Федеральное государственное образовательное учреждение

высшего профессионального образования

Санкт-Петербургский государственный университет

Институт «Высшая школа менеджмента»

Итоговый проект по курсу «Статистика»

Выполнили студенты 2-го курса бакалаврской программы «Менеджмент»

Номеровкина Ольга Алексеевна

Нигманова Милена Маратовна

Преподаватели

Марченко Ирина Владимировна

**Панкратова Ярославна Борисовна**

Санкт-Петербург

2024

Оглавление

[ВВЕДЕНИЕ 3](#_Toc185455685)

[Шаг 1. Поиск данных по выбранной теме 4](#_Toc185455686)

[Шаг 2. Выявление выбросов в наборе данных 5](#_Toc185455687)

[Шаг 3. Оценивание параметров и построение доверительных интервалов 6](#_Toc185455688)

[Оценка среднего возраста студентов, страдающих депрессией 6](#_Toc185455689)

[Построение доверительного интервала для доли болеющих 7](#_Toc185455690)

[Проверка гипотез 7](#_Toc185455691)

[Выводы по первой части 9](#_Toc185455692)

[Шаг 4. Изучение влияния разных факторов на развитие депрессии 10](#_Toc185455693)

[Нахождение коэффициентов корреляции – анализ тесноты связи 10](#_Toc185455694)

[Анализ тесноты связи между уровнем депрессии и количеством учебных часов 10](#_Toc185455695)

[Анализ тесноты связи между уровнем депрессии и возрастом 10](#_Toc185455696)

[Анализ тесноты связи между уровнем депрессии и уровнем академического давления 11](#_Toc185455697)

[Шаг 5. Построение линейной регрессии 13](#_Toc185455698)

[Зависимость уровня депрессии от количества учебных часов 13](#_Toc185455699)

[Зависимость уровня депрессии от возраста 13](#_Toc185455700)

[Зависимость уровня депрессии от уровня академического давления 14](#_Toc185455701)

[Шаг 6. Построение двухфакторной модели регрессионного анализа 16](#_Toc185455702)

[Зависимость уровня депрессии от академического давления и наследственности 16](#_Toc185455703)

[Зависимость уровня регрессии от академического давления и учебных часов 16](#_Toc185455704)

[ВЫВОДЫ 19](#_Toc185455705)

ВВЕДЕНИЕ

Депрессия – самое распространенное психическое расстройство, проявляющееся устойчивым снижением настроения, двигательной заторможенностью и нарушением мышления. В развитии депрессии участвует комплекс факторов — от генетической предрасположенности до сильного стресса, но что именно провоцирует нарушение, учёным так до сих пор и неизвестно.

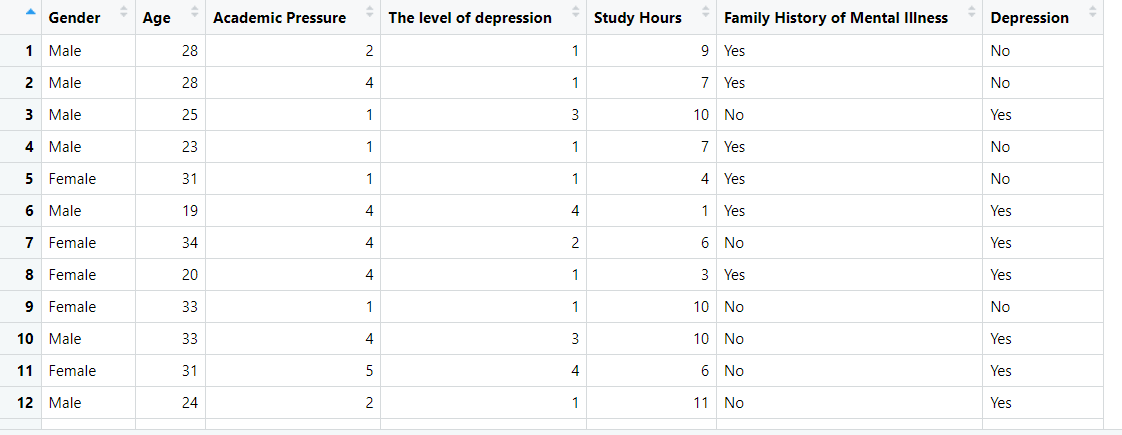
Мы являемся студентами и хотим выяснить, какие факторы оказывают влияние на развитие депрессии у наших коллег.

