Титульный лист

Оглавление

[Оглавление 1](#_Toc136608561)

[Задание №1 2](#_Toc136608562)

[Задание №2 4](#_Toc136608563)

[Задание №3 8](#_Toc136608564)

[Приложения 10](#_Toc136608565)

[Приложение 1 10](#_Toc136608566)

[Код задачи 1: 10](#_Toc136608567)

[Полученные графики зависимостей моделей: 11](#_Toc136608568)

[Приложение 2 12](#_Toc136608569)

[Код задачи 2 (2.1 и 2.2): 12](#_Toc136608570)

[Приложение 3 15](#_Toc136608571)

Задание №1

**Введение**

Объясняемая переменная: Education

Объясняющие переменные (регрессоры): Agriculture, Fertility

Обработать набор данных swiss, встроенный в R:

1. Оценить среднее значение, дисперсию и СКО объясняемой и объясняющих переменных
2. Построить зависимости вида , где y – объясняемая переменная,

x – регрессор

1. Оценить, насколько «хороша» модель по коэффициенту детерминации
2. Оценить, есть ли взаимосвязь между объясняемой переменной и объясняющей

переменной (по значению p-статистики, «количеству звездочек» у регрессора в

модели)

**Практическая часть**

1. Вычислим при помощи встроенных методов языка R (mean, var и sd) значения из задания

Таблица с полученными результатами средних значений, дисперсий и СКО:

Таблица 1

|  |  |  |  |
| --- | --- | --- | --- |
| **Переменная** | **Среднее значение** | **Дисперсия** | **СКО** |
| Education | 10.97872 | 92.45606 | 9.615407 |
| Agriculture | 50.65957 | 515.7994 | 22.71122 |
| Fertility | 70.14255 | 156.0425 | 12.4917 |

1. Построим зависимость вида Education~Agriculture. Будем это делать через встроенную функцию lm библиотеки “lmtest”. Тогда получим:

(1)

Построим зависимость вида Education~Fertility. Тогда:

(2)

1. У зависимости (1) коэффициент детерминации равен 0.3959 (около 40 %), что указывает нам на то, что модель хорошо отображает реальные данные.

У зависимости (2) коэффициент детерминации равен 0.4282 (около 43 %), что тоже показывает хорошую способность предсказывания модели.

1. Количество звездочек у регрессора в зависимости (1) равно 3 (\*\*\*). Это указывает на зависимость между объясняемой переменной (Education) и регрессором (Agriculture).

Количество звездочек у регрессора в зависимости (2) тоже равняется 3 (\*\*\*). Это указывает на зависимость между объясняемой переменной (Education) и регрессором (Fertility).

Код решения задачи и сведения о моделях приведены в [Приложении 1](#_Приложение_1).

**Вывод:**

Между объясняемой переменной и регрессорами есть взаимосвязь. При росте сельскохозяйственного населения или при увеличении рождаемости убывает количество “образованных” людей.

Задание №2

**Введение**

Набор данных: swiss

Объясняемая переменная: Infant.Mortality

Объясняющие переменные (регрессоры): Catholic, Agriculture, Education

**2.1**

Прочитав информацию из набора данных, выполнить задачи:

1. Проверить, что в наборе данных нет линейной зависимости (построить зависимости между переменными и проверить, что в каждой из них невысокий). В случае, если большой, один из таких столбцов можно исключить из рассмотрения.

2. Построить линейную модель зависимой переменной от указанных регрессоров по методу наименьших квадратов (команда lm пакета lmtest в языке R). Оценить, насколько хороша модель, согласно: 1) 2) p-значениям каждого коэффициента.

3. Ввести в модель логарифмы регрессоров (если возможно). Сравнить модели и выбрать наилучшую.

4. Ввести в модель всевозможные произведения пар регрессоров, в том числе квадраты регрессоров. Найти одну или несколько наилучших моделей по доле объяснённого разброса в данных .

**2.2**

Для зависимости, построенной при решении практического задания 2.1, оценить:

1. Доверительные интервалы для всех коэффициентов в модели, p = 95%.

2. Сделайте вывод о отвержении или невозможности отвергнуть статистическую

гипотезу о том, что коэффициент равен 0.

3. Доверительный интервал для одного прогноза (p = 95%, набор значений

регрессоров выбираете сами).

