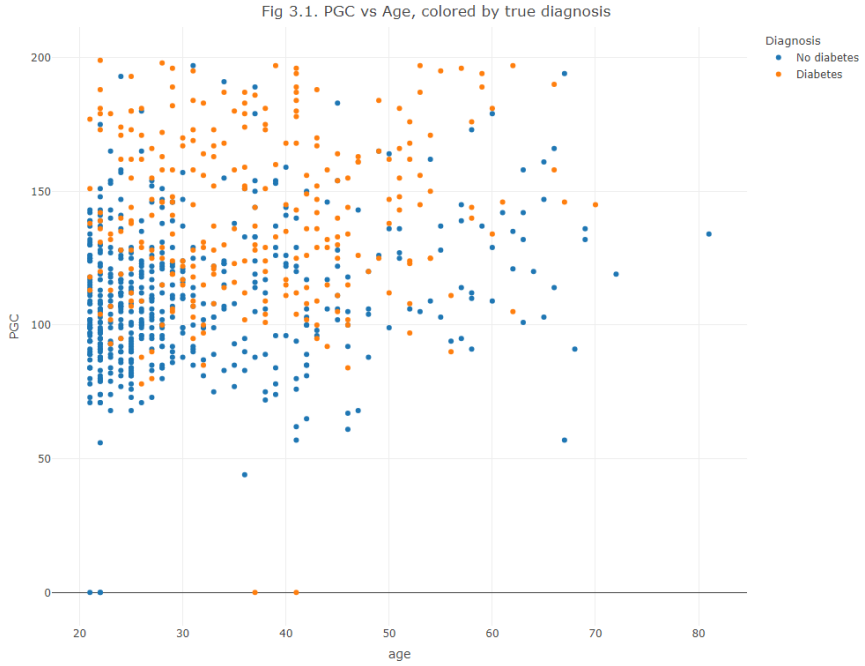
# Sim_notebook

Simon Jorstedt

2023-11-19

## Assignment 3

We are provided data covering the onset of diabetes within a five year period for a group of individuals. The data consists of nine variables including a binary response variable indicating diagnosis: presence of diabetes or not. In Figure 3.1, we plot Plasma Glucose Concentration (PGC) against age in years. Datapoints are colored by diagnosis.




Analysing Figure 3.1 we observe a large cluster of young people (ages ~20-30) that do not have diabetes, along with a significant number of outliers (among the non-diabetes people.) The people with diabetes however are much more spread out, with no clear clusters. It appears as though people with diabetes tend to have slightly larger Plasma Glucose Concentration (PGC) values than people without diabetes. Thus it does appear as though there is some explanatory power in the PGC values and Age, but it is likely not enough to achieve a highly accurate logistic regression (classification) predictor.

### Assignment 3.2 and 3.3

We will now fit a logistic regression (classification) model using the PGC and age features to predict the presence of diabetes. See *Machine Learning - A first course for engineers and scientists* (pp. 45-52) for a

discussion on logistic regression. We will initially use a classification threshold of $r = 0.5$. Mathematically, our model predictor $g(\boldsymbol{x})$ can be represented in the following way, where $\hat{y}(\boldsymbol{x}) = 1$ indicates presence of diabetes, and $\hat{y}(\boldsymbol{x} = 0)$ indicates absence.

$$\hat{y}(\boldsymbol{x}) = \begin{Bmatrix} 1 & \text{if} & g(\boldsymbol{x}) \geq 0.5 \\ 0 & \text{if} & g(\boldsymbol{x}) < 0.5 \end{Bmatrix}$$

