

**РЭУ им. Г.В. Плеханова**

**Факультет Математической экономики и информатики**

**Направление Экономика**

# Моделирование и прогнозирование заболеваемости индивида в зависимости от различных социальных факторов

Студент - Сергеев Д.А.

Научный руководитель - к.э.н., доцент Закревская Е.А.



## Цель

- **Выявить** социально-экономические детерминанты здоровья
- **Смоделировать** вероятности наличия заболеваний



## Задачи

- **Сбор** микроэкономических данных РМЭЗ
- **Анализ** связности выбранных факторов
- **Построение** логистических регрессий для моделирования вероятности заболевания



## Объект

- **Российский Мониторинг Экономического положения и Здоровья населения (РМЭЗ)**



## Предмет

- **Социально-экономические факторы выборки РМЭЗ за 2013 год**

# Российский Мониторинг Экономического положения и Здоровья населения (РМЭЗ)

- Серия ежегодных общенациональных репрезентативных опросов на базе вероятностной стратифицированной многоступенчатой территориальной выборки
- Опрашивается более 16000 индивидов
- Около 5000 социально-экономических характеристик
- Двухуровневая структура информации:
  - Уровень индивидов
  - Уровень домохозяйства (семьи)

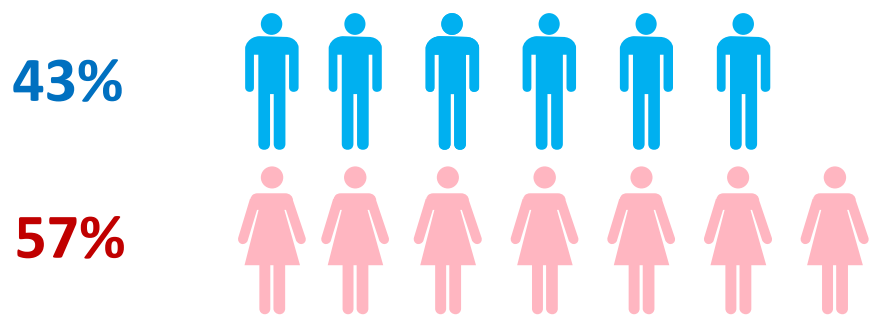
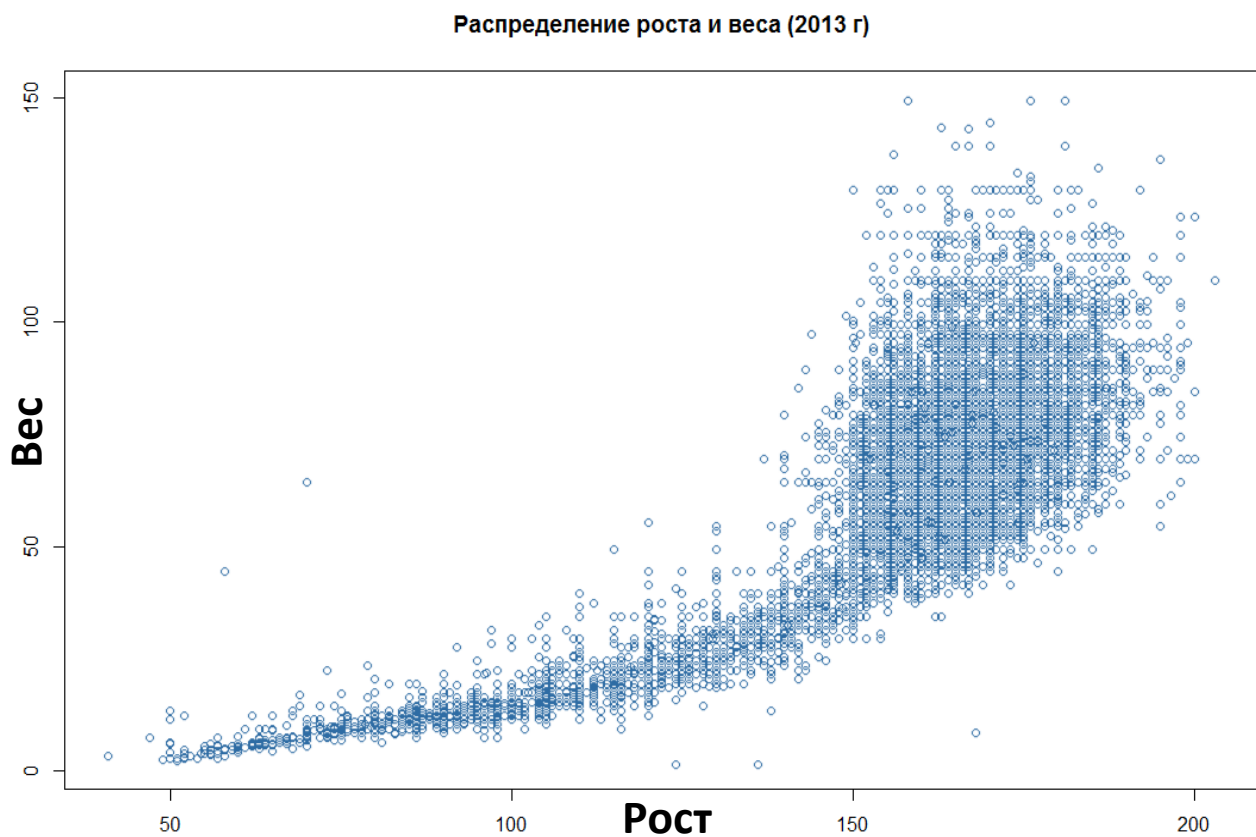
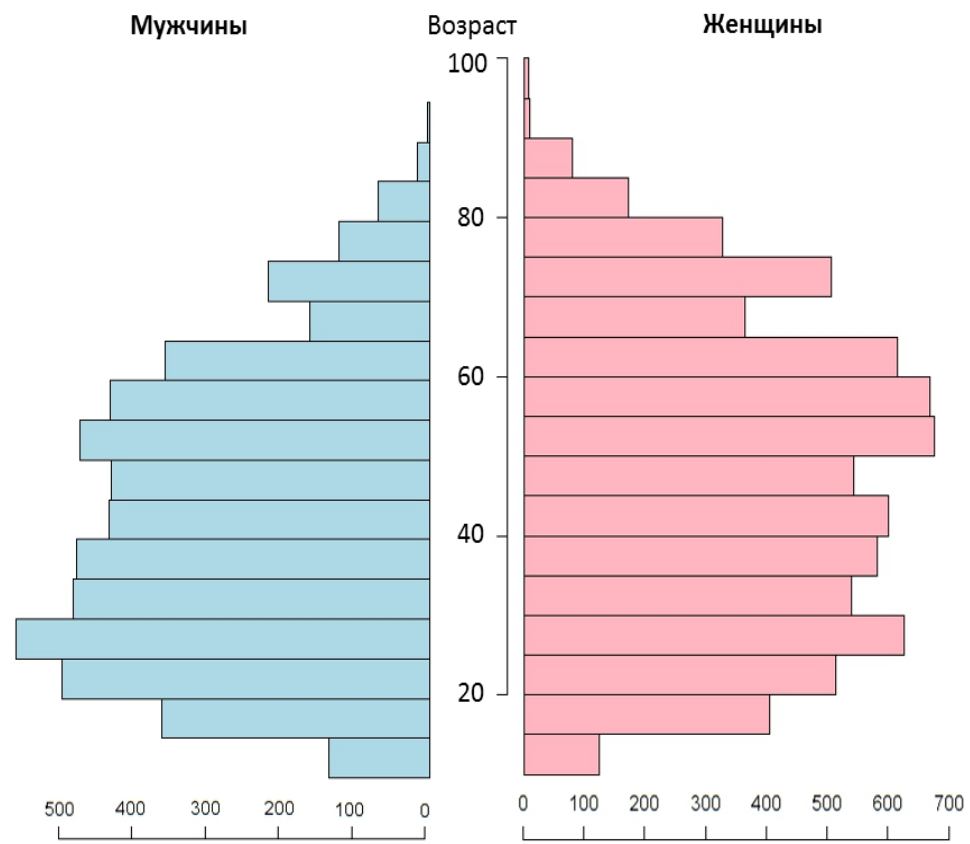


