Оглавление

[Задачи 3](#_Toc104907581)

[Задача 1 3](#_Toc104907582)

[Задача 2 5](#_Toc104907583)

[Задача 3 10](#_Toc104907584)

[Задача 4 16](#_Toc104907585)

[Заключение 16](#_Toc104907586)

[Список литературы 17](#_Toc104907587)

[**Приложения** 18](#_Toc104907588)

[Приложение 1 18](#_Toc104907589)

[Приложение 2 22](#_Toc104907590)

[Приложение 3 25](#_Toc104907591)

# Задачи

## Задача 1

Необходимо загрузить данные из указанного набора и произвести следующие действия.

Набор данных: Swiss

Объясняемая переменная: Examination

Регрессоры: Infant.Mortality, Education

1. Оцените среднее значение, дисперсию и СКО объясняемой переменной и регрессоров.

Среднее значение столбца переменной Examination примерно равно 17, то есть только 17% призывников получало наивысшую оценку на армейском экзамене. Довольно неплохое значение, но все равно низкое.

У регрессора Infant.Mortality среднее значение столбца почти равно 20, то есть в среднем каждый пятый ребёнок умирал до года, что говорит о крайне низком уровне медицины.

Среднее значение столбца регрессора Education равно почти 11. 11% призывников получали образование выше начального уровня. Это говорит о низком уровне образованности призывников.

Для оценки дисперсии и СКО переменных сначала произведём нормализацию и предобработку данных, а уже потом построим графики, на которых будут видны значения переменных и степень их отклонения от среднего.

Проанализируем такой график для переменной Examination (Рисунок 1.1). В среднем процент призывников, сдавших экзамен на отлично, примерно равен (меньше одного стандартного отклонения), что говорит о одинаковом уровне образования. Есть провинции с отклонением больше одного стандартного, но большинство в отрицательную сторону, то есть меньший процент сдавших на отлично.

Теперь для регрессора Infant.Mortality (Рисунок 1.2). Там все значения, кроме одного, не превышают двух стандартных отклонений (одно значение отличается от среднего на три отклонения). Это говорит о схожести провинций в уровне медицины.

Далее регрессор Education (Рисунок 1.3). Процент призывников, получивших образование выше начального, в большинстве провинций ниже среднего, но в пределах одного отклонения. Есть провинции, значение которых отличаются больше двух отклонений, и одна провинция с высоким процентом образования (отклонение выше 4 стандартных).

1. Постройте зависимости вида y = a + bx, где y – объясняемая переменная, x – регрессор (для каждого варианта по две зависимости).
2. Оцените, насколько «хороша» модель по коэффициенту детерминации R2?
3. Оцените, есть ли взаимосвязь между объясняемой переменной и объясняющей переменной (по значению p-статистики, «количеству звездочек» у регрессора в модели).

Для начала построим зависимость переменной Examination от Education.

Коэффициент детерминации R2 данной модели равен 47.64%, это модель предсказывает примерно половину значений объясняемой переменной. P-статистика равна 4.81e-08, "звёздочек" в модели у регрессора три, а значит взаимосвязь между Examination и Education есть, что довольно логично. Значения коэффициентов модели, их стандартные ошибки, p-статистика и уровень значимости приведены в таблице 1.1.

Таблица 1.1. Характеристики модели зависимости Examination от регрессора Education в наборе данных Swiss.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр/Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | 10.12748 | 1.28589 | 7.876 | 5.23e-10 | \*\*\* |
| Education | 0.57947 | 0.08852 | 6.546 | 4.81e-08 | \*\*\* |

Теперь проанализируем то же самое для зависимости переменной Examination от Infant.Mortality.

Коэффициент детерминации R2 равен 0.8932%, это значит, что данная модель абсолютно не объясняет зависимости данных в наборе Swiss. P-статистика равна 0.45, "звёздочек" в модели у регрессора нет, а значит и взаимосвязи между Examination и Infant.Mortality практически нет. Значения коэффициентов модели, их стандартные ошибки, p-статистика и уровень значимости приведены в таблице 1.2.

Таблица 1.2. Характеристики модели зависимости Examination от регрессора Infant.Mortality в наборе данных Swiss.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр/Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | 22.7175 | 8.1736 | 2.779 | 0.00792 | \*\* |
| Infant.Mortality | -0.3123 | 0.4056 | -0.770 | 0.44538 |  |

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

**Вывод:** были построены модели зависимости отличной аттестации призывников (Examination), от уровня обучения (Education) и смертности до года (Infant.Mortality) в провинциях Франции по данным 1888 года из набора Swiss. По итогу проведенной работы с помощью анализа P-статистики и стандартной ошибки коэффициентов перед регрессорами было установлено, что между Examination и Education есть сильная зависимость, но коэффициент детерминации R2 данной модели равен 47.64%, что указывает на низкую предсказательную способность модели.

## Задача 2

Необходимо загрузить данные из указанного набора и произвести следующие действия.

Набор данных: mtcars

Объясняемая переменная: mpg

Регрессоры: wt, qsec, hp, drat

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.

Проверим линейную регрессию wt ~ qsec + hp + drat: в этой зависимости R2 = 69.1%, то есть существует достаточно сильная зависимость между регрессорами, скорее всего, придется убрать один из регрессоров, но пока переменную wt оставим и попробуем использовать ее в последующих регрессиях.

Далее построим линейную зависимость qsec ~ wt + hp + drat, где R2 = 61.51%, это меньше, чем в предыдущей модели, но все равно наблюдается сильная зависимость, такой зависимостью вряд ли можно пренебречь, рассмотрим с параметром qsec следующие модели и тогда сделаем окончательные выводы.

Конечно, проверим третью модель hp ~ wt + qsec + drat. R2 регрессии равен 77.51%, это очень высокий показатель, поэтому можно сделать вывод, что переменная hp линейно зависима от регрессоров этой модели. Значит, параметр hp стоит убрать и больше не использовать в построении математических моделей.

Также проверим линейную регрессию drat ~ wt + qsec + hp, в этой зависимости R2 = 45.61%, то есть наблюдается средняя зависимость, по сравнению с предыдущими значениями R2 такой зависимостью можно пренебречь, поскольку уже было принято решение убрать регрессор hp. Тогда переменную drat можно использовать в последующих регрессиях.

В конечном итоге я пришел к вводу о том, что регрессор hp очень связан с остальными, поэтому его стоит исключить из рассмотрения последующих моделей. Остальные же регрессоры можно будет использовать с остальными для построения моделей линейных регрессий, но нужно будет внимательно следить за ними, поскольку прослеживается довольно значительная зависимость.

1. Постройте линейную модель зависимой переменной от указанных в варианте регрессоров по методу наименьших квадратов. Оценить, насколько хороша модель, согласно: 1) R2, 2) p-значениям каждого коэффициента.

Построим модель mpg ~ wt + qsec + drat. Значения коэффициентов данной модели, их стандартные ошибки, p-статистика и уровень значимости приведены в таблице 2.1.

Оценим модель: коэффициент детерминации R2 = 81.96%, это очень высокий показатель, значит, модель очень хороша и объясняет данные в наборе mtcars. P-значение регрессора wt очень маленькое (5.01е-07), и у регрессоров qsec и drat тоже маленькие, но уже больше 0.001 и 0.19 соответственно. У wt 3 звёздочки, у qsec 2 звёздочки, а у регрессора drat их нет. Также стоит отметить, что у всех регрессоров достаточно велика стандартная ошибка, а особенно у регрессора drat, она равна 1.23, поэтому, скорее всего, именно drat наиболее незначимый в модели параметр. VIF у каждого регрессора находится в пределах от 1.03 до 2.08, что говорит о независимости регрессоров.

Заключение: данная математическая модель достаточно хорошая, в ней сконцентрированы нужные и важные регрессоры, но возможно еще не все регрессоры стоит использовать, так, например, при удалении регрессора drat из модели R2 падает всего на 0.5%. Дальше попробуем найти не очень нужные регрессоры, которые можно было бы исключить без большого вреда для R2, и тем самым улучшить нашу модель.

