The Connection between Male and Female Military Sexual Assault

3 June 2021

Benjamin Anderson
Stanford University
Dept. of Computer Science
banders9@stanford.edu

Abstract — Building on a substantial literature documenting the prevalence, causes, and health effects of sexual assault in the United States Armed Forces, I study the relationship between the risk of sexual assault for male and female members of the military. Using linear regression, I identify a clear positive relationship between the sexual assault risk faced by women and that faced by men at a given military installation. Then, I apply non-parametric regression methods, which more convincingly model the relationship when evaluated with the leave-one-out cross-validation score.

1. Introduction

Sexual violence has a long history in the United States Armed Forces [1], but it is only in the past two decades that this topic has received significant public attention. A series of high-profile sexual assault scandals at training facilities in the 1990s and early 2000s drew increased scrutiny from the public, the media, and the military itself. [2] In 2004, a Department of Defense Task Force on sexual assault concluded that a centralized organization was needed to address the issue, leading to the founding of the Sexual Assault Prevention and Response Office, or SAPRO. [3] Since then, sexual violence in the military has been studied in a variety of academic disciplines: historians have traced connected experiences of sexual violence with the changing gender dynamics of the U.S. military, while psychologists and medical researchers have studied the traumatic effects of sexual violence on members of the armed forces.

In this paper, I study the relationship between the risk of assault for male and female members of the United States Armed Forces across different military bases. Using linear regression on a dataset from the RAND Corporation [4], and then improving on linear regression with non-parametric approaches, I demonstrate a generally positive relationship between sexual assault risk for men and women. In other words, military bases where women are at higher risk for sexual assault are also more dangerous for men. This is a key finding—both practically for those in the military working to combat sexual violence, and for researchers working to understand the root causes of the epidemic of sexual violence in the armed forces.

2. Related Work

Since the 1990s and early 2000s, as sexual assault in the military has attracted more public scrutiny, there has been substantial work on the problem in the humanities and social science literature. I will focus on the social science literature here, since it is more closely to the present study. Empirical research on sexual assault in the military has generally focused on three main topics: (a) prevalence of sexual assault; (b) health effects of sexual assault; and (c) correlates and root causes of sexual assault. Researchers have estimated the rate of sexual assault for various groups, including active-duty female personnel (7.3%), [5] female VA patients

(23%), [6] male and female military reserve members (1.6% and 13.1% respectively), [7] and transgender veterans (17.2%). [8]

Many of the same studies that estimated the prevalence or frequency of sexual assault also documented health effects. Researchers have found, using various approaches (including logistic regression), that experiences of military sexual assault are associated with deleterious mental health effects, including depression and PTSD. [7][8][9] Research focused on root causes and correlates has found that deployed men had lower risk for military sexual trauma (MST) than non-deployed men, and that veterans reporting combat exposure had greater risk of MST than those without. Among women, those in the Marines and Navy had greater risk for MST than Air Force veterans. Finally, MST was significantly higher among veterans who reported using Veterans Affairs healthcare [10]. Numerous theories have been advanced to qualitatively explain the prevalence of military sexual assault, including religious and cultural influence, gender stereotypes, alcohol, and male entitlement. [11]

In this paper, I build on the existing social science literature by analyzing the association between sexual assault against men and against women across different military locations. This analysis may shed light on whether the same factors that make a location riskier for women also make it more dangerous for men. It may lend strengthen certain theories about the causes of military sexual assault, and weaken others. For instance, theories that discuss male entitlement to women's bodies would have to account somehow for why men also suffer increased victimization at locations where women are more at risk. A strong relationship here would seem to imply that some factors related to the military installation—whether it is the rules, leadership, or culture specific to that location—can either work to keep both men and women safe from sexual violence, or put them both at risk. On the other hand, the lack of a clear relationship might suggest that military sexual assault against men and women have different causes, and should be analyzed separately.

