Федеральное государственное автономное образовательное учреждение высшего образования

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ**

**«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»**

###### Факультет экономических наук

образовательная программа «Экономика»

предмет «Эконометрика-1»

##### Проект

«Эконометрическое исследование зависимости высоты главной новогодней елки города от показателей региона»

|  |  |
| --- | --- |
|  | Выполнили: |
|  | Студенты группы БЭК2110 |
|  | Некрасова Мария Алексеевна  Серочкин Егор Сергеевич  Руководитель:  Станкевич Иван Павлович |
|  |  |
|  |  |

Москва 2023г.

Содержание

1. Введение…………………………………………………………………………………………3
2. Экономическая модель…………………………………………………………………………4
3. Предварительный анализ данных…………………………………………………………..…5
4. Оценка моделей и выводы……………………………………………………………………..10
5. Приложение…………………………………………………………………………………….12
6. **Введение**

Цель исследования: создать модель, прогнозирующую высоту главной елки города.

Задачи исследования:

1. Обозначить актуальность работы;
2. Собрать cross-section данные;
3. Придумать и доказать или опровергнуть гипотезы;
4. Проанализировать данные и оценить модель.

Актуальность работы, на наш взгляд, заключается в нескольких причинах.

Во-первых, на носу Новый Год.

Во-вторых, уже с ноября каждый уважающий себя человек должен постепенно наращивать новогоднее настроение и предвкушение праздника, чтобы 31 декабря не обнаружить себя неспособным радоваться наступлению Нового Года и весь вечер и всю ночь сидеть за столом с кислой миной. О последствиях такого печального исхода событий знает каждый: как Новый Год встретишь, так его и проведешь. Значит, без новогоднего настроения и детского блеска в глазах, 2024 год станет для нас настоящим апогеем серости и грусти, а за ним и 2025, 2026, 2027 и так далее.

При этом нельзя отрицать тот факт, что психоэмоциональное благополучие населения напрямую влияет на экономический рост в стране[[1]](#footnote-1). Учитывая, что Новый Год является важным праздником у большинства людей в большинстве стран мира, вклад его в наше счастье далеко не мал. Значит, всем нам необходимо эмоционально готовиться к этому прекрасному празднику и встречать его полными радости и воодушевления во имя процветания Родины.

Никто не поспорит, что важнейшим атрибутом Нового Года является елка. Именно благодаря сверкающим огонькам гирлянд и ярким краскам елочных игрушек наше новогоднее настроение поднимается в разы! А чем выше будет эта самая елка, тем праздничнее и радостнее будет наше настроение! Именно по этой причине просто необходимо пристально отслеживать и прогнозировать высоту главной новогодней елки в каждом городе, ведь этот показатель напрямую отражает, насколько светлым и счастливым будет будущее страны.

1. **Экономическая модель**

Зависимая переменная: Высота новогодней елки на главной площади города в 2022 году (метры)

Объясняющие переменные (все данные за 2022 год):

* Индекс коррупции по странам (шкала от 0 до 100, где 0 – сплошная коррупция, 100 – коррупции в стране инет)
* Город-миллионник (дамми: 0 – нет, 1 – да)
* Основная религия – христианство (дамми: 0 – да, 1 – нет)
* Продолжительность новогодних каникул в стране (дни)
* Уровень развития страны (дамми: 0 – развивающаяся, 1 – развитая)
* ВВП страны (млн $ США)
* Количество туристов в стране (тысяч человек)

Источники данных:

а) Индекс коррупции[[2]](#footnote-2)

б) Количество туристов в 2022 году[[3]](#footnote-3)

в) ВВП в 2022 году[[4]](#footnote-4)

г) Основная религия, продолжительность новогодних каникул, уровень развития страны, города-миллионники: эта информация была найдена методом поиска в Интернете (например: "Какая основная религия в стране Х" или "Продолжительность новогодних каникул в 2022 году в стране Х" или "Является ли город Х миллионником")

Мы верим в то, что данный набор переменных позволит построить рабочую модель по предсказыванию высоты главной ели, хотя данных у нас не так уж и много. Пройдемся по каждой переменной, чтобы объяснить наше видение.

