Adding Non-Linearity To The Right Hand Side Of An Equation for Categorical Data

true

2021-12-16

Contents

1 Introduction

Logistic regression models the *log odds* of an outcome as a function of a set of covariates:

$$\text{ln}\left(\frac{p(\text{outcome})}{1-p(\text{outcome})}\right)=\beta_0+\beta_1x_1+\beta_2x_2$$

Here $p(\text{outcome})$ is the probability of the outcome.

$\frac{p(\text{outcome})}{1-p(\text{outcome})}$ is the *odds* of the outcome.

Hence, ln $\left(\frac{p(\text{outcome})}{1-p(\text{outcome})}\right)$ is the *log odds*.

> It is plausible to think about adding non-linear functions of the covariates–e.g. ln(x), x^2–to the right hand side of our logistic regression equation.

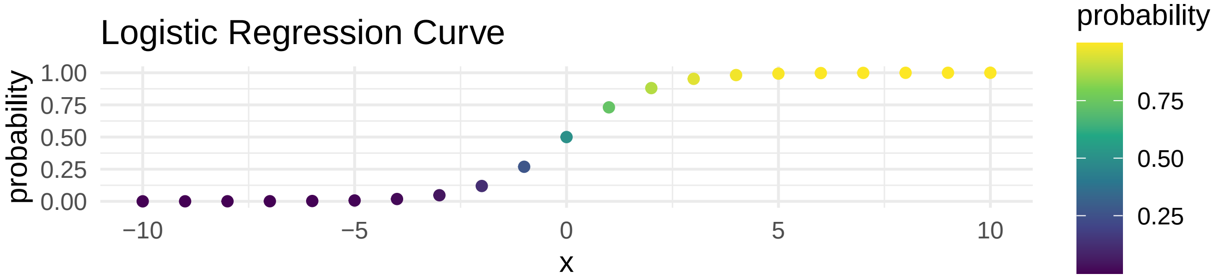
2 Consider Again The Equation For Logistic Regression

$$\text{ln}\left(\frac{p(\text{outcome})}{1-p(\text{outcome})}\right)=\beta_0+\beta_1x_1+\beta_2x_2$$

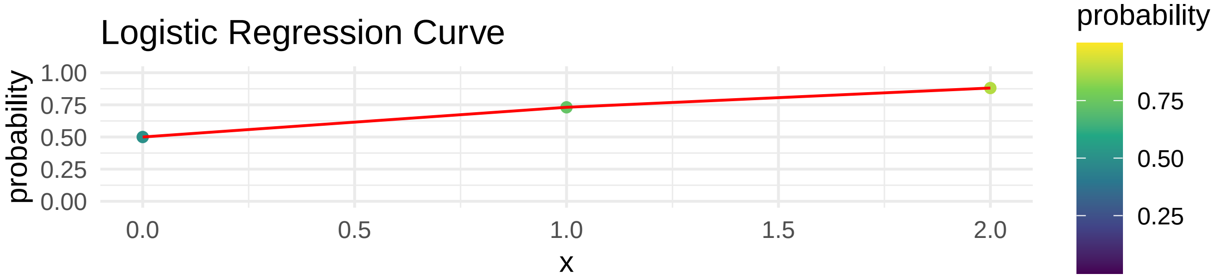
> A logistic regression is **already** a **non-linear** model because of the transformed y variable.
>
> A logistic regression creates a **non-linear** model of probabilities by being a **linear** model of the log-odds.

3 Visual Considerations

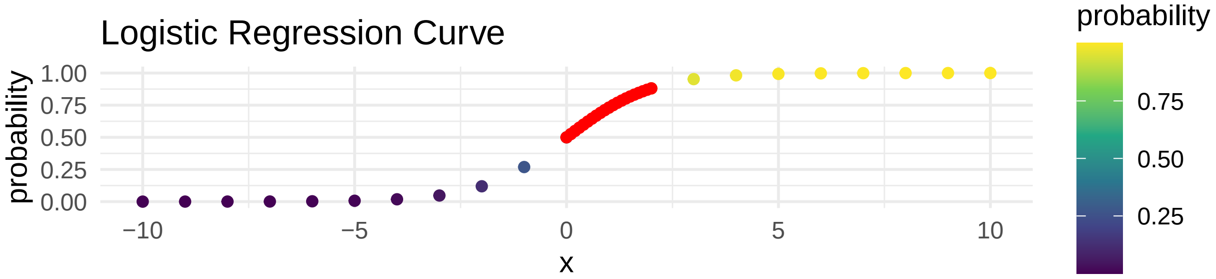
Plotting a logistic regression curve helps us to see the non-linearity of the equation.



It may sometimes appear that the plotted curve is linear.



But this is only a result of the fact that we are only using a portion of the logistic regression curve for a particular analysis.



4 Conclusion

The basic logistic regression equation is:

$$\ln\left(\frac{p(\text{outcome})}{1-p(\text{outcome})}\right) = \beta_0+\beta_1x_1+\beta_2x_2$$

The model is already *non-linear* because of the *transformed y variable*. We can certainly add non-linear terms to the right hand side of the model (e.g. x^2) but this will add non-linearity **on top of the already existing non-linearity** that is due to the *transformed dependent variable*.

We may indeed find that these non-linear terms on the right hand side of the equation are statistically significant, but will need to think carefully about the conceptual and substantive implications of the model given the *potentially multiple layers of non-linearity*.