Math 455 Class 33 April 17

For Midterm 2: be prepared for questions on the interpretation of regression coefficients of "dummy variables". Notes will be posted.

Today: some ideas and code from Ch. 9.

We discussed the case of a predictor variable with non-numerical values "categorical variable". This is handled automatically by R: a "reference level" and "dummy variables" are introduced.

What if the response Y is non-numerical? Many natural instances of this, e.g. winner of an election or sporting event. Health or education outcomes also a possible case.

We can still potentially make use of the ideas of this course! (Or, more accurately, the next one: Math 532 treats "Logistic Regression")

Suppose that we have two outcomes: "Win" and "Lose".

Idea 1: Instead of predicting "W" or "L" we will predict the probability p of W. where $0 < p < 1$.

Problem: Linear functions produce values y with $-\infty < y < \infty$.

Predicting that W has probability 2.5 or -3 doesn't make sense.

Idea 2: "Odds of W" by definition is

$$\frac{p}{1-p} \leftarrow \begin{matrix} \text{prob}(W) \\ \text{prob}(L). \end{matrix}$$

Terminology: if $p = \frac{2}{3}$ our team is a 2:1 favorite to win because $\frac{\frac{2}{3}}{1-\frac{2}{3}} = \frac{2}{1}$.

Note Odds(W) is in $(0, \infty)$

We consider the log-odds of W which is

$$\log\left(\frac{p}{1-p}\right)$$ and note

$$-\infty < \log\left(\frac{p}{1-p}\right) < \infty.$$

So now we have something which could reasonably be the output of a linear function.

Then we take our predictors X_1, X_2 etc. and try to produce a function

$$f(X_1, \ldots, X_{p-1}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}.$$

such that it is our prediction of

$$\log\left(\frac{p}{1-p}\right).$$

Problem: In our data, we don't see any probabilities. We only see wins and losses.

If we interpret W as $p=1$ and L as $p=0$, $\log\left(\frac{1}{1-0}\right) = +\infty$, $\log\left(\frac{0}{1-0}\right) = -\infty$.

These can't be used in a linear regression.

Idea 3: log-loss.

We give, using our values of $\beta_0, \ldots, \beta_{p-1}$ predictions of probabilities $p_1, \ldots, p_n$ for our n rows of data.

For each row of data we have a result "W" corresp. to $p=1$ or "L" ($p=0$).

We want to give a penalty in each row such that if our prediction is $p=0.95$ and the result is W, the penalty is small. if our prediction is $p=0.13$ and the result is W, the penalty should be large.

This penalty is called the "log-loss".

Assume our data of "W" and "L" are in the form $y_1, \ldots, y_n$ where

$$y_i = \begin{cases} 1 & \text{if } i\text{th game is W} \\ 0 & \text{\_\_\_\_\_ L.} \end{cases}$$

The log-loss penalty is $-\sum y_i \log(p_i) + (1-y_i)\log(1-p_i)$

$$= \sum_{i=1}^{n} -y_i \log(p_i) + (y_i - 1)\log(1-p_i).$$

We choose the coeff. est. $\hat{\beta}_0, \ldots, \hat{\beta}_{p-1}$

by minimizing the log-loss.

(This turns out to be equivalent to a max-likelihood method.)

In Faraway, we get one sentence about how we should consider a transformation.

$\log\left(\frac{y}{1-y}\right)$. ← the log-odds transform of logistic regression.

other transformations, e.g. $\log(y+\alpha)$

or $0.5 \log\left(\frac{1+y}{1-y}\right)$ are also possible

but less common in practice.

Another idea: we need not insist on a linear function. We can have any sort of function we can parametrize.

This means we can have e.g. a cubic polynomial, because in

$$a_0 + a_1 X + a_2 X^2 + a_3 X^3$$

a_0, a_1, a_2, a_3 are parameters we choose with least-squares.

We could instead have e.g. a piecewise linear function, or a piecewise linear continuous function. Or a piecewise polynomial function or a piecewise polyn. function with continuous 1st derivative.

Another idea: "roughness penalty".

We try to minimize

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx.$$

We assume that a form for the function (e.g. cubic polyn. above) where there is a $f''(x)$, and we force the optimization to prefer smoothness by adding the penalty-term $\lambda \int (f''(x))^2 dx$.