3. Data

3.1 Data Collection

Data for the present analysis comes from a 2014 study conducted by the RAND corporation. [4] That study reports prevalence of sexual assault (as a rate) broken down by gender, location, and branch of the military. The original study displayed this information as tables in a Portable Document Format, but the data was reposted by the Military Times in an HTML table more amenable to scraping. [12] I converted the data in the HTML table into a CSV, and used this as the starting point for my analysis.

3.2 Data Preparation

The original dataset has separate records for men and women: that is, the data is organized in a "long" format, with the columns "Service" (i.e. Army, Navy, etc.), "Location" (e.g. Aberdeen Proving Grounds), "Gender" (Male or Female), and "Sexual Assault Risk" (a percent). Because the comparison of interest is to compare male and female sexual assault risk at the same location, I used a "pivot" operation so that each location-service pair became one record, with associated male and female sexual assault risk. Because not all location-service records have both male and female information (probably because some branches of the military at some locations are single-sex), I removed all records with missing values for male or female sexual assault risk, leaving $n = 333$ data points, with the variables of interest being the female sexual assault rate (X) and the male sexual assault rate (Y) for a given military installation and branch.

3.3 Data Exploration & Visualization

A scatter plot of those X and Y variables is shown in Figure 1. It is quite clear that this is a positive relationship: a higher sexual assault risk for women tends to correspond to a higher risk for men.

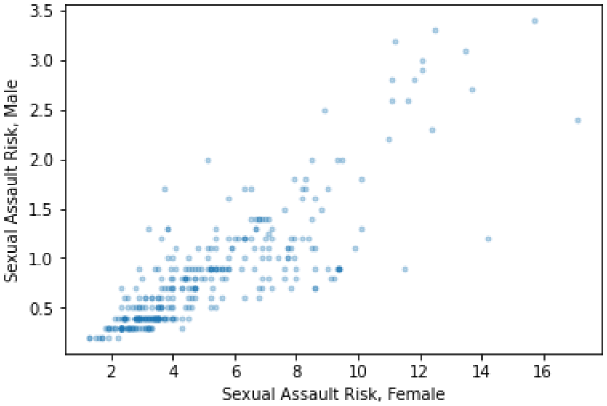

Figure 1: Scatterplot of X and Y variables

The relationship appears *roughly* linear, but we will explore both linear and non-linear models in this analysis. The mean and median rates of sexual assault for women across all military bases are 5.1% and 4.7% respectively, while the mean and median for men are 0.7% and 0.8%. (Of course, we should not read into these measures of central tendency for these aggregate rates too much, as different military installations are different sizes, ranging from around 100 people to over 200,000). In order to understand the relationship between these two variables of interest, I conduct a regression analysis, using methods explained in the following section.

4. Methods

The setting for this analysis is *regression*: I assume there is some relationship between two random variables, namely the sexual assault risk for women (X), and the sexual assault risk for men (Y), captured by some regression function $r(x)$, such that $r(x) = \mathbb{E}[Y \mid X = x]$. The goal of regression analysis is to estimate $r(x)$ by some other function $\hat{r}(x)$, based on the data, $\mathcal{D} = \{(X_1, Y_1), \ldots, X_n, Y_n\}$. Because the data look somewhat linear, I

4.1 Leave-One-Out Cross-Validation

In order to evaluate and compare different choices for the “form” of $\hat{r}(x)$, and to tune hyperparameters (such as the bandwidth in local polynomial regression) it is necessary to select some metric for the “goodness” of a given estimator, $\hat{r}(x)$ fit with the dataset $\mathcal{D}$. A naive approach would be to simply measure the prediction error on the same dataset $\mathcal{D}$ used to fit the model, with some metric like mean absolute error (MAE) or mean squared error (MSE). However, this approach does not work well when the form of the estimator can be arbitrarily complex, since a sufficiently expressive model can perfectly predict the training data (e.g. by just “memorizing” it). This means that many different models can all appear “perfect” by this metric, and worse, this measure of performance does not guarantee that the model will work well on unseen data.

