Домашняя работа

1. Описание данных и выбор регрессоров

Целью данной работы является исследование факторов, влияющих на размер заработной платы. Анализ проведен по данным РМЭЗ за 2013 г. (репрезентативная выборка по индивидам, волна 22).

В нашем наборе данных представлено 16087 наблюдений по 857 переменным. Зависимой переменной по условиям задания является среднемесячная зарплата в тыс. руб. (wage), на основе переменной rj13.2.

Ниже перечислены переменные, отобранные в качестве регрессоров, и приведены коды переменных, лежащих в основе выбранных регрессоров:

- Возраст (age). Рассчитывался как разница между 2013 (год проведения опроса) и датой рождения индивида (rh6);
- Пол (sex). Фиктивная переменная на основе rh5: для индивидов мужского пола принимает значение 1, для индивидов женского пола — 0.
- Образование (lower, mid, midspecial, high). 4 фиктивных переменных на основе r_diplom: lower принимает значение 1 при наличии незаконченного среднего образования, mid — законченного среднего образования, midspecial — законченного среднего специального образования, higher — законченного высшего образования и выше.
- Владение иностранным языком, помимо языков бывших республик СССР (lang). Фиктивная переменная на основе rj260: принимает значение 1 при владении респондентом иностранным языком.
- Рост в сантиметрах (height). Переменная rm2.

Переменная lang выбрана для проверки гипотезы о влиянии знания иностранного языка на размер среднемесячной зарплаты. Предполагается положительное влияние, так как специалист со знанием иностранного языка является более квалифицированным и востребованным на рынке труда.

Переменная height выбрана для проверки гипотезы о наличии дискриминации в отношении работников с более низким ростом. Предполагается положительное влияние роста на размер среднемесячной зарплаты вследствие наличия определенных психологических эффектов — более высокие работники воспринимаются как более уверенные в себе, к их мнению могут чаще прислушиваться и так далее.

Ниже приведена таблица с описательными статистиками всех отобранных переменных, включая фиктивные.

Таблица 1: Описательные статистики.

	min	max	mean	median	se	n
wage	0	300	19.7912	16	0.2044	6080
age	0	100	40.0691	40	0.1774	16087
sex	0	1	0.4317	0	0.0039	16087
lower	0	1	0.2029	0	0.0034	13595
mid	0	1	0.3227	0	0.0040	13595
midspecial	0	1	0.2359	0	0.0036	13595
high	0	1	0.2360	0	0.0036	13595
lang	0	1	0.1836	0	0.0033	13595
height	41	203	160.0321	165	0.1778	15724

2. Гистограммы

На рис. 1 и 2 представлены гистограммы дохода и возраста. Следует отметить, что в модель попадет только 6080 из 16087 наблюдений, так как число индивидов, указавших свой среднемесячный доход (6080), меньше общего числа наблюдений и является наименьшим по сравнению с другими регрессорами.

