

**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГБОУ «Тувинский государственный университет»
ФИЗИКО-МАТЕМАТИЧЕСКИЙ ФАКУЛЬТЕТ
КАФЕДРА МАТЕМАТИКИ И МЕТОДИКИ ПРЕПОДАВАНИЯ МАТЕМАТИКИ**

Выпускная квалификационная работа
(бакалаврская работа)

**Прогнозирование и расчет пассажирооборота Республики Тыва
до 2025 года**

Работа допущена к защите

И.о.зав.кафедры _____

Танзы М.В., к.ф-м.н., доцент

(фамилия, и.о., должность, уч.степень и звание)

Работа защищена «__» _____ 20__ г.

С оценкой _____

Председатель ГЭК _____

(подпись)

Сенашов В.И. д.ф-м.н., профессор
ведущий научный сотрудник Института
вычислительного моделирования
СО РАН, Красноярск

Члены комиссии _____

(подписи)

Студентки 4 курса 5 группы

направление подготовки 01.03.01

профили «Математика»

очной формы обучения

Шунней Айжана Радиевна
(Ф.И.О)

(подпись)

«__» _____ 20__ г.

Научный руководитель: _____

(подпись)

Сотников А.И.к.ф-м.н.,ст.преподаватель

(фамилия, и.о., должность, уч.степень и звание)

Кызыл – 2020.

СОДЕРЖАНИЕ

Введение.....	3
Глава 1. Теоретические и методологические основы применения регрессионного анализа в эконометрике	
1.1. Основные положения регрессионного анализа.....	5
1.2. Оценка параметров парной регрессионной модели.....	7
1.3. Оценка значимости уравнения регрессии и особенности применения коэффициента детерминации.....	12
Глава 2. Практическое применение регрессионного анализа в эконометрике	
Задача.....	16
Линейная модель.....	18
Квадратичная модель.....	34
Заключение.....	46
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ.....	47

Введение

Актуальность выпускной квалификационной работы состоит в том, что в эконометрике широко используются методы статистики. Во многих практических задачах прогнозирования, изучая различного рода связи в экономических, производственных системах, необходимо на основании экспериментальных данных выразить зависимую переменную в виде некоторой математической функции от независимых переменных – регрессоров, то есть построить регрессионную модель. Регрессионный анализ позволяет:

- производить расчет регрессионных моделей путем определения значений параметров – постоянных коэффициентов при независимых переменных – регрессорах, которые часто называют факторами;
- проверить гипотезу об адекватности модели имеющимся наблюдениям;
- использовать модель для прогнозирования значений зависимой переменной при новых или ненаблюдаемых значениях независимых переменных.

Цель работы состоит в построении нескольких регрессионных моделей (линейных и квадратичных) для прогноза пассажирооборота Республики Тыва до 2025 года, их анализе и выборе наиболее точной модели. Для достижения поставленной цели были решены следующие **задачи**:

- изучение основных положений регрессионного анализа
- рассмотрение оценки параметров парной регрессионной модели
- исследование оценки значимости уравнения регрессии и особенностей применения коэффициента детерминации

· провести прогнозирование пассажирооборота Республики Тыва до 2025 года

Предметом исследования является математико-статистические методы в экономических исследованиях.

Объект исследования дипломной работы – практическая задача по применению регрессионного анализа в эконометрике.

Практическая значимость выпускной квалификационной работы заключается в исследовании пассажирооборота Республики Тыва и прогнозирования.

Структура работы. Работа состоит из введения, двух глав, заключения, списка использованной литературы.

Во введении обосновывается актуальность выбранной темы, формируется цель, предмет и объект исследования, практическая значимость, раскрывается структура исследования.

В первой главе рассмотрена теория, требуемая для построения моделей регрессии и последующего их анализа.

Во второй главе рассматривается практическая часть работы, в которой производятся вычисления коэффициентов регрессионного уравнения, оценка качества моделей.

В заключении построены модели для прогнозирования пассажирооборота Республики Тыва до 2025 года, и вычислены ошибки прогноза, сделан вывод о том, какая из моделей является наиболее точной.

Глава 1. Теоретические и методологические основы применения регрессионного анализа в эконометрике

1.1. Основные положения регрессионного анализа

В эконометрике широко используется регрессионный анализ как метод выявления уравнения связи между зависимыми и независимыми переменными, наилучшим способом дающий оценку истинного соотношения между этими переменными.

Регрессия – зависимость между независимыми (объясняющими) переменными и условным математическим ожиданием зависимой (объясняемой) переменной.

Регрессионный анализ – один из наиболее известных статистических методов для решения следующей основной задачи: можно ли по значению одного признака сделать какие-либо выводы о предполагаемом значении другого?

Регрессионная модель

Основная цель регрессионного анализа состоит в определении связи между некоторой характеристикой Y наблюдаемого явления или объекта и величинами $x_1, x_2, \dots, x_n$, которые обуславливают, объясняют изменения Y .

К задачам регрессионного анализа относятся [1, с. 75]:

- установление формы зависимости между переменными;
- оценка модельной функции (модельного уравнения) регрессии;
- оценка неизвестных значений (прогноз значений) зависимой переменной.

Переменная Y называется зависимой переменной (откликом), а объясняющие переменные $x_1, x_2, \dots, x_n$, называются регрессорами или факторами. Одному и тому же конкретному набору $(x_1, x_2, \dots, x_n)$ значений факторов в различных наблюдениях могут соответствовать несколько различных (но не произвольных) значений y объясняемой переменной Y . В них может наблюдаться некоторая закономерность, они могут иметь определенный закон распределения. То есть каждому набору значений факторов в разных конкретных реализациях могут соответствовать не строго заранее определенные соответствующие значения объясняемой переменной, а её условные математические ожидания. Такая зависимость называется *регрессионной*.

Уравнение регрессии в общем виде выглядит:

$$y = f(x_1, x_2, \dots, x_n) \quad (1.1.1)$$

$$\text{или иначе } y = f x_1, x_2, \dots, x_n + \varepsilon \quad (1.1.2)$$

где y – *теоретический вид модели*, формула функции, которую задает исследователь, исходя из своих предположений о характере связи между переменными; ε – *случайная часть* (ошибка, возмущения).

Причины появления случайной составляющей ε :

- неучитывание в используемой модели некоторых факторов, от которых также существенно зависит Y ;
- не достаточно точно выбранный вид модели;
- наличие ошибок исходных данных и другие.

Классификация эконометрических моделей:

1. По количеству объясняющих переменных:
 - a. Модель парной регрессии;
 - b. Модель множественной регрессии.

2. По виду функции регрессии:
 - a. Модель линейной регрессии;
 - b. Модель нелинейной регрессии.
3. По типу используемых данных
4. По количеству эконометрических уравнений.

В предоставленной работе рассмотрена модель парной регрессии. До этого только из только кружка причин, оказывающих большое влияние на продуктивный знак, нужно отметить более значительно действующие причины. Парильная регрессия достаточна, ежели наличествует преобладающий причина, кой и употребляется соединяющей переменной. Уравнение обычный регрессии ассоциация меж двумя переменными, имеет место быть как некая регулярность только в среднем в целом сообразно надзоров. В уравнении регрессии корреляционная сообразно сущности ассоциация показателей ведётся в облике многофункциональной взаимосвязи, проявленной соответственной математической функцией.

1.2. Оценка параметров парной регрессионной модели

Парная линейная регрессионная модель – наиболее простой вид регрессионной модели, в которой рассматривается зависимость объясняемой переменной Y , только от одной объясняющей переменной X , причем эта зависимость *линейная*.

Линейная регрессия сводится к нахождению уравнения вида:

$$y_x = a + b * x \text{ или } y_x = a + b * x + \varepsilon. \quad (1.2.1)$$

Уравнение вида $y_x = a + b * x$ позволяет по заданным значениям фактора x находить теоретические значения результативного признака, подставляя в него фактические значения фактора x .

В парной регрессии выбор вида может быть осуществлен тремя методами [3, с. 25-26]: графическим, аналитическим, экспериментальным.

Парильная регрессия имеет возможность отдать неплохой итог при прогнозировании, ежели воздействием остальных причин, воздействующих на предмет изучения, разрешено пренебречь. Ежели ведь сииим воздействием пренебречь невозможно, то в данном случае надлежит попробовать обнаружить воздействие остальных причин, введя их в модель, т.е. выстроить уравнение множественной регрессии.

Модель множественной линейной регрессии в общем виде [2, с 85-86]:

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p + \varepsilon \quad (1.2.2)$$

$$\text{или в матричной форме } y = X * b + \varepsilon$$

где $Y=(y_1y_2 \dots y_n)^T$ – матрица-столбец наблюдаемых значений объясняемой переменной, $(\dots)^T$ – операции транспонирования матрицы.

Основная цель множественной регрессии – построить модель с большим числом факторов, определив при этом влияние каждого из них в отдельности, а также совокупное их воздействие на моделируемый показатель.

