Федоров Макар

Описание данных

В файле `diabetes.csv` содержатся показатели здоровья женщин индейского племени Пима, предоставленные Национальным институтом по изучению диабета, расстройств пищеварительной системы и почек (США). Известно, что среди представителей этого племени был зафиксирован самый высокий процент заболеваемости диабетом второго типа в мире.

Показатели в файле:

- `Pregnancies`: количество беременностей;
- `Glucose`: уровень глюкозы в плазме крови;
- `BloodPressure`: диастолическое (нижнее) кровяное давление в миллиметрах ртутного столба.
- `SkinThickness`: толщина кожной складки над трицепсом в миллиметрах;
- `Insulin`: уровень инсулина после двухчасовой углеводной нагрузки;
- `BMI`: индекс массы тела;
- `DiabetesPedigreeFunction: наследственная склонность к диабету;
- `Age`: возраст;
- `Outcome`: индикатор того, болеет ли человек диабетом или нет.

Задача 1

Загрузите данные из файла `diabetes.csv` и сохраните их в датафрейм `diabet`. Выведите описательные статистики по всем столбцам датафрейма. Если среди описательных статистик встречаются заведомо невозможные значения (например, давление, равное 0), удалите соответствующие им строки из датафрейма.

```
diabet <- read_csv("diabetes2.csv")
```

```
##
## ── Column specification ────────────────────────────────────────────────────────
## cols(
##   Pregnancies = col_double(),
##   Glucose = col_double(),
##   BloodPressure = col_double(),
##   SkinThickness = col_double(),
##   Insulin = col_double(),
##   BMI = col_double(),
##   DiabetesPedigreeFunction = col_double(),
##   Age = col_double(),
##   Outcome = col_double()
## )
```

```
summary(diabet)
```

```
##   Pregnancies        Glucose      BloodPressure    SkinThickness
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     Insulin           BMI        DiabetesPedigreeFunction      Age
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780           Min.   :21.00
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437           1st Qu.:24.00
##  Median : 30.5   Median :32.00   Median :0.3725           Median :29.00
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719           Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200           Max.   :81.00
##     Outcome
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.349
##  3rd Qu.:1.000
##  Max.   :1.000
```

```
diabet %>%
  filter(Insulin != 0) %>%
  filter(BloodPressure != 0) %>%
  filter(SkinThickness != 0) %>%
  filter(BMI != 0) %>%
  filter(Glucose != 0) -> diabet2
```

Мы удалили заведомо неправдоподобные значения, когда инсулин, кровяное давление, толщина кожи или глюкоза были равны нулю (одно из этих значений или любая их комбинация).

Задача 2

Постройте регрессионную модель, которая объясняет, каким образом заболеваемость диабетом зависит от уровня глюкозы в крови, кровяного давления, уровня инсулина, индекса массы тела и возраста человека. Выведите описание этой модели.

a. Какие переменные в модели оказались статистически значимыми? Укажите их.

b. Используя полученные в модели коэффициенты, объясните:

- как в среднем изменяются шансы человека заболеть диабетом при увеличении индекса массы тела на единицу;
- как в среднем изменяются шансы человека заболеть диабетом при увеличении нижнего давления при увеличении индекса массы тела на единицу.

```
logit_model <- glm(data = diabet2, Outcome ~ Glucose + BloodPressure +
Insulin + BMI + Age, family = "binomial")

summary(logit_model)

##
## Call:
## glm(formula = Outcome ~ Glucose + BloodPressure + Insulin + BMI +
##     Age, family = "binomial", data = diabet2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6242  -0.6774  -0.3827   0.6760   2.4729
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -9.7208551  1.1674617  -8.326  < 2e-16 ***
## Glucose        0.0379725  0.0057177   6.641 3.11e-11 ***
## BloodPressure -0.0027008  0.0114185  -0.237 0.813025
## Insulin       -0.0007491  0.0012984  -0.577 0.563979
## BMI            0.0812731  0.0212529   3.824 0.000131 ***
## Age            0.0550397  0.0138266   3.981 6.87e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 353.99  on 386  degrees of freedom
## AIC: 365.99
##
## Number of Fisher Scoring iterations: 5

exp(coef(logit_model))

##   (Intercept)       Glucose BloodPressure       Insulin           BMI
##  6.001866e-05  1.038703e+00  9.973029e-01  9.992512e-01  1.084667e+00
##           Age
##  1.056583e+00
```

Глюкоза, возраст и индекс массы тела оказались значимыми при уровне доверия < 0.001. Кровяное давление и инсулин оказались статистически незначимыми (при уровне значимости < 0.1). При увеличении индекса массы тела на одну единицу шансы заболеть диабетом увеличиваются в среднем примерно в 1.084 в раз. При увеличении кровяного давления на одну единицу, шансы заболеть диабетом увеличиваются примерно в 9.973 раз. В то же время, кровяное давление не является статистически значимым (p-value = 0.813025)

Задача 3

Посчитайте как минимум два показателя качества полученной модели. Используя полученные результаты, сделайте выводы о качестве модели.

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
roc(diabet2$Outcome, logit_model$fitted.values)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
##
## Call:
## roc.default(response = diabet2$Outcome, predictor =
logit_model$fitted.values)
##
## Data: logit_model$fitted.values in 262 controls (diabet2$Outcome 0) < 130
cases (diabet2$Outcome 1).
## Area under the curve: 0.8485
```

```
install.packages("pscl")
```

```
## Installing package into
'/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
```

```
library(pscl)
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
```

```
pR2(logit_model)
```

```
## fitting null model for pseudo-r2
```

```
##          llh      llhNull           G2     McFadden         r2ML
r2CU
```

```
## -176.9962819 -249.0489027  144.1052414    0.2893111    0.3076166
0.4276295
```

Во-первых, мы использовали ROC кривую. Площадь под кривой = 0.8485, что показывает хорошую работу модели. Во-вторых, мы используем McFadden's pseudo R^2 тест, который показывает сравнение данной модели с моделью, использующей только "Intercept" для предсказание. Чем ближе значения к единице, тем лучше (значения от 0.2 уже являются хорошими). Показатели McFadden's pseudo R^2 = 0.2893111, что говорит о хорошей работе модели