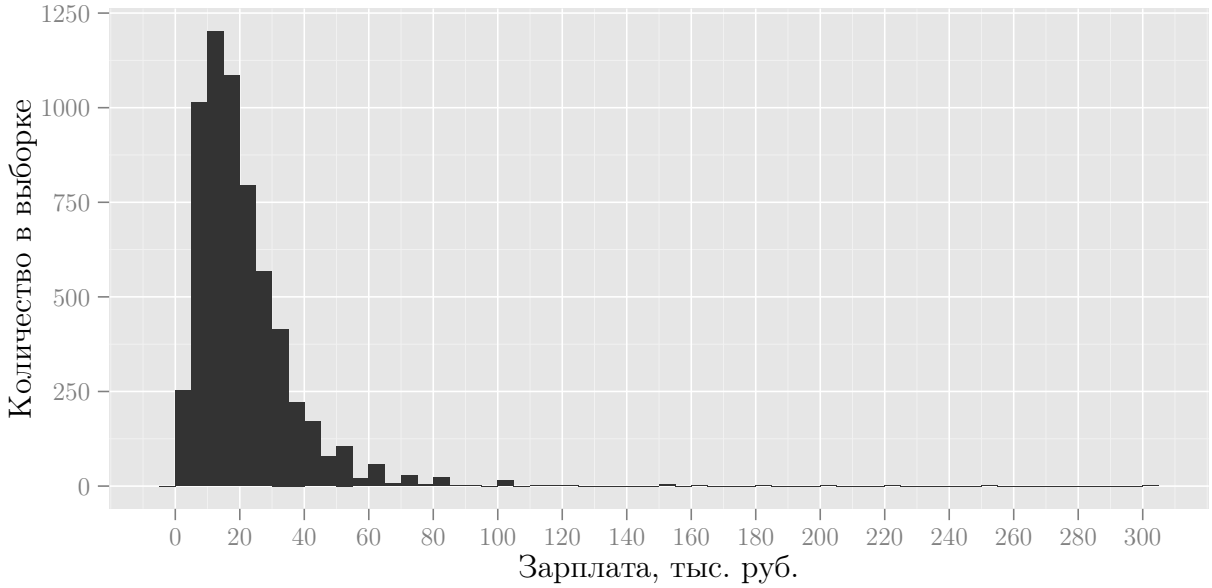

Рис. 1: Гистограмма дохода.

По гистограмме можно определить, что большинство индивидов получают среднемесячную зарплату менее 65 тыс. руб., и лишь немногие — выше 100 тыс. руб.

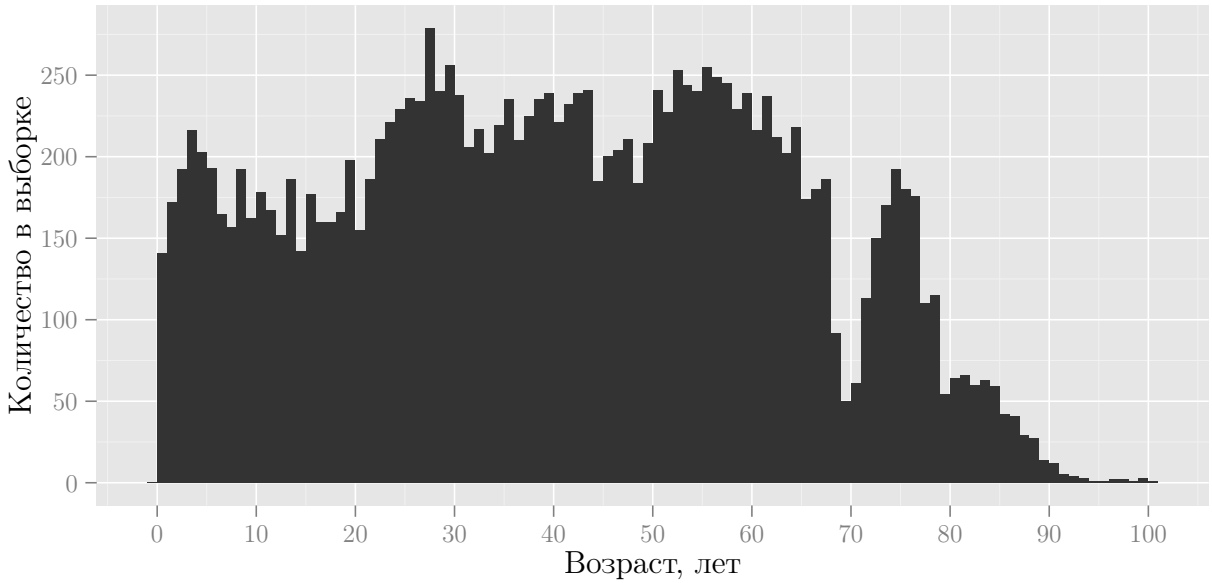

Рис. 2: Гистограмма возраста.