Построение регрессии сводится к оценке её параметров a и b [4, с 24-22]. Классический подход к оцениванию параметров линейной регрессии основан на методе наименьших квадратов (МНК). МНК позволяет получить такие оценки этих параметров, при которых сумма квадратов отклонений фактических значений y_i результативного признака (отклика) от теоретических (полученных по уравнению регрессии) y_i минимальна:

$$\sum_{i=1}^n y_i - y_{x_i} = \sum_{i=1}^n \varepsilon_i^2 \rightarrow \min \quad (1.2.3)$$

На практике невозможно определить сами *ошибки* ε_i , в расчетах имеем дело с их оценками, называемыми *остатками* ε_i , полученными для конкретной выборки. Задача состоит в получении таких оценок коэффициентов уравнения регрессии a и b , чтобы $\sum_{i=1}^n \varepsilon_i^2 \rightarrow \min$.

$$b = \frac{\text{cov}(x,y)}{\delta_x^2}, \quad a = y - b * x \quad (1.2.4)$$

где $\text{cov}(x, y) = y * x - y * x$ – ковариация признаков x и y ,

$\delta_x^2 = x^2 - x^2$ – дисперсия признака x

Параметр b называется коэффициентом регрессии. Его величина показывает среднее изменение отклика при изменении фактора на одну единицу.

Формально a – значение y при $x = 0$. Если фактор x не имеет и не может иметь нулевого значения, то какая трактовка параметра a не имеет смысла.

Остатки представляют собой взятые с соответствующим знаком разности экспериментальных y_i и оценочных y_i («практических» и «теоретических») значений объясняемой переменной.

Парная нелинейная модель регрессии

Близость абсолютной величины линейного коэффициента корреляции к нулю еще не говорит об отсутствии любой связи между признаками, а лишь отсутствие линейной корреляционной связи. На практике кроме линейной связи между показателями может наблюдаться более сложные – нелинейные регрессионные зависимости.

Основные типы нелинейных регрессий

- полиномиальные:

квадратичные $y = ax^2 + bx + c + \varepsilon$

кубические $y = ax^3 + bx^2 + cx + d + \varepsilon$

- гиперболическая $y = a + \frac{b}{x} + \varepsilon$
- степенная $y = a * x^b * \varepsilon$
- показательная $y = a * b^x * \varepsilon$
- экспоненциальная $y = e^{a+bx+\varepsilon}$
- логарифмическая $y = a + b * \ln x + \varepsilon$

Линеаризация некоторых нелинейных регрессий

Некоторые нелинейные регрессии путем математических преобразований могут быть сведены к соответствующим линейным регрессиям относительно других (производных) переменных. Параметры полученных линеаризованных моделей можно найти описанными выше методами.

а) *Показательная модель* $y = a * b^x * \varepsilon$

Логарифмируются левая и правая части уравнения и преобразуются в равенство:

$$\ln y = \ln(a * b^x * \varepsilon) = \ln a + \ln b^x + \ln \varepsilon = \ln a + x \ln b + \ln \varepsilon$$

Вводятся промежуточные переменные

$$\ln y = Y; \quad \ln a = A; \quad \ln b = B, \quad \ln \varepsilon = \varepsilon$$

С учетом введенных обозначений парная линейная регрессия имеет вид:

$$Y = A + Bx + E$$

Её отличие от исходной в том, что в качестве объясняемой переменной выступает не исходная y , а $Y = \ln y$. То есть в расчетах берется не исходный массив объясняемой переменной y , а массив предварительно полученных её логарифмов. И получены в результате расчётов будут не коэффициенты a и b ,

а их логарифмы $\ln a = A$ и $\ln b = B$. Из них требуемые значения можно получить:

$$a = e^{\ln a} = e^A, \quad b = e^{\ln b} = e^B$$

В конце подставляются полученные значения коэффициентов в уравнение исходной парной нелинейной регрессии $y = a * b^x * \varepsilon$.

б) *Квадратичная модель* $y = ax^2 + bx + c + \varepsilon$

Она также может быть линеаризована и сведена к линейной множественной модели.

Введем промежуточные переменные $x_1 = x^2$, $x_2 = x$.

Тогда имеем множественную линейную регрессию:

$$y = ax_1 + bx_2 + c + \varepsilon.$$

Число оцениваемых параметров для квадратичной модели $m=3$, а не 2, как у линейной, степенной и т.п.

Аналогично могут быть линеаризованы и другие нелинейные регрессии.

Характеристики парных нелинейных регрессий

Оценкой тесноты нелинейной связи служит *индекс корреляции*

$$\overline{R^2} = 1 - \frac{\sum_{i=1}^n y_i - \overline{y_i}^2}{\sum_{i=1}^n y_i - y_i^2} \quad (1.2.5)$$

Коэффициент детерминации, средняя относительная ошибка характеризуют качество не только линейного, но и нелинейного уравнения регрессии.

1.3 Оценка значимости уравнения регрессии и особенности применения коэффициента детерминации

После того как найдено уравнение линейной регрессии, проводится оценка значимости как уравнения в целом, так и отдельных его параметров [5, с 55].

Линейный коэффициент корреляции

Уравнение регрессии всегда дополняется показателем тесноты связи. При использовании линейной регрессии в качестве такого показателя выступает *линейный коэффициент корреляции*, r_{xy} характеризующий тесноту и направление связи между двумя коррелируемыми признаками в случае наличия между ними линейной зависимости. Расчет линейного коэффициента корреляции осуществляется по формулам:

$$r_{xy} = b * \frac{\delta_x}{\delta_y} = \frac{cov(x,y)}{\delta_x * \delta_y}, \quad (1.3.1)$$

где $\delta_x = \sqrt{\overline{x^2} - \bar{x}^2}$, $\delta_y = \sqrt{\overline{y^2} - \bar{y}^2}$ – среднеквадратическое отклонение признаков.

Линейный коэффициент корреляции находится в пределах $-1 \leq r_{xy} \leq 1$. Чем ближе по модулю значение r_{xy} к единице, тем сильнее *линейная связь* между признаками. При $r_{xy} = \pm 1$ имеем функциональную зависимость. Равенство нулю коэффициента r_{xy} означает полное отсутствие корреляционной связи.

Значимость линейного коэффициента корреляции определяется с помощью t-критерия Стьюдента, но можно воспользоваться и упрощенным правилом: если $|r| < 0.3$, то связь практически отсутствует, если $0.3 < |r| < 0.5$,

то связь слабая, если $0,5 < |r| < 0,7$, то связь достаточно сильная, если $|r| > 0,7$, то имеется высокая степень зависимости между признаками.

Коэффициент детерминации

Для оценки качества подбора линейной функции рассчитывается квадрат линейного коэффициента корреляции r_{xy}^2 , называемый *коэффициентом детерминации* R^2 . Коэффициент детерминации характеризует долю дисперсии результативного признака, объясняемую регрессией, в общей дисперсии результативного признака:

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (1.3.2)$$

где y_i – выборочные данные,

f_i – соответствующие им значения модели.

Также это квадрат корреляции Пирсона между двумя переменными. Он выражает количество дисперсии, общей между двумя переменными. Коэффициент принимает значения из интервала $[0;1]$. Чем ближе значение к 1 тем лучше построенная модель и тем ближе линия регрессии к экспериментальным точкам на графике. В случае парной линейной регрессионной модели коэффициент детерминации равен квадрату коэффициента корреляции, то есть $R^2 = r^2$.

Показателям тесноты связи можно дать качественную оценку (шкала Чеддока):

Количественная мера тесноты связи	Качественная характеристика силы связи
0,1 - 0,3	Слабая
0,3 - 0,5	Умеренная
0,5 - 0,7	Заметная
0,7 - 0,9	Высокая

0,9 - 0,99	Весьма высокая
------------	----------------

Чтобы иметь общее суждение о качестве модели из относительных отклонений по каждому наблюдению, определяют среднюю ошибку аппроксимации:

$$A = \frac{1}{n}. \quad A_i = \frac{1}{n} \sum \frac{y - y_x}{y} * 100\% \quad (1.3.3)$$

Средняя ошибка аппроксимации не должна превышать 8–10%.

Критерий Фишера

Оценка значимости (надежности) уравнения или же модели регрессии в целом производится на основе *F-критерия Фишера*, который состоит в проверке гипотезы H_0 : статистической незначимости уравнения регрессии и коэффициента детерминации ($R^2 = r_{xy}^2 = 0$).

$$F = \frac{r_{xy}^2}{1 - r_{xy}^2} * (n - 2) \quad (1.3.4)$$

Фактическое значение F-критерия Фишера сравнивается с табличным значением $F_{\text{табл}}(\alpha; k_1; k_2)$ при уровне значимости α и степенях свободы $k_1 = m$ и $k_2 = n - m - 1$ (n – число наблюдений, m – число параметров при переменной x). При этом:

- Если $F_{\text{факт}} > F_{\text{табл}}$, то гипотеза H_0 отклоняется, признается статистическая *значимость* уравнения в целом, то есть построенная модель «лучше» прогноза по среднему.
- Если $F_{\text{факт}} < F_{\text{табл}}$, то гипотеза H_0 принимается, уравнение *незначимо*, качество построенной модели сравнимо с точностью прогноза по среднему.

