САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ – ПРОЦЕССОВ УПРАВЛЕНИЯ
КАФЕДРА МАТЕМАТИЧЕСКОЙ ТЕОРИИ МИКРОПРОЦЕССОРНЫХ СИСТЕМ УПРАВЛЕНИЯ

Горбунова Мария Николаевна

Выпускная квалификационная работа бакалавра

Анализ и прогнозирование денежных доходов населения

010400

Прикладная математика и информатика

Заведующий кафедрой,
доктор физ.-мат. наук,
профессор Малафеев О. А.

Научный руководитель,
доктор физ.-мат. наук,
профессор Потапов Д. К.

Рецензент,
кандидат физ.-мат. наук,
доцент Колпак Е. П.

Санкт-Петербург
2016

Оглавление

Введение

К числу наиболее значимых направлений исследования социально-экономического положения государства относится анализ доходов населения. Грамотный анализ денежных доходов населения и их динамики является одним из основных моментов исследования уровня жизни населения страны.

Изучение экономического поведения и динамики населения позволяет получить объективную информацию об условиях жизни населения, покупательной способности доходов населения, степени дифференциации доходов и уровне бедности. Таким образом, статистические данные о доходах и их анализ необходимы государственным органам, принимающим решения, для определения мер адекватной экономической, налоговой и социальной политики.

В современном мире для принятия такого рода решений необходимо также направлять анализ в будущее и использовать средства прогнозирования. Это помогает выявить наиболее эффективные варианты долгосрочных, среднесрочных и текущих путей развития, объясняет причины основных направлений экономической политики, предсказывает результаты принимаемых решений. Поэтому прогнозирование уровня денежных доходов позволяет точнее делать выводы и является важным фактором, влияющим на достижения поставленных целей. Этим и определяется актуальность выбранной темы исследования.

Цель работы

Целью работы является построение и анализ эконометрической модели прогнозирования денежных доходов населения на примере г. Санкт-Петербурга.

Постановка задачи исследования

Для работы взята тема «Анализ и прогнозирование денежных доходов населения».

Имеются данные о среднедушевых доходах населения (в месяц; в рублях) за 1998-2014 годы. Были собраны статистические данные по пяти факторам за этот же период, которые, теоретически, влияют на выбранную выше величину.

В ходе исследования необходимо изучить теоретические основы статистического анализа денежных доходов населения;

Проанализировать структуру доходов населения, факторы, влияющие на доходы населения;

Подтвердить правильность предположения о влиянии данных факторов с использованием математической модели и статистических данных. В итоге чего выявляется статистическая значимость выбранных факторов.

Составить и проанализировать модель, проверить ее адекватность модели реальной ситуации на данных в среде Statistica.

Осуществить краткосрочный прогноз.

Методы исследования: динамические ряды, группировка, дисперсионный анализ, корреляционно-регрессионный анализ.

Глава 1. Денежные доходы населения

В данной главе будет изучена предметная область, исследована структура доходов населения и выбраны факторы, теоретически влияющие на доходы населения.

§1. Определение понятия денежных доходов населения

Под доходами населения понимается сумма денежных средств и материальных благ, полученных или произведенных домашними хозяйствами за определенный промежуток времени.

Различают денежные и натуральные доходы. *Денежные доходы* населения складываются из поступлений денег в форме оплаты труда, социальных трансфертов, доходов от собственности, предпринимательской деятельности, продажи продукции личного подсобного хозяйства и др. — алиментов, гонораров, благотворительной помощи и т.д. Денежные доходы населения подразделяются на первичные и располагаемые. *Первичные доходы* населения включают все полученные поступления, а *располагаемые доходы* населения — результат перераспределительных процессов. Они рассчитываются с помощью добавленных к первичным доходам социальных трансфертов и вычитания обязательных платежей и сборов.

К доходам населения относятся также средства, взятые в долг. В связи с этим выделяют конечные и общие доходы населения. *Конечные доходы* населения — это располагаемые доходы плюс чистые долги населения. *Номинальные доходы* включают всю сумму конечных доходов.

