**Задание 8**

*Построение и применение однофакторной*

*логистической регрессионной модели*

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

1. Найдите коэффициенты логистической регрессионной модели.
2. Постройте логистическую кривую и облако точек.
3. Выполните прогноз переменной отклика для нескольких новых объектов и нескольких объектов обучающей выборки.
4. Проиллюстрируйте на графике результаты прогнозирования.
5. Сделайте выводы о качестве логистической регрессионной модели.
6. Результаты оформите в виде Word-файла и прикрепите его здесь (Весь R-код поместите в конец файла как приложение).

Были использованы данные исследования 3,168 образцов голоса мужчин и женщин. Сразу же скажем, что в качестве переменной отклика мы выберем пол человека.

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

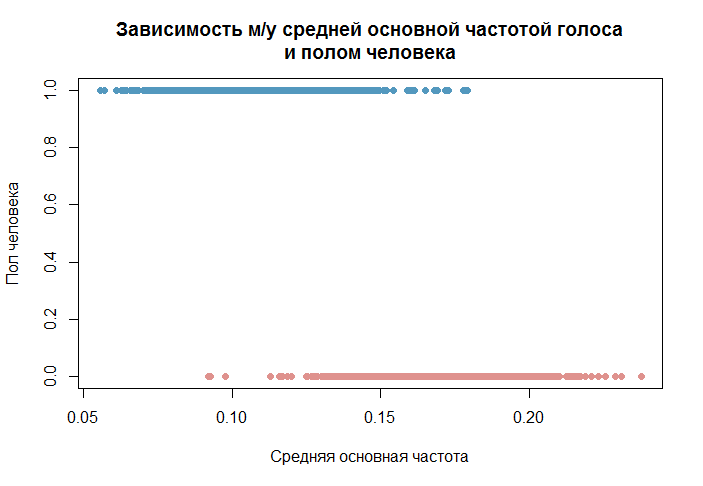
Данные содержат следующие характеристики[[1]](#footnote-1):

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*.*

В качестве фактора выберем характеристику 13. — основную частоту[[2]](#footnote-2).

Характеристика выбрана опытным путем: построив облако точек для характеристики, была предпринята попытка предугадать, как поведет себя логистическая кривая («распрямится» или же наоборот примет форму сигмоида).

В случае основной частоты точки красиво рассосредоточились[[3]](#footnote-3):



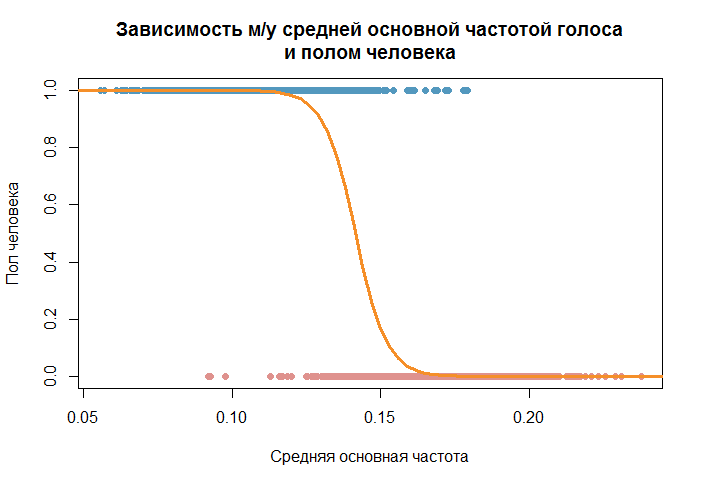
Для точек *голубого* цвета — мужчин, характерны частоты ниже, чем для точек *розового* цвета — женщин.

