Ивановская Вероника Юрьевна
канд. экон. наук, доцент
Плотникова Юлия Александровна
канд. физ.-мат. наук, доцент
Ивановская Алена Леонидовна
магистрант
ФГБОУ ВО «Вологодская государственная
молочнохозяйственная академия им. Н.В. Верещагина»
г. Вологда, Вологодская область

ВЫБОР НАИЛУЧШЕЙ МОДЕЛИ, ОПИСЫВАЮЩЕЙ ЗАВИСИМОСТЬ ИНДЕКСА ЧЕЛОВЕЧЕСКОГО РАЗВИТИЯ ОТ ВЛИЯЮЩИХ ФАКТОРОВ

Аннотация: *в статье проводится анализ зависимости индекса человеческого развития от 6 факторов с использованием моделей парной регрессии. В работе проанализировано понятие «многофакторные модели», представлена цель исследования.*

Ключевые слова: *факторы, модель парной регрессии, коэффициент детерминации, статистика Стьюдента.*

Многофакторные модели – основа прогнозирования экономических результативных признаков, а также средство оценки влияния каждого взятого в отдельности фактора в объяснении вариации результатов, поэтому особое место в эконометрике уделяется выбору наилучшей модели. Одна из проблем состоит в том, что заранее нельзя определить наилучший тип модели; нужно изучить несколько моделей перед тем, как выбрать лучшую из них.

Основная цель работы – провести анализ зависимости индекса человеческого развития (величина y) от 6 факторов (x_1 – ВВП, x_2 – расходы на конечное потребление в текущих ценах, x_3 – расходы домашних хозяйств, x_4 – расходы на текущее потребление в текущих ценах, x_5 – суточная калорийность питания населения, x_6 – ожидаемая продолжительность жизни при рождении).

Сначала было изучено влияние факторов x_i ($i = 1, 2, \ldots, 6$) на величину y и был отобран фактор x, наиболее сильно влияющий на y. После чего было построено нескольких нелинейных моделей, описывающих зависимость y от выбранного фактора x, после чего из них была выбрана в определенном смысле наилучшая. Для исследования были взяты данные по 25 странам за 1997 г. [1]. Расчеты проводились с использованием MS Excel. При подготовке работы использовались материалы из учебных пособий [2; 3].

Отбор фактора x, наиболее сильно влияющего на индекс человеческого развития, проводился тремя способами: с использованием моделей парной регрессии; с использованием модели множественной регрессии; с применением процедуры последовательного исключения факторов. Во всех случаях выбранным фактором оказалась ожидаемая продолжительность жизни при рождении в 1997 г. (фактор x_6).

Отбор фактора x с использованием моделей парной регрессии проводился следующим образом. Сначала по исходным данным были построены линейные модели парной регрессии, связывающие результат y с каждым из факторов x_i ($i = 1, 2, \ldots, 6$). На рис. 1–2 представлены поля корреляции (поля наблюдений) индекса человеческого развития с каждым из факторов. Уравнения регрессии, а также коэффициенты линейной парной корреляции приведены в таблице 1.

