

Федеральное государственное автономное образовательное учреждение высшего образования

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет экономических наук, образовательная программа «Экономика»

**Домашняя работа по курсу «Эконометрика - 1»**

**Как цена страховки жизни зависит от различных факторов**

Выполнили:

Черников Олег Владимирович, БЭК2111

Томинец Александр Владимирович, БЭК2111

Преподаватель:

Демьяненко Артем Владимирович

Москва 2023

1. **Введение**

В современном мире практически каждый человек имеет какую-либо страховку, будь то страховка имущества или страховка здоровья ([Population coverage for health care](https://www.oecd-ilibrary.org/docserver/health_glance-2017-24-en.pdf?expires=1702117536&id=id&accname=guest&checksum=28DB78C2FEEB95048E08F964380E735D)). По этой причине мы решили провести соответствующее исследование – как стоимость страховки зависит от различных фактов о человеке.

* 1. **Данные**

Данные были взять с сайта Kaggle ([Medical Cost Personal Datasets (kaggle.com)](https://www.kaggle.com/datasets/mirichoi0218/insurance)), которые были собраны в 2016 году Мири Чой. Описания того, как и где были собраны эти данные, к сожалению, приведены не были. В датасете 1338 наблюдений.

* 1. **Экономическая модель**

1. Список объясняющих переменных:
2. age – Возраст человека. Числовая переменная.
3. sex – Пол человека (male / female). Категориальная переменная.
4. bmi – индекс массы тела (кг/м^2). Числовая переменная.
5. сhildren – количество детей (от 0 до 5). Категориальная переменная.
6. smoker – курит ли человек (yes / no). Категориальная переменная.
7. region – регион бенефициара в Соединенных Штатах (southwest, southeast, northwest, northeast). Категориальная переменная.
8. Почему был выбран именно этот набор объясняющих переменных:

Во статьях, которые мы просмотрели, используется практически идентичный набор переменных ([How Age Affects Life Insurance Rates](https://www.investopedia.com/articles/personal-finance/022615/how-age-affects-life-insurance-rates.asp), [Ahmed I. Taloba, 2022](https://www.hindawi.com/journals/jhe/2022/7969220/)), за исключением некоторых отличий, таких как добавление, например, информации о хобби и профессии.

1. **Формулировка гипотез**

* Во-первых, на стоимость страховки должен положительно влиять возраст приобретающего, в приведенной статье возраст увеличивает стоимость страховки примерно на 8-10% ([How Age Affects Life Insurance Rates](https://www.investopedia.com/articles/personal-finance/022615/how-age-affects-life-insurance-rates.asp)).
* Во-вторых курящие – группа риска, которые так же должны заплатить дополнительную премию за риск ([Difference in lifetime medical expenditures between male smokers and non-smokers](https://www.sciencedirect.com/science/article/abs/pii/S0168851009002279?fr=RR-2&ref=pdf_download&rr=832cce390af67a64)). Поэтому мы решили проверить соответствующие гипотезы.

1. **Предварительный анализ данных**

В нашем наборе данных большинство независимых переменных – дамми-переменные, поэтому в начале обработаем их. Использовалась следующая схема кодировки:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
| children\_0 | 1 | 0 | 0 | 0 | 0 | 0 |
| children\_1 | 1 | 1 | 0 | 0 | 0 | 0 |
| children\_2 | 1 | 1 | 1 | 0 | 0 | 0 |
| children\_3 | 1 | 1 | 1 | 1 | 0 | 0 |
| children\_4 | 1 | 1 | 1 | 1 | 1 | 0 |
| children\_5 | 1 | 1 | 1 | 1 | 1 | 1 |

В данном случае отсутствие детей, когда признак children принимает значение 0, принимается, как базовая категория.

Такая кодировка позволяет интерпретировать коэффициенты регрессии относительно предыдущей дамми-переменной, а не относительно базовой.