*Таблица 1*

Объясняющие переменные и их влияние на зависимую переменную (предположения)

|  |  |  |
| --- | --- | --- |
| **Объясняющая переменная** | **Предполагаемое влияние объясняющей переменной на зависимую переменную** | **Предполагаемый вид влияния на зависимую переменную** |
| Индекс коррупции в стране | Чем выше индекс коррупции, тем больше вероятность, что несмотря на большой потенциал наличия елки выше среднего, она будет более низкой, т.е. более дешевой, ведь бюджет на новогодние украшения города разворуют. | **-** |
| Город-миллионник | В больших городах бюджет больше, чем в маленьких, соответственно, и елка должна быть выше. | **+** |
| Основная религия – христианство | Елка – атрибут христианской рождественской символики, поэтому скорее всего, выше будут елки в христианских странах. | **+** |
| Продолжительность новогодних каникул | Чем более важным праздником в стране является Новый год / Рождество, тем дольше длятся выходные, тем выше должна быть елка. | **+** |
| Уровень развития страны | В более развитых странах скорее всего больше свободных средств в бюджете на украшения к праздникам, тем у развивающихся стран. Но это не точно. | **?** |
| ВВП страны | Чем выше ВВП страны, тем выше должна быть елка, так как ВВП отражает уровень экономической активности и качество жизни людей. Чем выше качество жизни, тем выше должна быть елка. Но это не точно. | **+** |
| Количество туристов в стране | Здесь, вероятно, есть зеркальная причинно-следственная связь: чем больше туристов, тем выше доход от туризма в стране, тем больше средств в бюджете на елку. С другой стороны, чем выше елка, тем больше человек приедут на нее посмотреть. | **+** |

Гипотезы:

**А**) Скорее всего индекс коррупции будет выше у развивающихся стран, поэтому мы хотим проверить, будет ли коэффициент перед переменной *индекс коррупции\*(уровень развития страны)* положительным. То есть, в развивающихся странах чем ниже индекс коррупции (т.е. в стране ее больше), тем меньше высота елки.

**Б**) Высота елки отрицательно зависит от переменной, отражающей религию в стране (Christ\_count\_0): если основная религия – не христианство (Christ\_count\_0 = 1), высота елки уменьшается по сравнению с христианскими странами (Christ\_count\_0 = 0).

1. **Предварительный анализ данных**

* Размерность данных: 50х9
* Анализ и интерпретация описательных статистик

*Таблица 2*

Минимальные, максимальные, средние и медианные значения некоторых переменных

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Переменная** | **Min** | **Max** | **Mean** | **Median** | **Интерпретация** |
| Высота елки, метры | 0 | 82,3 | 23,04 | 22,5 | Минимальное значение, равное нулю, говорит об отсутствии традиции ставить елки в наблюдаемой стране. Максимальное значение превышает среднее практически в 4 раза, что может свидетельствовать о наличии выбросов. |
| Индекс коррупции | 28 | 90 | 57,64 | 59,5 | Минимальное и максимальное значения примерно равно удалены от среднего, т.е. выбросов скорее всего нет. |
| Продолжительность новогодних каникул, дни | 1 | 17 | 5,7 | 4,5 | Кажется, что минимальное и максимальное значение намекают на наличие выбросов, т.к. разброс существенный. |
| ВВП страны, млн $ | 14.212 | 20.952.694 | 1.618.999 | 451.977 | Разброс данных ужасный, было решено использовать ln(ВВП), т.к. изначально (Рис. 1) распределение ВВП было похоже на логнормальное, а если взять логарифм, то станет нормальным. Так и произошло (Рис. 2) |
| Количество туристов, тыс. | 605 | 104.968 | 20.911 | 7.422 | Очевидно, тоже имеют место выбросы, но что поделать… И изначальное распределение тоже похоже на логнормальное (Рис. 3), попробуем взять ln(Кол-во туристов) (Рис. 4) |

* Анализ столбчатых диаграмм

|  |  |
| --- | --- |
| Рис. 1 «Гистограмма, похожая на логнормальное распределение ВВП» | Рис. 2 «Гистограмма, похожая на нормальное распределение ln(ВВП)» |

|  |  |
| --- | --- |
| Рис. 3 «Гистограмма, похожая на логнормальное распределение количества туристов по странам» | Рис. 4 «Гистограмма, похожая на нормальное распределение ln(Количество туристов)» |

* Проверка на выбросы

Уже из прошлого пункта, глядя на графики, становится ясно, что в данных есть выбросы. Проверим целевую и не-дамми-переменные на наличие выбросов с помощью ящиков с усами.