**Практическая часть**

**2.1**

1. Строим все различные модели из пар регрессоров Catholic~Agriculture, Catholic~Education и Agriculture~Education для проверки на линейную независимость. У каждой пары получаем по коэффициенту детерминации :

Таблица 2

|  |  |  |
| --- | --- | --- |
| **Регрессор 1** | **Регрессор 2** |  |
| Catholic | Agriculture | 0.1422 |
| Catholic | Education | 0.001976 |
| Agriculture | Education | 0.3959 |

По полученной таблице становиться ясно, что Agriculture и Education являются зависимыми переменными (>30%).

1. Строим модель вида Infant.Mortality~Catholic + Agriculture + Education. Получаем таблицу значений:

Таблица 3

|  |  |  |
| --- | --- | --- |
| **Регрессор** | **pr** | **Зависимость** |
| Catholic | 0.0954 | “.” |
| Agriculture | 0.0957 | “.” |
| Education | 0.1472 | “ ” |

Коэффициент детерминации .

Исходя из таблицы и из можно сказать, что есть зависимость между Infant.Mortality и регрессорами Catholic и Agriculture, но она минимальна. Между Infant.Mortality и Education нет никакой зависимости.

1. Чтобы улучшить зависимость, попробуем ввести в нашу модель логарифмы от регрессоров:

Таблица 4

|  |  |
| --- | --- |
| **Логарифмическая модель** |  |
| Infant.Mortality~I(log(Catholic)) + Agriculture + Education | 0.05165 |
| Infant.Mortality~I(log(Agriculture)) + Catholic + Education | -0.02406 |
| Infant.Mortality~I(log(Education)) + Catholic + Agriculture | -0.0001944 |
| Infant.Mortality~I(log(Catholic + Agriculture + Education) | -0.004418 |

Нетрудно заметить, что наилучшим коэффициентом детерминации обладает Infant.Mortality~I(log(Catholic)) + Agriculture + Education, которая даже превосходит изначальной модели (0.05165>).

1. По аналогии с пунктом 3 попробуем ввести в нашу модель различные пары произведений регрессоров:

Таблица 5

|  |  |
| --- | --- |
| **Модель из произведений** |  |
| Infant.Mortality~Agriculture^2+Catholic+Education | 0.03409 |
| Infant.Mortality~Agriculture+Catholic^2+Education | 0.03409 |
| Infant.Mortality~Agriculture+Catholic+Education^2 | 0.03409 |
| Infant.Mortality~Agriculture\*Catholic\*Education | 0.1114 |
| Infant.Mortality~Agriculture\*Catholic+Education | 0.1587 |
| Infant.Mortality~Agriculture+Catholic\*Education | 0.04858 |
| Infant.Mortality~Agriculture\*Education+Catholic | 0.02573 |

Из таблицы наилучшими моделями (по сравнению с исходной) можно считать Infant.Mortality~Agriculture\*Catholic\*Education, Infant.Mortality~Agriculture\*Catholic+Education и Infant.Mortality~Agriculture+Catholic\*Education.

**Вывод:**

Не смотря на то, что при помощи произведений и логарифмов от регрессий мы смогли улучшить зависимость кол-ва детской смертности от объясняющих переменных, наша модель все еще остается плохой, так как коэффициент детерминации меньше 30%

**2.2**

* 1. Оценим доверительные интервалы для всех коэффициентов в исходной модели. Для этого вычислим критерий Стьюдента t при 95 % (p = 0.975) попадании в промежуток и следуя кол-ву степеней свободы в нашей модели (k = 43 – 4 = 39). Получим t = 2.022691. Найдем наши доверительные интервалы:

Таблица 6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Переменная** | **Estimate** | **Error** | **Доверительный интервал** | **Может ли коэффициент β = 0?** |
| Intercept | 22.36903 | 1.75389 | [18.82145, 25.91661] | Нет |
| Catholic | 0.01904 | 0.01117 | [-0.00355, 0.04163] | Да |
| Agriculture | -0.04490 | 0.02636 | [-0.09822, 0.00842] | Да |
| Education | -0.08520 | 0.05771 | [-0.96873, -0.73527] | Нет |

1. Найдем доверительный интервал для прогноза Infant.Mortality. Чтобы это сделать, зададим собственные наборы значений регрессоров (Catholic = 30, Agriculture = 20, Education = 20) и воспользуемся встроенным функционалом predict. Получим прогнозное значение fit = 20.33831, нижнее значение lwr = 18.89036 и верхнее upr = 21.78626.