Для парной линейной регрессии $m = 1$ поэтому:

$$F = \frac{S_{\text{факт}}^2}{S_{\text{ост}}^2} = \frac{y_x - y^2}{y - y_x^2} * (n - 2) \quad (1.3.5)$$

В парной линейной регрессии оценивается значимость не только уравнения в целом, но и отдельных его параметров. С этой целью по каждому из параметров определяется его стандартная ошибка: m_b и m_a .

Стандартная ошибка коэффициента регрессии определяется по формуле:

$$m_b = \frac{\overline{S_{\text{ост}}^2}}{x - x^2} = \frac{S_{\text{ост}}}{\delta_x * \bar{n}} \quad (1.3.6)$$

где $S_{\text{ост}}^2 = \frac{\sum (y - y_x)^2}{n - 2}$ – остаточная дисперсия на одну степень свободы.

Для оценки существенности коэффициента регрессии его величина сравнивается с его стандартной ошибкой, то есть определяется фактическое значение t -критерия Стьюдента: $t_b = \frac{b}{m_b}$, которое затем сравнивается с табличным значением при определенном уровне значимости α и числе степеней свободы $(n - 2)$.

Доверительный интервал для коэффициента регрессии определяется как $b \pm t_{\text{табл}} * m_b$.

Стандартная ошибка параметра a определяется по формуле:

$$m_a = \overline{S_{\text{ост}}^2} * \frac{x^2}{n * \sum (x - \bar{x})^2} \quad (1.3.7)$$

Процедура оценивания существенности данного параметра не отличается от рассмотренной выше для коэффициента регрессии.

Глава 2. Практическое применение регрессионного анализа в эконометрике

Пассажиروоборот — показатель отражения объёма перевозок пассажиров в пассажиро-километрах и исчисляется как произведение количества пассажиров на расстояние перевозок по каждому виду транспорта.

Пользуясь приведенными исходными данными на таблице о пассажируобороте автобусов общего пользования Республики Тыва:

Год	2003	2004	2005	2006	2007	2008	2009	2010
Пассажируоборот, млн пасс, %	511	514	565	627	652	666	656	657
Год	2011	2012	2013	2014	2015	2016	2017	2018
Пассажируоборот, млн пасс, %	649	658	654	676	705	657	642	687

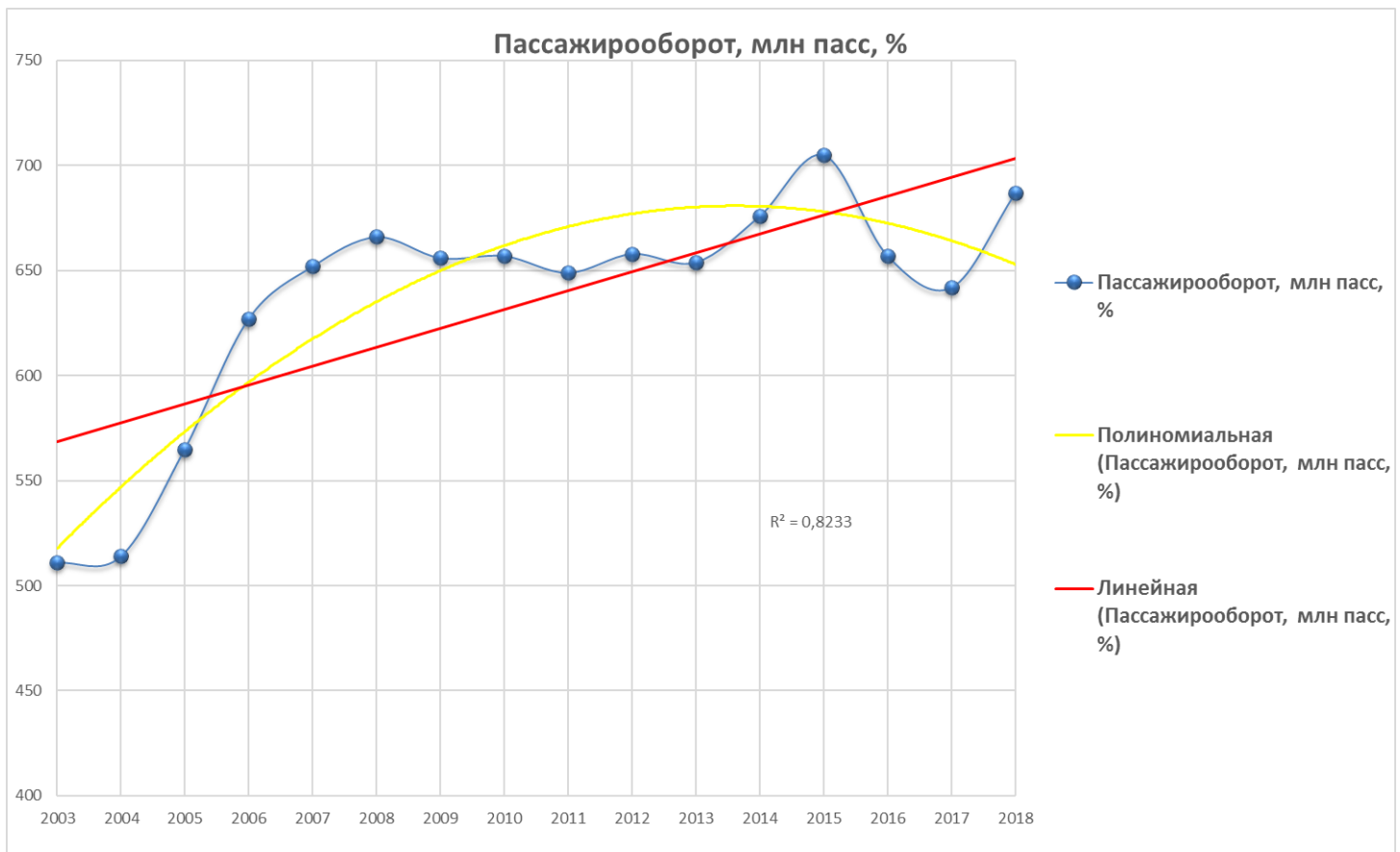
1. Построить поле корреляции.
2. Найти уравнения линейной, квадратичной регрессии.
3. Оценить для каждой модели: тесноту связи с помощью показателей корреляции и детерминации; с помощью средней ошибки аппроксимации качество уравнений; сделать прогноз до 2025г.
4. По значениям характеристик выбрать лучшее уравнение регрессии и дать обоснование.

Решение:

Использование графического метода.

Этот метод применяют для наглядного изображения формы связи между изучаемыми показателями. Для этого в прямоугольной системе координат строят график, по оси ординат откладывают индивидуальные значения результативного признака Y , а по оси абсцисс - индивидуальные значения факторного признака X .

На основании поля корреляции можно выдвинуть гипотезу о том, что связь между всеми возможными значениями X и Y носит линейный характер.



Линейная модель

Линейное уравнение регрессии имеет вид $y = bx + a$

Для расчета параметров регрессии построим расчетную таблицу

x	y	x ²	y ²	x*y
2003	511	4012009	261121	1023533
2004	514	4016016	264196	1030056
2005	565	4020025	319225	1132825
2006	627	4024036	393129	1257762
2007	652	4028049	425104	1308564
2008	666	4032064	443556	1337328
2009	656	4036081	430336	1317904
2010	657	4040100	431649	1320570
2011	649	4044121	421201	1305139
2012	658	4048144	432964	1323896
2013	654	4052169	427716	1316502
2014	676	4056196	456976	1361464
2015	705	4060225	497025	1420575
2016	657	4064256	431649	1324512
2017	642	4068289	412164	1294914
2018	687	4072324	471969	1386366
32168	10176	64674104	6519980	20461910

Для наших данных система уравнений имеет вид

$$16a + 32168b = 10176$$

$$32168a + 64674104b = 20161910$$

$$-32168a - 64673764b = -20458848$$

$$32168a + 64674104b = 20461910$$

$$340b = 3062$$

$$b = 9.0059$$

Теперь найдем коэффициент «a» из уравнения (1):

$$16a + 32168 \cdot b = 10176$$

$$16a + 32168 \cdot 9.0059 = 10176$$

$$16a = -279525.224$$

$$a = -17470.3265$$

Уравнение регрессии (эмпирическое уравнение регрессии):

$$y = 9.0059 x - 17470.3265$$

Параметры уравнения регрессии

Выборочные средние

$$\bar{x} = \frac{\sum x_i}{n} = \frac{32168}{16} = 2010.5$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{10176}{16} = 636$$

$$\overline{xy} = \frac{\sum x_i y_i}{n} = \frac{20461910}{16} = 1278869.375$$

Выборочные дисперсии:

$$S^2(x) = \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{64674104}{16} - 2010.5^2 = 21.25$$

$$S^2(y) = \frac{\sum y_i^2}{n} - \bar{y}^2 = \frac{6519980}{16} - 636^2 = 3002.75$$

Среднеквадратическое отклонение

$$S(x) = \sqrt{S^2(x)} = \sqrt{21.25} = 4.61$$

$$S(y) = \sqrt{S^2(y)} = \sqrt{3002.75} = 54.797$$

Коэффициент корреляции b можно находить по формуле, не решая систему непосредственно:

$$b = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{S^2(x)} = \frac{1278869.375 - 2010.5 \cdot 636}{21.25} = 9.0059$$

$$a = \bar{y} - b \cdot \bar{x} = 636 - 9.0059 \cdot 2010.5 = -17470.3265$$

Коэффициент корреляции.

