Федеральное государственное автономное образовательное учреждение высшего образования
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Работу выполнила
Группа 27

Догот Игнат
Ломакин Артемий
Лукашенко Анжелика
Петрова Екатерина
Потихонов Андрей
Шалашов Андрей

Преподаватели:
Лозина П.С.
Пырлик В. Н.
Терещенко Д. С.

Проект по дисциплине:

«Эконометрика»

Тема:

«Количество пропущенных по болезни дней на работе в зависимости от уровня заработной платы»

Санкт-Петербург 2021

Содержание

1 Введение

1.1 Постановка исследовательского вопроса

Главный вопрос исследования — как уровень заработной платы человека связан с количеством рабочих дней, пропущенных по болезни? В работе будет изучаться зависимость между средним количеством пропущенных по болезни рабочих дней в месяц и средней заработной платой в месяц за последний год респондента при прочих равных условиях.

1.2 Актуальность исследования

На фоне принятия условий пандемии в повседневные трудовые взаимоотношения, работодатели вынуждены учитывать увеличившуюся вероятность пропусков по причине заболевания среди своих сотрудников. Стоит также учитывать, что некоторые работники совершают пропуски по болезни, хотя сами не болеют. При этом, некоторые работодатели готовы платить своим сотрудникам более высокую заработную плату, если они продолжат работать даже в случае болезни. Мы постараемся выявить эту закономерность и понять, действительно ли сотрудники, получающие более высокую заработную плату, менее склонны совершать пропуски.

Наша работа основана на данных до 2019 года, однако потенциально выявленная зависимость между уровнем заработной платы и пропусками работы, по нашей оценке, не зависит от времени и внешних условий, и, при наличии взаимосвязи, в условиях пандемии она только усилилась.

2 Обзор литературы и выдвижение гипотезы

2.1 Идея нашей работы

Теория нашей работы опирается на некоторые данные и факты из статьи “Risk factors for sick leave - general studies” Allebeck и Mastekaasa (2004) авторства Peter Allebeck и Arne Mastekaasa, в которой подчеркивается, что исследования и регулярно публикуемая статистика показывают связь между социально-экономическим статусом и отсутствием болезней, однако в данной работе не было выявлено достаточное количество причинно-следственных механизмов, которые объясняли бы экономическое положение человека и частоту его заболеваний. Поэтому, чтобы изучить тему подробнее, мы выдвинули гипотезу о том, что при росте среднего дохода, количество пропущенных по болезни дней будет уменьшаться в связи с улучшением качества жизни.

2.2 Первичная формулировка модели и гипотез

Модель, которую мы хотели бы оценить, является мультипликативной - при очень большой заработной плате люди болеют и пропускают намного меньше чем при маленьких. Взятие логарифмов позволяет оценивать модель с помощью линейной регрессии. Рассматривается насколько процентов меняется среднее число пропусков в месяц при изменении заработной платы на 1 процент, вместо дней и рублей, которые мало показательны.

Возможные причины данного эффекта: Люди с маленькими доходами склонны больше болеть и меньше заботятся о себе Предположительно, чем больше доход тем больше вероятность заниматься превентивной медициной до того, как болезнь начинает становится серьезной Предположительно, люди с маленькими доходами имеют больше детей Высокий уровень ответственности у людей с большими доходами, который имеет положительный эффект на посещение работы

Исходя из исследовательского вопроса главная нулевая гипотеза, которую мы попытаемся отвергнуть, формулируется следующим образом: при прочих равных среднее количество пропущенных по болезни рабочих дней в месяц не зависит от средней заработной платы в месяц респондента.

3 Данные

Данные для исследования были взяты из базы «Российский мониторинг экономического положения и здоровья населения НИУ ВШЭ (RLMS HSE)», проводимый Национальным исследовательским университетом "Высшая школа экономики" и ООО «Демоскоп» при участии Центра народонаселения Университета Северной Каролины в Чапел Хилле и Института социологии Федерального научно-исследовательского социологического центра РАН. (Российского мониторинга экономического положения и здоровья населения). Этот мониторинг является серией ежегодных общенациональных репрезентативных опросов, в которых собрана информация о домохозяйствах

и индивидах, их доходе, профессии, состоянии здоровья и т.п. Обширность и вариативность данных в базе позволяют произвести наиболее точный, полный и глубокий анализ информации.

3.1 Выборка

Ввиду того, что в исследовании не имеет большого значения репрезентативность населения России, данные взяты по полной выборке: наблюдения по всем индивидам-жителям России. Были рассмотрены волны с 2016 по 2019, после фильтрования на нужные нам переменные и непустые значения во всех из них, наибольшее количество наблюдений осталось в выборке 2019 года, 28-й волны, которую мы и взяли в рассмотрение.

3.2 Работа с данными

Ссылка на код. При обработке исходного датасета были проделаны следующие действия:

Были отобраны только необходимые для проведения исследования столбцы, описанные ниже.

Удалены пропущенные значения в будущей зависимой переменной xl90 (количество пропущенных по болезни дней) и главных регрессорах xj60 (доход за последний месяц) и xj13.2 (средняя месячная зарплата за год)

Столбцы датасета были приведены к формату numeric, так как изначально значения в них были других, неподходящих для дальнейшей обработки форматов, а в некоторых переменных записывались в виде, отличном от целых чисел (например, «1e+00»)

В столбце xj72.173 (количество совершеннолетних детей в семье) у людей, у которых нет детей, стоят пропущенные значения, поэтому мы заменили их на нули

В столбце xm20.614 (наличие хронических гинекологических заболеваний) у мужчин стоят пропущенные значения; мы решили заменить их на 2, поскольку у мужчин нет гинекологических заболеваний

Отфильтровали все столбцы, убрав все значения большие 9999990, так как они означают, что респондент не ответил на вопрос Переменную xl90 (количество пропущенных по болезни дней за год) мы поделили на 12, так как нас интересует среднее количество пропущенных дней по болезни за месяц, а не за год, и создали sickdays

В столбцах, связанных с хроническими заболеваниями, мы заменили все двойки (изначально обозначающие отсутствие заболеваний) на нули, показывая, что у человека нет хронических заболеваний, и создали новый столбец chronical, в котором посчитали все хронические заболевания, удалив все столбцы, использованные для этого

Столбцы, содержащие факторные данные в другом формате, мы привели к типу factor

Нулевые значения из столбцов, описывающих заработок людей, были убраны, так как они не значимы для анализа. Аналогично, из столбца, связанного с количеством рабочих часов в неделе, были удалены значения больше ста, чтобы оставить только реалистичные наблюдения

3.3 Описание данных

Исходный датасет изначально содержит 18061 наблюдение; при выборе только необходимых переменных это число не уменьшается. После фильтрации, при которой зависимая переменная и главный регрессор не пустые, наблюдений становится 1578. Также, в некоторых наблюдениях указывался вариант “99999999/8/7” при пропущенном ответе. Путем избавления от подобных некачественных данных для корректности модели мы получили 1324 наблюдения. И конечным шагом стало удаление нереалистичных наблюдений в переменных о часах работы и нулевых значений в доходах. Количество наблюдений сократилось с 18061 до 1307, что мы считаем достаточным для построения модели и выводов.