Таблица 2.1. Характеристики модели зависимости mpg от регрессоров wt, qsec, drat в наборе данных mtcars.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр/Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | 11.3945 | 8.0689 | 1.412 | 0.16892 |  |
| wt | -4.3978 | 0.6781 | -6.485 | 5.01e-07 | \*\*\* |
| qsec | 0.9462 | 0.2616 | 3.616 | 0.00116 | \*\* |
| drat | 1.6561 | 1.2269 | 1.350 | 0.18789 |  |

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

Для решения данного пункта задания я построил модели с использованием регрессоров: ln(wt), ln(qsec), ln(drat). При анализе построенных моделей я заметил, что добавление в модель логарифма регрессора, где уже есть сам регрессор ведёт к сильному возрастанию vif у пары этих данных. Так же мной было замечено, что регрессор wt уменьшает R2 в любой модели. Поэтому, после тестов всех комбинаций стало ясно, что ln(wt) стоит внести в модель с логарифмом, а сам wt стоит исключить из модели.

Итог: самой лучшей моделью среди моделей с добавлением натуральных логарифмов от регрессоров является эта mpg ~ qsec + drat + ln(wt). Значения коэффициентов модели, их стандартные ошибки, p-статистика и уровень значимости приведены в таблице 2.3.

Таблица 2.3. Характеристики модели зависимости mpg от регрессоров qsec, drat, ln(wt) в наборе данных mtcars.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр/Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | 18.0083 | 7.2144 | 2.496 | 0.018715 | \* |
| qsec | 0.9035 | 0.2245 | 4.025 | 0.000393 | \*\*\* |
| drat | 0.8219 | 1.0818 | 0.760 | 0.453743 |  |
| ln(wt) | -15.1557 | 1.8445 | -8.217 | 6.07e-09 | \*\*\* |

1. 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.

Вначале добавим к регрессорам первоначальной линейной модели всевозможные комбинации с произведениями пар данных регрессоров. Добавлять будем регрессоры: I(wt^2), I(qsec^2), I(drat ^2), I(wt\*qsec), I(wt\*drat), I(drat\*qsec).

Дальше по-отдельности рассмотрим наилучшие модели добавлением квадратов регрессоров и с добавлением произведений пар регрессоров. В первом случае оказалось, что регрессор drat уменьшает R2 в любой модели, а растёт R2 лучше всего при добавлении I(drat ^2). Лучшей моделью с R2 = 82.02% оказалась линейная регрессия с исключением drat и добавлением I(drat^2). Все её характеристики параметров приведены в таблице 2.4.

Таблица 2.4. Характеристики модели зависимости mpg от регрессоров wt, qsec, I(drat^2) в наборе данных mtcars.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр/Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | 14.3106 | 6.4819 | 2.208 | 0.03562 | \* |
| wt | -4.3845 | 0.6744 | -6.502 | 4.8e-07 | \*\*\* |
| qsec | 0.9436 | 0.2611 | 3.614 | 0.00117 | \*\* |
| I(drat^2) | 0.2304 | 0.1657 | 1.390 | 0.17546 |  |

Во втором случае оказалось, что регрессор wt уменьшает R2 в любой модели, а растёт R2 лучше всего при добавлении регрессора I(drat\*wt). Лучшей моделью с R2 = 84.46% оказалась линейная регрессия с исключением wt и добавлением регрессора I(drat\*wt). Все её характеристики параметров приведены в таблице 2.5.

Таблица 2.5. Характеристики модели зависимости mpg от регрессоров qsec, drat, I(drat\*wt) в наборе данных mtcars.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр/Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | 0.3507 | 6.2803 | 0.056 | 0.95586 |  |
| qsec | 0.8944 | 0.2437 | 3.670 | 0.00101 | \*\* |
| drat | 5.1880 | 0.8523 | 6.087 | 1.45e-06 | \*\*\* |
| I(drat\*wt) | -1.3276 | 0.1817 | -7.305 | 5.93e-08 | \*\*\* |

И теперь можем попытаться сравнить первоначальную модель с двумя полученными, первая из которых – лучшая модель с логарифмом, а вторая это лучшая модель с добавлением произведения регрессоров, поскольку модель с добавлением квадратов уже заведомо оказалось хуже (у нее R2 меньше).

Сравним первоначальную модель mpg ~ wt + qsec + drat, где R2 = 81.96% с лучшей моделью добавлением произведения регрессоров mpg ~ qsec + drat + I(drat\*wt), где R2 = 84.96%.

Наилучшая модель выявляется по значению R2, тогда это будет модель с добавлением произведения регрессоров mpg ~ qsec + drat + I(drat\*wt). VIF у регрессоров данной модели не превышают 1.18, а значит, серьезной зависимости между регрессорами нет. Значения коэффициентов данной модели, их стандартные ошибки, p-статистика и уровень значимости приведены в таблице 2.5.

Далее выполним задание 5 и 6 для этой модели.

1. Найти доверительные интервалы для всех коэффициентов в наилучшей модели, p = 95%. Сделать вывод о отвержении или невозможности отвергнуть статистическую гипотезу о том, что коэффициент равен 0.

Найдём значение t, необходимое для определения доверительных интервалов. С помощью функции qt (0.975, df = 28) получим, что t = 2.048407, округлим до двух знаков после запятой, t = 2.05.

Доверительный интервал для коэффициента регрессора *qsec* [0.89-2.05\*0.24;0.89+2.05\*0.24] = [0.4;1.38].

Значение "0" не попадает в доверительный интервал коэффициента перед регрессором, а значит, переменная *mpg* связанаспеременной *qsec.*

Доверительный интервал для коэффициента регрессора *drat* [5.19-2.05\*0.85;5.19+2.05\*0.85] = [3.45;6.93].

Значение "0" не входит в доверительный интервал коэффициента перед регрессором, а значит, *mpg* зависит от *drat.*

Доверительный интервал для коэффициента регрессора *I(drat \* wt)* [-1.33-2.05\*0.18;-1.33+2.05\*0.18] = [-1.7;-0.96].

Значение "0" не входит в доверительный интервал коэффициента перед регрессором, а значит, *mpg* зависит *от I(drat \* wt).*

1. Доверительный интервал для одного прогноза (p = 95%, набор значений регрессоров выбираем сами).

Найдём доверительный интервал для прогноза, в котором переменные будут иметь значения: *qsec* = 17.5, *drat* = 4, *wt* = 3.4. С помощью функции predict получим верхнюю и нижнюю границы доверительного интервала, а также его среднее прогнозируемое значение.

Нижняя граница (lwr в результате функции predict) равна 17.13805, верхняя граница (upr в результате функции predict) равна 20.25923, а среднее прогнозируемое значение (fit в результате функции predict) равно 18.69864.

Код решения задачи и сведения о проверенных моделях приведены в Приложении 2.

***Вывод:*** я проверил данные мне в задание регрессоры *wt, qsec, hp, drat* на линейную зависимость. В результате чего была обнаружена сильная линейная зависимость, а значит, нельзя было использовать все регрессоры вместе. Опираясь на значение R2 , было принято решение исключить параметр *hp* из математических моделей, поскольку R2 = 77.51%. Далее я проверил возможность использования остальных регрессоров вместе: R2 значительно снизился, но все же оставалась средняя зависимость, поэтому важно было следить за оставшимися регрессорами. Затем я построил линейную регрессию с помощью этих переменных и оценил её по величине R2 и по характеристикам коэффициентов перед регрессорами. Математическая модель *mpg ~ wt + qsec + drat,* гдеR2 = 81.96%, оказалась очень хорошей, и она хорошо объясняет данные в наборе mtcars. Несмотря на этот высокий показатель, я попытался улучшить ее с помощью введения натуральных логарифмов от регрессоров и их попарных произведений. Таким образом, проанализировав более 25 линейных регрессий, я смог выявить наилучшую модель с имеющимся набором регрессоров. Наилучшей моделью стала линейная регрессия с добавлением произведения регрессоров *mpg ~ qsec + drat + I(drat\*wt)*, где R2 = 84.96%, что также выше предыдущего значения, а также VIF у регрессоров данной модели не превышают 1.18, а значит, серьезной зависимости между регрессорами нет. В целом модель является очень хорошей, поскольку R2 > 80%, то зависимость определенно есть. С помощью наилучшей модели мне удалось найти доверительные интервалы для всех коэффициентов регрессии *mpg ~ qsec + drat + I(drat\*wt)* (p = 95%) и сделать вывод о том, может ли коэффициент равен нулю или нет. Получилось, что значение «0» не попадает ни в один из доверительных интервалов, поэтому переменная *mpg* зависит от всех 3 регрессоров (*qsec, drat, I(drat\*wt)).* Я рассчитал доверительный интервал для прогноза, выбрав при этом значения регрессоров равными: *qsec* = 17.5, *drat* = 4, *wt* = 3.4, тогда доверительный интервал получился [20.25923;18.69864].

