Modelling Followback Probability

Bryan Xu

November 10, 2021

1 Introduction

The underlying question is simple: given some factors that we can calculate before following someone (i.e. number of followers, mutual followers, age of account), how can we extrapolate the chance that said person will follow us back? We can formulate this in two ways:

- A binary classification question: simply output yes or no as to whether or not this person will follow us back.
- A probability question: output the probability that this person will follow us back.

Each has its pros and cons, and as such we can implement both methods to test our hypotheses.

2 A Primer on Regression and Prediction

2.1 Simple Linear Regression

The most simple model for regression is formulated the following way: let Y be our *response variable* (i.e. our variable of interest) that depends on X, known as the *predictor variables*. We can write

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon,$$

where β_i are regression coefficients and ϵ is an "error" term known as *random noise*. Linear is a bit of a misnomer, as the linearity assumption only holds for the *parameters* and not necessarily the *variables*. As such,

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots \beta_p x^p + \epsilon$$

is still a *linear* model.

What assumptions must we follow for this model to be valid, and which ones may be broken in our model?

- Our predictors are fixed and not random variables. This is ok.
- Linearity of parameters. This is not hard to achieve because we can arbitrarily transform predictor variables to fit this.
- Constant variance, meaning that the variance of the errors does not depend on the values of the predictor variables. This will be an issue - we shouldn't have the same error for a person with 2 followers and 3 following and a person with 2000 followers and 3000 following.
- Assumptions about error ($\mathbb{E}[\epsilon_i] = 0$, $\text{Var}[\epsilon_i] = \sigma^2$, $\epsilon_i \perp \epsilon_j$). Hmm... these are most likely satisfied (hopefully lol)

2.2 Uses

While seemingly simple, this model is extremely robust and can be applied to a variety of situations. We use regression to solve the following problems:

(i) **Inference**: establishing relationships between the dependent and independent variables. This can involve fitting estimates for a response: given n data points with an estimator

$$\hat{Y} = x\hat{\beta},$$

inference estimates existing observations: $\hat{Y}_1, \ldots, \hat{Y}_n$.

(ii) **Prediction**: predict the variable of interest given new values of the other variables. In the previous example, this would mean estimating new observations: $\hat{Y}_{n+1}, \ldots, \hat{Y}_N$.

We have two ways to interpret prediction:

- The estimate of the average value of the response: $\mathbb{E}[Y_{n+1}]$.
- The estimate of a specific value of the response: Y_{n+1}.

We almost always work with the former case, as the latter is not very relevant in our case. We shouldn't care enough to try an estimate a specific case - plus it is also much harder to find a specific value without some sort of machine learning.

2.3 Multiple Linear Regression

We can generalize simple linear regression to support multiple independent variables. In this case

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i.$$

This can be further generalized to *multivariable linear regression*, which incorporates correlations between predictors. This is given by

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{i1} + \beta_{2j} X_{2i} + \cdots + \beta_{pj} X_{ip} + \epsilon_{ij}.$$

2.4 General Linear Models

What if instead we had Y not be a scalar, but a vector? Then we would use a general linear model:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{U}.$$

This will probably not prove to be useful in our case, as we only need a scalar/binary data.

3 Generalized Linear Models (GLM)

Although named very similarly, this is a distinct model from the above. GLMs help model response variables that are bounded or discrete. This is especially useful for us in the binary classification question, as we only have binary choices (yes or no).

Generalized linear models have three components:

(i) Linear predictor

(ii) Link function

(iii) Probability distribution

Linear regression is a special case of GLMs, with the identity link function and the normal distribution as the probability distribution. The classic example is the Poisson regression model, formulated as such:

$$\begin{aligned} \underbrace{\ln}_{\text{link function}} \quad \lambda_i &= \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}_{\text{linear predictor}} \\ Y_i &\sim \underbrace{\text{Poisson}}_{\text{probability distribution}} (\lambda_i) \end{aligned}$$

However this case is not very relevant to our studies. We will tackle the class of *logistic regression models*.

3.1 Logistic Regression

Consider the following model:

$$\begin{aligned} z_i &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \\ q_i &= \frac{1}{1 + \exp(-z_i)} \\ Y_i &\sim \text{Bernoulli}(q_i) \end{aligned}$$

4 Estimation

The theory of estimation is vast and not very interesting. The following as all that you need to know: suppose that we have a random variable X following some probability distribution $p(\theta)$ with unknown parameter θ; that is, $X \sim p(\theta)$. We want to "estimate" the true value of θ. The following are desirable properties of estimators.

4.1 Properties of Estimators

Definition 1 (Bias)**.** Let θ be our target parameter and let $\hat{\theta}$ be an estimator of θ. Then the *bias* of θ is defined as

$$\mathcal{B}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta.$$

If $\mathcal{B}(\hat{\theta}) = 0$, then we call $\hat{\theta}$ *unbiased*.

Obviously we want an unbiased estimator otherwise we run the risk of extremely poor results.

Definition 2 (Mean-Squared Error)**.** Let θ be our target parameter and let $\hat{\theta}$ be an estimator of θ. Then the *mean-squared error* of θ is defined as

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta}(X) - \theta)^2].$$

Again, we should try and get the lowest amount of mean-squared error.

4.2 Least Squares

Many types of least-squares estimators exist: OLS, WLS, GLS... honestly just use statsmodels.

4.3 Maximum Likelihood Estimation

Do our errors follow a probability distribution? Hmm...

4.4 Bayesian Regression

I don't think we will be using this.

4.5 Mixed Models

Are our variables correlated? Almost certainly. I will look into this.