

RLMS-HSE: ОТ ЧЕГО ЗАВИСИТ ЗАРАБОТНАЯ ПЛАТА?

ВВЕДЕНИЕ

Российский мониторинг экономического положения и здоровья населения НИУ ВШЭ (RLMS-HSE) (hse.ru/rlms) представляет собой серию ежегодных общенациональных репрезентативных опросов населения, содержащий в себе комплексную информацию об индивидах и домохозяйствах: структуре доходов и расходов, материальном благосостоянии, инвестиционном поведении, структуре занятости, миграционном поведении, состоянии здоровья и структуре питания, о планировании семьи и образовательном поведении, о системе ценностей россиян, восприятии ими проводимых в стране преобразований и т.д. Изначально идея создания такого мониторинга состояла в том, чтобы получить достоверную картину последствий проводимых реформ для населения страны. Однако набор вопросов настолько широк, что позволяет тестировать различные экономические теории, анализировать социальные, демографические и другие взаимосвязи.

ОПИСАНИЕ ДАННЫХ

В представленном задании используются результаты опроса населения РМЭЗ НИУ ВШЭ в 2017 году (hse.ru/rlms/spss). Датасет представляет собой выборку из индивидов, находящихся в трудоспособном возрасте. В датасете представлена информация об их заработной плате, образовании, семейном положении и другие характеристики.

Одна строка в вашем наборе данных (единица наблюдения) — это один индивид, принимавший участие в опросе в 2017 г. Число наблюдений — 3871. Информация о нем содержится в следующих переменных (файл Excel Data_RLMS.xls):

- **Inwage** — логарифм заработной платы, полученной за последние 30 дней по основному месту работы;

- **public** — 1 если государственное предприятие (в государственной собственности), 0 если частное;
- **educ** — уровень образования, категориальная переменная (0 для индивидов без среднего образования)
 1. среднее образование
 2. ПТУ, техническое училище
 3. институт, университет, академия
- **urban** — 1 респондент проживает в городе, 0 в сельской местности;
- **male** — пол, 1 для мужчин, 0 для женщин;
- **age** — возраст в годах;
- **children** — количество детей;
- **industry** — отрасль занятости;
 1. ЛЕГКАЯ, ПИЩЕВАЯ ПРОМЫШЛЕННОСТЬ
 2. ГРАЖДАНСКОЕ МАШИНОСТРОЕНИЕ
 3. ВОЕННО-ПРОМЫШЛЕННЫЙ КОМПЛЕКС
 4. НЕФТЕГАЗОВАЯ ПРОМЫШЛЕННОСТЬ
 5. ДРУГАЯ ОТРАСЛЬ ТЯЖЕЛОЙ ПРОМЫШЛЕННОСТИ
 6. СТРОИТЕЛЬСТВО
 7. ТРАНСПОРТ, СВЯЗЬ
 8. СЕЛЬСКОЕ ХОЗЯЙСТВО
 9. ОРГАНЫ УПРАВЛЕНИЯ
 10. ОБРАЗОВАНИЕ
 11. НАУКА, КУЛЬТУРА

12. ЗДРАВООХРАНЕНИЕ
13. АРМИЯ, МВД, ОРГАНЫ БЕЗОПАСНОСТИ
14. ТОРГОВЛЯ, БЫТОВОЕ ОБСЛУЖИВАНИЕ
15. ФИНАНСЫ
16. ЭНЕРГЕТИЧЕСКАЯ ПРОМЫШЛЕННОСТЬ
17. ЖИЛИЩНО-КОММУНАЛЬНОЕ ХОЗЯЙСТВО
18. ОПЕРАЦИИ С НЕДВИЖИМОСТЬЮ
19. СОЦИАЛЬНОЕ ОБСЛУЖИВАНИЕ
20. ЮРИСПРУДЕНЦИЯ
21. ЦЕРКОВЬ
22. ХИМИЧЕСКАЯ ПРОМЫШЛЕННОСТЬ
23. ДЕРЕВООБРАБАТЫВАЮЩАЯ ПРОМЫШЛЕННОСТЬ, ЛЕ
24. СПОРТ, ТУРИЗМ, РАЗВЛЕЧЕНИЯ
25. УСЛУГИ НАСЕЛЕНИЮ
26. ИТ, ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
27. ЭКОЛОГИЯ, ЗАЩИТА ОКРУЖАЮЩЕЙ СРЕДЫ
28. ОРГАНИЗАЦИЯ ОБЩЕСТВЕННОГО ПИТАНИЯ
29. СМИ, ИЗДАТЕЛЬСТВО, ПЕЧАТЬ, ТЕЛЕКОММУНИК
30. РЕКЛАМА, МАРКЕТИНГ
31. ОБЩЕСТВЕННЫЕ ОРГАНИЗАЦИИ, СОВЕТ ВЕТЕРАН

➤ **internet** — приходилось ли респонденту в течение последних 12 месяцев пользоваться Интернетом, 1 если да, 0 в противном случае;

➤ **psu** — номер региона.

