**Задание 9**

*Построение и применение*

*множественной логистической регрессионной модели*

Найдите реальные данные для задачи множественной бинарной классификации и выполните для них следующее:

1. Найдите коэффициенты логистической регрессионной модели.
2. Выполните прогноз переменной отклика для нескольких новых объектов и нескольких объектов обучающей выборки.
3. Сделайте выводы о качестве логистической регрессионной модели.
4. Результаты оформите в виде Word-файла и прикрепите его здесь (Весь R-код поместите в конец файла как приложение).

Мной были использованы те же данные, что и для выполнения задания 8: данные исследования 3,168 образцов голоса мужчин и женщин. В качестве переменной отклика также выберем пол человека.

Источник данных: <https://www.kaggle.com/primaryobjects/voicegender>.

Данные содержат следующие характеристики:

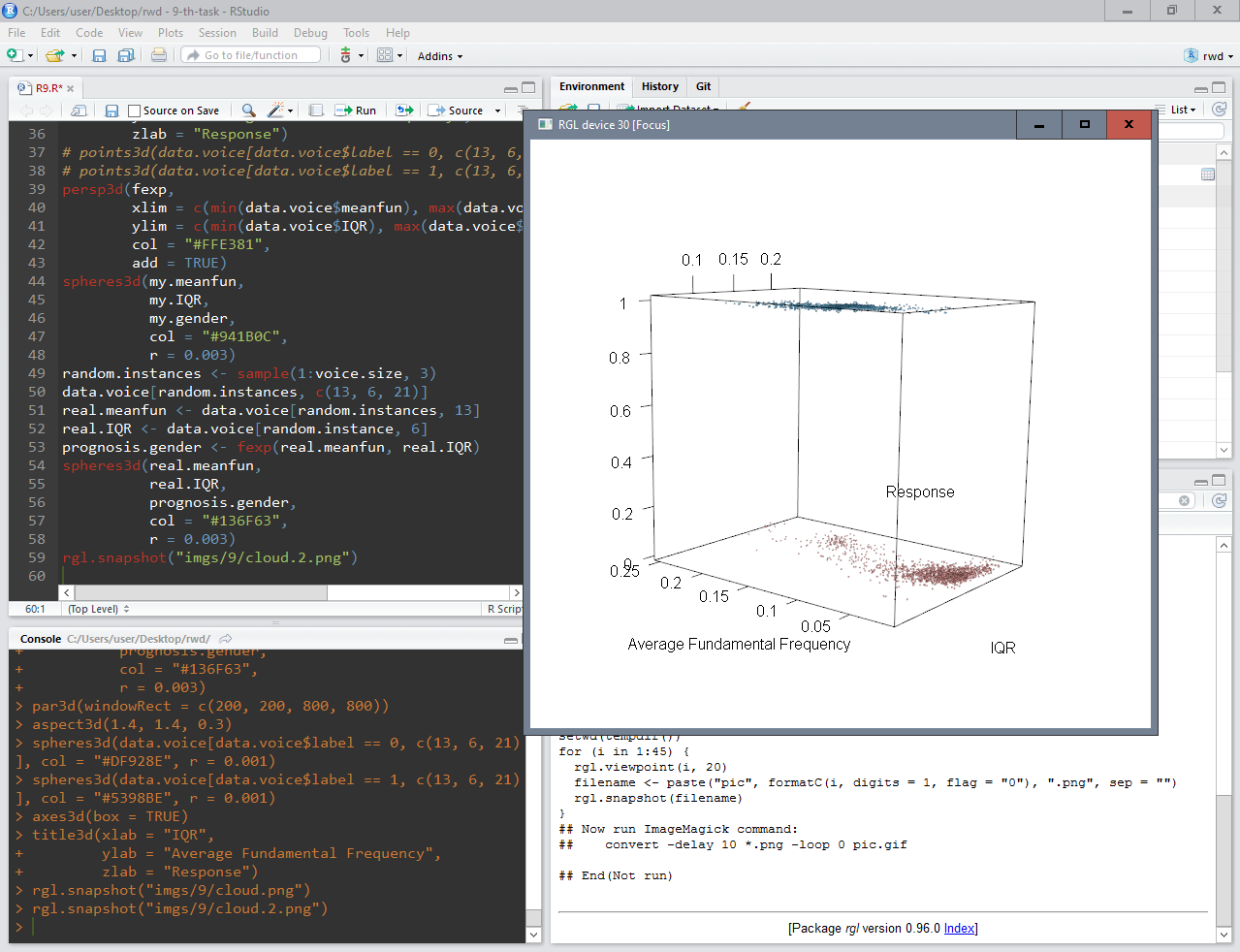
1. meanfreq — средняя частота звука, кГЦ;
2. sd — стандартное отклонение частоты;
3. median — медиана частоты, кГЦ;
4. Q25 — 25% квартиль, кГЦ;
5. Q75 — 75% квартиль, кГЦ;
6. IQR — интерквартильный размах, кГЦ;
7. skew — коэффициент ассиметрии;
8. kurt — коэффициент эксцесса;
9. sp.ent — спектральная энтропия;
10. sfm — спектральная плоскостность энтропия Винера;
11. mode — мода частоты;
12. centroid — центроид частоты;
13. meanfun — среднее основной частоты;
14. minfun — минимум основной частоты;
15. maxfun — максимум основной частоты;
16. meandom — среднее частоты волны самой высокой амплитуды;
17. mindom — минимум частоты волны самой высокой амплитуды;
18. maxdom — максимум частоты волны самой высокой амплитуды;
19. dfrange — разброс, девиация частоты волны самой высокой амплитуды;
20. modindx — индекс модуляции, отношение девиации частоты к частоте модулирующего сигнала;
21. label — переменная отклика, *“male”* — мужчина, *“female” —* женщина*.*