Таким образом, целью нашего исследования является изучение степени влияния разных факторов на развитие данного заболевания у студентов на основе статистического анализа с помощью программы R.

Перед началом исследования мы поставили перед собой следующие задачи:

1. Подготовить набор данных для исследования по выбранной теме;
2. Выдвинуть ряд гипотез для проверки;
3. Принять или опровергнуть выдвинутые гипотезы на основе статистического анализа данных;
4. Сформулировать выводы о полученных результатах.

Шаг 1. Поиск данных по выбранной теме

В представленной таблице указаны:

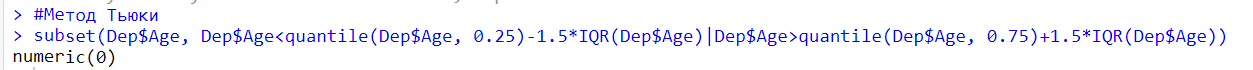
|  |  |
| --- | --- |
| **Название столбца** | **Принимаемые значения** |
| Gender (Пол студента) | Male (мужской)/Female(женский) |
| Age (Возраст студента) | От 18 до 34 лет |
| Academic Pressure (Академическое давление - уровень стресса, вызванный высокими требованиями к успеваемости) | От 1 до 5 (чем выше требования к успеваемости, тем выше оценка: 5 – очень высокое академическое давление) |
| Study Hours (Количество учебных часов) | От 0 до 12 часов |
| Family History of Mental Illness (Наличие психических расстройств у родственников) | Yes (есть родственники, страдающие депрессией)/No (нет родственников, страдающих депрессией) |
| Depression (Наличие депрессии) | Yes (наличие расстройства у студента)/No (отсутствие депрессии) |
| The level of depression (Уровень депрессии) | От 1 до 5 (1 - нет депрессии, 2 - легкая депрессия, 3 – средняя, 4 - умеренная, 5 – тяжелая) |

Шаг 2. Выявление выбросов в наборе данных

Для начала очистим данные от выбросов – элементов, значительно отличающихся от других элементов выборки. В наших данных единственный столбец, который может содержать выбросы – это возраст респондентов, т. к. в число студентов могут входить взрослые люди, возраст которых существенно выше остальных.

Найдём выбросы **методом Тьюки**:

Изображение выглядит как диаграмма, Прямоугольник, снимок экрана, линия

Автоматически созданное описание

**Итог:** выбросов нет, можно приступать к основной части работы.

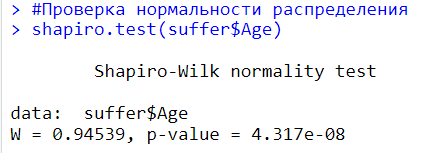
Шаг 3. Оценивание параметров и построение доверительных интервалов

Перед тем как приступить к основной части нашей работы, мы решили дать оценку среднему возрасту студентов, страдающих депрессией, построить доверительный интервал для доли болеющих, а также принять или отклонить выдвинутые нами гипотезы:

* Средний возраст болеющих студентов-девушек и студентов-юношей одинаков.
* Доли юношей и девушек, страдающих депрессией, равны.
* Доли студентов-девушек и студентов-юношей, страдающих депрессией, у которых есть в роду данное заболевание, равны.
* Доля болеющих студентов, у которых есть в роду депрессия, равна доле студентов, у которых в роду нет данного заболевания.

Оценка среднего возраста студентов, страдающих депрессией

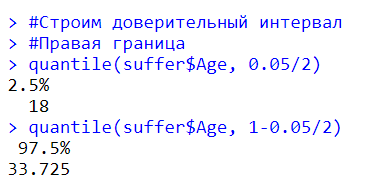
Шаг 1.

Шаг 2. Проверка нормальности распределения

**Итог:** p-value<0.05, значит, распределение ненормальное.

Шаг 3. Построение доверительного интервала

Распределение ненормальное, следовательно, чтобы дать оценку математическому ожиданию, необходимо построить доверительный интервал с помощью функции quantile().