На гистограмме видны демографические ямы 1941–1944 гг. (ВОВ), а также 1992 и 1998 года.

3. Оценка модели

Предположим, что зависимость выглядит следующим образом:

$wage_i = \beta_0 + \beta_1 age_i + \beta_2 sex_i + \beta_3 mid_i + \beta_4 midspecial_i + \beta_5 high_i + \beta_6 lang_i + \beta_7 height_i + \varepsilon_i.$

Оценим нашу модель:

$\widehat{wage}_i = \widehat{\beta}_0 + \widehat{\beta}_1 age_i + \widehat{\beta}_2 sex_i + \widehat{\beta}_3 mid_i + \widehat{\beta}_4 midspecial_i + \widehat{\beta}_5 high_i + \widehat{\beta}_6 lang_i + \widehat{\beta}_7 height_i.$

Ниже представлены результаты оценки модели.

Таблица 2: Первоначальные оценки параметров модели.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-19.0898	4.9997	-3.8182	0.0001
age	-0.0382	0.0158	-2.4245	0.0154
sex	5.9574	0.5357	11.1216	0.0000
mid	2.1278	0.7375	2.8852	0.0039
midspecial	4.0665	0.7568	5.3732	0.0000
high	9.3803	0.7699	12.1837	0.0000
lang	4.7962	0.5305	9.0404	0.0000
height	0.1895	0.0297	6.3880	0.0000

Все переменные (age, sex, mid, midspecial, high, lang, height) оказались значимыми на 5% уровне значимости. При этом коэффициент перед всеми регрессорами, за исключением возраста, положителен. Это означает, что:

- Образование положительно влияет на уровень среднемесячного дохода, при этом (при прочих равных):
  - Наличие законченного среднего образования приводит к росту дохода на 2.1 тыс. руб. в месяц;
  - Наличие законченного среднего специального образования приводит к росту дохода на 4.1 тыс. руб. в месяц;
  - Наличие законченного высшего образования приводит к росту дохода на 9.4 тыс. руб. в месяц;
- Знание иностранного языка позволяет при прочих равных получать ежемесячно на 4.8 тыс. руб. больше;
- Можно говорить о наличии половой дискриминации, так как при прочих равных мужчина получает на 6 тыс. руб. в месяц больше женщины;
- Присутствует и дискриминация по росту: независимо от пола индивид, который на 1 см выше, имеет среднемесячный доход на 190 руб. больше;
- Чем старше индивид, тем меньше его среднемесячный доход. Взросление на год приводит к падению дохода в среднем на 38 руб. в месяц.

Таким образом, можно сказать, что гипотеза о влиянии двух выбранных регрессоров не отвергается, и они были выбраны корректно.

4. Робастные ошибки

Оценим нашу модель, используя робастные ошибки:

Таблица 3: Оценки параметров при использовании робастных ошибок.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-19.0898	4.7286	-4.0371	0.0001
age	-0.0382	0.0141	-2.7112	0.0067
sex	5.9574	0.5573	10.6897	0.0000
mid	2.1278	0.4897	4.3447	0.0000
midspecial	4.0665	0.5290	7.6876	0.0000
high	9.3803	0.5971	15.7109	0.0000
lang	4.7962	0.6280	7.6369	0.0000
height	0.1895	0.0287	6.6126	0.0000

Мы видим, что оценки параметров не поменялись. Все переменные (age, sex, mid, midspecial, high, lang, height) остались значимы (на 5% уровне значимости), регрессор age теперь значим, как и остальные, на 1% уровне значимости. Ниже представлена таблица изменений в оценках параметров.