3.4 Разведывательный анализ данных

В этой части анализируется каждая переменная на выбросы, минимальные, средние и максимальные значения, достаточность наблюдений в каждой категории в факторных переменных. А также подробно описывается, как теоретически каждая переменная влияет на зависимую переменную.(Таблица 5)

Зависимая переменная: В качестве зависимой переменной мы выбрали среднее месячное значение количества дней пропущенных по болезни за последние 12 месяцев (sickdays). Респонденты пропускали в среднем месяц от 0,08 до 23 дней. Это несколько выбросов, где люди пропускали по болезни до 275 дней в год, что мы считаем наблюдениями, у людей, которые попадали сами или с детьми ложились за последний год в больницу на

долгое лечение (например, несколько можно видеть на графике ниже отмеченных красным). В среднем значение составляет 1,34 дня в месяц.

Для главного регрессора есть две переменные — “mincome - Сколько всего денег в течение последних 30 дней Вы лично получили, считая все: зарплату, пенсии, премии, прибыли, пособия, материальную помощь, случайные заработки и другие денежные поступления?”, которая учитывает все денежные поступления, но только за последний месяц при котором проводился опрос, и “awage - За последние 12 месяцев какова была Ваша среднемесячная зарплата на этом предприятии после вычета налогов - независимо от того, платят Вам ее вовремя или нет?”. Мы предполагаем, что чем больше уровень дохода, тем меньше дней будет пропускать индивид.

Переменная awage распределена от 3000 до 420000 рублей месяц, среднее значение — 30641 руб. Она будет более показательна ввиду того, что в ней усредненное значение за год (так же как и в зависимой переменной) и скорее всего люди точнее знали сумму их заработной платы в отличии от полного дохода, но эта переменная не учитывает дополнительные денежные поступления и поэтому не может полностью показывать благосостояние индивида. Значение без резких скачков достигает 420000. (Рис. 1)

Рис. 1:

Регрессор mincome принимает значения от 1800 до 2550000 рублей в месяц, в среднем люди получили 40875 руб. Его можно взять за обычный средний ежемесячный доход индивида за все 12 месяцев и принять за главный регрессор, но его значение не будет настолько показательным, потому что берется определенный месяц года, поэтому было принято решение использовать среднюю заработную плату. В данных о доходе также есть несколько выбросов, которые могут помешать оценке. (Рис. 9)

Контрольные переменные:

Социо-демографические переменные (Таблица 3):

status: тип населенного пункта - на случай если жизнь в деревне или в городе связана со здоровьем, заработком, культурой идти на работы даже если человек заболел или чувствует недомогание. Часто люди, которые проживают в мегаполисах подвержены шуму и стрессу, что может негативно влиять на общий уровень здоровья. Однако, есть и противоположное мнение, что, жизнь в деревне и регулярно подвергаться большим физическим нагрузкам (уход за огородом и тд), также ведет за собой некоторые осложнения. Количество человек в каждой группе достаточное для анализа. (Рис 11)

age: количество полных лет - люди старшего или пожилого возраста могут болеть и пропускать больше рабочих дней или в определенном возрасте могут быть дети, за которыми нужен уход по болезни, а молодые и амбициозные люди могут выходить на работу при некоторых симптомах болезни. Количество полных лет влияет на уровень здоровья человека и, по нашему мнению, имеет зависимость с количеством пропущенных дней по работе в связи с болезнью. Также вполне вероятно, что с наступлением определенного возраста человеку необходимо будет брать больничный по уходу за ребенком, из-за чего эта зависимость не может быть прямой. Больших выбросов нет, значения от 18 до 81. (Рис. 12.)

gender: пол респондента (1-мужской, 0-женский) - есть предположение, что женщины чаще берут больничный

если их ребенок заболел, поэтому обязательно необходимо иметь такую переменную в уравнении. Количество в каждой группе достаточное. (Рис. 13)

childcount: А сколько из них (детей) моложе 18 лет? и children: у Вас есть дети, родные или официально усыновленные? - люди с детьми берут больничный чаще, чем те, у которых их нет. Особенно берут, если за детьми нужен уход, поэтому эта переменная необходима к рассмотрению. Респондентов с детьми младше 18-ти выборке 715, один ребенок у 330 человек, а двое у 218-ти. Есть респонденты имеющие до 6 детей. (Рис. 2)

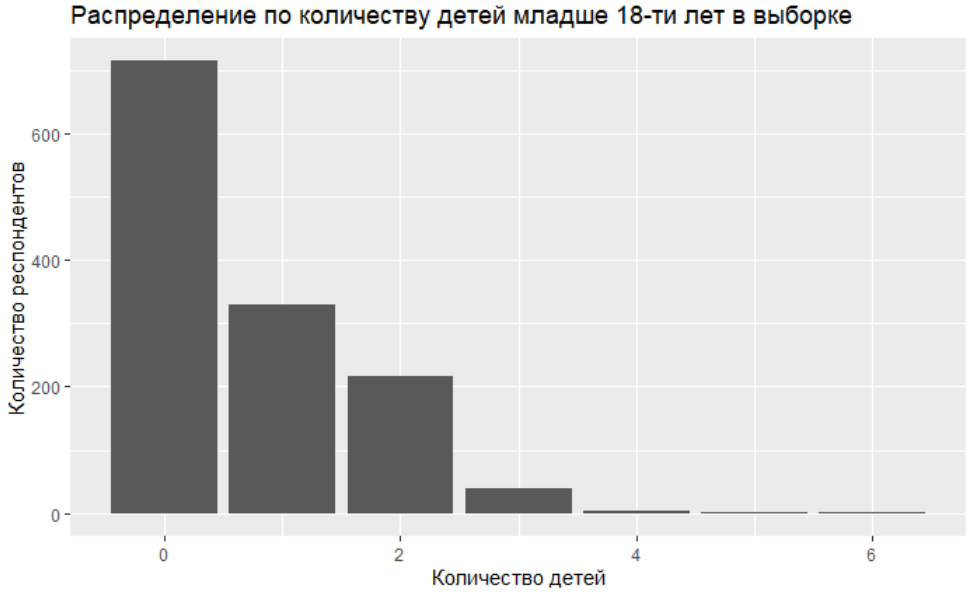

Рис. 2:

marst: Семейное положение - в литературе описано, что значительно чаще пропуски работы встречаются среди разведенных/разлученных и вдов/вдовцов, чем среди женатых людей; наблюдается высокий уровень пропусков по болезни среди разведенных мужчин, но не среди разведенных женщин или вдов; люди, которые развелись в предшествующем году, пропускали работу значительно чаще, чем другие; те, кто развелся в рассматриваемом году, подвергаются более высокому риску отсутствия по болезни в течение того же года. Больше всего людей в выборке состоящих в браке (783 человека), а людей не состоящих в браке, живущих с партнером, но не зарегистрированных и разведенных примерно одинаковое количество от 129 до 175 человек. (Таблица 3)

Связанные с условиями труда (Таблица 4): плохие, нагруженные, опасные для здоровья условия труда; разные более стрессовые сферы деятельности; большая ответственность и боязнь потерять работу могут влиять на здоровье и на то, будет ли человек брать больничные на работе.