**Вывод:**

Таким образом мы построили доверительный интервал для отдельных коэффициентов в заданной модели и для всей модели в частности

Код решения задачи и сведения о моделях приведены в [Приложении 2](#_Приложение_2).

# **Задание №3**

**Введение**

Номер волны выборки: 14

Файл: r14i\_os26b.csv

Подмножество для пункта 5: Женщины, не замужем; женщины, живущие в городе, разведённые

1. Построить линейную регрессию зарплаты на все параметры, которые были выделены из данных мониторинга. Оценить коэффициент вздутия дисперсии VIF.

2. Поэкспериментировать с функциями вещественных параметров: использовать логарифмы, степени (хотя бы от 0.1 до 2 с шагом 0.1), произведения вещественных регрессоров

3. Выделить наилучшие модели из построенных: по значимости параметров, включённых в зависимости, и по объяснённому с помощью построенных зависимостей разбросу adjusted .

4. Сделать вывод о том, какие индивиды получают наибольшую зарплату.

5. Оценить лучшие модели для подмножества индивидов, указанных в варианте. Сделать вывод о том, какие индивиды получают наибольшую зарплату.

**Практическая часть**

Для начала сортируем и оставляем только те данные, которые были обговорены в файле задания.

1. Строим линейную регрессию. Получаем модель (model1), у которой , что показывает не самую лучшую зависимость между зарплатой и объясняемыми переменными. Коэффициент вздутия дисперсии для каждой переменной:

Таблица 7

|  |  |
| --- | --- |
| **Переменная** | **VIF** |
| sex | 1.086182 |
| wed1 | 2.871808 |
| wed2 | 2.285566 |
| wed3 | 2.237277 |
| highter\_educ | 1.045986 |
| age | 1.294807 |
| city\_status | 1.017970 |
| worktime | 1.032658 |

Видим, что vif-ы меньше 5, что указывает на независимость перечисленных переменных.

* 1. Строим различные модели от логарифмов, степеней и произведений вещественных переменных. Получаем наилучшую модель с наибольшим коэффициентом детерминации , что незначительно, но лучше, чем в исходной модели.

Полученная модель: salary ~ sex + wed1 + wed2 + wed3 + highter\_educ + I(age^2) + city\_status + worktime

Таблица 8

|  |  |  |  |
| --- | --- | --- | --- |
| **Переменная** | **Estimate** | **Error** | **Зависимость** |
| (Intercept) | -6.527e-01 | 8.507e-02 | “\*\*\*” |
| sex | 4.826e-01 | 4.501e-02 | “\*\*\*” |
| wed1 | 1.052e-01 | 7.863e-02 | “ ” |
| wed2 | 2.699e-02 | 8.923e-02 | “ ” |
| wed3 | 3.903e-02 | 1.362e-01 | “ ” |
| highter\_educ | 4.811e-01 | 4.928e-02 | “\*\*\*” |
| I() | -1.363e-01 | 1.766e-02 | “\*\*\*” |
| city\_status | 4.351e-01 | 4.727e-02 | “\*\*\*” |
| worktime | 9.718e-06 | 2.016e-02 | “ ” |

Из полученной таблицы видно, что переменные wed1, wed2, wed3, worktime не влияют на зарплату. Судя по моделям также можно сказать, что на salary сильнее всего сказываются образование (положительная корреляция), пол (положительным образом), город (положительным образом) и менее значительно возраст (отрицательным образом).

1. Оцениваем подмножества индивидов. Получаем, что у женщин, которые не замужем, 0.1111.

У женщин, живущих в городе и которые разведены, 0.09088.

Отсюда можно сказать, что у первой подгруппы зависимость зарплаты и регрессоров выше, чем у второй.

**Вывод**: Полученные в ходе исследований модели улучшили зависимость между зарплатой и регрессорами, но не сильно. В конце концов модель все еще может считаться плохой, так как скорректированный коэффициент детерминации меньше 30 %.

Код решения задачи и сведения о моделях приведены в [Приложении 3](#_Приложение_3).