## Задача 3

Номер волны выборки РМЭЗ: 13

Подмножества для пункта 5: Городские жители, состоящие в браке; разведенные, без высшего образования

Необходимо загрузить данные из указанного набора и произвести следующие действия.

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

Для начала выберу набор параметров, который будет необходим, чтобы описать социально-экономическое положение граждан Российской Федерации. Столбца параметров, которые я выбрала: зарплата, пол, семейное положение, наличие высшего образование, возраст, населённый пункт, удовлетворенность условиями труда, продолжительность рабочей недели, опасность производства, причастность государства к владению предприятия и наличие второй работы. В предоставленных данных опроса это столбцы ij13.2, ih5, i\_marst, i\_educ, i\_age, status, ij1.1.2, ij6.2, ij21.3, ij23 и ij32.

Теперь преобразуем выбранные мной параметры по следующему принципу:

* Факторные переменные, «имеющие много значений», такие как: зарплата, длительность рабочей недели и возраст, - преобразуем в вещественные переменные и нормализуем их: вычтем среднее значение по этой переменной, разделим её значения на стандартное отклонение.
* Из параметра, отвечающего типу населённого пункта, создадим одну дамми-переменную city\_status со значением 1 для города или областного центра, 0 – в противоположном случае.
* Из параметра, отвечающего семейному положению, сделаем дамми-переменные 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.
* Из параметра, отвечающего за пол сделаем переменную sex, имеющую значение 1 для мужчин и равную 0 для женщин.
* Для каждого из остальных параметров сделаем отдельные дамми-переменные, которые будут иметь значение 1 для случая, когда выполняется условие в заданном вопросе и значение 0 для случая, когда ответом на вопрос будет «нет».

В итоге получим data2 с новыми переменными, с которыми уже можно работать и построить линейную зависимость зарплаты от остальных переменных: salary ~ sex, wed1, wed2, wed3, higher\_education, age, city\_status, satisfy, duration, dangerous, government, second\_job. Правда в этой модели есть проблема. Значение коэффициента, стандартная ошибка, t value и Pr(>|t|) регрессора wed3 равны NA, то есть wed3 абсолютно никак не связан с зарплатой. Также из-за него нельзя оценитьVIF, я уберу wed3 из модели.

Получим первоначальную рассматриваемую математическую модель salary ~ sex, wed1, wed2, higher\_education, age, city\_status, satisfy, duration, dangerous, government, second\_job.

R2 модели равен ~17%. Три "звёздочки" у регрессоров sex, higher\_education, age, city\_status, satisfy, duration и government, их p-статистика очень мала, это хороший показатель. Две "звёздочки" у регрессора dangerous, его p-статистика тоже невелика, регрессор хороший. И одна "звёздочка" у регрессора wed1, p-статистика = 0.03156, значит регрессор хоть что-то объясняет. VIF всех регрессоров лежит в диапозоне от 1.022587 до 1.668576, это означает, что регрессоры независимы между собой, их все можно использовать вместе в одной модели. Детально с характеристиками получившейся модели можно ознакомиться в таблице 3.1.

Таблица 3.1. Характеристики модели зависимости salary ~ sex, wed1, wed2, higher\_education, age, city\_status, satisfy, duration, dangerous, government и second\_job в наборе данных 13-ой волны исследования.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр/Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | -0.58521 | 0.05745 | -10.186 | < 2e-16 | \*\*\* |
| sex | 0.42510 | 0.03618 | 11.751 | < 2e-16 | \*\*\* |
| wed1 | 0.09489 | 0.04412 | 2.151 | 0.03156 | \* |
| wed2 | 0.04545 | 0.06046 | 0.752 | 0.45224 |  |
| higher\_education | 0.38312 | 0.03999 | 9.579 | < 2e-16 | \*\*\* |
| age | -0.06349 | 0.01894 | -3.353 | 0.00081 | \*\*\* |
| city\_status | 0.34918 | 0.03939 | 8.865 | < 2e-16 | \*\*\* |
| satisfy | 0.29840 | 0.03477 | 8.583 | < 2e-16 | \*\*\* |
| duration | 0.07125 | 0.01780 | 4.004 | 6.40e-05 | \*\*\* |
| dangerous | 0.15649 | 0.04830 | 3.240 | 0.00121 | \*\* |
| government | -0.29403 | 0.03645 | -8.066 | 1.05e-15 | \*\*\* |
| second\_job | -0.01746 | 0.07799 | -0.224 | 0.82286 |  |

1. Поэкспериментируйте с функциями вещественных параметров: используйте логарифм и степени (хотя бы от 0.1 до 2 с шагом 0.1).

Функциями вещественных параметров, кроме зарплаты, в моём случае являются возраст и продолжительность рабочей недели, поэкспериментируем с ними. Для начала будем добавлять в первоначальную линейную регрессию натуральные логарифмы от этих переменных, а именно ln(age) и ln(duration).

С логарифмами там могут быть только три различные вариации: в первый раз добавляем только ln(age), во вторую модель добавляем ln(duration), ну а в третью оба логарифма сразу.

После получения значения R2 и VIF для этих трёх зависимостей точно определяем, что лучшей из них будет третья модель salary ~ sex, wed1, wed2, higher\_education, age, city\_status, satisfy, duration, dangerous, government, second\_job, ln(age), ln(duration). Её коэффициент детерминации R2 равен 28.15%, а команда VIF показывает свои значения в пределах нормы. Подробная информация о значении коэффициентов, стандартных ошибках и значимости использующихся регрессоров приведена в таблице 3.2.

Таблица 3.2. Характеристики модели зависимости salary ~ sex, wed1, wed2, higher\_education, age, city\_status, satisfy, duration, dangerous, government, second\_job, ln(age) и ln(duration) в наборе данных 13-ой волны исследования.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр/Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | -3.924e-01 | 1.964e-01 | -1.998 | 0.046275 | \* |
| sex | 3.047e-01 | 7.652e-02 | 3.982 | 7.89e-05 | \*\*\* |
| wed1 | 1.133e-01 | 1.036e-01 | 1.093 | 0.274866 |  |
| wed2 | 8.890e-02 | 1.246e-01 | 0.714 | 0.475866 |  |
| higher\_education | 5.747e-01 | 8.422e-02 | 6.824 | 2.70e-11 | \*\*\* |
| age | -3.776e-01 | 1.219e-01 | -3.097 | 0.002069 | \*\* |
| city\_status | 4.245e-01 | 7.679e-02 | 4.854 | 1.64e-06 | \*\*\* |
| satisfy | 7.747e-02 | 7.228e-02 | 5.873 | 8.01e-09 | \*\*\* |
| duration | 3.120e-01 | 6.888e-02 | 1.125 | 0.261318 |  |
| dangerous | -2.755e-01 | 9.988e-02 | 3.124 | 0.001893 | \*\* |
| government | 3.120e-01 | 7.042e-02 | -3.912 | 0.000105 | \*\*\* |
| second\_job | -2.755e-01 | 1.439e-01 | -0.269 | 0.788197 |  |
| log(age) | 1.010e-01 | 6.573e-02 | 1.537 | 0.125061 |  |
| log(duration) | 6.958e-05 | 6.795e-02 | 0.001 | 0.999183 |  |

Поэкспериментируем со степенями функций вещественных переменных, то есть будем возводить регрессоры age и duration в разные степени и добавлять их к первоначальной модели порознь и вместе.

