STA 4241 Lecture, Week 3

September 9th, 2021

Overview of what we will cover

- Classification approaches
    - Logistic regression

Classification

- In the previous lecture we focused on predicting a quantitative response variable given a set of predictors
- Classification is the term used for when we want to predict a qualitative or categorical response
- Classification is not that different from regression methods that predict a quantitative response
    - First predict the probability of being in each class
        - Similar to predicting quantitative response
    - Then classify based on this probability

Classification

- The setup will be similar to before
- We observe $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$
- $\boldsymbol{X}_i$ represents a p-dimensional set of predictors for subject i
- We want to predict Y_i with as little error as possible
  - Especially focused on predicting outcomes for a test set
- Use the training data to build a classifier that predicts Y_i given $\boldsymbol{X}_i$

Introduction to GLMs

- Generalized linear models are used to relate a response variable Y to a set of predictors $\boldsymbol{X}$
  - Understand associations between predictors and outcome
  - Predict the outcome for given predictor levels
- Linear regression is actually a special case of GLMs
- GLMs are a broad class of models that can handle a variety of outcome types

Introduction to GLMs

- A GLM has three key components
  1. Random component
  2. Systematic component
  3. Link function
- There are a number of options for each component and we must choose one for each
- Thankfully there are standard choices for the random and link components

Random component

- The random component specifies the probability distribution for Y
- The three most common choices are
  - Normal distribution for continuous data
  - Binomial distribution for binary data
  - Poisson distribution for count data
- Many other probability distributions work with GLMs as well
- These three are by far the most commonly used
  - Adequate for most situations encountered

Systematic component

- This component specifies how the explanatory variables are related to the outcome
- Typically we use a linear function, such as

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- The above quantity is sometimes also known as the linear predictor
- This is the same as the manner in which we specify $f(X)$ in linear regression
    - Same issues must be addressed such as linearity, additivity, overfitting, etc.

Link function

- The link function specifies the functional relationship between the linear predictor and the mean of the outcome

- If we let $\mu = E(Y|X)$, then we specify

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- $g()$ is the link function used

- Each type of data has a corresponding link function that is most commonly used

Link function

- For the normal distribution, the canonical link is the identity link
  - Ordinary linear regression

$$g(\mu) = \mu = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- For the Poisson distribution, the canonical link is the log link

$$g(\mu) = \log(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- For the bernoulli distribution, the canonical link is called the logit or logistic link

$$g(\mu) = log\left[\frac{\mu}{1-\mu}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

GLMs for binary data

- Let's focus for now on binary outcomes
- The most common classification problem is one in which the outcome only takes two classes
- We will then see extensions to GLMs for multiple classes
- Will also cover classification methods that are not GLMs

GLMs for binary data

- Let's consider the idea of using the identity link (linear regression) for a binary outcome with a single covariate
- The identity link might work well for some X values
- We may obtain probabilities outside of 0 or 1 for some X values
  - Clearly incorrect
- Intuitively the identity link doesn't make a lot of sense for binary data for another reason
- We don't expect a change in X to have the same impact on $P(Y = 1)$ when $P(Y = 1)$ is close to zero or 1 vs. when $P(Y = 1)$ is close to 0.5

GLMs for binary data

- Due to these two issues, other link functions are more common
  - Logit link
  - Probit link
- Let's look at the logit link first

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

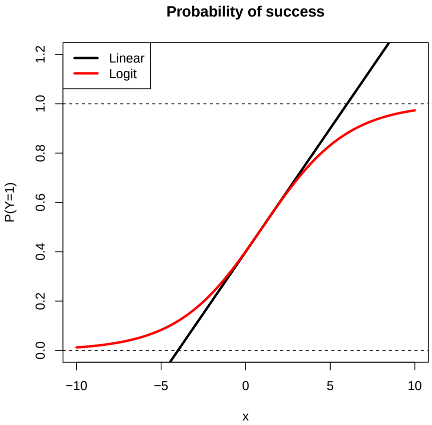
- Note these probabilities are constrained to lie in $(0, 1)$

GLMs for binary data

- GLMs with binary data and the logit link are called logistic regression
- Some key points about logistic regression
  - The parameter β_1 determines the rate of increase or decrease of the curve relating X to $p(X)$
  - When $\beta_1 > 0$, $p(X)$ increases as X increases
  - When $\beta_1 < 0$, $p(X)$ decreases as X increases
- How does this compare to the linear link?