# **Приложения**

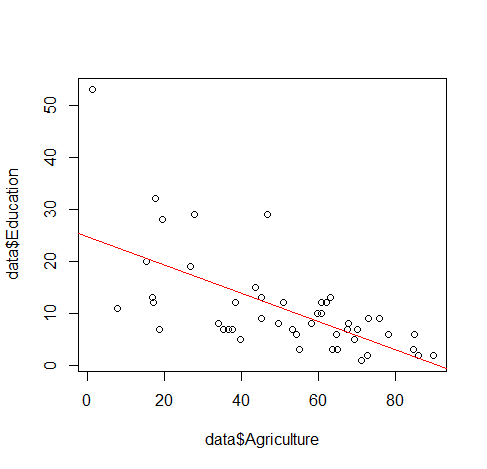
## **Приложение 1**

### Код задачи 1:

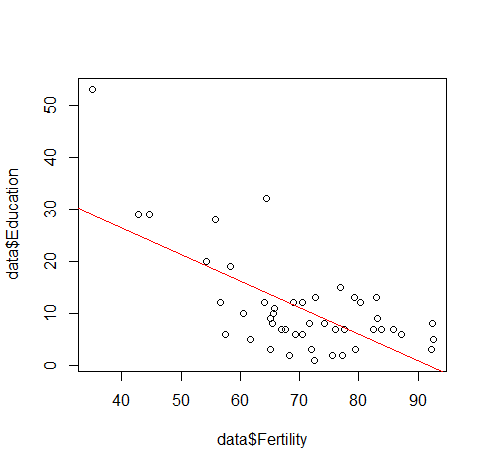
|  |
| --- |
| library("lmtest") #подключаем библиотеку lmtest для работы с линейной регрессией  data = swiss #подключаем набор данных swiss  #Объясняемая переменная: Education  #Объясняющие переменные (регрессоры): Agriculture, Fertility  #//////////////////////////////////////////////  #1) Вычислить средние значения, десперсии и СКО  #вычисляем средние значения Education, Agriculture, Fertility  mean(data$Education) # 10.97872  mean(data$Agriculture) # 50.65957  mean(data$Fertility) # 70.14255  #для иллюстрации наших зависимостей  plot(data$Agriculture, data$Education) + abline(a = 10.97872, b = 0, col = "red")  plot(data$Fertility, data$Education) + abline(a = 10.97872, b = 0, col = "red")  #вычисляем дисперсии Education, Agriculture, Fertility  var(data$Education) # 92.45606 (дисперсия малая, т.к. меньше 100)  var(data$Agriculture) # 515.7994 (большая дисперсия)  var(data$Fertility) # 156.0425 (большая дисперсия)  #вычисляем средние квадраты отклонений Education, Agriculture, Fertility  sd(data$Education) # 9.615407 (приемлемое СКО)  sd(data$Agriculture) # 22.71122 (большое)  sd(data$Fertility) # 12.4917 (большое)  #/////////////////////////////////////////////  #2) Построить зависимости вида y = a + bx (y - объясняемая переменная, x - регрессор)  model1 = lm(Education~Agriculture, data)  summary(model1)  #Education=24.69527-0.27076\*Agriculture (график наклонен вниз)  #R^2=0.3959 (хорошая зависимость)  #"\*\*\*" у регрессора (между переменными есть взаимосвязь)  model2 = lm(Education~Fertility, data)  summary(model2)  #Education=46.81788-0.51095\*Fertility (график наклонен вниз)  #R^2=0.4282 (хорошая зависимость)  #"\*\*\*" у регрессора (между переменными есть взаимосвязь)  #изображаем получившиеся графики наших зависимостей  #для model1  plot(data$Agriculture, data$Education) + abline(a = 24.69527, b = -0.27076, col = "red")  #для model2  plot(data$Fertility, data$Education) + abline(a = 46.81788, b = -0.51095, col = "red") |

### Полученные графики зависимостей моделей:

Зависимость Education~Agriculture:



Зависимость Education~Fertility:



