Week 8: Logistic Regression

03/11/2020

Jake Campbell

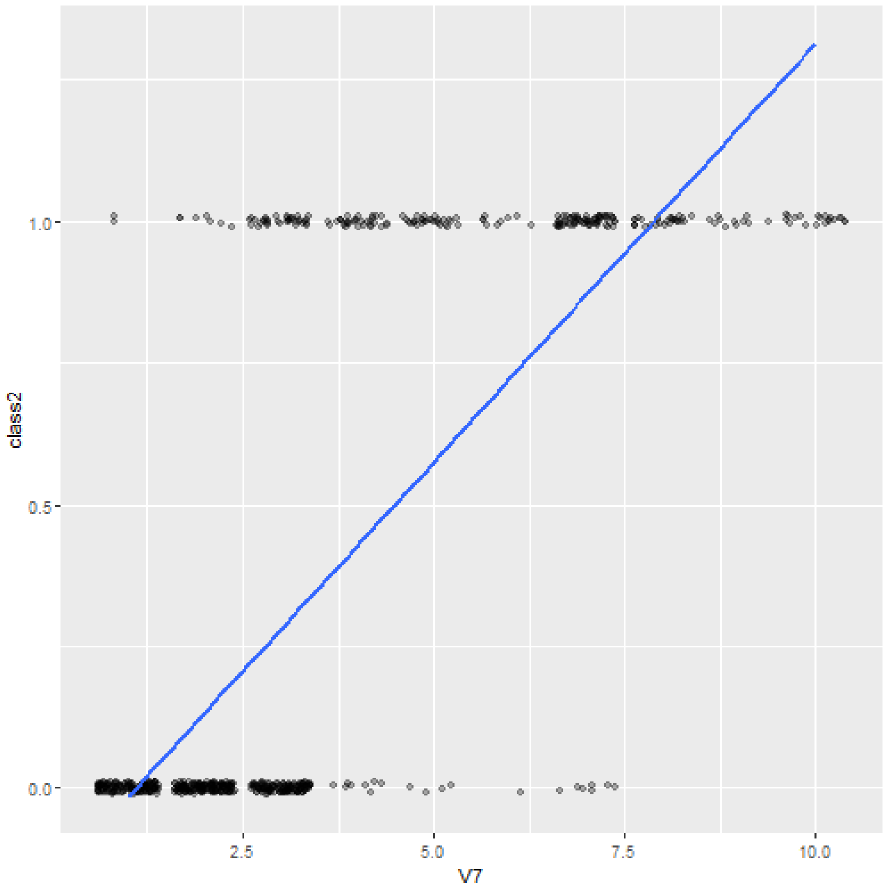
Moving On From Linear Problems

- Up to now, we've only looked at modeling continuous response variables
    - Miles per gallon, rate of crime, etc.
- There are several problems where we need to model a categorical response
    - Does a team win a game, does a student accept an offer of admission?
- These problems involving categorical response variables are classification problems

Issues With Modeling This Problem Linearly

- Let's say we were trying to model a yes/no problem
- We can treat all yes responses as `1` and no responses as `0` and predict the probability of either a `1` or `0`
- A simple linear model isn't really flexible to this scenario
- Predictions can go above and beyond `0` or `100%` probability

Issues With Modeling This Problem Linearly



4/12

Logistic Regression

- We can attack binary classification problems using the logistic function
  - Can take any input and return a value between 0 and 1

$$P = \frac{e^{(B_0 + B_x)}}{1 + e^{(B_0 + B_x)}}$$

- By using some simple math, we can edit the logistic function to get something similar to the linear regression equation
  - Instead of predicting y, we are predicting the log odds of y

$$log(\frac{P}{1 - P}) = B_0 + B_x$$

5/12

Odds and Probability

- We aren't directly predicting probability in logistic regression; we are predicting the log of the odds ratio
- Probability can be easily changed to odds and vice-versa
- The odds ratio is simply the odds of event `A` happening in the numerator and the odds of it not happening in the denominator
- For example, if a football team had an `80%` chance of winning, their odds of winning would be `.8 / (1 - .8)` or `4 / 1`
  - The team has `4` to `1` odds of winning; if the game was played `5` times, the team would win `4` times

6/12

Logistic Regression in R

- We will use `glm()` (generalized linear model) to fit logistic regression models
- `glm` can fit several linear and non-linear models
- Specify the argument `family = "binomial"` to perform logistic regression
- `summary()` can be used similarly to how we use it for `lm()` models

Logistic Regression Output

- Like in our `lm` output, we get coefficient estimates and p-values.
- Coefficient estimates are the increase in the log odds of `y` for a one unit increase in `x`
- Deviance is a measure of model fit
  - The lower the better
- Residual deviance is the model's deviance, while null deviance is the deviance with only an intercept
  - Similar to an overall F-test in `lm` output

Interpreting Logistic Regression Coefficients

- Remember that we are predicting the log odds of event `y` occurring
  - A one unit increase in `x`, leads to an increase in the log odds of event `y` occurring by its coefficient estimate
- We can take the exponential of our coefficient estimates to convert them to odds

```
exp(coef(tumor.glm))
```

```
##  (Intercept)           V1           V3           V4           V5           V6 
## 0.0000411763 1.7067054139 1.3745942848 1.3908643372 1.1008866850 1.4667887059 
##           V7           V8           V9 
## 1.5628015794 1.2368248052 1.7058506868
```

- For example, a one-unit increase in `V1` would up the odds of the tumor being malignant by about `70%`

9/12

Logistic Regression Assumptions

- Response is binary (two classes)
- Independent observations
- No multi-collinearity issues
- Large sample sizes
    - Logistic regression is fit by maximum likelihood which requires larger sample sizes

The Role of Polynomials

- We don't expect variables to have a linear relation with the response; we do expect variables to have a linear relation with the log odds
- The log odds of a model can be plotted against an independent variable
- Relationships that don't appear linear might be improved with the use of polynomials

Measuring Accuracy

- We can make class probability predictions using `predict()` and specifying `type = "response"`
  - Without specifying type, we get log odds predictions
- We can round the predicted probabilities to the nearest whole number and measure how accurate our model is
- Straightforward method of determining how good our model is