Ковариация.

$$\text{cov}(x, y) = \overline{x \cdot y} - \bar{x} \cdot \bar{y} = 1278869.375 - 2010.5 \cdot 636 = 191.38$$

Рассчитываем показатель тесноты связи. Таким показателем является

выборочный линейный коэффициент корреляции, который рассчитывается по формуле:

$$r_{xy} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{S(x) \cdot S(y)} = \frac{1278869.375 - 2010.5 \cdot 636}{4.61 \cdot 54.797} = 0.758$$

В нашем примере связь между признаком Y и фактором X высокая и прямая.

Кроме того, коэффициент линейной парной корреляции может быть определен через коэффициент регрессии b:

$$r_{x,y} = b \cdot \frac{S(x)}{S(y)} = 9.006 \frac{4.61}{54.797} = 0.758$$

Значимость коэффициента корреляции.

Выдвигаем гипотезы:

$H_0: r_{xy} = 0$, нет линейной взаимосвязи между переменными;

$H_1: r_{xy} \neq 0$, есть линейная взаимосвязь между переменными;

Для того чтобы при уровне значимости α проверить нулевую гипотезу о равенстве нулю генерального коэффициента корреляции нормальной двумерной случайной величины при конкурирующей гипотезе $H_1 \neq 0$, надо вычислить наблюдаемое значение критерия (величина случайной ошибки)

$$t_{\text{набл}} = r_{xy} \frac{\sqrt{n-2}}{\sqrt{1 - r_{xy}^2}}$$

и по таблице критических точек распределения Стьюдента, по заданному уровню значимости α и числу степеней свободы $k = n - 2$ найти критическую точку $t_{\text{крит}}$ двусторонней критической области. Если $t_{\text{набл}} < t_{\text{крит}}$ оснований отвергнуть нулевую гипотезу. Если $|t_{\text{набл}}| > t_{\text{крит}}$ — нулевую гипотезу отвергают.

$$t_{\text{набл}} = 0.758 \frac{\sqrt{14}}{\sqrt{1 - 0.758^2}} = 4.343$$

По таблице Стьюдента с уровнем значимости $\alpha=0.05$ и степенями свободы $k=14$ находим $t_{\text{крит}}$:

$$t_{\text{крит}}(n-m-1; \alpha/2) = t_{\text{крит}}(14; 0.025) = 2.145$$

где $m = 1$ - количество объясняющих переменных.

Если $|t_{\text{набл}}| > t_{\text{критич}}$, то полученное значение коэффициента корреляции признается значимым (нулевая гипотеза, утверждающая равенство нулю коэффициента корреляции, отвергается).

Поскольку $|t_{\text{набл}}| > t_{\text{крит}}$, то отклоняем гипотезу о равенстве 0 коэффициента корреляции. Другими словами, коэффициент корреляции статистически - значим

В парной линейной регрессии $t^2_r = t^2_b$ и тогда проверка гипотез о значимости коэффициентов регрессии и корреляции равносильна проверке гипотезы о существенности линейного уравнения регрессии.

Интервальная оценка для коэффициента корреляции (доверительный интервал).

$$(r - t_{\text{крит}} \sqrt{\frac{1-r^2}{n-2}}; r + t_{\text{крит}} \sqrt{\frac{1-r^2}{n-2}})$$

Доверительный интервал для коэффициента корреляции.

$$(0.758 - 2.145 \sqrt{\frac{1-0.758^2}{16-2}}; 0.758 + 2.145 \sqrt{\frac{1-0.758^2}{16-2}})$$
$$r \in (0.383; 1)$$

Уравнение регрессии (оценка уравнения регрессии).

$$y_x = r_{xy} \cdot \frac{x - \bar{x}}{S(x)} \cdot S(y) + \bar{y} = 0.758 \frac{x - 2010.5}{4.61} 54.797 + 636 = 9.006x - 17470.326$$

Линейное уравнение регрессии имеет вид $y = 9.006x - 17470.326$

Коэффициент регрессии $b = 9.006$ показывает среднее изменение результативного показателя (в единицах измерения y) с повышением или понижением величины фактора x на единицу его измерения. В данном примере с увеличением на 1 единицу y повышается в среднем на 9.006.

Коэффициент $a = -17470.326$ формально показывает прогнозируемый уровень y , но только в том случае, если $x=0$ находится близко с выборочными значениями.

Но если $x=0$ находится далеко от выборочных значений x , то

буквальная интерпретация может привести к неверным результатам, и даже если линия регрессии довольно точно описывает значения наблюдаемой выборки, нет гарантий, что также будет при экстраполяции влево или вправо.

Подставив в уравнение регрессии соответствующие значения x , можно определить выровненные (предсказанные) значения результативного показателя $y(x)$ для каждого наблюдения.

Связь между y и x определяет знак коэффициента регрессии b (если > 0 – прямая связь, иначе – обратная). В нашем примере связь прямая.

Коэффициент эластичности.

Коэффициенты регрессии (в примере b) нежелательно использовать для непосредственной оценки влияния факторов на результативный признак в том случае, если существует различие единиц измерения результативного показателя y и факторного признака x .

Для этих целей вычисляются коэффициенты эластичности и бета – коэффициенты.

Средний коэффициент эластичности E показывает, на сколько процентов в среднем по совокупности изменится результат y от своей средней величины при изменении фактора x на 1% от своего среднего значения.

Коэффициент эластичности находится по формуле:

$$E = \frac{\partial y}{\partial x} \frac{x}{y} = b \frac{\bar{x}}{\bar{y}}$$

$$E = 9.006 \frac{2010.5}{636} = 28.469$$

В нашем примере коэффициент эластичности больше 1. Следовательно, при изменении X на 1%, Y изменится более чем на 1%. Другими словами – X существенно влияет на Y .

Бета – коэффициент

Бета – коэффициент показывает, на какую часть величины своего среднего квадратичного отклонения изменится в среднем значение результативного признака при изменении факторного признака на величину

его среднеквадратического отклонения при фиксированном на постоянном уровне значении остальных независимых переменных:

$$\beta_j = b_j \frac{S(x)}{S(y)} = 9.006 \frac{4.61}{54.797} = 0.758$$

Т.е. увеличение x на величину среднеквадратического отклонения S_x приведет к увеличению среднего значения Y на 75.8% среднеквадратического отклонения S_y .

Ошибка аппроксимации.

Оценим качество уравнения регрессии с помощью ошибки абсолютной аппроксимации. Средняя ошибка аппроксимации - среднее отклонение расчетных значений от фактических:

$$\overline{A} = \frac{\sum |y_i - y_x| : y_i}{n} 100\%$$

Ошибка аппроксимации в пределах 5%-7% свидетельствует о хорошем подборе уравнения регрессии к исходным данным.

$$\overline{A} = \frac{0.801}{16} 100\% = 5.01\%$$

В среднем, расчетные значения отклоняются от фактических на 5.01%. Поскольку ошибка меньше 7%, то данное уравнение можно использовать в качестве регрессии.

Эмпирическое корреляционное отношение.

Эмпирическое корреляционное отношение вычисляется для всех форм связи и служит для измерения тесноты зависимости. Изменяется в пределах [0;1].

$$\eta = \sqrt{\frac{\sum (\overline{y} - y_x)^2}{\sum (y_i - \overline{y})^2}}$$

$$\eta = \sqrt{\frac{27576.012}{48044}} = 0.758$$

где

$$\sum (\overline{y} - y_x)^2 = 48044 - 20467.99 = 27576.012$$

Индекс корреляции.

Для линейной регрессии индекс корреляции равен коэффициенту корреляции $r_{xy} = 0.758$.

Полученная величина свидетельствует о том, что фактор x существенно влияет на y .

Для любой формы зависимости теснота связи определяется с помощью **множественного коэффициента корреляции:**

$$R = \sqrt{1 - \frac{\sum(y_i - \bar{y}_x)^2}{\sum(y_i - \bar{y})^2}}$$

Данный коэффициент является универсальным, так как отражает тесноту связи и точность модели, а также может использоваться при любой форме связи переменных. При построении однофакторной корреляционной модели коэффициент множественной корреляции равен коэффициенту парной корреляции r_{xy} .

В отличие от линейного коэффициента корреляции он характеризует тесноту нелинейной связи и не характеризует ее направление. Изменяется в пределах $[0;1]$.

Теоретическое корреляционное отношение для линейной связи равно коэффициенту корреляции r_{xy} .

Коэффициент детерминации.

Квадрат (множественного) коэффициента корреляции называется коэффициентом детерминации, который показывает долю вариации результативного признака, объясненную вариацией факторного признака.

Чаще всего, давая интерпретацию коэффициента детерминации, его выражают в процентах.

$$R^2 = 0.758^2 = 0.574$$

т.е. в 57.4% случаев изменения x приводят к изменению y . Другими словами - точность подбора уравнения регрессии - средняя. Остальные 42.6% изменения Y объясняются факторами, не учтенными в модели.