## **Приложение 2**

### Код задачи 2 (2.1 и 2.2):

|  |
| --- |
| library("lmtest") #подключаем библиотеку для выявления линейной зависимости  #Набор данных: swiss  #Объясняемая переменная: Infant.Mortality  #Регрессоры: Catholic, Agriculture, Education  data = swiss#подключаем набор данных swiss  #/////////////////////////////////////////////////////////  #2.1  #1) Проверяем регрессоры на линейную независимость  linmodel1 = lm(Catholic~Agriculture, data)  summary(linmodel1) #R^2 = 0.1422 - нет зависимиости между регрессорами (R^2<30%)  linmodel2 = lm(Catholic~Education, data)  summary(linmodel2) #R^2 = 0.001976 - нет зависимости между регрессорами (R^2<30%)  linmodel3 = lm(Agriculture~Education, data)  summary(linmodel3) #R^2 = 0.3959 - есть зависимость между регрессорами (R^2>30%)  #наибольшие зависимости наблюдаются там, где имеется параметр Agriculture, поэтому попробуем его исключить  modeltest1 = lm(Infant.Mortality~Catholic + Education, data)  summary(modeltest1) #R^2 = -0.007655 (R^2<30%)  #скоректированный коэффициент детерминации получился отрицательным, что указывает на плохую взаимосвязь между  #объясняемой переменной и регрессорами  #попробуем из модели исключить зависимую переменную Education  modeltest2 = lm(Infant.Mortality~Catholic + Agriculture, data)  summary(modeltest2) #R^2 = 0.008206 (R^2<30%)  #взимосвязь между Infant.Mortality и регрессорами улучшилась, но модель все еще считается плохой  #проверим модель со всеми переменными  modeltest3 = lm(Infant.Mortality~Catholic + Agriculture + Education, data)  summary(modeltest3) #R^2 = 0.03409 (R^2<30%)  #в этом случае коэффициент детерминации больше, поэтому оставим в дальнейшем исходную модель без изменений  #//////////////////////////////////////////////////////////////////////  #2) Проверим на зависимость объясняемую переменную от регрессоров  #из модели modeltest3 R^2 = 0.03409 (R^2<30%)  #и есть "." возле Catholic (pr = 0.0954) и Agriculrure (pr = 0.0957), но " " у Education (pr = 0.1472),  #из этого есть зависимость между Infant.Mortality и регрессорами (кроме Education), но она минимальна  #//////////////////////////////////////////////////////////////////  #3) Попробуем ввести в нашу модель логарифмы регрессоров  logmodel1 = lm(Infant.Mortality~I(log(Catholic)) + Agriculture + Education, data)  summary(logmodel1) #R^2 = 0.05165 (R^2<30%)  #значение R^2 лучше предыдущей модели  logmodel2 = lm(Infant.Mortality~I(log(Agriculture)) + Catholic + Education, data)  summary(logmodel2) #R^2 = -0.02406 (R^2<30%)  #R^2<0, плохая модель  logmodel3 = lm(Infant.Mortality~I(log(Education)) + Catholic + Agriculture, data)  summary(logmodel3) #R^2 = -0.0001944 (R^2<30%)  #R^2<0, плохая модель  logmodel4 = lm(Infant.Mortality~I(log(Catholic + Agriculture + Education)), data)  summary(logmodel4) #R^2 = -0.004418 (R^2<30%)  #R^2<0, плохая модель  #из представленных логарифмических моделей наилучшей