Отчет по Научно исследовательской работе.

Выполнил: Масленников Иван Андреевич

Студент РТУ МИРЭА института «ИИИ» группы КМБО-06-22 под руководством Петрусевича Дениса Андреевича.

Оглавление

[Задание №1 3](#_Toc134526360)

[Условие 3](#_Toc134526361)

[Решение 4](#_Toc134526362)

[Вывод 5](#_Toc134526363)

[Задание №2.1 7](#_Toc134526364)

[Условие 7](#_Toc134526365)

[Решение 8](#_Toc134526366)

[Вывод 9](#_Toc134526367)

[Задание №2.2 10](#_Toc134526368)

[Условие 10](#_Toc134526369)

[Решение 11](#_Toc134526370)

[Вывод 13](#_Toc134526371)

[Задание №3 14](#_Toc134526372)

[Условие 14](#_Toc134526373)

[Решение 15](#_Toc134526374)

[Вывод 20](#_Toc134526375)

[Список литературы 21](#_Toc134526376)

[Приложения 22](#_Toc134526377)

# Задание №1

Необходимо загрузить данные из указанного набора и произвести следующие действия.

Набор данных: *Swiss*.

Объясняемая переменная: *Education*.

Регрессоры: *Fertility, Examination.*

## Условие

1. Оцените среднее значение, дисперсию и СКО переменных, указанных во втором и

третьем столбце.

2. Постройте зависимости вида у = a + bx, где у - объясняемая переменная, х -

регрессор (для каждого варианта по две зависимости).

3. Оцените, насколько «хороша» модель по коэффициенту детерминации R2?

4. Оцените, есть ли взаимосвязь между объясняемой переменной и объясняющей

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

модели).

## Решение

Задание №1

Оценка среднего значения, дисперсии и СКО переменных использовалась функция mean(), которая выводит среднее значение по передаваемому ей списку элементов, а так же найдено СКО с помощью формулы: sqrt(sum((A - mean(A))^2/(length(A)-1))). Кроме того получена дисперсия благодаря функции var(). После выполнения анализа получим данные значения в таблице 1:

Таблица 1. Характеристики параметров *Examination, Agriculture, Fertility в наборе Swiss*

|  |  |  |  |
| --- | --- | --- | --- |
| Переменные | Среднее значение | Дисперсия | СКО |
| *Examination* | 16.48936 | 7.977883 | 63.64662 |
| *Agriculture* | 50.65957 | 22.71122 | 515.7994 |
| *Fertility* | 70.14255 | 12.4917 | 156.0425 |

Задание №2

Построение линейной зависимости используем команду lm():

lm(*Examination~Fertility+Agriculture, data* ), в результате получаем общую характеристику для модели в таблице 2.1 ,а так же побочные для зависмостей *Examination~Fertility* в таблице 2.2 и *Examination~Agriculture* в таблице 2.3:

Таблица 2.1. Общая характеристика модели зависимости *Examination* от регрессоров *Agriculture* и *Fertility*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Параметр\ Характеристики** | **Значение** | **Std. Error** | **t value** | **Pr(>|t|)** | **Уровень значимости** |
| *(Intercept)* | 46.45927 | 4.01858 | 11.561 | 6.33e-15 | \*\*\* |
| *Fertility* | -0.29438 | 0.06024 | -4.887 | 1.40e-05 | \*\*\* |
| *Agriculture* | -0.18400 | 0.03313 | -5.553 | 1.52e-06 | \*\*\* |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Таблица 2.2. Характеристика зависимости *Examination* от регрессора *Fertility.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Параметр\ Характеристики** | **Значение** | **Std. Error** | **t value** | **Pr(>|t|)** | **Уровень значимости** |
| *(Intercept)* | 45.42289 | 5.17670 | 8.774 | 2.65e-11 | \*\*\* |
| *Fertility* | -0.41250 | 0.07268 | -5.675 | 9.45e-07 | \*\*\* |

Имеем зависимость cat(*data$Examination\_Fertility*<-45.42 + (-0.412)\**data$Fertility*) из таблицы 2.2.

Таблица 2.3. Характеристика для зависимости *Examination* от регрессора *Agriculture.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Параметр\ Характеристики** | **Значение** | **Std. Error** | **t value** | **Pr(>|t|)** | **Уровень значимости** |
| *(Intercept)* | 28.70668 | 2.11001 | 13.605 | < 2e-16 | \*\*\* |
| *Agriculture* | -0.24117 | 0.03807 | -6.334 | 9.95e-08 | \*\*\* |
|  |  |  |  |  |  |

Имеем зависимость cat(*data$Examination\_Agriculter*<-28.70 + (-0.241)\**data$Agriculture*) из таблицы 2.3.

Из взаимосвязи вида y= a + bx можно сделать вывод, с ростом коэффициента *Agriculture* и *Fertility* уменьшается % призывников, получивших наивысшую оценку на армейском экзамене, следовательно взаимосвязи являются отрицательными.

Задание №3

Рассмотрение того на сколько «хороша» модель по R2 нужно использовать команду *summary(lm(Examination~Fertility+Agriculture, data))*. Из этого следует, что 64% наших данных описывается нашей моделью. Можем сделать вывод (*R-squared* = 64%), что модель относительно хороша: для такой зависимости (только две объясняемых переменных) коэффициент не мал, но для полного описания процесса нужно добавлять другие параметры. Также при рассмотрении зависимостей *Examination~Fertility*, R2 равен 40% и *Examination~Agriculture,* R2 равен 45%, что говорит о неплохом качестве моделей.

Задание №4

Оценка взаимосвязи между объясняемой переменной и объясняющей переменной, воспользуемся функцией summary.

При рассмотрении таблицы 2.1 значений *p* для переменной-предиктора *Fertility* равно 1.40e-05(\*\*\*). Поскольку это значение меньше 0,05, оно имеет статистически значимую связь с переменной отклика в модели.

Значение *p* для переменной-предиктора *Agriculture* равно 1.52e-06(\*\*\*). Поскольку это значение меньше 0,05, оно имеет статистически значимую связь с переменной отклика в модели.

## Вывод

Анализ проведен на данных пакета *swiss*, который содержит информацию об уровне образования, занятости в сельском хозяйстве, рождаемости и других факторах в 47 регионах Швейцарии. В результате анализа были получены следующие выводы:

* Были построены графики зависимости показателя *Examination от Agriculture* и Fertility с помощью графической библиотеки ggplot2.
* Были построены модели зависимости *Examination от Agriculture и Fertility*. При этом коэффициент детерминации для модели с двумя объясняющими переменными составил 64%, что говорит о хорошей адаптации модели.
* Были проведены статистические тесты, которые показали, что обе переменные, *Agriculture и Fertility*, имеют статистически значимую связь с переменной отклика *Examination*.
* Из взаимосвязи вида y = a + bx был сделан вывод, что при увеличении значений *Agriculture и Fertility* показатель *Examination* уменьшается, что говорит о наличии отрицательной корреляции между этими переменными.