Карта проведения опросов РМЭЗ

# Отобранные для анализа переменные

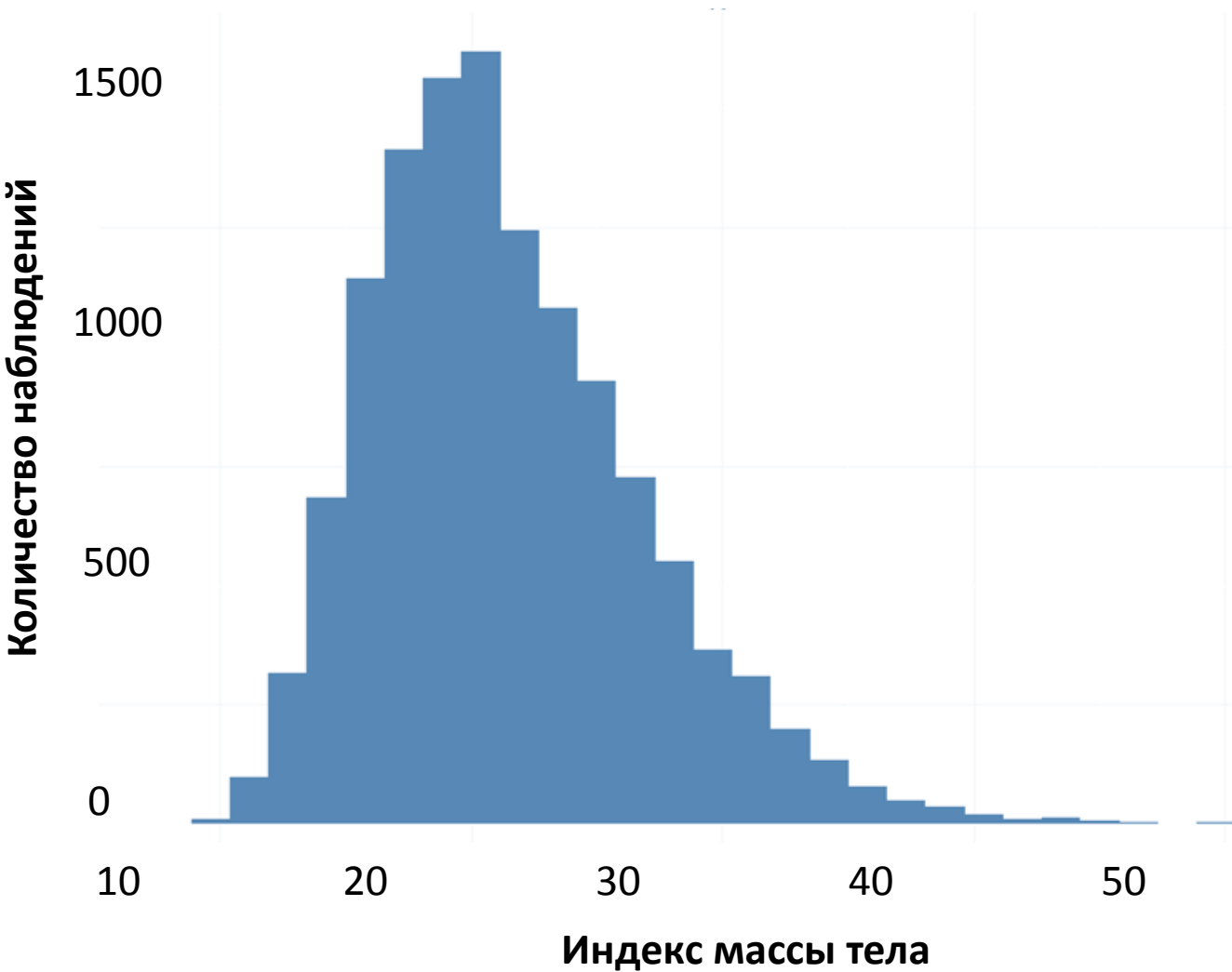
- Социальные характеристики Пол, возраст, семейное положение, место жительства и др.
- Здоровье Самооценка здоровья, наличие/отсутствие различных заболеваний, ИМТ, диеты и др.
- Работа Удовлетворенность работой/материальным положением, наличие работы, второй работы, отпуск и др.
- Вредные привычки Курит ли индивид, употребляет алкоголь, как часто, в каком объеме и др.
- Спорт Какими видами спорта занимается (бег, лыжи, тренажеры, плавание, аэробика и т.д)

# Характеристики выборки




Средний вес    Средний рост	
79.57 кг	174.7 см
71.76 кг	162.4 см

# Характеристики выборки




ИНДЕКС МАССЫ ТЕЛА	СООТВЕТСТВИЕ МЕЖДУ МАССОЙ ЧЕЛОВЕКА И ЕГО РОСТОМ
16 И МЕНЕЕ	Выраженный дефицит
16—18,5	Недостаточная (дефицит)
18,5—24,99	Норма
25—30	Избыточная(предожирение)
30—35	Ожирение первой степени
35—40	Ожирение второй степени
40 И БОЛЕЕ	Ожирение третьей степени

Средний ИМТ



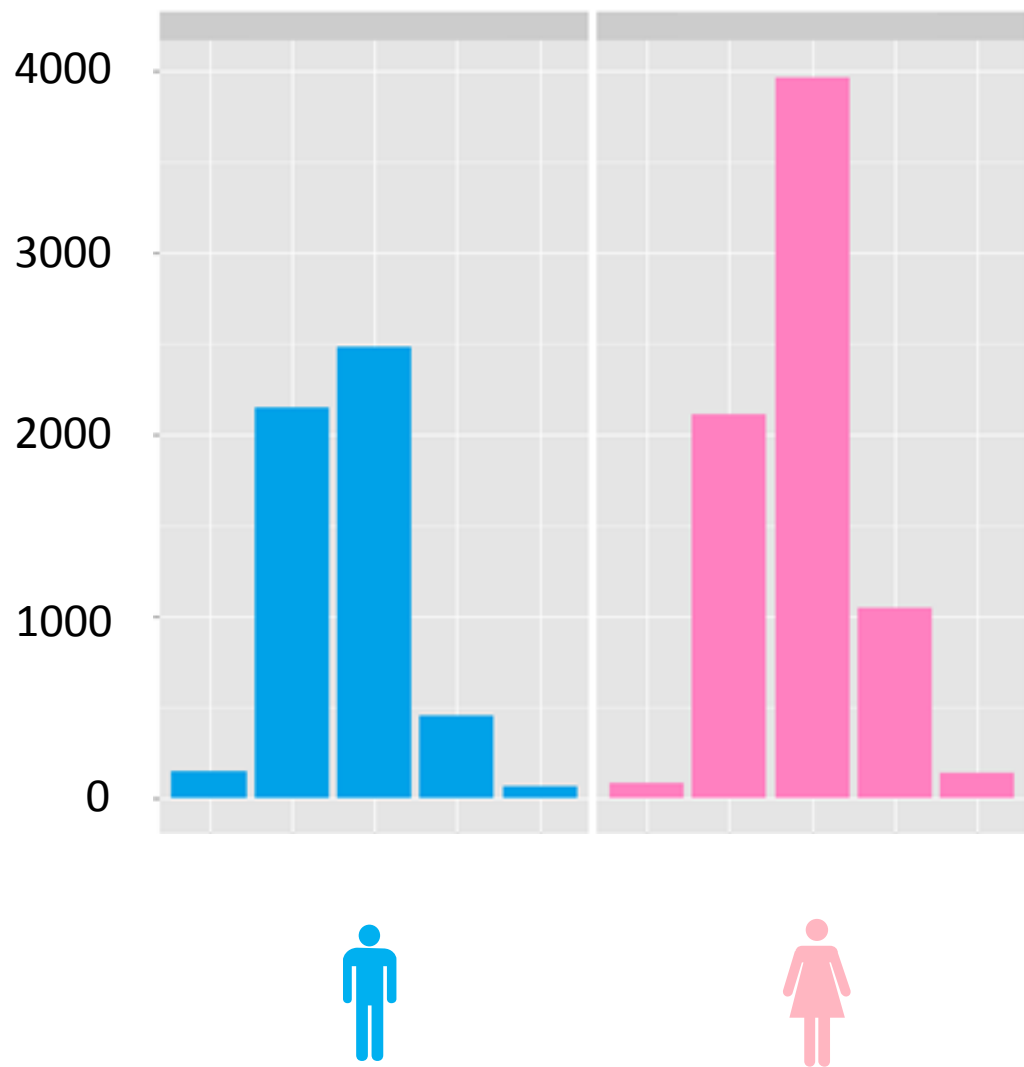
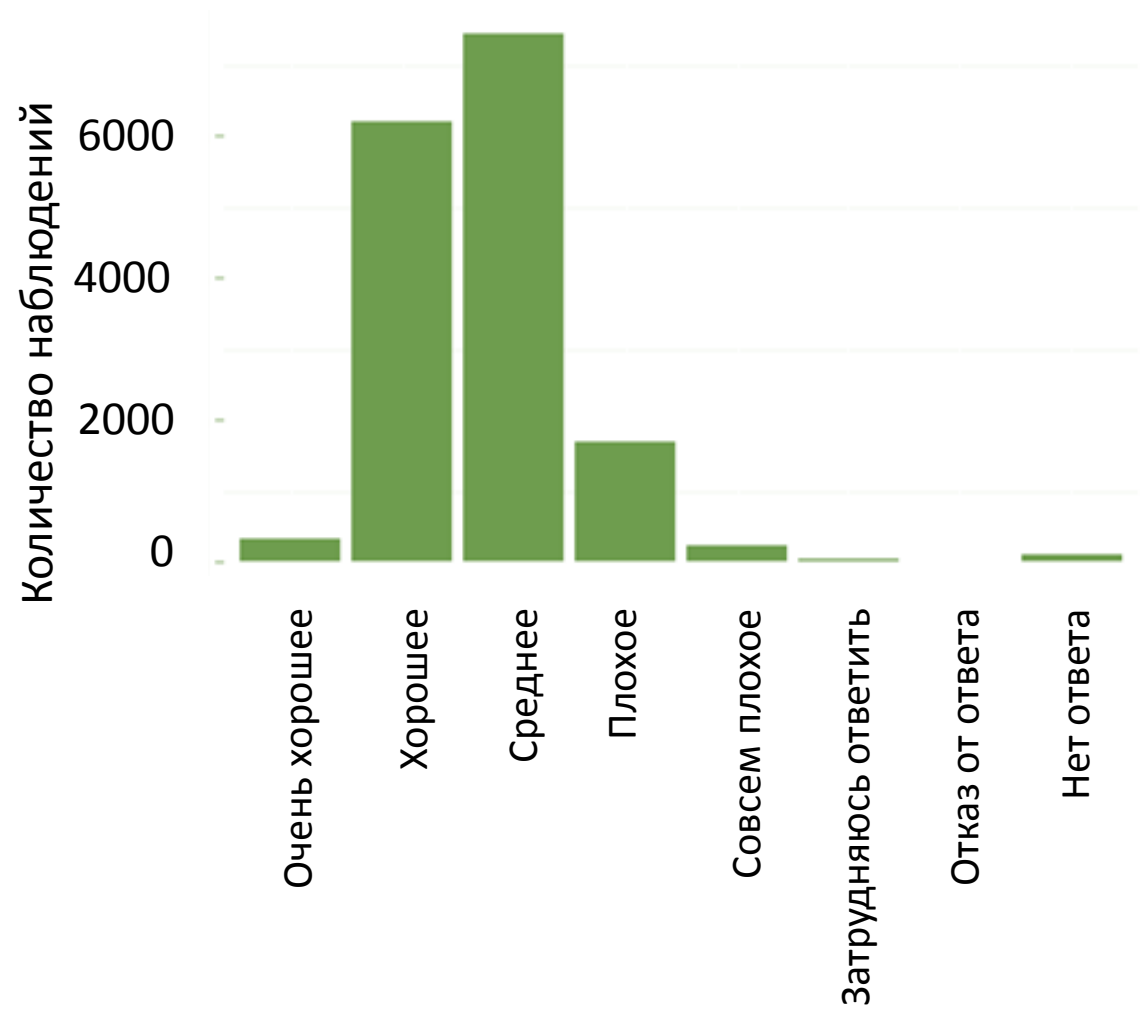
26.05



