PLSC 40502: Problem Set 1

YOUR NAME

January 8, 2025

This problem set is due at **11:59 pm on Wednesday, January 22nd**.

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.

```
# 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 k-length vector of coefficients β, n-length outcome vector $\mathbf{Y}$ and n by k design matrix $\mathbf{X}$ and returns the poisson log-likelihood evaluated for each individual observation.

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-White 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 `estfun()` functions from `maxLik` (and the `sandwich` library) useful for this part.

Problem 2

Now we'll modify the previous problem and consider a different data-generating process. Another distributional assumption that researchers often make for count variables is to assume that they have a **negative binomial** distribution. This distribution has **two** parameters - one version of the parameterization of this distribution has a mean parameter μ_i and a non-negative "shape" parameter θ.

$$Y_i \underset{\text{i.i.d.}}{\sim} \text{NegBin}(\mu_i, \theta)$$

where

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

The probability mass function for the Negative Binomial distribution given parameters μ and θ is:

$$f(y; \mu, \theta) = \frac{\Gamma(y + \theta)}{\Gamma(\theta)y!} \times \frac{\mu^y \theta^\theta}{(\mu + \theta)^{(y+\theta)}}$$

This PMF is implemented in the `pnbinom()` function in base R where θ is referred to as the `size` parameter and μ is referred to as `mu`.

Part A

Write an R function that takes as input a $k + 1$-length vector of parameters where the first k parameters are the regression coefficients β and the last parameter is the shape parameter θ (or some transformation of it), n-length outcome vector $\mathbf{Y}$ and n by k design matrix $\mathbf{X}$ and returns the negative binomial log-likelihood evaluated for each individual observation.

Hint: You may want to pass a transformed version of θ in the parameter vector to make optimization easier. By default, most numeric optimizers work best when optimizing over *unconstrained* parameters, but recall that the shape parameter θ must be non-negative. In other words, your function should apply some transformation to that $k + 1$th element of the parameter vector to convert it to θ before passing it to the (log) PMF.

Part B

Using the UCDP data, use your function from Part B and the `maxLik` R package to obtain the MLE for the coefficients of a negative binomial regression hat 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').

What is your maximum likelihood estimate of the "shape" or "dispersion" parameter θ?

Part C

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 $\alpha = .05$.

Compare this with your estimates from the Poisson regression.

Part D

Implement the "robust" Huber-White sandwich estimator for the variance-covariance matrix of your Negative Binomial parameters (again ignore clustering). Compare these standard errors to the conventional MLE standard errors and compare this difference to what you found for the Poisson regression. Comment on what you've found.

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 $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$

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.