оказалась logmodel1,  #поэтому мы можем заменить modeltest3 на нее  #/////////////////////////////////////////////////////////////////  #4) Введем в нашу модель всевозможные пары произведений регрессоров  mulmodel1 = lm(Infant.Mortality~Agriculture^2+Catholic+Education, data)  summary(mulmodel1) #R^2 = 0.03409  #нет изменений по сравнению с исходной моделью  mulmodel2 = lm(Infant.Mortality~Agriculture+Catholic^2+Education, data)  summary(mulmodel2) #R^2 = 0.03409  #нет изменений  mulmodel3 = lm(Infant.Mortality~Agriculture+Catholic+Education^2, data)  summary(mulmodel3) #R^2 = 0.03409  #нет изменений  mulmodel4 = lm(Infant.Mortality~Agriculture\*Catholic\*Education, data)  summary(mulmodel4) #R^2 = 0.1114  #R^2 лучше, чем в исходной модели  mulmodel5 = lm(Infant.Mortality~Agriculture\*Catholic+Education, data)  summary(mulmodel5) #R^2 = 0.1587  #R^2 лучше, чем в предыдущей модели  mulmodel6 = lm(Infant.Mortality~Agriculture+Catholic\*Education, data)  summary(mulmodel6) #R^2 = 0.04858  #R^2 лучше, чем в исходной модели, но хуже, чем в предыдущих 2-ух  mulmodel7 = lm(Infant.Mortality~Agriculture\*Education+Catholic, data)  summary(mulmodel7) #R^2 = 0.02573  #R^2 хуже, чем в исходной модели  #следовательно, mulmodel4, mulmodel5, mulmodel6 лучше отображают связь между  #объясняемой переменной и регрессорами, чем modeltest3  #//////////////////////////////////////////////////////////////////  #2.2  #1-2) Оценим доверительные интервалы для всех коэффициентов в модели modeltest3, p = 95%  modeltest3 = lm(Infant.Mortality~Catholic + Agriculture + Education, data)  summary(modeltest3) #R^2 = 0.03409 (R^2<30%)  #Intercept: est = 22.36903, error = 1.75389  #Catholic: est = 0.01904, error = 0.01117  #Agriculture: est = -0.04490, error = 0.02636  #Education: est = -0.08520, error = 0.05771  #p = 0.975 (95%)  #k = 43 - 4 = 39  t = qt(0.975, 39) #критерий Стьюдента  print(t)  #t = 2.022691  #Пределы ошибок при оценке каждой переменной  #Intercept: est = 22.36903+-1.75389\*2.022691 = [18.82145, 25.91661] (0 не попадает в данный интервал, статистическая гипотеза отвергнута)  #Catholic: est = 0.01904+-0.01117\*2.022691 = [-0.00355, 0.04163] (0 попадает в данный интервал, статистическая гипотеза верна)  #Agriculture: est = -0.04490+-0.02636\*2.022691 = [-0.09822, 0.00842] (0 попадает в данный интервал, статистическая гипотеза верна)  #Education: est = -0.08520+-0.05771\*2.022691 = [-0.96873, -0.73527] (0 не попадает в данный интервал, статистическая гипотеза отвергнута)  #////////////////////////////////////////////////////////////////  #3) Найдем доверительный интервал для прогноза Infant.Mortality  new.data = data.frame(Catholic = 30, Agriculture = 20, Education = 20) #набор значений регрессоров выбрали сами  predict(modeltest3, new.data, interval = "confidence")  #fit = 20.33831 (прогноз)  #lwr = 18.89036 (нижнее значение)  #upr = 21.78626 (верхнее значение) |

