Московский государственный университет им. М.В.Ломоносова

Зачетная работа по курсу

“Математическая статистика – 2”

студента Школы магистров

направление «Математические методы анализа экономики»

Туманова Андрея Анатольевича

Задание № 22

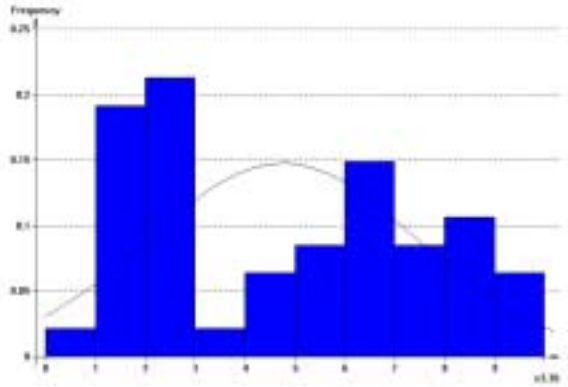
Москва
2001

Зачетная работа по курсу «Математическая статистика - 2»

1. Вариационные характеристики рядов:

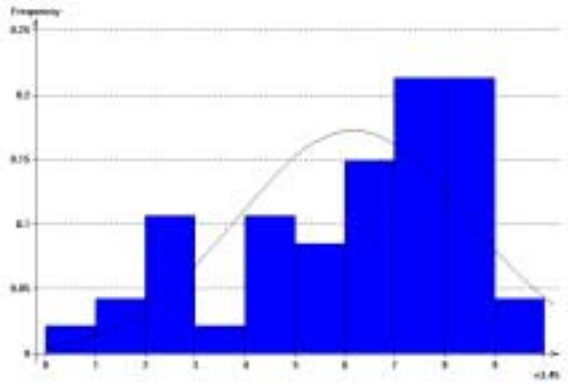
1.1. Коррупция – v3.36

Среднее	4.763	Минимум	0.92
Дисперсия	7.288	Максимум	9.34
Среднеквадр.		Размах	8.42
отклонение (ск.)	2.700	Медиана	4.64
Среднеквадр.		Квантиль 20%	1.99
отклонение (неск.)	2.671	Квантиль 80%	7.51
Коэффициент		Асимметрия	0.19
вариации	0.57	Эксцесс	-1.46



1.2. Имущественная безопасность – v3.45

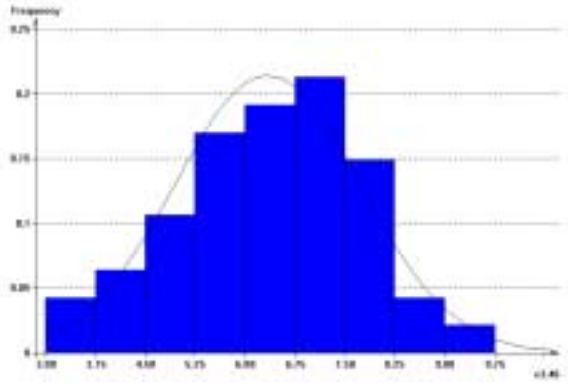
Среднее	6.161	Минимум	0.82
Дисперсия	5.350	Максимум	9.23
Среднеквадр.		Размах	8.41
отклонение (ск.)	2.313	Медиана	6.93
Среднеквадр.		Квантиль 20%	4.42
отклонение (неск.)	2.288	Квантиль 80%	8.30
Коэффициент		Асимметрия	0.81
вариации	0.38	Эксцесс	-0.29



1.3. Социальная интеграция – v3.46

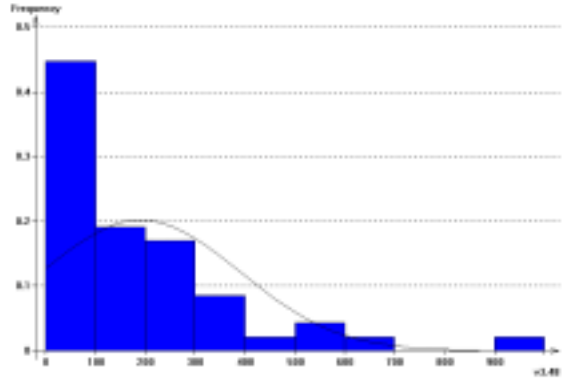
Среднее	6.335	Минимум	3.21
Дисперсия	1.947	Максимум	9.25
Среднеквадр.		Размах	6.04
отклонение (ск.)	1.395	Медиана	6.19
Среднеквадр.		Квантиль 20%	5.20
отклонение (неск.)	1.380	Квантиль 80%	7.57
Коэффициент		Асимметрия	-0.20
вариации	0.22	Эксцесс	-0.51

