполнить анализ результатов, построить прогноз уровня результата, указав, при каких условиях он будет возрастать и при каких – снижаться.

Таблица 1. Исходные данные

Субъекты РФ	y_1	x_1	x_2	x_3	x_4
Белгородская обл.	48,3	38,30	11,09	1,3	4,6
Брянская обл.	65,3	28,74	-0,14	3,3	3,2
Владимирская обл.	80,5	30,93	2,69	3,0	6,9
Воронежская обл.	107,6	58,81	2,67	1,8	11,0
Ивановская обл.	33,1	18,11	1,20	1,3	5,2
Калужская обл.	33,1	21,58	0,96	0,9	5,9
Костромская обл.	22,8	17,00	0,31	1,1	3,2
Курская обл.	65,0	28,84	-1,29	1,3	2,8
Липецкая обл.	39,8	33,26	5,05	0,7	4,3
Орловская обл.	34,3	20,45	1,51	1,5	2,6
Рязанская обл.	66,7	27,89	-0,38	0,7	6,4
Смоленская обл.	55,1	29,99	-1,44	1,3	2,4
Тамбовская обл.	67,4	29,98	-2,62	4,6	3,6
Тверская обл.	60,4	30,39	-0,31	0,9	5,7
Тульская обл.	43,4	41,08	-1,87	1,3	6,5
Ярославская обл.	52,0	41,81	1,53	0,9	7,1

Решение задачи

1**.Необходимо построить линейное уравнение множественной регрессии с полным перечнем заданных показателей и оценить его.**

Так как у нас после вывода результатов остались 2 надежные модели, то мы получим 2 линейных уравнения множественной регрессии для соответствующих моделей.

Уравнение первой модели выглядит следующим образом:

$$y=11,149+1,401x_1$$

Уравнение второй модели выглядит следующим образом:

$$y=-0,381+1,371x_1+7,703x_3$$

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. Ошибка	Бета		
1	(Константа)	11,149	12,922		,863	,403
	x1	1,401	,395	,687	3,542	,003
2	(Константа)	-,381	12,323		-,031	,976
	x1	1,371	,345	,673	3,968	,002
	x3	7,703	3,320	,393	2,320	,037

a. Зависимая переменная: y1

Величины a1=11,149 и a2=-0,381 оценивают агрегированное влияние прочих (кроме учтенных в моделях факторов x1 и x3) факторов на результат y.Величины b1 и b2 указывают, что с увеличением x1 и x2 на единицу их значений результат увеличивается соответственно на 1,401 и 1,371+7,703=9,074 соответственно. Сравнивать эти значения не следует, так как они зависят от единиц измерения каждого признака и потому несопоставимы между собой.

2. Необходимо провести исключение неинформативных переменных и получить модель только с информативными переменными для уровня значимости α = 10%.

В таблице 2 фиксируется процесс пошагового включения/исключения переменных в регрессионную модель.

Как видно из таблицы, очередность включения переменных такова: x1, x3. Переменные x2 и x4 не были включены, также не потребовалось исключать какую–либо переменную. Можно отметить следующие основания или критерии для такой приоритетности, хотя они и не являются полностью взаимно независимыми:

• Статистическая значимость, связанная с принятием данной переменной в регрессию. Значение критерия Фишера для включения каждой из этих переменных <0,1, для исключения > 0,1. Другими словами, нулевая гипотеза, состоящая в том, что результат действия случаен и статистически незначим, отвергается в первом случае и не отвергается во втором;

Таблица 2. Выходная информация множественной регрессии

Введенные или удаленные переменные[a]

Модель	Включенные переменные	Исключенные переменные	Метод
1	x1	.	Шаговый (критерий: вероятность F-включения <= ,050, F-исключения >= ,100).
2	x3	.	Шаговый (критерий: вероятность F-включения <= ,050, F-исключения >= ,100).

a. Зависимая переменная: y1

• Модель 1 (только переменная x1) –таблица 3– объясняет почти 48% вариации зависимой переменной (R^2= 0,473, скорректированныйR^2 = 0,435, что несущественно). Модель 2, где добавляется переменная x3, поднимает R^2, а значит, и уровень объяснения вариации до 0,627 (0,570) или чуть больше, чем на 15%. То есть основная доля вариации объясняется переменной x1.

Таблица 3. Сводка для моделей

Сводка для модели[c]

Модель	R	R-квадрат	Скорректированный R-квадрат	Стд. ошибка оценки
1	,687[a]	,473	,435	15,98774

2	,792[b]	,627	,570	13,95198

a. Предикторы: (конст) x1

b. Предикторы: (конст) x1, x3

c. Зависимая переменная: y1

3. Построить матрицу парных коэффициентов корреляции. Установить, какие факторы мультиколлинеарны.Рассчитать множественный коэффициент корреляции.

При построении уравнения множественной регрессии возникает проблема мультиколлинеарности факторов, их тесной линейной взаимозависимости. Мультиколлинеарность может проявляться в функциональной (явной) и стохастической (скрытой) формах.