Таким образом, можно сделать вывод, что в данных *swiss* существует статистически значимая зависимость между показателем *Examination* и переменными *Agriculture* и *Fertility*. При этом увеличение значений *Agriculture и Fertility* сопровождается снижением показателя *Examination*. Однако для более точного предсказания значения *Examination* необходимо включать в модель другие переменные.

Код решения задачи и сведения о проверенных моделях приведены в Приложении 1

# Задание №2.1

Набор данных: *Seatbelts*

Объясняемая переменная: *front*

Регрессоры: *law, kms, PetrolPrice*

## Условие

1. Проверьте, что в наборе данных нет линейной зависимости (построить зависимости между переменными, указанными в варианте, и проверить, что *R2* в каждой из них невысокий). В случае, если *R2* большой, один из таких столбцов можно исключить из рассмотрения.

2. Постройте линейную модель зависимой переменной от указанных в варианте регрессоров по методу наименьших квадратов (команда *lm* пакета *lmtest* в языке R). Оценить, насколько хороша модель, согласно: 1) *R2*, 2) p-значениям каждого коэффициента.

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

4. Введите в модель всевозможные произведения пар регрессоров, в том числе квадраты регрессоров. Найдите одну или несколько наилучших моделей по доле объяснённого разброса в данных *R2*.

## Решение

Задание №1

Проверка наличия в наборе данных линейной зависимости. В случае, если R2 большой, один из таких столбцов можно исключить из рассмотрения.

В результате построения зависимостей имеем, что :

* *model1*=(lm(*law*~*kms* + *PetrolPrice*, *data*))R^2 = 28% относительно независимые регрессоры
* *model1*=(lm(*kms*~*law* + *PetrolPrice*, *data*))R^2 = 27% относительно независимые регрессоры
* *model1*=(lm(*PetrolPrice*~*law* + *kms*, *data*))R^2 = 19% относительно независимые регрессоры

Из списка делаем вывод, что все регрессоры относительно независимы и их необязательно исключать из дальнейшего рассмотрения.

Задание№2

Построение модели, используя команду lm(). Рассмотрим *R2* и p-статистику, с помощью summary().

* *model2 = lm(front~law + kms + PetrolPrice, data)*
* summary(*model2*)

Модель неплохая, но есть невзаимосвязанная переменная kms (с отсутствием звездочек). *Adjusted R-squared* = 42%. По результатам зависимости *front*~law и *front*~*PetrolPrice* имеет три звезды.

Получим коэф. R^2 для каждого отдельно:

* summary(lm(*front~law, data*))$r.squared, *R^2* равен 31%.
* summary(lm(front~kms, data))$r.squared, *R^2* равен 12%.
* summary(lm(front~PetrolPrice, data))$r.squared, *R^2* равен 29%.

В результате пробуем исключить *kms* как наименее значимый параметр:

* summary(lm(*front~law + PetrolPrice, data*)) *Adjuster R-squared* = 43%.

Коэф. увеличился на 1% => исключение параметра не стало критичным. Следовательно, далее пользуемся моделью *front~law + PetrolPrice*, так как она имеет неплохой *R2*, и хорошие показатели p-статистики.

Задание№3

Введение в модель логарифмов регрессоров, не забывая про проверку vif(). Сравним модели и выберем наилучшую. Хочу заметить, что модель нельзя ввести log регрессоров без замены значений равных 0 в таблице *law* на 0.001(*data$law*[*data$law* == 0] <- 0.001).

Лучшей моделью в результате сравнений среди логарифмов оказалась:

* *model3\_1 = (lm(log(front)~ I(log(law)) + I(log(PetrolPrice)), data))*
* *model 3\_1* имеет коэффициент *R2 =* 0.4835, а так же неплохие значения vif.

Задание №4

Введение в модель всевозможных произведений пар регрессоров, в том числе квадратов регрессоров, не забывая о проверке значений vif() и найдем лучшую модель основываясь на доле объяснённого разброса в данных *R2*.

Лучшей моделью среди произведений пар регрессоров оказалась:

*model4\_1 = lm(front ~ law + I(PetrolPrice^2), data)*

*model4* имеет коэффициент R2 = 0.4306, а так же хорошие значения vif.

## Вывод

Из проведенного анализа данных следует, что зависимость между параметрами *"front", "law", "kms" и "PetrolPrice"* является сложной и линейной. Попытки использования всех регрессоров в одной модели показали, что они относительно независимы, что было подтверждено значениями VIF, но коэффициент детерминации (*R-squared*) был низким для каждого регрессора по отдельности.

Наилучшая модель получилась, когда были использованы только регрессоры *"law"* и *"PetrolPrice".* Коэффициент детерминации (*Adjusted R-squared*) улучшился, и статистические характеристики также улучшились.

Попытка ввести логарифмическое преобразование параметра *"front"* показала некоторое улучшение модели, но в целом не дало заметного прироста качества.

Попытки ввести все возможные произведения пар регрессоров не привели к улучшению модели, а лишь ухудшили ее показатели, поэтому было принято решение использовать только регрессоры *"law" и "PetrolPrice".*

Код решения задачи и сведения о проверенных моделях приведены в Приложении 2.1

# Задание №2.2

Набор данных: *Seatbelts*

Объясняемая переменная: *front*

Регрессоры: *law, kms, PetrolPrice*

## Условие

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).

# Решение

Задание№1

Доверительные интервалы для всех коэффициентов в модели*(model*), p = 95%.

*model = lm(front~law + kms + PetrolPrice, data)*

confint(*model*, level = 0.95)

И получим интервалы в таблице 1:

Таблица 1. Доверительные интервалы для общей модели

|  |  |  |
| --- | --- | --- |
|  | 2.5% | 97.5% |
| *(Intercept)* | 1.246773e+03 | 1.618217e+03 |
| *law* | -2.894007e+02 | 1.517127e+02 |
| *kms* | -8.353530e-03 | 6.856602e-03 |
| *PetrolPrice* | -7.118528e+03 | -3.644096e+03 |

*Для model3\_1:*

Доверительный интервал для прогноза лучшей модели(*model3\_1*):

Для подсчета вручную требуется найти число степеней свободы в модели: 189 - 3 = 186. После найдем для такого числа степеней свободы и p = 95% значение t-критерия Стьюдента: t = 1.97, с помощью cat(t\_critical = qt(0.975, df = 186)).