Для оценки уровня и динамики доходов населения следует различать номинальные, располагаемые и реальные доходы. Номинальные доходы характеризуют уровень денежных доходов независимо от налогообложения и изменения цен. Располагаемые доходы – это номинальные доходы за вычетом налогов и других обязательных платежей, т.е. средства используемые населением на потребление и сбережение. Реальные доходы

характеризуют номинальные доходы с учетом изменения розничных цен и тарифов.

В данной модели для анализа и прогнозирования будут использоваться показатели среднедушевого дохода. *Среднедушевой денежный доход* определяется путем деления суммы конечных доходов (т.е. номинальных доходов) на численность населения субъекта.

В качестве факторов, оказывающих непосредственное влияние на величину доходов населения, можно рассматривать сами источники формирования доходов, среди которых особо выделяют уровень заработной платы и социальные трансферты, а также такие социально-экономические показатели, как индекс потребительских цен, валовой региональный продукт, численность занятых в экономике и др.

Глава 2. Математический аппарат

В качестве математического аппарата будет использоваться регрессионный анализ - статистический метод исследования зависимости случайной величины y от независимых факторов x_i $(i = 1, \dots, p)$.

Отбор факторов, влияющих на зависимую переменную, является одним из важных моментов разработки модели.

Прежде всего, факторы проверяются на наличие линейной корреляции между ними, признаком наличия которой является условие:

$$\left|r_{x_i x_j}\right| \geq r_{\text{кр}},$$

где $r_{\text{кр}}$ выявленно эмпирическим путем и равняется $\sim 0{,}8 - 0{,}9$.

Существование тесной взаимосвязи между факторами приводит к получению некачественных параметров модели. От таких факторов избавляются методом исключения.

Не желательно вводить в модель большое количество независимых переменных, так как это может отрицательно повлиять на выявление закономерностей.

Различают следующие виды уравнений множественной регрессии: линейные, нелинейные, сводящиеся к линейным, и нелинейные, не сводящиеся к линейным (внутренне нелинейные).

Наиболее простая из моделей множественной регрессии - линейная модель:

$$y = a_0 + \sum_{i=1}^{p} a_i x_i + \varepsilon$$

где a_i - коэффициенты модели;

y, x_i - значения зависимой переменной и независимой;

ε - случайная ошибка регрессионной зависимости;

p - число факторных признаков.

Для оценки параметров уравнения множественной регрессии обычно применяется метод наименьших квадратов (МНК), согласно которому следует выбирать такие значения параметров a_i, при которых сумма квадратов отклонений фактических значений результативного признака от теоретических значений минимальна, т. е.:

$$S = \sum (\hat{y}_i - y_i) \rightarrow min$$

И тогда:

$$S = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_1 - \cdots - a_p x_p\right) = S(a_0, a_1, \ldots, a_p)$$

Для оценки качества полученного уравнения регрессии можно использовать коэффициент детерминации R^2. Он показывает, какая доля результативного признака изменяется под действием факторных признаков. Если значение коэффициента близко к единице, то качество уравнения регрессии высокое.

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)}{\sum_{i=1}^{n} (y_i - \bar{y}_i)}$$

Оценка статистической значимости уравнения регрессии осуществляется с помощью F-критерия Фишера:

$$F = \frac{\dfrac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{p}}{\dfrac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n - p - 1}} = \frac{R^2}{1 - R^2} \frac{n - p - 1}{p}$$

Где n – число измерений, p – число независимых переменных.

Если оно превосходит табличное значение, то данную модель можно использовать для анализа и прогнозирования.

Коэффициент множественной корреляции R – показывает, насколько зависимая переменная тесно связана с независимыми факторами. Чем ближе величина R к единице, тем теснее данная связь, тем лучше зависимость.

$$R = \sqrt{R^2}$$

Также определить значимость отобранных независимых факторов можно t-критерием Стьюдента. Для этого нужно найти значение $t_{\text{табл}}$ и сравнить его со значениями t-статистики.

$$t_{a_i} = \frac{a_i}{s_{a_i}}$$

Где s_{a_i} – стандартная ошибка коэффициентов регрессии, определяемая как:

$s_{a_i} = s_{\text{ост}}|(X'X)^{-1}|_{ii}$, а $s_{\text{ост}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{(n-p-1)}$

- Если $t_{ai} < t_{\text{табл}}$, то фактор X_i статистически не важен;
- Если $t_{ai} > t_{\text{табл}}$, то фактор X_i значим и информативен;

Оценить степень влияние отдельных независимых факторов на результативный признак можно при помощи коэффициентов эластичности, β и Δ.