27.25

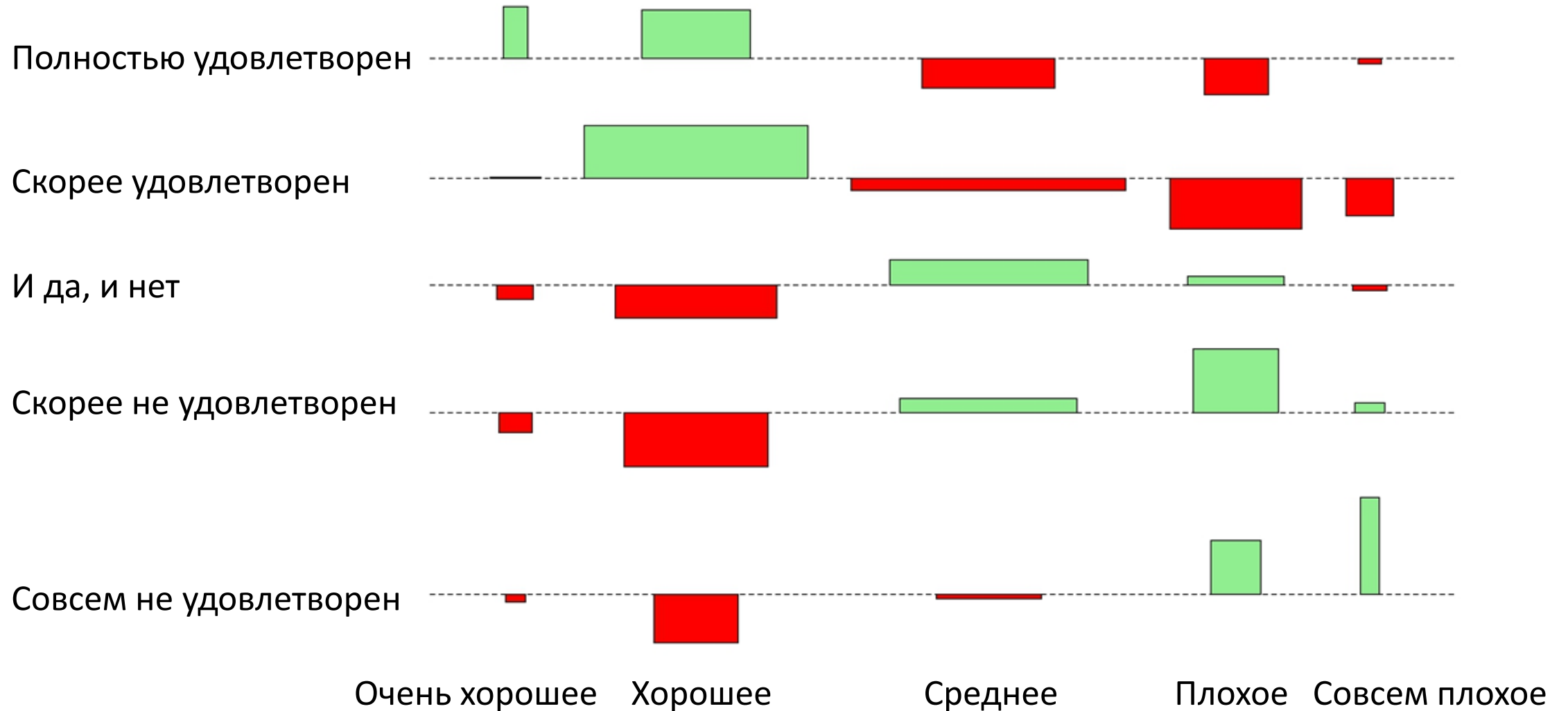
# Характеристики выборки

Самооценка состояния здоровья



Женщины более склонны негативно оценивать своё здоровье

# Связь между оценкой здоровья и уровнем удовлетворенности жизнью





# Характеристики выборки (болезни)

Болезнь	Количество наблюдений	Процент	Болезнь	Количество наблюдений	Процент
Заболевания сердца	2071	16,41	Неврология	952	7,54
Заболевания легких	949	7,52	Заболевания глаз	1716	13,60
Заболевания печени	1091	8,64	Аллергии	820	6,50
Заболевания почек	1083	8,58	Варикоз	1208	9,57
Заболевания пищеварения	2424	19,20	Заболевания кожного покрова	284	2,25
Заболевания позвоночника	2357	18,67	Онкологические заболевания	195	1,55
Диабет, повышенный сахар крови	1034	8,19	Гинекологические заболевания	618	4,90
Гипертоническая болезнь	3355	26,58	Заболевания мочеполовых органов	472	3,74
Заболевания суставов	2691	21,32	Инвалидность	1249	9,90
ЛОР-заболевания	1065	8,44			

# Инструментарий: Таблица сопряженности

Общий вид:

	1	...	$j$	...	$L$
1	$n_{11}$	...	$n_{1j}$	...	$n_{1L}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	...	$\vdots$
$i$	$n_{i1}$	...	$n_{ij}$	...	$n_{iL}$
$\vdots$	$\vdots$	...	$\vdots$	$\ddots$	$\vdots$
$K$	$n_{K1}$	...	$n_{Kj}$	...	$n_{KL}$

- $n_{ij} = \sum_{(x,y)} [x = i][y = j]$
- $n_i = \sum_j n_{ij}$
- $n_j = \sum_i n_{ij}$
- $n = \sum_i \sum_j n_{ij}$

Расчетное значение хи-квадрат статистики:

$$X^2 = \sum_{(i,j)} \frac{\left(n_{ij} - \frac{n_i n_j}{n}\right)^2}{\frac{n_i n_j}{n}}$$

$$X^2 \sim \chi^2_{(K-1)(L-1)}$$

# Инструментарий: Логистическая регрессия

Форма логистической модели:

- $\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$
- $g(x) = \ln \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \beta_0 + \beta_1 x$  Однофакторная логистическая модель
- $g(x) = \ln \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$  Многофакторная логистическая модель

$\pi(x) = E(Y|x)$  – условное математическое ожидание  
величины  $Y$  при условии наступления событий  $x$   
 $g(x)$  - логит-преобразование

# Инструментарий: Логистическая регрессия

Критерий Вальда для проверки значимости коэффициентов:

$$W = \hat{\beta}' [\widehat{Var}(\hat{\beta})]^{-1} \hat{\beta} = \hat{\beta}' (X \hat{V} X) \hat{\beta}$$

$$W \sim \chi^2_{(K-1)(L-1)}$$

где  $\hat{V} = \begin{bmatrix} \hat{\pi}_1(1 - \hat{\pi}_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{\pi}_n(1 - \hat{\pi}_n) \end{bmatrix}$

# Инструментарий: Логистическая регрессия (интерпретация)

Шанс того, что  $x = 1$  равен  $\frac{\pi(1)}{1-\pi(1)}$ ,  $x = 0$  равен  $\frac{\pi(0)}{1-\pi(0)}$