Теперь в качестве факторов будем использовать 2 характеристики: по предложению автора данных возьмем характеристики 13. и 6. — средняя основная частота и интерквартильный размах.

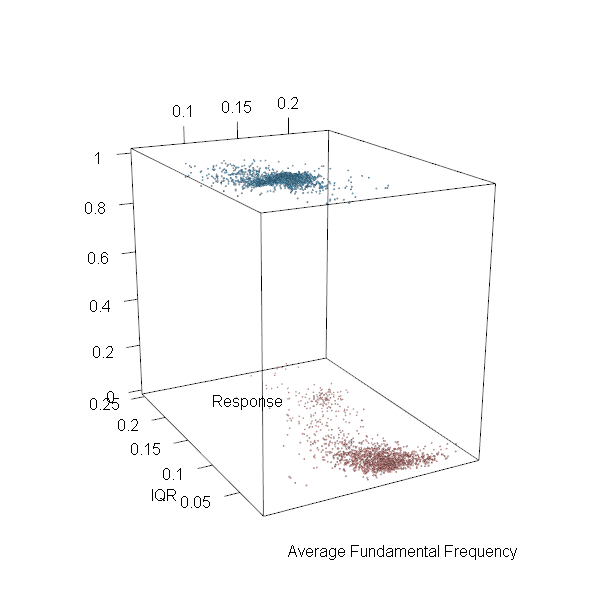
Так как характеристики всего две, мы можем наглядно представить наши данные на графике. Для трехмерных построений будем использовать пакет OpenGL для R — *“rgl”*.

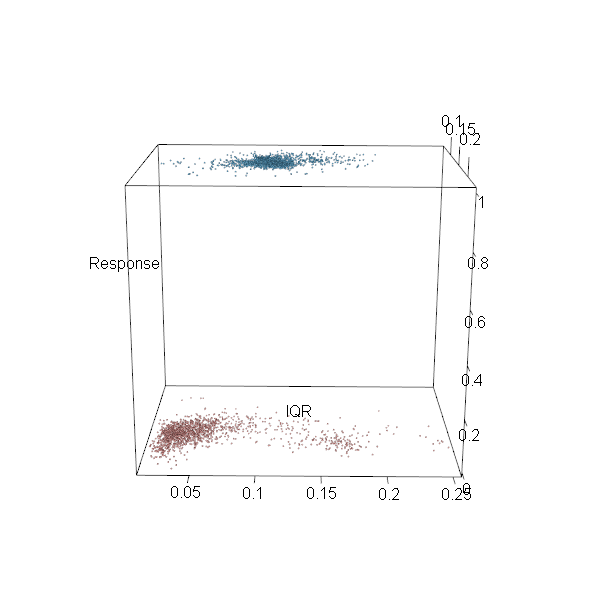
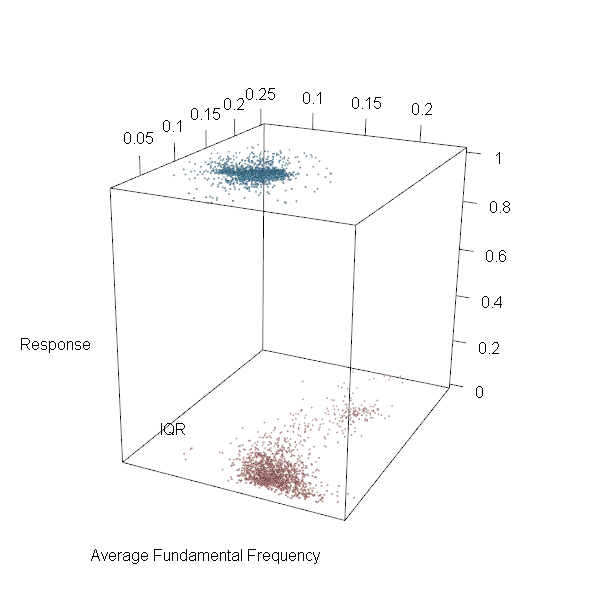
Построенные с помощью пакета графики открываются в отдельном окне, где пользователь может вращать его. Выбрав наиболее приятную точку обзора, с помощью функции rgl.snapshot() можно сохранить текущий видимый график.