С вероятностью в 95% данное распределение является нормальным (χ^2=2.09, число степ. свободы=7)



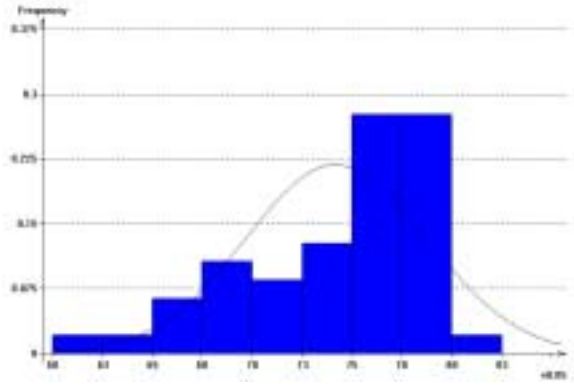
1.4. Серьезные преступления – v3.48

Среднее	191.07	Минимум	7.9
Дисперсия	39088.29	Максимум	951.8
Среднеквадр.		Размах	943.9
отклонение (ск.)	197.71	Медиана	109.1
Среднеквадр.		Квантиль 20%	53.5
отклонение (неск.)	195.59	Квантиль 80%	279.3
Коэффициент		Асимметрия	1.92
вариации	1.03	Эксцесс	4.19



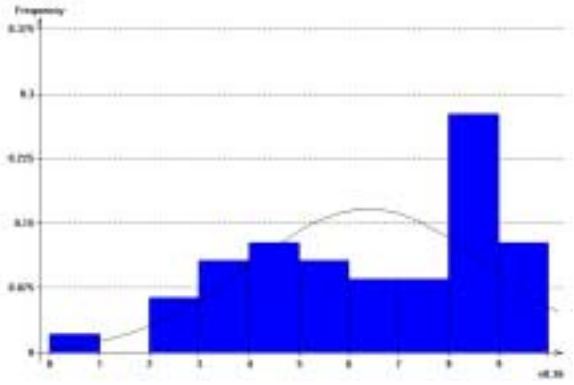
1.5. Средняя продолжительность жизни – v8.05

Среднее	74.23	Минимум	62.4
Дисперсия	20.83	Максимум	80.0
Среднеквадр.		Размах	17.6
отклонение (ск.)	4.56	Медиана	76.4
Среднеквадр.		Квантиль 20%	70.1
отклонение (неск.)	4.52	Квантиль 80%	78.1
Коэффициент		Асимметрия	-0.93
вариации	0.06	Эксцесс	-0.08



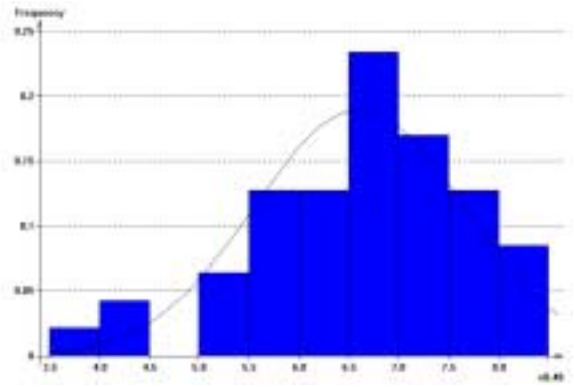
1.6. Качество жизни – v8.36

Среднее	6.394	Минимум	0.54
Дисперсия	5.738	Максимум	9.59
Среднеквадр.		Размах	9.05
отклонение (ск.)	2.395	Медиана	6.92
Среднеквадр.		Квантиль 20%	4.19
отклонение (неск.)	2.370	Квантиль 80%	8.73
Коэффициент		Асимметрия	-0.41
вариации	0.37	Эксцесс	-0.97



1.7. Злоупотребление алкоголем – v8.40

Среднее	6.607	Минимум	3.75
Дисперсия	1.108	Максимум	8.32
Среднеквадр.		Размах	4.57
отклонение (ск.)	1.053	Медиана	6.78
Среднеквадр.		Квантиль 20%	5.78
отклонение (неск.)	1.041	Квантиль 80%	7.57
Коэффициент		Асимметрия	-0.53
вариации	0.16	Эксцесс	0.25