harmwork: является ли производство, на котором Вы работаете, вредным или опасным, т.е. дающим Вам право на досрочное назначение трудовой пенсии, на дополнительные выплаты или льготы? Данная переменная показывает, насколько организм работающего подвержен негативному влиянию на работе и как следствие может вызвать ухудшение состояние организма и более продолжительные больничные. Достаточно человек ответили “да” (212), чтобы оставить эту переменную.(Таблица 4)

branch: в какой отрасли Вы работаете на этой работе? - Эта переменная близка к harmwork, однако, мы можем предположить, что в одинаковых отраслях, но на разных предприятиях или даже профессиях, вред, приносимый сотрудникам, будет существенно отличаться. Нет сильно превалирующих отраслей, но есть множество маленьких, в итоге нам кажется, что переменная будет показательна в регрессии, но может иметь эффект. (Рис. 14)

hours: сколько часов в среднем продолжается Ваша обычная рабочая неделя? - При неправильном подходе к циклу работа-отдых, организм может существенно перенапрягаться, что может повлечь за собой серьезные проблемы с организмом, этой переменной мы учитываем психологическое состояние индивида. Мы убрали все данные, где указано более 100 часов в неделю. В среднем работают 42 часа в неделю. (Таблица 4)

ofwork: Вы оформлены на этой работе официально, то есть по трудовой книжке, трудовому соглашению, контракту? Мы считаем, что если человек не будет оформлен по контракту, то он будет опасаться брать больничные, боясь потерять работу. Доля тех, кто работает по договору намного выше, поэтому эту переменную мы не будем учитывать в построении регрессии. (Рис. 3)

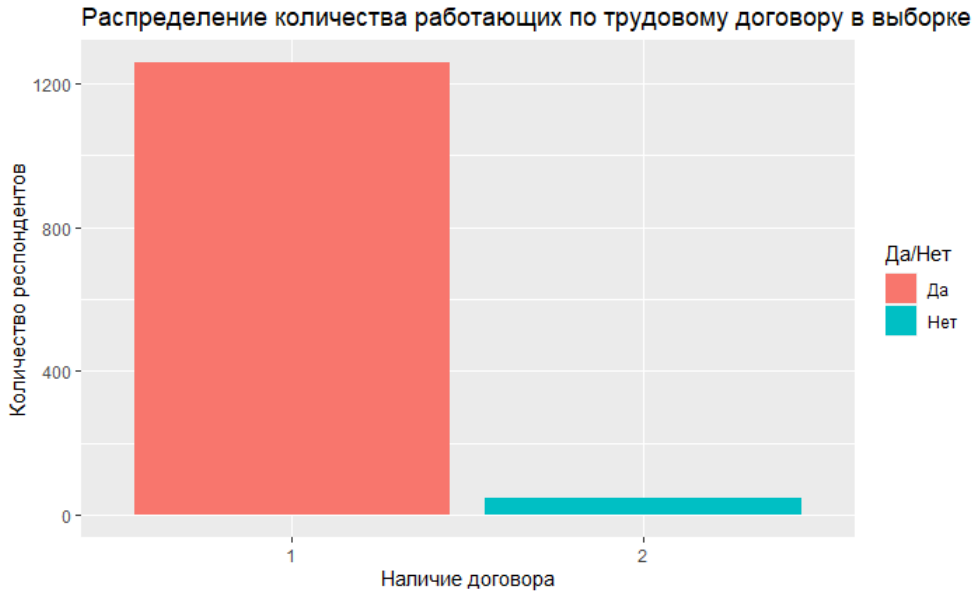

Рис. 3:

entrepreneurship: как Вы считаете, на этой работе Вы занимаетесь предпринимательской деятельностью? - Эту переменную мы оставили для анализа, так как часто предприниматели не могут уйти в отпуск, потому то они являются ключевыми персонами в бизнесе, тем более, если бизнес маленький. БОльшая часть респондентов не занимается предпринимательством, поэтому эту переменную мы тоже пропустим в регрессии. (Рис. 4)

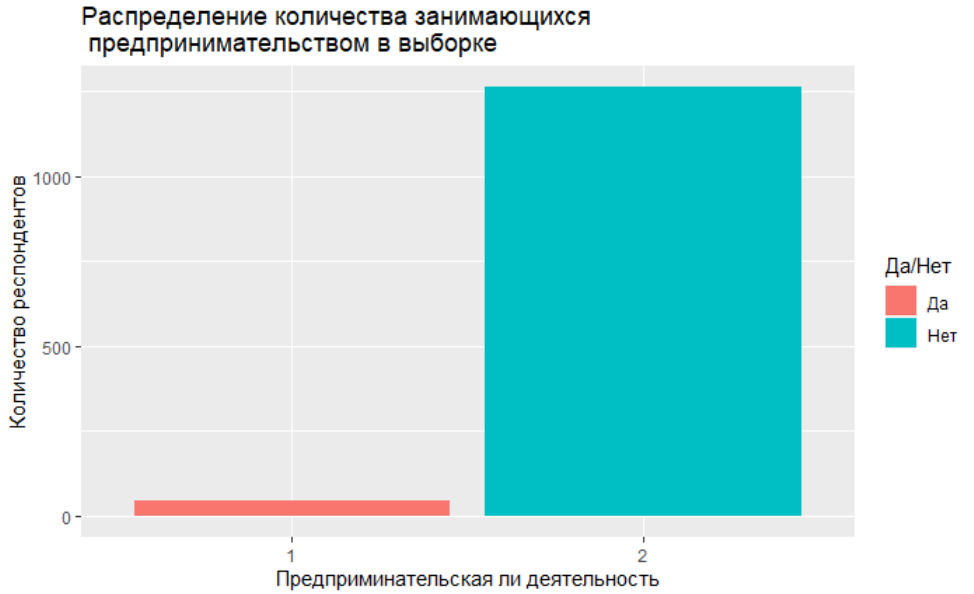

Рис. 4:

occup: Профессиональная группа - эта переменная помогает выявить уровень ответственности у людей. Предположительно, люди руководящих должностей не имеют возможность уйти на больничный и, например, при болезни детей, находят пути выйти на работу. Больше всего в этой переменной работников из офисных работников и сферы услуг - 291 и 288 человек. 185 квалифицированных работников сельского хозяйства. (Рис. 15)

Связанные со здоровьем (Таблица 4): Более здоровые люди с меньшим количеством хронических заболеваний, не имеющие инвалидность или не прошедшие через операции будут брать меньше больничных.

chronical: количество хронических заболеваний - нами создан новый столбец в котором указано сколько хро-

нических заболеваний из ниже представленных имеет индивид. Количество хронических заболеваний непосредственно влияет на количество дней, проведенных на больничном, потому что каждое хроническое заболевание может приходить в активную фазу в любое время (чаще всего в межсезонье) и влечет вероятность ухода на больничный. Всего максимум у человека 13 заболеваний, а в среднем 1.74.(Таблица 5)

disability: назначена ли Вам какая-нибудь группа по инвалидности? - Так как инвалидность вероятно влечет за собой какие-либо ограничения по работе и ослабленный организм, то эту переменную мы рассматриваем для анализа. Большая часть респондентов не имеют инвалидности, так что переменная не даёт правильной оценки в регрессии. (Рис. 5)