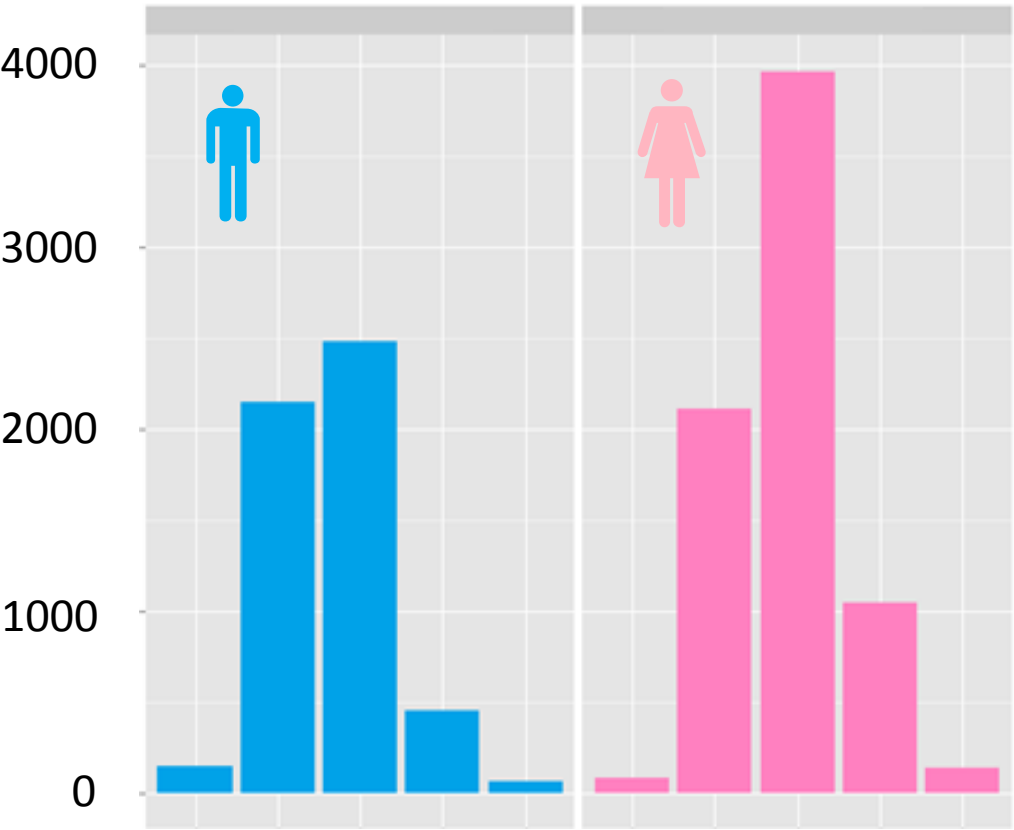
Отношение шансов (OR):

$$OR = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} = \frac{\frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}}{\frac{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)}{\left(\frac{1}{1+e^{\beta_0}}\right)}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{(\beta_0+\beta_1)-\beta_0} = e^{\beta_1}$$

# Анализ связности социально-экономических факторов

Состояние здоровья	Мужчины	Женщины	Сумма по строке
Доля в строке			
Очень хорошее	142	78	220
Доля в строке	0.645	0.355	0.017
Доля в столбце	0.027	0.011	
Доля в таблице	0.011	0.006	
Хорошее	2145	2113	4258
Доля в строке	0.504	0.496	0.337
Доля в столбце	0.406	0.288	
Доля в таблице	0.170	0.167	
Среднее	2480	3968	6448
Доля в строке	0.385	0.615	0.511
Доля в столбце	0.470	0.540	
Доля в таблице	0.196	0.314	
Плохое	452	1048	1500
Доля в строке	0.301	0.699	0.119
Доля в столбце	0.086	0.143	
Доля в таблице	0.036	0.083	
Очень плохое	59	137	196
Доля в строке	0.301	0.699	0.016
Доля в столбце	0.011	0.019	
Доля в таблице	0.005	0.011	
Сумма по столбцу	5278	7344	12622
Доля в столбце	0.418	0.582	

Таблица сопряженности пола респондента и его/её субъективной оценки своего состояния здоровья



● Женщины более склонны негативно оценивать своё здоровье

$\chi^2 = 299.96$        $p\text{-value} = 1.103309e-63$

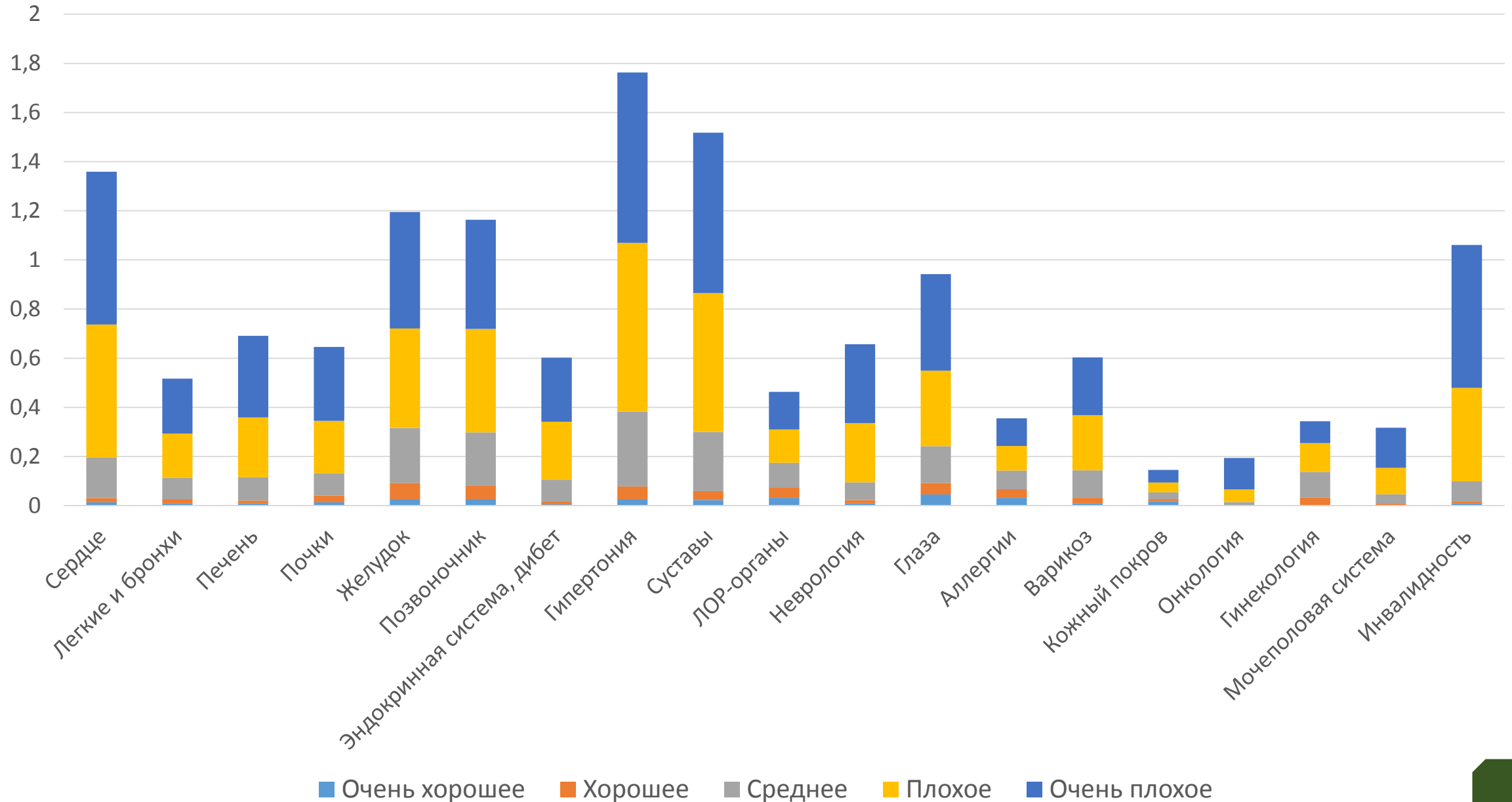
# Анализ связности социально-экономических факторов

	Статистика хи-квадрат	P-value
<i>Есть ли у индивида болезни сердца</i>	2568,783	0
<i>Есть ли у индивида болезни легких</i>	524,633	3,1471E-112
<i>Есть ли у индивида болезни печени</i>	941,285	1,8892E-202
<i>Есть ли у индивида болезни почек</i>	626,385	3,0154E-134
<i>Есть ли у индивида болезни желудка</i>	1064,005	4,7976E-229
<i>Есть ли у индивида болезни позвоночника</i>	1210,473	8,5454E-261
<i>Есть ли у индивида заболевания эндокринной системы, диабет или повышенный сахар крови</i>	866,138	3,6151E-186
<i>Есть ли у индивида гипертонические заболевания</i>	2665,903	0
<i>Есть ли у индивида заболевания суставов</i>	2208,380	0
<i>Есть ли у индивида заболевания ЛОР-органов</i>	190,824	3,5255E-40
<i>Есть ли у индивида неврологические заболевания</i>	1012,131	8,389E-218
<i>Есть ли у индивида заболевания глаз</i>	795,037	9,1298E-171
<i>Есть ли у индивида аллергии</i>	116,761	2,6255E-24
<i>Есть ли у индивида варикозные заболевания</i>	627,233	1,9766E-134
<i>Есть ли у индивида заболевания кожного покрова</i>	69,111	3,4967E-14
<i>Есть ли у индивида онкологические заболевания</i>	350,790	1,1839E-74
<i>Есть ли у индивида гинекологические заболевания</i>	114,380	8,4637E-24
<i>Есть ли у индивида заболевания мочеполовой системы</i>	404,893	2,4383E-86
<i>Есть ли у индивида какая-нибудь группа по инвалидности</i>	2285,181	0

