PLSC 40502: Problem Set 1

Solutions

January 8, 2022

This problem set is due at **11:59 pm on Friday, January 20th**.

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 addition to your solutions, please submit an annotated version of this `.rmd` file saved as "Yourlastname_Yourfirstinitial_pset1_feedback.rmd", noting the problems where you needed to consult the solutions and why along with any remaining questions or concerns about the material. 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 inverse-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

---

By independence

$$\ell(\beta|\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{N} \ell(\beta|X_i, Y_i)$$

Substituting in the poisson pmf and using our link function

$$\ell(\beta|\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{N} \log \left[\frac{\exp(X_i'\beta)^{Y_i} \exp(-\exp(X_i'\beta))}{Y_i!} \right]$$

Drop the proportionality constants (that depend on Y only) and take the log

$$\ell(\beta|\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{N} Y_i X_i'\beta - \exp(X_i'\beta)$$

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`).

```
pois_loglik <- function(beta, Y, X){
  return(Y*(X%*%beta) - exp(X%*%beta))
}
```

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 $\mathbf{X}$

First do the data preprocessing.

One major issue with numeric differentiation in GLMs is that when covariates that have very different ranges, you can often get negative Hessians that do not invert to be valid variance-covariance matrices (e.g. you get negative diagonals). One solution here would be to pass the gradient and Hessian manually, but often you can solve issues in estimation by re-centering or re-scaling the variables. Year is a common culprit here.

While we still want the coefficients to be interpretable in terms of a "single year" increase, which year is year "0" is ultimately arbitrary. So let's recenter to year 2000 (if we don't we'll encounter some big problems). You could also recenter to the means

```
library(maxLik)
```

```
## Loading required package: miscTools
```

```
##
## Please cite the 'maxLik' package as:
## Henningsen, Arne and Toomet, Ott (2011). maxLik: A package for maximum likelihood estimation in R. C
##
## If you have questions, suggestions, or comments regarding the 'maxLik' package, please use a forum o
## https://r-forge.r-project.org/projects/maxlik/
```

```
# Make the model matrix
X_matrix <- model.matrix(bd_best ~ I(year-2000) + I(type_of_conflict == 2) + I(type_of_conflict == 4),
```

Next obtain the point estimates

```
pois_reg <- maxLik(pois_loglik, start = rep(0, ncol(X_matrix)),
                   method="BFGS",
                   Y = ucdp$bd_best,
                   X = X_matrix)
```

Print the coefficients

```
pois_coef <- coef(pois_reg)
pois_coef
```

```
## [1]  6.659550434 -0.009684781  1.490152103  1.241732363
```

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.

Let's invert the negative hessian to get the MLE variance-covariance matrix

```
pois_vcov <- solve(-hessian(pois_reg))
```

Our standard errors are

```
pois_se <- sqrt(diag(pois_vcov))
```

And so our 95% CI is

```
year_95CI <- c(pois_coef[2] - qnorm(.975)*pois_se[2], pois_coef[2] + qnorm(.975)*pois_se[2])
year_95CI
```

```
## [1] -0.009843355 -0.009526206
```

Our 95% confidence interval for the coefficient on year is $[-0.0098, -0.0095]$. As this does not cover zero, we would reject the null that the coefficient is zero at the $\alpha = .05$ level.

We can calculate a p-value using our usual expression for the test statistic and that p-value is far below .05 that it is below the level of numerical precision in R.

```
year_pval <- 2*pnorm(-abs((pois_coef[2]/pois_se[2])))
```

Transforming to the effect on the CEF instead of the log-CEF

```
exp(pois_coef[2])
```

```
## [1] 0.990362
```

We predict that a one-unit increase in the year will reduce the expected number of battle deaths in the conflict by a factor of .99 (the effect is additive on the log-scale which means that it is multiplicative on the actual scale)

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.

Our prediction for the CEF is

```
predicted_counts <- exp(pois_coef%*%c(1, 18, 1, 0)) # Remember we re-centered year to 2000
predicted_counts
```

```
##          [,1]
## [1,] 2908.453
```

We predict that the battle deaths for this conflict will be, in expectation, 2908

Our standard error using the delta method

```
predicted_se <- sqrt(numericGradient(function(x) exp(x%*%c(1, 18, 1, 0)), t0=pois_coef)%*%pois_vcov%*%t
predicted_se
```

```
##         [,1]
## [1,] 8.59549
```

And the 95% CI is $[2891.606, 2925.300]$

```
c(predicted_counts - qnorm(.975)*predicted_se, predicted_counts + qnorm(.975)*predicted_se)
```

```
## [1] 2891.606 2925.300
```

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?