Пример использования *rgl* для трехмерных построений:



Построим облако точек наших данных:





К сожалению, пакет, по всей видимости, не позволяет использовать различные кодировки в названиях осей: оставим оригинальные названия характеристик.

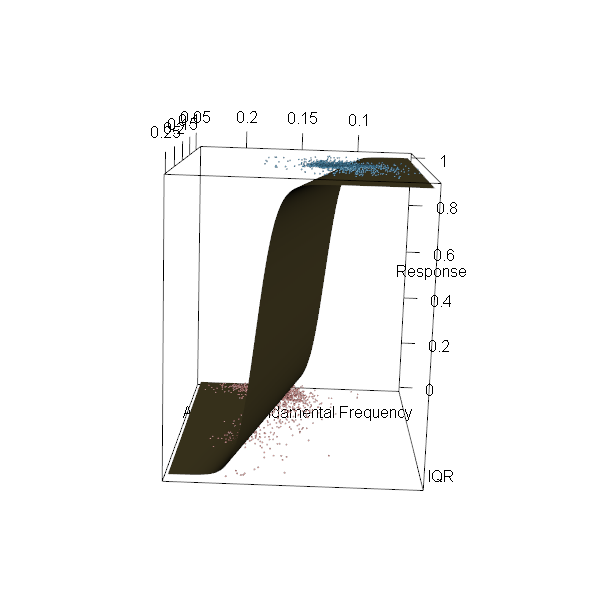
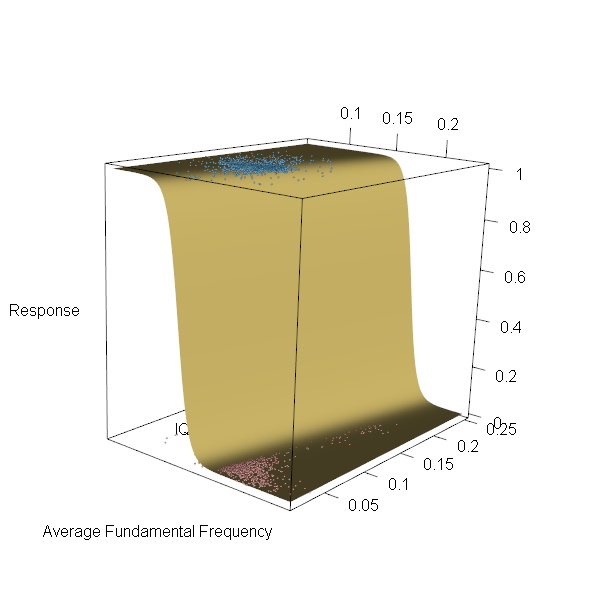
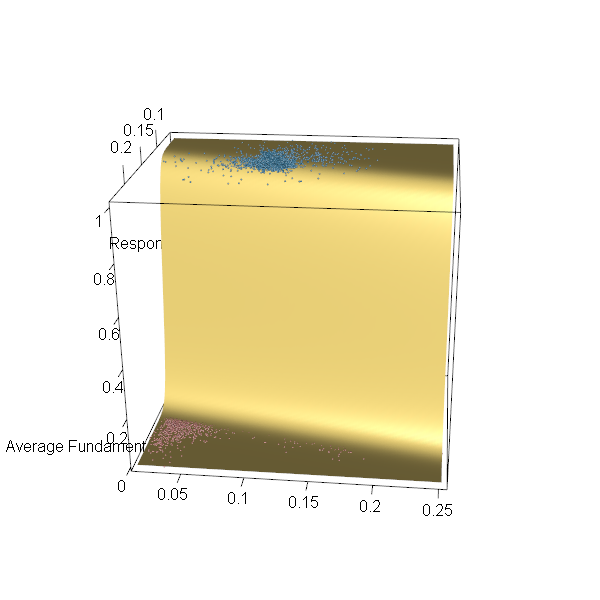
Для точек *голубого* цвета — мужчин, средняя основная частота ниже, чем для точек *розового* цвета — женщин, а интерквартильный размах выше.