Здесь за базовый регион принимается northeast.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | |
| Statistic | N | Mean | St. Dev. | Min | Median | Max |
|  | | | | | | |
| age | 1,338 | 39.207 | 14.050 | 18 | 39 | 64 |
| bmi | 1,338 | 30.663 | 6.098 | 15.960 | 30.400 | 53.130 |
| charges | 1,338 | 13,270.420 | 12,110.010 | 1,121.874 | 9,382.033 | 63,770.430 |
| sex\_male | 1,338 | 0.505 | 0.500 | 0 | 1 | 1 |
| children\_1 | 1,338 | 0.571 | 0.495 | 0 | 1 | 1 |
| children\_2 | 1,338 | 0.329 | 0.470 | 0 | 0 | 1 |
| children\_3 | 1,338 | 0.149 | 0.357 | 0 | 0 | 1 |
| children\_4 | 1,338 | 0.032 | 0.176 | 0 | 0 | 1 |
| children\_5 | 1,338 | 0.013 | 0.115 | 0 | 0 | 1 |
| smoker\_yes | 1,338 | 0.205 | 0.404 | 0 | 0 | 1 |
| region\_northwest | 1,338 | 0.243 | 0.429 | 0 | 0 | 1 |
| region\_southeast | 1,338 | 0.272 | 0.445 | 0 | 0 | 1 |
| region\_southwest | 1,338 | 0.243 | 0.429 | 0 | 0 | 1 |
|  | | | | | | |

В таблице выше представлены основные описательные статистики по всем, в том числе закодированным, признакам. Из нее мы получаем много полезной информации о признаках. Например, распределение age похоже на равномерное со средним равным 40. Это значит, что в нашей выборке представлены примерно одинаково представители всех возрастных категорий от 18 до 64 лет. Интересно заметить, что среднее и медианное значение bmi находиться в районе 30, что свидетельствует об ожирении ([Body Mass Index](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4890841/)). Можно предполагать, следовательно, что выборка смещена в сторону более полных людей. Среднее значение и медиана у целевой переменной charges отличаются, что свидетельствует о вероятном наличии выбросов в данных. Однако в анализе расходов или доходов это проблема всегда присутствует. Вероятно, есть смысл логарифмировать данные, чтобы приблизить распределение к нормальному. Из строчки с sex\_male мы узнаем из среднего значения, что в выборке представлено равное количество мужчин и женщин. Следующие пять строк касаются количества детей у бенефициариев. Из-за особенности кодировки переменных отвечающих за количество детей результаты описательных статистик интерпретируются особым образом. Так, у 57% респондентов есть 1 или более детей, у 33% два ребенка или больше и так далее. 43% людей не имеют детей. Следующая строчка свидетельствует о том, что курящих в данных в 4 раза меньше, чем некурящих, всего 20%. Последние 3 строчки посвящены признакам, связанным с местоположением бенефициариев. В наборе данных представлено примерно равное количество людей из всех 4 регионов.

Далее давайте посмотрим на корреляционную матрицу и выясним, как наши признаки связаны между собой.

Изображение выглядит как текст, снимок экрана, прямоугольный, Красочность

Автоматически созданное описание

Из корреляционной матрицы видно, что между переменной smoker\_yes и переменной charges наблюдается высокая положительная корреляция. Так как это зависимая и независимая переменная то не возникает никаких проблем, а лишь подтверждается одна из сформулированных ранее гипотез. Из-за того, что мы кодировали переменную children с накопительным эффектом, дамми-переменные сильно коррелируют между собой, но проблем с этим не должно возникнуть. Остальные регрессоры имеют несущественную корреляцию, либо не имеют ее вообще.

Теперь посмотрим на основные переменные и зависимости между ними более подробно, используя диаграммы рассеяния, гистограммы и ящики с усами.

Изображение выглядит как текст, диаграмма, линия, График

Автоматически созданное описание

Изображение выглядит как диаграмма, линия, График, текст

Автоматически созданное описание

Первым делом посмотрим на распределение целевой переменной charges. Первая гистограмма показывает, что без какой-либо обработки charges имеет очень длинный хвост и своим видом напоминает экспоненциальное распределение. Следовательно, в данных наблюдаются выбросы. Чтобы данные стали походить на нормально распределенные их можно прологарифмировать. Результат показан на 2 графике выше. Полученное распределение стало более «нормальным», хотя имеет что-то похожее на вторую моду в районе между 10 и 11. Есть смысл оценить модель с логарифмом от зависимой переменной.