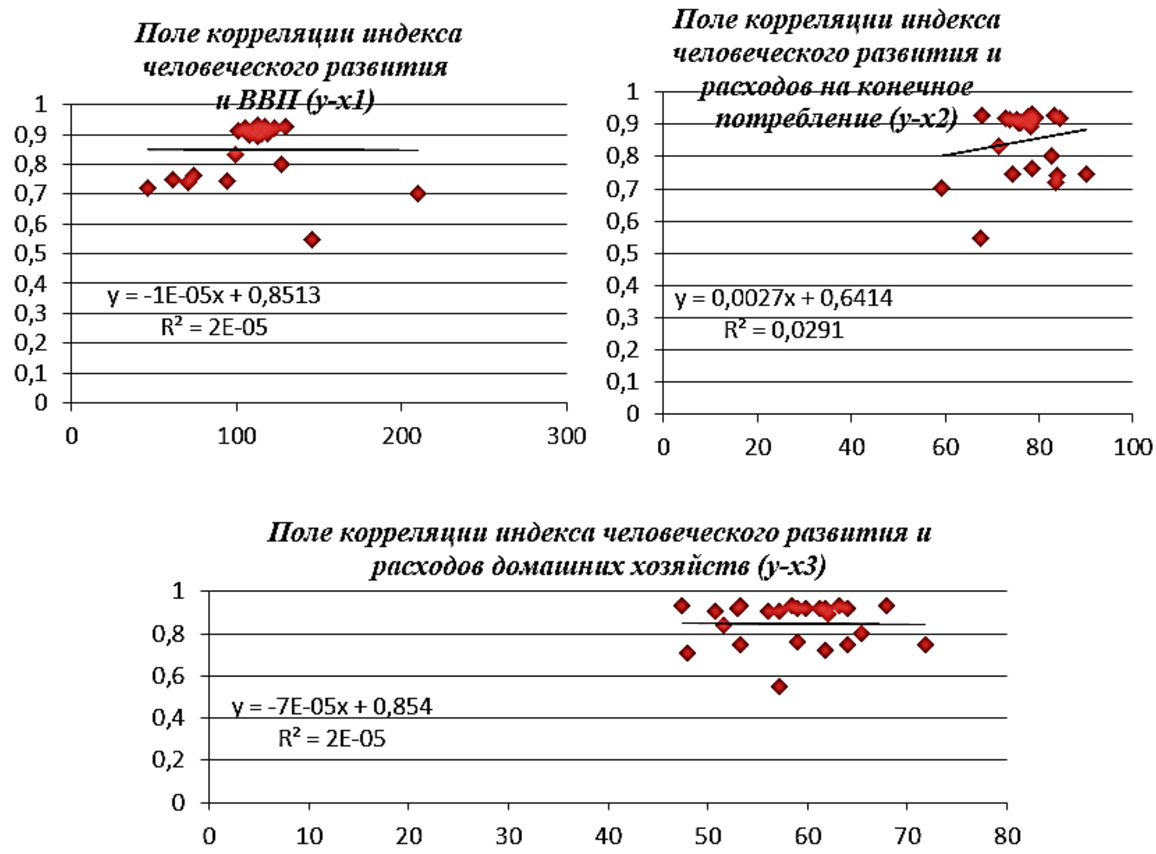
Поле корреляции индекса человеческого развития и ВВП (y-x1)
y = -1E-05x + 0,8513
R² = 2E-05
Поле корреляции индекса человеческого развития и расходов на конечное потребление (y-x2)
y = 0,0027x + 0,6414
R² = 0,0291
Поле корреляции индекса человеческого развития и расходов домашних хозяйств (y-x3)
y = -7E-05x + 0,854
R² = 2E-05

Рис. 1

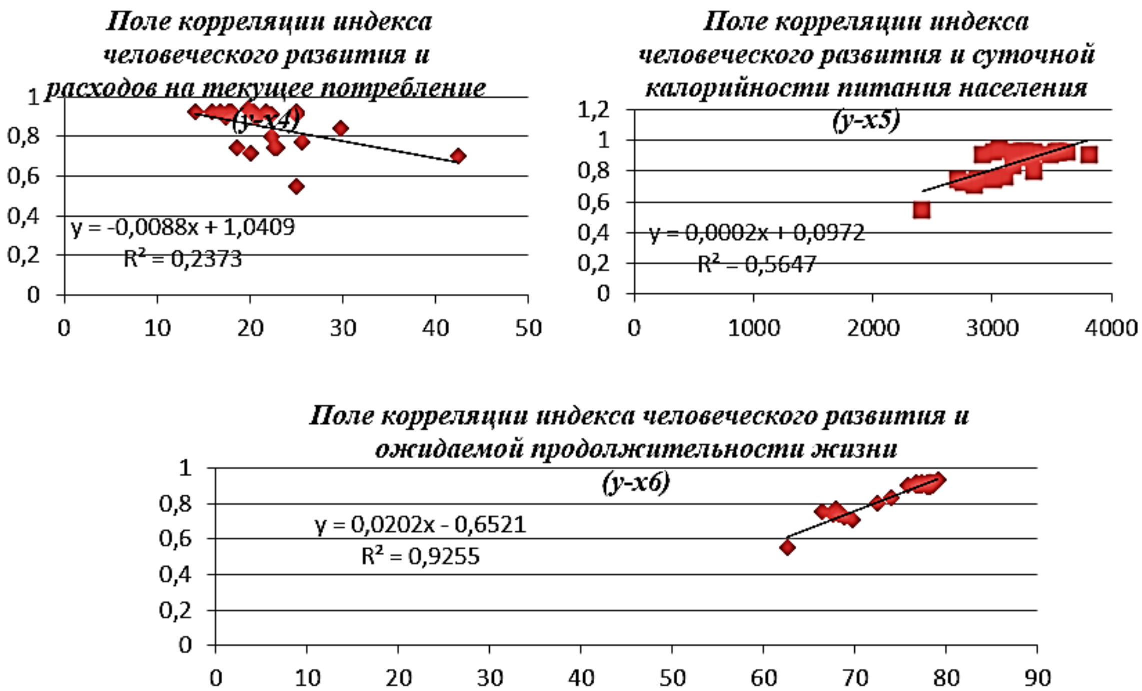
Поле корреляции индекса человеческого развития и расходов на текущее потребление (y-x4)
y = -0,0088x + 1,0409
R² = 0,2373
Поле корреляции индекса человеческого развития и суточной калорийности питания населения (y-x5)
y = 0,0002x + 0,0972
R² = 0,5647
Поле корреляции индекса человеческого развития и ожидаемой продолжительности жизни (y-x6)
y = 0,0202x - 0,6521
R² = 0,9255