A better metric measures how a fitted model performs on *unseen* data, i.e. samples of X_j, Y_j pairs from the same distribution which were not used to fit the model. When a lot of data is available, then the dataset can be partitioned into “train” and “test” datasets, where the first is used to fit the model and the second is used to validate its performance. However, when data is scarce, it can be costly to waste a substantial fraction of the dataset for validation. One data-efficient alternative to the train-test split is *cross-validation*. Cross-validation is based on the same principle of testing on unseen data, but rather than designating one subset of the data as the "testing" dataset, cross-validation breaks the data into k groups (or "folds"), and then fits k different models. Each model uses all but one of the folds as training data, and the remaining fold to measure the model's performance. The results from all folds are then averaged (or summed) to produce an aggregate performance metric.

Cross-validation with a number of folds equal to the number of data points is known as *leave-one-out cross-validation*, as the model is trained on all but one of the data points, and then evaluated on the one remaining "left-out" data point. This form of cross-validation not only provides a good measure of the error (as a model trained on more of the data ought to be a better approximation of a model trained on all of it), but for some models, including the non-parametric models discussed below, there are computational shortcuts that make calculating the leave-one-out cross-validation score very efficient (relative to doing it naively). [13]

For the regression methods described in this paper, I evaluate performance using the predicted residual sum of squares (PRESS) statistic, a form of leave-one-out cross-validation which calculates the sum of squared errors for each "left-out" point. [14]

$$PRESS = \sum_{i=1}^{n} (Y_i - \hat{r}_{(-i)}(X_i))^2$$

Here, $\hat{r}_{(-i)}$ refers to a model fit on all data points other than the i-th data point.

4.2 Ordinary Least Squares

Linear regression, or ordinary least squares (OLS), assumes that the relationship between the covariate(s) X, and the response variable Y, is a linear function, plus some noise ϵ which is independently and identically distributed according to a Gaussian distribution with zero mean and fixed variance. That is, for each pair in the dataset, (X_i, Y_i), we have:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad\qquad \epsilon_i \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$$

If these conditions are met, then the choice of β_0, β_1 which minimizes the residual sum of squares (RSS), or equivalently the mean squared error (MSE), is also the maximum-likelihood estimator for β_0, β_1 given the observed data. [15]

4.3 Non-Parametric Regression and Linear Smoothers

Not all data exhibit a linear relationship, or meet the assumptions for linear regression. In such cases, a non-linear model can better explain the relationship between the covariates and the response variable. In our case, it is quite obvious from the scatter plot in Figure 1 that the data do not exhibit uniform variance for all values of X (i.e. homoskedasticity), which is one of the assumptions of OLS. (It is possible to modify linear regression to account for this.) Moreover, it's difficult to tell just from a scatter plot whether the data really are linear, or whether a non-linear model would be a better fit. In the following sections, I discuss two non-linear, non-parametric approaches for regression analysis.

Some non-linear models are *parametric*, just as linear regression is: a certain functional form $r(x)$ is assumed, and then some objective function (such as RSS) is used to select parameters to fit an estimator $\hat{r}(x)$. Other non-linear regression models are *non-parametric*, which means they make fewer assumptions, and in particular do not assume a particular functional form (e.g. quadratic, logistic, exponential) for the regression function $r(x)$; instead, they assume something more general, such as a certain degree of smoothness.

One class of non-parametric models called *linear smoothers*, which have the property that the prediction function $\hat{r}(x)$ can be written as a linear combination of all Y_i's in the dataset, by coefficients $\ell_1(x), \ldots, \ell_n(x)$, where ℓ is some function of x (which does not need to be linear). [13] The following non-parametric approaches fall into the category of linear smoothers.