**Итог:** с вероятностью 0.95 средний возраст студента, страдающего депрессией, лежит в промежутке от 18 до (примерно) 34 лет. Обосновать полученный результат затруднительно ввиду большой длины интервала.

Построение доверительного интервала для доли болеющих



**Итог:** с вероятностью 0.95 доля студентов, страдающих депрессией, находится в промежутке от 0.46 до 0.55. Следовательно, около половины студентов страдают данным психическим расстройством.

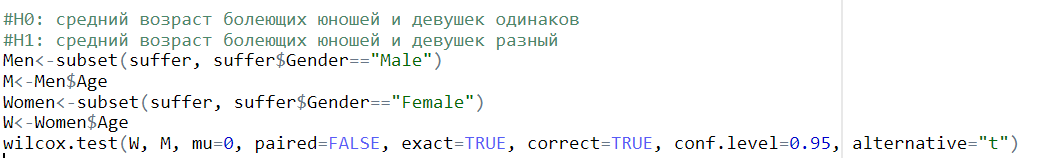
Проверка гипотез

Гипотеза 1:

* Средний возраст болеющих студентов-девушек и студентов-юношей одинаков.

Шаг 1. Делаем выборку юношей, страдающих депрессией и выборку девушек, страдающих депрессией.

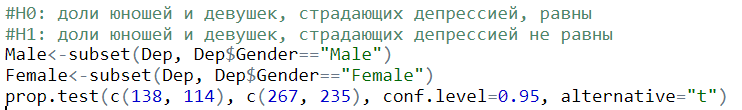
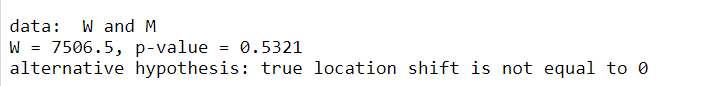
***Шаг 2.*** С помощью критерия Уилкоксона о значении медианы разности двух генеральных совокупностей решаем, принимать или отклонять гипотезу.

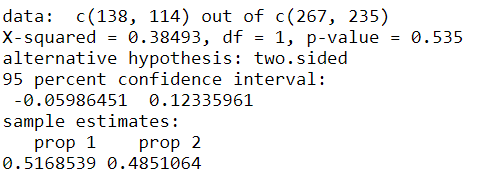
**Итог:** p-value>0.05, гипотезу принимаем.

**Гипотеза 2:**

* Доли юношей и девушек, страдающих депрессией равны.

Шаг 1. Находим общее количество девушек в данных и общее количество юношей.

****Шаг 2. С помощью prop.test() проверяем гипотезу о равенстве долей.

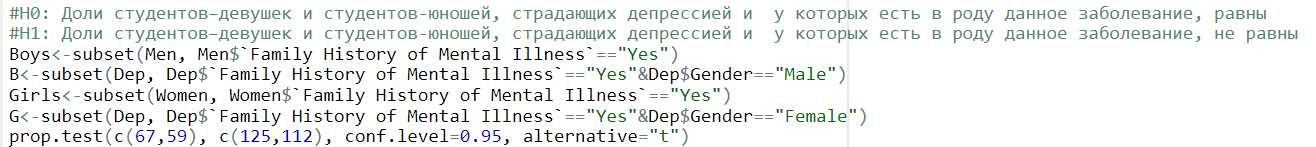


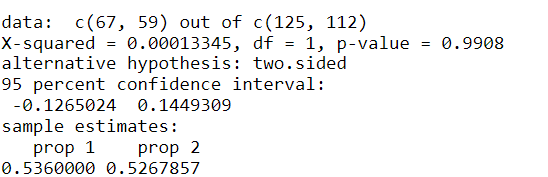
**Итог:** p-value>0.05, гипотезу принимаем. Следовательно, депрессия примерно одинаково распространена среди студентов-юношей и студентов-девушек.

Гипотеза 3:

* По наследству депрессия в основном передается юношам (это наша альтернативная гипотеза).

***Шаг 1.*** Находим количество юношей-студентов, которые страдают депрессией, и у которых в роду есть родственники с данным заболеванием. Аналогичную процедуру проводим для девушек.

