Анализ реальных данных на примере РЛМС

AleevaV

22 01 2020

Исследователь Вовочка решил попробовать свои силы и провести исследование, чтобы выявить факторы, влияющие на доход людей, на реальных данных. Но один он не может справиться, поэтому Вам необходимо помочь Вовочке в этом.

В рамках данного задания мы будем проводить небольшое эконометрическое исследование на данных Российского мониторинга экономического положения и здоровья населения (РМЭЗ, он же RLMS). Опишем предварительные шаги, которые Вовочке и Вам надо будет сделать, прежде чем приступать к работе с данными.

1. Первый шаг — это получить данные. Данные РМЭЗ являются открытыми и после простой процедуры регистрации доступны для скачивания на сайте Высшей Школы Экономики. В рамках этого задания Вовочке посоветовали взять репрезентативную выборку по индивидам за 2013 год - волна 22 (нужный файл называется r22i\_os\_31.sav).

NB! Если на сайте РМЭЗ при скачивании репрезентативной выборки, скачалась полная (r22iall\_31.sav), то не стоит паниковать, так как из полной выборки сделать репрезентативную можно одной командой, отобрав наблюдения по условию origsm=1 :) 2) Для загрузки данных в R Вовочке посоветовали воспользоваться пакетом rlms. Описание и инструкция по установке. Пакет rlms автоматически выполняет несколько процедур по очистке данных. Если загружать данные без его использования, то необходимо перекодировать все пропуски в данных (“нет ответа”, “затрудняюсь ответить” и т.д.) в NA самостоятельно.

Замечание. Если функция rlms\_read выдаёт ошибку при чтении файла, откатите пакет haven до более старой версии. Версия 1.1.1. пакета haven перестала читать некоторые sav файлы. Разработчики пакета haven в курсе этой ошибки и мы надеемся, что её в ближайшем будущем поправят. Откатиться можно с помощью команды: devtools::install\_version(“haven”, version = “1.1.0”). Либо скачать архивную версию пакета руками, разархивировать её, и установить командой devtools::install\_local(“папка\_с\_разархивированным\_пакетом”).

1. Для исследования Вовочка составил массив данных с отобранными переменными:

Заработной платой на основе вопроса rj13.2 Возрастом на основе переменной rh6 и года проведения опроса Полом на основе переменной rh5 Образованием на основе переменной r\_diplom Типом населенного пункта на основе переменной status Удовлетворенностью работой в целом на основе переменной rj1.1.1. Чтобы получить такой же массив, есть ряд моментов, нужно сделать следующие шаги:

1. Подгрузить пакеты и файл.
2. Отобрать только 6 переменных, описанных выше.
3. Создать переменную возраст. Год рождения респондента (переменная rh6) - числовая переменная, чтобы перейти от года рождения к возрасту нужно, как несложно догадаться, вычесть из года проведения опроса (2013) год рождения респондента (возможными расхождениями, связанными с месяцем рождения, можно пренебречь).
4. Отобрать только два типа населённого пункта: город и областной центр (должно остаться 10692 записи).
5. Отобрать только две категории степени удовлетворенности работой в целом: полностью удовлетворен и скорее удовлетворен (должно остаться 9089 записей).
6. Отобрать только тех людей, у кого образование входит в одну из следующих категорий:

Категория 1: окончил 0 - 6 классов; Категория 2: незаконченное среднее образование (7 - 8 кл); Категория 3: незаконченное среднее образование (7 - 8 кл) + что-то еще; Категория 4: законченное среднее образование; Категория 5: законченное среднее специальное образование; Категория 6: законченное высшее образование и выше; То есть таким образом Вы выбросите варианты “Затрудняюсь ответить”, “Отказ от ответа” и “Нет ответа” (должно остаться 9089 записей).

1. Из переменной тип населенного пункта сделать дамми-переменную, равную 1 для города и 0 для областного центра.
2. Из переменной удовлетворённость работой сделать дамми-переменную, равную 1 для полностью удовлетворен и 0 для скорее удовлетворен.
3. Из переменной пол сделать дамми-переменную, равную 1 мужчин и 0 для женщин.
4. Нет оснований считать, что эффект, к примеру, получения образования в виде 7-8 классов по сравнению с 0-6 классами такой же, как у получения высшего образования по сравнению со средним специальным образованием (оба примера - соседние категории), поэтому переменную образование необходимо превратить в набор фиктивных переменных. Использовать будем следующие категории:

Незаконченное среднее образование (категории от 1 до 3 из пункта 6) Законченное среднее образование (категория 4 из пункта 6) Законченное среднее специальное образование (категория 5 из пункта 6) Законченное высшее образование (категория 6 из пункта 6) В итоге Вы должны получить 4 фиктивные переменные, отвечающие за принадлежность респондента к одной из этих категорий.

1. В полученном массиве данных должно быть 9089наблюдений.

## [1] 9089 9

1. Создать массив данных, очищенный от пропущенных наблюдений, NA. Таким образом, у Вас должно получиться 2984 наблюдения в массиве без NA! Если оно другое, значит, где-то ошибка и надо пересмотреть предыдущие пункты.