Для отображения точек на графике мы предварительно заменили факторы переменной отклика *“male”* и *“female”* на 1 и 0 соответственно.

По методу максимизации функции максимального правдоподобия найдем коэффициенты для нашей модели:

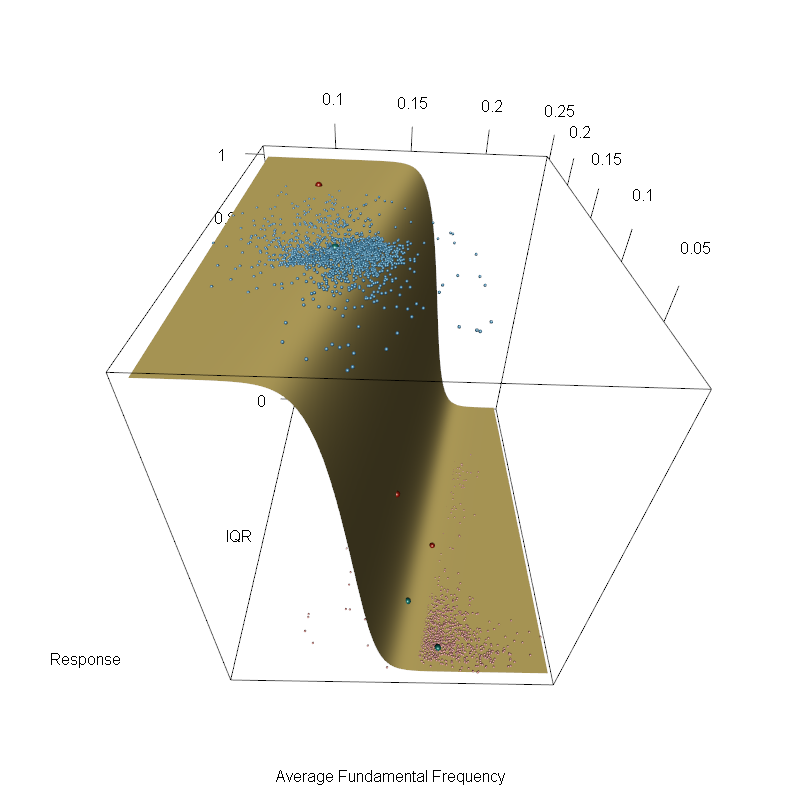
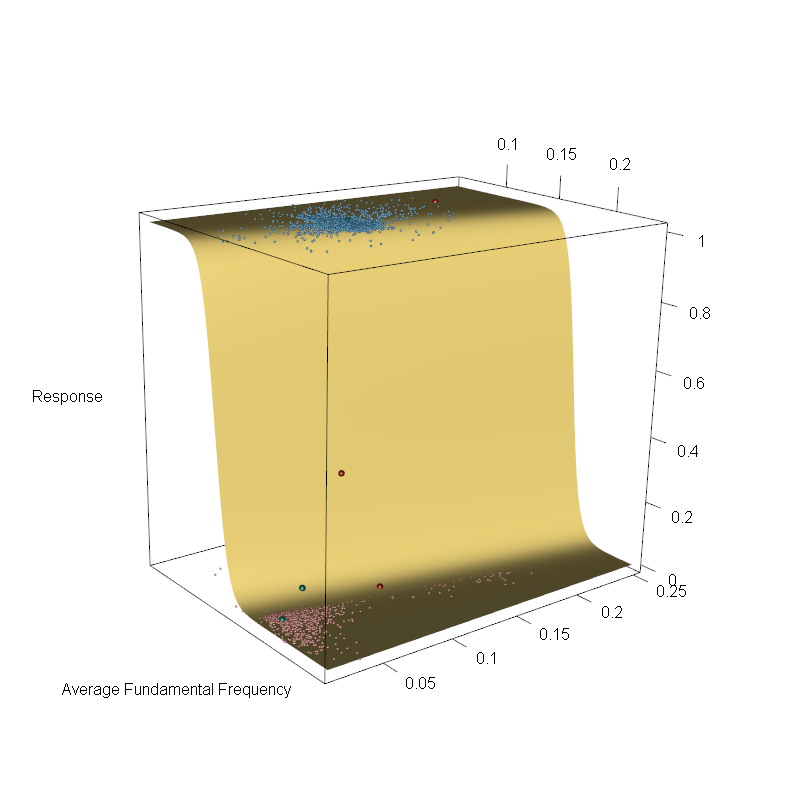
|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

Построим логистическую поверхность:



Нетрудно заметить, что на сцене с 3D-объектами присутствует источник света. Его местоположение также можно изменять и настраивать.