4.3 Locally-Weighted Linear Regression

Locally-weighted linear regression is a type of local polynomial regression, which is a non-parametric regression method. The motivation for this approach is straightforward: rather than assuming that the regression function r is globally linear, as OLS does, we essentially assume it is *locally* linear, i.e. linear over a small window. As an analogy, imagine trying to approximate a circle, using only straight lines. If you could only draw four lines, you'd have a square—not very circular. But if you could draw a bunch of small line segments, e.g. 30 or 50 or 1000, then you'd get something that looked much closer to a circle, since over a very small segment of the curve, a circle is close-ish to a line.

Local linear regression takes this idea to the extreme, essentially fitting a locally-weighted line at every possible x. Of course, this is never computed at every possible value of x; it is only necessary to compute $\hat{r}(x)$ at points of interest. For this reason, local linear regression is a form of "lazy learning", as there is no real "fitting" process; the training data is just memorized. [16] In order to make a prediction for some input x, a line is fit to the data, with more importance given to points nearer to x. More formally, each point in the training data, X_i, is given a weight, $w(X_i)$, based on its proximity to x (I use a Gaussian kernel with bandwidth h). Then, a line is selected to minimize the *weighted* sum of squares:

$$\hat{\beta}_0, \hat{\beta}_1 = \underset{\beta_0, \beta_1}{\text{argmin}}\ J(\beta_0, \beta_1) = \underset{\beta_0, \beta_1}{\text{argmin}}\ \sum_{i=1}^{n} w(X_i) \cdot (Y_i - [\beta_0 + \beta_1 \cdot X_i])^2$$

Finally, that line is used to make a prediction, *only* at that x: $\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$. This is considered a non-parametric method, because the computation depends on the number of training data points. (In fact, it is also a linear smoother.) [13] I wrote the code to do these computations from scratch in Python. In order to choose the bandwidth h (which determines the "size" of the neighborhood to which the local line is fit), I compare the PRESS statistic (cross-validation score) for different, equally-spaced choices of bandwidth, and select the bandwidth that achieves the lowest PRESS.

4.4 Smoothing Spline

The smoothing spline is a different non-parametric method, with slightly different assumptions than local linear regression. Rather than assuming some measure of local linearity, which is the same across the entire dataset, the motivation for smoothing splines is an objective function which both rewards fitting the data well, and penalizes lack of smoothness (which formally depends on the first and second derivatives of the regression function). In general terms, we aim to find a function which minimizes the following equation:

$$M(\lambda) = \sum_{i} (Y_i - \hat{r}(X_i))^2 + \lambda J(r)$$

...where $J(r)$ is some "roughness penalty." [13] The precise result depends on exactly the type of penalization used, but one common approach is to fit a continuous, piecewise-cubic function that interpolates the data, and has continuous first and second derivatives. I used the `UnivariateSpline` package from Python's `scipy` library, which achieves a similar effect, although the implementation does not exactly match the penalized regression form here. (Instead of a smoothness penalty, the hyperparameter `s` controls the number of "knots", which determines the number of cubic "pieces"). [17] Just as the roughness penalty controls the tradeoff between fitting the data well vs. a smoother function, a larger number of knots allows a more complex function that fits the data better (reducing bias), while a smaller number of knots constrains the function more (reducing variance). In order to choose the best value for `s`, I compare the PRESS statistic for different, equally-spaced choices of `s`, and select the the one achieves the lowest PRESS.

5 Results

5.1 Linear Regression

Linear regression achieved a PRESS statistic of 35.8. In the linear regression setting, it also makes sense to compute R^2, the coefficient of determination, which is intepreted as the fraction of the variation in the data that is explained by the linear relationship. I also calculated R, (Pearson's correlation) which measures both the direction and the strength of the relationship between the covariate and response variables. These results, along with a plot of the fitted line, are shown in Figure 2.

