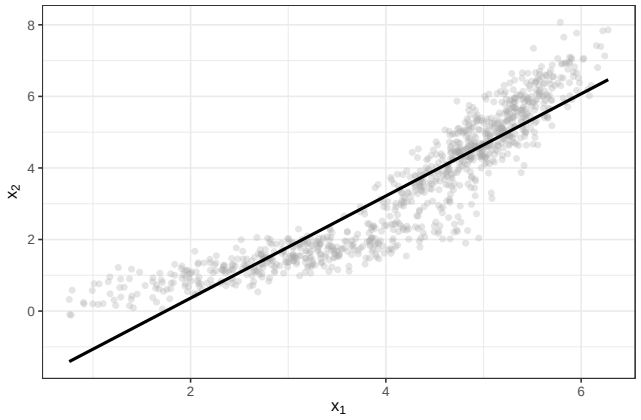
1 Predicting tree biomass

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.



Solution.

- Feature: tree diameter (x_1); target: biomass (x_2); task: regression.
- The fitted slope is roughly 1.5: increasing diameter by 1 m raises expected biomass by about 1.5 t. This works as a simple explanatory or coarse predictive baseline.

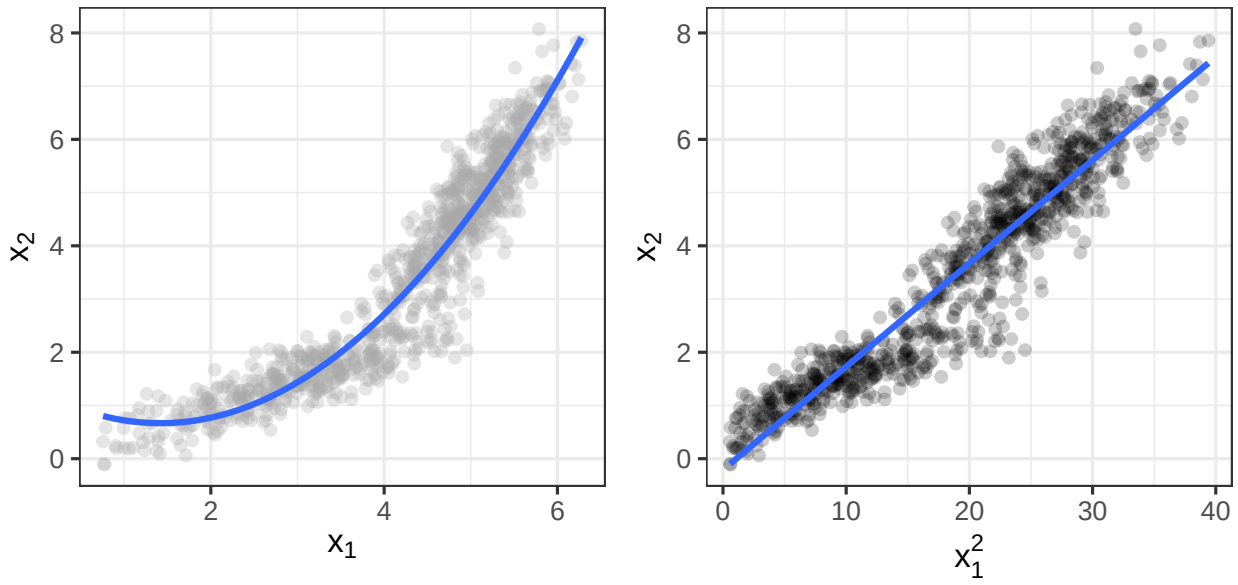
---

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?

Solution.

- The linear model underestimates the target for large x_1 (negative residuals on the right).
- Transforming the predictor helps; e.g., use x_1^2. The scatter of (x_1^2, x_2) is much closer to linear (see additional code), so the line fits high-diameter trees better.
- Estimation stays ordinary least squares but on the transformed predictor; the slope now quantifies the expected change in biomass per unit increase in x_1^2.