Обычно считается, что две переменные явно коллинеарны или находятся между собой в линейной зависимости, если их коэффициент корреляции > 0,7. Однако по величине парных коэффициентов корреляции обнаруживаетсялишь явная коллинеарность факторов. Наибольшие трудности при использовании аппарата множественной регрессии возникают при наличии стохастической (скрытой) мультиколлинеарности: чем она сильнее, тем менее надежна оценка распределения суммы объясненной вариации по отдельным факторам с использованием R^2. Точных количественных критериев для определения наличия или отсутствия скрытой коллинеарности не существует – можно говорить лишь о некоторых эвристических подходах к ее выявлению.

Для оценки мультиколлинеарности факторов используется определитель матрицы парных коэффициентов корреляции между факторами: чем ближе он к нулю, тем сильнее мультиколлинеарность факторов и ненадежнее результаты множественной регрессии, и наоборот.

Для уравнения регрессии (где в качестве зависимой переменной выступает y1), сравним различные варианты набора независимых переменных:

- x1, x3;
- x1, x3, x2;

- x1, x2, x3, x4.

Для каждого из этих вариантов построим матрицу парных коэффициентов корреляции и рассчитаем ее определитель.

Откроем исходный файл и выполним последовательность команд **Анализ ► Снижение размерности ► Факторный анализ.** В диалоговом окне **Факторный анализ** зададим сначала переменные x1и x3, после чего нажмем **кнопку Дескриптивные**.

Затем в окне **Факторный анализ: Дескриптивные** активизируем позиции для корреляционной **матрицы Коэффициенты и Детерминант.** Снова скомандуем **ОК.** Повторим эту процедуру еще два раза, задавая последовательно в качестве переменных x1, x3, x2 и x1, x3, x2,x4.

В таблицах 4, 5, 6 показаны результаты – матрицы парных коэффициентов корреляции и значения детерминантов для каждого из перечисленных вариантов.

Таблица 4. Корреляционная матрица для двух переменных

Матрица корреляций[a]

		x1	x3
Корреляция	x1	1,000	,038
	x3	,038	1,000

a. Детерминант = ,999

Таблица 5. Корреляционная матрица для трех переменных

Матрица корреляций[a]

		x1	x3	x2
Корреляция	x1	1,000	,038	,246
	x3	,038	1,000	-,234
	x2	,246	-,234	1,000

a. Детерминант = ,879

Таблица 6. Корреляционная матрица для четырех переменных

Матрица корреляций[a]

		x1	x3	x2	x4
Корреляция	x1	1,000	,038	,246	,703

	x3	,038	1,000	-,234	-,122
	x2	,246	-,234	1,000	,168
	x4	,703	-,122	,168	1,000

а. Детерминант = ,424

Как видно из этих таблиц, в первом варианте (определитель равен 0,999) мультиколлинеарность невыражена, и практически равна 1, что говорит о надежности результатов множественной регрессии. Во втором варианте (0,879) с добавлением новой независимой переменной x2происходит некоторое появление, но определитель все равно остается достаточно большим. Но для третьего варианта, последобавления переменной x4, можно говорить о появлении достаточно выраженной мультиколлинеарности – определитель уменьшается в 2 раза (0,424). Отсюда следует, что данная переменная практически линейно не связана с другими.

4. Дать оценку полученного уравнения на основе коэффициента детерминации и общего *F*-критерия Фишера.

Максимально возможное значение критерия под влиянием случайных факторов при данных степенях свободы и уровне значимости. , где уровень значимости берется равным 0,05. Вычисляется критическое значения с помощью функции Excel*FРАСПОБР*, которая возвращает обратное значение для одностороннего F-распределения вероятностей. Степени свободы берутся соответственно 1 и 14.

При сравнении фактического и критического значений *F*-критерия Фишера для оценивания статистической надежности результатов регрессионного моделирования, выясняется, что лучшей моделью является первая модель (F_1факт $= 12{,}546 > F_1$табл $= 4{,}6$) , но вторая модель () также является надежной. Это означает, что гипотеза о случайной природе оцениваемых характеристик отклоняется и признается их статистическая значимость и надежность.

Квадрат коэффициента корреляции *R*есть индекс детерминации, определяющий качество построенной модели. Чем ближе R^2к единице, тем лучше регрессия описывает связь между независимыми и зависимой переменной. В нашем случае, как видно из таблицы 3, коэффициенты детерминации в обеих моделях не высокий, но достаточно значимый, во второй модели достигает практически 0,6. Значения R^2и скорректированного R^2существенно различаются, это говорит о том, что используется слишком много независимых переменных при недостаточном объеме выборки. В таком случае скорректированныйR^2заслуживает большего доверия.

По своему математическому смыслу R^2характеризует долю от общей дисперсии зависимой переменной *Y,* объясняемую регрессией. Иначе говоря, коэффициент детерминации определяется как отношение дисперсии, обусловленной регрессией, к общей регрессии.