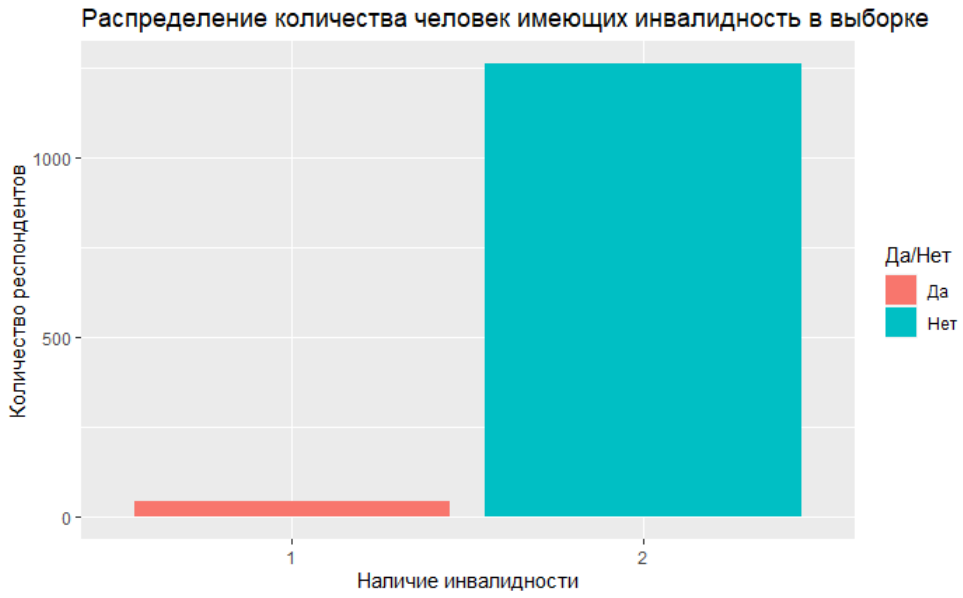

Рис. 5:

operations: в течение последних 12 месяцев Вам делали хирургические операции? - Хирургические вмешательства влекут за собой восстановительный период, то есть человек будет находиться на больничном, что отразиться на нашей модели. Тех, кто делал хирургические операции тоже мало, но мы всё ещё предполагаем, что много людей, которые пропустили большую часть года имели такие операции, хоть мы убрали переменную из уравнения регрессии, нам кажется она понадобится для анализа в дальнейшем.(Таблица 4)

3.5 Проверка регрессоров на эндогенность и мультиколлинеарность

Эндогенность. Анализ регрессоров на одновременность определения их значения с зависимой переменной (как по одной из причин эндогенности). Очевидно, что количество пропущенных дней из-за болезни не влияет на возраст, тип населенного пункта индивида, пол, на то является ли его его работа вредной или опасной для здоровья, количество детей и хронических болезней - следующие экзогенные переменные: age, status, gender, harmwork, branch, childcount, chronical. (Рис. 6)

Также в нашей модели имеются еще три переменные (mincome, awage, hours), которые в отличие от вышеописанных можно отнести к эндогенным. На интуитивном уровне, количество пропущенных дней по болезни сказывается на экономическом положении индивида. А также очевидно, что чем больше человек пропускает, тем меньше он работает в неделю, поэтому эти переменные тоже взаимовлияют друг на друга.

Мультиколлинеарность. Корреляционный анализ показал, что все переменные имеют слабую связь, кроме переменной age и childcount, а также age и chronical с со значениями -0,339 и -0,39, что по модулю не так близко к 1. Существует риск мультиколлинеарности, из-за связи возраста и количества детей несовершеннолетнего возраста, потому что в определенный возраст у людей больше всего детей меньше 18ти лет, поэтому есть зависимость. А также с возрастом количество хронических заболеваний тоже увеличивается, так что они связаны. Но мультиколлинеарность слабая и, чтобы избежать проблемы пропущенных переменных, мы решили оставить переменную возраста, потому что она также влияет на настрой респондента как работника в среднем, его готовность и возможность работать при легком недомогании.

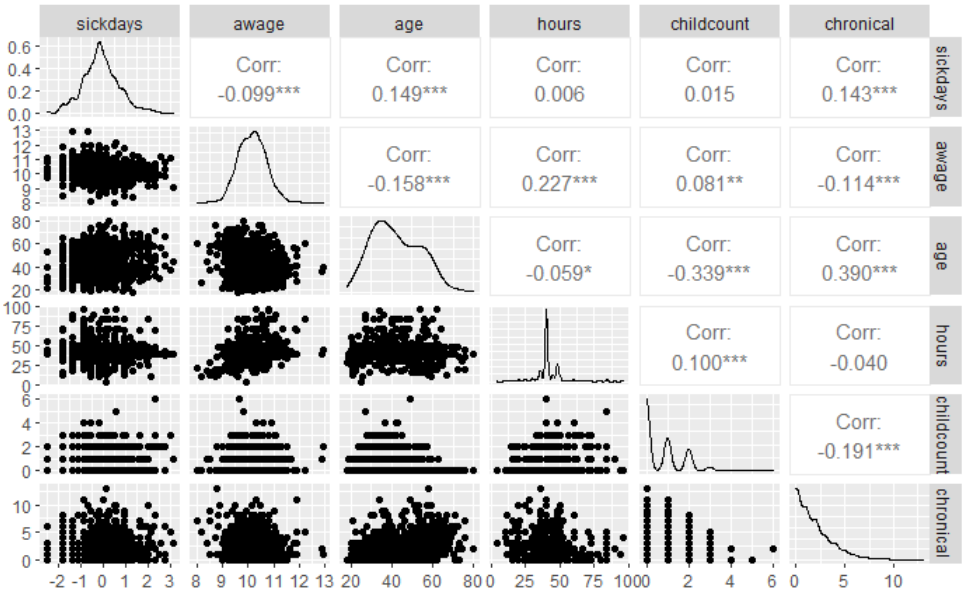

Рис. 6:

Также мультиколлинеарность могут вызывать переменные harmwok и branch. Такое предположение возникает из-за того, что отрасли работы и классификация их по вредоносности очень зависимыми между собой. Чтобы убедиться в этом, была выдвинута нулевая гипотеза об их независимости и проведен Хи-квадрат тест, в результате которого p-value получилось меньше 2.2e-16. Поэтому отвергается нулевая гипотеза и делается вывод о том, что эти две переменные являются зависимыми. Было принято решение об удалении переменной branch регрессии.

4 Регрессионный анализ

Проанализировав регрессоры на их влияние на зависимую переменную был сделан вывод, что у числовых переменных (кроме переменной age) скорее линейный эффект. С увеличением количества рабочих часов, несовершеннолетних детей, хронических заболеваний индивид начинает пропускать больше рабочих дней. Мы решили, что возраст имеет нелинейный эффект в связи с периодами в жизни людей, когда их дети достаточно взрослые, чтобы позаботиться о себе, в то время как сами они ещё достаточно здоровы и мало болеют.

4.1 Описание регрессионного анализа

Конечный список переменных можно посмотреть в табл 6.