Тогда для каждого из коэффициентов будут справедливы уравнения:

Таблица 2. Доверительные интервалы для *model3\_1*

|  |  |  |
| --- | --- | --- |
|  | 2.5% | 97.5% |
| *law\_con* | -0.05899181 | -0.03742819 |
| *PetrolPRice\_con* | -0.821111 | -0.415815 |

*Для model4\_1:*

Доверительный интервал для прогноза лучшей модели(*model4\_1*):

Для подсчета вручную требуется найти число степеней свободы в модели: 189 - 3 = 186. После найдем для такого числа степеней свободы и p = 95% значение t-критерия Стьюдента: t = 1.97, с помощью cat(t\_critical = qt(0.975, df = 186)).

Тогда для каждого из коэффициентов будут справедливы уравнения:

Таблица 3. Доверительные интервалы для *model4\_1*

|  |  |  |
| --- | --- | --- |
|  | 2.5% | 97.5% |
| *law\_con* | -285.3494 | -159.1906 |
| *PetrolPRice\_con* | -34529.65 | -18175.49 |

Задание№2

Выводы о отвержении или невозможности отвергнуть статистическую гипотезу о том, что коэффициент равен 0 для параметров из таблицы 1:

*Для model:*

* Для *Intercept*, *B* не может равняться 0. Следовательно отвергаем гипотезу о том, что этот коэффициент может быть равен 0, на уровне значимости 5%.
* Для *law*, *B* не может равняться 0. Следовательно отвергаем гипотезу о том, что этот коэффициент может быть равен 0, на уровне значимости 5%.
* Для kms, *B* может равняться 0. Следовательно подтверждаем гипотезу о том, что этот коэффициент может быть равен 0, на уровне значимости 5%.
* Для *PetrolPrice*, B не может равняться 0. Следовательно отвергаем гипотезу о том, что этот коэффициент может быть равен 0, на уровне значимости 5%.

Выводы о отвержении или невозможности отвергнуть статистическую гипотезу о том, что коэффициент равен 0 для параметров из таблицы 2 и таблицы 3:

*Для model3\_1:*

* Для *law\_con*, B не может равняться 0. Следовательно отвергаем гипотезу о том, что этот коэффициент может быть равен 0, на уровне значимости 5%
* Для *PetrolPrice\_con*, B не может равняться 0. Следовательно отвергаем гипотезу о том, что этот коэффициент может быть равен 0, на уровне значимости 5%.

*Для model4\_1:*

* Для *law\_con*, B не может равняться 0. Следовательно отвергаем гипотезу о том, что этот коэффициент может быть равен 0, на уровне значимости 5%
* Для *PetrolPrice\_con*, B не может равняться 0. Следовательно отвергаем гипотезу о том, что этот коэффициент может быть равен 0, на уровне значимости 5%.

Задание№3

Анализ оценки доверительного интервала для *model3\_1* и *model4\_1* прогноза

(p = 95%, с набором значений: *law* = 20, *PetrolPrice* = 10), используя функцию predict:

*Таблица 4 для model3\_1:*

|  |  |  |
| --- | --- | --- |
| **fit** | **lwr** | **upr** |
| 3.439129 | 2.543131 | 4.335128 |

Получим прогноз для *model3\_1* в котором:

* прогноз модели = 3.439129
* нижняя граница доверительного интервала = 2.543131
* верхняя граница доверительного интервала = 4.335128

*Таблица 5 для ля model4\_1:*

|  |  |  |
| --- | --- | --- |
| **fit** | **lwr** | **upr** |
| -2638552 | -3456750 | -1820354 |

Получим прогноз для *model4\_1* в котором:

* прогноз модели = -2638552
* нижняя граница доверительного интервала = -3456750
* верхняя граница доверительного интервала = -1820354

## Вывод

В результате данной работы была построена общая модель регрессии, описывающая зависимость переменной *"front" от "law", "kms" и "PetrolPrice"*. Были рассчитаны доверительные интервалы для всех коэффициентов в модели с уровнем доверия 95%. Также была проверена гипотеза о том, может ли коэффициент равняться 0, на уровне значимости 5%. В результате проверки гипотезы оказалось, что коэффициенты *"Intercept",* *"law" и "PetrolPrice"* не могут равняться 0, а коэффициент *"kms"* может равняться 0.

Далее была построена лучшая модель с использованием переменных *"law" и "PetrolPrice"* и были рассчитаны доверительные интервалы для коэффициентов этой модели. На основании результатов рассчетов было выявлено, что коэффициенты *"law" и "PetrolPrice"* не могут равняться 0 на уровне значимости 5%.

Был также рассчитан доверительный интервал для первого прогноза на основе лучшей модели. Предсказанный прогноз составил 3.439129, а нижняя и верхняя границы доверительного интервала равны соответственно 2.543131 и 4.335128.

Таким образом, на основании данной работы можно сделать вывод о том, что переменные *"law" и "PetrolPrice"* значимо влияют на переменную *"front"*, а переменная *"kms"* не имеет статистически значимого влияния на *"front".* Кроме того, можно использовать лучшую модель для прогнозирования значения переменной *"front"* на основе значений *"law"* и *"PetrolPrice"*, при этом уровень доверия составляет 95%.

Код решения задачи и сведения о проверенных моделях приведены в Приложении 2.2

# Задание №3

Набор данных: r12i\_os26c.scv

Объясняемая переменная: *salary*

Регрессоры:

1) переменная *wed1* имеет значение 1 в случае, если респондент женат, 0 – в противном случае

2) *wed2*=1, если респондент разведён или вдовец

3) *wed3* = 1, если респондент никогда не состоял в браке

4) переменная *sex* – пол

5) переменная *age* – возраст

6) переменная *higher\_educ -* наличие высшего образования

7) переменная *city\_status – место проживания*

*8)* переменная *working\_hours –* среднее число рабочих часов в неделю

## Условие

1. Постройте линейную регрессию зарплаты на все параметры, которые вы выделили из данных мониторинга. Не забудьте оценить коэффициент вздутия дисперсии VIF.

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

3. Выделите наилучшие модели из построенных: по значимости параметров, включённых в зависимости, и по объяснённому с помощью построенных зависимостей разбросу *adjusted R2 - R2 adj.*

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

5. Оцените лучшие модели для подмножества индивидов, указанных в варианте(Не состоявшие в браке мужчины, с высшим образованием; городские жители, состоящие в браке, женщины). Сделайте вывод о том, какие индивиды получают наибольшую зарплату.