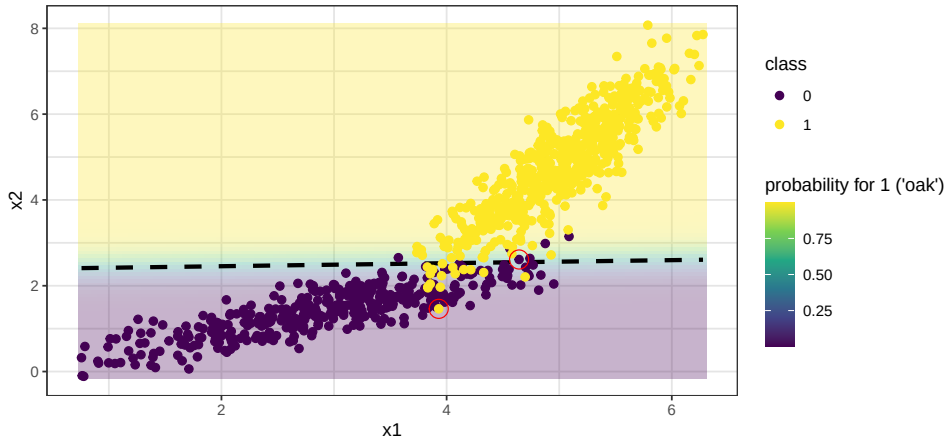
```
## ADDITIONAL CODE FOR IN-CLASS DEMO -------------------------------------
```



---

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?



Solution.

- Both highlighted points are misclassified, so each has positive logistic loss. Recall that the *logistic loss* for a point with true label $y \in \{0, 1\}$ and predicted probability p for class 1 is given by:

$$L(y, p) = -\left[y \log(p) + (1 - y) \log(1 - p)\right]$$

  For a misclassified point, the predicted probability p for the true class is less than 0.5, hence the loss is greater than $-\log(0.5) \approx 0.693$, and increases the more confidently incorrect the prediction is. The logistic loss penalizes not just errors but especially confident misclassifications.

  Mini calculation examples:

– *Misclassified true 0*: Suppose $y = 0$ (true class "beech") and the model predicts $p = 0.6$ for class 1. The loss is:

$$L(0, 0.6) = -\left[0 \cdot \log(0.6) + 1 \cdot \log(1 - 0.6)\right] = -\log(0.4) \approx 0.916$$

– *Misclassified true 1*: Suppose $y = 1$ (true class "oak") and the model predicts $p = 0.3$ for class 1. The loss is:

$$L(1, 0.3) = -\left[1 \cdot \log(0.3) + 0 \cdot \log(0.7)\right] = -\log(0.3) \approx 1.204$$

In both cases, the loss is large because the model was confidently wrong.

Thus, both highlighted points contribute positively to the total loss, with loss increasing the more confidently an observation is misclassified.

- The class-1 point (oak) lies farther from the boundary, so its predicted probability for class 1 is even lower, making its contribution to the loss larger than that of the barely misclassified class-0 point (beech).

---

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:

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

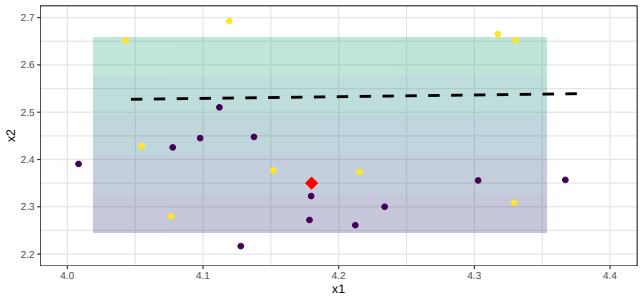
Solution.

- From the boundary line $x_2 = b\,x_1 + a$, predict "oak" if $x_2 > b\,x_1 + a$ and "beech" otherwise.
- Equivalently, use the sign test $x_2 - b\,x_1 - a > 0$ for "oak" (this coincides with $\pi(\mathbf{x} \mid \boldsymbol{\theta}) > 0.5$ for class "oak").
- In compact form, the rule can be written as

$$y = \begin{cases} \text{"oak"} & \text{if } x_2 > b\,x_1 + a, \\ \text{"beech"} & \text{otherwise.} \end{cases}$$

---

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 $k = 5$?



Solution.

- For $k = 3$, the three nearest neighbors contain two oaks and one beech, so the majority vote classifies the point as oak.
- For $k = 5$, the five nearest neighbors contain three beeches and two oaks, so the majority vote classifies the point as beech.