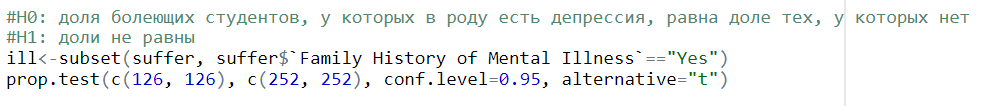
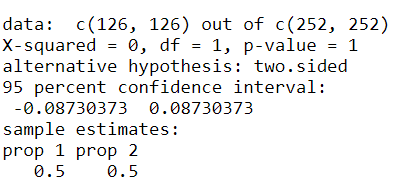
***Шаг 2.***С помощью prop.test() проверяем гипотезу о равенстве долей.



**Итог:** p-value>0.05, следовательно, принимаем нулевую гипотезу. Можно утверждать, что депрессия передается по наследству вне зависимости от пола, так как доли с вероятностью 0.95 равны.

Гипотеза 4:

* Доля болеющих студентов, у которых есть в роду депрессия, равна доле студентов, у которых в роду нет данного заболевания.



**Итог:** p-value>0.05, гипотезу принимаем,

Выводы по первой части

Сделать выводы относительно среднего возраста болеющих студентов и предположить возможные причины и пути их решения затруднительно, поскольку распределение ненормальное, доверительный интервал слишком большой. Как выяснилось, с вероятностью 0.95 можно утверждать, что около половины студентов страдают депрессией, что делает актуальным наш последующий анализ о влиянии факторов на развитие данного заболевания.

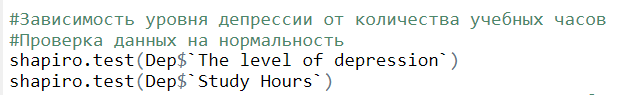
В ходе проверки гипотез мы выяснили, что юноши и девушки подвержены депрессии в равной степени, причем доля студентов девушек и юношей, у которых есть родственники, страдающие данным заболеванием, одинаковы. Кроме того, с вероятностью 0.95 можно утверждать, что депрессия не передается по наследству, так как мы приняли гипотезу о том, что доля болеющих студентов, у которых есть родственники с данным заболеванием, равна доле студентов, у которых таких родственников нет.

Шаг 4. Изучение влияния разных факторов на развитие депрессии

Нахождение коэффициентов корреляции – анализ тесноты связи

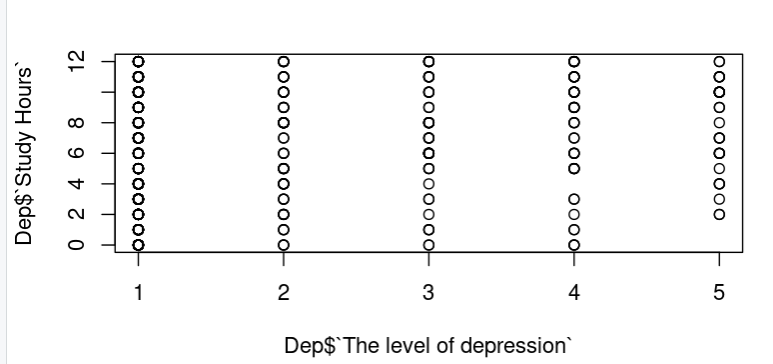
Анализ тесноты связи между уровнем депрессии и количеством учебных часов

Шаг 1. Проверка данных на нормальность.



**Итог:** в обоих случаях p-value<0.05, распределения ненормальные.

Шаг 2. Построение графика и нахождение коэффициента корреляции.



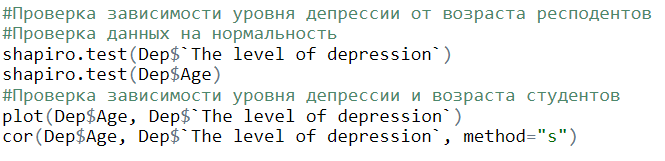
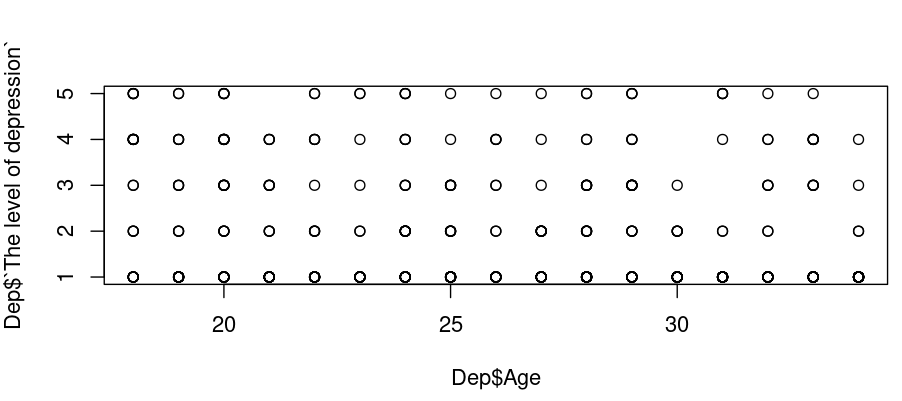