5. Выполнить анализ результатов, построить прогноз уровня результата, указав, при каких условиях он будет возрастать и при каких – снижаться.

Используя аналитическую систему SPSS, я получил следующие данные по коэффициентам:

Таблица 7. Коэффициенты

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.	95,0%% доверительный интервал для B	
		B	Стд. Ошибка	Бета			Нижняя граница	Верхняя граница
1	(Константа)	11,149	12,922		,863	,403	-16,567	38,864
	x1	1,401	,395	,687	3,542	,003	,553	2,249
2	(Константа)	-,381	12,323		-,031	,976	-27,004	26,242
	x1	1,371	,345	,673	3,968	,002	,624	2,117
	x3	7,703	3,320	,393	2,320	,037	,531	14,876

$t_{табл}$ для числа степеней свободы df=n-2=16-2=14 и α =0,05 по таблице составит 2,144.

Для первой модели:

Случайные ошибкиm_a, m_b:

$m_a = 12{,}922, m_b = 0{,}395.$

$$t_a = 0{,}863, t_b = 3{,}542 .$$

Фактическое значение не превосходит табличное, поэтому гипотеза не откланяется, правда t_b превосходит, т.е. b статистически значим.

Доверительные интервалы:

$$\gamma_{a_{\min}} = -16{,}567 ;$$

$$\gamma_{a_{\max}} = 38{,}864 ;$$

$$\gamma_{b_{\min}} = 0{,}553 ;$$

$$\gamma_{b_{\max}} = 2{,}249 ;$$

Анализ границ доверительных интегралов приводит к выводу о том, чтотолько параметрb_1является статистически значимым.

Если прогнозные значения факторов составят 102,9% от их среднего уровня, то$x_p = 31{,}1 \cdot 1{,}029 = 32$, то $\hat{y}_p = 11{,}149 + 1{,}401 \cdot 32 = 56$

Ошибка прогноза:

$$m_{\hat{y}_p} = 16 \cdot \sqrt{1 + \frac{1}{16} + \frac{(32 - 31{,}1)^2}{1634}} = 16{,}5$$

Предельная ошибка прогноза:

$$\Delta_{\hat{y}_p} = 2{,}144 \cdot 16{,}5 = 35{,}4 .$$

Доверительный интервал прогноза:

$$\gamma_{\hat{y}_p} = 56 \pm 35{,}4 ;$$

$$\gamma_{\hat{y}_{p_{\min}}} = 56 - 35{,}4 = 20{,}6 ;$$

$$\gamma_{\hat{y}_{\max}} = 56 + 35{,}4 = 91{,}4 ;$$

Диапазон границ доверительного интервала:

$$D_\gamma = \frac{91,4}{20,6} = 4,43 .$$

Для второй модели:

Случайные ошибки m_a, m_{b_1}, m_{b_2}:

$$m_a = -0,381, m_{b_1} = 1,371, m_{b_2} = 7,703.$$

$$t_a = -0,031, t_{b_1} = 3,968, t_{b_2} = 2,320 .$$

Доверительные интервалы:

$$\gamma_{a_{\min}} = -27,004 ;$$

$$\gamma_{a_{\max}} = 26,242 ;$$

$$\gamma_{b_{1\min}} = 0,624 ;$$

$$\gamma_{b_{1\max}} = 2,117 ;$$

$$\gamma_{b_{2\min}} = 0,531 ;$$

$$\gamma_{b_{2\max}} = 14,876 ;$$

Анализ границ доверительных интегралов приводит к выводу о том, чтотолько параметры b_1 и b_2 являются статистически значимыми.

Если прогнозные значения факторов составят 102,9% от их среднего уровня, то $x_{p_1} = 31,1 \cdot 1,029 = 32$, $x_{p_2} = 1,6 \cdot 1,029 = 1,64$, то $\hat{y}_p = -0,381 + 1,371 \cdot 32 + 7,703 \cdot 1,64 = 56$

Ошибка прогноза:

$$m_{\hat{y}_p} = 16 \cdot \sqrt{1 + \frac{1}{16} + \frac{(32 - 31,1)^2}{1634} + \frac{(1,64 - 1,6)^2}{17,7}} = 16,5$$

Предельная ошибка прогноза:

$$\Delta_{\hat{y}_p} = 2,144 \cdot 16,5 = 35,4 .$$

Доверительный интервал прогноза:

$$\gamma_{\hat{y}_p} = 56 \pm 35,4 ;$$

$$\gamma_{\hat{y}_{p\min}} = 56 - 35,4 = 20,6 ;$$

$$\gamma_{\hat{y}_{\max}} = 56 + 35,4 = 91,4 ;$$

Диапазон границ доверительного интервала:

$$D_Y = \frac{91,4}{20,6} = 4,43$$.

Выводы:

- я рассмотрел множественную линейную регрессию и ее характеристики;
- я закрепил навыки решением типовой задачи на основе использованияIBMSPSSStatistics.