*Таблица 3*

Ящичковые диаграммы целевой переменной и non-dummy переменных

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

Выбросы наблюдаются в самой целевой переменной (высота елки), а также в переменной продолжительность каникул.

* Проверка корреляционных зависимостей с выбросами

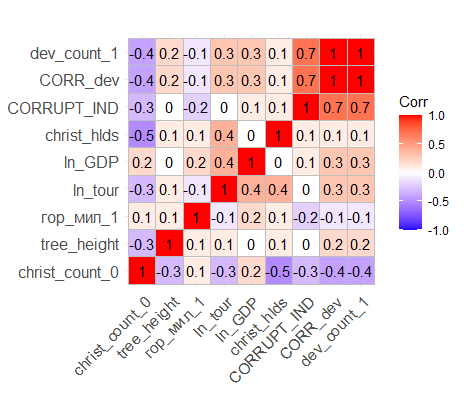


Рис. 5 «Корреляционная матрица до удаления выбросов»

Самым важным выводом из корреляционной матрицы является взаимосвязь объясняющих переменных с высотой елки (tree\_height). Из матрицы видно, что индекс коррупции в стране (CORRUPT\_IND) и ВВП (ln\_GDP) не влияют на высоту елки. Более всего на высоту елки оказывает влияние дамми-переменная, показывающая, является ли основная религия в стране христианством, причем высотка дерева зависит от нее отрицательно. И это логично, ведь за базовое значение (ноль) мы брали христианство, а за единицу – другие религии. Значит, в нехристианских странах, елка ниже, чем в христианских.

Дополнительно была введена переменная CORR\_dev (= CORRUPT\_IND \* dev\_count\_1), чтобы проверить гипотезу А. Видно, что высота елки положительно зависит от этой переменной, хотя от самой коррупции не зависит. Значит, в развивающихся странах чем выше коррупция (индекс коррупции ближе к нулю), тем ниже елка. Переменные город-миллионник, количество туристов, продолжительность каникул имеют положительное значение корреляции с высотой елки.

Для дальнейшего анализа и оценки гипотез были удалены следующий наблюдения:

1. Продолжительность каникул > 15 дней
2. Высота елки > 40 метров

* Проверка корреляционных зависимостей без выбросов

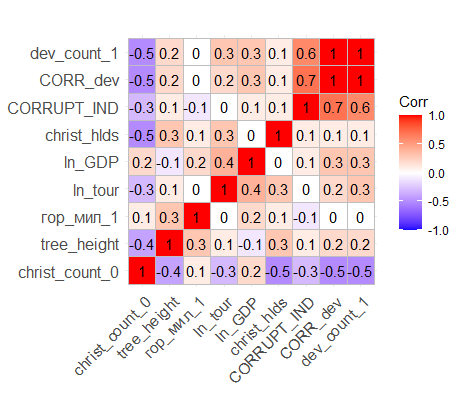


Рис. 6 «Корреляционная матрица после удаления выбросов»

Сразу видно, что после удаления выбросов взаимосвязь многих переменных с высотой елки увеличилась, а именно: религия-христианство (christ\_count\_0), город-миллионник (гор\_мил\_1), ВВП (ln\_GDP), продолжительность каникул (christ\_hlds), индекс коррупции (CORRUPT\_IND). С большой вероятностью можно сказать, что любая модель будет лучше предсказывать данные без выбросов.

Примечательно, что высота елки после очищения выборки от выбросов стала хоть и незначительно, но отрицательно зависеть от ln(ВВП) и положительно от индекса коррупции.