Приведенные выше характеристики рядов свидетельствуют о том, что практически все ряды имеют достаточно большой разброс. Использование для сравнения разброса ряда дисперсию (или среднеквадратическое отклонение), а уж тем более размах ряда, не правильно, т.к. ряды не нормированы (т.е. не приведены к одной шкале измерения). Поэтому в качестве такого показателя можно использовать коэффициент вариации, т.е. отношения среднеквадратического отклонения к среднему значению. Чем меньше данный показатель, тем менее вероятно, что ряд будет влиять на объясняемый показатель, т.е. на «Личное поручительство». При превышении коэффициента вариации уровня в 33% (0.33), совокупность считается не однородной. В данной задаче к рядам, имеющим однородную совокупность можно

отнести следующие: v3.46 «Социальная интеграция», v8.05 «Средняя продолжительность жизни» и v8.40 «Злоупотребление алкоголем».

Значения асимметрии и эксцесса приведены для анализа распределения значений ряда. Эксцесс отражает «крутизну» кривой распределения, а эксцесс – «скос» значений.

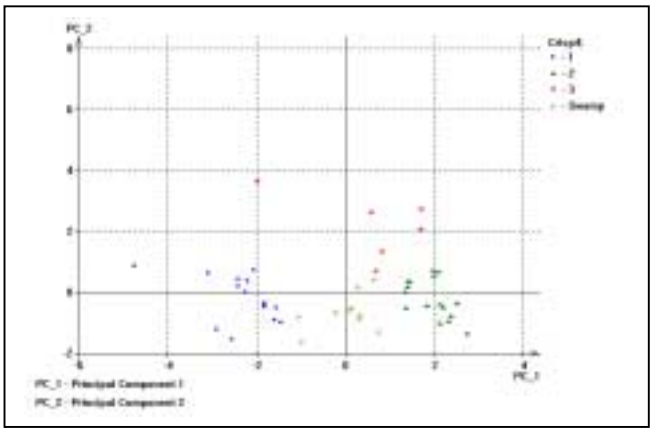
На графиках линией показано нормальное распределение.

2. Кластерный анализ объясняющих переменных

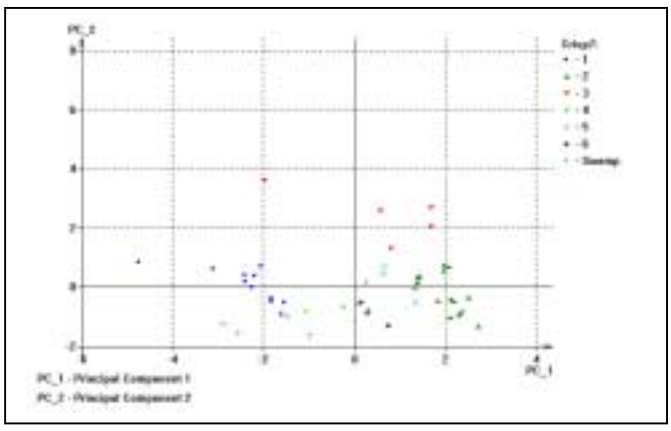
В результате кластерного анализа методом k-средних была получена следующая зависимость между количеством классов и одним из главных показателей – доли объясненной дисперсии.

Кол-во классов	Доля объясненной дисперсии, %	Прирост, %
2	48.17	-
3	54.17	6.00
4	**63.95**	**9.78**
5	67.80	3.85
6	70.46	2.66
7	**72.83**	**2.37**
8	73.29	0.46
9	**77.15**	**3.86**
10	78.00	0.85
11	78.19	0.19
12	78.42	0.23
13	**83.81**	**5.39**
14	85.35	1.54
15	**87.87**	**2.52**
16	88.69	0.82
17	90.55	1.86
18	90.24	-0.31
19	91.77	1.53
20	92.64	0.87
Auto	78.42	= 12