Таблица 4: Различия в оценках параметров.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0000	-0.2711	-0.2189	-0.0001
age	0.0000	-0.0017	-0.2867	-0.0086
sex	0.0000	0.0216	-0.4319	0.0000
mid	0.0000	-0.2477	1.4595	-0.0039
midspecial	0.0000	-0.2278	2.3144	0.0000
high	0.0000	-0.1729	3.5272	0.0000
lang	0.0000	0.0975	-1.4036	0.0000
height	0.0000	-0.0010	0.2245	0.0000

Применение робастных ошибок привело к снижению (а не повышению) стандартных ошибок у всех регрессоров, кроме sex и lang. Почему это произошло?

Протестируем нашу модель на гетероскедастичность, используя тесты Уайта и Гольдфельда-Квандта:

Таблица 5: Тест Уайта

Test statistic	df	P value
35.4171	7	0 * * *

Таблица 6: Тест Гольдфельда-Квандта, переменная age.

Test statistic	df1	df2	P value
0.7408	2402	2401	1

Таблица 7: Тест Гольдфельда-Квандта, переменная height.

Test statistic	df1	df2	P value
2.4586	2402	2401	0 * * *

Тест Уайта показал наличие гетероскедастичности; тест Гольдфельда-Квандта не отверг гипотезу об условной гомоскедастичности в случае зависимости остатков от переменной age и показал наличие гетероскедастичности для случая зависимости остатков от переменной height.

Посмотрим на эту зависимость:

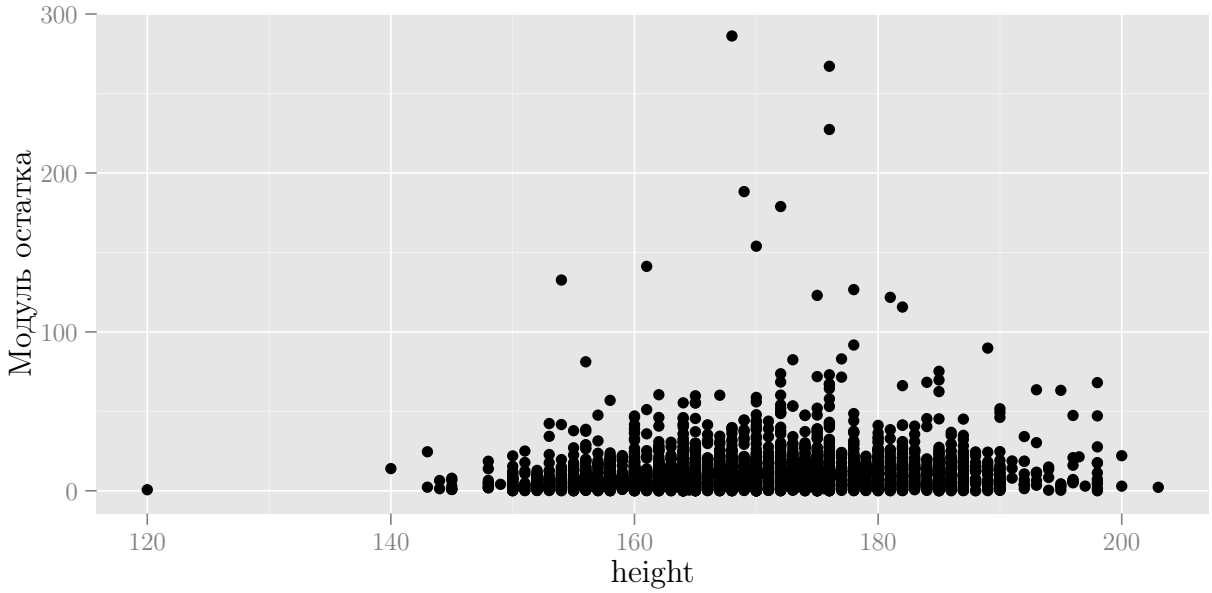

Рис. 3: Зависимость остатков от переменной height.