Далее следуют дамми переменные, идентифицирующие принадлежность к регионам и отраслям.

Регионы:

1. Санкт-Петербург (id1)
2. Москва (id2)
3. Московская область (id3)
4. Республика Коми, Сыктывкар (id4)
5. Республика Коми, Усинск и Усинский район (id5)
6. Ленинградская область, Волосовский район (id6)
7. Смоленск (id7)
8. Тверская область, Ржев и Ржевский район (id8)
9. Тула (id9)
10. Калужская область, Куйбышевский район (id10)
11. Нижний Новгород (id11)
12. Республика Чувашия, Шумерля и Шумерлинский район (id12)
13. Пензенская область, Земетчинский район (id13)
14. Липецк (id14)
15. Тамбовская область, Уварово и Уваровский район (id15)
16. Казань (id16)
17. Саратов (id17)
18. Саратовская область, Вольск и Вольский район (id18)
19. Волгоградская область, Руднянский район (id19)
20. Кабардино-Балкария (id20)

21. Ростовская область, Батайск (id21)
22. Краснодар (id22)
23. Ставропольский край, Георгиевск и Георгиевский район (id23)
24. Краснодарский край, Куцевский район (id24)
25. Челябинск (id25)
26. Курган (id26)
27. Удмуртская Республика, Глазов и Глазовский район (id27)
28. Оренбургская область, Орск (id28)
29. Пермская область, Соликамск и Соликамский район (id29)
30. Челябинская область, Октябрьский район (id30)
31. Томск (id31)
32. Новосибирская область, Бердский район (id32)
33. Алтайский край, Бийск и Бийский район (id33)
34. Алтайский край, Курьинский район (id34)
35. Красноярск (id35)
36. Владивосток (id36)
37. Красноярский край, Назарово и Назаровский район (id37)
38. Амурская область, Тамбовский район (id38)

Укрупненные регионы:

39. Москва и Санкт-Петербург (id39)
40. Москва и Московская область (id40)
41. Санкт-Петербург и Ленинградская область (id41)

42. Республика Коми (id42)
43. Саратов и Саратовская область (id43)
44. Краснодар и Краснодарский край (id44)
45. Алтайский край (id45)
46. Красноярск и Красноярский край (id46)

Информация об отраслях занятости также дублируется в дамми-переменных:

47. легкая и пищевая промышленность (id47)
48. гражданское машиностроение (id48)
49. ВПК (id49)
50. нефтегазовая промышленность (id50)
51. другие отрасли добывающей промышленности (id51)
52. строительство (id52)
53. транспорт, связь (id53)
54. сельское хозяйство (id54)
55. органы управления (id55)
56. образование (id56)
57. наука, культура (id57)
58. здравоохранение (id58)
59. армия, МИД, служба безопасности (id59)
60. торговля (id60)
61. финансы (id61)
62. энергетическая промышленность (id62)

63. ЖКХ (id63)

Укрупненные отрасли:

64. Промышленность (id64)

65. Образование и наука (id65)

66. Образование и здравоохранение (id66)

67. Торговля и финансы (id67)

68. ВПК и армия (id68)

69. Строительство и транспорт (id69)

70. Органы управления и образование (id70)

71. Нефтегазовая и другие добывающие (id71)

72. Легкая, пищевая и гражданская промышленность (id72)

73. Нефтегазовая, другие добывающие и энергетическая (id73)

74. Образование, наука и здравоохранение (id74)

75. Сельское хозяйство и торговля (id75)

ЗАДАНИЕ

Используя имеющиеся данные, вам необходимо ответить на вопрос:

От чего зависит заработная плата?

Вы можете конкретизировать поставленный вопрос. Например, вот так: Как влияет уровень образования на заработную плату? Как влияет сфера занятости? Какие другие факторы влияют или наоборот не оказывают влияния на заработную плату? Есть ли разница между проживающими в городской и сельской местности? Также можно описать различия для разных возрастов, регионов, отраслей и др., которые считаете нужными.

Рекомендации по проведению исследования

Ниже приведен примерный план вашего исследования, на который вы можете опираться, и подробные комментарии к пунктам плана. Не обязательно выполнять все пункты и соблюдать приведенный порядок, но работа будет оцениваться в соответствии с критериями, с которыми вам обязательно стоит ознакомиться. Постарайтесь выполнить пункты этого плана так, чтобы набрать как можно больше баллов в сумме по критериям.