Нами было рассмотрено 3 немного отличающиеся модели предсказания количества пропущенных по болезни дней от различных параметров, а именно: от заработной платы, возраста, количества детей, количества хронических заболеваний, пола и т.д. В некоторых случаях мы провели дополнительные операции над регрессорами, чтобы улучшить качество предсказаний. Например, во всех трёх моделях мы не берём месячную зарплату в чистом виде, а перед этим ищем его натуральный логарифм. Тоже самое, во всех трёх случаях, мы делаем с количеством пропущенных дней.

Формула регрессии model0:

$\log(sickdays_i) = \beta_0 + \beta_1 \log(awage_i) + u_i, i \in [1, 1307]$ (График 7)

Формула регрессии model1:

$\log(sickdays_i) = \beta_0 + \beta_1 \log(awage_i) + \beta_2 age_i + \beta_3 hours_i + \beta_4 chronical_i + \beta_6 harmwork_i + \beta_7 gender_i + \beta_9 age_i^2 + u_i, i \in [1, 1307]$

Формула регрессии model2:

$\log(sickdays_i) = \beta_0 + \beta_1 \log(awage_i) + \beta_2 age_i + \beta_3 hours_i + \beta_4 childcount_i + \beta_5 chronical_i + \beta_6 harmwork_i + \beta_7 status_i + \beta_8 gender_i + \beta_9 marst_i + \beta_{10} occup_i + \beta_{11} agei^2 + u_i, i \in [1, 1307]$

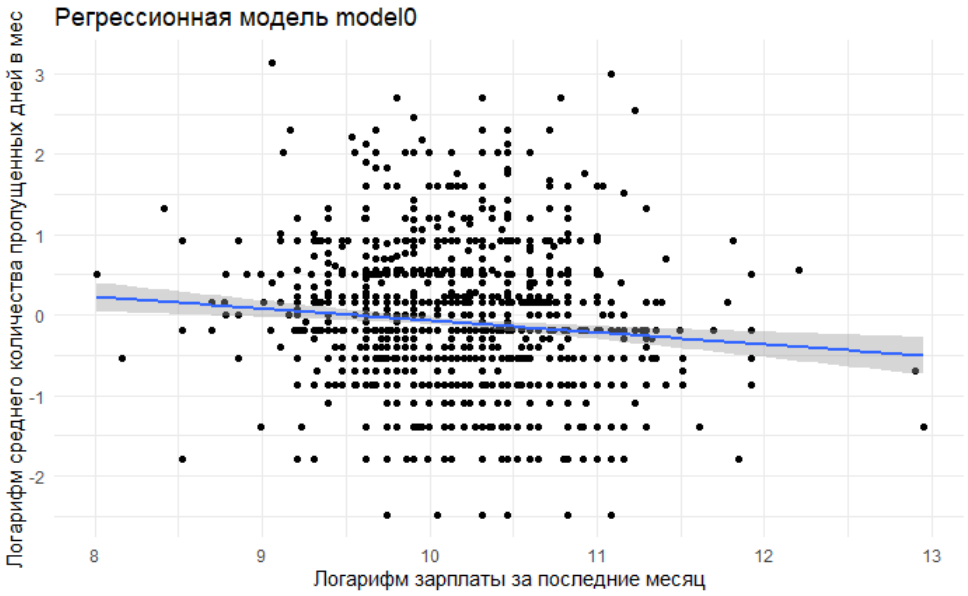

Рис. 7:

4.2 Ожидаемые результаты

Мы предполагаем, что с увеличением дохода работник будет пропускать по болезни всё меньше рабочих дней.

5 Результаты

Первая модель (model0) связывает напрямую прологарифмированное количество больничных дней с прологарифмированным месячной заработной платой. После проведённой линейной регрессии главный регрессор оказался значимым. Но хоть и прослеживается ожидаемая отрицательная корреляция между зарплатой и количеством больничных, присутствует очень слабое теоретическое обоснование валидности модели.

Остальные две модели (model1 и model2) отличаются тем, что в первую были добавлены контрольные переменные, которые в какой-то степени связаны со здоровьем человека, а во вторую все остальные, которые были анализированы ранее. Объясняющая сила каждой из них примерно одинаковы. В итоговом варианте model2 более теоретически оправдана при проверке гипотез. Все значимые коэффициенты при контрольных переменных совпадают с теоретическими предположениями: эффект заработной платы на среднее количество пропущенных рабочих дней отрицателен, а эффекты количества детей и хронических заболеваний положительны.

5.1 Интерпритация результатов

В среднем чем выше заработная плата человека, тем меньше вероятность того, что он будет пропускать работу по болезни. И с увеличением его зарплаты на 1%, уровень пропусков будет снижаться на 0.117%, что небольшая величина, которая с маленькой вероятностью будет сказываться на решении работодателей.

А также то, в каком типе населенного пункта живет человек каким то образом имеет влияние на его пропуски на работе. При увеличении хронических заболеваний на 1, количество пропусков по болезни увеличится на 4,2%, а при одном дополнительном ребенке на 6,8%.

5.2 Ответ на содержательный вопрос

Как среднее количество пропущенных по болезни рабочих дней в месяц зависит от средней заработной платы в месяц за последний год респондента при прочих равных условиях?:

Тест коэффициентов с бесконечным уровнем свободы показал, что среднее изменение количества пропущенных по болезни рабочих дней в месяц респондентом при изменении средней заработной платы в месяц за последний год на 1% равно -0.117%, что значит, что при прочих равных условиях зависит отрицательно.

Таблица 1: Результаты регрессии

	Зависимая переменная:		
	log(sickdays)		
	(1)	(2)	(3)
Логарифм заработной платы	-0.148^{***}	-0.148^{***}	-0.117^{**}
	(0.042)	(0.044)	(0.048)
Количество детей			0.068^{**}
			(0.034)
Количество хрон. заболеваний		0.045^{***}	0.042^{***}
		(0.014)	(0.014)
Работа вредна для здоровья		0.109^{*}	0.063
		(0.058)	(0.059)
Город			0.182^{***}
			(0.056)
ПГТ			0.223^{**}
			(0.110)
Село			0.165^{***}
			(0.063)
Наблюдения	1,305	1,305	1,305

Заметка: *p<0.1; **p<0.05; ***p<0.01

В скобках даны робастные стандартные ошибки вычисленные методом Уайта

Односторонний t-тест показал, что вероятность, что коэффициент при главном регрессоре равен 0, равна 0.0012, будучи меньше 1%, можно утверждать, что коэффициент точно имеет отрицательное влияние на зависимую переменную.

Были также получены ответы на следующие вопросы:

Есть ли разница в пропусках работы у жителей разных видов населенных пунктов?

При проверке нулевой гипотезы равенства всех коэффициентов при дамми-переменных, обозначающих тип населенного пункта, нулю была получена её вероятность равная 0.00233, при которой можно отвергнуть нулевую гипотезу.

Влияет ли количество хронических заболеваний и пол респондента на количество пропущенных дней одновременно?

При проверке нулевой гипотезы равенства коэффициентов при дамми-переменных, обозначающих пол и количество хронических заболеваний, нулю было получено p-value равное 0.03918, при котором можно отвергнуть нулевую гипотезу.