## Решение

Необходимо считать данные исследования из файла и убрать значения NA, преобразовав их в удобном виде:

1. Переменная *sex* принимает значение 1 для мужского пола, 0 – для женского
2. *age* ­– переменная с нормализованным возрастом (формула нормализации значения: data2["*age*"] = (*age* - mean(*age*)) / sqrt(var(*age*))
3. Семейное положение:
   1. *wed1* = 1, если человек состоит в зарегистрированном браке, иначе 0
   2. *wed2* = 1, если человек разведён или вдовец, иначе 0
   3. *wed3* = 1, если человек никогда не был в браке, иначе 0
   4. Проверим, что между *wed1, wed2, wed3* нет линейной зависимости
4. *higher\_edu*c = 1, если у человека есть высшее образование, иначе 0
5. *city\_status* = 1, если человек живёт в городе, иначе 0
6. *working\_hours* – переменная с нормализованным числом рабочих часов в неделю (формула для нормализации значения: (*working\_hours* - mean(*working\_hours*)) / sqrt(var(*working\_hours*)))
7. *salary* – переменная с нормализованной зарплатой (формула для нормализации значения: ((*salary* - mean(*salary*)) / sqrt(var(*salary*)))

Задание №1

Линейная регрессия зарплаты на все параметры и оценим vif:

*model1 = lm(data = data2, salary ~ sex + age + wed1 + wed2 + wed3 + higher\_educ + city\_status + working\_hours*

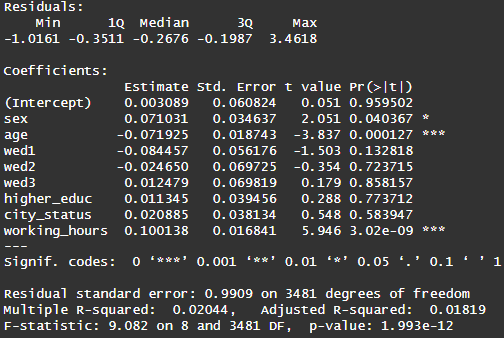


Рисунок №1. Характеристики model1.

Оценим vif:

Рисунок №1.1. vif(model1)

Из рисунка №1 и №1.1 видим, что переменные *wed1* имеет плохую p-статистику и vif. Уберём ее и посмотрим, как изменится *R2*:

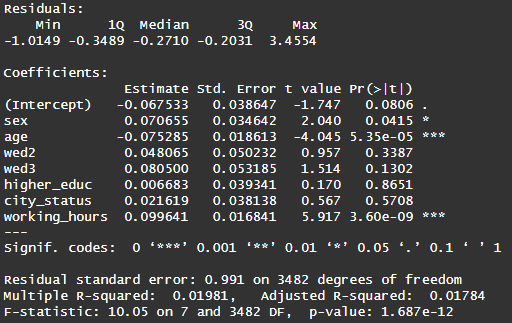


Рисунок №2. Результат работы *model1* с исключенным регрессором *wed1.*

Оценим vif:

Рисунок №2.1. Результат vif(*model1*) без регрессора *wed1*.

Из рисунка №2 и №2.1 видим, что переменная *higher\_educ* имеет плохую p-статистику. Уберём ее и посмотрим, как изменится *R2*:

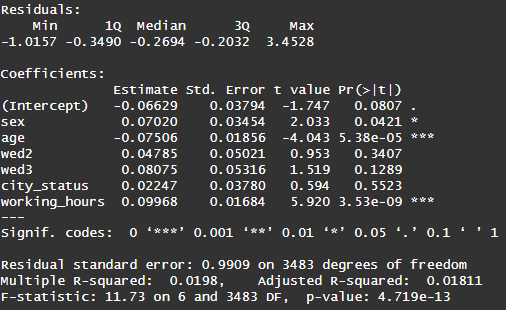


Рисунок №3. Результат работы *model1* с исключенным регрессором *wed1* и *higher\_educ.*

Оценим vif:

Рисунок №3.1. Результат vif(*model1*) без регрессора *wed1 и higher\_educ.*

Из рисунка №3 и №3.1 видим, что p-статистика чуть улучшилась и параметры vif стабилизировались.

Задание №2

Эксперимент с функциями вещественных параметров, используя логарифмы (логарифмы и степени имеет смысл вводить только для параметров *age* и *working\_hours*, так как остальные принимают только значения 0 или 1.) и степени (хотя бы от 0.1 до 2 с шагом 0.1).

Модель с log:

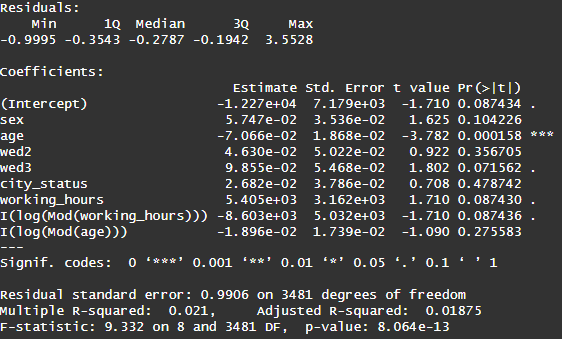
*model1 = lm( salary ~ sex + age + wed2 + wed3 + city\_status + working\_hours + I(log(Mod(working\_hours))) + I(log(Mod(age))), data=data2)*

Рисунок №4. Результат работы *model1* c I*(log(Mod(working\_hours))) + I(log(Mod(age))).*

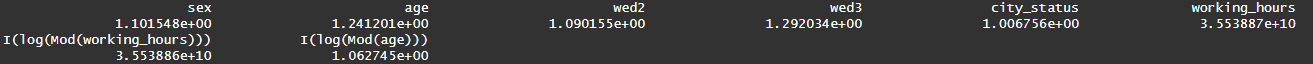
Оценим vif:

Рисунок №4.1. Результат работы vif(*model1)* c I*(log(Mod(working\_hours))) + I(log(Mod(age))).*

Из рисунка №4 и №4.1 видим, что p-статистика плохая и параметры vif ухудшились для *working\_hours* и *I(log(Mod(working\_hours))),* исключаю *working\_hours.*

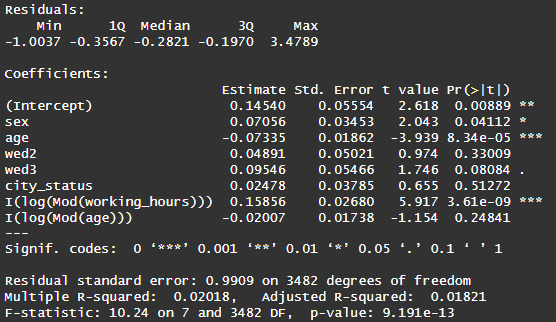
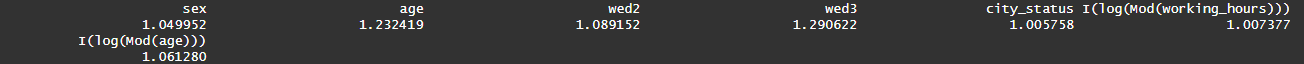


Рисунок №5. Результат работы *model1* с исключенным параметром *working\_hours.*

Оценим vif:

Из рисунка №5 и №5.1 видим, что p-статистика чуть улучшилась и параметры vif стабилизировались. Дальнейшее исключение регрессоров не приведет к значительным результатам, следовательно дальше можно работать с набором регрессоров:

(*sex + age + wed2 + wed3 + city\_status + I(log(Mod(working\_hours)))+ I(log(Mod(age))))* ).