Рис. 2

Таблица 1

Линейные модели парной регрессии и значения коэффициентов детерминации для этих моделей

Зависимость индекса человеческого развития от:	Линейная модель	Коэффициент детерминации
1) ВВП 1997 г. (в % 1990 г.);	$y = –0,00001x_1 + 0,8513$	$R^2 = 0,00002$
2) расходов на конечное потребление в текущих ценах (в % к ВВП);	$y = 0,0027x_2 + 0,6414$	$R^2 = 0,0291$
3) расходов домашних хозяйств (в % к ВВП);	$y = –0,00007x_3 + 0,8540$	$R^2 = 0,00002$
4) расходов на текущее потребление в текущих ценах (в % к ВВП);	$y = –0,0088x_4 + 1,0409$	$R^2 = 0,2373$
5) суточной калорийности питания населения (в ккал на душу населения);	$y = 0,0002x_5 + 0,0972$	$R^2 = 0,5647$
6) ожидаемой продолжительности жизни (в годах) при рождении в 1997 г.	$y = 0,0202x_6 – 0,6521$	$R^2 = 0,9255$

Из таблицы 1 видно, что наибольший коэффициент детерминации 0,9255 имеет последняя модель. Это означает, что с точки зрения моделирования взаимосвязи индекса человеческого развития с каждым фактором по отдельности самой тесной оказывается связь с фактором x_6.

Отбор фактора x с использованием с использованием модели множественной регрессии происходил так. Была построена линейная модель множественной регрессии, связывающая y со всеми факторами x_i ($i = 1, 2, …, 6$) в совокупности. Модель имеет следующий вид:

$$y = –0,63627 – 0,00048\, x_1 + 0,00131\, x_2 – 0,00141\, x_3 + \\ + 0,00015\, x_4 + 0,00004\, x_5 + 0,01854\, x_6. \qquad (1)$$

Коэффициент множественной детерминации построенной модели равен 0,964. В таблице 2 приведены значения t-статистики (статистики Стьюдента), оценивающие статистическую значимость параметров множественной регрессии при каждом из факторов.

Таблица 2

Значения статистики Стьюдента для параметров множественной регрессии

Фактор	Значение t-статистики
x_1	–2,08803

x_2	0,89460
x_3	–1,14140
x_4	0,10685
x_5	2,11404
x_6	11,65416

Из таблицы видно, что значение t при факторе x_6 по модулю превосходит остальные. Это означает, что данный фактор является самым значимым с точки зрения статистики Стьюдента. Отбор фактора x с использованием процедуры последовательного исключения факторов осуществлялся следующим образом [4]. Из модели (1) последовательно исключались (по одному на каждом шаге) факторы с наименьшим по модулю значением t-статистики (выделены красным в таблице 3).

Таблица 3

Последовательное исключение факторов

Фактор	Значение t-статистики				
	шаг 1	шаг 2	шаг 3	шаг 4	шаг 5
x_1	–2,0880	–2,3171	–3,7056	–3,6969	–3,5128
x_2	0,8947	0,9192			
x_3	–1,1414	–1,2132	–0,7956		
x_4	0,1068				
x_5	2,1140	2,1899	2,3373	2,2422	
x_6	11,6541	14,0008	14,2620	14,6373	20,9519

Оставшийся фактор x_6 можно считать наиболее влияющим на результат с точки зрения процедуры последовательного исключения факторов.

Таким образом, ожидаемую продолжительность жизни при рождении в 1997 г. можно считать наиболее влияющим на индекс человеческого развития среди 6 предложенных. Представляется разумным добавить к линейной модели, описывающей зависимость y от x, некоторые нелинейные модели, а затем выбрать из всех моделей наилучшую.