- Частный коэффициент эластичности - показывает, на сколько % изменится среднее значение результативного признака, если среднее значение конкретного факторного признака изменится на 1%

$$Э_i = a_i * \frac{\bar{X}_i}{\bar{Y}}$$

- β-коэффициент - показывает, на какую величину изменится среднее квадратическое отклонение результативного признака, если среднее квадратическое отклонение конкретного факторного признака изменится на 1 единицу.

$$\beta_i = a_i * \frac{\sigma_{xi}}{\sigma_y}$$

- Δ-коэффициент - показывает удельный вес влияния конкретного факторного признака в совместном влиянии всех факторных признаков на результативный показатель.

$$\Delta_i = \beta_i * \frac{r_{xiy}}{R^2}$$

Проверка выполнения предпосылок МНК

Использование вычислительной процедуры по методу наименьших квадратов с целью получения оценок коэффициентов модели предполагает выполнение ряда условий:

- независимые переменные представляют собой неслучайный набор чисел, их средние значения и дисперсии конечны;
- случайные ошибки ε_j — имеют нулевую среднюю и конечную дисперсию

$$M(\varepsilon_j) = 0; M(\varepsilon_j \varepsilon_j) = \sigma_\varepsilon^2 < \infty$$

- между независимыми переменными отсутствует корреляция и автокорреляция;
- случайная ошибка не коррелированна с независимыми переменными;
- случайная ошибка подчинена нормальному закону распределения.

Глава 3. Результаты расчетов на примере статистических данных по Санкт-Петербургу

По данным, представленным в таблице 1, изучается зависимость средне душевого денежного дохода в мес, Y (руб.) от следующих переменных: X_1 - среднемесячная номинальная начисленная заработная плата (руб.), X_2 – среднемесячные назначенные пенсии (руб.), X3 – экономически активное население (тыс. чел.), X4 – валовой региональный продукт (млн. руб.)

Год	Y	X1	X2	X3	X4	X5	X6
1998	1178,8	1147,9	429,4	2316	92029	178	10,1
1999	1838	1687,3	546,9	2426	150727,3	141,1	10,2
2000	2589,6	2511,5	879,8	2417	205399,1	123,5	6,3
2001	3468	3695,3	1284	2437	275442,6	118,1	3,9
2002	4497	5434,7	1663,5	2518	367198	114,7	3,4
2003	6851	6467,5	1969	2516	435683	112,2	4,1
2004	8855,1	7931,1	2265,4	2553	518885	112,7	2,7
2005	12266,1	10133,9	2846,2	2570	666393	112	2,2
2006	14097,7	13033,2	3302,4	2645	825102	108,67	2,4
2007	16876,4	17552	4364,9	2709	1119660	108,44	2,1
2008	17648,7	22473,4	5354,6	2704	1431840	113,13	2
2009	22132,6	23884,4	7249,4	2718	1473348	115,83	4,1
2010	24593,8	27189,5	8824,5	2660	1699486	106,44	2,7
2011	25994,7	29522	9573,7	2858	2071757	106,86	2
2012	27795	32930	10060	2896	2291993	106,01	1,1
2013	31407	36848	10547	2849	2496549	105,04	1,5
2014	34724	40697	11470	2593,1	2652050	105,46	1,4

Для использования методов регрессионного анализа необходимо убедиться, что среди факторов нет линейно связанных. Иначе возникает явление мультиколлинеарности и происходит искажение коэффициентов регрессии. Чаще всего мультиколлинеарность является следствием того, что независимые факторы характеризуют одно свойство изучаемого процесса или являются частями одного признака. Для исследования выбранных признаков на мультиколлинеарность будет использоваться метод корреляции. Для этого нужно построить и проанализировать матрицу парных корреляций. Ее можно построить, используя инструмент Сервис ->

Анализ данных -> Корреляция в Excel. Считается, что показатели линейно зависимы, если их парный коэффициент корреляции превосходит по абсолютной величине 0,8