Количество хронических заболеваний и пол респондента также влияют на количество пропусков.

6 Критический анализ и возможные расширения исследования

6.1 Внутренняя валидность

Оценка причинно-следственных связей: может быть неточной ввиду слабой теоретической подкованности в темах здоровья и болезней и небольшого количества подходящей качественной литературы. Некорректная спецификация функциональной формы: возможна квадратичная нелинейная зависимость из-за, например, фактора мотивации людей с низкими доходами в продвижении по карьерной лестнице для увеличения доходов, тогда предельный эффект для некоторых уровней доходов может быть положительным. Неидеальность регрессора: ввиду недостаточности данных было принято решение взять за главный регрессор заработную плату, хотя данные о доходах были бы более показательны из-за эффекта пассивных доходов на здоровье и решения идти на работу. Эндогенность: возможная эндогенность главного регрессора и переменной hours из-за одновременной причинности с количеством пропущенных дней, проблема могла бы быть устранена с помощью инструментальных переменных. Возможные пропущенные переменные: Переменные про контроль присутствия на работе и факт физического присутствия в конкретном месте работы связаны с пропусками в течение рассматриваемого года, но для анализа отсутствия в течение следующих лет они имеют небольшое прогностическое значение Переменные, отражающие образ жизни, поведение, личные установки и привычки человека очень трудно или невозможно выделить Нет достаточных доказательств взаимосвязи между переменными о работе в психосоциальной рабочей среде и прогулах при отсутствии заболевания Поэтому можно сказать, что есть незначительный риск проблемы пропущенных переменных.

Тесты гипотез, тем не менее, имеют высокий уровень значимости. В данной работе использованы робастные стандартные ошибки для состоятельности при гетероскедастичности.

6.2 Внешняя валидность

Различия в популяциях: отсутствуют, поскольку используется выборка данных исключительно среди индивидов в рамках одного государства Различия в контексте данных (параметрах): считается, что обобщение результатов исследования при различных параметрах (в нашем случае, например, уровень должности) является некорректным. В доступной нам выборке не содержится информации о сфере деятельности и занимаемой должности индивидов, хоть мы и предполагаем, что, например, высшие государственные чины, директора и топ-менеджмент крупных компаний и прочие менее склонны (и имеют меньше возможностей) совершать прогулы, чем рядовые сотрудники предприятий. Наличие подобных данных могло бы уточнить и разнообразить наше исследование. Кроме того, данные содержали информацию о чувстве ответственности индивидов и уровню их доверия к другим людям. Мы оцениваем эти данные как субъективные, и сравнение их друг с другом может быть некорректным.

7 Заключение

В ходе проведенного исследования было выявлено, что основными факторами, влияющими на количество пропущенных дней по болезни, являются количество детей, хронических заболеваний и тип населенного пункта.

Несмотря на выводы из изученной литературы, в наших данных ни переменная о семейном положении индивида, ни информация о разводах не стала достаточно значимой.

Заработная плата в рассматриваемой выборке оказывает значимое, но незначительное влияние на количество пропускаемых дней. Поэтому работодатели должны будут решать вопрос об увеличении заработной платы для улучшения производительности отталкиваясь от того, сколько тот или иной работник приносит прибыли.

8 Оценка вкладов членов команды в групповую работу

Таблица 2: Оценка вкладов членов команды в групповую работу

Член команды	**Вклад**	**Оценка**
Петрова Екатерина	Коммуникация. Организация работы. Построение регрессионной модели. Редактура. Анализ данных. Изучение регрессионной модели. Анализ результатов. Работа над разведывательным анализом. Редактура.	10
Догот Игнат	Анализ данных. Изучение научной литературы. Форматирование отчета. Работа над разведывательным анализом.	10
Потихонов Андрей	Анализ данных. Форматирование отчета. Формирование теоретической базы работы. Анализ результатов. Работа над разведывательным анализом.	10
Лукашенко Анжелика	Коммуникация. Анализ данных. Форматирование отчета. Анализ ожидаемых результатов. Работа над разведывательным анализом. Написание заключения. Редактура	10
Ломакин Артемий	Анализ данных. Форматирование отчета. Построение регрессионной модели. Модификация датасета.	10
Шалашов Андрей	Анализ данных. Построение и анализ регрессионной модели. Модификация датасета.	10

Список литературы

[1] Peter Allebeck и Arne Mastekaasa. «Chapter 5. Risk factors for sick leave-general studies». В: *Scandinavian Journal of Public Health* 32.63_suppl (2004), с. 49—108.

«Российский мониторинг экономического положения и здоровья населения НИУ ВШЭ (RLMS HSE)», проводимый Национальным исследовательским университетом "Высшая школа экономики" и ООО «Демоскоп» при участии Центра народонаселения Университета Северной Каролины в Чапел Хилле и Института социологии Федерального научно-исследовательского социологического центра РАН.

9 Приложения

Рис. 8:

status	gender	children	marst
1: 637	0:776	1:1049	1:152
2: 388	1:529	2: 256	2:783
3: 57			3:175
4: 223			4:129
			5: 59
			6: 7

Таблица 3: Описательная статистика социально-демографических факторных переменных

harmwork	branch	ofwork	entrepreneurship	disability	operations	occup
0:1094	14 :242	1:1257	1: 44	1: 43	0:1194	3 :291
1: 211	10 :159	2: 48	2:1261	2:1262	1: 111	2 :288
	7 :132					5 :185
	12 :101					7 :166
	1 : 87					8 :156
	6 : 74					4 : 79
	(Other):510					(Other):140

Таблица 4: Описательная статистика факторных переменных относящихся к здоровью и работе

Рис. 9:

Рис. 10:

	variable	mean	median	sd	min	max
1	age	42.11	41.00	11.82	18.00	80.00
2	awage	30640.60	25000.00	23929.01	3000.00	420000.00
3	childcount	0.69	0.00	0.89	0.00	6.00
4	chronical	1.74	1.00	1.95	0.00	13.00
5	hours	42.01	40.00	9.81	4.00	96.00
6	mincome	40875.13	30000.00	94936.35	1800.00	2550000.00
7	sickdays	1.35	0.83	1.74	0.08	22.92