Добавим модели со степенями для набора регрессоров с шагом 0.1 от 0.1 до 2.0:

current\_pow=0.1

*model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))*

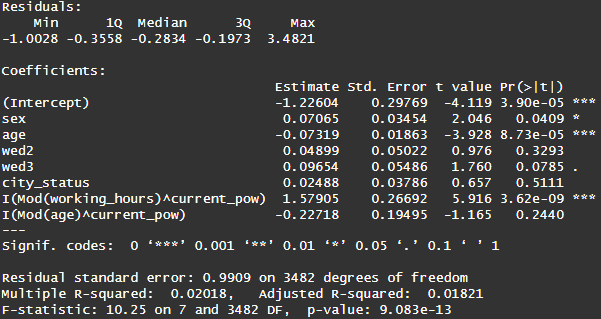


Рисунок №6. Результат работы *model1* в степени 0.1.

Проверим vif:



Рисунок 6.1. Результат работы vif(*model1*) в степени 0.1.

Из рисунка №6 и №6.1 видим, что p-статистика плохая, параметры vif хорошие, дальнейшую работу проводим с имеющейся моделью с шагом степени 0.1, так как исключение параметров не даст значимых изменений*.* Подмечу что дальнейшее увеличение степени приведут к ухудшению p-статистики и постепенному уменьшению *R2*, а так же мало колебанию значений vif, что не особо скажется на точности моделей.

Задание №3

Наилучшие модели из построенных: по значимости параметров, включённых в зависимости, и по объяснённому с помощью построенных зависимостей разбросу *adjusted R2 – R2\_adj.*

Наилучшими по значению R2 из всех моделей являются модели для степеней 0.1, 0.2, 0.3. Разброс R2 - R2\_adj одинаковый, аналогично модели являются лучшими по p-статистике у регрессоров. Из этих трёх моделей лучшей является с *current\_pow*=0.1 показателем R2 ~ 0.01821:

*model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))*

Задание №4

Анализ лучшей модели показал, что больше всего зарабатывают молодые мужчины с высшим образованием, которые разведены или не когда не были женаты, проживающие в городах, работающие большое число часов в неделю.

Задание №5

Оценка регрессии для подмножества индивидов:

а) Не состоявшие в браке мужчины, с высшим образованием:

data3 = subset(data2, wed3 == 1)

data3 = subset(data3, higher\_educ == 1)

data3 = subset(data3, sex == 1)

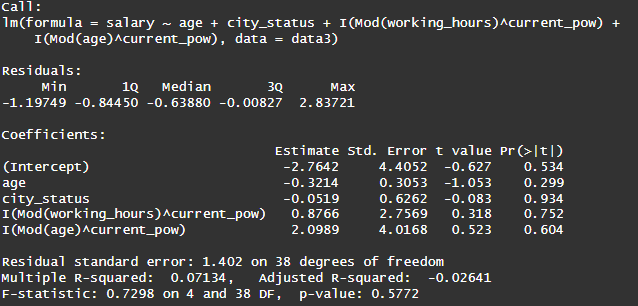


Рисунок №7. Результат вычисления модели *model1 = lm(data = data3, salary ~ age + city\_status + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))* для не состоявших в браке мужчин, с высшим образованием.

Вывод: R^2 ~ 0.07134. Наибольшая зарплата у мужчин с высшим образованием старшего возраста, живущих в городе, не состоявшие в браке, работающих много.

б) Городские жители, состоящие в браке, женщины:

data3 = subset(data2, sex == 0)

data3 = subset(data3, city\_status == 1)

data3 = subset(data3, wed1 == 1)

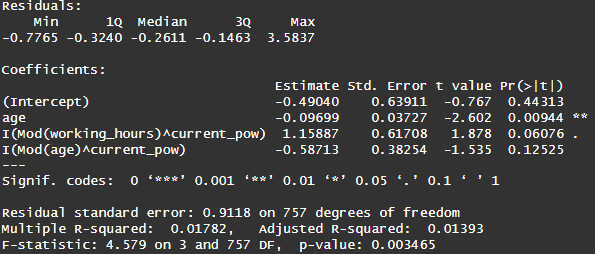


Рисунок №7. Результат вычисления модели *model1 = lm(data = data3, salary ~ age + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))* для городских жителей, состоящих в браке, женщин.

Вывод: *R^2* ~ 0.01782. Наибольшая зарплата у работающих много, молодого возраста.

## Вывод

Из данной работы была получена база данных с выбранными параметрами для анализа, которая была нормализована и проанализирована на наличие линейной зависимости между семейными положениями. После этого была построена линейная регрессия зарплаты на все параметры, которые были выделены из данных мониторинга. Поскольку полученные результаты были неудовлетворительными, были использованы функции вещественных параметров, в том числе логарифм и степени (хотя бы от 0.1 до 2 с шагом 0.1). Исходя из полученных результатов, можно сделать вывод, что ни одна из построенных моделей не дал удовлетворительных результатов, так как значения коэффициента детерминации R^2 оказались низкими. Следует отметить, что при использовании логарифма и степеней вещественных параметров для улучшения результатов, были достигнуты небольшие улучшения. Однако, в целом, даже после применения таких методов, результаты остались неудовлетворительными.

Так же стоит отметить, что из всей статистики больше всего зарабатывают мужчины с высшим образованием, старшего возраста, живущие в городах, работающие большое число часов в неделю.

Среди не состоявших в браке мужчин, с высшим образованием, наибольшая зарплата у мужчин с высшим образованием старшего возраста, живущих в городе, не состоявшие в браке, работающих много.

Среди городских жителей, состоящих в браке, женщин, наибольшая зарплата у работающих много, молодого возраста.

Код решения задачи и сведения о проверенных моделях приведены в Приложении

Заключение

В ходе курсовой работы был проведен анализ нескольких пакетов данных, таких как: Swiss, Seatbelts, Построение регрессии по данным RLMS(r12i\_os26b.csv). А также составлены модели регрессионного анализа и анализирование данных на основе базовых метрик. На основе проведенного анализа были выявлены основные параметры каждого задания.

Помимо этого, были сделаны выводы по каждой задаче и выявлены изъяны при помощи построений моделей, интервалов, а также коэффициент инфляции отклонения и расчета описательной статистики и сводки результатов моделей. По результатам анализа был сделан вывод, …..