where

$$g(\boldsymbol{x}) = \frac{1}{1 + e^{-z}}$$

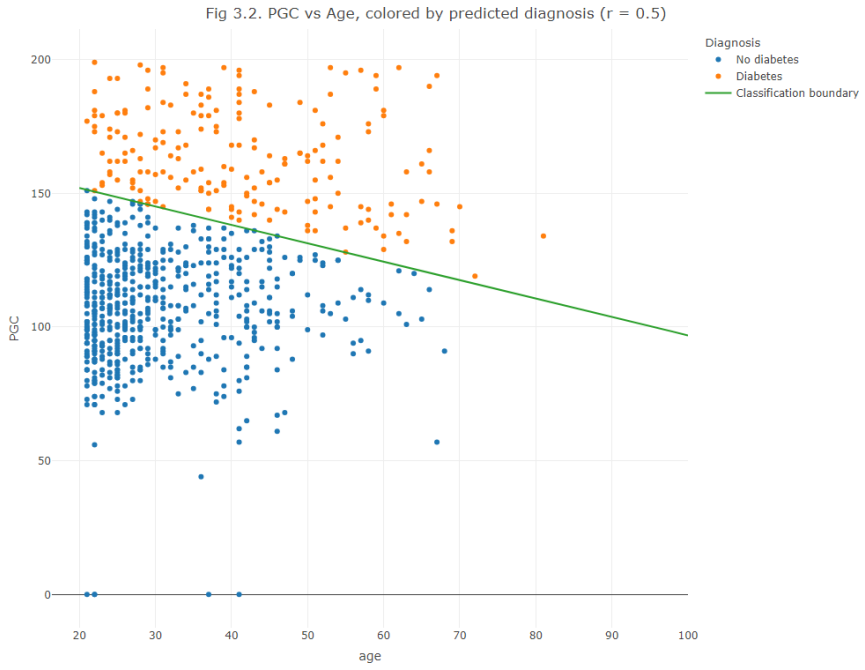
where

$$z = \boldsymbol{\theta}^T \boldsymbol{x} = (\theta_0, \theta_1, \theta_2) \cdot (1, x_{\text{pg}}, x_{\text{age}})^T.$$

The decision boundary for this model will be the set of points that satisfy the following equation:

$$g(\boldsymbol{x}) = \frac{e^{\boldsymbol{\theta}^T \boldsymbol{x}}}{1 + e^{\boldsymbol{\theta}^T \boldsymbol{x}}} = \frac{1}{2} \implies \boldsymbol{\theta}^T \boldsymbol{x} = 0 \implies \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 = 0 \implies x_2 = \frac{-\theta_0}{\theta_2} - \frac{\theta_1}{\theta_2} x_1.$$

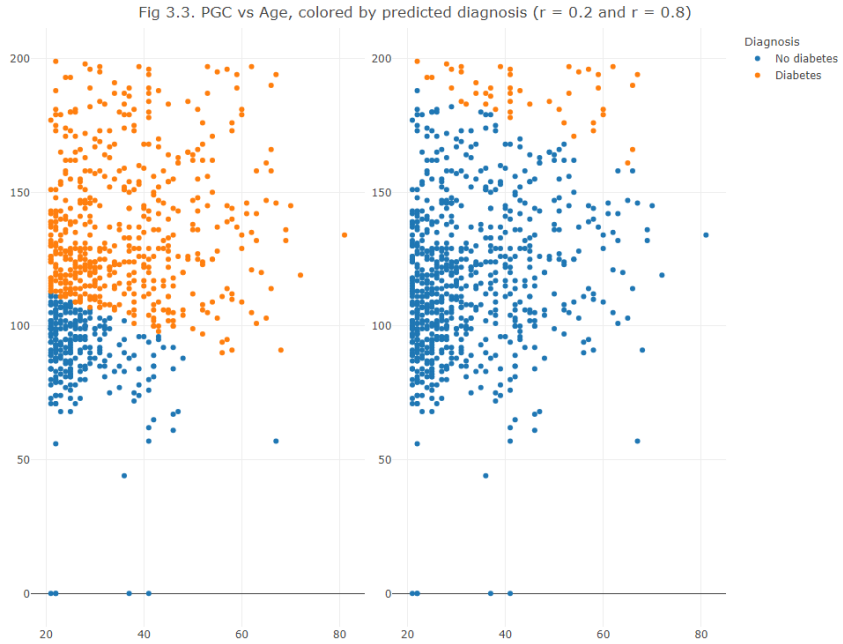
In Figure 3.2 below, we plot again PGC values against Age, but datapoints are colored by predicted diagnosis (using $r = 0.5$). The discussed linear decision boundary of the model is also included.




When comparing Figure 3.1 and 3.2 we see that the quality of the classification is decent at best. The resulting predictions are visually precisely what could be expected when analysing Figure 3.1. The model essentially implements what was discussed previously: diabetes is common among people with high PGC values. The missclassification error 0.266 reflects our model assessment, as it is good, but far from excellent. The decision boundary echoes this, as it clearly divides the data into high and low PGC values, but it also increases the "diabetes region" as Age increases.

## Assignment 3.4

In Figure 3.3 we plot again PGC against Age values, but this time color points based on classification thresholds $r = 0.2$ and $r = 0.8$. To clarify, this is *almost* the same model as discussed previously,




Comment on Figure 3.3.

