STATS 200: Homework 8

Due Friday, December 9, at 5PM

1. **Fitting a Bradley-Terry model.** The file `NBA_record.csv` contains the results of all 1230 NBA games from the 2015–2016 regular season. The 30 teams are encoded numerically from 1 to 30; the key for this encoding is provided in the file `teams.txt`. Each row of `NBA_record.csv` indicates the home team, away team, and outcome Y for one game, where $Y = 1$ if the home team won and $Y = 0$ otherwise.

For parts (a) and (b), you may *not* use an existing software implementation of the Bradley-Terry or logistic regression model; however, you may use any generic optimization or equation-solving routine (or you may implement the Newton-Raphson iterations yourself, if you are brave).

(a) Fit the Bradley-Terry model, with an intercept term α for the home-court advantage, to this data set. What are the 8 teams (in ranked order) with the highest Bradley-Terry scores? How much greater is the log-odds of winning for the home team than for the away team?

One approach to do this in R is to use the generic optimization function `optim`. To do this, first define a function

```
loglik = function(theta,Home,Away,Y) {
  ...
}
```

that returns the log-likelihood for the Bradley-Terry model given inputs $\theta = (\alpha, \beta_2, \ldots, \beta_k)$ (constraining $\beta_1 \equiv 0$), Home $= (i_1, \ldots, i_n)$, Away $= (j_1, \ldots, j_n)$, and Y $= (Y_1, \ldots, Y_n)$, where i_m and j_m are the home and away teams for game m. To then read the data file and maximize the log-likelihood:

```
table = read.csv('NBA_record.csv')
result = optim(theta0,loglik,Home=table$Home,Away=table$Away,Y=table$Y,
    method='BFGS',control=list('fnscale'=-1))
```

Here `theta0` is any initialization for θ (for example the all 0's vector). The method will use the BFGS algorithm, and `'fnscale'=-1` indicates that it should perform maximization rather than minimization.

(b) Fit the Bradley-Terry model without an intercept term. (You may do this in R by defining a new function `loglik_noalpha = function(theta,Home,Away,Y)` where now $\theta = (\beta_2, \ldots, \beta_k)$, and using `optim` as before.) Evaluate the log-likelihoods at the full model and sub-model MLEs, and carry out a generalized likelihood ratio test of the null hypothesis of no home court advantage, $H_0 : \alpha = 0$. What is the p-value that you obtain for your test?

(c) For the mth game, suppose we define 30 covariates $x_{m,1}, \ldots, x_{m,30}$ in the following way: Let $x_{m,1} = 1$ always. Let $x_{m,i} = 1$ if team i is the home team of this game and $i \neq 1$, and let $x_{m,j} = -1$ if team j is the away team of this game and $j \neq 1$. Let $x_{m,k} = 0$ for all other k. Explain why logistic regression for Y_m using the covariates $x_{m,1}, \ldots, x_{m,30}$ is equivalent to the Bradley-Terry model, where we constrain the Bradley-Terry score of team 1 to be $\beta_1 \equiv 0$. If we were to run this logistic regression, what would be the interpretation of the fitted coefficient for the first covariate $x_{m,1}$? For the 10th covariate $x_{m,10}$?

(d) Fit the logistic regression in part (c) using any standard regression software, and verify that the fitted coefficients match (up to reasonable numerical accuracy) your estimated parameters from part (a).

To do this in R, you may construct a matrix X of size 1230 × 30 containing the covariates as defined in part (c), and then fit the regression using

```
model = glm.fit(X,table$Y,family=binomial())
coefs = model$coefficients
```

2. **A heteroskedastic linear model.** Consider observed response variables $Y_1, \ldots, Y_n \in \mathbb{R}$ that depend linearly on a single covariate $x_1, \ldots, x_n$ as follows:

$$Y_i = \beta x_i + \varepsilon_i.$$

Here, the ε_i's are independent Gaussian noise variables, but we do not assume they have the same variance. Instead, they are distributed as $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ for possibly different variances $\sigma_1^2, \ldots, \sigma_n^2$. The unknown parameter of interest is β.

