ИСПОЛЬЗОВАНИЕ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ ДЛЯ ИССЛЕДОВАНИЯ ДЕМОГРАФИЧЕСКОЙ СИТУАЦИИ В ПЕНЗЕНСКОЙ ОБЛАСТИ

Е.А. Симакова

Территориальный орган Федеральной службы государственной статистики по Пензенской области, главный специалист-эксперт, г. Пенза, Россия

58.SimakovaEA@rosstat.gov.ru

Одной из наиболее важных проблем социально-экономического развития регионов является уменьшение численности населения. Любые экономические процессы определяются большим числом одновременно действующих факторов. В связи с этим возникает задача исследования взаимосвязи между зависимой переменной Y и несколькими независимыми переменными $X_1, X_2, \dots, X_n$. Данная задача решается с использованием множественного регрессионного анализа.

Регрессионный анализ предполагает нахождение коэффициентов при независимых переменных и построение уравнения множественной регрессии на основе исходных данных следующего вида:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip},$$

где $i, p = 1, 2, \dots, n$ [1; 2].

Рассмотрим таблицу исходных данных (таблица 1), в которую вошли статистические данные за 10 лет по следующим показателям:

предикторы:

X_1 – ожидаемая продолжительность жизни при рождении (число лет);

X_2 – коэффициенты естественного прироста (убыли) населения (на 1000 человек населения);

X_3 – коэффициенты миграционного прироста (убыли) (на 10000 человек населения);

зависимая переменная:

Y – численность населения (тысяч человек).

Таблица 1. Исходные данные

Год	X_1	X_2	X_3	Y
2014	71,6	-4,0	-3,5	1356,6
2015	72,0	-4,2	-18,9	1350,8
2016	72,4	-4,4	-18,6	1342,6
2017	73,2	-5,2	-31,6	1334,3
2018	73,0	-5,9	-55,7	1323,1
2019	73,4	-6,2	-42,6	1308,1
2020	71,1	-10,1	-20,3	1294,5
2021	69,8	-12,7	-11,8	1278,8
2022	72,1	-8,9	-26,6	1261,1
2023	72,6	-8,2	-2,6	1246,6

Проверим с помощью регрессионного анализа влияние предикторов модели X_1, X_2, X_3 на зависимую переменную Y.

В таблицах (таблица 2-4) приведены результаты выполнения регрессионного анализа в MS Excel.

Таблица 2. Регрессионная статистика

Множественный *R*	0,958
R-квадрат	0,917
Нормированный *R*-квадрат	0,875
Стандартная ошибка	13,565
Наблюдения	10

Таблица 3. Дисперсионный анализ

	df	*SS*	*MS*	*F*	Значимость *F*
Регрессия	3	12183,801	4061,267	22,070	0,001
Остаток	6	1104,104	184,017		
Итого	9	13287,905			

Таблица 4. Коэффициенты регрессии

	Коэффициенты	Стандартная ошибка	*t*-статистика	*P*-Значение
Y-пересечение	3717,756	505,059	7,361	0,000
X_1	-32,040	6,923	-4,628	0,004
X_2	17,288	2,175	7,950	0,000
X_3	-1,003	0,345	-2,911	0,027

Проведем анализ полученных результатов. Множественный *R* – это коэффициент множественной корреляции. Он выражает степень зависимости отклика от всех предикторов модели. Значение коэффициента находится в промежутке
от 0 до 1. Чем ближе значение к 1, тем лучше связь предикторов и зависимой переменной. В данной модели множественный *R* равен 0,958, что говорит о высокой тесноте связи.

Для оценки адекватности модели используется коэффициент детерминации или *R*-квадрат. Он показывает долю влияния всех предикторов модели на дисперсию зависимой переменной. Значение *R*-квадрата находится в промежутке от 0 до 1. Значение 0 указывает на то, что зависимая переменная никак не может быть объяснена предикторами. Значение 1 указывает на то, что результативный признак полностью объясняется независимыми переменными. В полученной модели *R*-квадрат равен 0,917, следовательно, уравнением регрессии объясняется 91,7% дисперсии зависимой переменной. То есть модель не учитывает только 8,3% факторов, влияющих на численность населения Пензенской области.

Проверим полученную модель по *F*-критерию Фишера. Для данной модели из 10 наблюдений, 3 оцененных параметров регрессии и уровня значимости 0,05, находим табличное значение *F*-критерия равное 4,35. Так как наблюдаемое значение *F*-критерия (22,070) больше табличного, то расхождения между вычисленными дисперсиями несущественные и носят неслучайный характер.

Для оценки качества модели следует обратить внимание на значимость *F*. Если значимость меньше 0,05, то это указывает на то, что предикторы в модели

имеют статистически значимую связь с откликом, а если значимость больше или равна 0,05, то связь – статистически незначима. В модели значимость *F* равна 0,001, следовательно, независимые переменные имеют статистически значимую связь с результативным признаком, то есть с численностью населения.

Для того, чтобы полученную модель регрессионного анализа можно было использовать для прогнозирования, нужно отдельно проанализировать значимость каждого предиктора. Это выполняется с помощью проверки значения вероятности и сравнения значений *t*-статистики с критическим значением статистики Стьюдента для каждого коэффициента регрессии.