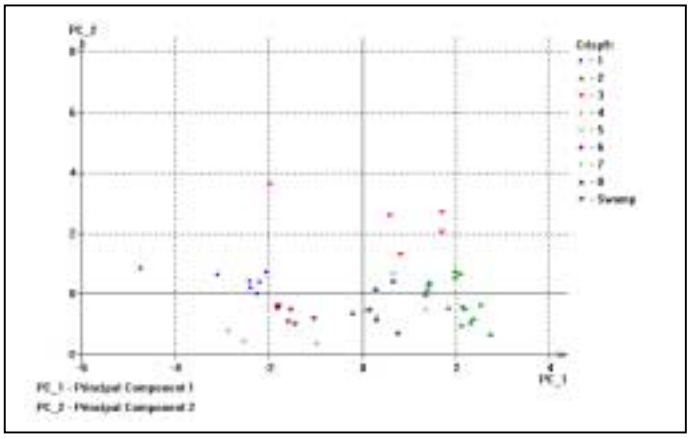
Исходя из анализа приращений доли объясненной дисперсии при увеличении числа классов, можно сделать следующий вывод: в зависимости от стоящих перед исследователем задач, оптимальное количество классов может составить 4 (в случае разбиения на группы типа «развивающиеся», «развитые» и страны «с переходной экономикой»), 7 или 9 (если требуется более детальный анализ) или 13 (15). Рассмотрим эти случаи более подробно:



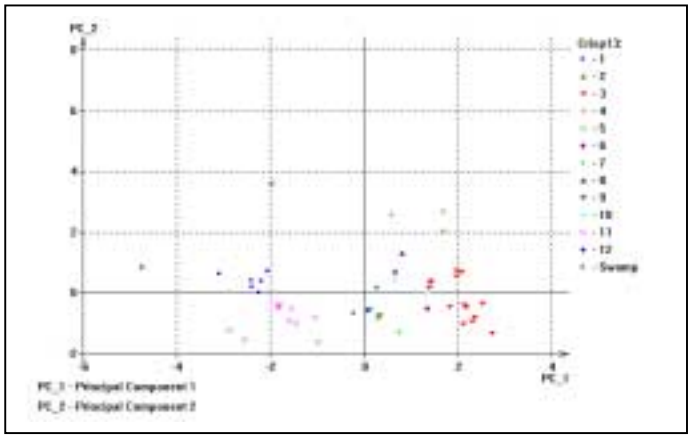
4 кластера



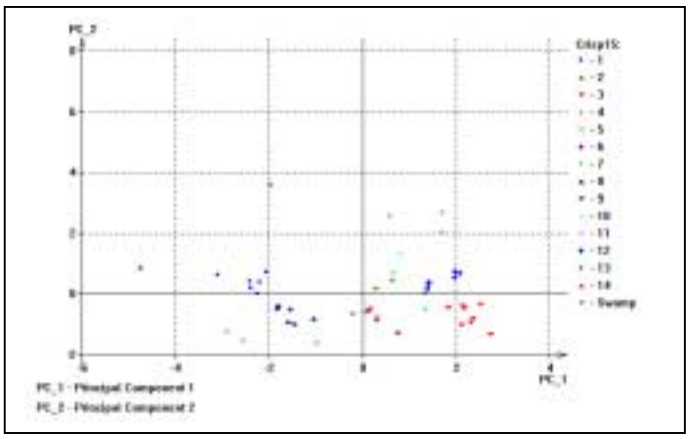
7 кластеров



9 кластеров



13 кластеров



15 кластеров

Для данного исследования разбиение стран на 4 класса – неприемлемо из-за незначительной доли объясненной дисперсии. 15 классов – это слишком дробное разбиение, затрудняющее анализ. Выбор между 7 и 9 классами решается в пользу 9 за счет существенной добавки в объясненной дисперсии. Анализируя структуру разбиения на 9 и 13 классов, приведенную ниже, можно сделать выбор в пользу 13 классов, т.к. при разбиении на 9 классов в один кластер попадают такие разные страны, как, например, США, Австралия, Бельгия и Новая Зеландия с одной стороны и Южная Африка с другой. В связи с этим дальнейший анализ будет проводиться для разбиения на 13 классов.

9 кластеров		13 кластеров	
№	Состав класса	№	Состав класса
1	Венгрия Польша Колумбия Бразилия Венесуэлла Словения Россия	1	Венгрия Польша Колумбия Бразилия Венесуэлла Словения Россия
2	Швейцария Норвегия Нидерланды Финляндия Германия Швеция Канада Испания Австрия Дания Франция Исландия	2	Швейцария Норвегия Нидерланды Финляндия Германия Швеция Канада Испания Австрия Дания Франция Исландия

	Люксембург Сингапур		Люксембург Сингапур
3	США Австралия Бельгия Новая Зеландия Южная Африка	3	Австралия США Новая Зеландия
4	Индонезия Индия Турция	4	Индонезия Индия Турция
5	Израиль Гонконг	5	Израиль Гонконг
6	Греция Италия Япония	6	Греция Италия Япония
7	Португалия Малайзия	7	Малайзия
8	Великобритания Чили Ирландия	8	Великобритания Ирландия
9	Чехия	9	Мексика