GLMs for binary data

- We see the logit link respects the $(0, 1)$ boundary lines
- Changes in X have smaller effects when $P(Y = 1)$ is closer to 0 or 1

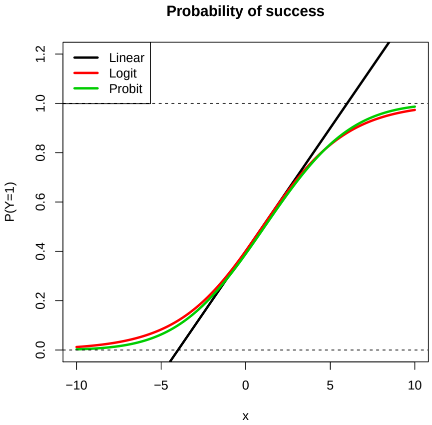


GLMs for binary data

- An alternative link is the probit link
- If we let $\Phi()$ be the CDF of a standard normal distribution, then the link function is
$$g(\cdot) \equiv \Phi^{-1}(\cdot)$$
which implies
$$p(X) = \Phi(\beta_0 + \beta_1 X)$$
- Note this again ensures that the probability is inside $(0, 1)$

GLMs for binary data

- We see that the probit looks similar to the logit
- Typically give similar answers



GLMs for binary data

- Let's look at an example using data from the Challenger disaster in 1986
- Is the temperature associated with the probability of a failure of at least one primary O-ring?

Flight	Temperature	Failure
1	66	0
2	70	1
3	69	0
4	68	0
5	67	0
6	72	0
⋮	⋮	⋮

GLMs for binary data

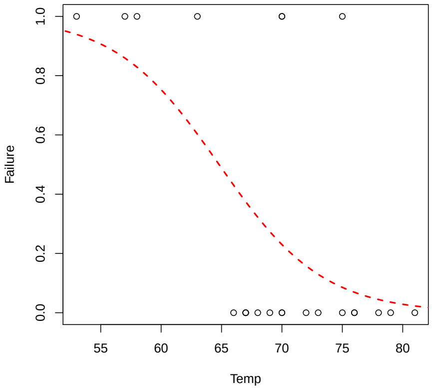
- If we fit a logistic regression model we estimate

$$\begin{aligned}\log\left[\frac{p(X)}{1-p(X)}\right] &= \widehat{\beta}_0 + \widehat{\beta}_1 \times \text{temperature} \\ &= 15.04 - 0.23 \times \text{temperature}\end{aligned}$$

- Lower temperatures seem to increase the probability of failure
- The p-value for the test of $H_0 : \beta_1 = 0$ is 0.03

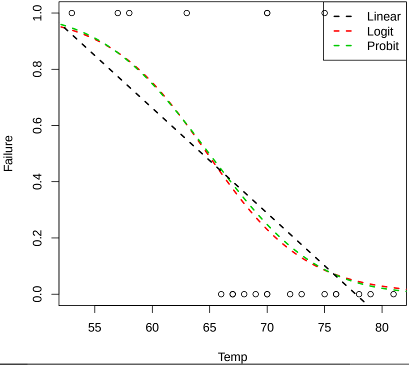
GLMs for binary data

- Let's look at the predicted probabilities of failure
- Our model predicts the probability of failure was over 0.99 for the temperature found on the day of the disaster (36 degrees)



GLMs for binary data

- Let's try to fit the probit and linear models as well
- Very similar results between probit and logit
- Linear does ok, but predicts that the probability the challenger would fail to be > 1



GLMs for binary data

- Be careful about extrapolating results to new input values that were not observed in the data
- The model may fit your observed data well, but not necessarily at predictor values outside the range of your data
- The challenger data point was 15 degrees lower than any point in our observed data

Logistic regression

- Logistic regression is far more commonly used in applied research than probit regression
- This is mostly due to interpretation, not due to the ability to predict outcomes better
- We will focus on logistic regression here, but all ideas apply directly to probit regression
  - With the exception of interpretation!

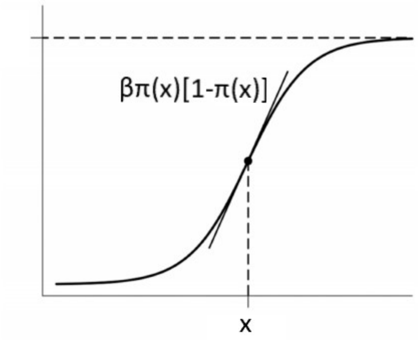
Logistic regression

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

- β_0 represents the log odds of success for subjects with $X = 0$
- e^{β_0} is the odds of success for subjects with $X = 0$
- β_1 represents the strength of association between X and Y
    - $\beta_1 > 0$, then $p(X) \uparrow$ as $X \uparrow$
    - $\beta_1 < 0$, then $p(X) \downarrow$ as $X \uparrow$
    - $\beta_1 = 0$, then $p(X) = e^{\beta_0}/(1 + e^{\beta_0})$ which is a constant, with $p(X) > 0.5$ when $\beta_0 > 0$