(a) Suppose that the error variances $\sigma_1^2, \ldots, \sigma_n^2$ are all known. Show that the MLE $\hat{\beta}$ for β, in this case, minimizes a certain weighted least-squares criterion, and derive an explicit formula for $\hat{\beta}$.

(b) Show that the estimate $\hat{\beta}$ in part (a) is unbiased, and derive a formula for the variance of $\hat{\beta}$ in terms of $\sigma_1^2, \ldots, \sigma_n^2$ and $x_1, \ldots, x_n$.

(c) Compute the Fisher information $I_{\mathbf{Y}}(\beta) = -\mathbb{E}_\beta[l''(\beta)]$ in this model (still assuming $\sigma_1^2, \ldots, \sigma_n^2$ are known constants). Show that the variance of $\hat{\beta}$ that you derived in part (b) is exactly equal to $I_{\mathbf{Y}}(\beta)^{-1}$.

In the remaining parts of this question, denote by $\tilde{\beta}$ the usual (unweighted) least-squares estimator for β, which minimizes $\sum_i (Y_i - \beta x_i)^2$. In practice we might not know the values of $\sigma_1^2, \ldots, \sigma_n^2$, so we might still estimate β using $\tilde{\beta}$.

(d) Derive an explicit formula for $\tilde{\beta}$, and show that it is also an unbiased estimate of β.

(e) Derive a formula for the variance of $\tilde{\beta}$ in terms of $\sigma_1^2, \ldots, \sigma_n^2$ and $x_1, \ldots, x_n$. Show that when all error terms have the same variance σ_0^2, this coincides with the general formula $\sigma_0^2(X^T X)^{-1}$ for the linear model, from Lecture 25.

(f) Using the Cauchy-Schwarz inequality $(\sum_i a_i^2)(\sum_i b_i^2) \geq (\sum_i a_i b_i)^2$ for any positive numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$, compare your variance formulas from parts (b) and (e) and show directly that the variance of $\tilde{\beta}$ is always at least the variance of $\hat{\beta}$. Explain, using the Cramer-Rao lower bound, why this is to be expected given your finding in part (c).

(g) Estimating each σ_i^2 by the squared residual $(Y_i - \tilde{\beta} x_i)^2$, derive a plugin estimate for the standard error of $\tilde{\beta}$ that is robust to possible differences in the variances $\sigma_1^2, \ldots, \sigma_n^2$.

(h) Perform a simulation that compares your standard error estimate in (g) to the usual standard error estimate of $\tilde{\beta}$ obtained by linear regression software packages, as follows: Let $(x_1, x_2, \ldots, x_{100}) = (0.01, 0.02, \ldots, 1)$ and $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_{100}^2) = (0.01^2, 0.02^2, \ldots, 1^2)$. In each of $B = 10000$ simulations, generate $Y_i = x_i + \varepsilon_i$ for $i = 1, \ldots, 100$, and fit the linear model $Y = x + \varepsilon$ using any standard linear regression package to obtain the estimate $\tilde{\beta}$ and an estimated standard error for $\tilde{\beta}$. Compute also the estimated standard error of $\tilde{\beta}$ using your method in part (g).

Report the true (empirical) standard deviation of $\tilde{\beta}$ across the B simulations, and plot two histograms of the estimated standard errors using the two different methods. Summarize briefly your findings.

In R, given data vectors `x` and `Y`, you may run the regression $Y = x + \varepsilon$ using the command

```
model = lm(Y ~ x + 0)
```

(The +0 means that the regression should not be fit with an intercept term.) The least-squares estimate $\tilde{\beta}$ is obtained as

```
summary(model)[["coefficients"]][["x","Estimate"]]
```

and the estimated standard error of $\tilde{\beta}$ is obtained as

```
summary(model)[["coefficients"]][["x","Std. Error"]]
```