	Мексика Филиппины Таиланд Корея Аргентина Тайвань Китай		Чехия Филиппины Таиланд Корея Аргентина Китай
		10	Южная Африка
		11	Бельгия
		12	Португалия
		13	Тайвань Чили

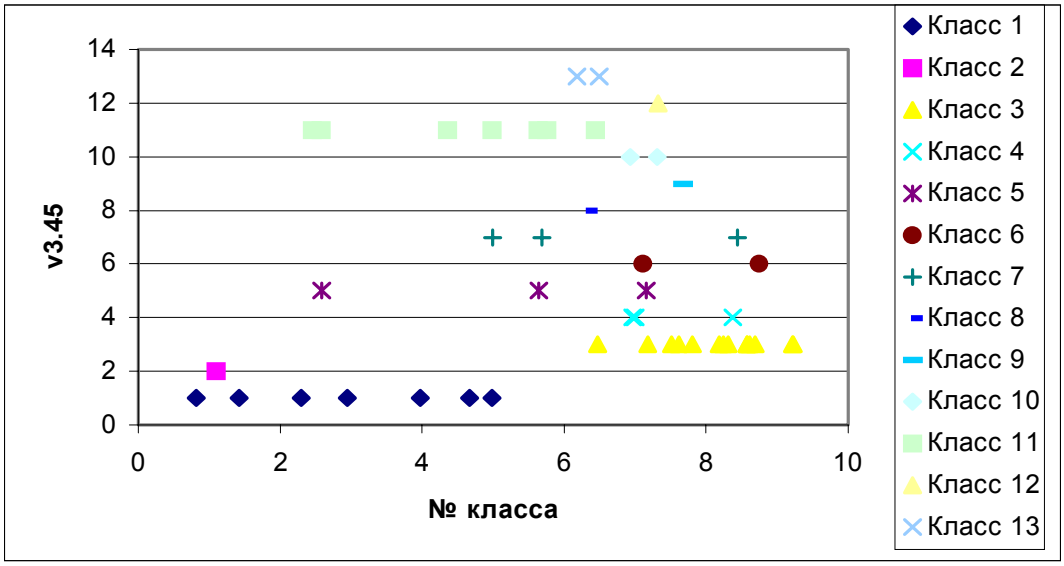
Характеристика центров классов:

Перем.	1	2	3	4	5	6	7	8	9	10	11	12	13
v3.36	2.01	2.28	7.52	7.66	1.73	6.88	3.36	4.21	3.73	6.71	1.93	4.27	5.57
v3.46	4.48	5.75	7.70	6.52	5.30	5.35	5.92	6.15	8.33	7.40	5.57	7.46	6.59
v3.48	192.26	951.80	170.43	614.13	23.03	299.75	65.50	499.70	52.90	126.30	65.94	64.60	129.00
v8.05	69.83	65.20	77.66	77.40	65.50	78.25	78.80	77.30	72.00	76.85	71.21	75.40	74.50
v8.36	3.08	5.09	8.83	8.79	3.44	6.65	6.79	8.45	7.16	7.94	4.51	5.56	5.59
v8.40	5.05	5.38	7.34	5.95	7.46	7.98	7.44	7.01	6.57	5.78	6.30	6.31	6.48

Среднее значение зависимой переменной в каждом из 13 классов и диапазон ее изменения