Между собой сравнивались модели из шести функциональных классов (линейные, степенные, экспоненциальные, логарифмические, полиномиальные 2-й и полиномиальные 3-й степени модели). Для этого отдельно для каждого из шести классов с помощью классического метода наименьших квадратов (МНК) была определена оптимальная модель, а затем из полученных 6 моделей из разных классов была выбрана наилучшая. Выбор осуществлялся с помощью сравнения значений скорректированных коэффициентов детерминации (R^2_{adj}) построенных моделей.

Хорошо известно, что скорректированный коэффициент детерминации позволяет сравнивать с точки зрения качества модели с разным числом параметров [4]. Классический коэффициент детерминации позволяет лишь оценить качество приближения результата регрессионной зависимостью от факторов. Так, включая дополнительный фактор в модель множественной регрессии или дополнительный параметр в модель парной регрессии, мы можем только увеличить значение R^2. Так, если точки поля корреляции некоторой выборки не лежат на одной прямой или на одной параболе, то, для коэффициентов детерминации $R^2_{2\ \text{степ}}$ и $R^2_{3\ \text{степ}}$ полиномиальных моделей 2-й и 3-й степени, соответственно, имеет место неравенство $R^2_{2\ \text{степ}} < R^2_{3\ \text{степ}}$.

То есть с точки зрения коэффициента детерминации модель 3-й степени всегда окажется как минимум не хуже, чем модель 2-й степени. Но это совсем не означает, что она окажется лучшей с точки зрения возможностей анализа и прогнозирования. Скорректированный же коэффициент детерминации учитывает не только качество приближения результата регрессионной зависимостью от факторов, но и число параметров модели, беря «штраф» за введение дополнительных параметров. Напомним что R^2_{adj} связан с классическим коэффициентом детерминации R^2 следующей формулой:

$$R^2_{adj} = 1\text{-}(1\text{-}R^2)(n\text{-}1)/(n\text{-}m\text{-}1), \ (3)$$

где n – объем выборки, m – число объясняющих параметров модели. Для моделей парной регрессии значения m следующие: m = 1 для линейной, степенной, экспоненциальной и логарифмической моделей, m = 2 – для полиномиальной 2-й степени модели, m = 3 – для полиномиальной 3-й степени модели [4].

Таблица 4

Модели зависимостей индекса человеческого развития
от ожидаемой продолжительности жизни при рождении в 1997 г.
и значения скорректированного коэффициента детерминации

Тип модели	Уравнение модели	R^2	R^2_{adj}
Линейная	y = 0,0202x – 0,6521	0,9255	0,9223
Степенная	y = 0,0003$x^{1,8738}$	0,9070	0,9030
Экспоненциальная	y = 0,1224$e^{0,0259x}$	0,8981	0,8937
Логарифмическая	y = 1,4549ln x – 5,4174	0,9307	0,9277
Полиномиальная 2-й степени	y = –0,0006x^2 + 0,1092x – 3,844	0,9378	0,9321
Полиномиальная 3-й степени	y = 0,00005x^3 – 0,0114x^2 + 0,8694x – 21,675	0,9397	0,9311

В итоге наилучшей с точки зрения величины скорректированного коэффициента детерминации моделью из шести выбранных классов является полиномиальная модель 2-й степени.

Обратим внимание, что, сравнивая между собой полученные с помощью МНК полиномиальные модели 2-й степени и 3-й степени, можно видеть, что коэффициент детерминации до корректировки оказался выше (0,9397 < 0,9378) у модели 3-й степени, как и должно быть, согласно (2). Это означает, что полиномиальная модель 3-й степени лучше приближает результат y с помощью фактора x, чем такая же модель 2-й степени. Но скорректированный коэффициент детерминации R^2_{adj} оказался выше у полиномиальной модели 2-й степени. Значит, добавление в модель еще одного параметра привело, с одной стороны, к улучшению качества приближения, а с другой – к ухудшению качества модели с точки зрения ее использования для экономического анализа и прогнозирования.

Список литературы

1. Елисеева И.И. Практикум по эконометрике: Учебное пособие / И.И. Елисеева [др.]; под ред. И.И. Елисеевой. – М.: Финансы и статистика, 2003. – 192 с.

2. Плотников М.Г. Практикум по эконометрике. – Вологда – Молочное: Вологодская ГМХА, 2016. – 48 с.

3. Плотников М.Г. Теория вероятностей и математическая статистика: учебное пособие в формате электронного учебника. – Вологда – Молочное: Вологодская ГМХА, 2015. – 222 с.

4. Елисеева И.И. Эконометрика: Учебник пособие / И.И. Елисеева [др.]; под ред. И.И. Елисеевой. – М: Проспект, 2011. – 288 с.