**Итог:** коэффициент корреляции равен 0.18, степень связи между количеством учебных часов и уровнем депрессии очень слабая.

Анализ тесноты связи между уровнем депрессии и возрастом

Шаг 1. Проверка данных на нормальность.

Шаг 2. Построение графика и нахождение коэффициента корреляции.

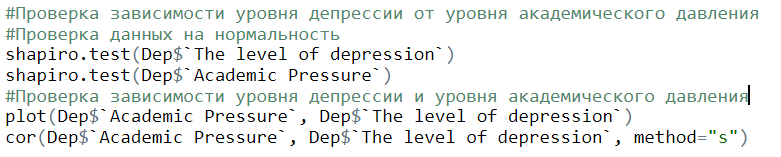


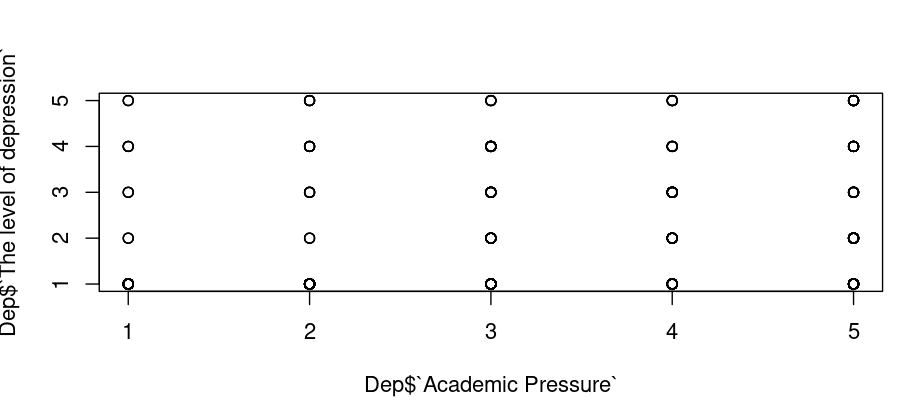
**Итог:** коэффициент корреляции равен -0.2, т. е. чем старше студент, тем меньше он страдает от депрессии. Возможно, это связано с тем, что чем дольше студенты учатся, тем грамотнее они распределяют своё время на отдых и учебу, а также в большей степени понимают предъявляемые требования. Однако зависимость очень слабая, поэтому делать какие-либо предположения не имеет особого смысла.

Анализ тесноты связи между уровнем депрессии и уровнем академического давления

Шаг 1. Проверка данных на нормальность.

Шаг 2. Построение графика и нахождение коэффициента корреляции.

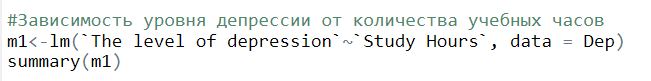


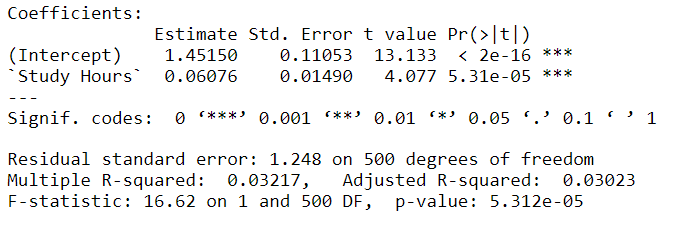


**Итог:** коэффициент корреляции равен 0.3. Чем больше академическая нагрузка, тем обычно выше уровень депрессии. Скорее всего, это связано с тем, что высокие требования к успеваемости студентов заставляют их больше переживать по поводу предстоящих работ и результатов, что негативно сказывается на их настроении и ментальном самочувствии. Чтобы снизить данную зависимость, студентам необходимо научиться грамотно планировать свое расписание, оставляя время на отдых, и контролировать свои тревожные мысли. Регулярные сессии с психологом могут пойти на пользу.