Для оценки качества параметров регрессии построим расчетную таблицу

x	y	y(x)	$(y_i - y_{cp})^2$	$(y - y(x))^2$	$(x_i - x_{cp})^2$	$ y - y_x : y$
2003	511	568.456	15625	3301.178	56.25	0.112
2004	514	577.462	14884	4027.396	42.25	0.123
2005	565	586.468	5041	460.86	30.25	0.038
2006	627	595.474	81	993.918	20.25	0.0503
2007	652	604.479	256	2258.206	12.25	0.0729
2008	666	613.485	900	2757.794	6.25	0.0789
2009	656	622.491	400	1122.841	2.25	0.0511
2010	657	631.497	441	650.4	0.25	0.0388
2011	649	640.503	169	72.2	0.25	0.0131
2012	658	649.509	484	72.1	2.25	0.0129
2013	654	658.515	324	20.383	6.25	0.0069
2014	676	667.521	1600	71.9	12.25	0.0125
2015	705	676.526	4761	810.742	20.25	0.0404
2016	657	685.532	441	814.095	30.25	0.0434
2017	642	694.538	36	2760.266	42.25	0.0818
2018	687	703.544	2601	273.708	56.25	0.0241
32168	10176	10176	48044	20467.988	340	0.801

Оценка параметров уравнения регрессии.

Анализ точности определения оценок коэффициентов регрессии.

Несмещенной оценкой дисперсии возмущений является величина:

$$S^2 = \frac{\sum (y_i - y_x)^2}{n - m - 1}$$

$$S^2 = \frac{20467.988}{14} = 1461.999$$

$S^2 = 1461.999$ - необъясненная дисперсия или дисперсия ошибки регрессии

(мера разброса зависимой переменной вокруг линии регрессии).

$$S = \sqrt{S^2} = \sqrt{1461.999} = 38.24$$

$S = 38.24$ - стандартная ошибка оценки.

Стандартная ошибка регрессии рассматривается в качестве меры разброса данных наблюдений от смоделированных значений. Чем меньше значение стандартной ошибки регрессии, тем качество модели выше.

S_a - стандартное отклонение случайной величины a .

$$S_a = S \cdot \frac{\sqrt{\sum x^2}}{n S(x)}$$

$$S_a = 38.24 \cdot \frac{\sqrt{64674104}}{16 \cdot 4.61} = 4169.071$$

S_b - стандартное отклонение случайной величины b .

$$S_b = \frac{S}{\sqrt{n \cdot S(x)}}$$

$$S_b = \frac{38.24}{\sqrt{16 \cdot 4.61}} = 2.074$$

Доверительные интервалы для зависимой переменной.

Экономическое прогнозирование на основе построенной модели предполагает, что сохраняются ранее существовавшие взаимосвязи переменных и на период упреждения. Для прогнозирования зависимой переменной результативного признака необходимо знать прогнозные значения всех входящих в модель факторов.

Прогнозные значения факторов подставляют в модель и получают точечные прогнозные оценки изучаемого показателя.

$$(a + bx_p \pm \epsilon)$$

Рассчитаем границы интервала, в котором будет сосредоточено 95% возможных значений Y при неограниченно большом числе наблюдений и $X_p = 2212$

$$t_{\text{крит}}(n-m-1; \alpha/2) = t_{\text{крит}}(14; 0.025) = 2.145$$

$$y(2212) = 9.006 \cdot 2212 - 17470.326 = 2450.685$$

Вычислим ошибку прогноза для уравнения $y = bx + a$

$$\varepsilon = t_{\text{крит}} S \sqrt{\frac{1}{n} + \frac{(\bar{x} - x_p)^2}{\sum (x_i - \bar{x})^2}}$$

$$\varepsilon = 2.145 \cdot 38.236 \sqrt{\frac{1}{16} + \frac{(2010.5 - 2212)^2}{340}} = 896.5$$

или

$$\varepsilon = t_{\text{крит}} S \sqrt{\frac{1}{n} + \frac{(\bar{x} - x_p)^2}{n(\bar{x}^2 - \bar{x}^2)}}$$

$$\varepsilon = 2.145 \cdot 38.236 \sqrt{\frac{1}{16} + \frac{(2010.5 - 2212)^2}{16(4042131.5 - 2010.5^2)}} = 896.5$$

$$2450.685 \pm 896.5$$

$$(1554.19; 3347.18)$$

С вероятностью 95% можно гарантировать, что значения Y при неограниченно большом числе наблюдений не выйдет за пределы найденных интервалов.

Вычислим ошибку прогноза для уравнения $y = bx + a + \varepsilon$

$$\varepsilon = t_{\text{крит}} S \sqrt{1 + \frac{1}{n} + \frac{(\bar{x} - x_p)^2}{\sum (x_i - \bar{x})^2}}$$

$$\varepsilon = 2.145 \cdot 38.236 \sqrt{1 + \frac{1}{16} + \frac{(2010.5 - 2212)^2}{340}} = 900.24$$

$$2450.685 \pm 900.24$$

$$(1550.44; 3350.93)$$

Индивидуальные доверительные интервалы для Y при данном значении X .

$$(a + bx_i \pm \varepsilon)$$

где

$$\varepsilon = t_{\text{крит}} S \sqrt{1 + \frac{1}{n} + \frac{(\bar{x} - x_i)^2}{\sum (x_i - \bar{x})^2}}$$

$$\varepsilon = 2.145 \cdot 38.24 \sqrt{1 + \frac{1}{16} + \frac{(2010.5 - x_i)^2}{340}}$$

$$t_{\text{крит}} (n-m-1; \alpha/2) = (14; 0.025) = 2.145$$

x_i	$y = -17470.33 + 9.01x_i$	ε_i	$y_{\min} = y - \varepsilon_i$	$y_{\max} = y + \varepsilon_i$
2003	568.456	90.884	477.571	659.34
2004	577.462	89.348	488.114	666.809
2005	586.468	88.009	498.459	674.477
2006	595.474	86.878	508.596	682.351
2007	604.479	85.962	518.517	690.441
2008	613.485	85.269	528.217	698.754
2009	622.491	84.803	537.688	707.295
2010	631.497	84.57	546.927	716.067
2011	640.503	84.57	555.933	725.073
2012	649.509	84.803	564.705	734.312
2013	658.515	85.269	573.246	743.783
2014	667.521	85.962	581.559	753.483
2015	676.526	86.878	589.649	763.404
2016	685.532	88.009	597.523	773.541
2017	694.538	89.348	605.191	783.886
2018	703.544	90.884	612.66	794.429

С вероятностью 95% можно гарантировать, что значения Y при неограниченно большом числе наблюдений не выйдет за пределы найденных интервалов.

Проверка гипотез относительно коэффициентов линейного уравнения регрессии.

1) t-статистика. Критерий Стьюдента.

С помощью МНК мы получили лишь оценки параметров уравнения регрессии, которые характерны для конкретного статистического наблюдения

(конкретного набора значений x и y).

Для оценки статистической значимости коэффициентов регрессии и корреляции рассчитываются t -критерий Стьюдента и доверительные интервалы каждого из показателей. Выдвигается гипотеза H_0 о случайной природе показателей, т.е. о незначимом их отличии от нуля.

Чтобы проверить, значимы ли параметры, т.е. значимо ли они отличаются от нуля для генеральной совокупности используют статистические методы проверки гипотез.

В качестве основной (нулевой) гипотезы выдвигают гипотезу о незначимом отличии от нуля параметра или статистической характеристики в генеральной совокупности. Наряду с основной (проверяемой) гипотезой выдвигают альтернативную (конкурирующую) гипотезу о неравенстве нулю параметра или статистической характеристики в генеральной совокупности. Проверим гипотезу H_0 о равенстве отдельных коэффициентов регрессии нулю (при альтернативе H_1 не равно) на уровне значимости $\alpha=0.05$.

$H_0: b = 0$, то есть между переменными x и y отсутствует линейная взаимосвязь в генеральной совокупности;

$H_1: b \neq 0$, то есть между переменными x и y есть линейная взаимосвязь в генеральной совокупности.

В случае если основная гипотеза окажется неверной, мы принимаем альтернативную. Для проверки этой гипотезы используется *t-критерий Стьюдента*.

Найденное по данным наблюдений значение t -критерия (его еще называют наблюдаемым или фактическим) сравнивается с табличным (критическим) значением, определяемым по таблицам распределения Стьюдента (которые обычно приводятся в конце учебников и практикумов по статистике или эконометрике).

Табличное значение определяется в зависимости от уровня значимости (α) и числа степеней свободы, которое в случае линейной парной регрессии равно $(n-2)$, n -число наблюдений.

Если фактическое значение t -критерия больше табличного (по модулю), то основную гипотезу отвергают и считают, что с вероятностью $(1-\alpha)$ параметр или статистическая характеристика в генеральной совокупности значимо отличается от нуля.