● Каждое заболевание статистически значимо связано с субъективной оценкой состояния здоровья индивида

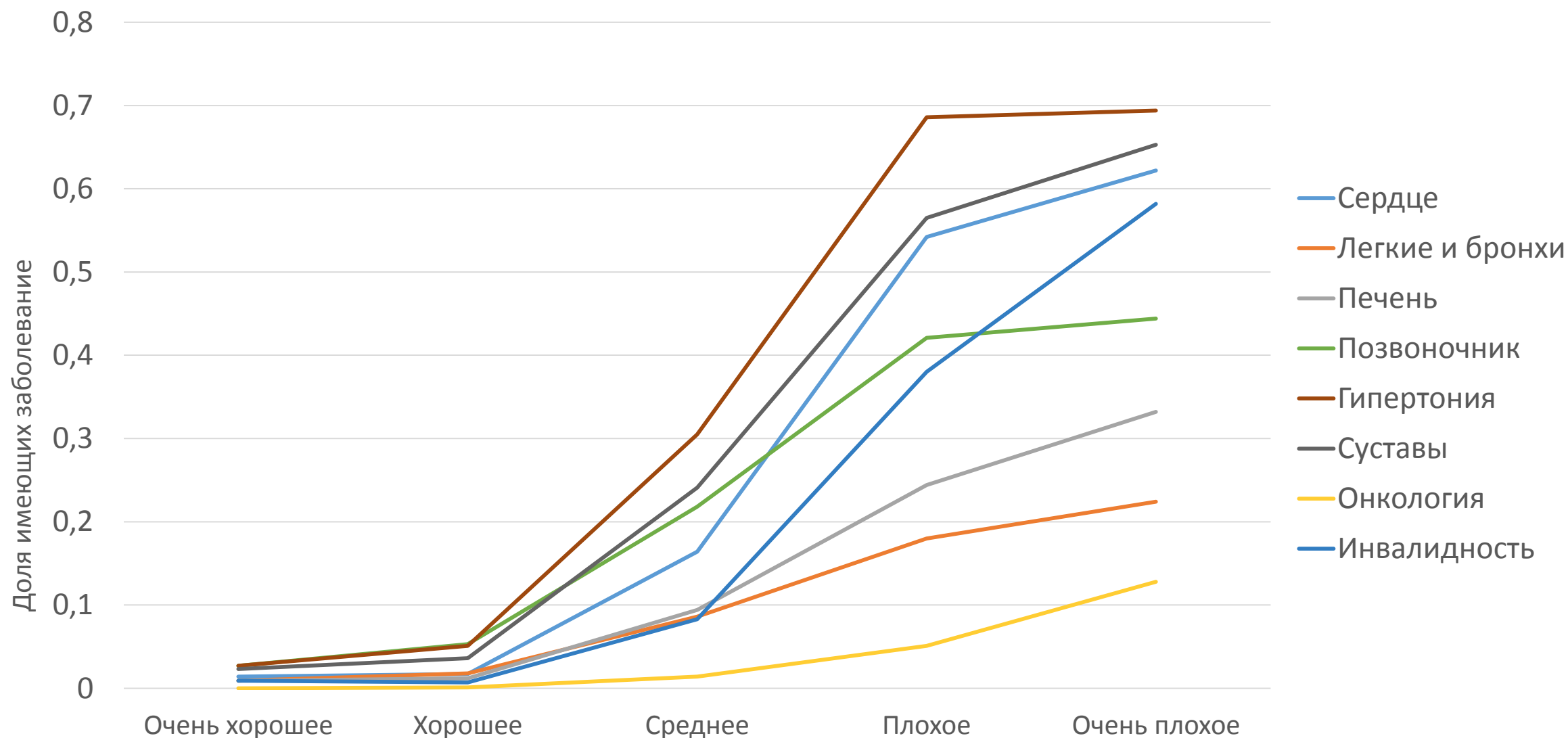
● Респонденты, имеющие определенные заболевания, более склонны считать своё состояние здоровья «Средним» или «Плохим»

# Доля имеющих заболевание и оценка здоровья





# Доля имеющих заболевание и оценка здоровья



# Анализ связности социально-экономических факторов

Результаты расчетов по таблицам сопряженности (спорт и самооценка здоровья)

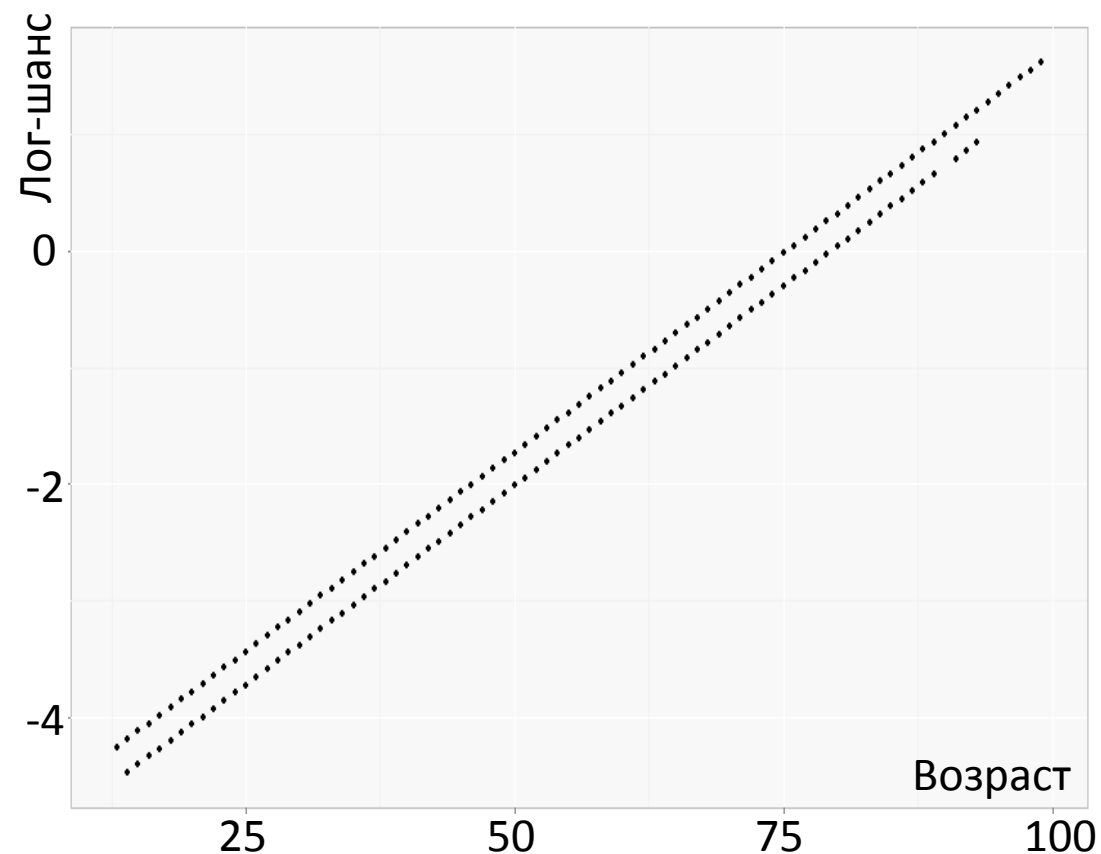
	<i>Хи-квадрат</i>	<i>P-value</i>
<i>Бег</i>	128.5195	2.2e-16
<i>Тренажеры</i>	169.5104	2.2e-16
<i>Прогулочная ходьба</i>	16.162	0.002809
<i>Велосипед</i>	59.3896	3.897e-12
<i>Плавание</i>	51.0728	2.155e-10
<i>Танцы, аэробика, шейпинг</i>	69.2877	3.209e-14
<i>Волейбол, футбол, баскетбол, хоккей</i>	345.1271	2.2e-16
<i>Борьба, бокс</i>	102.2679	2.2e-16

Для каждого вида спорта прослеживалась четкая тенденция снижения доли респондентов, отвечавших негативно о состоянии своего здоровья при занятиях любым из перечисленных видов физической активности