We get very similar predictions from a linear regression (they differ by only about 200)

```
lin_mod <- lm(bd_best ~ I(year-2000) + I(type_of_conflict == 2) + I(type_of_conflict == 4), data=ucdp)
pred_lin <- predict(lin_mod, newdata = data.frame(year = 2018, type_of_conflict = 2))
pred_lin
```

```
##        1
## 3166.026
```

Using a linear fit we predict the number of battle deaths will be 3,166.

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.

Robust SEs (uncorrected "HC0")

Actually, `estfun` is what we need here since that's observation-wise (but this is explained in the documentation for `gradient`)

```
library(sandwich)
pois_vcov_robust <- solve(-hessian(pois_reg))%*%(t(estfun(pois_reg))%*%(estfun(pois_reg)))%*%solve(-hess
pois_vcov_robust
```

```
##               [,1]          [,2]         [,3]         [,4]
## [1,]  0.0180196830  4.368316e-04 -0.017615828 -0.023774516
## [2,]  0.0004368316  9.419221e-05 -0.001930652 -0.001095333
## [3,] -0.0176158279 -1.930652e-03  0.293007598  0.038132320
## [4,] -0.0237745155 -1.095333e-03  0.038132320  0.047977802
```

```
pois_se_robust <- sqrt(diag(pois_vcov_robust))
```

Compare the two sets of standard errors

```
pois_se
```

```
## [1] 1.139498e-03 8.090675e-05 2.958143e-03 1.746948e-03
```

```
pois_se_robust
```

```
## [1] 0.134237413 0.009705267 0.541301762 0.219038358
```

What's the ratio of one to the other?

```
pois_se_robust/pois_se
```

```
## [1] 117.8040 119.9562 182.9870 125.3834
```

The conventional standard errors for the poisson regression are comically small and clearly the result of misspecification of the variance term. The robust standard errors are orders of magnitude larger than the conventional ones - roughly 100 to 200 times larger! Luckily, QMLE-Poisson is still consistent for the true CEF under our modeling assumption that the log-CEF is linear, so we can use the robust standard errors to conduct valid inference. In this case, we would not conclude that there is a statistically significant relationship between year and the battle deaths count.

```
year_pval_robust <- 2*pnorm(-abs((pois_coef[2]/pois_se_robust[2])))
year_pval_robust
```

```
## [1] 0.3183331
```

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 $p(y_i|\beta_0, \beta_1, \sigma^2)$

$$p(y_i|\beta_0, \beta_1, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2 \right\}$$

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).

By independence

$$\ell(\beta_0, \beta_1, \sigma^2|\mathbf{Y}) = \sum_{i=1}^{N} \log p(y_i|\beta_0, \beta_1, \sigma^2)$$

Properties of logs

$$\ell(\beta_0, \beta_1, \sigma^2|\mathbf{Y}) = \sum_{i=1}^{N} \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2$$

More properties of logs and dropping additive constants

$$\ell(\beta_0, \beta_1, \sigma^2|\mathbf{Y}) = -N\log(\sigma) - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2$$

Let's use $(a/b)^2 = a^2/b^2$ and pull the denominator out of the sum since it doesn't depend on i

$$\ell(\beta_0, \beta_1, \sigma^2|\mathbf{Y}) = -n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - \beta_0 - \beta_1 X_i)^2$$

Expand the expression further

$$\ell(\beta_0, \beta_1, \sigma^2|\mathbf{Y}) = -n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i^2 + \beta_0^2 + \beta_1^2 X_i^2 - 2Y_i\beta_0 - 2Y_i X_i\beta_1 + 2\beta_0\beta_1 X_i)$$

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.

First take the partial derivative with respect to β_0

$$\frac{\partial}{\partial \beta_0} \ell(\beta_0, \beta_1, \sigma^2 | \mathbf{Y}) = \frac{1}{2\sigma^2} \sum_{i=1}^{N} (2\beta_0 - 2Y_i + 2\beta_1 X_i)$$

Set equal to 0 and solve for $\hat{\beta_0}$

$$0 = \frac{1}{2\sigma^2} \sum_{i=1}^{N} (2\hat{\beta_0} - 2Y_i + 2\hat{\beta_1} X_i)$$