Если фактическое значение t -критерия меньше табличного (по модулю), то нет оснований отвергать основную гипотезу, т.е. параметр или статистическая характеристика в генеральной совокупности незначимо отличается от нуля при уровне значимости α .

$$t_{\text{крит}}(n-m-1; \alpha/2) = t_{\text{крит}}(14; 0.025) = 2.145$$

$$t_b = \frac{b}{S_b}$$

$$t_b = \frac{9.006}{2.074} = 4.34$$

Поскольку $4.34 > 2.145$, то статистическая значимость коэффициента регрессии b подтверждается (отвергаем гипотезу о равенстве нулю этого коэффициента).

$$t_a = \frac{a}{S_a}$$

$$t_a = \frac{-17470.326}{4169.071} = 4.19$$

Поскольку $4.19 > 2.145$, то статистическая значимость коэффициента регрессии a подтверждается (отвергаем гипотезу о равенстве нулю этого коэффициента).

Доверительный интервал для коэффициентов уравнения регрессии.

Определим доверительные интервалы коэффициентов регрессии, которые с надежностью 95% будут следующими:

$$(b - t_{\text{крит}} S_b; b + t_{\text{крит}} S_b)$$

$$(9.01 - 2.145 \cdot 2.074; 9.01 + 2.145 \cdot 2.074)$$

$$(4.558; 13.454)$$

С вероятностью 95% можно утверждать, что значение данного параметра будут лежать в найденном интервале.

$$(a - t_{\text{крит}} S_a; a + t_{\text{крит}} S_a)$$

$$(-17470.326 - 2.145 \cdot 4169.071; -17470.326 + 2.145 \cdot 4169.071)$$

$$(-26412.984; -8527.669)$$

С вероятностью 95% можно утверждать, что значение данного параметра будут лежать в найденном интервале.

2) F-статистика. Критерий Фишера.

Коэффициент детерминации R^2 используется для проверки существенности уравнения линейной регрессии в целом.

Проверка значимости модели регрессии проводится с использованием F-критерия Фишера, расчетное значение которого находится как отношение дисперсии исходного ряда наблюдений изучаемого показателя и несмещенной оценки дисперсии остаточной последовательности для данной модели.

Если расчетное значение с $k_1=(m)$ и $k_2=(n-m-1)$ степенями свободы больше табличного при заданном уровне значимости, то модель считается значимой.

$$R^2 = 1 - \frac{\sum(y_i - y_x)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{20467.99}{48044} = 0.574$$

где m – число факторов в модели.

Оценка статистической значимости парной линейной регрессии производится по следующему алгоритму:

1. Выдвигается нулевая гипотеза о том, что уравнение в целом статистически незначимо: $H_0: R^2=0$ на уровне значимости α .

2. Далее определяют фактическое значение F-критерия:

$$F = \frac{R^2 (n - m - 1)}{1 - R^2 m}$$

$$F = \frac{0.574 (16-1-1)}{1 - 0.574 \cdot 1} = 18.862$$

или по формуле:

$$F = \frac{\sum(y_x - \bar{y})^2}{\sum(y_i - y_x)^2} \cdot \frac{(n - m - 1)}{m} = \frac{27576.0118}{20467.99} \cdot \frac{(16-1-1)}{1} = 18.862$$

где

$$\sum(y_x - \bar{y})^2 = 48044 - 20467.99 = 27576.0118$$

где $m=1$ для парной регрессии.

3. Табличное значение определяется по таблицам распределения Фишера для заданного уровня значимости, принимая во внимание, что число степеней свободы для общей суммы квадратов (большей дисперсии) равно 1 и число степеней свободы остаточной суммы квадратов (меньшей дисперсии) при линейной регрессии равно $n-2$.

$F_{\text{табл}}$ - это максимально возможное значение критерия под влиянием случайных факторов при данных степенях свободы и уровне значимости α . Уровень значимости α - вероятность отвергнуть правильную гипотезу при условии, что она верна. Обычно α принимается равной 0,05 или 0,01.

4. Если фактическое значение F-критерия меньше табличного, то говорят, что нет основания отклонять нулевую гипотезу.

В противном случае, нулевая гипотеза отклоняется и с вероятностью $(1-\alpha)$ принимается альтернативная гипотеза о статистической значимости уравнения в целом.

Табличное значение критерия со степенями свободы $k_1=1$ и $k_2=14$, $F_{\text{табл}} = 4.6$

Поскольку фактическое значение $F > F_{\text{табл}}$, то коэффициент детерминации статистически значим (найденная оценка уравнения регрессии статистически надежна).

Связь между F-критерием Фишера и t-статистикой Стьюдента выражается равенством:

$$t_r^2 = t_b^2 = \sqrt{F} = \sqrt{18.862} = 4.34$$

Прогнозирование до 2025 года

Годы	Фактические значения	Годы	Прогноз
2003	511	2019	678
2004	514	2020	676
2005	565	2021	677
2006	627	2022	683
2007	652	2023	687
2008	666	2024	691
2009	656	2025	684
2010	657		
2011	649		
2012	658		
2013	654		
2014	676		
2015	705		
2016	657		
2017	642		
2018	687		



Выводы

Изучена зависимость Y от X . На этапе спецификации была выбрана парная линейная регрессия. Оценены её параметры методом наименьших квадратов. Статистическая значимость уравнения проверена с помощью коэффициента детерминации и критерия Фишера. Установлено, что в исследуемой ситуации 57.4% общей вариабельности Y объясняется изменением X . Установлено также, что параметры модели статистически значимы. Возможна экономическая интерпретация параметров модели - увеличение X на 1 ед.изм. приводит к увеличению Y в среднем на 9.006 ед.изм. Полученные оценки уравнения регрессии позволяют использовать его для прогноза.

Квадратичная модель

Уравнение регрессии имеет вид $y = cx^2 + bx + a$

Находим параметры уравнения методом наименьших квадратов.

Система уравнений МНК:

$$an + b\sum x + c\sum x^2 = \sum y$$

$$a\sum x + b\sum x^2 + c\sum x^3 = \sum yx$$

$$a\sum x^2 + b\sum x^3 + c\sum x^4 = \sum yx^2$$

x	y	x^2	y^2	x y	x^3	x^4	$x^2 y$
2003	511	4012009	261121	1023533	8036054027	16096216216081	2050136599
2004	514	4016016	264196	1030056	8048096064	16128384512256	2064232224
2005	565	4020025	319225	1132825	8060150125	16160601000625	2271314125
2006	627	4024036	393129	1257762	8072216216	16192865729296	2523070572

2007	652	4028049	425104	1308564	8084294343	16225178746401	2626287948
2008	666	4032064	443556	1337328	8096384512	16257540100096	2685354624
2009	656	4036081	430336	1317904	8108486729	16289949838561	2647669136
2010	657	4040100	431649	1320570	8120601000	16322408010000	2654345700
2011	649	4044121	421201	1305139	8132727331	16354914662641	2624634529
2012	658	4048144	432964	1323896	8144865728	16387469844736	2663678752
2013	654	4052169	427716	1316502	8157016197	16420073604561	2650118526
2014	676	4056196	456976	1361464	8169178744	16452725990416	2741988496
2015	705	4060225	497025	1420575	8181353375	16485427050625	2862458625
2016	657	4064256	431649	1324512	8193540096	16518176833536	2670216192
2017	642	4068289	412164	1294914	8205738913	16550975387521	2611841538
2018	687	4072324	471969	1386366	8217949832	16583822760976	2797686588

Для наших данных система уравнений имеет вид

$$16a + 32168b + 64674104c = 10176$$

$$32168a + 64674104b + 130028653232c = 20461910$$

$$64674104a + 130028653232b + 261426730288328c = 41145034174$$

Получаем $c = -1.448$, $b = 5832.135$, $a = -5871139.831$

Уравнение регрессии:

$$y = -1.448x^2 + 5832.135x - 5871139.831$$

Эмпирические коэффициенты регрессии a и b являются лишь оценками теоретических коэффициентов β_i , а само уравнение отражает лишь общую тенденцию в поведении рассматриваемых переменных.

Ошибка аппроксимации.

Оценим качество уравнения регрессии с помощью средней относительной ошибки аппроксимации.

$$\overline{A} = \frac{\sum |y_x - y_i| : y_i}{n} 100\%$$

Ошибка аппроксимации в пределах 5%-7% свидетельствует о хорошем подборе уравнения регрессии к исходным данным.

$$\overline{A} = \frac{0.5118}{16} 100\% = 3.2\%$$

Поскольку ошибка меньше 7%, то данное уравнение можно использовать в качестве регрессии.

Для определения размеров погрешности или точности прогноза показателя Y рассчитаем коэффициент несоответствия Тейла по формуле:

$$K_T = \frac{\sqrt{\sum (y_i - \overline{y})^2}}{\sqrt{\sum y_i^2}}$$

$$K_T = \frac{8488.649}{6519980} = 0.0013$$

Этот показатель изменяется от 0 до 1. Чем ближе его значение к нулю, тем лучше результаты прогнозирования.