* Анализ диаграмм рассеивания

|  |  |
| --- | --- |
| Рис. 7 «Диаграмма рассеяния высоты елки и ln(ВВП)» | Рис. 8 «Диаграмма рассеяния высоты елки и ln(количество туристов)» |
| Рис. 9 «Диаграмма рассеяния высоты елки и CORR\_dev\*город-миллионник» | Рис. 10 «Диаграмма рассеяния высоты елки и индекса коррупции» |
| Рис. 11 «Диаграмма рассеяния высоты елки и продолжительности каникул» | Рис. 12 «Диаграмма рассеяния высоты елки и ln(количество туристов)\*город-миллионник» |

Диаграммы рассевания помогут понять, как различные переменные и их комбинации влияют на высоту елки, и есть ли между ними явная взаимосвязь (например, линейная, полиномиальная, логарифмическая и т.д.). Ввиду специфики данных, на диаграммах рассеяния сложно сходу определить тип взаимосвязи переменных. Также во многих данных наблюдается мультиколлинеарность.

1. **Оценка моделей**

Методом бесконечно долгого подбора потенциально значимых сочетаний переменных было обнаружено следующее: в замудренных моделях R2 больше, но в простеньких моделях с меньшим числом коэффициентов больше их значимость. В Приложении представлены некоторые результаты наших экспериментов.

Так как умные люди вообще не переживают об R2, мы решили последовать их примеру и выбрать для рассмотрения модель попроще, а именно:

*Высота елки ~ log(индекс коррупции\*(уровень развития страны) + 1)*

*+ религия-христианство\*город-миллионник + log(количество туристов)\*город-миллионник*

Оценив коэффициенты модели, получили следующий результат:

*Таблица 4*

Коэффициенты одной из моделей

|  |  |
| --- | --- |
|  | |
|  | Dependent variable: |
|  |  |
|  | tree\_height |
|  | |
| CORR\_dev \* гор\_мил\_1 | -0.058 |
|  | (0.055) |
|  |  |
| christ\_count\_0 \* гор\_мил\_1 | -11.416\*\*\* |
|  | (3.647) |
|  |  |
| гор\_мил\_1\* ln\_tour | 1.398\*\*\* |
|  | (0.378) |
|  |  |
| Constant | 15.884\*\*\* |
|  | (2.250) |
|  |  |
|  | |
| Observations | 46 |
| R2 | 0.300 |
| Adjusted R2 | 0.251 |
| Residual Std. Error | 8.880 (df = 42) |
| F Statistic | 6.014\*\*\* (df = 3; 42) |
|  | |
| Note: | \*p\*\*p\*\*\*p<0.01 |

Проинтерпретировать модель можно следующим образом:

* + В городах-миллионниках развивающихся стран чем выше коррупция, тем ниже елка, но это незначительно. *Гипотеза А подтвердилась!*
  + В городах-миллионниках стран, где христианство не является основной религией, высота елки будет на 11 метров ниже, чем в странах, где христианство – основная религия. *Гипотеза Б подтвердилась!*
  + Чем больше количество туристов в городе-миллионнике, тем выше главная елка этого города.
  + В городах, не являющихся миллионниками, высота елки составляет 15.8 метров.

Безусловно, предсказания модели далеко не идеальны. Например, у нас так и не получилось максимально приблизиться к тому, чтобы суметь предсказать высоту елки, равную 0, то бишь ее отсутствие. Возможно, при большем наборе данных и каких-нибудь других объясняющих переменных, этого удалось бы достичь.