Из рис. 3 видно, что наибольшая дисперсия у остатков для тех наблюдений, значение переменной height которой близко к среднему значению. Это логично, ведь число индивидов и со слишком низким, и со слишком высоким ростом мало по сравнению с числом индивидов среднего роста. Следовательно, у индивидов среднего роста более высок разброс среднемесячной зарплаты. А отрицательная корреляция ε_i^2 и $(x_i - \bar{x})^2$ эквивалентна завышению стандартных ошибок, получаемых с помощью стандартного МНК, по сравнению с робастными ошибками.

5. Приложение. Использованные команды

5.0. Обработка данных

```
setwd("~/Documents/Coursera/Econometrics/Homework")
data1 <- read.rlms("r22i_os25a.sav")
rlms_sav2rds("~/Documents/Coursera/Econometrics/Homework")
```

5.1. Описание данных и выбор регрессоров

```
library("memisc")
library("lmtest")
library("psych")
library("sandwich")
library("glmnet")
library("ggplot2")
library("car")
library("dplyr")
library("broom")
library("foreign")
library("vcd")
library("devtools")
library("hexbin")
library("pander")
library("sjPlot")
library("knitr")
library("rlms")
library("tikzsetup")
tikzsetup()
data <- readRDS("r22i_os25a.Rds")
df <- select(data, rj13.2, rh6, rh5, r_diplom, rj260, rm2)
df$age <- 2013-df$rh6
df$wage <- df$rj13.2 / 1000
df$sex <- as.integer(df$rh5=="МУЖСКОЙ")
df$lower <- as.integer(df$r_diplom=="окончил 0 - 6 классов" |
                       df$r_diplom=="незаконч среднее образование (7 - 8 кл)" |
                       df$r_diplom=="незаконч среднее образование (7 - 8 кл) + что-то еще")
df$mid <- as.integer(df$r_diplom=="законч среднее образование")
df$midspecial <- as.integer(df$r_diplom=="законч среднее специальное образование")
df$high <- as.integer(df$r_diplom=="законч высшее образование и выше")
df$lang <- as.integer(df$rj260=="Да")
df$height <- df$rm2
d <- select(df, wage, age, sex, lower, mid, midspecial, high, lang, height)
table1 <- psych::describe(d)
table <- select(table1, min, max, mean, median, se, n)
panderOptions('digits', 6)
panderOptions('round', 4)
panderOptions('keep.trailing.zeros', TRUE)
table <- data.frame(table)
pander(table, split.table = Inf, caption = "Описательные статистики.")
```

5.2. Гистограммы

```
qplot(data=d, wage, xlab = "Зарплата, тыс. руб.", ylab = "Количество в выборке", binwidth = 5) +
  scale_x_continuous(breaks=seq(0,300,20))
qplot(data=d, age, xlab = "Возраст, лет", ylab = "Количество в выборке", binwidth = 1) +
  scale_x_continuous(breaks=seq(0,100,10)) + scale_y_continuous(breaks=seq(0,300,50))
```

5.3. Оценка модели

```
model1 <- lm(data=d, wage~age+sex+mid+midspecial+high+lang+height)
table2 <- coeftest(model1)[1:8,]
table3 <- coeftest(model1, vcov. = vcovHC(model1))[1:8,]
pander(table2, caption = "Первоначальные оценки параметров модели.")
```

5.4. Робастные ошибки

```
pander(table3, caption = "Оценки параметров при использовании робастных ошибок.")
table4 <- table3
for (i in 1:4){
  table4[,i] <- table3[,i] - table2[,i]
}
pander(table4, caption = "Различия в оценках параметров.")
pander(bptest(model1), caption = "Тест Уайта")
d <- augment(model1, d)
pander(gqtest(model1, order.by = ~age, data=d, fraction=0.2),
       caption = "Тест Гольдфельда-Квандта, переменная age.")
pander(gqtest(model1, order.by = ~height, data=d, fraction=0.2),
       caption = "Тест Гольдфельда-Квандта, переменная height.")
qplot(data=d, height, abs(.resid), xlab = "height", ylab = "Модуль остатка")
```