Изображение выглядит как диаграмма, текст, снимок экрана, линия

Автоматически созданное описание

Следующим посмотрим на зависимость между charges и smoker, то есть как курение влияет на страховые расходы. Из ящика с усами очевидно, что курение очень сильно влияет на результирующую сумму. Медиана charges у курящих находиться намного выше, чем у некурящих. Этот вывод также подтверждается высокой корреляцией, которая обсуждалась выше. Отсюда также следует, что наши данные неоднородны, потому что разница между выборками с курящими и некурящими очень существенная.

Изображение выглядит как диаграмма, График, линия, снимок экрана

Автоматически созданное описание

График выше показывает распределение charges для 2 групп: курящих и некурящих в отдельности. Видно, что среднее и медианное значение для распределения курящих будет больше, чем для некурящих. Интересно также заметить, что распределение расходов для курящих является бимодальным. Скорее всего, на это влияет какой-то 3 признак, например, возраст или bmi, то есть для более старых и курящих людей более высокие ставки, чем для молодых курящих.

Изображение выглядит как линия, диаграмма, График, снимок экрана

Автоматически созданное описание

Из точечной диаграммы зависимости bmi и log(charges) видно, что облако точек курящих сильно выше такового для некурящих, и их тренды также сильно разняться. Опять же подтверждается теория о неоднородности наших данных.

Изображение выглядит как текст, линия, График, снимок экрана

Автоматически созданное описание

Точечная диаграмма age и log(charges) показывает те же самые результаты, что и диаграмма с bmi.

Остальные диаграммы не показали значимых результатов.

1. **Оценка моделей**

Поскольку эмпирическим путем была замечена неоднородность данных, будут оценены три отдельные модели, а также будет проведен тест Чоу для статистического выявления принадлежности выборок к одной генеральной совокупности.

Модель для всей совокупности имеет следующую формулу:

Модель только для курильщиков:

Модель только для некурящих:

|  |  |  |  |
| --- | --- | --- | --- |
| **Regression Results** | | | |
|  | | | |
|  | Dependent variable: | | |
|  |  | | |
|  | log(charges) | | |
|  | Full | Smokers only | No Smokers |
|  | | | |
| age | 0.034\*\*\* | 0.009\*\*\* | 0.042\*\*\* |
|  | (0.002) | (0.001) | (0.001) |
|  |  |  |  |
| bmi | 0.011\*\*\* | 0.049\*\*\* | 0.001 |
|  | (0.004) | (0.002) | (0.002) |
|  |  |  |  |
| sex\_male | 0.020 | -0.003 | -0.102\*\*\* |
|  | (0.042) | (0.024) | (0.025) |
|  |  |  |  |
| children\_1 | 0.126\*\* | -0.028 | 0.206\*\*\* |
|  | (0.053) | (0.031) | (0.032) |
|  |  |  |  |
| children\_2 | 0.203\*\*\* | 0.050 | 0.153\*\*\* |
|  | (0.065) | (0.037) | (0.040) |
|  |  |  |  |
| children\_3 | 0.0002 | -0.005 | -0.033 |
|  | (0.079) | (0.041) | (0.049) |
|  |  |  |  |
| children\_4 | 0.066 | -0.060 | 0.265\*\*\* |
|  | (0.165) | (0.118) | (0.096) |
|  |  |  |  |
| children\_5 | -0.206 | 0.189 | -0.051 |
|  | (0.237) | (0.226) | (0.134) |
|  |  |  |  |
| region\_northwest | -0.115\* | -0.017 | -0.080\*\* |
|  | (0.060) | (0.035) | (0.036) |
|  |  |  |  |
| region\_southeast | -0.085 | -0.060\* | -0.175\*\*\* |
|  | (0.061) | (0.033) | (0.037) |
|  |  |  |  |
| region\_southwest | -0.167\*\*\* | -0.014 | -0.183\*\*\* |
|  | (0.060) | (0.036) | (0.036) |
|  |  |  |  |
| Constant | 7.378\*\*\* | 8.463\*\*\* | 7.100\*\*\* |
|  | (0.125) | (0.068) | (0.076) |
|  |  |  |  |
|  | | | |
| Observations | 1,338 | 274 | 1,064 |
| R2 | 0.312 | 0.761 | 0.695 |
| Adjusted R2 | 0.306 | 0.751 | 0.692 |
| Residual Std. Error | 0.766 (df = 1326) | 0.193 (df = 262) | 0.413 (df = 1052) |
| F Statistic | 54.707\*\*\* (df = 11; 1326) | 76.034\*\*\* (df = 11; 262) | 218.128\*\*\* (df = 11; 1052) |
|  | | | |