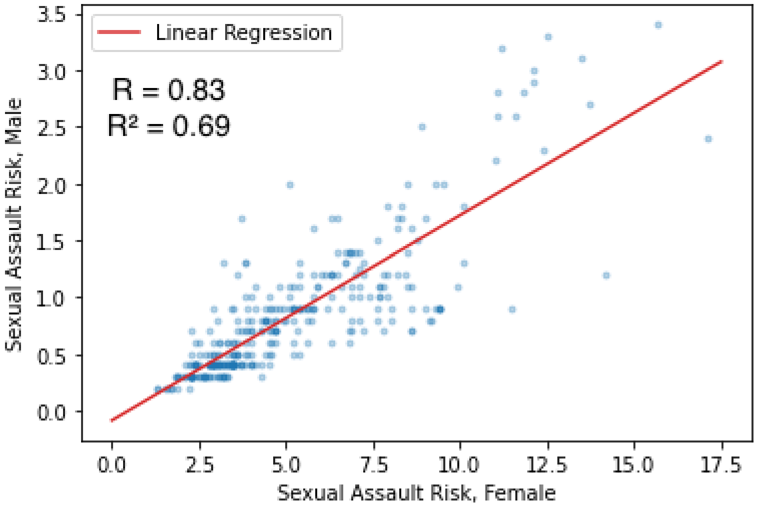

Figure 2: Linear Regression

The residuals for this model are shown in Figure 3. If it was not already clear from the plot of the fitted line, there is considerable heteroskedasticity. Depending on the setting, this may call for transforming a variable, or using a more robust linear regression model. However, this is beyond the scope of the current analysis, since I am primarily treating the linear model as a baseline against which to compare non-parametric regression models.

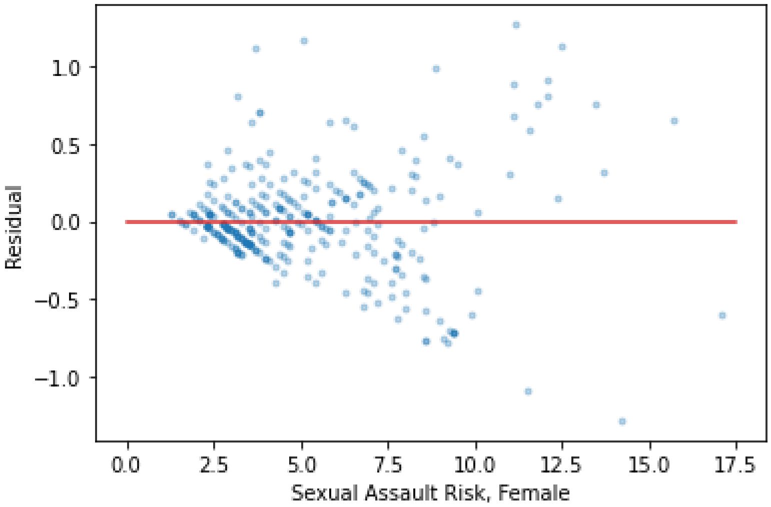

Figure 3: Residuals for Linear Regression Model

5.2 Non-Parametric Regression

As described in the Methods section, locally-weighted linear regression and cubic splines were fit to the data, with hyperparameters h and s respectively tuned by cross-validation (PRESS statistic). The results are summarized in the table below.

Method	Hyperparameter	PRESS
Linear Regression	N/A	35.8
Locally-Weighted Linear Regression	$h = 2.03$	34.8
Cubic Smoothing Spline	$s = 12.07$	32.2

Figure 4: Table, Linear vs. Non-Parametric Models

Both of the non-parametric models perform better than linear regression, with the cubic smoothing spline performing best of all. Whether this performance gain is *meaningful* is another question. The PRESS statistic is the sum of squared errors over *all* data points, so a reduction by 3.2 corresponds to a reduction of around 0.01 (3.2 divided by 333 data points) in the predicted squared error for a single data point. We can also inspect the fitted models, shown in Figure 4, to assess the goodness of fit.