	Столбец 1	Столбец 2	Столбец 3	Столбец 4	Столбец 5	Столбец 6	Столбец 7
Столбец 1	1						
Столбец 2	0,993524	1					
Столбец 3	0,985854	0,990661	1				
Столбец 4	0,834584	0,824245	0,820913	1			
Столбец 5	0,988459	0,997408	0,991461	0,833701	1		
Столбец 6	-0,61142	-0,58063	-0,56686	-0,69117	-0,57395	1	
Столбец 7	-0,7047	-0,67826	-0,64968	-0,7523	-0,67225	0,899854	1

Рис. 1

Далее мультиколлинеарность устраняется путем удаления одного из коррелирующих признаков, причем удаляется тот признак, у которого коэффициент корреляции с результативным признаком меньше. Из данных на рис. 1 видно, что коррелирующими факторами являются ВРП (X4), з/п (X1) и пенсии (X2). Следовательно, оставляем фактор X3, X5, X6

Далее на закладке Данные выберем строку Анализ данных и в качестве инструмента данных - Регрессия - ок. В открывшемся окне Регрессии зададим Входной интервал Yи X (рис.2,3).

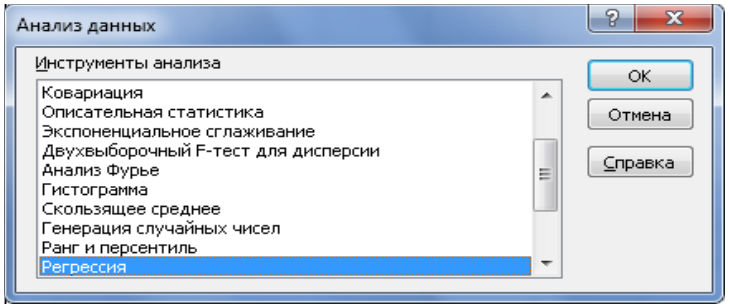

Рис. 2. Окно Анализ данных

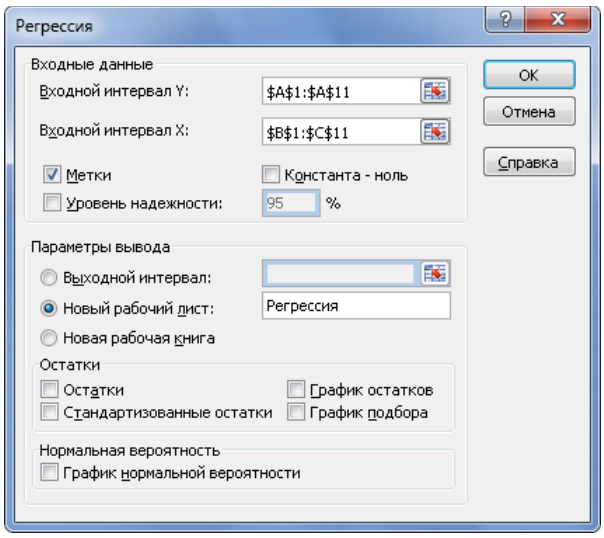

Рис. 3. Окно Регрессия.

Получим результаты регрессионного анализа на новом листе Регрессия (Рис. 4)

Регрессионная статистика								
Множественный R	0,846024							
R-квадрат	0,715756							
Нормированный R-квадра	0,650161							
Стандартная ошибка	6563,261							
Наблюдения	17							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*ачимость F*			
Регрессия	3	1,41E+09	4,7E+08	10,9118	0,000742			
Остаток	13	5,6E+08	43076392					
Итого	16	1,97E+09						
	Коэффициен	*артная о*	*татистик*	*-Значение*	*ижние 95%*	*ерхние 95%*	*ижние 95,0*	*рхние 95,0%*
Y-пересечение	-115393	47062,96	-2,45189	0,029109	-217067	-13719,9	-217067	-13719,9
Переменная X 1	47,07975	14,96337	3,146334	0,007726	14,75336	79,40614	14,75336	79,40614
Переменная X 2	106,0835	209,1701	0,507164	0,620533	-345,801	557,9681	-345,801	557,9681
Переменная X 3	-1319,18	1498,876	-0,88011	0,394776	-4557,31	1918,941	-4557,31	1918,941

Рис. 4. Лист Регрессия.