Multiply by σ^2

$$0 = \sum_{i=1}^{N} (\hat{\beta_0} - Y_i + \hat{\beta_1} X_i)$$

Split the sums

$$0 = N\hat{\beta_0} - \sum_{i=1}^{N} Y_i + \hat{\beta_1} \sum_{i=1}^{N} X_i$$

Rearrange

$$N\hat{\beta_0} = \sum_{i=1}^{N} Y_i - \hat{\beta_1} \sum_{i=1}^{N} X_i$$

Divide by N and let $\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$ and $\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$ denote the means

$$\hat{\beta_0} = \bar{Y} - \hat{\beta_1} \bar{X}$$

Next take the partial derivative with respect to β_1

$$\frac{\partial}{\partial \beta_1} \ell(\beta_0, \beta_1, \sigma^2 | \mathbf{Y}) = \frac{1}{2\sigma^2} \sum_{i=1}^{N} (2\beta_1 X_i^2 - 2Y_i X_i + 2\beta_0 X_i)$$

Set equal to 0

$$0 = \frac{1}{2\sigma^2} \sum_{i=1}^{N} (2\hat{\beta_1} X_i^2 - 2Y_i X_i + 2\hat{\beta_0} X_i)$$

Same steps as before

$$0 = \hat{\beta}_1 \sum_{i=1}^{N} X_i^2 - \sum_{i=1}^{N} Y_i X_i + \hat{\beta}_0 \sum_{i=1}^{N} X_i$$

Substituting our expression for $\hat{\beta}_0$

$$0 = \hat{\beta}_1 \sum_{i=1}^{N} X_i^2 - \sum_{i=1}^{N} Y_i X_i + (\bar{Y} - \hat{\beta}_1 \bar{X}) \sum_{i=1}^{N} X_i$$

Rearranging and combining some sums

$$0 = \hat{\beta}_1 \sum_{i=1}^{N} \left(X_i^2 - \bar{X} X_i \right) - \sum_{i=1}^{N} \left(Y_i X_i - \bar{Y} X_i \right)$$

Rearranging

$$\hat{\beta}_1 \sum_{i=1}^{N} \left(X_i^2 - \bar{X} X_i \right) = \sum_{i=1}^{N} \left(Y_i X_i - \bar{Y} X_i \right)$$

Dividing

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} \left(Y_i X_i - \bar{Y} X_i \right)}{\sum_{i=1}^{N} \left(X_i^2 - \bar{X} X_i \right)}$$

And factoring

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} \left(X_i (Y_i - \bar{Y}) \right)}{\sum_{i=1}^{N} \left(X_i (X_i - \bar{X}) \right)}$$

You can stop here, but to show it in the most familiar form, we need subtract $\sum_{i=1}^{N} \bar{X}(Y_i - \bar{Y})$ from the numerator and $\sum_{i=1}^{N} \bar{X}(X_i - \bar{X})$ from the denominator

We can show that both of those terms equal zero since

$$\sum_{i=1}^{N} \bar{X}(Y_i - \bar{Y}) = \bar{X} \sum_{i=1}^{N} (Y_i - \bar{Y}) = \bar{X}(N\bar{Y} - N\bar{Y}) = 0$$

$$\sum_{i=1}^{N} \bar{X}(X_i - \bar{X}) = \bar{X} \sum_{i=1}^{N} (X_i - \bar{X}) = \bar{X}(N\bar{X} - N\bar{X}) = 0$$

Then,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} \left(X_i (Y_i - \bar{Y}) - \bar{X}(Y_i - \bar{Y}) \right)}{\sum_{i=1}^{N} \left(X_i (X_i - \bar{X}) - \bar{X}(X_i - \bar{X}) \right)}$$

And finally,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

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?

The MLE is the OLS estimator, which we know by the Gauss-Markov theorem is also unbiased for the true regression parameters β_0 and β_1

Part E (Optional Bonus!):

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

Define $\hat{Y}_i$ as $\beta_0 + \beta_1 X_i$

You should find that $\hat{\sigma^2}$ ends up being

$$\hat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

which we know from the properties of the OLS estimator is biased but consistent due to the presence of N rather than $N - K$ in the denominator ("Bessel's correction"). Replacing that denominator with the "degrees-of-freedom" corrected sample size would yield an unbiased estimator of the population variance parameter.

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 $\mathcal{L}(M|\mathbf{X})$

$$\mathcal{L}(M|\mathbf{X}) = \prod_{i=1}^{N} \frac{1}{M} \times \mathbb{1}(X_i \leq M)$$

where $\mathbb{1}(X_i \leq M)$ denotes an indicator that the observation X_i is less than or equal to M and is zero otherwise.