Шаг 5. Построение линейной регрессии

Зависимость уровня депрессии от количества учебных часов





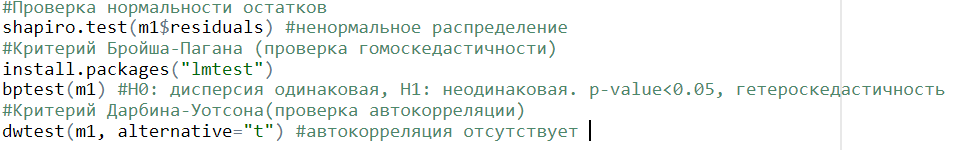
**Итог:**

The level of depression = 1.45 + 0.06\* `Study Hours`+ **ε**.

P-value обоих коэффициентов меньше 0.05, следовательно, коэффициенты значимы. Следовательно, при увеличении количества учебных занятий на час уровень депрессии среди студентов увеличивается на 0.06.

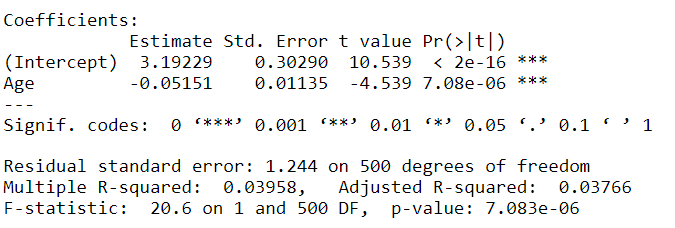
Коэффициент детерминации значимый, т. к. p-value<0.05. Можно сделать вывод о том, что изменение уровня депрессии у студентов лишь на 3% объясняется изменением количества учебных часов.

Чтобы модели можно было доверять, необходимо проверить остатки на нормальность распределения (распределение должно быть нормальным с математическим ожиданием равном нулю), гомоскедастичность (дисперсии должны быть одинаковыми), независимость (не должно быть автокорреляции остатков).



Таким образом, модели нельзя доверять, т. к. не соблюдается два условия.

Зависимость уровня депрессии от возраста

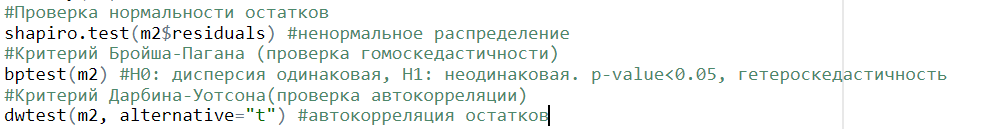


**Итог:**

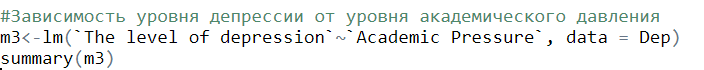
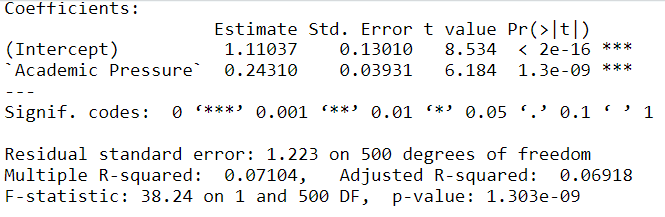
The level of depression = 3.19 - 0.05\*Age+ **ε**

**Коэффициенты значимы, т. к. p-value<0.05. Чем старше студент, тем ниже уровень депрессии, причем при увеличении возраста студента на 1 год уровень депрессии снижается на 0.05.**

**Коэффициент детерминации значимый, следовательно, изменение уровня депрессии лишь на 3% объясняется изменением возраста.**



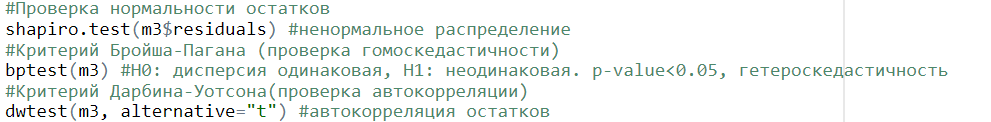
Однако модели нельзя доверять, т. к. остатки не соответствуют условиям хорошей модели.