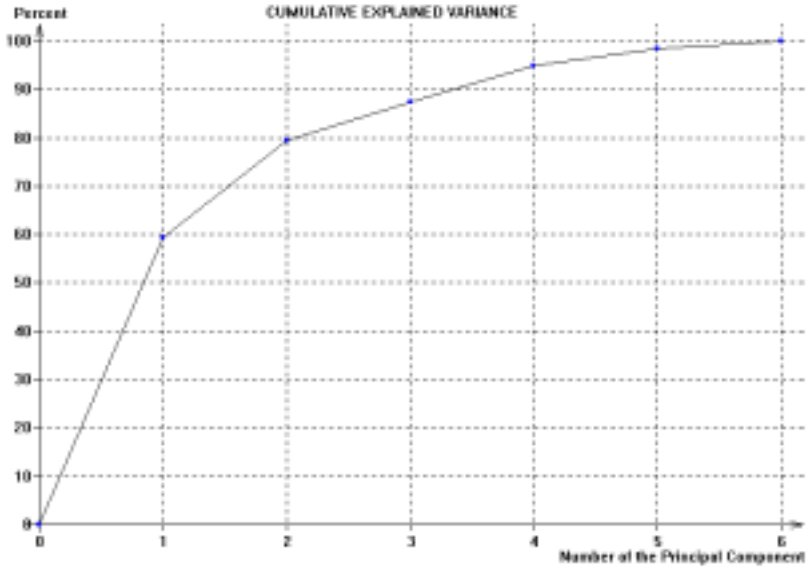
№	Среднее	Min	Max	Кол-во в классе
1	3.02	0.82	4.98	7
2	1.10	1.10	1.10	1
3	8.16	6.47	9.23	14
4	7.45	6.96	8.38	3
5	5.13	2.59	7.16	3
6	7.93	7.11	8.75	2
7	6.37	4.99	8.44	3
8	6.33	6.33	6.33	1
9	7.67	7.67	7.67	1
10	7.12	6.93	7.31	2
11	4.60	2.45	6.44	7
12	7.32	7.32	7.32	1
13	6.34	6.18	6.49	2

К сожалению, как видно из приведенного ниже графика, одно и тоже значение зависимой переменной может принадлежать разным классам. Поэтому предсказывать принадлежность страны к определенному классу по значению зависимой переменной в общем случае нельзя.Это можно сделать только на некоторых интервалах. Кроме того, можно попытаться предсказать, в какие классы страна не попадет.



3. Построение главных компонент для объясняющих переменных

Анализируя результаты расчета видно, что доля объясненной дисперсии ведет себя следующим образом:



Соответственно, первые две главные компоненты объясняют около 80% общего разброса. Точное значение, полученное из таблицы, приведенной ниже, составляет 79.41%

Главная компонента	Дисперсия	%	Накопленный разброс, %
1	3.56	59.83	59.83
2	*1.20*	*20.03*	***79.41***
3	0.47	7.90	87.31
4	0.46	7.70	95.01
5	0.20	3.41	98.42

6	0.09	1.58	100.00

Коэффициенты линейного уравнения, отражающего зависимость между значениями главных компонент их значениями исходных переменных, могут быть получены из матрицы значений главных компонент (Loads).

$$\begin{cases} z_1 = 0.48 \cdot x_{3.36} + 0.44 \cdot x_{3.46} + 0.06 \cdot x_{3.48} + 0.45 \cdot x_{8.05} + 0.50 \cdot x_{8.36} + 0.34 \cdot x_{8.40} \\ z_2 = 0.13 \cdot x_{3.36} - 0.11 \cdot x_{3.46} + 0.85 \cdot x_{3.48} + 0.04 \cdot x_{8.05} + 0.15 \cdot x_{8.36} - 0.47 \cdot x_{8.40} \\ z_3 = 0.11 \cdot x_{3.36} + 0.33 \cdot x_{3.46} - 0.45 \cdot x_{3.48} + 0.11 \cdot x_{8.05} + 0.11 \cdot x_{8.36} - 0.81 \cdot x_{8.40} \\ z_4 = 0.03 \cdot x_{3.36} - 0.70 \cdot x_{3.46} - 0.18 \cdot x_{3.48} + 0.69 \cdot x_{8.05} + 0.04 \cdot x_{8.36} - 0.09 \cdot x_{8.40} \\ z_5 = 0.85 \cdot x_{3.36} - 0.30 \cdot x_{3.46} - 0.12 \cdot x_{3.48} - 0.37 \cdot x_{8.05} - 0.19 \cdot x_{8.36} - 0.01 \cdot x_{8.40} \\ z_6 = 0.14 \cdot x_{3.36} + 0.33 \cdot x_{3.46} + 0.15 \cdot x_{3.48} + 0.42 \cdot x_{8.05} - 0.82 \cdot x_{8.36} + 0.01 \cdot x_{8.40} \end{cases}$$