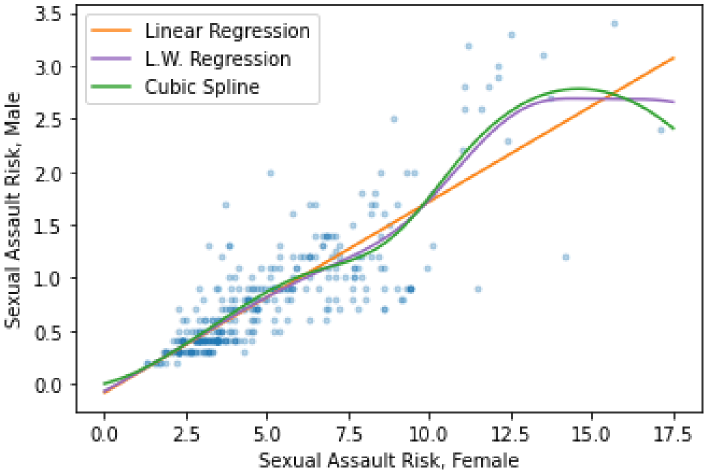

Figure 5: Graph, Linear vs. Non-Parametric Models

It seems, at least from visual inspection, that the non-parametric models fit the data more closely, without introducing too much variance (or "wiggliness"). It appears that the relationship really is basically linear for X between 0 and 7.5, but then for larger X, there are some gains from non-linearity. It is also notable that, although cubic splines outperform locally-weighted linear regression when measured by the PRESS statistic, the fitted models look nearly identical. I suspect that the smoothing spline may have lower variance, which would cause its performance to suffer less when a point is omitted for cross-validation. Overall, the non-parametric models still support the same general conclusion as the linear regression model, namely that a higher rate of sexual assault among female service members is associated with a higher rate of sexual assault among male service members.

Discussion

The main conclusion from the experiments conducted in this analysis is that there is a clear positive relationship between sexual assault risk for female service members, and male service members at the same location. This is reflected in the correlation coefficient (Pearson's $R = 0.83$), the coefficient of determination ($R^2 = 0.69$), and the fitted models, both linear and non-linear/non-parametric, which all generally indicate a positive relationship.

As briefly discussed before, this conclusion is a significant addition to existing work on the subject of military sexual assault. Much of the existing literature discusses causes and effects of military sexual assault, and even some work on gender-specific health outcomes, but there are, to my knowledge, no studies that analyze the significant variation in the risk of sexual violence across military installations, and that identify the relationship between the sexes with respect to sexual assault risk demonstrated in this paper. This finding is important, because it adds to the understanding of both officers working to eradicate sexual assault in the military, and scholars working to understand its root causes and possible solutions.

Based on these results, I would suggest that there may be some underlying factors that make the *same* location riskier for both men and women. Some work on sexual assault in the military describes how the rigid hierarchy can make accountability difficult. Superior officers have a great deal of authority over those under their command, and this authority can be abused, used to sweep complaints under the rug. It is therefore plausible that leadership makes a big difference—and that military installations under the eye of officers who take the issue more seriously have lower rates of sexual assault for both men and women. On the other hand, locations with leadership who would rather cover it up and ignore the problem would result in worse outcomes for both men and women.

Of course, there are plenty of other possible causes that might be a common factor behind the victimization of both men and women—for instance, beyond the leadership at the very top, the *culture* at a military base (i.e. a culture of violence vs. a culture of respect). More work would be required to confirm or cast doubt on any particular theory, but my work suggests that research into root causes ought to consider *common* causes that explain why the victimization of men and women go hand in hand.

Code

You can find all data, data preparation code, and analysis code at https://github.com/andersonbcdefg/stats205-final-project.

References

[1] Meyer, Leisa D. (1998). Creating GI Jane: Sexuality and Power in the Women's Army Corps during World War 2. Columbia U.P.

[2] Turchik, J. A., & Wilson, S. M. (2010). Sexual assault in the U.S. military: A review of the literature and recommendations for the future. Aggression and Violent Behavior, 15(4), 267–277. doi:10.1016/j.avb.2010.01.005.

[3] United States Department of Defense Sexual Assault Prevention and Response (2021). Mission and History. https://sapr.mil/mission-history.

