МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ АГЕНТСТВО ЖЕЛЕЗНОДОРОЖНОГО ТРАНСПОРТА
Государственное бюджетное образовательное учреждение
высшего образования
«ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ПУТЕЙ СООБЩЕНИЯ ИМПЕРАТОРА АЛЕКСАНДРА I»

Кафедра «ИНФОРМАЦИОННЫЕ И ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ»

Дисциплина: «Информатика»

О Т Ч Е Т
по лабораторной работе № 2

Выполнил студент Шефнер А.
Факультета *АИТ*
Группы *ИВБ-211*

Санкт-Петербург
2023

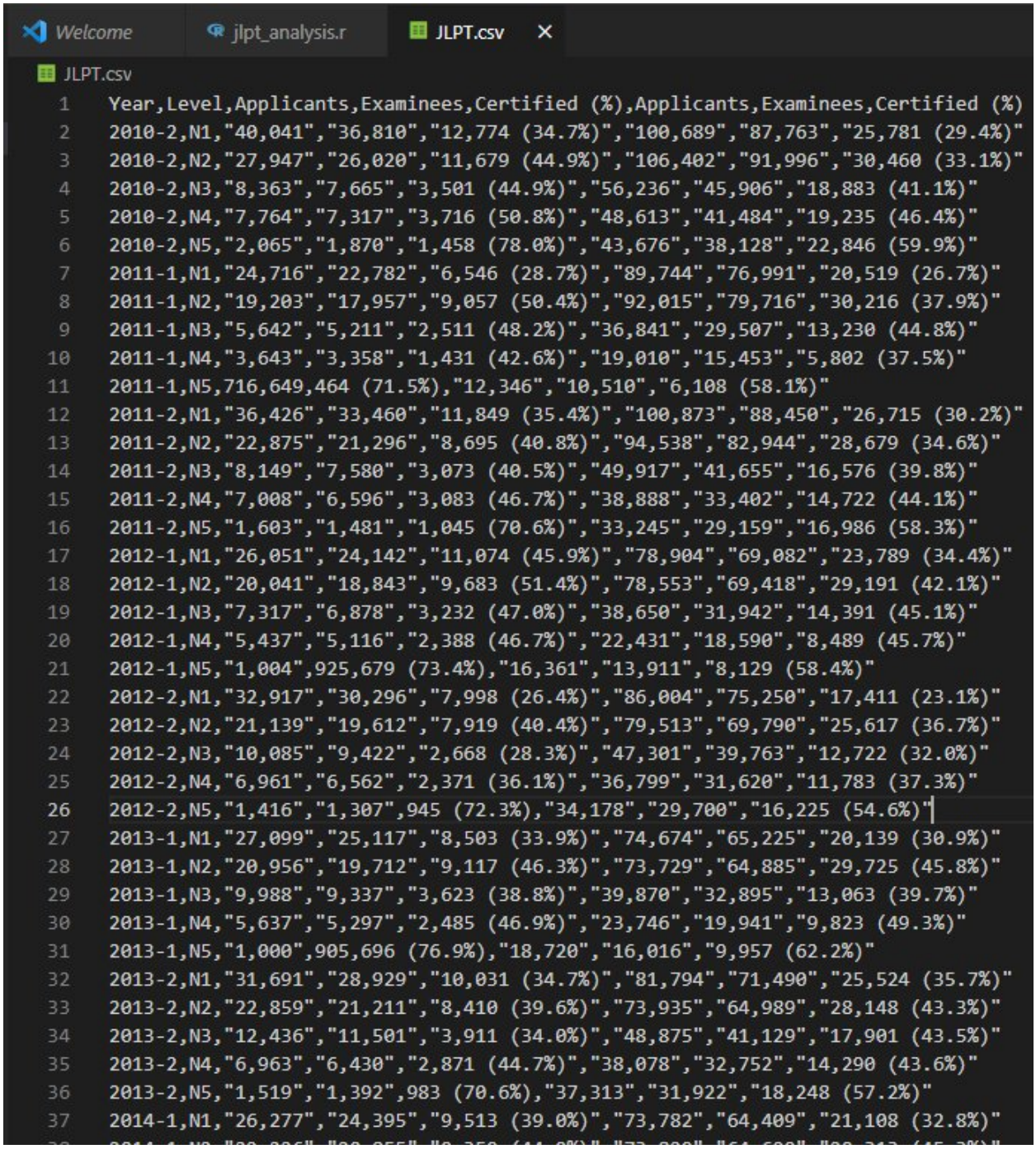
Постановка задачи

Необходимо создать R-скрипт, в котором реализовать модель линейной регрессии. Скрипт должен содержать:

1. Чтение данных формате csv.
2. Обучение модели и нахождение коэффициентов.
3. Вывод результатов на экран.
4. Графическое изображение исходных данных и найденной прямой линии.

