🔷 Problem: Rare events and biased sampling

When an event (like “default” or “disease”) is rare, a normal training sample might contain very few positive cases.
To make modeling possible, researchers often oversample the rare cases — for example, making a training set where 40% of samples are “rare events.”

But this causes bias:

- The model learns a distribution that doesn’t match the real-world frequency.
- The predicted probabilities become too large, since the training set exaggerates how common the event is.

———

🔷 Solution: Adjust the intercept (β_0)

The slope coefficients (β_i) remain correct because they describe relationships (e.g., how cholesterol affects risk), but the intercept (β_0) must be corrected to restore the true base rate of the event.

The correction formula is:

$$\beta_0^* = \beta_0 + \log\left(\frac{p_{\text{rare}}}{1 - p_{\text{rare}}}\right) - \log\left(\frac{p_{\text{set}}}{1 - p_{\text{set}}}\right)$$

Where:

- p_{rare}: true probability of the rare event in the population
- p_{set}: observed probability of the rare event in the training set

This adjusts only the intercept so that predicted probabilities reflect real-world rates again.

———

🔷 Control vs. Case sample size

- To reduce variance in estimates, we use more control (non-event) samples than cases.
- A control-to-case ratio of 5–6 to 1 is usually enough; adding more controls beyond that brings little benefit.

Example:

- 100 “case” patients (disease)
- 500–600 “control” patients (no disease) output again

Here’s the polished explanation again, formatted for clarity and completeness:

🔷 Problem: Rare Events and Biased Sampling

When an event (like **default** or **disease**) is rare, a typical training sample may contain very few positive cases.
To make modeling feasible, researchers often **oversample** the rare cases — for instance, creating a training set where **40%** of the samples are rare events.

However, this introduces **bias**:

- The model learns a **distorted distribution** that doesn’t match real-world frequencies.
- The **predicted probabilities** become **too high**, because the model “thinks” the event is much more common than it actually is.

🔷 Solution: Adjust the Intercept (β_0)

The **slope coefficients (β_i)** remain valid because they describe **relationships** between predictors and the outcome (e.g., how cholesterol affects disease risk).
But the **intercept (β_0)** must be adjusted to restore the **true base rate** of the event in the population.

✅ Intercept correction formula

$$\beta_0^* = \beta_0 + \log\left(\frac{p_{\text{rare}}}{1 - p_{\text{rare}}}\right) - \log\left(\frac{p_{\text{set}}}{1 - p_{\text{set}}}\right)$$

Where:

- p_{rare}: true probability of the rare event in the population
- p_{set}: observed probability of the rare event in the (biased) training set

This adjustment recalibrates the **predicted probabilities** so they match real-world frequencies.

🔷 Control vs. Case Sample Size

To reduce **variance** in the estimates:

- Use **more controls (non-events)** than cases.
- A **control-to-case ratio** of **5–6 to 1** usually provides good efficiency.
  Adding more controls beyond that yields **diminishing returns**.

Example
If you have:

- 100 "case" patients (with disease)
- 500–600 "control" patients (without disease)

→ That's typically sufficient for stable model estimation.

Would you like me to show an **R example** that demonstrates oversampling bias and how to apply this intercept correction in practice?