1. **Предварительный анализ.** Опишите предоставленные вам данные: сколько в них наблюдений, как они распределены по разным подгруппам, какие в них есть особенности (наличие выбросов, асимметричность распределений, пропуски, неравномерность по подгруппам и тому подобное). С чем могут быть связаны эти особенности? Какие зависимости/паттерны вы наблюдаете?
2. **Постановка гипотезы.** Выявленные зависимости помогут навести вас на предположение, которое можно проверить в рамках исследования на предоставленных данных. Подумайте о том, что вам интересно исследовать и почему это может быть интересно кому-то кроме вас. Подумайте, какие экономические теории или модели могут служить основой вашего исследования.

Например, подумайте: От каких факторов зависит заработная плата? Правда ли что, если дольше учиться, то заработная плата будет больше? В каких отраслях вероятно более высокая заработная плата? Есть ли гендерное неравенство в заработных платах?

Обратите внимание, что ответы на вопросы могут быть разными для разных групп индивидов. Попробуйте исследовать отдельно группы, которые можете выделить. Посмотрите на данные в разбивке на отрасли/регионы/места проживания: есть ли между ними различия? Можете ли порекомендовать что-то, что поможет увеличить заработную плату?

В результате вами должно быть высказано предположение, которое будет проверяться в рамках исследования, — это ваша гипотеза. Объясните, почему ваша гипотеза может выполняться. Какие причинно-следственные взаимосвязи стоят за гипотезой (механизм гипотезы)?

Помните, что проверка вашей гипотезы должна быть возможна на предоставленных данных. Поэтому учитывайте предоставленный набор переменных при составлении гипотезы.

- 3. Выбор переменных.** Выберите из представленного набора данных факторы, влияние которых вы считаете важным, и переменные, влияние которых рассматриваете. Сформулируйте гипотезу о том, положительно или отрицательно факторы будут связаны с заработной платой. Расскажите подробно, почему ожидаете именно такую зависимость. Обратите внимание на единицы измерения выбранных переменных при исследовании их совместного распределения.
- 4. Разведывательный анализ.** Проиллюстрируйте с помощью графиков распределение выбранных переменных. Посчитайте основные описательные статистики для них. Какие особенности вы наблюдаете? Как их можно интерпретировать, то есть объяснить? Охарактеризуйте форму распределений (скошенность, симметричность/ асимметричность, вариативность).

В ходе разведывательного анализа учитывайте наличие выбросов и пропусков в выбранных вами переменных. Стоит обратить внимание на неоднородность данных: в разных категориях распределения одной и той же переменной могут различаться. Не забывайте, что анализ можно проводить не только по исходным переменным, но и по преобразованным данным. Например, можете агрегировать участников по возрасту, полу и другим характеристикам.

- 5. Проверка гипотезы на данных.** Используйте математические методы и статистику, чтобы проверить выполнение гипотезы. Начните с визуального анализа и изобразите взаимосвязь на графике или нескольких графиках, если это необходимо. Учитывайте, что для разных типов переменных и взаимосвязей оптимальные визуализации различаются, и выбирайте те, что наиболее четко демонстрируют наличие или отсутствие связи.

Проверьте, устойчивы ли ваши выводы о выполнении или невыполнении гипотезы. Посмотрите то же самое по группам — стабильно ли подтверждается результат? Если результаты получаются разными на разных подвыборках данных, предложите интерпретацию, почему так может быть.

- 6. Выводы и ограничения.** Расскажите о том, как можно применить ваши результаты. Какие вы можете предложить действия для увеличения заработной платы и для кого? Какие получили важные знания?

Какие полученные вами выводы могут быть полезны для социальной политики? Если можете, предложите конкретные меры, основанные на полученных вами выводах.

Есть ли какие-то факторы, которые вы не смогли учесть в исследовании? Могут ли быть альтернативные объяснения выполнения взаимосвязи кроме того, что вы предложили? Подумайте, какие есть ограничения в применении результатов и откуда они следуют: каких именно данных вам не хватило, чтобы быть более уверенными в результате, и как именно вы могли бы их использовать.

Расскажите, для кого можно применить полученный результат? Для жителей каких стран /регионов/городов? Можно ли расширить на людей других стран?

Удачи в подготовке исследования! Помните, что идеальных исследований не бывает, используйте предоставленные данные по максимуму для получения нетривиальных выводов, но не забывайте про рамки, в которых эти выводы могут быть применимы.