Если *P*-значение меньше 0,05, то предиктор является статистически значимым, а если *P*-значение больше или равно 0,05, то коэффициент статистически незначим. В данном случае можно говорить о том, что переменные X_1, X_2, и X_3 являются статистически значимыми.

В данной модели из 10 наблюдений, 3 оцененных параметров регрессии и уровня значимости 0,05, находим значение статистики Стьюдента равное 2,365. Значение *t*-статистики для переменных X_1, X_2, и X_3 по модулю больше *t*-критического, следовательно, переменные статистически значимы.

Уравнение множественной регрессии для данной модели имеет вид:

$$Y = 3717{,}756 - 32{,}040X_1 + 17{,}288X_2 - 1{,}003X_3.$$

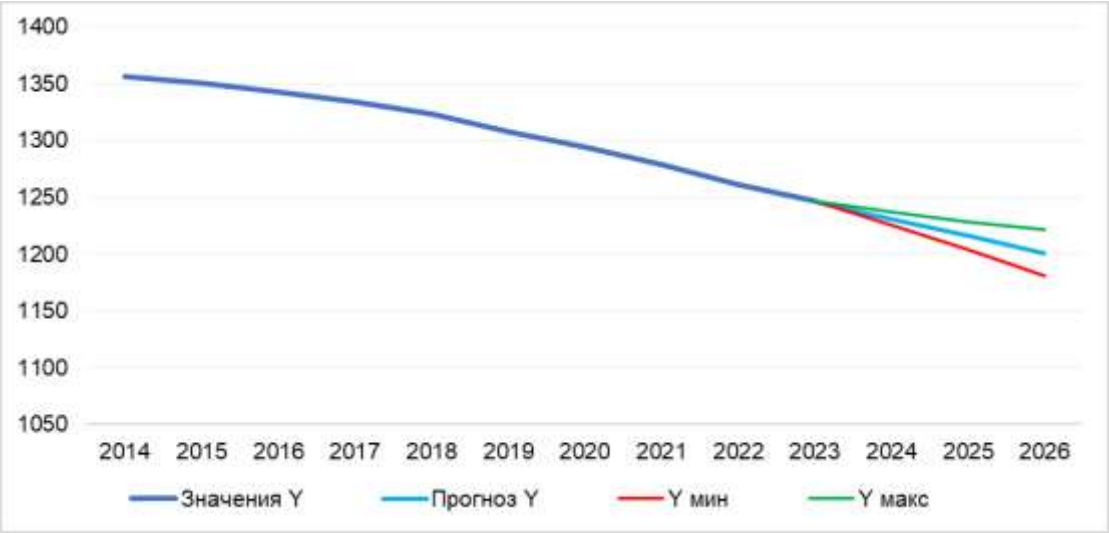
Проведем анализ полученных коэффициентов при независимых переменных. Положительный коэффициент при переменной X_2 означает, что с возрастанием переменной, значение отклика также возрастает. Отрицательный коэффициент при переменных X_1, и X_3 означает, что с возрастанием этих переменных, значение зависимой переменной убывает [3].

Множественная регрессия используется для предсказания зависимой переменной Y по известным X_1, X_2, X_3, путем подстановки значений предикторов в уравнение регрессии.

Для данной модели спрогнозируем Y (численность населения) за 2024-2026 годы. Так как переменные X_1, X_2, и X_3 за эти года неизвестны, то уравнением множественной регрессии мы воспользоваться не можем. В данном случае для предсказания Y применим функцию ПРЕДСКАЗ в MS Excel. Данная функция позволяет спрогнозировать Y, а также его доверительный интервал с вероятностью 95% (самое низкое и самое высокое значение Y из возможных).

Таблица 5. Результаты прогноза

Год	*Y*	$Y_{мин}$	$Y_{макс}$
2014	1356,6	-	-
2015	1350,8	-	-
2016	1342,6	-	-
2017	1334,3	-	-
2018	1323,1	-	-
2019	1308,1	-	-
2020	1294,5	-	-
2021	1278,8	-	-
2022	1261,1	-	-
2023	1246,6	-	-
2024	1231,3	1225,3	1237,2
2025	1216,3	1204,0	1228,5
2026	1201,2	1181,1	1221,4



Результаты прогноза

В результате исследований, можно сделать вывод о том, что данная модель является статистически значимой и может успешно применяться на практике.

Проанализировав полученные предсказанные значения Y, видно, что даже в самом благоприятном случае, численность населения Пензенской области будет уменьшаться.

Список использованных источников:

1. Кремер, Н. Ш. Теория вероятностей и математическая статистика: учебник и практикум для академического бакалавриата / Н. Ш. Кремер. – 4-е изд., перераб. и доп. – М.: Издательство Юрайт, 2015. – 514 с.

2. Краснов, А. Ю. Статистические методы в инженерных исследованиях: Учебно-методическое пособие / А. Ю. Краснов. – СПб.: Университет ИТМО, 2022. – 119 с.

3. Наследов, А. IBM SPSS Statistics 20 и AMOS: профессиональный статистический анализ данных / А. Наследов. – СПб.: Питер, 2013. – 416 с.