# Список литературы

1. руководство пользователя R // An introduction to R URL: https://translated.turbopages.org/proxy\_u/en-ru.ru.3189212b-645772ba-4ef88081-74722d776562/https/cran.rstudio.com/doc/manuals/R-intro.html (дата обращения: 10.04.2023).
2. Hadley Wickham ggplot2: Elegant Graphics for Data Analysis. - Second Edition изд. - Houston, Texas, USA: Springer, 2016. - 260 с.
3. dplyr // A Grammar Of Data Manipulation URL: https://dplyr.tidyverse.org/ (дата обращения: 15.04.2023).
4. Rob J. Hyndman, George Athanasopoulos Forecasting: Principles and Practice. - Second Edition изд. - Monash University, Australia: OTexts, 2018. - 504 с.
5. Julian J. Faraway Practical Regression and Anova using R. - 1-е изд. - Monash University, Australia: University of Bath, July 2002. - 210 с.
6. J. Fox, S. Weisberg An R Companion to Applied Regression. - 3-е изд. - University of California, Davis: SAGE Publications, 2018. - 608 с.
7. Артамонов Н.В. Эконометрика с R. - 3-е изд. - Россия, 119002, Москва: МЦНМО, 2019 - 256 с.

Приложения

Приложение к задаче №1

|  |
| --- |
| library("lmtest")  library("GGally")  library('tidyverse')  data=swiss  summary(data)  ##Среднее значение элементов - " Education, Examination, Fertility пакета swiss.  #data$Examination = 16.48936  #data$Agriculture = 50.65957  #data$Fertility = 70.14255  cat(mean(data$Examination))  cat(mean(data$Agriculture))  cat(mean(data$Fertility))  ##Среднее квадратичное отклонение - "Examination, Agriculture, Fertility пакета swiss.  #standard\_deviation\_Examination: 7.977883  #standard\_deviation\_Agriculture: 22.71122  #standard\_deviation\_Fertility: 12.4917  standard\_deviation\_Examination<-sqrt(sum((data$Examination - mean(data$Examination))^2/(length(data$Examination)-1)))  standard\_deviation\_Agriculture<-sqrt(sum((data$Agriculture - mean(data$Agriculture))^2/(length(data$Agriculture)-1)))  standard\_deviation\_Fertility<-sqrt(sum((data$Fertility - mean(data$Fertility))^2/(length(data$Fertility)-1)))  cat("standard\_deviation\_Examination: " , standard\_deviation\_Examination)  cat("standard\_deviation\_Agriculture: " ,standard\_deviation\_Agriculture)  cat("standard\_deviation\_Fertility: " , standard\_deviation\_Fertility)  ##Дисперсия - Education, Examination, Fertility пакета swiss.  #data$Examination = 63.64662  #data$Agriculture = 515.7994  #data$Fertility = 156.0425  var(data$Examination)  var(data$Agriculture)  var(data$Fertility)  ##Зависимости вида y= a + bx.  summary(lm(Examination~Fertility, data))  summary(lm(Examination~Agriculture, data))  cat(data$Examination\_Fertility<-45.42 + (-0.412)\*data$Fertility)  cat(data$Examination\_Agriculter<-28.70 + (-0.241)\*data$Agriculture)  #График зависимости data$Agriculture  ggplot(data=swiss, aes(x = data$Examination, y = data$Examination\_Agriculter)) +    geom\_point(color= "red" , shape=25, fill= "blue" , size = 2.2, alpha = 0.5) +    geom\_smooth(method = 'loess',size=1.3, level=0.80)  #График зависимости data$Fertility  ggplot(data=swiss, aes(x = data$Examination, y = data$Examination\_Fertility)) +    geom\_point(color= "red" , shape=25, fill="blue" , size =2.2, alpha= 0.5) +    geom\_smooth(method = 'loess' ,size=1.3, level=0.80)  ##Коэффициент детерминации равен 64%, из этого следует, что 64% наших данных описывается нашей моделью.  #Можем сделать вывод (R-squared = 64%), что модель относительно хороша: для такой  #зависимости (только две объясняющая переменная) коэффициент не мал, но для  #полного описания процесса нужно добавлять другие параметры.  ##P-начение p для переменной-предиктора Fertility равно 1.40e-05.(\*\*\*)  #Поскольку это значение меньше 0,05, оно имеет статистически значимую связь с переменной отклика в модели.  #P-начение p для переменной-предиктора Agriculture равно 1.52e-06.(\*\*\*)  #Поскольку это значение меньше 0,05, оно имеет статистически значимую связь с переменной отклика в модели.  model = lm(Examination~Fertility+Agriculture, data )  model  summary(model)  ## Из взаимосвязи вида y= a + bx можно сделать вывод, с ростом коэффициента Agriculture и Fertility уменьшается  # % призывников, получивших наивысшую оценку на армейском экзамене, следовательно взаимосвязи являются отрицательными.  cat(data$Examination\_Fertility<-45.42 + (-0.412)\*data$Fertility)  cat(data$Examination\_Agriculter<-28.70 + (-0.241)\*data$Agriculture)  #Вывод: обе взаимосвязи являются отрицательными. |