# Моделирование вероятности заболеваний: Сердце

$$g(x) = \beta_0 + \beta_1 \times \text{возраст} + \beta_2 \times \text{пол}$$

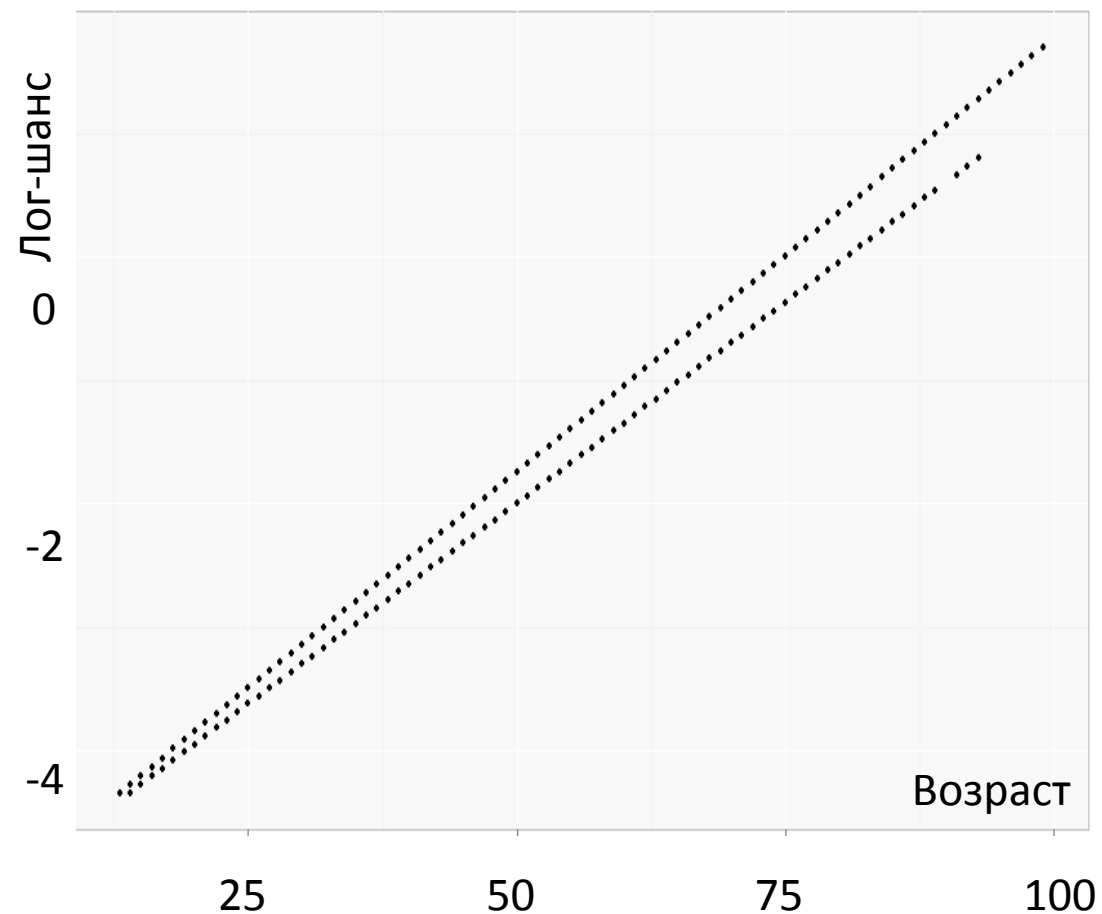
	Коэффициент	Станд. Ошибка	Z-статистика	Pr(> Z )
Константа	-5.4346	0.1085	-50.08	<0.0001
Возраст	0.0684	0.0017	39.42	<0.0001
Пол (0 – муж.)	0.2838	0.0566	5.02	<0.0001



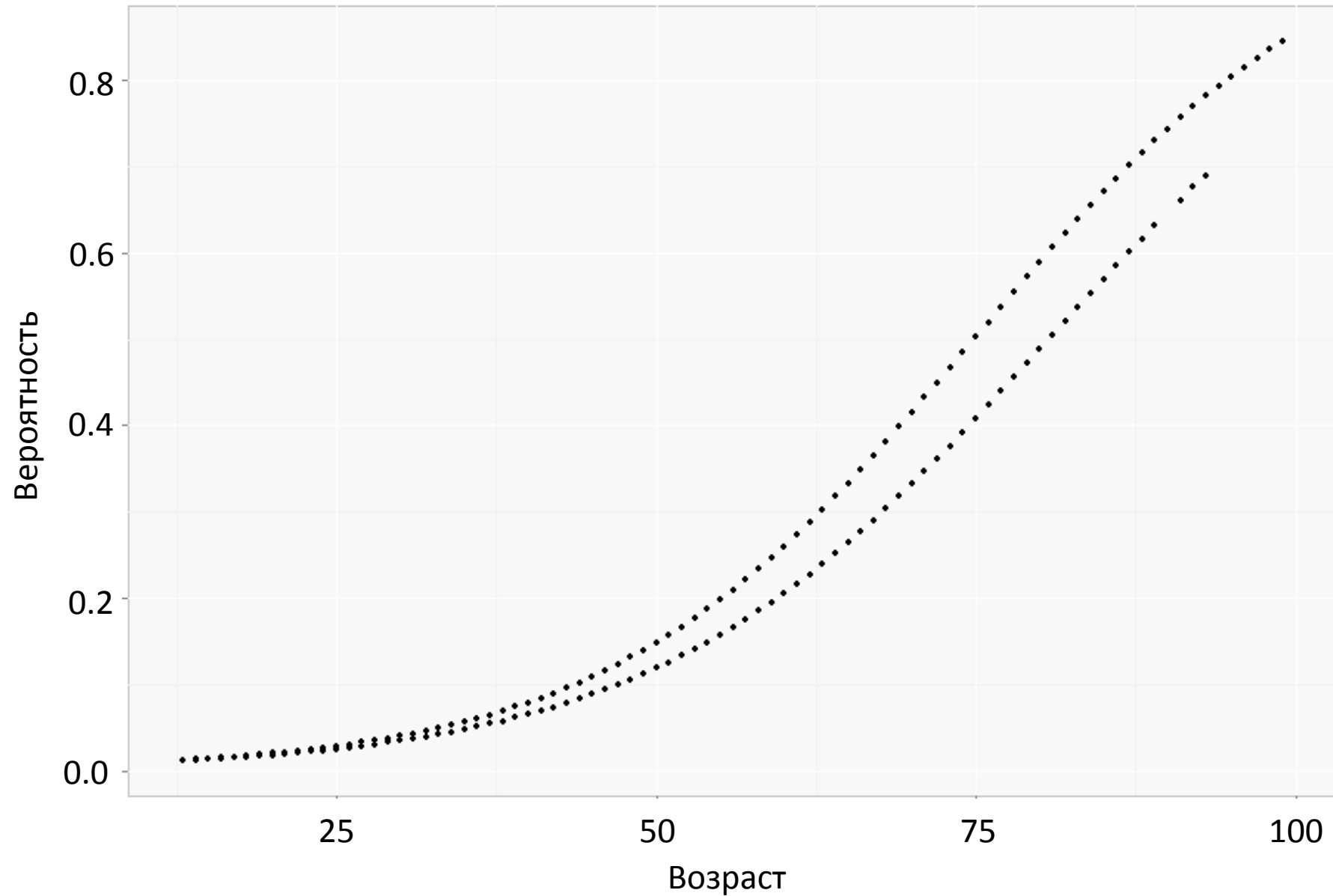
# Моделирование вероятности заболеваний: Сердце

$$g(x) = \beta_0 + \beta_1 \times \text{возраст} + \beta_2 \times \text{пол} \times \text{возраст}$$

	Коэфф.	Станд. Ошибка	Z-статистика	Pr(> z )
Конст.	-5.253	0.1051	-49.97	< 2e-16
Возраст	0.065	0.0018	34.26	< 2e-16
Возраст * Пол (0 – муж.)	0.005	0.0009	5.37	7.77e-08



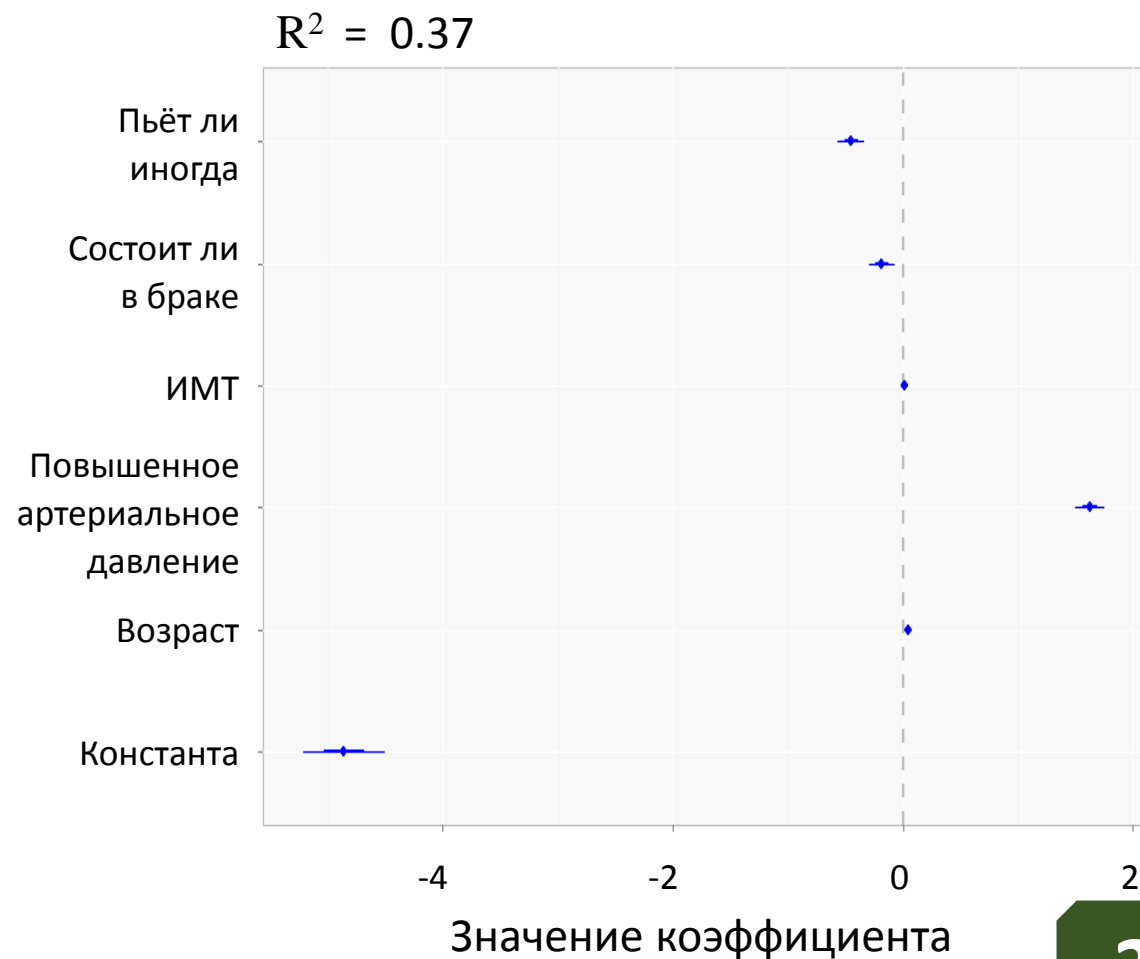
# Моделирование вероятности заболеваний: Сердце