## **Приложение 3**

|  |
| --- |
| #install.packages("devtools")  #devtools::install\_github("bdemeshev/rlms")  #library("rlms")  library("lmtest")  library("data.table")  library("car")  #Считываемые данные с таблицы  #зарплата, пол, семейное положение, наличие высшего образования,  #возраст, тип населенного пункта, длительность рабочей недели  data <- read.csv("C:\\Users\\trudo\\YandexDisk\\NIR\\r14i\_os26b.csv")  sex = ifelse(data$jh5 == 2, 0, 1) #пол (44)  wed1 <- ifelse(data$j\_marst == 2, 1, 0)  wed2 <- ifelse(data$j\_marst == 4 | data$j\_marst == 5, 1, 0)  wed3 <- ifelse(data$j\_marst == 1, 1, 0)  familystatus = data.frame(wed1, wed2, wed3) #семейное положение (34)  education = data$j\_diplom #наличие высшего образования (38)  highter\_educ <- ifelse(education == 6, 1, 0)  locality = data$status #тип населенного пункта (28)  city\_status <- ifelse(locality == 1 | locality == 2, 1, 0)  salary = data$jj13.2 #зарплата (83 столбец)  age = data$j\_age #возраст (47)  worktime = data$jj6.1a #длительность рабочей недели (72)  ourdata1 <- data.frame(salary,  sex,  familystatus,  highter\_educ,  age,  city\_status,  worktime)  #удаляем строки, где есть NA или данные числа  ourdata1 <- na.omit(ourdata)  ourdata1 <- subset(ourdata, salary != 99999997 & salary != 99999998 & salary != 99999999 &  worktime != 99999996 & worktime != 99999997 & worktime != 99999998 & worktime != 99999999)  ourdata = ourdata1  ourdata$salary <- (ourdata$salary - mean(ourdata$salary))/sqrt(var(ourdata$salary))  ourdata$age = (ourdata$age - mean(ourdata$age))/sqrt(var(ourdata$age))  ourdata$worktime = (ourdata$worktime - mean(ourdata$worktime))/sqrt(var(ourdata$worktime))  print(str(ourdata))  model1 = lm(salary~sex+wed1+wed2+wed3+highter\_educ+age+city\_status+worktime, data=ourdata)#R^2 = 0.1465  summary(model1)  model1R2 = summary(model1)$adj.r.squared  #summary(model1)$adj.r.squared  vif(model1) #vif хороший(<5), переменные можно изпользовать вместе  #логарифмы  logmodel1 = lm(salary~sex+wed1+wed2+wed3+highter\_educ+I(log(abs(age)))+city\_status+worktime, data=ourdata) #R^2 = 0.1465  summary(logmodel1)  logmodel2 = lm(salary~log(abs(sex+wed1+wed2+wed3+highter\_educ+age+city\_status+worktime)), data=ourdata) #R^2 = 0.03141  summary(logmodel2)  #логарифмы не особо помогают улучшить зависимость между переменными  degbettermodel = model1  degree = 0  R2 = model1R2  **for**(i **in** 1:20){  n <- 0.1\*i  degree = n  modeltime <- lm(salary~I(sex^n)+wed1+wed2+wed3+highter\_educ+age+city\_status+worktime, data=ourdata)  R2time <- summary(modeltime)$adj.r.squared  **if**(R2time > R2){  degbettermodel <- modeltime  R2 <- R2time  }    modeltime <- lm(salary~sex+wed1+wed2+I(wed3^n)+highter\_educ+age+city\_status+worktime, data=ourdata)  R2time <- summary(modeltime)$adj.r.squared  **if**(R2time > R2){  degbettermodel <- modeltime  R2 <- R2time  }    modeltime <- lm(salary~sex+wed1+wed2+wed3+I(highter\_educ^n)+age+city\_status+worktime, data=ourdata)  R2time <- summary(modeltime)$adj.r.squared  **if**(R2time > R2){  degbettermodel <- modeltime  R2 <- R2time  }    modeltime <- lm(salary~sex+wed1+wed2+wed3+highter\_educ+I(age^n)+city\_status+worktime, data=ourdata)  R2time <- summary(modeltime)$adj.r.squared  **if**(R2time > R2){  degbettermodel <- modeltime  R2 <- R2time  }    modeltime <- lm(salary~sex+wed1+wed2+wed3+highter\_educ+age+I(city\_status^n)+worktime, data=ourdata)  R2time <- summary(modeltime)$adj.r.squared  **if**(R2time > R2){  degbettermodel <- modeltime  R2 <- R2time  }  }  print(degree)  summary(degbettermodel)  vif(degbettermodel)  #наилучшая модель lm(formula = salary ~ sex + wed1 + wed2 + wed3 + highter\_educ + I(age^2) + city\_status + worktime, data = ourdata)  #она имеет скоректированный коэффициент детерминации R^2 = 0.1815, что незначительно, но лучше исходной модели model1  #произведения  mulmodel1 = lm(salary~wed3+highter\_educ+age+I(city\_status\*sex)+wed1+wed2+worktime, data=ourdata)#R^2 = 0.1375  summary(mulmodel1)  mulmodel2 = lm(salary~wed3+highter\_educ+I(age\*city\_status)+sex+wed1+wed2+worktime, data=ourdata)#R^2 = 0.1127  summary(mulmodel2)  mulmodel3 = lm(salary~wed3+highter\_educ+I(age\*city\_status\*sex)+wed1+wed2+worktime, data=ourdata)#R^2 = 0.05111  summary(mulmodel3)  mulmodel4 = lm(salary~wed3+I(highter\_educ\*age)+city\_status+sex+wed1+wed2+worktime, data=ourdata)#R^2 = 0.1066  summary(mulmodel4)  mulmodel5 = lm(salary~wed3+I(highter\_educ\*age)+I(city\_status\*sex)+wed1+wed2+worktime, data=ourdata)#R^2 = 0.09523  summary(mulmodel5)  mulmodel6 = lm(salary~wed3+I(highter\_educ\*age\*city\_status)+sex+wed1+wed2+worktime, data=ourdata)#R^2 = 0.06295  summary(mulmodel6)  mulmodel7 = lm(salary~wed3+I(highter\_educ\*age\*city\_status\*sex)+wed1+wed2+worktime, data=ourdata)#R^2 = 0.008498  summary(mulmodel7)  mulmodel8 = lm(salary~I(wed3\*highter\_educ)+age+city\_status+sex+wed1+wed2+worktime, data=ourdata)#R^2 = 0.1103  summary(mulmodel8)  mulmodel9 = lm(salary~I(wed3\*highter\_educ)+age+I(city\_status\*sex)+wed1+wed2+worktime, data=ourdata)#R^2 = 0.09931  summary(mulmodel9)  mulmodel10 = lm(salary~I(wed3\*highter\_educ)+I(age\*city\_status)+sex+wed1+wed2+worktime, data=ourdata)#R^2 = 0.06791  summary(mulmodel10)  mulmodel11 = lm(salary~I(wed3\*highter\_educ)+I(age\*city\_status\*sex)+wed1+wed2+worktime, data=ourdata)#R^2 = 0.01325  summary(mulmodel11)  mulmodel12 = lm(salary~I(wed3\*highter\_educ\*age)+city\_status+sex+wed1+wed2+worktime, data=ourdata)#R^2 = 0.1046  summary(mulmodel12)  mulmodel13 = lm(salary~I(wed3\*highter\_educ\*age)+I(city\_status\*sex)+wed1+wed2+worktime, data=ourdata)#R^2 = 0.09315  summary(mulmodel13)  mulmodel14 = lm(salary~I(wed3\*highter\_educ\*age\*city\_status)+sex+wed1+wed2+worktime, data=ourdata)#R^2 = 0.06166  summary(mulmodel14)  mulmodel15 = lm(salary~I(wed3\*highter\_educ\*age\*city\_status\*sex)+wed1+wed2+worktime, data=ourdata)#R^2 = 0.008929  summary(mulmodel15)  #по полученным R^2 ясно, что ни одна из моделей не является лучше исходной  #исходя из вышеперечисленных моделей, наилучшей можно считать модель вида lm(formula = salary ~ sex + wed1 + wed2 + wed3 + highter\_educ + I(age^2) + city\_status + worktime, data = ourdata)  #она обладает скоректированным коэффициентом детерминации большим, чем в исходной модели (0.1815 > 0.1465)  #если же смотреть на кол-во зведочек возле регрессоров, то у полученной модели подле wed1 они отсутствуют,  #в остальном исходная и полученная модели схожи  #Судя по моделям также можно сказать, что на зарплату сильнее всего влияют образование, пол, город и менее значительно возраст  #есть также хоть и не такая значительная кореляция с тем, что человек никогда не состоял в браке  #остальные переменные, как оказалось, не особо коррелируют с уровнем зарплаты  #5)  newdata <- ourdata  newdata <- subset(newdata, sex == 0) #оставляем из выборки только женщин  newdata$salary <- (newdata$salary - mean(newdata$salary))/sqrt(var(newdata$salary))  newdata$age = (newdata$age - mean(newdata$age))/sqrt(var(newdata$age))  newdata$worktime = (newdata$worktime - mean(newdata$worktime))/sqrt(var(newdata$worktime))  singlewoman <- newdata  singlewoman <- subset(newdata, wed1 == 0) #женщины, ни разу не бывшие в браке  singlewomanmodel = lm(salary ~ highter\_educ + I(age^2) + city\_status + worktime, data = singlewoman)#R^2 = 0.1211  summary(singlewomanmodel) #R^2 = 0.1111  divwoman <- newdata #женщины, разведенные, живущие в городе  divwoman <- subset(newdata, wed2 == 0 & city\_status == 1)  divwomanmodel = lm(salary ~ highter\_educ + I(age^2) + worktime, data = divwoman)#R^2 = 0.1211  summary(divwomanmodel) #R^2 = 0.09088  #у singlewoman корреляция у R^2 выше, чем у divwoman (0.1111 > 0.09088)  #соответственно, зарплата выше у женщин, не бывавших в браке |