Very Large Constants, Coefficients, OR's or RR's or IRR's

Possible causes and solutions

Andy Grogan-Kaylor

2025-05-27

Table of contents

One sometimes sees *extremely large β coefficients, OR's or RR's or IRR's*.

1. Is this due to very low *base rates* of one category of the outcome? Those can inflate coefficients, RR's, OR's, and IRR's.
2. What are the *units of the independent variables*? For example if your independent variable is a percentage expressed on a 0.0 to 1.0 scale, the scale of your independent variable can radically inflate your estimates for β. Re-scaling the independent variable sometimes helps.
3. Should some of your *continuous independent variables be centered*? Should you *collapse categories for some of your categorical independent variables*?

> 💡 **What Does β_0 Mean?**
>
> Remember that the β_0 term is the mean outcome for an individual with a score of 0 on all of the independent variables. This may sometimes provide reasonable estimates, but it may sometimes be necessary to center some or all of your independent variables. If independent variables are not centered, β_0 may represent the value of the outcome for a combination of independent variables well outside of the data range. See the figure below for an illustration of this idea.

4. Are there interactions (e.g. $x \times z$) in your model? Remember that when there is an interaction (e.g. $x \times z$) in a model, the coefficient for the main effect (e.g. βx) represents the coefficient for individuals for whom z is 0. Perhaps z should be centered? (Remember also that interactions in nonlinear models are *complicated* and require *special consideration*, an issue discussed elsewhere on this site.)
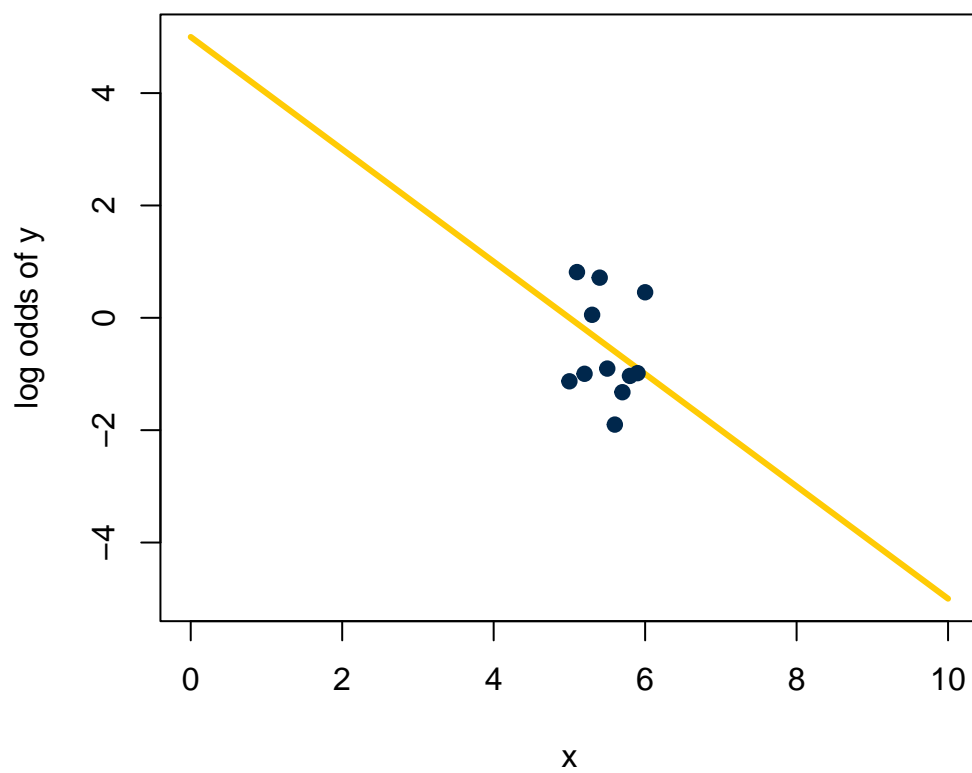
Figure 1: intercept outside range of observed data

5. Is this an issue of *complete, or near complete separation*? i.e. one variable almost perfectly predicts the outcome? That too can lead to very large OR's or RR's.
6. Are there *very small cell (subsample) sizes*? Small sample sizes (cell sizes) can create anomalous estimates i.e. estimates from small samples (small table cells) can be very *noisy* estimates.

There may be other possibilities, but if one of the above strategies does not solve the problem, I would suggest deleting the variable from your model if possible.