Csv-файл:

Файл, который я использовал для анализа содержит в себе данные о количестве заявок, проведённых и сданных экзаменов JLPT разных уровней в Японии и за рубежом. Данные взяты с сайта Wikipedia. Выглядит он примерно так:



Welcome | jlpt_analysis.r | JLPT.csv

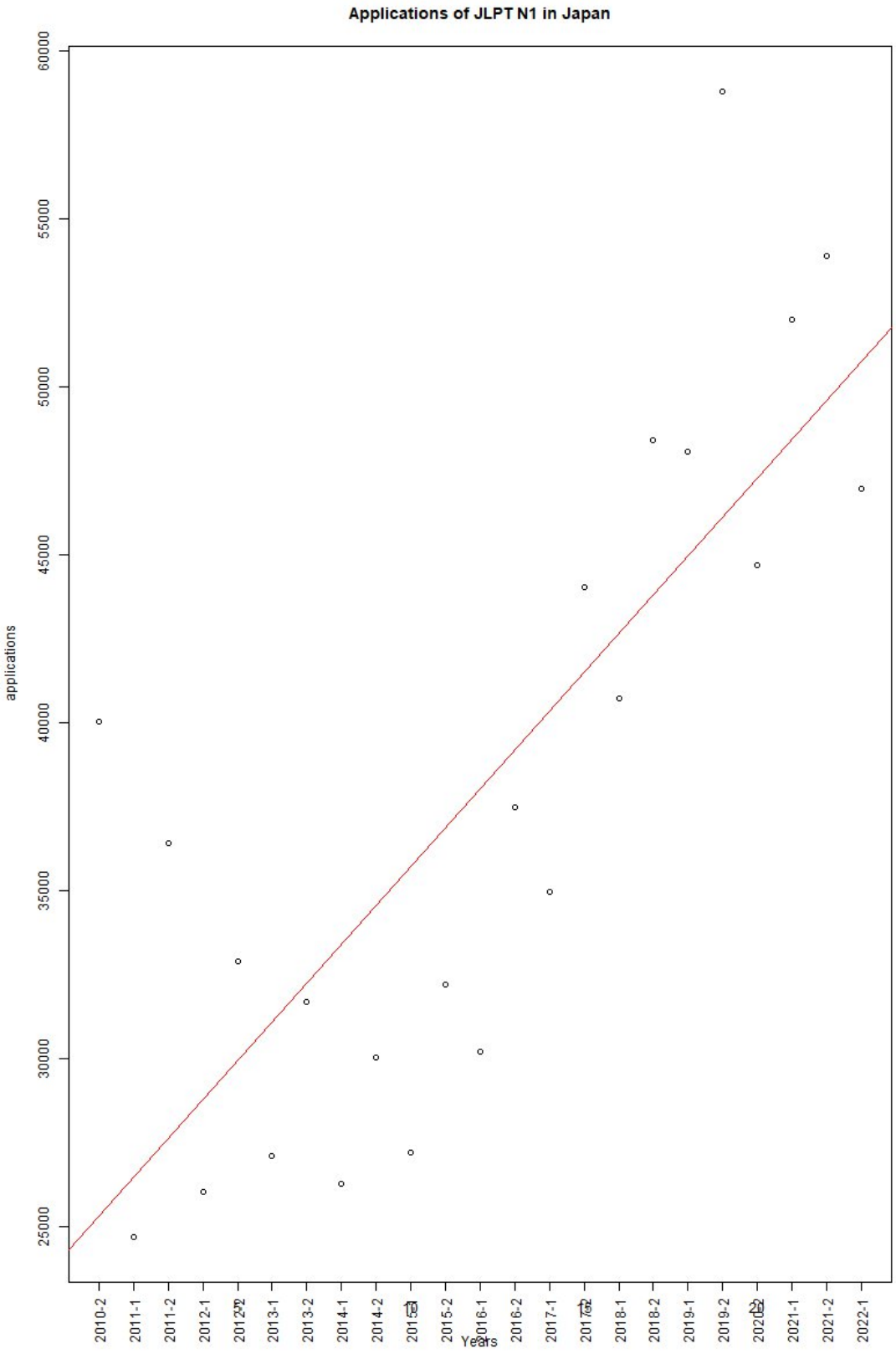
JLPT.csv

```
Year,Level,Applicants,Examinees,Certified (%),Applicants,Examinees,Certified (%)
2010-2,N1,"40,041","36,810","12,774 (34.7%)","100,689","87,763","25,781 (29.4%)"
2010-2,N2,"27,947","26,020","11,679 (44.9%)","106,402","91,996","30,460 (33.1%)"
2010-2,N3,"8,363","7,665","3,501 (44.9%)","56,236","45,906","18,883 (41.1%)"
2010-2,N4,"7,764","7,317","3,716 (50.8%)","48,613","41,484","19,235 (46.4%)"
2010-2,N5,"2,065","1,870","1,458 (78.0%)","43,676","38,128","22,846 (59.9%)"
2011-1,N1,"24,716","22,782","6,546 (28.7%)","89,744","76,991","20,519 (26.7%)"
2011-1,N2,"19,203","17,957","9,057 (50.4%)","92,015","79,716","30,216 (37.9%)"
2011-1,N3,"5,642","5,211","2,511 (48.2%)","36,841","29,507","13,230 (44.8%)"
2011-1,N4,"3,643","3,358","1,431 (42.6%)","19,010","15,453","5,802 (37.5%)"
2011-1,N5,716,649,464 (71.5%),"12,346","10,510","6,108 (58.1%)"
2011-2,N1,"36,426","33,460","11,849 (35.4%)","100,873","88,450","26,715 (30.2%)"
2011-2,N2,"22,875","21,296","8,695 (40.8%)","94,538","82,944","28,679 (34.6%)"
2011-2,N3,"8,149","7,580","3,073 (40.5%)","49,917","41,655","16,576 (39.8%)"
2011-2,N4,"7,008","6,596","3,083 (46.7%)","38,888","33,402","14,722 (44.1%)"
2011-2,N5,"1,603","1,481","1,045 (70.6%)","33,245","29,159","16,986 (58.3%)"
