POLS/CSSS 510:
Maximum Likelihood Methods for the Social Sciences

Problem Set 3

Professor: Christopher Adolph, Political Science and CSSS

Fall Quarter 2020

Due in class on Tuesday 10 November 2020

General instructions for homeworks: Homework can be handwritten or typed. For any exercises done with R or other statistical packages, you should attach all code you have written and all (interesting) output. Materials should be stapled together in order by problem. The most readable and elegant format for homework answers incorporates student comments, code, output, and graphics into a seamless narrative, as one would see in a textbook.

Problem 1: Modeling Cy Young winners using logit

The dataset `cyyoung.csv` contains information on selected North American baseball pitchers from 1980 to 2002. Pitchers' performance can be measured in several ways: their record of games won or lost, the number of runs (points) they allowed the other team to score per game, the number of players they "struck out," the number of players they "walked," and the number of innings they pitched. At the end of the season, two pitchers (one from the American League, and one from the National League) are voted the best pitchers of the year.

a. [5 points] Fit a logistic regression to the variable `cy` with `era` and `winpct` as the only covariates. Report the estimated parameters, their standard errors, and the log likelihood at its maximum. Perform this fit using `optim()`, then replicate the fit using `glm()`.

b. [10 points] Without using a special package like simcf, calculate the probability a pitcher receives the Cy Young award given

$$\texttt{era} = \{1.50, 1.75, 2.00, \ldots, 4.75, 5.00\},$$

with winpct held at its mean value. Now, calculate the probability again, for the same range of era's, given *either* winpct = 0.5 *or* winpct = 0.9. You should end up with $3 \times 15 = 45$ probabilities. Plot these estimated probabilities nicely (the tile package works well for this graphic and the next problem, but for this part, even a matrix works well).

c. [15 points] Calculate or simulate 95 percent confidence intervals for each of the probabilities plotted in b. (You may now use any package you wish.) Design a graphic (not a table) to incorporate these confidence intervals. Interpret your findings.

d. [30 points] Find a "better model" of cy. You may add other variables from the dataset, remove variables already in the model, and/or transform or any variables you wish, except cy. Whatever choice you make you should justify in some fashion. Fit your new model, and show whether your fit has improved using **(i)** a likelihood ratio test, **(ii)** AIC and/or BIC, **(iii)** in-sample ROC curves, **(iv)** in-sample Actual versus Predicted plots, and **(v)** cross-validation using the metric(s) of your choice.

e. [20 points] Suppose your model from d. has the following form:

$$\begin{aligned} \mathsf{cy}_i &\sim \text{Bernoulli}(\pi_i) && (1) \\ \pi_i &= \text{logit}^{-1}(\alpha + z_i\gamma + \mathbf{x}_i\beta) && (2) \end{aligned}$$

where z_i is a covariate of particular interest and $\mathbf{x}_i$ is a vector of additional covariates. We are interested in understanding how the conditional expectation of cy changes as z changes *given* particular (fixed) values of each covariate in $\mathbf{x}$.

Choose a particular hypothetical change in z such that z changes from z_{pre} to z_{post}.[1] Then simulate the first difference in the probability (or, if you prefer,

1 If z is binary, this will likely be a change from 0 to 1; otherwise, the choice of *ex ante* and *ex post* values of z will require some thought.

the relative risk) of receiving a Cy Young given the change in z, holding other covariates $\mathbf{x}$ constant (that is, make sure $\mathbf{x}_{\text{pre}} = \mathbf{x}_{\text{post}}$).

Then repeat the exercise, simulating the first difference in the probability (or, if you prefer, the relative risk) of receiving a Cy Young given the change in z, but with some different value(s) of $\mathbf{x}_{\text{pre}} = \mathbf{x}_{\text{post}}$.[2] Display your results neatly in a plot or table, and be sure to show the uncertainty in these estimates.

f. [10 points] Does logistic regression offer a defensible probability model here? What assumptions of this model might be violated by the variable cy?

g. [10 points] Suppose the pitchers selected for inclusion in the dataset were all considered "contenders" for the Cy Young award by knowledgable experts. Pitchers whom the experts considered unlikely to win the award were excluded. How might this fact affect your findings?

h. [Bonus: +10 points; may be turned in anytime during the quarter] In d., you computed the first difference (or relative risk) in cy given a change in z and fixed x for hypothetical values of each covariate. This is useful for understanding relationships that might hold in the population from which our sample is drawn. But if we are more interested in our particular sample – as well as heterogeneity within that sample – there is a better way to explore first differences and relative risks: in-sample counterfactuals.

To use an in-sample counterfactual, we consider *for each sampled individual* the expected change in that individual's outcome, cy_i, given the same counterfactual amount of change in z_i, the factual (observed) $z_{i,\text{pre}}$, and the factual values of the observed $\mathbf{x}_i$.[3] This leads to a different predicted first difference or relative risk for each individual i.[4]

Your task is to compute the full set of in-sample first differences (or relative risks) in cy for the change in z you considered in d., and present these in some fashion

2 Make sure that for any particular scenario, $\mathbf{x}$ has the same *ex ante* and *ex post* values.

3 If z has an unbounded scale, it's enough to "start" each observation at $z_{i,\text{pre}}$. If z is bounded (e.g., a binary variable), then you may have to set $z_{i,\text{pre}}$ to a counterfactual level to keep all hypothetical $z_{i,\text{post}}$'s to the logically required range.

4 *Intuition check:* in-sample counterfactual first differences from a logistic regression will be different for each i, but for "plain vanilla" linear regression with unbounded linear covariates, linear response, and no interactions, first differences (but not relative risks) will be the same for every i. Why?

(a table, or a density plot, or a dotplot, for example). Then compute the average in-sample first difference or relative risk. Discuss the advantages of this approach compared to other presentations of logistic regression results.

Contents of `cyyoung.csv`

Variable	Description
`year`	The year of the observation; also called a "season"
`pitcher`	The name of the pitcher
`natleag`	$= 1$ if the pitcher played in the National League; $= 0$ if he played in the American League
`wins`	The number of games the pitcher personally won
`winpct`	The percentage of games which the pitcher personally won
`era`	The number of runs the pitcher allows per 9 innings; low values are better
`strikeout`	The number of strikeouts the pitcher collected over a season; high numbers are better
`innings`	The number of innings (periods) a pitcher played during the season
`cy`	Whether the pitcher won a Cy Young award for the season (one awarded per league per season)
`walks`	The number of walks the pitcher collected over a season; low numbers are better
`team`	The name of the pitcher's team
`twinpct`	The percentage of the games the pitcher's team won during the season, regardless of whether the pitcher played
`trank`	The rank of the team within its division at the end of the season
`playoffs`	$= 1$ if the team advanced to the playoffs; $= 0$ if the team did not

Hints

Strikeouts (technically) and walks can happen any positive number of times per inning of play.

A hypothesis is that pitchers from good teams have a better chance of winning the Cy Young award, all else equal.

Because of differences in rules, it is harder for American League pitchers to achieve low ERAs.