[4] Morral, Andrew R., Terry L. Schell, Matthew Cefalu, Jessica Hwang, and Andrew Gelman (2018). Sexual Assault and Sexual Harassment in the U.S. Military: Annex to Volume 5. Tabular Results from the 2014 RAND Military Workplace Study for Installation- and Command-Level Risk of Sexual Assault and Sexual Harassment. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/research_reports/RR870z8.html.

[5] WOLFE, J., SHARKANSKY, E. J., READ, J. P., DAWSON, R., MARTIN, J. A., & OUIMETTE, P. C. (1998). Sexual Harassment and Assault as Predictors of PTSD Symptomatology Among U.S. Female Persian Gulf War Military Personnel. Journal of Interpersonal Violence, 13(1), 40–57. doi:10.1177/088626098013001003.

[6] SKINNER, K. M., KRESSIN, N., FRAYNE, S., TRIPP, T. J., HANKIN, C. S., MILLER, D. R., & SULLIVAN, L. M. (2000). The Prevalence of Military Sexual Assault Among Female Veterans' Administration Outpatients. Journal of Interpersonal Violence, 15(3), 291–310. doi:10.1177/088626000015003005.

[7] Street, A. E. (2008). Sexual harassment and assault experienced by reservists during military service: Prevalence and health correlates. The Journal of Rehabilitation Research and Development, 45(3), 409–420. https://doi.org/10.1682/jrrd.2007.06.0088.

[8] Beckman, K., Shipherd, J., Simpson, T., & Lehavot, K. (2018). Military Sexual Assault in Transgender Veterans: Results From a Nationwide Survey. Journal of Traumatic Stress, 31(2), 181–190. https://doi.org/10.1002/jts.22280.

[9] Schry, A. R., Hibberd, R., Wagner, H. R., Turchik, J. A., Kimbrel, N. A., Wong, M., Elbogen, E. E., Strauss, J. L., & Brancu, M. (2015). Functional correlates of military sexual assault in male veterans. Psychological Services, 12(4), 384–393. https://doi.org/10.1037/ser0000053.

[10] Barth, S. K., Kimerling, R. E., Pavao, J., McCutcheon, S. J., Batten, S. V., Dursa, E., Peterson, M. R., & Schneiderman, A. I. (2016). Military Sexual Trauma Among Recent Veterans. American Journal of Preventive Medicine, 50(1), 77–86. https://doi.org/10.1016/j.amepre.2015.06.012.

[11] Castro, C. A., Kintzle, S., Schuyler, A. C., Lucas, C. L., & Warner, C. H. (2015). Sexual Assault in the Military. Current Psychiatry Reports, 17(7). https://doi.org/10.1007/s11920-015-0596-7.

[12] Copp, Tara. "Sexual Assault Risk at Your Military Base: Here's a Searchable Database." (2018). Military Times. www.militarytimes.com/news/your-military/2018/09/25/sexual-assault-risk-at-your-military-base-heres-a-searchable-database/.

[13] Wasserman, Larry (2006). Nonparametric Regression. In: All of Nonparametric Statistics (pp. 61–123). Springer New York. https://doi.org/10.1007/0-387-30623-4_5.

[14] Wikipedia contributors. (2021, January 11). PRESS statistic. In Wikipedia, The Free Encyclopedia. Retrieved June 3, 2021, from https://en.wikipedia.org/w/index.php?title=PRESS_statistic&oldid=999789869.

[15] Wasserman, Larry (2010). Linear and Logistic Regression. In: All of Statistics: A Concise Course in Statistical Inference (pp. 209–229). Springer New York. https://doi.org/10.1007/978-0-387-21736-9.

[16] Atkeson, C. G., Moore, A. W., & Schaal, S. (1997). Artificial Intelligence Review, 11(1/5), 11–73. https://doi.org/10.1023/a:1006559212014.

[17] SciPy 1.0 Contributors. (2020) Univariate Spline. In: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17(3), 261-272. https://docs.scipy.org/doc/scipy/reference/generated/scipy.inter