2012-1,N1,"26,051","24,142","11,074 (45.9%)","78,904","69,082","23,789 (34.4%)"
2012-1,N2,"20,041","18,843","9,683 (51.4%)","78,553","69,418","29,191 (42.1%)"
2012-1,N3,"7,317","6,878","3,232 (47.0%)","38,650","31,942","14,391 (45.1%)"
2012-1,N4,"5,437","5,116","2,388 (46.7%)","22,431","18,590","8,489 (45.7%)"
2012-1,N5,"1,004",925,679 (73.4%),"16,361","13,911","8,129 (58.4%)"
2012-2,N1,"32,917","30,296","7,998 (26.4%)","86,004","75,250","17,411 (23.1%)"
2012-2,N2,"21,139","19,612","7,919 (40.4%)","79,513","69,790","25,617 (36.7%)"
2012-2,N3,"10,085","9,422","2,668 (28.3%)","47,301","39,763","12,722 (32.0%)"
2012-2,N4,"6,961","6,562","2,371 (36.1%)","36,799","31,620","11,783 (37.3%)"
2012-2,N5,"1,416","1,307",945 (72.3%),"34,178","29,700","16,225 (54.6%)"
2013-1,N1,"27,099","25,117","8,503 (33.9%)","74,674","65,225","20,139 (30.9%)"
2013-1,N2,"20,956","19,712","9,117 (46.3%)","73,729","64,885","29,725 (45.8%)"
2013-1,N3,"9,988","9,337","3,623 (38.8%)","39,870","32,895","13,063 (39.7%)"
2013-1,N4,"5,637","5,297","2,485 (46.9%)","23,746","19,941","9,823 (49.3%)"
2013-1,N5,"1,000",905,696 (76.9%),"18,720","16,016","9,957 (62.2%)"
2013-2,N1,"31,691","28,929","10,031 (34.7%)","81,794","71,490","25,524 (35.7%)"
2013-2,N2,"22,859","21,211","8,410 (39.6%)","73,935","64,989","28,148 (43.3%)"
2013-2,N3,"12,436","11,501","3,911 (34.0%)","48,875","41,129","17,901 (43.5%)"
2013-2,N4,"6,963","6,430","2,871 (44.7%)","38,078","32,752","14,290 (43.6%)"
2013-2,N5,"1,519","1,392",983 (70.6%),"37,313","31,922","18,248 (57.2%)"
2014-1,N1,"26,277","24,395","9,513 (39.0%)","73,782","64,409","21,108 (32.8%)"
```