Сначала рассмотрим такие модели со степенями регрессоров age и duration от 0.1 до 1. Прежде всего я заметила, что лучшими моделями однозначно будут модели с введением степеней этих переменных вместе, чтобы оба регрессора в степени входили в модель. Так в линейной регрессии salary ~ sex, wed1, wed2, higher\_education, age, city\_status, satisfy, duration, dangerous, government, second\_job, I(age0.1) R2 = 26.48%; в модели salary ~ sex, wed1, wed2, higher\_education, age, city\_status, satisfy, duration, dangerous, government, second\_job, I(duration0.1) R2 = 21.36%; а в модели salary ~ sex, wed1, wed2, higher\_education, age, city\_status, satisfy, duration, dangerous, government, second\_job, I(age0.1), I(duration0.1) коэффициент детерминации R2 уже равен 28.18%.

Дальше я начала увеличивать степени с шагом 0.1 и проверять величину R2 и VIF каждой из моделей. С повышением степеней вплоть до единицы R2 растёт, но вот VIF становился только хуже и уже к возведению в степень 0.4 VIF стал недопустимым для того, чтобы выполнялось условие независимости регрессоров модели. Остановимся на модели, где присутствуют оба регрессора, возведённые в степень 0.3(там ещё VIF соответствует норме): salary ~ sex, wed1, wed2, higher\_education, age, city\_status, satisfy, duration, dangerous, government, second\_job, I(age0.3), I(duration0.3). Её adjusted R2 равен 28.23%, а VIF не превышает ~ 5 для любого из коэффициентов регрессоров. Подробные характеристики данной математической модели приведены в таблице 3.3.

Таблица 3.3. Характеристики модели зависимости salary ~ sex, wed1, wed2, higher\_education, age, city\_status, satisfy, duration, dangerous, government, second\_job, I(age0.3) и I(duration0.3) в наборе данных 13-ой волны исследования.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр/Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | -1.06172 | 0.37331 | -2.844 | 0.00465 | \*\* |
| sex | 0.30362 | 0.07648 | 3.970 | 8.29e-05 | \*\*\* |
| wed1 | 0.11497 | 0.10350 | 1.111 | 0.26720 |  |
| wed2 | 0.09132 | 0.12441 | 0.734 | 0.46330 |  |
| higher\_education | 0.57518 | 0.08417 | 6.834 | 2.54e-11 | \*\*\* |
| age | -0.49739 | 0.17886 | -2.781 | 0.00564 | \*\* |
| city\_status | 0.37420 | 0.07676 | 4.875 | 1.49e-06 | \*\*\* |
| satisfy | 0.42519 | 0.07224 | 5.886 | 7.47e-09 | \*\*\* |
| duration | 0.05069 | 0.10383 | 0.488 | 0.62560 |  |
| dangerous | 0.31230 | 0.09981 | 3.129 | 0.00186 | \*\* |
| government | -0.27638 | 0.07038 | -3.927 | 9.86e-05 | \*\*\* |
| second\_job | -0.03951 | 0.14382 | -0.275 | 0.78367 |  |
| I(age^0.3) | 0.72652 | 0.43537 | 1.669 | 0.09583 |  |
| I(duration^0.3) | 0.10644 | 0.38596 | 0.276 | 0.78285 |  |

Перейдя к степеням от 1 до 2 стала заметна своя зависимость, она немного непохожа на то, что уже было рассмотрено. Здесь чем выше становилась степень, в которую возводились регрессоры age и duration, тем VIF становился лучше(меньше), а R2 ниже. В итоге допустимым для построения адекватной модели VIF стал только к степени 2, то есть к возведению функций вещественных параметров в квадрат. Лучшей из этих моделей является salary ~ sex, wed1, wed2, higher\_education, age, city\_status, satisfy, duration, dangerous, government, second\_job, I(age2), I(duration2). Но её R2 совсем чуть-чуть превышает R2 первоначальной модели, он равен 17.71%. Такая модель не может претендовать на звание лучшей, а именно такую нам надо найти для пункта 3.

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

Сравню двух претендентов на лучшую модель из всех рассмотренных мною. Первая модель - это лучшая модель среди регрессий с введением логарифмов переменных (её характеристики по-прежнему в таблице 3.2), а вторая это лучшая модель среди всех регрессий с введением степеней переменных, а именно модель, где возведение идёт в степень 0.3 (её характеристики можн посмотреть в таблице 3.3).

У первой модели R2 = 28.15%, значимо примерно половина регрессоров. У второй модели R2 = 28.23%, значимо тоже примерно половина регрессоров. Разницы между коэффициентами детерминации почти нет. Посмотрим тогда какие VIF у этих двух регрессий. У модели с логарифмами максимальное значение VIF ~ 4, в то время как у модели со степенью 0.3 VIF достигает почти 9 у двух регрессоров.

Значит, я могу заключить, что лучшей математической моделью из всех рассмотренных является модель salary ~ sex, wed1, wed2, higher\_education, age, city\_status, satisfy, duration, dangerous, government, second\_job, ln(age), ln(duration). На данных этой модели будем делать дальнейшие выводы.

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

Рассмотрим коэффициенты перед значимыми регрессорами в нашей лучшей модели, полученные с помощью команды summary и отображённые в таблице 3.2.

sex: 0.3 - положительный

higher\_education: 0.58 - положительный

age: -0.38 - отрицательный

city\_status: 0.37 - положительный

satisfy: 0.42 - положительный

dangerous: 0.31 - положительный

government: -0.28 – отрицательный

Вывод о том, какие индивиды получают большую зарплату: большую зарплату получают в большинстве своём мужчины, люди с высшим образованием (это самый важный показатель), молодые люди, также люди, проживающие в городе и индивиды, удовлетворённые своими условиями труда. Также большую зарплату получают люди, работающие на опасных или вредных производствах и люди, работающие в негосударственных компаниях. Семейное положение, длительность рабочей недели и наличие второй работы не влияли на уровень заработанной платы в 2004 году.

1. Оцените регрессии для подмножества индивидов, указанных в варианте.

Выделим подмножество городских жителей, состоящих в браке с помощью применения функции subset дважды. Построим выбранную мной лучшую модель зависимости зарплаты от других параметров, но теперь учтём то, что мы находимся в подмножестве городских жителей, состоящих в браке (в этом подмножестве переменные city\_status и wed1 равны единице). salary ~ sex, higher\_education, age, satisfy, duration, dangerous, government, second\_job, log(age), log(duration). R2 = 29%, незначительны переменные duration, second\_job, log(age) и log(duration); VIF нормален (максимум ~ 4). Значения коэффициентов модели, их стандартные ошибки, p-статистика и уровень значимости приведены в таблице 3.4.

Таблица 3.4. Характеристики модели зависимости salary ~ sex, higher\_education, age, satisfy, duration, dangerous, government, second\_job, log(age) и log(duration) в наборе данных 13-ой волны исследования.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр/Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | 0.04408 | 0.28811 | 0.153 | 0.878546 |  |
| sex | 0.36429 | 0.11229 | 3.244 | 0.001367 | \*\* |
| higher\_education | 0.67109 | 0.12049 | 5.570 | 7.61e-08 | \*\*\* |
| age | -0.43755 | 0.16291 | -2.686 | 0.007803 | \*\* |
| satisfy | 0.43347 | 0.11018 | 3.934 | 0.000113 | \*\*\* |
| duration | 0.13038 | 0.12207 | 1.068 | 0.286674 |  |
| dangerous | 0.33540 | 0.13345 | 2.513 | 0.012698 | \* |
| government | -0.34116 | 0.10859 | -3.142 | 0.001917 | \*\* |
| second\_job | -0.20639 | 0.20087 | -1.027 | 0.305354 |  |
| log(age) | 0.10841 | 0.09424 | 1.150 | 0.251265 |  |
| log(duration) | -0.04599 | 0.10795 | -0.426 | 0.670519 |  |

Среди городских жителей в браке всё те же индивиды получают зарплату больше других, а также в этом подмножестве наличие высшего образование играет ещё более важную роль (коэффициент равен 0.67).

Выделим подмножество разведённых жителей без высшего образования с помощью применения функции subset дважды. Построим выбранную мной лучшую модель зависимости зарплаты от других параметров, учитывая, что мы находимся в подмножестве разведённых жителей без высшего образования (в этом подмножестве переменная wed2 равны единице, а переменная higher\_education равна нулю). salary ~ sex, age, city\_status, satisfy, duration, dangerous, government, second\_job, log(age), log(duration). R2 = 7.12% - очень маленький, все переменные, кроме satisfy незначительны (у satisfy 1 "звёздочка"), VIF нормален (максимум ~5.5). Значения коэффициентов модели, их стандартные ошибки, p-статистика и уровень значимости приведены в таблице 3.5.