*Приложение 1*

Сравнение некоторых замудренных моделей

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | | | | |
|  | Dependent variable: | | | | |
|  |  | | | | |
|  | tree\_height | | | | |
|  | (1) | (2) | (3) | (4) | (5) |
|  | | | | | |
| CORRUPT\_IND \* christ\_hlds | 0.008 | 0.004 | 0.004 |  |  |
|  | (0.019) | (0.019) | (0.019) |  |  |
|  |  |  |  |  |  |
| (christ\_count\_0 - 1) \* ln\_GDP |  | 0.543 | 0.891 | 0.886 | 0.533 |
|  |  | (0.681) | (2.468) | (2.398) | (0.670) |
|  |  |  |  |  |  |
| гор\_мил\_1 \* ln\_tour | 1.118 | 1.562 | 1.658 | 1.973 | 1.570 |
|  | (3.676) | (3.565) | (3.678) | (3.567) | (3.514) |
|  |  |  |  |  |  |
| гор\_мил\_1 \* ln\_GDP | 1.072 | 4.529 | 4.317 | 3.769 | 4.570 |
|  | (2.899) | (3.509) | (3.843) | (3.716) | (3.453) |
|  |  |  |  |  |  |
| log(CORR\_dev + 1) \* гор\_мил\_1 | -3.234\* | -3.901\*\* | -3.801\* | -3.036\* | -3.837\*\* |
|  | (1.819) | (1.784) | (1.933) | (1.755) | (1.728) |
|  |  |  |  |  |  |
| гор\_мил\_1 |  | -47.128 | -45.609 | -48.977 | -48.040 |
|  |  | (32.358) | (34.431) | (33.252) | (31.562) |
|  |  |  |  |  |  |
| (christ\_count\_0 + 1) \* christ\_hlds \* гор\_мил\_1 | -1.103 | -1.025 | -1.010 |  | -1.019 |
|  | (0.897) | (0.876) | (0.896) |  | (0.863) |
|  |  |  |  |  |  |
| christ\_count\_0 \* гор\_мил\_1 | -13.205 | -16.412\* | -16.065 | -13.786 | -16.004\* |
|  | (9.751) | (9.476) | (9.904) | (9.282) | (9.110) |
|  |  |  |  |  |  |
| christ\_count\_0 | 4.191 |  | -4.909 | -9.303 |  |
|  | (9.228) |  | (33.333) | (32.288) |  |
|  |  |  |  |  |  |
| CORR\_dev | 0.122 | 0.161\* | 0.160\* | 0.127\* | 0.166\*\* |
|  | (0.084) | (0.085) | (0.087) | (0.073) | (0.080) |
|  |  |  |  |  |  |
| ln\_tour | -0.415 | -0.393 | -0.449 | -0.589 | -0.409 |
|  | (3.381) | (3.301) | (3.373) | (3.288) | (3.253) |
|  |  |  |  |  |  |
| ln\_GDP | -1.689 | -4.219 | -3.811 | -3.425 | -4.267 |
|  | (2.884) | (3.014) | (4.128) | (4.012) | (2.961) |
|  |  |  |  |  |  |
| christ\_hlds | 0.596 | 0.751 | 0.748 |  | 0.961 |
|  | (1.413) | (1.390) | (1.411) |  | (0.868) |
|  |  |  |  |  |  |
| Constant | 31.516\*\* | 68.995\*\* | 68.936\*\* | 71.916\*\* | 69.641\*\* |
|  | (13.882) | (28.284) | (28.716) | (27.767) | (27.689) |
|  |  |  |  |  |  |
|  | | | | | |
| Observations | 46 | 46 | 46 | 46 | 46 |
| R2 | 0.391 | 0.432 | 0.432 | 0.408 | 0.431 |
| Adjusted R2 | 0.195 | 0.225 | 0.201 | 0.239 | 0.247 |
| Residual Std. Error | 9.205 (df = 34) | 9.029 (df = 33) | 9.166 (df = 32) | 8.947 (df = 35) | 8.900 (df = 34) |
| F Statistic | 1.988\* (df = 11; 34) | 2.090\*\* (df = 12; 33) | 1.873\* (df = 13; 32) | 2.414\*\* (df = 10; 35) | 2.342\*\* (df = 11; 34) |
|  | | | | | |
| Note: | \*p\*\*p\*\*\*p<0.01 | | | | |