Logistic regression

- By taking derivatives, we can see that the rate of change in $p(X)$ is $\beta_1 p(X)(1 - p(X))$
- Note that this rate of change is maximized when $p(X) = 0.5$



Rate of change

Logistic regression

- β_1 has a nice interpretation in terms of odds ratios
  - Main reason why the logit link is preferred over the probit link
- e^{β_1} is the odds ratio for a one unit change in X
- The odds of success at X are

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

- At $X+1$ the odds are

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

Logistic regression

- If we take the ratio of these two, we can see that the odds ratio for $X+1$ versus X is simply

$$\mathrm{OR} = \frac{p(X+1)/\left[1-p(X+1)\right]}{p(X)/\left[1-p(X)\right]} = e^{\beta_1}$$

- β_1 represents the log odds ratio comparing $X+1$ to X
- Very useful quantity for epidemiologists and other applied researchers
- Probit regression coefficients don't have such a nice interpretation

Logistic regression

- Just as in linear regression, the extension to multiple covariates is straightforward
- Model is now

$$\log\left(\frac{p(\boldsymbol{X})}{1 - p(\boldsymbol{X})}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

- The inclusion of categorical covariates, nonlinear terms, or interaction terms is the same as before

Logistic regression

- The main difference lies in the interpretation of coefficients
- Now, the individual coefficients must be interpreted conditionally on the remaining covariates
- β_1 is the conditional log odds ratio for a one unit change in X_1 while conditioning on $X_2, \ldots, X_p$
- The odds ratio between X_1 and Y is the same at all possible levels of $X_2, \ldots X_p$
- Is this assumption reasonable?
  - Maybe not, but this model can provide a reasonable approximation to the truth in many cases
  - Bias-variance trade-off

Estimation in logistic models

- Previously, we used the least squares criterion to estimate the parameters in linear regression
- This criterion doesn't directly apply in the binary outcome setting
- Maximum likelihood is the preferred approach for estimating $\boldsymbol{\beta}$
- Interestingly, for ordinary least squares, the least squares and maximum likelihood approaches coincide
  - Our least squares estimate was also the maximum likelihood estimate (MLE)

Estimation in logistic models

- The first step to finding the MLE of $\boldsymbol{\beta}$ is writing down the likelihood of the data
- Our outcome is binary and therefore follows a bernoulli distribution
- If Y is bernoulli, then its PMF is given by

$$P(Y = y) = p^y(1-p)^{(1-y)}, \quad y = 0, 1$$

where $p = P(Y = 1)$

Estimation in logistic models

- In our setting, we have parameterized this probability as a function of $\boldsymbol{X}_i$
$$P(Y_i = 1|\boldsymbol{X}_i) = p(\boldsymbol{X}_i) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}$$
- Since our data are independent, we can write our likelihood as

$$\begin{aligned} L(\boldsymbol{Y}|\boldsymbol{X}) &= \prod_{i=1}^{n} L(Y_i|\boldsymbol{X}_i) \\ &= \prod_{i=1}^{n} p(\boldsymbol{X}_i)^{Y_i}(1 - p(\boldsymbol{X}_i))^{(1-Y_i)} \end{aligned}$$

Estimation in logistic models

- Note this is a function of $\boldsymbol{\beta}$
- To find the MLE, we simply need to maximize $L(\boldsymbol{Y}|\boldsymbol{X})$ with respect to $\boldsymbol{\beta}$
- Typically this involves taking the derivative with respect to $\boldsymbol{\beta}$, setting equal to zero, and solving for $\boldsymbol{\beta}$
- Unfortunately, this does not have a nice closed-form solution

Estimation in logistic models

- One way to proceed is through optimization techniques to find the minimum numerically
  - Newton-Raphson method, many others
  - Might have trouble finding the maximum as the dimension of $\boldsymbol{\beta}$ grows
- Another approach is through iteratively re-weighted least squares
  - Successively fit weighted least squares to the model until it converges
  - Used in the glm function in R

Classification in logistic models

- Once we have an estimate $\boldsymbol{\beta}$, we can proceed with classification
- As discussed in previous lectures, classification should proceed by choosing the most likely class for the outcome
- With only two classes, this amounts to predictions of the form

$$\widehat{Y}_i = \begin{cases} 1, & \widehat{p}(\boldsymbol{X}_i) \geq 0.5 \\ 0, & \widehat{p}(\boldsymbol{X}_i) < 0.5 \end{cases}$$

Classification in logistic models

- This can also be framed in terms of the linear predictor of the model

$$\widehat{Y}_i = \begin{cases} 1, & \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j X_{ij} \geq 0 \\ 0, & \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j X_{ij} < 0 \end{cases}$$

- If our model is correctly specified (we have the correct distribution of $Y|X$), this should approximate the Bayes classifier
  - Best we could hope to do

- We will see later that other classification techniques lead to similar classifiers