Код программы:

```
jlpt_csv <- read.csv("JLPT.csv")
years <- subset(jlpt_csv, jlpt_csv$Level == "N1")$Year
applications <- as.integer(gsub(",", "", (c(subset(jlpt_csv, jlpt_csv$Level ==
"N1")$Applicants)))) # nolint
xx <- 1:23
line <- lm(applications ~ xx)
plot(applications, xlab = "Years",
    main = "Applications of JLPT N1 in Japan")
abline(line, col = "red")
axis(1, at = seq_along(applications), labels = years, las = 2)
print(summary(applications))
```

Результат работы скрипта:




```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 24716   30140   36426   38044   45831   58799
```

Выводы:

Я исследовал статистику заявок на JLPT N1 в Японии с помощью линейной регрессии и языка R.

Контрольные вопросы:

1. *Описать модель линейной регрессии и вывести соответствующие математические формулы. Приведите пример задачи, которую можно изучать с помощью предлагаемых методов.* Линейная регрессия некоторой зависимой переменной y на набор независимых переменных $x = (x_1, \ldots, x_r)$, где r – это число предсказателей, предполагает, что линейное отношение между y и x: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + \varepsilon$. Это уравнение регрессии. β_0, β_1, …, β_r – коэффициенты регрессии, и ε – случайная ошибка. Цель линейной регрессии заключается в нахождении оптимальных значений коэффициентов β_0 и β_1, которые минимизируют сумму квадратов ошибок (например, методом наименьших квадратов). Пример задачи, которую можно решить такими способами был у меня на физике. Я измерял растяжение пружины при различной массе грузов. После этого нужно было отметить эти точки на графике от массы и количества делений динамометра и провести прямую по этим точкам. Линейная регрессия дала бы мне наиболее оптимальную прямую.

2. *Откуда были взяты Ваши данные и что они означают?* Эти данные были взяты с Википедии. Они означают количество заявок на сдачу экзамена JLPT в Японии и за рубежом. Линейная регрессия применялась для того, чтобы исследовать динамику количества заявок.
3. *Какие трудности возникли у Вас при написании скрипта.* Мне было сложно понять для чего нужен R, если есть python. Для чтения CSV файла я использовал функций read.csv. Поскольку файл был взят с конвертера таблицы, пришлось немного его изменить, чтобы он нормально прочитался. Далее необходимо было преобразовать строки таблицы в числа. После этого без проблем получилось вычислить прямую функцей lm, вывести данные с помощью plot и нарисовать прямую с помощью abline. Так же я вывел в консоль summary вектора заявок.
4. *Дать возможные рекомендации для усовершенствования и усложнения изучаемой модели.* Я думаю, для этих данных подошла бы множественная логическая регрессия, поскольку в некоторых графах таблицы зависимость была нелинейная и не экспоненциальная.
5. *Рассмотреть в первом приближении более сложные алгоритмы машинного обучения (скажем методы наивного байесовского классификатора, опорных векторов, кластеризации и пр.). Здесь достаточно минимального ответа, особенно в случае, если Вы плохо поняли эту теорию.* Метод опорных векторов - набор схожих алгоритмов

обучения с учителем, использующихся для задач классификации и регрессионного анализа. Основная идея метода — перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с наибольшим зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы. Разделяющей гиперплоскостью будет гиперплоскость, создающая наибольшее расстояние до двух параллельных гиперплоскостей. Алгоритм основан на допущении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.