Таблица 3.5. Характеристики модели зависимости salary ~ sex, age, city\_status, satisfy, duration, dangerous, government, second\_job, log(age) и log(duration) в наборе данных 13-ой волны исследования.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр/Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | -0.12249 | 0.60114 | -0.204 | 0.839 |  |
| sex | 0.13484 | 0.29440 | 0.458 | 0.648 |  |
| age | -0.39153 | 0.43324 | -0.904 | 0.370 |  |
| city\_status | 0.18605 | 0.22037 | 0.844 | 0.402 |  |
| satisfy | 0.55548 | 0.24710 | 2.248 | 0.028 | \* |
| duration | 0.12251 | 0.15806 | 0.775 | 0.441 |  |
| dangerous | 0.20898 | 0.39646 | 0.527 | 0.600 |  |
| government | -0.29888 | 0.21337 | -1.401 | 0.166 |  |
| second\_job | -0.41414 | 0.38274 | -1.082 | 0.283 |  |
| log(age) | 0.09048 | 0.25214 | 0.359 | 0.721 |  |
| log(duration) | 0.05031 | 0.15452 | 0.326 | 0.746 |  |

У людей без высшего образования работа скорее всего предусматривает что-то простое, без возможности карьерного роста, поэтому нет зависимости ни от каких переменных, работать в таких специальностях может каждый.

Код решения задачи и сведения о проверенных моделях приведены в Приложении 3.

**Вывод:** я проанализировала данные опроса НИУ ВШЭ о материальном состоянии граждан России за 2004 год. Мне предстояло найти зависимость между некоторыми параметрами из данных и зарплаты людей. Первым делом мною были выделены параметры, которые, по моему мнению, могли влиять на уровень зарплаты индивидов. Туда попали зарплата, пол, семейное положение, наличие высшего образование, возраст, населённый пункт, удовлетворенность условиями труда, продолжительность рабочей недели, опасность производства, причастность государства к владению предприятия и наличие второй работы. Конечно, для начала я построила линейную регрессию зарплаты на все выбранные мною параметры (salary ~ sex, wed1, wed2, higher\_education, age, city\_status, satisfy, duration, dangerous, government, second\_job) и оценила её: R2 модели равен ~17%, VIF всех регрессоров лежит в диапозоне от 1.022587 до 1.668576. Дальше мною были построены около 30-ти математических моделей, в которых я экспериментировала с добавлением логарифмов функций от вещественных переменных (возраст и продолжительность рабочей недели) и их степеней от 0.1 до 2. У каждой полученной модели я проверяла R2, VIF, и количество значимых регрессоров, таким образом была отобрана лучшая модель salary ~ sex, wed1, wed2, higher\_education, age, city\_status, satisfy, duration, dangerous, government, second\_job, ln(age), ln(duration). Оценив коэффициенты перед объясняющими переменными данной регрессии, я сделала вывод о том, какие же индивиды получают большую зарплату, то есть какие условия должны были быть выполнены, чтобы в 2004 году скорее всего человек получал большую зарплату, чем другие. И в конце задания я выделила два разных подмножества (городские жители, состоящие в браке; разведенные, без высшего образования), построила на этих подмножествах мою «лучшую» получившуюся модель и так же оценила, у каких индивидов больше шанс на получение большей заработанной платы. В итоге можно сказать, что большую зарплату получают в большинстве своём мужчины, люди с высшим образованием, молодые люди, проживающие в городе и индивиды, удовлетворённые своими условиями труда. А также большую зарплату получают люди, работающие на опасных или вредных производствах и люди, работающие в негосударственных компаниях.

## Задача 4

# Заключение

1. В задаче №1 я построила графики стандартных отклонений некоторых столбцов набора данных Swiss. А также я построила две математические модели с одним

регрессором в каждой и оценила их по нескольким параметрам.

1. В задаче №2 я строила большое количество (около 20-ти) линейных зависимостей одной переменной от трёх других, одновременно проверяя регрессоры моделей на независимость. Выделила из регрессий лучшую по доле объяснения данных набора Swiss. Для неё я нашла доверительные интервалы для всех коэффициентов, а далее рассчитала доверительных интервал для одного прогноза.
2. В задаче №3 я тщательно искала зависимости между большим количеством параметров из набора данных. Мне пришлось построить более 30-ти линейных регрессий, чтобы найти самую лучшую из них. Тем самым найти самую оптимальную зависимость параметров между собой. Оценив коэффициенты перед объясняющими переменными данной регрессии, я сделала вывод о том, какие индивиды получают большую зарплату. Также я выделила 2 разных подмножества и на них тоже ответила на вопрос о том, какие индивиды больше зарабатывают.

# Список литературы

1. Introduction to Econometrics with R/Christoph Hanck, Martin Arnold, Alexander Gerber, Martin Schmelzer. - Essen, Germany: University of Duisburg-Essen, 2021.
2. Айвазян, С.А. Основы эконометрики/С.А. Айвазян, В.С. Мхитарян – Москва: Изд. объединение «ЮНИТИ», 1998. – 1005 с.
3. Вербик, М. Путеводитель по современной эконометрике/М. Вербик – Москва: «Научная книга», 2008. – 616 с.
4. Доугерти, К. Введение в эконометрику/К. Доугерти – Москва: ИНФРА-М, 2009. – 465 с.
5. Магнус, Я.Р. Эконометрика. Начальный курс/Я.Р. Магнус, П.К. Катышев, А.А. Пересецкий – Москва: Изд-во «ДЕЛО», 2004. – 576 с.

**Приложения**

Приложение 1

library("lmtest")

data = swiss

help(swiss)

data

mean(data$Examination)

mean(data$Infant.Mortality)

mean(data$Education)

data["Examination1"] = data$Examination - mean(data$Examination)

data["Education1"] = data$Education - mean(data$Education)

data["Infant.Mortality1"] = data$Infant.Mortality - mean(data$Infant.Mortality)

data["Examination2"] = (data$Examination - mean(data$Examination))/sqrt(var(data$Examination))

data["Education2"] = (data$Education - mean(data$Education))/sqrt(var(data$Education))

data["Infant.Mortality2"] = (data$Infant.Mortality - mean(data$Infant.Mortality))/sqrt(var(data$Infant.Mortality))

plot(data$Examination2) + abline(a = 0, b = 0, col = "red")

plot(data$Infant.Mortality2) + abline(a = 0, b = 0, col = "blue")

plot(data$Education2) + abline(a = 0, b = 0, col = "green")

model1 = lm(Examination~Education, data)

model1

summary(model1)

plot(data$Examination, data$Education) + abline(a = 10.13, b = 0.58, col = "red")

model2 = lm(Examination~Infant.Mortality, data)

model2

summary(model2)

plot(data$Examination, data$Infant.Mortality) + abline(a = 22.72, b = -0.31, col = "red")