Коэффициенты обратной зависимости X от главных компонент могут быть получены из матрицы корреляций между переменными и главными компонентами:

$$\begin{cases} x_{3.36} = 0.91 \cdot z_1 + 0.14 \cdot z_2 + 0.08 \cdot z_3 + 0.03 \cdot z_4 + 0.38 \cdot z_5 + 0.04 \cdot z_6 \\ x_{3.46} = 0.83 \cdot z_1 - 0.12 \cdot z_2 + 0.23 \cdot z_3 - 0.47 \cdot z_4 - 0.14 \cdot z_5 + 0.10 \cdot z_6 \\ x_{3.48} = 0.11 \cdot z_1 + 0.93 \cdot z_2 - 0.31 \cdot z_3 - 0.12 \cdot z_4 - 0.06 \cdot z_5 + 0.05 \cdot z_6 \\ x_{8.05} = 0.85 \cdot z_1 + 0.05 \cdot z_2 + 0.07 \cdot z_3 + 0.47 \cdot z_4 - 0.17 \cdot z_5 + 0.13 \cdot z_6 \\ x_{8.36} = 0.95 \cdot z_1 + 0.16 \cdot z_2 + 0.08 \cdot z_3 + 0.03 \cdot z_4 - 0.09 \cdot z_5 - 0.25 \cdot z_6 \\ x_{8.40} = 0.65 \cdot z_1 - 0.51 \cdot z_2 - 0.56 \cdot z_3 - 0.06 \cdot z_4 - 0.01 \cdot z_5 + 0.00 \cdot z_6 \end{cases}$$

Как видно из корреляционной матрицы, первая главная компонента наиболее сильно связана с четырьмя показателями: коррупция, социальная интеграция, средняя продолжительность жизни и качество жизни. Связь со всеми показателями положительна, т.е. при увеличении показателей, значение главной компоненты увеличивается.

	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
v3.36	**0.91**	0.14	0.08	0.02	0.38	0.04
v3.46	**0.83**	-0.12	0.23	-0.47	-0.14	0.10
v3.48	0.11	0.93	-0.31	-0.12	-0.06	0.05
v8.05	**0.85**	0.05	0.07	0.47	-0.17	0.13
v8.36	**0.95**	0.16	0.08	0.03	-0.09	-0.25
v8.40	0.65	-0.51	-0.56	-0.06	-0.01	0.00

На первый взгляд казавшаяся положительная связь главной компоненты и коррупции на самом деле таковой не является: чем выше значение показателя коррупции, тем ниже ее уровень

в стране (правильнее назвать данный показатель «отсутствие коррупции»). Таким образом, первую главную компоненту можно интерпретировать как интегральный показатель социально-политической ситуации в стране, которая выражается в социальной интеграции общества, отсутствия в нем коррупции, высокой продолжительности и качестве жизни. Необходимо отметить, что показатель v8.05 «Средняя продолжительность жизни» достаточно тесно связан с показателями v3.26 «Коррупция» и v8.36 «Качество жизни» (соответствующие коэффициенты корреляции, представленные в расчетах больше 0.82).

Вторая главная компонента практически полностью объясняется показателем v3.48 «Серьезные преступления», поэтому ее название можно не менять.

Рассчитанный коэффициент ранговой корреляции Спирмена между значениями первой главной компоненты и значениями зависимой переменной равен 0.80. Такое большое значение данного коэффициента позволяет говорить о наличии устойчивой связи между рангами первой главной компоненты и зависимой переменной, т.е. упорядочивание, произведенное по значениям главной компоненты достаточно точно отражает упорядочивание зависимой переменной, и, значит, можно использовать классический линейную модель многомерной регрессии для оценки зависимой переменной не только через объясняющие переменные, но и через главные компоненты. Логично будет предположить, что в соответствующем уравнении регрессии по главным компонентам, значимыми окажутся максимум три первые главные компоненты, чей вклад в общую дисперсию превышает 87%.

4. Модель многомерной линейной регрессии для зависимой переменной.

1. Модель в общем виде (зависимость от исходных признаков)

Модель многомерной линейной регрессии от всех исходных переменных с константой приведена в приложении 7. Графа «Вероятность» показывает вероятность того, что коэффициент окажется незначимо отличным от 0. Из уравнения видно, что незначимыми при критическом уровне в 5% являются константа и переменные v3.46, v8.05 и v8.40. Достаточно высокий коэффициент детерминации свидетельствует о хорошем качестве уравнения, которое в целом является значимым (т.е. все коэффициенты значимо отличны от 0 – см. значение F – статистики).

Показатели R^2 и R^2(adj.) отражают долю объясненной дисперсии, т.е степень тесноты связи между зависимой переменной и объясняющими переменными. Чем данный показатель выше, тем меньше будет разница между фактическими значениями зависимой переменной и значениями, полученными из уравнения регрессии.