Построим прогноз для пользовательских данных и некоторых данных из обучающей выборки:



Точки *красного* цвета — пользовательские данные, *зеленого —* данные обучающей выборки. Выберем пороговое значение вероятности . Для всех объектов обучающей выборки модель сделала верный прогноз.

Нельзя, однако, одну из точек, заданных пользовательскими данными, отнести однозначно (в случае остальных точек визуально можно было определить класс объекта) к какому-либо классу (средняя основная частота — , интерквартильный размах — ). Согласно выбранного порогового значения вероятности отнесем данный объект к женщинам.

Таблицу с прогнозами можно найти [здесь](#_Таблица_прогнозов).

Приложение

# Скрипт программы на R

data.voice <- read.csv("data/voice.csv",

header = TRUE,

sep = ",")

data.voice <- na.omit(data.voice)

# перекодируем значение переменной отклика

# 0 — женщина, 1 — мужчина

levels(data.voice$label) <- c(0, 1)

data.voice$label <- as.numeric(levels(data.voice$label)[data.voice$label])

voice.size <- length(data.voice[, 1])

voice.regression <- glm(formula = label ~ meanfun + IQR,

family = binomial,

data = data.voice)

th <- coef(voice.regression)

fexp <- function(x, y) {

return (1 / (1 + exp(-(th[1] + th[2] \* x + th[3] \* y))))

}

th

# зададим пользовательские значения факторов

my.meanfun <- c(0.173, 0.095, 0.146)

my.IQR <- c(0.1, 0.2, 0.09)

my.prob <- fexp(my.meanfun, my.IQR)

voices.my <- **data.frame**(my.meanfun,

my.IQR,

my.prob)

library("rgl")

# зададим размеры окна графика

par3d(windowRect = c(200, 200, 1000, 1000))

# зададим соотношение осей

aspect3d(1.4, 1.4, 0.3)

# изобразим точки, на самом деле маленькие сферы радиуса 0.001

# класса женщин

spheres3d(data.voice[data.voice$label == 0, c(13, 6, 21)], col = "#DF928E", r = 0.001)

# изобразим точки, на самом деле маленькие сферы радиуса 0.001

# класса мужчин

spheres3d(data.voice[data.voice$label == 1, c(13, 6, 21)], col = "#5398BE", r = 0.001)

# отобразим оси и зададим им лейблы

axes3d(box = TRUE)

title3d(xlab = "Average Fundamental Frequency",

ylab = "IQR",

zlab = "Response")

# построим логистическую поверхность

persp3d(fexp,

xlim = c(min(data.voice$meanfun), max(data.voice$meanfun)),

ylim = c(min(data.voice$IQR), max(data.voice$IQR)),

col = "#FFE381",

add = TRUE)

# изобразим точки прогноза пользовательских значений

spheres3d(voices.my,

col = "#941B0C",

r = 0.003)

voices.my

# выберем случайны объекты из выборки

random.instances <- sample(1:voice.size, 3)

voice.prognosis <- data.voice[random.instances, c(13, 6, 21)]

voice.prognosis$prob <- fexp(voice.prognosis$meanfun,

voice.prognosis$IQR)

voice.prognosis

# изобразим точки прогноза значений из обучающей выборки

spheres3d(voice.prognosis$meanfun,

voice.prognosis$IQR,

voice.prognosis$prob,

col = "#136F63",

r = 0.003)

# ф-ия для сохранения скриншота текущего видимого графика

rgl.snapshot("imgs/9/logistic.prognosis.2.png")

# Таблица прогнозов

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *№* | *Основная частота* | *Интерквартильный размах* | *Вероятность* | *Прогноз класса* | *Истинный класса* |
| *1* | 0.173 | 0.10 | 0.006938688 | 0 |  |
| *2* | 0.095 | 0.20 | 0.999993221 | 1 |  |
| *3* | 0.146 | 0.09 | 0.320529374 | 0 |  |
| *4* | 0.1551075 | 0.05533597 | 0.0277005995 | 0 | 0 |
| *5* | 0.1122564 | 0.11703953 | 0.9974105484 | 1 | 1 |
| *6* | 0.1732985 | 0.02856536 | 0.0004916886 | 0 | 0 |