Зависимость уровня депрессии от уровня академического давления

**Итог:**

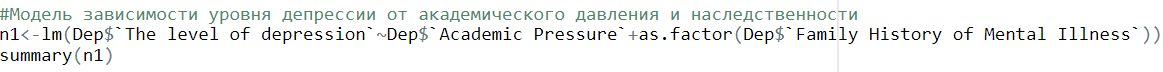
The level of depression = 1.11 + 0.24\*`Academic Pressure` + **ε**

Коэффициенты значимы, p-value<0.05. Чем выше академическое давление (требования к успеваемости студента), тем выше уровень депрессии.

Коэффициент детерминации значимый, изменение уровня депрессии на 7% объясняется изменением уровня академического давления.

Однако и этой модели нельзя доверять, поскольку не соблюдаются условия для остатков.

Шаг 6. Построение двухфакторной модели регрессионного анализа

Зависимость уровня депрессии от академического давления и наследственности

Изображение выглядит как текст, Шрифт, снимок экрана

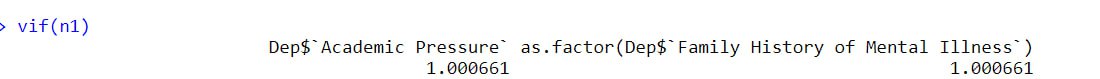
Автоматически созданное описание

**Итог:**

The level of depression = 0.98 + 0.25 \* `Academic pressure` + 0.26 \* `Family History of Mental Illness` + **ε**

Переменная `Family History of Mental Illness` принимает значение 1, если в роду у студента была депрессия, и 0 – если не было.

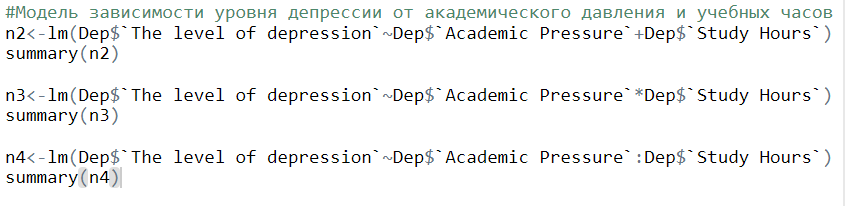
Коэффициенты данной модели являются значимыми. Можно сделать следующие выводы: при увеличении академического давления на 1 единицу уровень депрессии увеличивается на 0.25, при условии, что семейная история заболеваний студента остается неизменной. При неизменном академическом давлении наличие депрессии у родственников увеличивает уровень депрессии студента на 0.26 по сравнению с теми, у кого в роду депрессии нет.

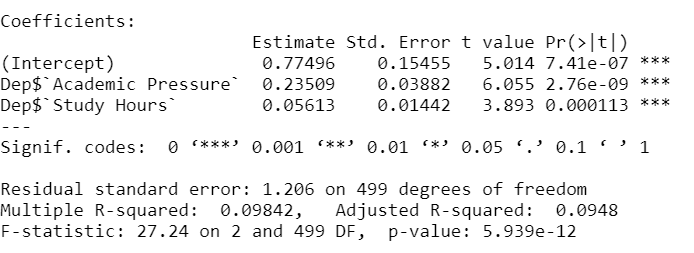
В соответствии с коэффициентом детерминации (Adjusted R-squared) только 8% изменения уровня депрессии объясняется изменением академического давления.

Мультиколлинеарность небольшая, качество модели хорошее.

Зависимость уровня регрессии от академического давления и учебных часов

Мы построили 3 модели.

**Модель n2** показывает зависимость уровня депрессии от двух независимых переменных, т. е. коэффициенты перед этими переменными показывают, как изменение каждой из переменных влияет на уровень депрессии, если другая переменная остается неизменной.