## [1] 2984 9

Теперь, когда данные скачаны, загружены в R и отобраны в массив, самое сложное уже позади! Вы можете приступать к выполнению задания, но не забудьте загрузить необходимые пакеты! :)

Задания 1 и 2 нужно делать по массиву данных с 9089 наблюдениями, то есть с массивом, где есть NA :)

1. Исследователь Вовочка захотел более подробно изучить возрастную структуру полученных им данных и обнаружил, что максимальный возраст респондента в его выборке равен

## [1] 100

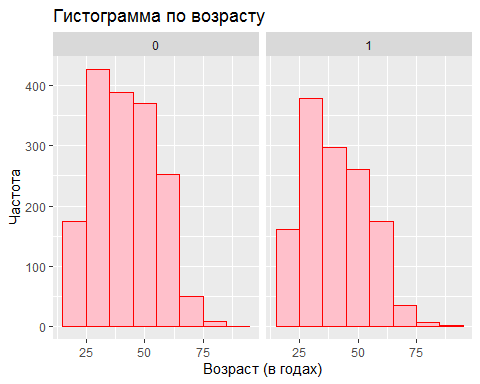
1. Как настоящий исследователь Вовочка решил проверить, сколько пропущенных значений в переменной, которая представляет для него главный интерес, в заработной плате. Оно равно

## [1] 6082

Поняв, что так дело не пойдет, Вовочка как любой неопытный исследователь, решил бороться с пропусками простым и робастным методом, просто очистил данные от них :) Сделаем так и мы. То есть здесь и далее будем работать только с массивом из 2984 наблюдений, где нет NA.

3.Вовочка решил показать на графике, какое же распределение имеет заработная плата. Посмотрев на гистограмму, Вовочка сделал следующий вывод

* Абсолютное большинство респондентов имеет доход до 50 тысяч рублей

4.Ниже представлены два графика, которые изобразил Вовочка. Деление на 0 и 1 Вовочка сделал по переменной пол. Но он забыл подписать ось X.  5. Теперь Вовочка решил оценить влияние возраста, пола, 3 фиктивных переменных на образование (не включая фиктивную переменную на самый низкий уровень образования), фиктивной переменной на проживание в городе и удовлетворенности работой на заработную плату респондентов. Сначала он решил проверить регрессию на значимость в целом и получил следующее значение тестовой статистики:

## value numdf dendf   
## 59.08497 7.00000 2976.00000

6.Какой вывод сделал Вовочка относительно значимости регрессии в целом? - На любом разумном уровне значимости гипотеза о равенстве всех коэффициентов нулю отвергается. То есть как минимум один из коэффициентов значим и оказывает влияние на доход.

1. При проверке гипотезы о значимости коэффициента при переменной пол Вовочка хочет сравнить значение наблюдаемой статистики с критическим значением для уровнем значимости 5%, какое значение наблюдаемой статистики Вовочка должен использовать?

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 16023.79597 1657.61298 9.666790 8.692698e-22  
## sex 9507.86026 636.04718 14.948357 8.997800e-49  
## age -67.60084 23.57581 -2.867381 4.167942e-03  
## educ\_m 2255.56558 1388.22524 1.624784 1.043146e-01  
## educ\_ms 4588.67910 1399.35093 3.279148 1.053196e-03  
## educ\_h 11796.24920 1359.31435 8.678088 6.510636e-18  
## city -2803.14793 636.33841 -4.405121 1.094562e-05  
## satisf 1326.23648 704.54135 1.882411 5.987770e-02

1. Сравнив наблюдаемое значение статистики с различными критическими значениями, Вовочка пришел к следующему выводу:

-На уровне значимости 1%, 5% и 10% коэффициент при дамми на пол значим. Мужчины получают в среднем выше доход, чем женщины

9.Вовочка решил посмотреть робастное стандартное отклонение для дамми на тип населённого пункта. Оно равно:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | 16024 | 1227 | 13.06 | 5.823e-38 |
| **sex** | 9508 | 697.6 | 13.63 | 4.454e-41 |
| **age** | -67.6 | 20.89 | -3.236 | 0.001225 |
| **educ\_m** | 2256 | 787.8 | 2.863 | 0.004224 |
| **educ\_ms** | 4589 | 860.4 | 5.333 | 1.036e-07 |
| **educ\_h** | 11796 | 968 | 12.19 | 2.266e-33 |
| **city** | -2803 | 602.7 | -4.651 | 3.446e-06 |
| **satisf** | 1326 | 759.9 | 1.745 | 0.08104 |

-602.7

1. Также Вовочка провел тест на значимость коэффициента при дамми на удовлетворённость. Он сделал вывод, что коэффициент значим на 10%. Но Вовочка затруднятся с интерпретацией результата. Какова она?

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16023.796 1226.667 13.0629 < 2.2e-16 \*\*\*  
## sex 9507.860 697.591 13.6296 < 2.2e-16 \*\*\*  
## age -67.601 20.889 -3.2361 0.001225 \*\*   
## educ\_m 2255.566 787.792 2.8631 0.004224 \*\*   
## educ\_ms 4588.679 860.370 5.3334 1.036e-07 \*\*\*  
## educ\_h 11796.249 968.000 12.1862 < 2.2e-16 \*\*\*  
## city -2803.148 602.689 -4.6511 3.446e-06 \*\*\*  
## satisf 1326.236 759.892 1.7453 0.081037 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* Те, у кого удовлетворённость выше, получают заработную плату на 1326.24 рублей выше.