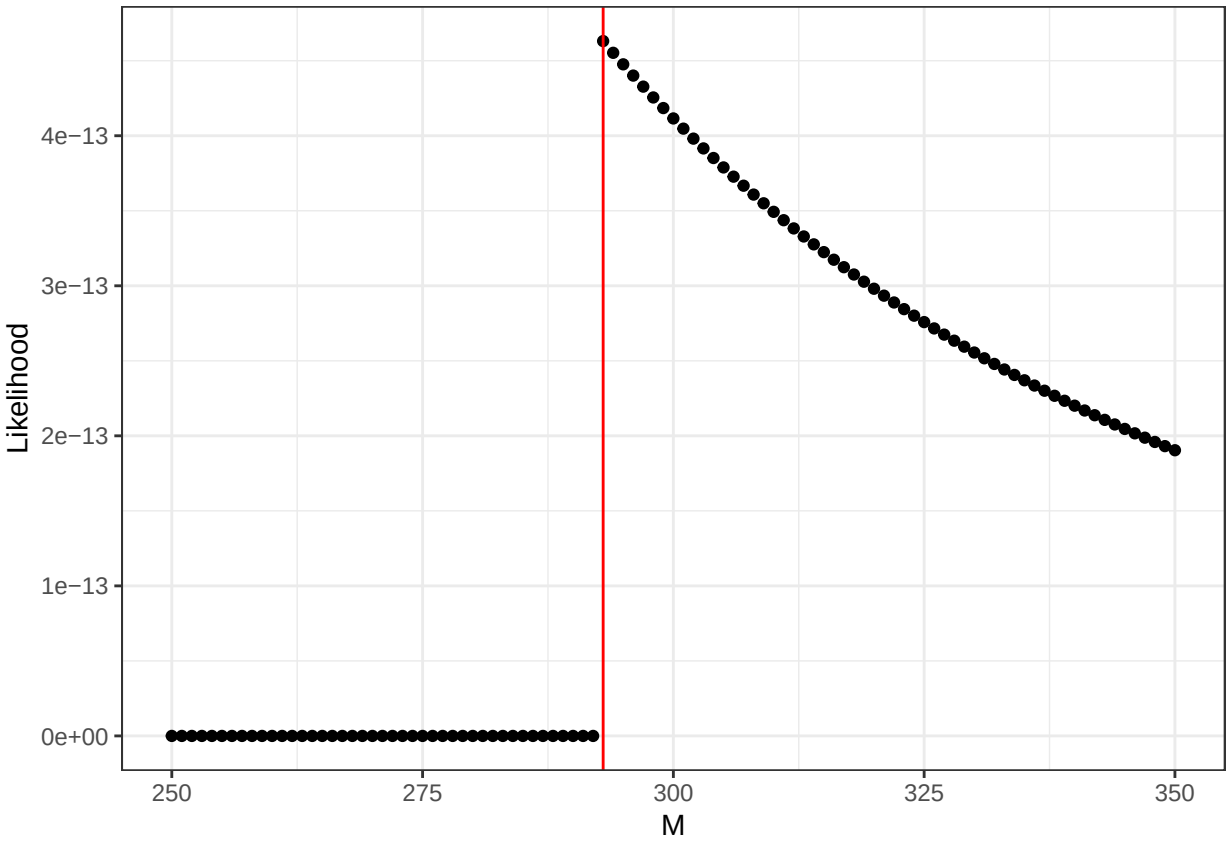
Part B

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

```
likelihood_fun <- function(M, X){
  prod((1/M)*as.integer(X <= M))
}

X_obs <- c(10, 30, 78, 293, 43)

lik_plot <- data.frame(x=250:350) %>% ggplot(aes(x=x)) + xlim(250, 350) + geom_point(aes(y=sapply(x, fu
lik_plot
```



The likelihood is zero up to $M = 292$, takes on a non-zero value at 293, and then is strictly decreasing in M.

Part C

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

From inspection of the graph in part B, we see that the maximum is a boundary point, the MLE is $\hat{M} = \max\{\mathbf{X}\}$ and in this case, our ML estimate is 293.

Part D

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

Intuitively, if the MLE is not the true parameter, then it will always be less than M. Therefore, the expectation of $\hat{M} < M$ and the MLE is biased downwards.

However, because the probability of drawing an observation $X_i = M$ goes to 1 as n goes to infinity, the MLE is consistent as the chances of drawing a sample where the sample maximum is the true maximum approach 1 as the sample size gets arbitrarily large.