Средние значения

$$\overline{x} = \frac{\sum x_i}{n} = \frac{32168}{16} = 2010.5$$

$$\overline{x^2} = \frac{\sum x_i^2}{n} = \frac{64674104}{16} = 4042131.5$$

$$\overline{y} = \frac{\sum y_i}{n} = \frac{10176}{16} = 636$$

$$\overline{x \cdot y} = \frac{\sum x_i y_i}{n} = \frac{20461910}{16} = 1278869.375$$

Дисперсия

$$D(x) = \frac{\sum x_i^2}{n} - \overline{x}^2 = \frac{64674104}{16} - 2010.5^2 = 21.25$$

$$D(x_{1,2}) = \frac{\sum x_{1,2}^2}{n} - \overline{x}^2 = \frac{261426730288328}{16} - 4042131.5^2 = 343579728.25$$

$$D(y) = \frac{\sum y_i^2}{n} - \overline{y}^2 = \frac{6519980}{16} - 636^2 = 3002.75$$

Среднеквадратическое отклонение

$$\sigma(x) = \sqrt{D(x)} = \sqrt{21.25} = 4.6098$$

$$\sigma(y) = \sqrt{D(y)} = \sqrt{3002.75} = 54.7974$$

Коэффициент эластичности.

Коэффициент эластичности представляет собой показатель силы связи фактора x с результатом y , показывающий, на сколько процентов изменится значение y при изменении значения фактора на 1%.

$$E = \frac{\partial y}{\partial x} \frac{x}{y} = (b + 2a \overline{x}) \frac{\overline{x}}{y(\overline{x})}$$

$$E = (5832.135 + 2 \cdot -1.448 \cdot 2010.5) \frac{2010.5}{1278869.375} = 27.155$$

Коэффициент эластичности больше 1. Следовательно, при изменении x на 1%, Y изменится более чем на 1%. Другими словами - изменение x существенно влияет на Y .

Эмпирическое корреляционное отношение.

Эмпирическое корреляционное отношение вычисляется для всех форм связи и служит для измерения тесноты зависимости. Изменяется в пределах [0;1].

$$\eta = \sqrt{\frac{\sum(\bar{y} - y_x)^2}{\sum(y_i - \bar{y})^2}} = \sqrt{\frac{39555.351}{48044}} = 0.907$$

где

$$\sum(\bar{y} - y_x)^2 = 48044 - 8488.649 = 39555.351$$

В отличие от линейного коэффициента корреляции он характеризует тесноту нелинейной связи и не характеризует ее направление. Изменяется в пределах [0;1].

Связи между признаками могут быть слабыми и сильными (тесными). Полученная величина свидетельствует о том, что изменение x существенно влияет на y .

Индекс детерминации.

$$R^2 = 1 - \frac{\sum(y_i - \bar{y}_x)^2}{\sum(y_i - \bar{y})^2}$$

$$R^2 = 1 - \frac{8488.649}{48044} = 0.823$$

т.е. в 82.33% случаев x влияет на изменение y . Другими словами - точность подбора уравнения регрессии - высокая.

Для оценки качества параметров уравнения построим расчетную таблицу

x	y	$y(x)$	$(y_i - y_{cp})^2$	$(y_i - y(x))^2$	$(y_i - y(x)) : y_i$
2003	511	517.77	15625	45.828	0.0132

2004	514	547.05	14884	1092.302	0.0643
2005	565	573.434	5041	71.133	0.0149
2006	627	596.922	81	904.704	0.048
2007	652	617.513	256	1189.351	0.0529
2008	666	635.208	900	948.148	0.0462
2009	656	650.007	400	35.921	0.00914
2010	657	661.909	441	24.097	0.00747
2011	649	670.915	169	480.254	0.0338
2012	658	677.024	484	361.921	0.0289
2013	654	680.237	324	688.401	0.0401
2014	676	680.554	1600	20.741	0.00674
2015	705	677.975	4761	730.37	0.0383
2016	657	672.499	441	240.211	0.0236
2017	642	664.126	36	489.581	0.0345
2018	687	652.858	2601	1165.687	0.0497
		10176	48044	8488.649	0.512

Стандартные ошибки коэффициентов:

$$S_a = \sqrt{-7471118496438.2} = \text{NaN}$$

$$S_b = \sqrt{1848313.5097} = 1359.527$$

$$S_c = \sqrt{0.1143} = 0.338$$

Интервальный прогноз.

Определим среднеквадратическую ошибку прогнозируемого показателя.

$$Uy = y_{n+L} \pm K$$

где

$$K = t_a \cdot Sy \cdot \sqrt{1 + \frac{1}{n} + \frac{3(n+2L-1)^2}{n(n^2-1)}}$$

L - период упреждения; y_{n+L} - точечный прогноз по модели на $(n + L)$ -й момент времени; n - количество наблюдений во временном ряду; Sy - стандартная ошибка прогнозируемого показателя; $T_{\text{табл}}$ - табличное значение критерия По таблице Стьюдента находим $T_{\text{табл}}$

$$T_{\text{табл}}(n-m-1; \alpha/2) = (;) = 2.16$$

Точечный прогноз, $t = 2019$: $y(2019) = -1.448 \cdot 2019^2 + 5832.135 \cdot 2019 - 5871139.831 = 638.69$

$$K_1 = 2.16 \cdot 25.55 \cdot \sqrt{1 + \frac{1}{16} + \frac{3(16+2 \cdot 1-1)^2}{16(16^2-1)}} = 62.32$$

$$638.69 - 62.32 = 576.37 ; 638.69 + 62.32 = 701.01$$

Интервальный прогноз:

$t = 2019$: (576.37; 701.01)

Точечный прогноз, $t = 2020$: $y(2020) = -1.448 \cdot 2020^2 + 5832.135 \cdot 2020 - 5871139.831 = 621.63$

$$K_2 = 2.16 \cdot 25.55 \cdot \sqrt{1 + \frac{1}{16} + \frac{3(16+2 \cdot 2-1)^2}{16(16^2-1)}} = 63.6$$

$$621.63 - 63.6 = 558.03 ; 621.63 + 63.6 = 685.23$$

Интервальный прогноз:

t = 2020: (558.03;685.23)

Точечный прогноз, t = 2021: $y(2021) = -1.448 \cdot 2021^2 + 5832.135 \cdot 2021 - 5871139.831 = 601.67$

$$K_3 = 2.16 \cdot 25.55 \sqrt{1 + \frac{1}{16} + \frac{3(16+2 \cdot 3-1)^2}{16(16^2 - 1)}} = 65$$

$$601.67 - 65 = 536.67 ; 601.67 + 65 = 666.67$$

Интервальный прогноз:

t = 2021: (536.67;666.67)

Точечный прогноз, t = 2022: $y(2022) = -1.448 \cdot 2022^2 + 5832.135 \cdot 2022 - 5871139.831 = 578.82$

$$K_4 = 2.16 \cdot 25.55 \sqrt{1 + \frac{1}{16} + \frac{3(16+2 \cdot 4-1)^2}{16(16^2 - 1)}} = 66.5$$

$$578.82 - 66.5 = 512.32 ; 578.82 + 66.5 = 645.32$$

Интервальный прогноз:

t = 2022: (512.32;645.32)

Точечный прогноз, t = 2023: $y(2023) = -1.448 \cdot 2023^2 + 5832.135 \cdot 2023 - 5871139.831 = 553.07$

$$K_5 = 2.16 \cdot 25.55 \sqrt{1 + \frac{1}{16} + \frac{3(16+2 \cdot 5-1)^2}{16(16^2 - 1)}} = 68.1$$

$$553.07 - 68.1 = 484.97 ; 553.07 + 68.1 = 621.17$$

Интервальный прогноз:

t = 2023: (484.97;621.17)

Точечный прогноз, t = 2024: $y(2024) = -1.448 \cdot 2024^2 + 5832.135 \cdot 2024 - 5871139.831 = 524.42$

$$K_6 = 2.16 \cdot 25.55 \sqrt{1 + \frac{1}{16} + \frac{3(16+2 \cdot 6-1)^2}{16(16^2 - 1)}} = 69.78$$

$$524.42 - 69.78 = 454.64 ; 524.42 + 69.78 = 594.2$$

Интервальный прогноз:

$$t = 2024: (454.64; 594.2)$$

Точечный прогноз, $t = 2025$: $y(2025) = -1.448 \cdot 2025^2 + 5832.135 \cdot 2025 - 5871139.831 = 492.88$

$$K_7 = 2.16 \cdot 25.55 \sqrt{1 + \frac{1}{16} + \frac{3(16+2 \cdot 7-1)^2}{16(16^2 - 1)}} = 71.56$$

$$492.88 - 71.56 = 421.32 ; 492.88 + 71.56 = 564.44$$

Интервальный прогноз:

$$t = 2025: (421.32; 564.44)$$

Проверка гипотез относительно коэффициентов линейного уравнения регрессии.

1) t-статистика. Критерий Стьюдента.

По таблице Стьюдента находим $T_{\text{табл}}$

$$T_{\text{табл}}(n-m-1; \alpha/2) = (13; 0.025) = 2.16$$

Поскольку < 2.16 , то статистическая значимость коэффициента регрессии a_0 не подтверждается (принимая гипотезу о равенстве нулю этого коэффициента).

$$t_{a1} = \frac{5832.1347}{1359.5269} = 4.2898$$

Поскольку $4.2898 > 2.16$, то статистическая значимость коэффициента регрессии a_1 подтверждается (отвергаем гипотезу о равенстве нулю этого коэффициента).

$$t_{a2} = \frac{|-1.4482|}{0.3381} = 4.2832$$

Поскольку $4.2832 > 2.16$, то статистическая значимость коэффициента регрессии a_2 подтверждается (отвергаем гипотезу о равенстве нулю этого коэффициента).