# Моделирование вероятности заболеваний: Сердце

$$g(x) = \beta_0 + \beta_1 \times \text{возраст} + \beta_2 \times \text{повышенное артериальное давление} + \beta_3 \times \text{ИМТ} + \beta_4 \times \text{Состоит ли в браке} + \beta_5 \times \text{Пьёт ли иногда}$$

	Коэффициент	Станд. Ошибка	Z-статистика	Pr(> Z )
Константа	-4.8581	0.1793	-27.10	<0.0001
Возраст	0.0457	0.0019	23.64	<0.0001
Повышенное артериальное давление	1.6302	0.0635	25.66	<0.0001
ИМТ	0.0169	0.0052	3.26	0.0011
Состоит ли в браке	-0.1843	0.0570	-3.23	0.0012
Пьёт ли иногда	-0.4500	0.0581	-7.75	<0.0001



# Выводы по модели

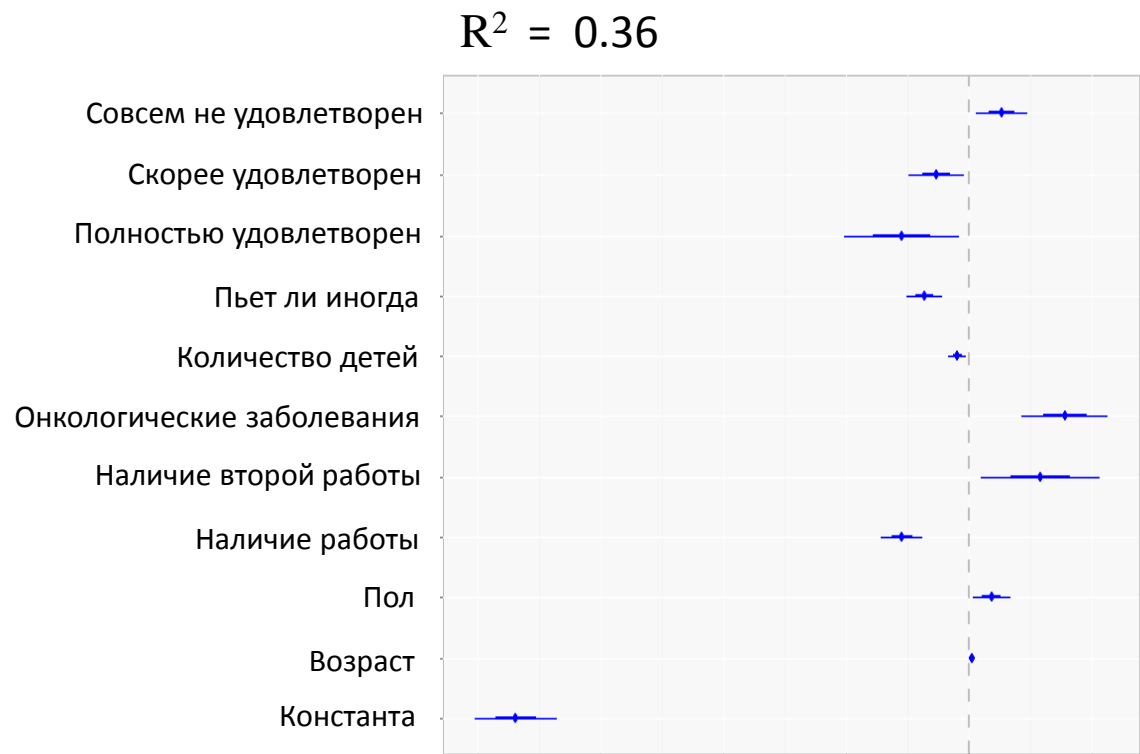
## Для заболеваний сердца:

- Повышенное артериальное давление значительно увеличивает шанс развития болезни
- С увеличением индекса массы тела вероятность заболевания увеличивается
- Вероятность наличия заболевания сердца у респондентов, состоящих в браке, ниже, чем у холостых
- Умеренное употребление алкоголя значительно снижает вероятность наличия болезни

# Моделирование вероятности заболеваний: Неврология

$$g(x) = \beta_0 + \beta_1 \times \text{возраст} + \beta_2 \times \text{пол} + \beta_3 \times \text{Наличие работы} + \beta_4 \times \text{Наличие второй работы} + \beta_5 \times \text{Наличие онкологических заболеваний} + \beta_6 \times \text{Количество детей} + \beta_7 \times \text{Пьёт ли иногда} + \beta_8 \times \text{Полностью удовлетворен материальным положением} + \beta_9 \times \text{Скорее удовлетворен м. п.} + \beta_{10} \times \text{Совсем не удовлетворен м. п.}$$

	Коэффициент	Станд. Ошибка	Z-статистика	Pr(> Z )
Константа	-3.6872	0.1675	-22.01	<0.0001
Возраст	0.0314	0.0022	14.12	<0.0001
Пол (0 – муж.)	0.1883	0.0762	2.47	0.0134
Наличие работы	-0.5433	0.0844	-6.44	<0.0001
Наличие второй работы	0.5854	0.2424	2.41	0.0158
Онкологические заболевания	0.7838	0.1758	4.46	<0.0001
Количество детей	-0.0917	0.0366	-2.50	0.0123
Пьёт ли иногда	-0.3590	0.0725	-4.95	<0.0001
Полностью удовлетворен	-0.5462	0.2345	-2.33	0.0199
Скорее удовлетворен	-0.2622	0.1132	-2.32	0.0206
Совсем не удовлетворен	0.2703	0.1047	2.58	0.0098





# Выводы по модели

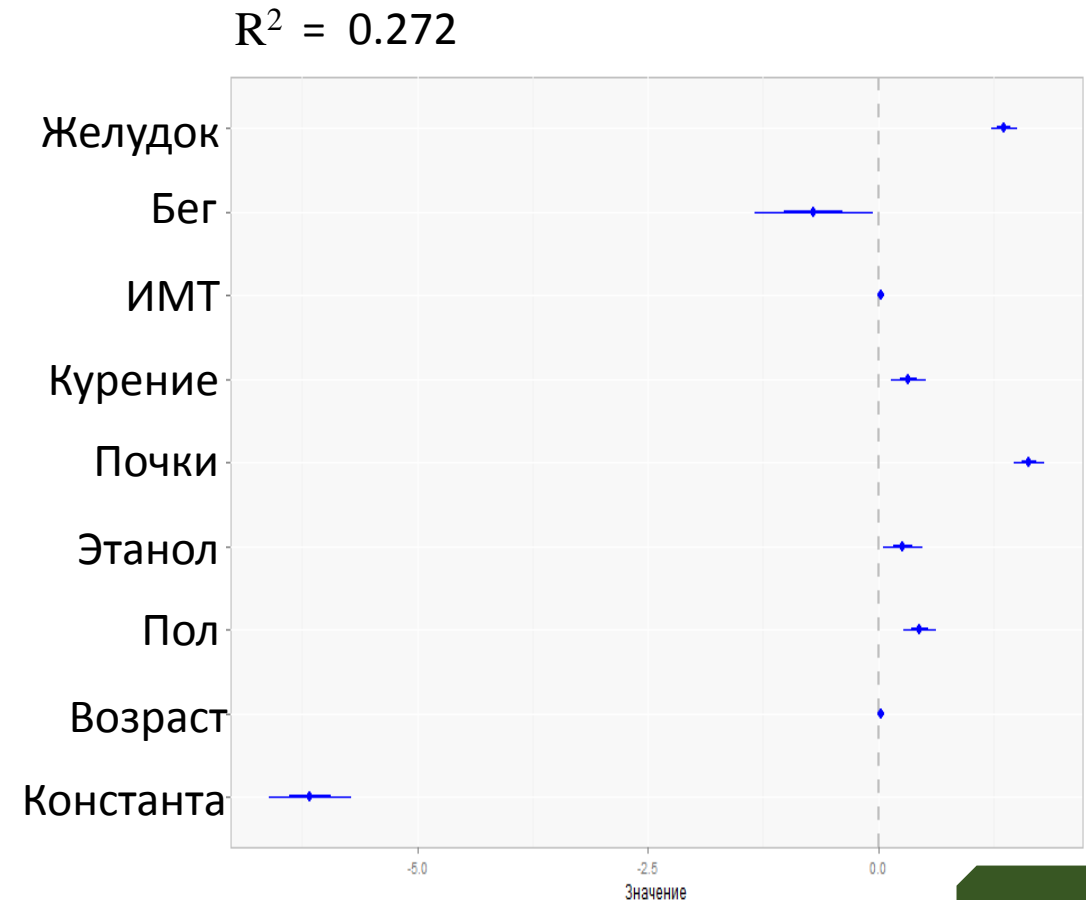
## Для неврологических заболеваний:

- Наличие работы снижает вероятность наличия неврологического заболевания, в то время как наличие второй работы значительно увеличивает её
- Онкологические заболевания значительно повышают шанс наличия неврологических заболеваний
- Наличие детей снижает вероятность развития заболевания
- Умеренное употребление алкоголя снижает шанс наличия болезни
- Удовлетворенность материальным положением значительно снижает вероятность наличия болезни, в то время как неудовлетворенность – повышает

# Моделирование вероятности заболеваний: Печень

$g(x) = \beta_0 + \beta_1 \times \text{возраст} + \beta_2 \times \text{пол} + \beta_3 \times \text{ИМТ} + \beta_4 \times \text{Употребляет больше 200 грамм этанола в месяц}$   
 $\beta_5 \times \text{Занимается бегом, коньками или лыжами} + \beta_6 \times \text{Наличие заболеваний почек} +$   
 $\beta_7 \times \text{Наличие заболеваний желудка}$

	Коэфф.	Станд. ошибка	Z-статистика	Pr(> Z )
Константа	-6.1625	0.2216	-27.81	<0.0001
Возраст	0.0326	0.0023	14.41	<0.0001
Пол (0 – муж.)	0.4483	0.0889	5.04	<0.0001
Употребляет больше 200 грамм этанола в месяц	0.2669	0.1059	2.52	0.0118
Наличие заболеваний почек	1.6358	0.0823	19.86	<0.0001
Курит ли респондент	0.3263	0.0954	3.42	0.0006
ИМТ	0.0347	0.0062	5.56	<0.0001
Бег, коньки, лыжи	-0.7016	0.3203	-2.19	0.0285
Наличие заболеваний желудка	1.3651	0.0719	18.99	<0.0001



# Выводы по модели

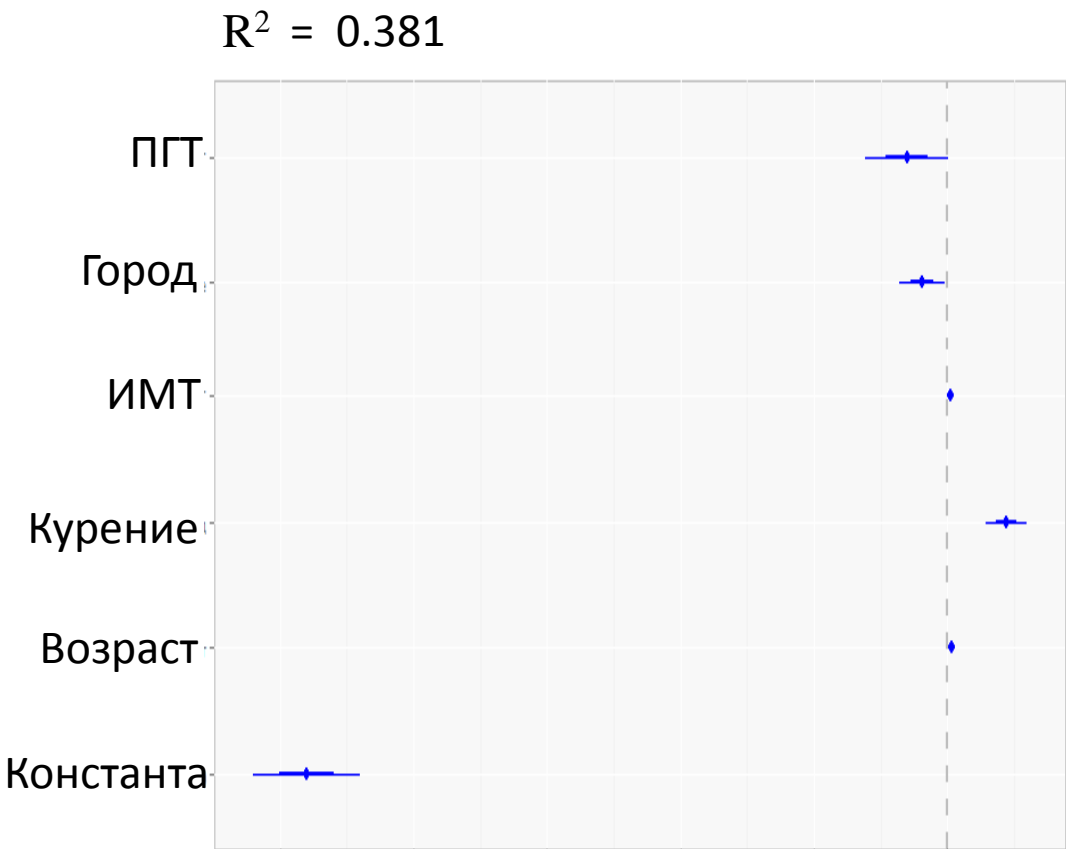
## Для заболеваний печени:

- Употребление более 200 грамм этанола в месяц сильно увеличивает шанс развития заболевания печени,
- Курение статистически значимо способствует развитию болезни
- Наличие заболеваний почек и желудка значительно повышает шансы наличия болезней печени
- Занятия аэробными видами спорта, такими как бег трусцой, коньки и лыжи сильно понижают вероятность наличия болезни печени

# Моделирование вероятности заболеваний: Легкие

$$g(x) = \beta_0 + \beta_1 \times \text{возраст} + \beta_2 \times \text{Курение} + \beta_3 \times \text{ИМТ} + \beta_4 \times \text{Проживает в городе} + \beta_6 \times \text{Проживает в ПГТ}$$

	Коэффициент	Станд. Ошибка	Z-статистика	Pr(> Z )
Константа	-4.8001	0.2013	-23.84	<0.0001
Возраст	0.0314	0.0021	15.22	<0.0001
Курит ли индивид	0.4423	0.0770	5.74	<0.0001
ИМТ	0.0238	0.0063	3.78	0.0002
Проживает в городе	-0.1900	0.0871	-2.18	0.0292
Проживает в ПГТ	-0.3026	0.1562	-1.94	0.0527



# Выводы по модели

## Для заболеваний легких:

- Курение значительно увеличивает шанс наличия заболеваний легких
- Возраст и ИМТ повышают вероятность наличия заболевания
- Место проживания респондента также статистически значимо влияет на вероятность заболевания

# Моделирование вероятности заболеваний: Желудок

$$g(x) = \beta_0 + \beta_1 \times \text{возраст} + \beta_2 \times \text{пол} + \beta_3 \times \text{Курение} + \beta_4 \times \text{Наличие диабета} + \beta_5 \times \text{Печень} + \beta_6 \times \text{Неврология} + \beta_7 \times \text{Диета} + \beta_8 \times \text{Областной центр} + \beta_9 \times \text{Село}$$

	Коэффициент	Станд. Ошибка	Z-статистика	Pr(> Z )
Константа	-3.0596	0.0969	-31.58	<0.0001
Возраст	0.0213	0.0014	15.11	<0.0001
Пол (0 – муж.)	0.2621	0.0557	4.70	<0.0001
Курение	0.2156	0.0599	3.60	0.0003
Диабет	0.3046	0.0787	3.87	0.0001
Сидел ли на диете	0.3373	0.0793	4.25	<0.0001
Печень	1.3984	0.0707	19.78	<0.0001
Неврология	0.7577	0.0774	9.79	<0.0001
Областной центр	0.2112	0.0599	3.52	0.0004
Село	-0.1447	0.0699	-2.07	0.0384



# Выводы по модели

## Для заболеваний желудка:

- Имеется статистически значимое превышение вероятности наличия заболевания у женщин над мужчинами
- Курение увеличивает шанс наличия болезни
- Наличие диабета, болезней печени и неврологических заболеваний значимо повышают вероятность наличия заболеваний желудка
- Место проживания также статистически значимо влияет на вероятность заболевания

# Заключение

- В работе были предложены методы оценки вероятности наличия заболевания у индивида в зависимости от социальных факторов
- Проведен анализ связности социально-экономических факторов
- Выявлены социально-экономические детерминанты здоровья
- Построены многофакторные логистические регрессии для моделирование вероятности наличия заболевания
- По результатам построения моделей была проведена интерпретация полученных коэффициентов



Спасибо за внимание!