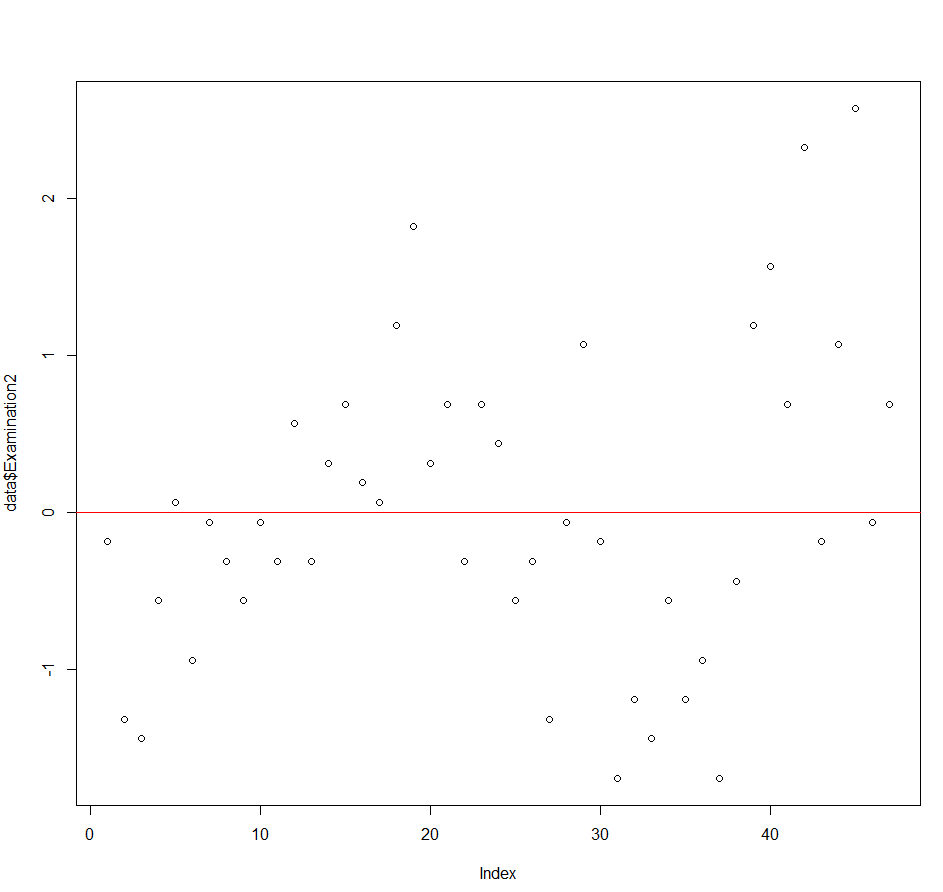


Рисунок 1.1. Отклонения значений переменной Examination от своего среднего значения.

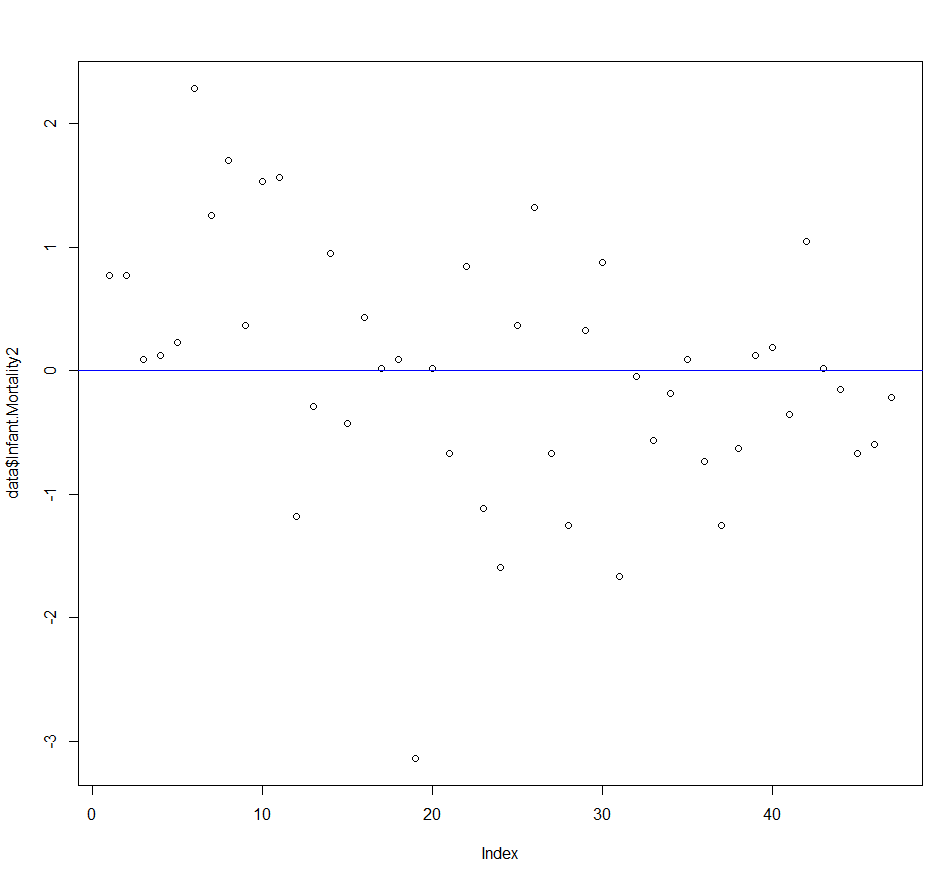


Рисунок 1.2. Отклонения значений переменной Infant.Mortality от своего среднего значения.

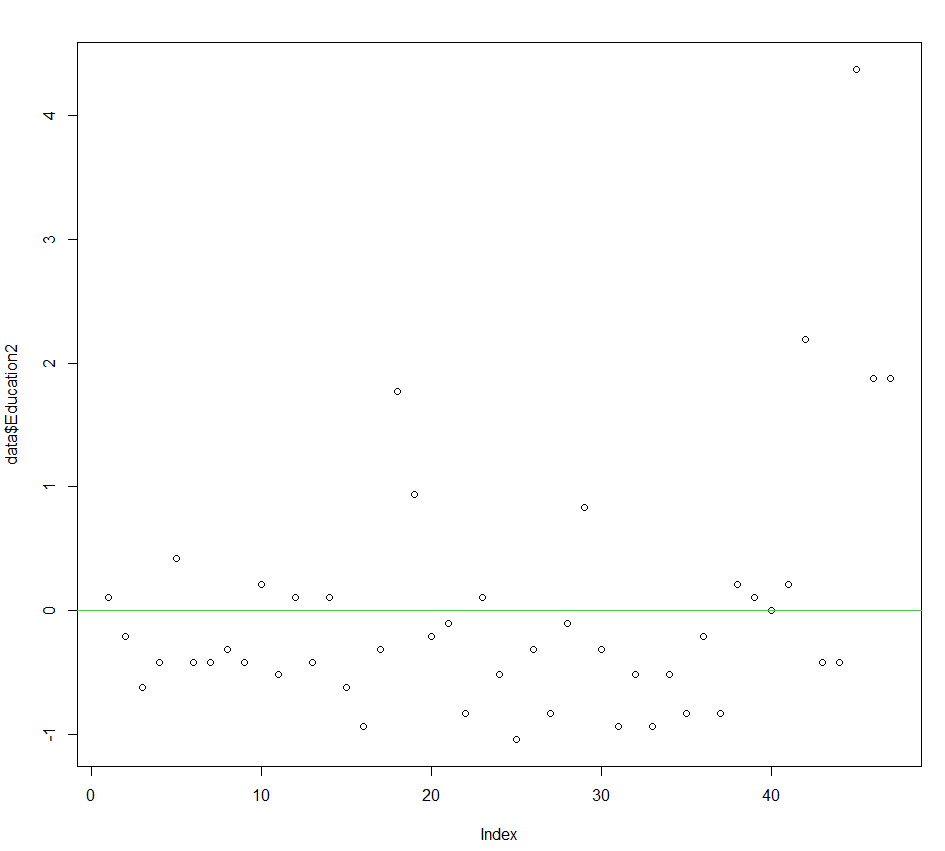


Рисунок 1.3. Отклонения значений переменной Education от своего среднего значения.