*Приложение 2*

Сравнение более простых моделей

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | |
|  | Dependent variable: | | | | | | |
|  |  | | | | | | |
|  | tree\_height | | | | | | |
|  | (1) | (2) | (3) | (4) | (5) | (6) | (7) | |
|  | | | | | | | |
| I(log(CORR\_dev + 1) \* гор\_мил\_1) | -1.050 | -1.044 | -0.374 |  | -0.941 | -0.971 |  | |
|  | (0.976) | (0.965) | (0.926) |  | (0.966) | (0.965) |  | |
|  |  |  |  |  |  |  |  | |
| гор\_мил\_1 | 3.901 |  |  |  | 13.097\*\*\* | 14.481\*\*\* | 12.195\*\* | |
|  | (12.188) |  |  |  | (3.783) | (5.221) | (4.701) | |
|  |  |  |  |  |  |  |  | |
| I((christ\_count\_0 + 1) \* christ\_hlds) |  |  |  |  | -0.011 |  |  | |
|  |  |  |  |  | (0.420) |  |  | |
|  |  |  |  |  |  |  |  | |
| I((christ\_count\_0 + 1) \* christ\_hlds \* гор\_мил\_1) |  |  |  |  |  | -0.186 | -0.144 | |
|  |  |  |  |  |  | (0.489) | (0.488) | |
|  |  |  |  |  |  |  |  | |
| I(christ\_count\_0 \* гор\_мил\_1) | -10.354\* | -10.120\* |  |  | -12.231\*\*\* | -12.504\*\*\* | -10.469\*\*\* | |
|  | (5.445) | (5.336) |  |  | (3.928) | (3.941) | (3.383) | |
|  |  |  |  |  |  |  |  | |
| I(christ\_count\_0 \* christ\_hlds) | -0.505 | -0.465 | -2.200\*\* |  |  |  |  | |
|  | (1.292) | (1.271) | (0.910) |  |  |  |  | |
|  |  |  |  |  |  |  |  | |
| I(гор\_мил\_1 \* ln\_tour) | 0.975 | 1.370\*\*\* | 0.945\*\*\* |  |  |  |  | |
|  | (1.301) | (0.403) | (0.345) |  |  |  |  | |
|  |  |  |  |  |  |  |  | |
| CORRUPT\_IND |  |  |  | 0.183 |  |  |  | |
|  |  |  |  | (0.179) |  |  |  | |
|  |  |  |  |  |  |  |  | |
| I(CORRUPT\_IND \* christ\_hlds) |  |  |  | -0.031 |  |  |  | |
|  |  |  |  | (0.035) |  |  |  | |
|  |  |  |  |  |  |  |  | |
| christ\_hlds |  |  |  | 1.150 |  |  |  | |
|  |  |  |  | (2.557) |  |  |  | |
|  |  |  |  |  |  |  |  | |
| I(christ\_hlds \* ln\_tour) |  |  |  | 0.121 |  |  |  | |
|  |  |  |  | (0.218) |  |  |  | |
|  |  |  |  |  |  |  |  | |
| Constant | 16.107\*\*\* | 16.221\*\*\* | 17.379\*\*\* | 7.686 | 15.796\*\*\* | 15.736\*\*\* | 15.736\*\*\* | |
|  | (2.525) | (2.472) | (2.469) | (9.486) | (3.315) | (2.332) | (2.333) | |
|  |  |  |  |  |  |  |  | |
|  | | | | | | | |
| Observations | 46 | 46 | 46 | 46 | 46 | 46 | 46 | |
| R2 | 0.306 | 0.304 | 0.243 | 0.093 | 0.291 | 0.293 | 0.276 | |
| Adjusted R2 | 0.219 | 0.236 | 0.189 | 0.004 | 0.222 | 0.224 | 0.224 | |
| Residual Std. Error | 9.065 (df = 40) | 8.966 (df = 41) | 9.239 (df = 42) | 10.236 (df = 41) | 9.049 (df = 41) | 9.033 (df = 41) | 9.034 (df = 42) | |
| F Statistic | 3.522\*\*\*  (df = 5; 40) | 4.474\*\*\*  (df = 4; 41) | 4.489\*\*\*  (df = 3; 42) | 1.046  (df = 4; 41) | 4.205\*\*\*  (df = 4; 41) | 4.256\*\*\*  (df = 4; 41) | 5.336\*\*\*  (df = 3; 42) | |
|  | | | | | | | |
| Note: | \*p\*\*p\*\*\*p<0.011 |

1. <https://ojs.stanford.edu/ojs/index.php/intersect/article/download/2668/1577/9680> [↑](#footnote-ref-1)
2. <https://images.transparencycdn.org/images/Report_CPI2022_English.pdf> [↑](#footnote-ref-2)
3. <https://www.unwto.org/tourism-statistics/key-tourism-statistics> [↑](#footnote-ref-3)
4. <https://data.worldbank.org/indicator/NY.GDP.MKTP.CD> [↑](#footnote-ref-4)