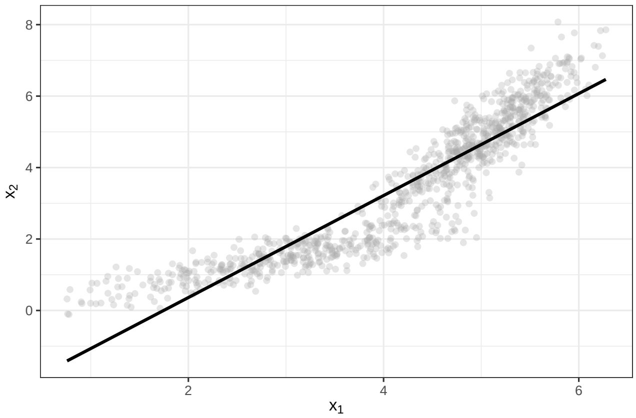
# 1 Predicting tree biomass

Estimating the biomass of trees is essential for assessing forest carbon stocks, but direct measurement is destructive and labor-intensive. Since tree diameter at breast height (DBH) is easy to record and closely related to total wood mass, it serves as a key variable for predicting biomass. Consider the following data on above-ground tree `biomass` ($x_2$) in $t$ and trunk `DBH` ($x_1$) in $m$.

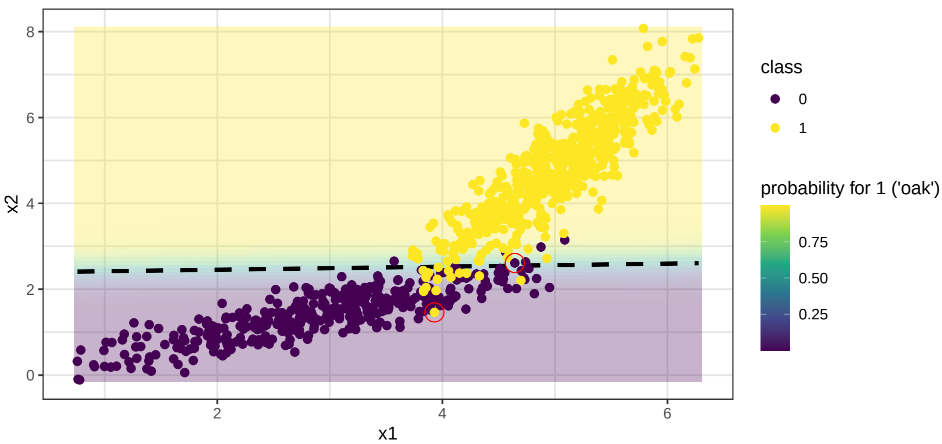
a) Explain which variable plays the role of *feature* and *target*, respectively. Assume we obtain the following linear model. Estimate the coefficient associated with $x_1$ from the plot and interpret it.




b) It seems that the model does not fit the data too well. What is the model's tendency for observations with large $x_1$ value? Come up with a suitable transformation to $x_1$ that might improve the model fit. Describe how the transformed point cloud and model will look. How do the computation of the $x_1$ coefficient and its interpretation change?

# 2 Predicting tree species

a) Assume now that we want to use both tree `DBH` ($x_1$) and `biomass` ($x_2$) to predict a third variable, tree `species` ($y$; 0 = "beech", 1 = "oak"). A *logistic regression* model yields the *decision boundary* pictured below (dashed line). What can you say for the respective values of the training loss function at the two highlighted points?



b) Use the parameters of the decision boundary (intercept $a = 2.38$, slope $b = 0.04$) to derive a decision rule for classifying a tree. The rule should be of the following form, where the conditions depend on $x_1$ and $x_2$:

$$y = \begin{cases} \text{“oak”} & \text{if} \ldots \\ \text{“beech”} & \text{if} \ldots \end{cases}$$

c) Focusing on a small region near the decision boundary, how would you classify the highlighted point (red diamond) if you were to use *k-nearest neighbors* with Euclidean distance and $k = 3$? What if $k = 5$?