По данным регрессионной статистики мы получили следующие данные:

Множественный R - это $\sqrt{R^{2,}}$ где R^2 - коэффициент детерминации.

R-квадрат - это R^2. В нашем примере значение R^2=0,715756 свидетельствует о том, что изменения зависимой переменной Y на 71,5756%

можно объяснить изменениями включенных в модель объясняющих переменных – X3,X5,X6 (экономически активное население, уровень безработицы, ИПЦ). И на 28,4244% среднедушевые доходы зависят от других неучтённых факторов. Такое значение свидетельствует об адекватности модели.

Нормированный R-квадрат - поправленный (скорректированный по числу степеней свободы) коэффициент детерминации.

Стандартная ошибка регрессии $S=\sqrt{S^2}$, где $S^2=\sum (e_i^2/ (n-m))$ - необъясненная дисперсия (мера разброса зависимой переменной вокруг линии регрессии); n - число наблюдений (в нашем случае 7), m - число объясняющих переменных (в нашем примере равно 3).

F - расчетное значение F-критерия Фишера. Если нет табличного значения, то для проверки значимости уравнения регрессии в целом можно посмотреть Значимость F. На уровне значимости $\alpha=0,05$ уравнение регрессии признается значимым в целом, если Значимость F<0,05, и незначимым, если Значимость F≥0,05. В нашем случае значимость 0,00074.

Для нашего примера имеем следующие значения:

Таблица 2

Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	3	1410122418	470040805,8	10,91179604	0,0007417
Остаток	13	559993098,7	43076392,21		
Итого	16	1970115516			

В нашем случае расчетное значение F-критерия Фишера составляет 10,912. Значимость F меньше 0,05. Таким образом, полученное уравнение в целом значимо.

В последней таблице приведены значения параметров (коэффициентов) модели, их стандартные ошибки и расчетные значения t-критериев Стьюдента для оценки значимости отдельных параметров модели.

Таблица 4

	Коэффициент	тандартная ошиб	t-статистика	P-Значение	Нижние 95%	ерхние 95%	жние 95,0	рхние 95,0
Y-пересечение	-115393	47062,9626	-2,451891708	0,029109317	-217066,6371	-13719,9	-217067	-13719,9
Переменная X 1	47,07975	14,96336903	3,146333539	0,007725979	14,7533564	79,40614	14,75336	79,40614
Переменная X 2	106,0835	209,1701028	0,50716389	0,620533414	-345,8010108	557,9681	-345,801	557,9681
Переменная X 3	-1319,18	1498,875666	-0,880114701	0,394776349	-4557,306516	1918,941	-4557,31	1918,941

Анализ данной таблицы позволяет сделать вывод о том, что на уровне значимости α=0,05 значимым оказывается лишь коэффициент при факторе X_3, так как лишь для него Р-значение меньше 0,05. Таким образом, фактор X_5 b X6 не существенны и их включение в модель не целесообразно. Поскольку коэффициент регрессии в экономических исследованиях имеют четкую экономическую интерпретацию, то границы доверительного интервала для коэффициента регрессии не должны содержать противоречивых результатов, как например, - 345,8≤ b_1≤ 557,96. Такого рода запись указывает, что истинное значение коэффициента регрессии одновременно содержит положительные и отрицательные величины и даже ноль, чего не может быть. Это также подтверждает вывод о статистической незначимости коэффициентов регрессии при факторах X5 и X6. Таким образом, целесообразно исключить несущественные факторы. Но мы оставим эти факторы факторы, так как в случае его исключения, модель не будет многофакторной. Поэтому мы будем иметь ввиду, что факторы X6,X_5 малозначим и построим уравнение зависимости Y от значимой объясняющей переменной X_2 и незначимых.

.

Оценим точность и адекватность полученной модели.

Согласно проведенной регрессионной статистики мы видим следующие результаты:

1. Коэффициент множественной корреляции (*множественный R*) равен 0,846. Следовательно, связь между факторами весьма тесная (по шкале Чудока)

2. *Значение* $R^2=0,7158$ свидетельствует о том, что вариация зависимой переменной Y в основном можно объяснить вариацией включенных в модель объясняющих переменных X3, X5, X6. Это свидетельствует об адекватности модели.