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 кривую:

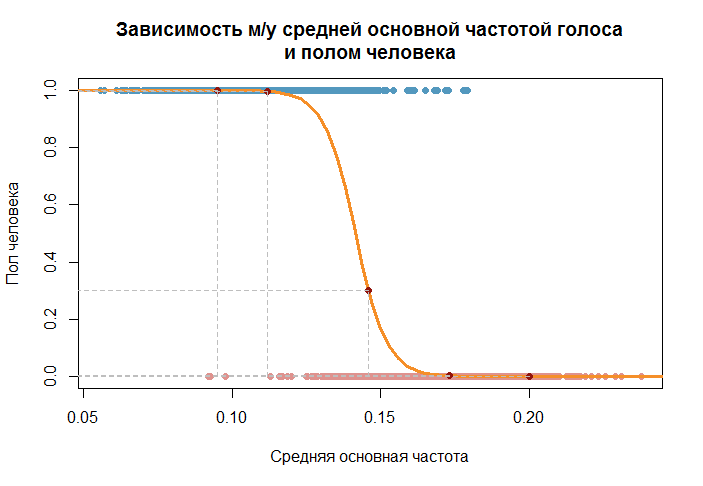


На графике хорошо видно, что пороговое значение основной частоты можно установить на уровне .

Пользовательские данные:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *№* | 1 | 2 | 3 | 4 | 5 |
| *Основная частота* | 0.173 | 0.2 | 0.095 | 0.112 | 0.146 |

Построим точки спрогнозированных значений переменной отклика:



Таким образом для установленного порогового значения, номера 3 и 4 окажутся мужчинами, а 1 и 2 — женщинами. Нельзя однозначно отнести номер 5 к какому либо классу, однако, согласно нашему пороговому значению, отнесем его к женщинам.

Как и указывалось авторами данных, величина основной частоты является определяющим фактором в классификации пола человека.

Приложение

# Скрипт программы на R

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

header = TRUE,

sep = ",")

# 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])

plot(data.voice$meanfun,

data.voice$label,

type="n",

main="Зависимость м/у средней основной частотой голоса\nи полом человека",

xlab = "Средняя основная частота",

ylab = "Пол человека")

for (i in 1:voice.size)

{

colour <- ifelse(data.voice$label[i] == 1,

"#5398BE",

"#DF928E")

points(data.voice$meanfun[i],

data.voice$label[i],

type="p",

pch=16,

col=colour,

add="TRUE")

}

voice.regression <- glm(formula = data.voice$label ~ data.voice$meanfun,

family = binomial,

data = data.voice)

th <- coef(voice.regression)

fexp <- function(x){ return(1 / (1 + exp(-(th[1] + th[2] \* x))))}

th

curve(fexp(x), 0, 0.3, col="#F58F29", lwd=3, add=TRUE)

my.meanfun <- c(0.173, 0.2, 0.095, 0.112, 0.146)

my.gender <- fexp(my.meanfun)

points(my.meanfun,

my.gender,

type="p",

pch=16,

col="#92140C",

add="TRUE")

for(i in 1:length(my.meanfun))

{

segments(my.meanfun[i], 0, my.meanfun[i], my.gender[i], col="gray", lty = 2, add="TRUE")

segments(my.meanfun[i], my.gender[i], 0, my.gender[i], col="gray", lty = 2, add="TRUE")

}

1. К сожалению, мне не удалось точно описать все характеристики данных в виду некомпетентности в вопросах физики и звуковых волн. [↑](#footnote-ref-1)
2. Термин «основная частота» спорный и найден мной на просторах Интернета. Оригинальное название характеристики — **“fundamental frequency”,** подробную информацию можно найти по [ссылке](https://en.wikipedia.org/wiki/Fundamental_frequency). [↑](#footnote-ref-2)
3. Уже после выполнения задания я обнаружил, что в источнике данных говорилось об «основной частоте» как об определяющем факторе пола человека. Также по [ссылке](http://www.primaryobjects.com/2016/06/22/identifying-the-gender-of-a-voice-using-machine-learning/) можно найти исследование определения пола человека по характеристикам его голоса от автора данных. [↑](#footnote-ref-3)