Доверительный интервал для коэффициентов уравнения регрессии.

F-статистика. Критерий Фишера.

Коэффициент детерминации.

$$R^2 = 1 - \frac{\sum(y_i - \bar{y}_x)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{8488.6493}{48044} = 0.8233$$

$$F = \frac{R^2 (n - m - 1)}{1 - R^2 m} = \frac{0.8233 (16-2-1)}{1 - 0.8233 \cdot 2} = 30.2887$$

Находим из таблицы $F_{кр}(2;13;0.05) = 3.81$

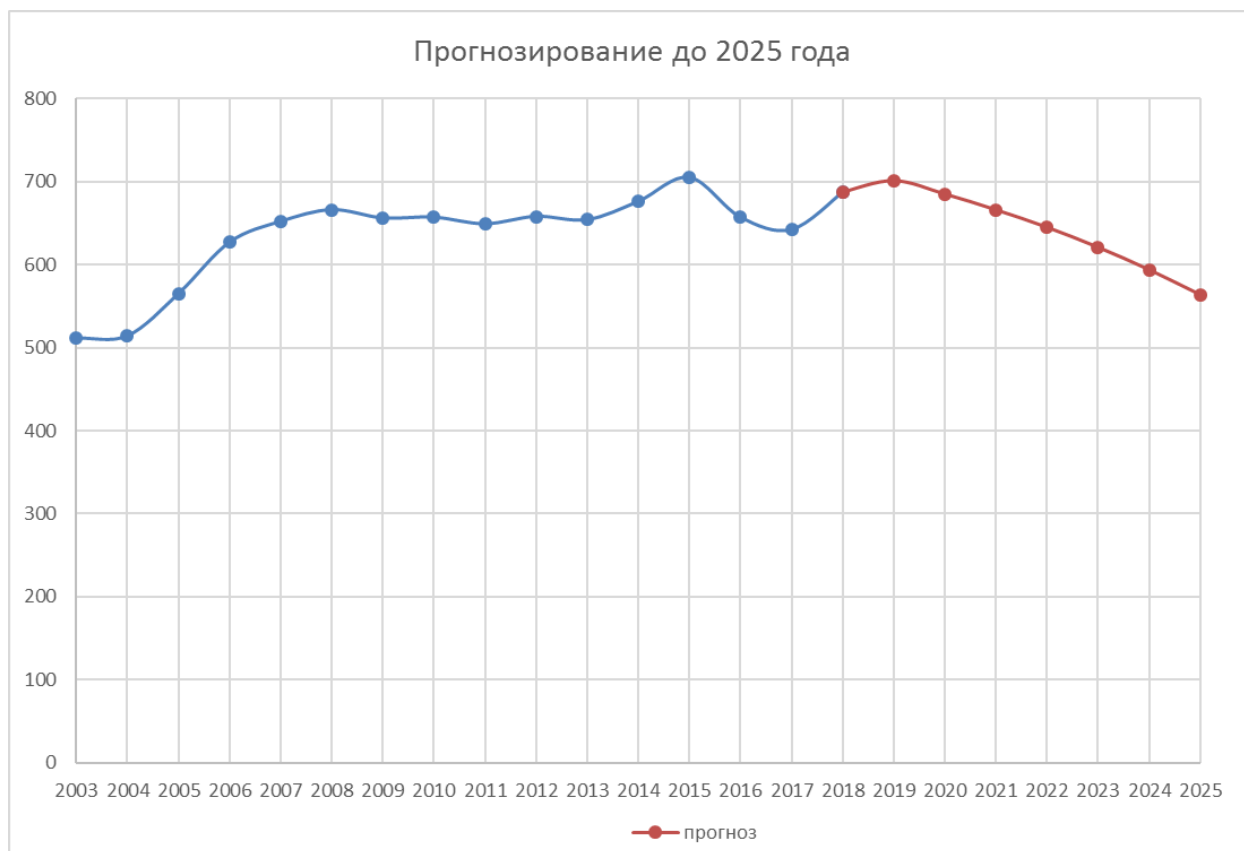
где m - количество факторов в уравнении регрессии ($m=2$).

Поскольку $F > F_{кр}$, то коэффициент детерминации (и в целом уравнение регрессии) статистически значим.

Прогнозирование до 2025 года

Годы	Фактические значения	Годы	Прогноз
2003	511	2019	701
2004	514	2020	685
2005	565	2021	666
2006	627	2022	645
2007	652	2023	621
2008	666	2024	594
2009	656	2025	564
2010	657		
2011	649		
2012	658		
2013	654		
2014	676		
2015	705		

2016	657		
2017	642		
2018	687		



Изучена зависимость Y от X . На этапе спецификации была выбрана парная параболическая регрессия. Оценены её параметры методом наименьших квадратов. Статистическая значимость уравнения проверена с помощью коэффициента детерминации и критерия Фишера. Установлено, что в исследуемой ситуации 82.33% общей вариабельности Y объясняется изменением X . Установлено также, что параметры модели статистически не значимы. Полученные оценки уравнения регрессии позволяют использовать его для прогноза.

Вывод

Регрессия	E	R ²	$\bar{A}$
Линейная	28,469	0,574	5,01%
Квадратичная	27,15	0,823	3,2%

По представленным результатам таблицы наилучшими характеристиками в данной задаче обладает квадратичная модель регрессии, так как коэффициент детерминации имеет наибольшее значение, а средняя ошибка аппроксимации и коэффициент эластичности принимают наименьшее из перечисленных моделей регрессии.

Заключение

В работа была рассмотрена одна из актуальнейших задач анализа данных – задача прогнозирования, одним из основных методом решения которой является регрессионный анализ. Поставленная в данном исследовании цель - с помощью регрессионного анализа построить модели пригодные для прогнозирования пассажирооборота Республики Тыва– была достигнута в полном объеме путем решения следующих задачи:

1. Изучение теории регрессионного анализа.
2. Обработка статистических данных для последующего анализа.
3. Обучение регрессионных моделей.
4. Оценка адекватности моделей.

После выполнения данной работы мы на практике можем применять навыки, полученные в ходе выполнения работы. Знаем основные методы построения моделей линейной и множественной регрессии, основные признаки неудовлетворительного качества экономических моделей, основные приемы повышения качества экономических моделей, методы прогнозирования значений моделируемой величины. Развили умение применять на практике математический аппарат регрессионного анализа для оценки параметров финансово-экономических моделей, проверять качество эконометрических моделей и анализировать их с помощью аппарата статистической проверки гипотез; определять ошибки спецификации моделей и возможности для повышения качества моделей, устранять причины, приводящие к низкому качеству моделей, использовать построенные эконометрические модели финансово-экономических объектов для прогнозирования экономической ситуации и выработки рекомендаций.

С помощью анализа значений различных функционалов ошибок для каждой из построенных моделей, сделан вывод о том, что квадратичная модель выдает наиболее точный прогноз на исследованном наборе данных. Результаты данного исследования могут быть использованы для предсказания пассажирооборота Республики Тыва.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Алексеев А.Р. Экономическая статистика : учебник для вузов / Алексеев А.Р., Воробьев А.Н., Громыко Г.Л ; под ред. Ю.Н. Иванова. - М.: ИНФРА-М, 2007. - 734 с. Текст: непосредственный.
2. Басовский Л.Е., Прогнозирование и планирование в условиях рынка, учебное пособие.- М.: ИНФРА-М, - 2002.-260с. Текст: непосредственный.
3. Бабешко Л.О. Основы эконометрического моделирования : учеб. пособие / Л. О. Бабешко. - Изд. 4-е. - М. : КомКнига, 2010. - 428 с. Текст: непосредственный. Загл. с экрана.-URL: <http://bookre.org/reader?file=489994> (дата обращения:15.12.2019). Текст: электронный.
4. Елисеева И.И., Эконометрика: учебник, 2е изд.- М.: Финансы и статистика, 2005.-576с. Текст: непосредственный.
5. Салин В.Н. Социально-экономическая статистика : учебник / В.Н.Салин, Е.П. Шпаковская. - М. : Юристъ, 2001. - 457 с. Текст: непосредственный.
6. Тихомиров Н. П. Эконометрика : учебник для вузов / Н. П. Тихомиров, Е. Ю. Дорохина. - М. : Экзамен, 2003. - 512 с. Загл. с экрана. -URL: <http://bookre.org/reader?file=813969> (дата обращения: 02.02.2020). Текст: электронный.
7. Эконометрика : учебник для вузов / под ред. Ю.Н. Иванова. - М. : ИНФРА-М, 2008. - 735 с. Текст: непосредственный.
8. Управление Федеральной службы государственной статистики по Красноярскому краю, Республике Хакасия и Республике Тыва: официальный сайт. -Красноярск. URL: <https://krasstat.gks.ru/about> (дата обращения: 23.12.2019)