Notes on Linear Regression

Earl Patrick B. Macalam

December 11, 2020

1 Why Not Linear Regression?

Note here that our response variable takes on categories.

1. Coding used in linear regression implies an ordering on the outcomes.
2. If the response variable's values did take on a natural ordering, such as mild, moderate, and severe, and we felt the gap between mild and moderate was similar to the gap between moderate and severe, then a 1, 2, 3 coding would be reasonable.

2 Logistic Regression

- Logistic regression models the probability that Y belongs to a particular category.

$$Pr(default = Yes|Balance)$$

for any given value of balance, a prediction can be made for default. For example, one might predict default = Yes for any individual for whom $p(balance) > 0.5$, or to be conservative, $p(balance) > 0.1$.

3 The Logistic Model

$$p(X) = \frac{\exp^{\beta_0+\beta_1 X}}{1 + \exp^{\beta_0+\beta_1 X}}.$$

By taking the log-odds we can have similar interpretation as the linear regression.

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

- To fit the model we use a method called maximum likelihood. The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows: we seek estimates for β_0 and β_1 such that the predicted probability $\hat{p}(x_i)$ of default for each individual corresponds as closely as possible to the individual's observed default status.

4 Interpretation on Coefficients and p-value

- We see that $\beta_1 = 0.0055$; this indicates that an increase in balance is associated with an increase in the probability of default. To be precise, a one-unit increase in balance is associated with an increase in the log odds of default by 0.0055 units.
- Since the p-value associated with balance is tiny, we can reject H_0. In other words, we conclude that there is indeed an association between balance and probability of default.
- The estimated intercept is typically not of interest; its main purpose is to adjust the average fitted probabilities to the proportion of ones in the data.

5 Making Predictions

- After obtaining the β estimates, we can make predictions by plugging the estimates to our model along with its predictor value.

$$\hat{p}(X) = \frac{\exp^{\beta_0+\beta_1 X}}{1+\exp^{\beta_0+\beta_1 X}} = \frac{e^{-10.6513+0.0055\cdot 1000}}{1+e^{-10.6513+0.0055\cdot 1000}} = 0.00576$$

  we predict that the default probability for an individual with a balance of \$1000 is 0.00576.

- One can use qualitative predictors with the logistic regression model using the dummy variable approach. As an example, the Default data set contains the qualitative variable student. To fit the model we simply create a dummy variable that takes on a value of 1 for students and 0 for non-students.

6 Multiple Logistic Regression

- Have multiple predictors.

$$p(X) = \frac{\exp^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p}}{1+\exp^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p}}.$$

- When the coefficient for the variable is negative, this might indicate less likely to chances to default or something.
- MLE is use for estimation.
- Similar way of making predictions as to simplee logistic regression.