Приложение к задаче №2.1

|  |
| --- |
| library("lmtest")  library("GGally")  library("car")  library("dplyr")  data(Seatbelts)  data = na.omit(Seatbelts)  data <- as.data.frame(data)  ##проверим возможность использования всех зявленных регрессоров в одной модели  model1=(lm(law~kms + PetrolPrice, data))  vif(model1)  summary(model1) #R^2 = 28%. относительно независимые регрессоры  model1=(lm(kms~law + PetrolPrice, data))  vif(model1)  summary(model1) #R^2 = 27%. относительно независимые регрессоры  model1=(lm(PetrolPrice~law + kms, data))  vif(model1)  summary(model1) #R^2 = 19%. относительно независимые регрессоры  ##построим общую модель регрессии  model2 = lm(front~law + kms + PetrolPrice, data)  vif(model2)  summary(model2)  #модель неплохая, но есть невзимосвязанная переменная kms (с отсутствием звездочек)  #Adjusted R-squared = 42%. По результатам зависимости front~law и front~PetrolPrice имеет три звездды.  #р-значение у kms большое.  ##Получим коэф. R^2 для каждого отдельно  summary(lm(front~law, data))$r.squared #R^2 равен 31%.  summary(lm(front~kms, data))$r.squared #R^2 равен 12% Попробуем исключить.  summary(lm(front~PetrolPrice, data))$r.squared #R^2 равен 29%.  summary(lm(front~law + PetrolPrice, data)) #Adjuster R-squared = 43%. Коэф. увеличился на 1% и улучшилась p-статистика, следовательно далее пльзуемся этой моделью.  ##Введем логарифмы  #В модель нельзя ввести log регрессоров без замены значений равных 0 в таблице law на 0.001  data$law[data$law == 0] <- 0.001  #log регрессоров  model3=(lm(log(front)~ I(log(law)) + I(log(PetrolPrice)), data))  vif(model3)  summary(model3) #Лучшая модель log(front)~I(log(law), Adjusted R-squared: 0.4835  summary(lm(log(front)~I(log(law)), data))$r.squared #R^2 равен 39%.  model3 = lm(front ~ I(log(law)) + I(log(PetrolPrice)), data)  vif(model)  summary(model) #Лучшая модель log(front)~I(log(law), Adjusted R-squared: 0.4299  summary(lm(log(front)~I(log(law)), data))$r.squared #R^2 равен 39%.  model3 = lm(front ~ I(log(law)) + PetrolPrice, data)  vif(model)  summary(model) #Лучшая модель log(front)~I(log(law), Adjusted R-squared: 0.4309  summary(lm(log(front)~I(log(law)), data))$r.squared #R^2 равен 39%.  model3 = lm(front ~ law + I(log(PetrolPrice)), data)  vif(model)  summary(model) #Лучшая модель log(front)~I(log(law), Adjusted R-squared: 0.4299  summary(lm(log(front)~I(log(law)), data))$r.squared #R^2 равен 39%.  ##Введем в модель всевозможные произведения пар регрессоров  model4 = lm(front ~ law + PetrolPrice + I(law^2) + I(PetrolPrice^2) + I(law\*PetrolPrice), data)  vif(model4) # есть линейная зависимость, уберем регрессоры с максимальным VIF  summary(model4)#Лучшая модель front~law, Adjusted R-squared: 0.428  model4 = lm(front ~ law + PetrolPrice + I(PetrolPrice^2) + I(law\*PetrolPrice), data)  vif(model4) # есть линейная зависимость, уберем регрессоры с максимальным VIF  summary(model4)#Лучшая модель front~law, Adjusted R-squared: 0.428  model4 = lm(front ~ law + PetrolPrice + I(PetrolPrice^2), data)  vif(model4) # есть линейная зависимость, уберем регрессоры с максимальным VIF  summary(model4)#Лучшая модель front~law, Adjusted R-squared: 0.4278  model4 = lm(front ~ law + I(PetrolPrice^2), data)  vif(model4) # линейной зависимости нет  summary(model4)#Лучшая модель front~law, Adjusted R-squared: 0.4306  # Наилучшей из них будет следующая модель:  model3\_1 = (lm(log(front)~ I(log(law)) + I(log(PetrolPrice)), data)) # R^2 = 0.4835  ##вывод: исходя из данного исследования самой лучшей моделью будет model3\_1 со значением Adjusted R-squared: 0.4835,  #самая сильная связь у log(front)~I(log(law) со значением Adjusted R-squared: 0.3911, но также существует хорошая зависимость front~law. |

Приложение к задаче №2.2

|  |
| --- |
| library("lmtest")  library("GGally")  library("car")  library("dplyr")  data(Seatbelts)  data = na.omit(Seatbelts)  data <- as.data.frame(data)  ##Построим общую модель регрессии  model = lm(front~law + kms + PetrolPrice, data)  summary(model)  ##Доверительные интервалы для всех коэффициентов в модели, p = 95%.  confint(model, level = 0.95)  ##проверим гипотезу о том может ли B=0.  #(Intercept) [1.246773e+03 , 1.618217e+03], B не может равняться 0. Следовательно отвергаем гипотезу о том, что этот коэффициент может быть равен 0, на уровне значимости 5%.  #law [-2.894007e+02 ,-1.517127e+02], B не может равняться 0. Следовательно отвергаем гипотезу о том, что этот коэффициент может быть равен 0, на уровне значимости 5%.  #kms [-8.353530e-03 , 6.856602e-03], B может равняться 0. Следовательно подтверждаем гипотезу о том, что этот коэффициент может быть равен 0, на уровне значимости 5%.  #PetrolPrice [-7.118528e+03 ,-3.644096e+03], B не может равняться 0. Следовательно отвергаем гипотезу о том, что этот коэффициент может быть равен 0, на уровне значимости 5%.  ##Для подсчета доверительного интервала для лучшей модели(model3\_1):  data$law[data$law == 0] <- 0.001  model3\_1 = (lm(log(front)~ I(log(law)) + I(log(PetrolPrice)), data))  summary(model3\_1) #Лучшая модель log(front)~I(log(law), Adjusted R-squared: 0.4835  #найдем Число степеней свободы в модели: 189 - 3 = 186.  #После найдем для такого числа степеней свободы и p = 95% значение t-критерия Стьюдента: t = 1.97  cat(t\_critical = qt(0.975, df = 186))  #Тогда для каждого из коэффициентов будут справедливы уравнения:  cat(law\_con <- cbind(((-0.048210) - 1.97\*0.005473), ((-0.048210) + 1.97\*0.005473)))  #law\_con [-0.05899181 -0.03742819], B не может равняться 0. Следовательно отвергаем гипотезу о том, что этот коэффициент может быть равен 0, на уровне значимости 5%.  cat(PetrolPRice\_con <- cbind(((-0.618463) - 1.97\* 0.102867), ((-0.618463) + 1.97\* 0.102867)))  #PetrolPRice\_con [-0.821111 -0.415815], B не может равняться 0. Следовательно отвергаем гипотезу о том, что этот коэффициент может быть равен 0, на уровне значимости 5%.  ##Доверительный интервал для первого прогноза.  new.data = data.frame(law = 20, PetrolPrice = 10)  predict(model3\_1, new.data, interval = "confidence")  #прогноз модели = 3.439129  #нижняя граница доверительного интервала = 2.543131  #верхняя граница доверительного интервала = 4.335128 |