* 1. **Тест Чоу**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| RSSFull | RSS\_smoker\_yes | RSS\_smoker\_no | n | k | Fstat | Fcrit | p\_val |
| 777.58 | 9.78 | 179.47 | 1338 | 12 | 340.42 | 1.76 | 0 |

* 1. **Интерпретация коэффициентов моделей**

1. Выборка из курящих

* Age – при увеличении на 1 год, charges увеличиваются на 0,9%
* Bmi – при увеличении на 1, charges увеличиваются на 4,9%
* Sex\_male – неинтерпритируемая
* Children\_1, Children\_2, Children\_3, Children\_4 , Children\_5 – неинтерпритируемы

1. Выборка из некурящих

* Age – при увеличении на 1 год, charges увеличиваются на 4,2%
* Bmi – неинтерпритируемая
* Sex\_male – сharges меньше на 10,2% для мужчин
* Children\_1 – charges на 20,6% выше для людей с одним ребенком относительно бездетных
* Children\_2 – charges на 15,3% выше для людей с двумя детьми относительно людей с одним ребенком
* Children\_3 – неинтерпретируем
* Children\_4 – charges на 26,5% выше для людей с четырьмя детьми относительно людей с тремя детьми
* Children\_5 – неинтерпретируем
* Region\_northwest – charges меньше на 8% для жителей северо-запада относительно жителей северо-востока
* Region\_southeast – charges меньше на 17,5% для жителей юго-восток относительно жителей северо-востока
* Region\_southwest – charges меньше на 18,3% для жителей юго-запад относительно жителей северо-востока
  1. **Выводы по проделанной работе**

1. Гипотеза о влияние возраста на стоимость страховки

Результаты модели подтвердили нашу гипотезу о влияние возраста, так как возраст значим во всех моделях, то мы можем проинтерпретировать полученные коэффициенты. Однако полученные нами результаты разняться для 2 групп, курящих и некурящих. Для первой группы возраст оказывает очень слабое влияние, всего 0.9%, на расходы, вероятно, страховые компании, получая информацию о статусе курильщика, не так сильно обращают внимание на возраст, как на факторы здоровья, что подтверждается значимостью и высоким влиянием bmi. В тоже время возраст играет важную роль в модели для некурящих. Оценка регрессии показывает, что для данной категории увеличение возраста на 1 год ведет к увеличению charges на 4.2%, что существенно выше в сравнении с курильщиками. Оба результата, тем не менее, не соответствуют описанным в статье ([How Age Affects Life Insurance Rates](https://www.investopedia.com/articles/personal-finance/022615/how-age-affects-life-insurance-rates.asp)). Предполагаемые 8-10% намного выше наблюдаемых в наших моделях.

1. Гипотеза о курящих

Неоднородность, выявленная в ходе работы, подтвердила нашу гипотезу о том, что курящие вынуждены доплачивать дополнительную премию за риск страховым компаниям.

1. Влияние детей

В ходе оценки модели для некурящих респондентов была выявлена неоднозначность в результатах относительно количества детей. Коэффициенты при одном ребенке и двух детях являются значимыми и положительными, что говорит о «штрафе за материнство», то есть большее количество детей сопровождается большими расходами по страховке. Однако следующая переменная, отвечающая за наличие троих детей, оказывается незначима. В то же время четвертый коэффициент оказывается положительным и значимым. Последний пятый опять незначим. Отсюда можно сделать предположение о том, что страховые компании разбивают количество детей на категории по типу 1 ребенок, 2 ребенка и многодетные.

1. Общий вывод по моделям

Исходя из всего вышенаписанного можно сделать финальное заключение о том, что страховые компании, оценивая размер страхового взноса для курящих людей, обращают внимание в первую очередь на показатели здоровья. В это же время для некурящих эти факторы играют куда меньшую роль в определении размеров выплат. Для последней группы существенны такие факторы, как возраст, количество детей, пол и регион проживания.