Для построения модели, в котором все бы параметры оказались бы значимы, необходимо поочередно исключать из модели переменные, коэффициенты которых незначимо отличаются от 0, начиная с того из них, чья t – статистика наиболее близка к 0. Так как первым претендентом на исключение является свободный член уравнения, то можно пойти двумя путями:

1. Строить линейную множественную модель регрессии без свободного члена. Тогда у коэффициента детерминации не будет никакой интерпретации, т.к. он уже не всегда будет отражать долю объясненной дисперсии. Данный показатель уже нельзя будет применять для сравнения различных вариантов модели регрессии, но, в результате, все переменные модели будут значимы. Построение модели по этому методу представлены на стр. 11 – 13 приложения.
2. Оставить в модели свободный член, но не интерпретировать его. Такой подход лишен недостатков предыдущего способа, но коэффициенты, полученные в результате расчетов будут смещены по сравнению с полученными первым способом. Результаты построения модели данным способом представлены на стр. 7 – 10 приложения.

Какой метод выбрать – это целиком зависит от целей исследования. Проведенные расчеты обеими способами показали, что состав переменных практически не меняется. Если руководствоваться критерием максимизации коэффициента детерминации, то лучший вариант представлен на стр. 12 (без константы). В нем нет константы и не все коэффициенты являются значимыми, что не позволяет считать этот результат наилучшей моделью.

Если определять значимость коэффициентов при 6.3% уровне, то набор оставшихся переменных в обоих вариантах совпадает. Исключение переменной v8.40 из модели на стр. 9 делает все коэффициенты модели (кроме константы) значимо отличными от 0 практически при любом уровне значимости, но взамен происходит уменьшение R^2 на 0.01. Что лучше – зависит опять же от целей анализа.

Принимая за окончательные модели, представленные на стр. 9 и 13, можно сделать следующие выводы: уровень личной безопасности в стране определяется коррупцией, серьезными преступлениями, качеством жизни и злоупотреблением алкоголем. От всех переменных, кроме серьезных преступлений, показатель личной безопасности зависит положительно, т.е. чем лучше страна по параметру, тем выше в ней объясняемый показатель. Отрицательная зависимость личной и имущественной безопасности от серьезных преступлений вызывает некоторые подозрения в неправильности модели, т.к. ожидаемый результат был обратный: чем меньше преступлений, тем безопасней жить. Полученную же зависимость можно попытаться объяснить следующим образом: чем меньше в стране серьезных преступлений, тем меньше люди сами думают о своей безопасности и потенциально, уровень безопасности в стране низкий. Вполне возможно, что на результат повлияли гораздо большие единицы измерения

данного показателя. Для более детального объяснения необходимо иметь больше информации о способах сбора и представления данных показателей.

2. Модель от главных компонент

Другой способ построения зависимости переменной «Личная безопасность» – это проведение расчетов не по исходным переменным, а по полученным главным компонентам.

v3.45 = **6.616 + 1.067 * PC_1 - 0.416 * PC_2** + 0.285 * PC_3 - 0.138 * PC_4 + 0.541 * PC_5 - 0.660 * PC_6
t-стат. **(40.75) (13.18)** **(-2.99)** (1.28) (-0.61) (1.60) (-1.33)
ст. ош. (0.151) (0.081) (0.139) (0.222) (0.225) (0.338) (0.497)

Качество уравнения регрессии:

R^2 (adj.) = 0.80 R^2 = 0.825 F = 31.497 Число степеней свободы = 40

$F_{крит}(6,40, 0.05) = 2.34$ $t_{крит}(40,0.05) = 2.021$ при уровне значимости 0.05

Так как главные компоненты некоррелированы, то можно сразу откинуть все незначимые переменные.

v3.45 = **6.616 + 0.003 * PC_1 - 0.416 * PC_2**
t-стат. **(39.70) (12.84)** **(-2.91)**
ст. ош. (0.155) (0.083) (0.143)

Качество уравнения регрессии:

R^2 (adj.) = 0.79 R^2 = 0.797 F = 86.66 Число степеней свободы = 44

$F_{крит}(6,44, 0.05) = 2.34$ $t_{крит}(44,0.05) = 2.021$ при уровне значимости 0.05

Как и предполагалось выше, в модели оказались значимы лишь первые две главные компоненты, которые к тому же и легко интерпретируемы. Полученная зависимость от главных компонент так же обладает достаточно хорошими прогностическими возможностями, но коэффициент детерминации несколько ниже, чем в модели от исходных показателей, поэтому для анализа лучше использовать линейную многомерную модель регрессии по исходным переменным.