Приложение к задаче №3

|  |
| --- |
| library("lmtest")  library("GGally")  library("car")  library("foreign")  library("dplyr")  library("sandwich")  library("rlms")  library("rstatix")  library("haven")  library("devtools")  data <- read.csv("C:/Users/Иван/Desktop/уник/R/r12i\_os26b.csv",sep=",", dec = ".", header=TRUE)  glimpse(data)  data = select(data, hh5, h\_age, h\_marst, h\_diplom, status, hj13.2, hj6.2)  data = na.omit(data)  glimpse(data)  ##Обновленная база данных для нормализованных значений  data2 = select(data)  glimpse(data)  #Пол  data2["sex"] = 0  data2$sex[which(data$hh5 == 1)] <- 1  #Возраст  age = data$h\_age  data2["age"] = (age - mean(age)) / sqrt(var(age))  ##Семейное положение:  #Состоит ли в зарегестрированном браке?  data2$wed1 = 0  data2$wed1[which(data$h\_marst==2)] <- 1  data2$wed1[which(data$h\_marst==6)] <- 1  #Разведен или вдовец?  data2$wed2 = 0  data2$wed2[which(data$h\_marst==4)] <- 1  data2$wed2[which(data$h\_marst==5)] <- 1  #Никогда не состоял в браке?  data2$wed3 = 0  data2$wed3[which(data$h\_marst==1)] <- 1  #Проверим, что отсутствует линейная зависимость между семейными положениями  vif(lm(data$hj13.2 ~ data2$wed1 + data2$wed2 + data2$wed3))  #Наличие высшего образования  data2$higher\_educ = 0  data2$higher\_educ[which(data$h\_diplom==6)] <- 1  #Живет ли в городе  data2$city\_status = 0  data2$city\_status[which(data$status==1)] <- 1  data2$city\_status[which(data$status==2)] <- 1  #Нормализованное среднее число рабочих часов в неделю  working\_hours = data$hj6.2  data2$working\_hours = (working\_hours - mean(working\_hours)) / sqrt(var(working\_hours))  #Нормализованная средняя зарплата  salary = data$hj13.2  data2$salary = (salary - mean(salary)) / sqrt(var(salary))  #####Постройте линейную регрессию зарплаты на все параметры, которые вы выделили из данных мониторинга. Не забудьте оценить коэффициент вздутия дисперсии VIF.  model1 = lm(data = data2, salary ~ sex + age + wed1 + wed2 + wed3 + higher\_educ + city\_status + working\_hours)  vif(model1)  summary(model1)  #R^2 ~ 0.01819, плохая p-статистика и vif, исключаю wed1  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + higher\_educ + city\_status + working\_hours)  vif(model1)  summary(model1)  #R^2 ~ 0.01784, плохая p-статистика, исключаю higher\_educ, увеличевается R^2.  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status + working\_hours)  vif(model1)  summary(model1)  #R^2 ~ 0.01811, плохая p-статистика.  #####Поэкспериментируйте с функциями вещественных параметров: используйте логарифм и степени (хотя бы от 0.1 до 2 с шагом 0.1).  #Логарифмы и степени имеет смысл вводить только для параметров age и working\_hours, так как остальные принимают только значения 0 или 1.  #с логарифмами:  model1 = lm( salary ~ sex + age + wed2 + wed3 + city\_status + working\_hours + I(log(Mod(working\_hours))) + I(log(Mod(age))), data=data2)  vif(model1)  summary(model1)  #R^2 ~ 0.01875  #плохая p-статистика и vif для working\_hours и I(log(Mod(working\_hours))), исключаю working\_hours  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status + I(log(Mod(working\_hours)))+ I(log(Mod(age))))  vif(model1)  summary(model1)  #R^2 ~ 0.01821  #чуть улучшилась p-статистика и хороший vif.  #Со степенями:  current\_pow=0.1  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  vif(model1)  summary(model1)  #R^2 ~ 0.01821  #Хорошие значения vif  current\_pow = 0.2  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  vif(model1)  summary(model1)  #R^2 ~ 0.01821  #Хорошие значения vif  current\_pow = 0.3  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  vif(model1)  summary(model1)  #R^2 ~ 0.0182  #Хорошие значения vif  current\_pow = 0.4  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  vif(model1)  summary(model1)  #R^2 ~ 0.01818  #Хорошие значения vif  current\_pow = 0.5  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  vif(model1)  summary(model1)  #R^2 ~ 0.01816  #Хорошие значения vif  current\_pow = 0.6  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  vif(model1)  summary(model1)  #R^2 ~ 0.01813  #Хорошие значения vif  current\_pow = 0.7  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  vif(model1)  summary(model1)  #R^2 ~ 0.0181  #Хорошие значения vif  current\_pow = 0.8  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  vif(model1)  summary(model1)  #R^2 ~ 0.01807  #Хорошие значения vif  current\_pow = 0.9  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  vif(model1)  summary(model1)  #R^2 ~ 0.01803  #Хорошие значения vif  #Для степени 1 результат мы уже имеем  current\_pow = 1.1  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status +I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  vif(model1)  summary(model1)  #R^2 ~ 0.01797  #Хорошие значения vif  current\_pow = 1.2  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  vif(model1)  summary(model1)  #R^2 ~ 0.01794  #Хорошие значения vif  current\_pow = 1.3  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  vif(model1)  summary(model1)  #R^2 ~ 0.01791  #Хорошие значения vif  current\_pow = 1.4  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  vif(model1)  summary(model1)  #R^2 ~ 0.01789  #Хорошие значения vif  current\_pow = 1.5  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  vif(model1)  summary(model1)  #R^2 ~ 0.01787  #Далее в период с current\_pow = 1.5 до current\_pow = 2.0: чуть уменьшгается R^2 везде хорошие значения VIF и везде плохая p-статистика.  #####Выделите наилучшие модели из построенных: по значимости параметров, включенных в зависимости , и по объясненному с помощью построенных зависимостей разбросу adjusted R2 - R2adj.  #Сравним лучшие модели:  current\_pow=0.1  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  vif(model1)  summary(model1)  #Multiple R-squared: 0.02018, Adjusted R-squared: 0.01821  current\_pow = 0.2  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status +I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  vif(model1)  summary(model1)  #Multiple R-squared: 0.02018, Adjusted R-squared: 0.01821  current\_pow = 0.3  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status +I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  vif(model1)  summary(model1)  #Multiple R-squared: 0.02017, Adjusted R-squared: 0.0182  #Разброс R2 - R2\_adj везде одинаковый, а R^2 больше для степени 0.2  #Итого, среди моделей без линейной зависимости параметров с хорошими показателями p-статистики у регрессоров лучшей по R^2 оказалась модель для степени 0.1:  current\_pow=0.1  model1 = lm(data = data2, salary ~ sex + age + wed2 + wed3 + city\_status + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  vif(model1)  summary(model1)  #####Сделайте вывод о том, какие индивиды получают наибольшую зарплату.  #Учитывая лучшую модель больше всего зарабатывают молодые мужчины с высшим образованием, которые разведены или не когда не были женаты, проживающие в городах,  #работающие большое число часов в неделю.  #####Оцените регрессии для подмножества индивидов: а) Не состоявшие в браке мужчины, с высшим образованием; б )городские жители, состоящие в браке, женщины.  current\_pow = 0.1  #Не состоявшие в браке мужчины, с высшим образованием  data3 = subset(data2, wed3 == 1)  data3 = subset(data3, higher\_educ == 1)  data3 = subset(data3, sex == 1)  model1 = lm(data = data3, salary ~ age + city\_status + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  #для добавления проверки vif требуется убрать значения NA по этому я удалил wed2 из списка регрессоров модели.  vif(model1) #хорошие значения vif  summary(model1)  # R^2 ~ 0.07134  #Наибольшая зарплата у мужчин с высшим образованием старшего возраста, живущих в городе , не состоявшие в браке, работающих много.  #городские жители, состоящие в браке, женщины  data3 = subset(data2, sex == 0)  data3 = subset(data3, city\_status == 1)  data3 = subset(data3, wed1 == 1)  model1 = lm(data = data3, salary ~ age + I(Mod(working\_hours)^current\_pow) + I(Mod(age)^current\_pow))  #для добавления проверки vif требуется убрать значения NA по этому я удалил wed2 и wed3 из списка регрессоров модели.  vif(model1) #хорошие значения vif  summary(model1)  # R^2 ~ 0.01782  #Наибольшая зарплата у работающих много, молодого возраста. |