*More complicated boundary.* When it comes to the expanded model, the decision boundary can be found as the solution to the same kind of equation as the simpler model, but the equation is now more complicated. It can however be solved numerically using the R `polyroot` function. The equation is

$$\theta_0 + \theta_1 x_1 + \theta_2 \cdot x_2 + \theta_3 \cdot x_1^4 + \theta_4 \cdot x_1^3 x_2 + \theta_5 \cdot x_1^2 x_2^2 + \theta_6 \cdot x_1 x_2^3 + \theta_7 \cdot x_2^4 = 0$$

## 3.5

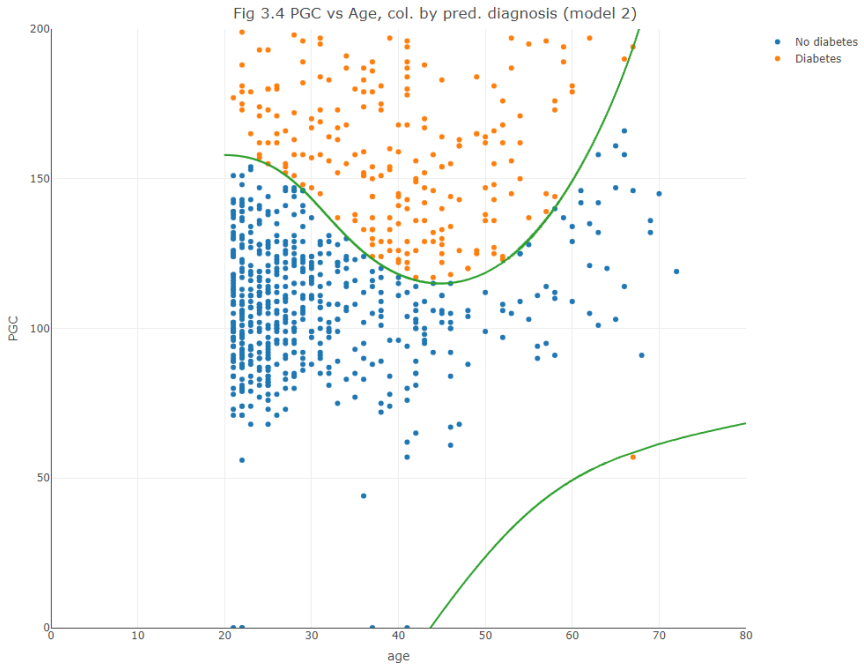
```r
# Create new variables z_1, ... z_5
diab_data <- diab_data %>%
  mutate(z1 = PGC^4 * age^0,
         z2 = PGC^3 * age^1,
         z3 = PGC^2 * age^2,
         z4 = PGC^1 * age^3,
         z5 = PGC^0 * age^4)

# New model
model2 <- glm(formula = diabetes ~ PGC + age + z1 + z2 + z3 + z4 + z5,
    family = binomial(link = "logit"),
    data = diab_data)

summary(model2)
```

```
##
## Call:
## glm(formula = diabetes ~ PGC + age + z1 + z2 + z3 + z4 + z5,
##     family = binomial(link = "logit"), data = diab_data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.279e+00  1.129e+00  -8.217  < 2e-16 ***
## PGC          3.772e-02  9.473e-03   3.981 6.85e-05 ***
## age          1.453e-01  2.072e-02   7.014 2.32e-12 ***
## z1           1.266e-08  5.610e-09   2.257  0.02402 *
## z2          -1.760e-07  7.638e-08  -2.304  0.02122 *
## z3           8.424e-07  3.439e-07   2.450  0.01430 *
## z4          -1.682e-06  6.317e-07  -2.662  0.00776 **
## z5           8.045e-07  4.056e-07   1.983  0.04732 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 991.38  on 766  degrees of freedom
## Residual deviance: 741.07  on 759  degrees of freedom
## AIC: 757.07
##
## Number of Fisher Scoring iterations: 5
```

```r
diab_data_newmodel <- diab_data %>%
  mutate(predicted_z_05 = as.integer(fitted(model2) >= 0.5))
```



## Extra / old code

$$z = \boldsymbol{\theta}^T \boldsymbol{x} = (-5.8979, 0.0356, 0.0245) \cdot (1, x_{\text{pg}}, x_{\text{age}})^T$$