## Приложение 2

library("lmtest")

library("car")

data = swiss

help(swiss)

data

#проверим отсутствие линейной зависимости между регрессорами

modelr1 = lm(Catholic~Agriculture+Infant.Mortality, data)

modelr2 = lm(Agriculture~Infant.Mortality+Catholic, data)

modelr3 = lm(Infant.Mortality~Agriculture+Catholic, data)

summary(modelr1)

summary(modelr2)

summary(modelr3)

modelmain = lm(Education~Catholic+Agriculture+Infant.Mortality, data)

summary(modelmain)

vif(modelmain)

#Введём логарифмы в модель

model1 = lm(Education~Catholic+Agriculture+Infant.Mortality+log(Catholic), data)

summary(model1)

vif(model1)

#R^2 = 45.17%

model2 = lm(Education~Catholic+Agriculture+Infant.Mootality+log(Agriculture), data)

summary(model2)

vif(model2)

#R^2 = 53.84%

model3 = lm(Education~Catholic+Agriculture+Infant.Mortality+log(Infant.Mortality), data)

summary(model3)

vif(model3)

#R^2 = 40.67%

model4 = lm(Education~Catholic+Agriculture+Infant.Mortality+log(Catholic)+log(Agriculture), data)

summary(model4)

vif(model4)

#R^2 = 56.82% - это наилучшая модель из всех возможных моделей с введёнными логарифмами от регрессоров

model5 = lm(Education~Catholic+Agriculture+Infant.Mortality+log(Agriculture)+log(Infant.Mortality), data)

summary(model5)

vif(model5)

#R^2 = 52.9%

model6 = lm(Education~Catholic+Agriculture+Infant.Mortality+log(Catholic)+log(Infant.Mortality), data)

summary(model6)

vif(model6)

#R^2 = 43.84%

model7 = lm(Education~Catholic+Agriculture+Infant.Mortality+log(Catholic)+log(Agriculture)+log(Infant.Mortality), data)

summary(model7)

vif(model7)

#R^2 = 55.98%

#Введём всевозможные произведения пар в модель

model8 = lm(Education~Catholic+Agriculture+Infant.Mortality+I(Catholic\*Agriculture), data)

summary(model8)

vif(model8)

#R^2 = 41.31%

model9 = lm(Education~Catholic+Agriculture+Infant.Mortality+I(Agriculture\*Infant.Mortality), data)

summary(model9)

vif(model9)

#R^2 = 40.05%

model10 = lm(Education~Catholic+Agriculture+Infant.Mortality+I(Infant.Mortality\*Catholic), data)

summary(model10)

vif(model10)

#R^2 = 40.71%

model11 = lm(Education~Catholic+Agriculture+Infant.Mortality+I(Catholic\*Agriculture)+I(Agriculture\*Infant.Mortality), data)

summary(model11)

vif(model11)

#R^2 = 40.14%

model12 = lm(Education~Catholic+Agriculture+Infant.Mortality+I(Agriculture\*Infant.Mortality)+I(Infant.Mortality\*Catholic), data)

summary(model12)

vif(model12)

#R^2 = 41.09%

model13 = lm(Education~Catholic+Agriculture+Infant.Mortality+I(Catholic\*Agriculture)+I(Infant.Mortality\*Catholic), data)

summary(model13)

vif(model13)

#R^2 = 42.2%

model14 = lm(Education~Catholic+Agriculture+Infant.Mortality+I(Catholic\*Agriculture)+I(Agriculture\*Infant.Mortality)+I(Infant.Mortality\*Catholic), data)

summary(model14)

vif(model14)

#R^2 = 43.36% - это наилучшая модель из всех возможных моделей с введёнными произведениями пар регрессоров

#Введём квадраты регрессоров в модель

model15 = lm(Education~Catholic+Agriculture+Infant.Mortality+I(Catholic^2), data)

summary(model15)

vif(model15)

#R^2 = 53.69%

model16 = lm(Education~Catholic+Agriculture+Infant.Mortality+I(Agriculture^2), data)

summary(model16)

vif(model16)

#R^2 = 42.12%

model17 = lm(Education~Catholic+Agriculture+Infant.Mortality+I(Infant.Mortality^2), data)

summary(model17)

vif(model17)

#R^2 = 40.46%

model18 = lm(Education~Catholic+Agriculture+Infant.Mortality+I(Catholic^2)+I(Agriculture^2), data)

summary(model18)

vif(model18)

#R^2 = 55.71% - это наилучшая модель из всех возможных моделей с введёнными квадратами регрессоров

model19 = lm(Education~Catholic+Agriculture+Infant.Mortality+I(Agriculture^2)+I(Infant.Mortality^2), data)

summary(model19)

vif(model19)

#R^2 = 41.27%

model20 = lm(Education~Catholic+Agriculture+Infant.Mortality+I(Catholic^2)+I(Infant.Mortality^2), data)

summary(model20)

vif(model20)

#R^2 = 52.62%

model21 = lm(Education~Catholic+Agriculture+Infant.Mortality+I(Catholic^2)+I(Agriculture^2)+I(Infant.Mortality^2), data)

summary(model21)

vif(model21)

#R^2 = 54.79%

model4=

lm(Education~Catholic+Agriculture+Infant.Mortality+log(Catholic)+log(Agriculture), data)

summary(model4)

vif(model4)

modelbest = lm(Education~Agriculture+Infant.Mortality+log(Catholic)+log(Agriculture), data)

summary(modelbest)

vif(modelbest)

t\_critical = qt(0.975, df = 42)

#t\_critical = 2.018082

#Доверительный интервал для одного прогноза

new.data = data.frame(Agriculture = 42,Infant.Mortality = 15, Catholic = 35)

predict(modelbest, new.data, interval = "confidence")

# fit lwr upr

# 13.01212 7.957007 18.06723

Таблица 2.2. Характеристики модели зависимости Education от регрессоров Catholic, Agriculture, Infant.Mortality, ln(Catholic), ln(Agriculture) и ln(Infant.Mortality) в наборе данных Swiss.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр/Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | 84.76697 | 91.22439 | 0.929 | 0.35835 |  |
| Catholic | -0.17014 | 0.09856 | -1.7260 | 0.0920 |  |
| Agriculture | 0.06897 | 0.10523 | 0.655 | 0.51594 |  |
| Infant.Mortality | 0.91567 | 2.60812 | 0.351 | 0.72737 |  |
| log(Catholic) | 5.31734 | 2.70096 | 1.969 | 0.05595 |  |
| log(Agriculture) | -9.62253 | 2.74205 | -3.509 | 0.00113 | \*\* |
| log(Infant.Mortality) | -22.8181 | 48.39734 | -0.471 | 0.63986 |  |

Приложение 3

library("rlms")

data <- rlms\_read("r13i\_os26b.sav")

library("lmtest")

library("dplyr")

library("GGally")

library("car")

library("sandwich")

data1 = select(data, ij13.2, ih5, i\_marst, i\_educ, i\_age, status, ij1.1.2, ij6.2, ij21.3, ij23, ij32)

data1 = na.omit(data1)

glimpse(data)

#зарплата c элементами нормализации

sal = as.numeric(data1$ij13.2)

sal1 = as.character(data1$ij13.2)

sal2 = lapply(sal1, as.integer)

sal = as.numeric(unlist(sal2))

mean(sal)

data1["salary"] = (sal - mean(sal)) / sqrt(var(sal))

#пол

data1["sex"]=data1$ih5

data1["sex"] = lapply(data1["sex"], as.character)

data1$sex[which(data1$sex !='1')] <- 0

data1$sex[which(data1$sex =='1')] <- 1

data1$sex = as.numeric(data1$sex)

#семейное положение

data1["wed"]= data1$i\_marst

data1["wed"] = lapply(data1["wed"], as.character)

data1["wed1"]= data1$i\_marst

data1$wed1 = 0

data1$wed1[which(data1$wed=='2')] <- 1

data1$wed1 = as.numeric(data1$wed1)

data1["wed2"]= data1$i\_marst

data1$wed2 = 0

data1$wed2[which(data1$wed=='4')] <- 1

data1$wed2[which(data1$wed=='5')] <- 1

data1$wed2 = as.numeric(data1$wed2)

data1["wed3"]= data1$i\_marst

data1$wed3 = 0

data1$wed3[which(data1$wed=='1')] <- 1

data1$wed3[which(data1$wed=='3')] <- 1

data1$wed3 = as.numeric(data1$wed3)

#наличие высшего образования

data1["h\_educ"] = data1$i\_educ

data1["h\_educ"] = lapply(data1["h\_educ"], as.character)

data1["higher\_education"] = data1$i\_educ

data1["higher\_education"] = 0

data1$higher\_education[which(data1$h\_educ=='21')] <- 1

data1$higher\_education[which(data1$h\_educ=='22')] <- 1

data1$higher\_education[which(data1$h\_educ=='23')] <- 1

#возраст c элементами нормализации

age1 = as.character(data1$i\_age)

age2 = lapply(age1, as.integer)

age3 = as.numeric(unlist(age2))

data1["age"]= (age3 - mean(age3)) / sqrt(var(age3))