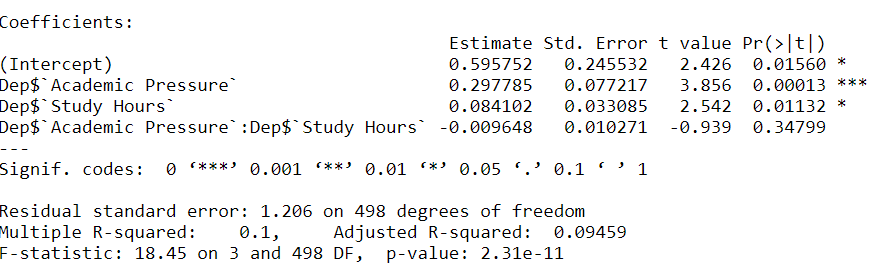


The level of depression = 0.77 + 0.24 \* `Academic pressure` + 0.06 \* `Study Hours` + **ε**

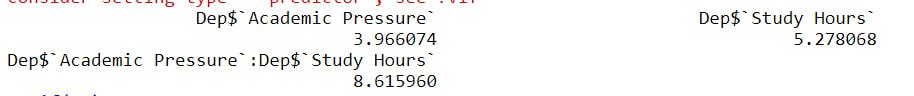


Мультиколлинеарность небольшая, качество модели хорошее.

**Модель n3** показывает зависимость уровня депрессии от академического давления и количества учебных часов как отдельно (т. е. влияние двух независимых переменных), так и при учете влияния этих переменных друг на друга.

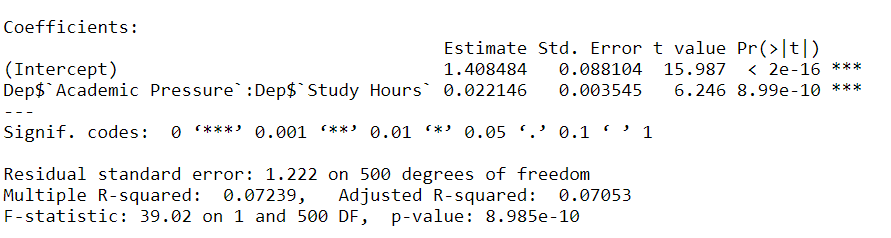


The level of depression = 0.59 + 0.3 \* `Academic pressure` + 0.08 \* `Study Hours - 0.01 \* `Academic pressure` \* `Study Hours` + **ε**

****

**Умеренная коллинеарность, качество модели низкое.**

**Модель n4** показывает зависимость уровня депрессии от переменных, влияющих друг на друга, т. е. не учитывается отдельное влияние учебных часов и академического давления.



The level of depression = 1.41 + 0.02 \* `Academic pressure` \* `Study Hours` + **ε**

**Итог:** коэффициент детерминации больше всего у модели n2, следовательно, данная модель лучше всего объясняет вариацию уровня депрессии, а именно на 9%.

**Уравнение для модели:**

The level of depression = 0.77 + 0.24 \* `Academic pressure` + 0.06 \* `Study Hours` + **ε**

При неизменных учебных часах повышение требований к успеваемости студентов на 1 поднимает уровень депрессии на 0.24, а при неизменном академическом давлении увеличение учебных часов на 1 увеличивает уровень депрессии студента на 0.06.

ВЫВОДы

В нашем случае безрезультативность – это тоже результат.

В ходе исследования мы выяснили, что с вероятностью 0.95 депрессией страдает половина студентов, причем заболевание одинаково распространено среди юношей и девушек.

Мы построили двухфакторные модели регрессионного анализа с целью изучить влияние факторов, таких как академическое давление, количество учебных часов, возраст на уровень депрессии. Однако ни одной из получившихся моделей нельзя доверять.

Кроме того, мы построили многофакторные модели регрессионного анализа, но вариация уровня депрессия объясняется на небольшое число процентов изменениями внутри моделей, следовательно, модели не имеют особой значимости.

Таким образом, мы не смогли на основе статистического анализа выяснить, какие факторы в большей степени влияют на развитие депрессии у студентов. Причины развития депрессии в целом так и остаются загадкой. Мы предполагаем, что это связано с тем, что психика каждого человека уникальна, и ряд факторов, который может спровоцировать данное заболевание, как у студента, так и у любого человека, особенный.

Наша единственная рекомендация студентам – это беречь себя и не болеть!