PLSC 40502: Problem Set 1

YOUR NAME

January 5, 2024

This problem set is due at **11:59 pm on Wednesday, January 17th**.

Please upload your solutions as a .pdf file saved as "Yourlastname_Yourfirstinitial_pset1.pdf"). In addition, an electronic copy of your .Rmd file (saved as "Yourlastname_Yourfirstinitial_pset1.Rmd") must be submitted to the course website at the same time. We should be able to run your code without error messages. In order to receive credit, homework submissions must be substantially started and all work must be shown.

Problem 1

In "The Declining Risk of Death in Battle," Lacina et. al. (2006) study whether there has been a downward time trend in fatalities in armed conflict over time. This paper fits into a broader empirical literature on the study of armed conflict and its consequences. One feature of the datasets used in many of these empirical papers is that the outcomes of interest are often non-negative integers (such as event counts). As such, researchers often make use of count regression models to test hypotheses.

While Lacina et. al. (2006) examine a wide range of conflicts from 1900 onward, this problem will use a different dataset that focuses on the post-Cold War period exclusively. The UCDP Battle-Related Deaths Dataset (v. 22.1) provides annual estimates of battle-related deaths in armed conflicts from 1989-2021. We will examine whether there has been a downward time trend in battle deaths in conflict during this time period via a Poisson regression.

```r
# Read in the UCDP data
ucdp <- read_csv("data/ucdp-brd-conf-221.csv")
```

The data-generating process for the Poisson regression assumes that the outcome has a distribution

$$Y_i \underset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda_i)$$

with linear predictor and link function

$$\lambda_i = \exp(X_i'\beta)$$

One way of interpreting the regression parameters is that they are additive on the *log* scale. In other words, we are assuming a linear model for the log CEF:

$$\log(E[Y_i|X_i]) = X_i'\beta$$

Part A

Write down the log-likelihood $\ell(\beta|\mathbf{X}, \mathbf{Y})$ for the Poisson GLM regression parameters

Part B

Write an R function that takes as input a vector of coefficients β, outcome vector $\mathbf{Y}$ and design matrix $\mathbf{X}$ and returns the poisson log-likelihood.

Hint: You can have it return the log-likelihood as a scalar or a vector of the individual log-likelihood for each observation – the latter will be useful (and works with `maxLik`).

Part C

Using the UCDP data, use your function from Part B and the `maxLik` R package to obtain the MLE for the coefficients of a poisson GLM that regresses the total number of battle deaths in a given conflict-year (using the "best" estimate: `bd_best`) on the year of observation, an indicator for whether the conflict is "interstate", and an indicator for whether the conflict is "internationalized intrastate" (the "left out" group for these dummies is 'intrastate').

Hint: Check the codebook for the `type_of_conflict` variable to find our how to generate the correct dummy variables for conflict type.

Hint: Don't forget the intercept when making your design matrix $\mathbf{X}$

Part D

Obtain an estimate of the variance-covariance matrix under the assumption that the model is correctly specified. Provide a 95% confidence interval for the coefficient on `year`. Conduct a hypothesis test for the null that the coefficient on `year` equals zero with $\alpha = .05$.

Provide a substantive interpretation of the coefficient on `year` in terms of battle deaths in a conflict-year.

Part E

What does the model predict will be the expected count of battle deaths for an interstate conflict in the year 2018? Construct a 95% confidence interval for this prediction using the delta method and your variance-covariance matrix from D.

Hint: You may find the `numericGradient()` function from `maxLik` useful for this part.

Part F

Compare your prediction from E to the same prediction from a linear regression model using the same variables. Do the two models give meaningfully different results for the CEF?

Part G

Implement the "robust" Huber sandwich estimator for the variance-covariance matrix of your Poisson regression coefficient (ignore clustering for now, just implement the "heteroskedasticity"-robust version). Compare these standard errors to the conventional MLE standard errors. What does this tell you about the modeling assumptions that you've made in previous parts of this problem?

Hint: You may find the `gradient()` and `hessian()` functions from `maxLik` useful for this part.

Problem 2

In this problem you will derive a closed form expression for the MLE of the "Normal" regression model. We will focus on the simple case of one regressor and one intercept. Assume the following data-generating process for n observations $\mathbf{Y} = \{Y_1, Y_2, Y_3, \dots Y_n\}$:

$$Y_i \underset{\text{i.i.d.}}{\sim} \text{Normal}(\beta_0 + \beta_1 X_i, \sigma^2)$$

In other words, each observation Y_i is independent of the others and is assumed to come from a normal distribution with mean $\beta_0 + \beta_1 X_i$ and variance σ^2.

Part A

Write down the probability density function for a single observation $p(y_i|\beta_0, \beta_1, \sigma^2)$

Part B

Write down the log-likelihood $\ell(\beta_0, \beta_1, \sigma^2|\mathbf{Y})$. Simplify as much as you can and drop any additive constants (this will help in the next part).

Part C

Find the MLE for β_0 and β_1

Hint: Express one MLE in terms of the MLE of the other and substitute.

Part D

What familiar estimator is the MLE of β_0 and β_1. What does this tell you about the bias of the MLE for these parameters?

Part E (Optional Bonus!):

Find the MLE for σ^2. Is this MLE unbiased?

Problem 3

In this problem we will consider estimating the maximum of a uniform distribution using i.i.d. samples. The discrete uniform version is sometimes referred to as the "German Tank Problem" as it arose during WWII as Allied forces attempted to estimate the extent of German tank manufacturing using the observed serial numbers from captured tanks.

Consider a setting with n i.i.d. observations $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$

For each, assume $X_i \underset{\text{i.i.d}}{\sim} \text{DiscreteUniform}(1, M)$. In other words, each observation is independently and identically distributed uniformly on the integers between 1 and M.

The Discrete Uniform Distribution on integers 1 to M has a probability mass function of:

$$P(X_i = x) = \begin{cases} \frac{1}{M} & \text{if } 1 \leq x \leq M \\ 0 & \text{otherwise} \end{cases}$$

Part A

Write down the likelihood $\mathcal{L}(M|\mathbf{X})$

Part B

Suppose we observe 5 observations: $\mathbf{X} = \{10, 30, 78, 293, 43\}$. Make a graph of the likelihood function.

Part C

Find the MLE of M, $\hat{M}$.

Part D

Is the MLE unbiased? Is the MLE consistent? Explain why or why not.