#населенный пункт

data1["status1"]=data1$status

data1["status1"] = lapply(data1["status1"], as.character)

data1["city\_status"] = 0

data1$city\_status[which(data1$status1=='1')] <- 1

data1$city\_status[which(data1$status1=='2')] <- 1

data1$city\_status = as.numeric(data1$city\_status)

#удовлетворенность условиями труда

data1["sat"]=data1$ij1.1.2

data1["sat"] = lapply(data1["sat"], as.character)

data1["satisfy"] = 0

data1$satisfy[which(data1$sat=='1')] <- 1

data1$satisfy[which(data1$sat=='2')] <- 1

data1$satisfy = as.numeric(data1$satisfy)

#продолжительность рабочей недели с элементами нормализации

dur1 = as.character(data1$ij6.2)

dur2 = lapply(dur1, as.integer)

dur3 = as.numeric(unlist(dur2))

data1["duration"] = (dur3 - mean(dur3)) / sqrt(var(dur3))

#опасное ли производство, на котором работает человек

data1["dan"] = data1$ij21.3

data1["dan"] = lapply(data1["dan"], as.character)

data1["dangerous"] = data1$ij21.3

data1["dangerous"] = 0

data1$dangerous[which(data1$dan=='1')] <- 1

#государство - владелец или совладелец предприятия

data1["gov"] = data1$ij23

data1["gov"] = lapply(data1["gov"], as.character)

data1["government"] = data1$ij23

data1["government"] = 0

data1$government[which(data1$gov=='1')] <- 1

#есть ещё работа

data1["w2"] = data1$ij32

data1["w2"] = lapply(data1["w2"], as.character)

data1["second\_job"] = data1$ij32

data1["second\_job"] = 0

data1$second\_job[which(data1$w2=='1')] <- 1

data2 = select(data1, salary, sex, wed1, wed2, wed3, higher\_education, age, city\_status, satisfy, duration, dangerous, government, second\_job)

#линейная регрессия зарплаты на все параметры

model\_test = lm(salary~sex+wed1+wed2+wed3+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job, data2)

summary(model\_test)

vif(model\_test)

model\_original = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job, data2)

summary(model\_original)

vif(model\_original)

#логарифмы

model1 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+log(age), data2)

summary(model1)

vif(model1)

#R^2 = 26.46%, vif в норме

model2 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+log(duration), data)

summary(model2)

vif(model2)

#R^2 = 21.36%, vif в норме

model3 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+log(age)+log(duration), data2)

summary(model3)

vif(model3)

#R^2 = 28.15%, vif в норме, лучшая модель с логарифмами

#степени

model4 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(age^0.1), data2)

summary(model4)

vif(model4)

#R^2 = 26.48%, vif в норме

model5 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(duration^0.1), data2)

summary(model5)

vif(model5)

#R^2 = 21.36%, vif в норме

model6 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(age^0.1)+I(duration^0.1), data2)

summary(model6)

vif(model6)

#R^2 = 28.18%, vif в норме

model7 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(age^0.3), data2)

summary(model7)

vif(model7)

#R^2 = 26.5%, vif в норме

model8 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(duration^0.3), data2)

summary(model8)

vif(model8)

#R^2 = 21.36%, vif в норме

model9 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(age^0.3)+I(duration^0.3), data2)

summary(model9)

vif(model9)

#R^2 = 28.23%, vif в норме

#как видно из этих 6 моделей, наилучшую модель мы получаем, возводя в степень оба регрессора вещественных переменных

model10 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(age^0.4), data2)

summary(model10)

vif(model10)

#R^2 = 26.52%, vif у I(age^0.4) и age равен ~12

model11 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(duration^0.4), data2)

summary(model11)

vif(model11)

#R^2 = 21.36%, vif у duration и I(duration^0.4) равен ~13

model12 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(age^0.4)+I(duration^0.4), data2)

summary(model12)

vif(model12)

#R^2 = 28.29%, vif у I(age^0.4) и age равен ~12; у duration и I(duration^0.4) равен ~13

model13= lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_jobI(age^0.8), data2)

summary(model13)

vif(model13)

#R^2 = 26.56%, vif очень плохой у I(age^0.8) и age = 127

model14 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(duration^0.8), data2)

summary(model14)

vif(model14)

#R^2 = 21.36%, vif очень плохой у duration и I(duration^0.8) = 137

model15 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(age^0.8)+I(duration^0.8), data2)

summary(model15)

vif(model15)

#R^2 = 28.37%, vif очень плохой у I(age^0.8) и age = 127; очень плохой у duration и I(duration^0.8) = 137

#вывод для степеней 0<n< 1: чем ближе степень к единице, тем хуже становится vif. Начиная со степени 0.4 vif уже

#становится недопустимым, модели становятся только хуже, значит здесь лучшая модель это model9

model16 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(age^1.1), data2)

summary(model16)

vif(model16)

#R^2 = 26.59%, vif ужасный у I(age^1.1) и age = 588

model17 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(duration^1.1), data2)

summary(model17)

vif(model17)

#R^2 = 21.36%, vif ужасный у duration и I(duration^1.1) = 615

model18 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(age^1.1)+I(duration^1.1), data2)

summary(model18)

vif(model18)

#R^2 = 28.43%, vif ужасный у I(age^1.1) и age = 550; у duration и I(duration^1.1) = 621

model19 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(age^1.4), data2)

summary(model19)

vif(model19 )

#R^2 = 26.6%, vif очень плохой у I(age^1.4) и age = 42

model20 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(duration^1.4), data2)

summary(model20)

vif(model20)

#R^2 = 21.36%, vif очень плохой у duration и I(duration^1.4) = 42

model21 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(age^1.4)+I(duration^1.4), data2)

summary(model21)

vif(model21)

#R^2 = 28.46%, vif очень плохой у I(age^1.4) и age = 39; у duration и I(duration^1.4) = 44

model22 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(age^1.7), data2)

summary(model22)

vif(model22)

#R^2 = 26.61%, vif плохой у I(age^1.7) и age = 16

model23 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(duration^1.7), data2)

summary(model23)

vif(model23)

#R^2 = 21.37%, vif плохой у duration и I(duration^1.7) = 16

model24 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(age^1.7)+I(duration^1.7), data2)

summary(model24)

vif(model24)

#R^2 = 28.48%, vif плохой у I(age^1.7) и age = 4.5; у duration и I(duration^1.7) = 17

model25 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(age^1.9), data2)

summary(model25)

vif(model25)

#R^2 = 26.61%, vif плохой у I(age^1.9) и age = 10.5

model26 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(duration^1.9), data2)

summary(model26)

vif(model26)

#R^2 = 21.37%, vif плохой у duration и I(duration^1.9) = 10.5

model27 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(age^1.9)+I(duration^1.9), data2)

summary(model27)

vif(model27)

#R^2 = 28.48%, vif почти в норме у I(age^1.9) и age = 9.5; у duration и I(duration^1.9) = 11

model28 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(age^2), data2)

summary(model28)

vif(model28)

#R^2 = 17.7%, vif в норме

model29 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(duration^2), data2)

summary(model29)

vif(model29)

#R^2 = 17%, vif в норме

model30 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+I(age^2)+I(duration^2), data2)

summary(model30)

vif(model30)

#R^2 = 17.71%, vif в норме

model3 = lm(salary~sex+wed1+wed2+higher\_education+age+city\_status+satisfy+duration+dangerous+government+second\_job+log(age)+log(duration), data2)

summary(model3)

vif(model3)

#подмножество городских жителей, состоящих в браке

data3 = subset(data2, city\_status == 1)

data3

data4 = subset(data3, wed1 == 1)

data4

model\_subset1 = lm(data = data4, salary~sex+higher\_education+age+satisfy+duration+dangerous+government+second\_job+log(age)+log(duration))

summary(model\_subset1)

vif(model\_subset1)

#подмножество разведённых людей без высшего образования

data5 = subset(data2, wed2 == 1)

data5

data6 = subset(data5, higher\_education == 0)

data6

model\_subset2 = lm(data = data6, salary~sex+age+city\_status+satisfy+duration+dangerous+government+second\_job+log(age)+log(duration))

summary(model\_subset2)

vif(model\_subset2)