Таблица 5: Описательная статистика численных переменных

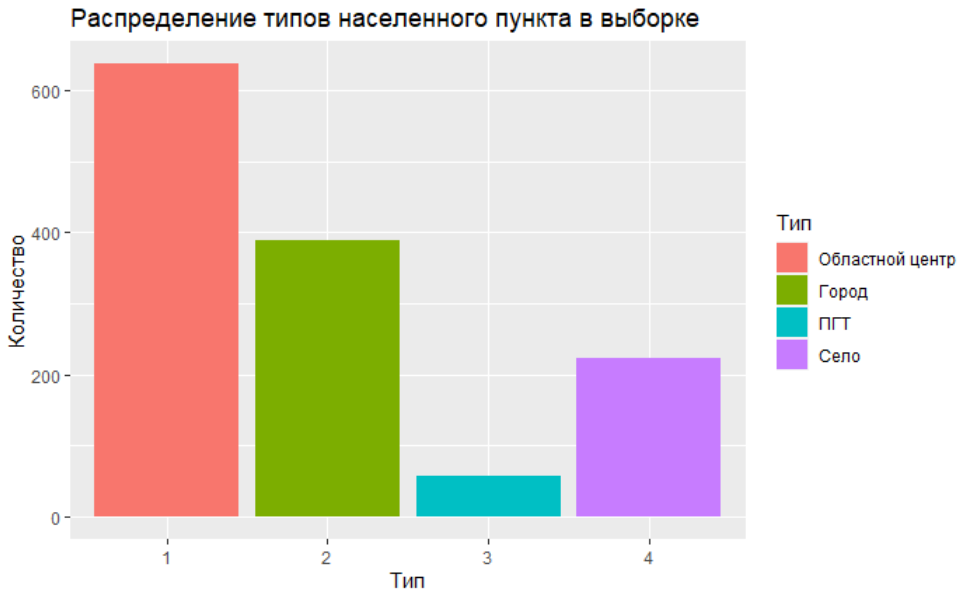
Распределение типов населенного пункта в выборке
Количество
600
400
200
0
1
2
3
4
Тип
Тип
Областной центр
Город
ПГТ
Село

Рис. 11:

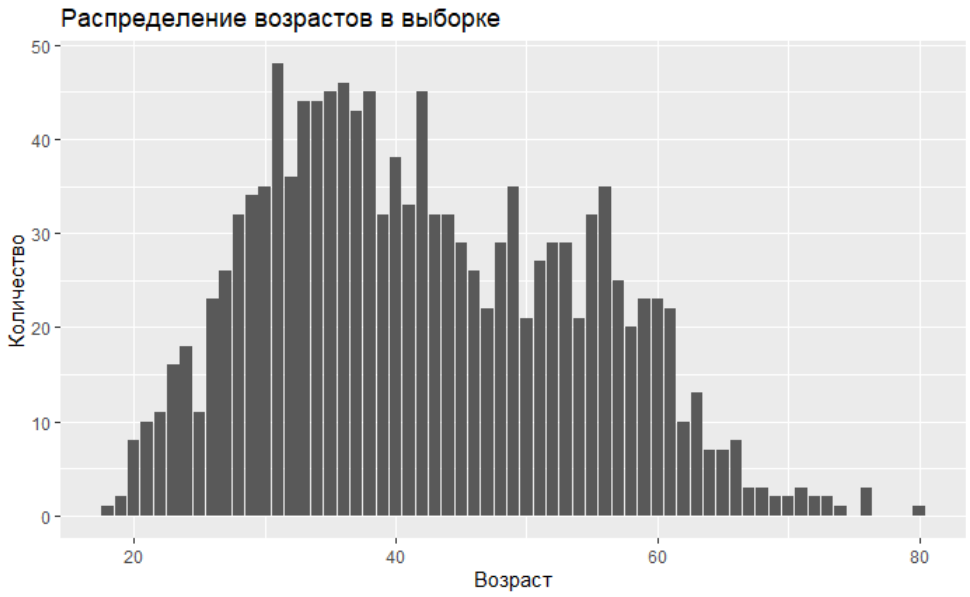
Распределение возрастов в выборке
Количество
50
40
30
20
10
0
20
40
60
80
Возраст

Рис. 12:

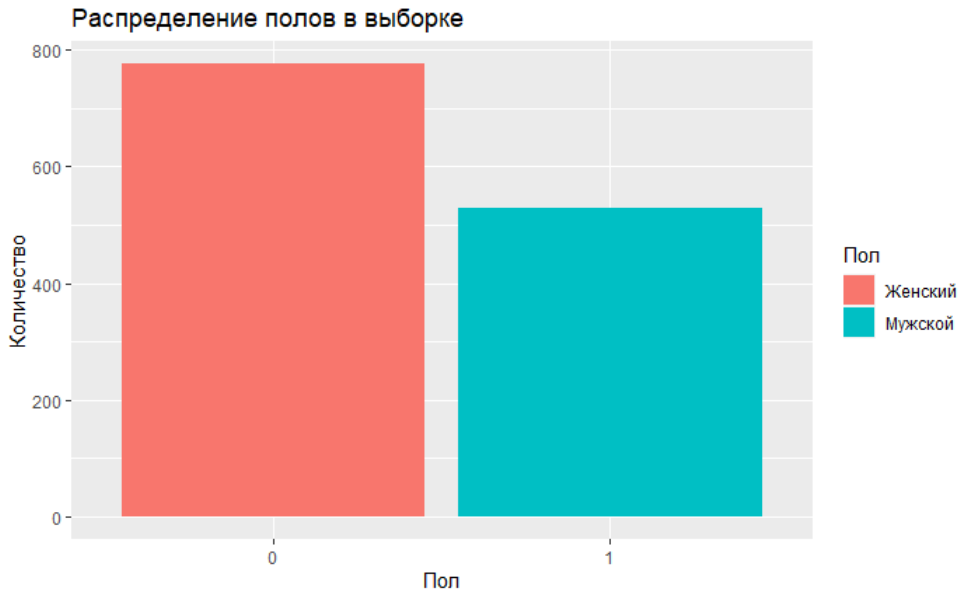

Рис. 13:

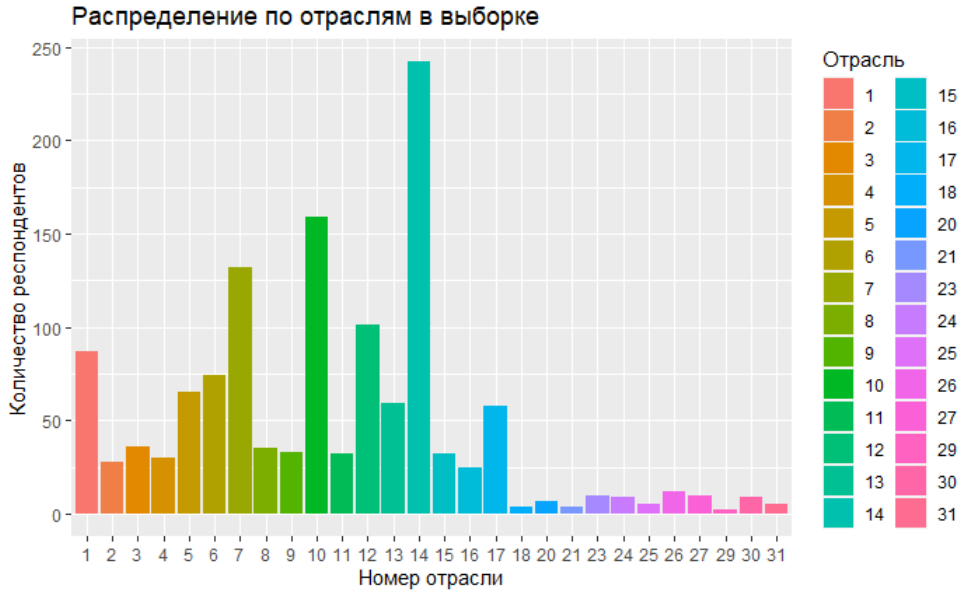

Рис. 14: (1 ЛЕГКАЯ, ПИЩЕВАЯ ПРОМЫШЛЕННОСТЬ, 2 ГРАЖДАНСКОЕ МАШИНОСТРОЕНИЕ 3 ВОЕННО-ПРОМЫШЛЕННЫЙ КОМПЛЕКС 4 НЕФТЕГАЗОВАЯ ПРОМЫШЛЕННОСТЬ 5 ДРУГАЯ ОТРАСЛЬ ТЯЖЕЛОЙ ПРОМЫШЛЕННОСТИ 6 СТРОИТЕЛЬСТВО 7 ТРАНСПОРТ, СВЯЗЬ 8 СЕЛЬСКОЕ ХОЗЯЙСТВО 9 ОРГАНЫ УПРАВЛЕНИЯ 10 ОБРАЗОВАНИЕ 11 НАУКА, КУЛЬТУРА 12 ЗДРАВООХРАНЕНИЕ 13 АРМИЯ, МВД, ОРГАНЫ БЕЗОПАСНОСТИ 14 ТОРГОВЛЯ, БЫТОВОЕ ОБСЛУЖИВАНИЕ 15 ФИНАНСЫ 16 ЭНЕРГЕТИЧЕСКАЯ ПРОМЫШЛЕННОСТЬ 17 ЖИЛИЩНО-КОММУНАЛЬНОЕ ХОЗЯЙСТВО 18 ОПЕРАЦИИ С НЕДВИЖИМОСТЬЮ 20 СОЦИАЛЬНОЕ ОБСЛУЖИВАНИЕ 21 ЮРИСПРУДЕНЦИЯ 22 ЦЕРКОВЬ 23 ХИМИЧЕСКАЯ ПРОМЫШЛЕННОСТЬ 24 ДЕРЕВООБРАБАТЫВАЮЩАЯ ПРОМЫШЛЕННОСТЬ, ЛЕСНОЕ ХОЗЯЙСТВО 25 СПОРТ, ТУРИЗМ, РАЗВЛЕЧЕНИЯ 26 УСЛУГИ НАСЕЛЕНИЮ 27 IT, ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ 28 ЭКОЛОГИЯ, ЗАЩИТА ОКРУЖАЮЩЕЙ СРЕДЫ 29 ОРГАНИЗАЦИЯ ОБЩЕСТВЕННОГО ПИТАНИЯ 30 СМИ, ИЗДАТЕЛЬСТВО, ПЕЧАТЬ, ТЕЛЕКОММУНИКАЦИИ 31 РЕКЛАМА, МАРКЕТИНГ 32 ОБЩЕСТВЕННЫЕ ОРГАНИЗАЦИИ, СОВЕТ ВЕТЕРАНОВ И ПР.)

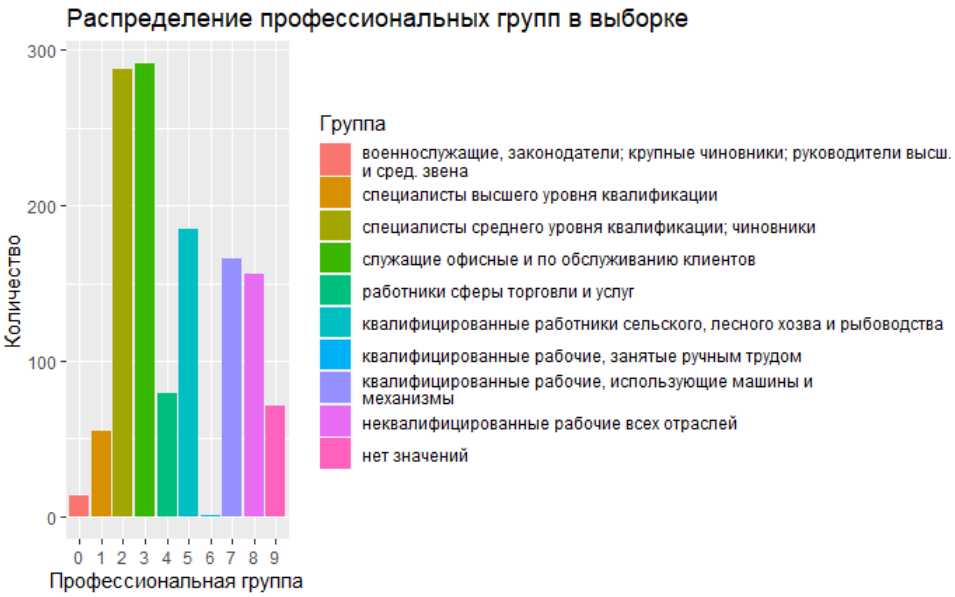
Распределение профессиональных групп в выборке
Количество
300
200
100
0
0 1 2 3 4 5 6 7 8 9
Профессиональная группа
Группа
военнослужащие, законодатели; крупные чиновники; руководители высш. и сред. звена
специалисты высшего уровня квалификации
специалисты среднего уровня квалификации; чиновники
служащие офисные и по обслуживанию клиентов
работники сферы торговли и услуг
квалифицированные работники сельского, лесного хозва и рыбоводства
квалифицированные рабочие, занятые ручным трудом
квалифицированные рабочие, использующие машины и механизмы
неквалифицированные рабочие всех отраслей
нет значений

Рис. 15:

Таблица 6: Список переменных

Название переменной	Описание переменной	Тип переменной	Вид переменной
sickdays	Сколько в среднем дней респондент пропустил в месяц за последний год	числовая	Зависимая переменная
awage	Среднемесячная зарплата на предприятии после вычета налогов за последние 12 месяцев - независимо от того, платят Вам ее вовремя или нет?	числовая	Главный регрессор
age	Количество полных лет	числовая	Контрольная переменная
harmwork	Является ли производство, на котором респондент работает, вредным или опасным, т.е. дающим Вам право на досрочное назначение трудовой пенсии, на дополнительные выплаты или льготы?	бинарная (0-нет, 1-да)	Контрольная переменная
status	Тип населенного пункта	Факторная (1-Областной центр, 2-Город, 3-ПГТ, 4-Село)	Контрольная переменная
gender	Пол респондента	Бинарная (0-нет, 1-да)	Контрольная переменная
hours	Сколько часов в среднем продолжается обычная рабочая неделя?	Числовая	Контрольная переменная
childcount	Сколько у респондента детей моложе 18 лет?	Числовая	Контрольная переменная
chronical	Количество хронических заболеваний	Числовая	Контрольная переменная
occup	Профессиональная группа	Факторная (10 разных групп)	Контрольная переменная
marst	Семейное положение	Факторная (6 разных видов)	Контрольная переменная
ofwork	Респондент оформлен на этой работе официально, то есть по трудовой книжке, трудовому соглашению, контракту?	Бинарная (1-да, 2-нет)	Только для анализа
entrepreneurship	Занимается ли респондент предпринимательской деятельностью (по его мнению)?	Бинарная (1-да, 2-нет)	Только для анализа
disability	Назначена ли респонденту какая-нибудь группа по инвалидности?	Бинарная (1-да, 2-нет)	Только для анализа
operations	В течение последних 12 месяцев респонденту делали хирургические операции?	Бинарная (1-да, 2-нет)	Только для анализа
branch	В какой отрасли респондент работает на этой работе?	Факторная (32 разных отрасли)	Только для анализа