Duke University Synthetic Data Project
Threshold Verification Measure
Effect of Partition Size on Distribution of Measures Observed in Regression Models
1/10/2018, T. Balmat

In computing the threshold verification measure for regression models using authentic OPM data, observations are partitioned such that a uniform number of employees are represented in each partition and each employee appears in one partition only. A single model is fit to each partition and the proportion of partitions with parameter of interest having an estimate beyond the specified threshold value is reported.[1] A model that has been used in race pay disparity analysis is

$$y = \beta_0 + \beta_{race} + \beta_{age} \times age + \beta_{age^2} \times +age^2 + \beta_{ed} \times ed_{years} + \beta_{bureau} + \beta_{occ} + \beta_{year} \quad (1)$$

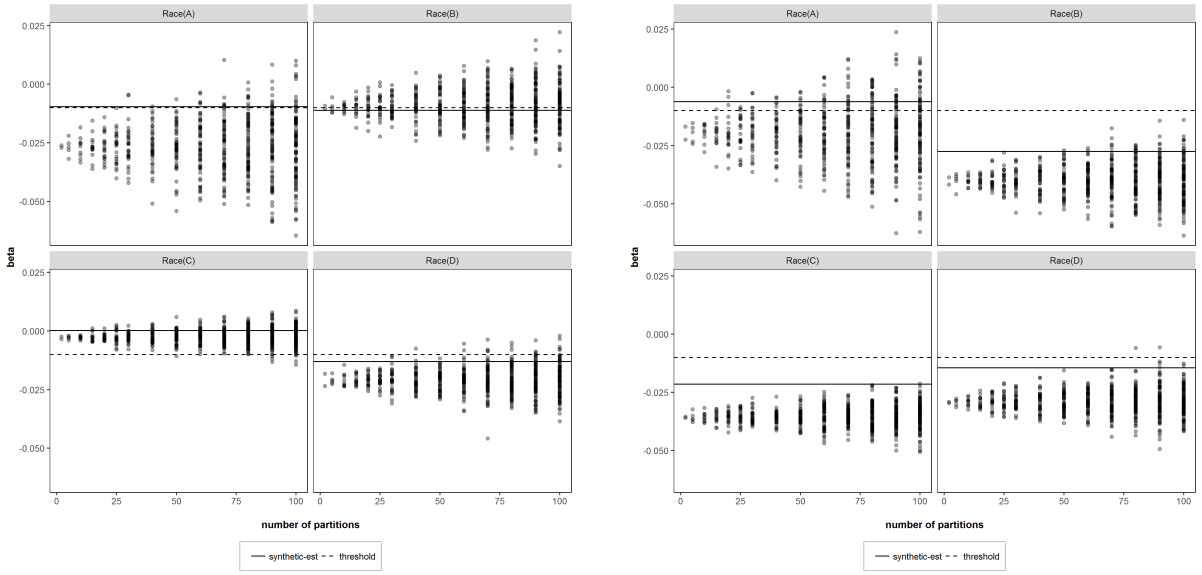
where y is the logarithm of *basic pay* and *race*, *sex*, *age*, *education*, *bureau*, *occupation*, and *year* are independent covariates. *Bureau*, *occupation*, and *year* are used as controlling fixed effects. To account for interaction between *sex* and *race*, separate models are fit to female and male subsets of observations.

Recall that, assuming homoskedasticity of residuals, the variances of regression parameter estimates are on the diagonal of the covariance matrix, $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Assuming that employees are randomly assigned to partitions and that assignment does not bias independent variable distribution, a k-fold increase in number of partitions causes a $1/k$ decrease in number of employees per partition and a corresponding $1/k$ reduction in the sums of independent variable products in $\boldsymbol{X}'\boldsymbol{X}$.[2] This reduction requires a k-fold increase in $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ in order for the identity $(1/k)\boldsymbol{X}'\boldsymbol{X}[k(\boldsymbol{X}'\boldsymbol{X})^{-1}] = \mathbf{I}$ to be maintained. Given that σ^2 is constant, then, a k-fold increase in parameter estimate variance is expected to occur with a k-fold increase in partitions. A result of this is that the number of partitions observed beyond threshold is dependent to some extent on choice of partition count. Consider a case where all partition parameter estimates are near, but within, the specified threshold. Doubling the number of partitions and refitting the model would double parameter estimate variance, possibly causing a number of partitions to be observed with estimates beyond threshold. Figure 1 shows, for race disparity model (1), female and male, the effect of partition count on the computed verification measure (partitions beyond threshold). As described, an approximate doubling of variance is observed with each doubling of partitions, giving a $\sqrt{2}$, or 1.4, increase in standard error and dispersion.

The threshold measure reports proportion partitions at threshold, not the count. However, variance bias due to choice of partition count also biases proportion at threshold. Figure 2 shows a general pattern of decreasing proportion at threshold with increasing number of partitions. This is consistent with the increased dispersion of parameter estimates with increase in number of partitions. The important point is that using the same data set and model we observe different results by changing an algorithmic parameter of the analysis (number of partitions).

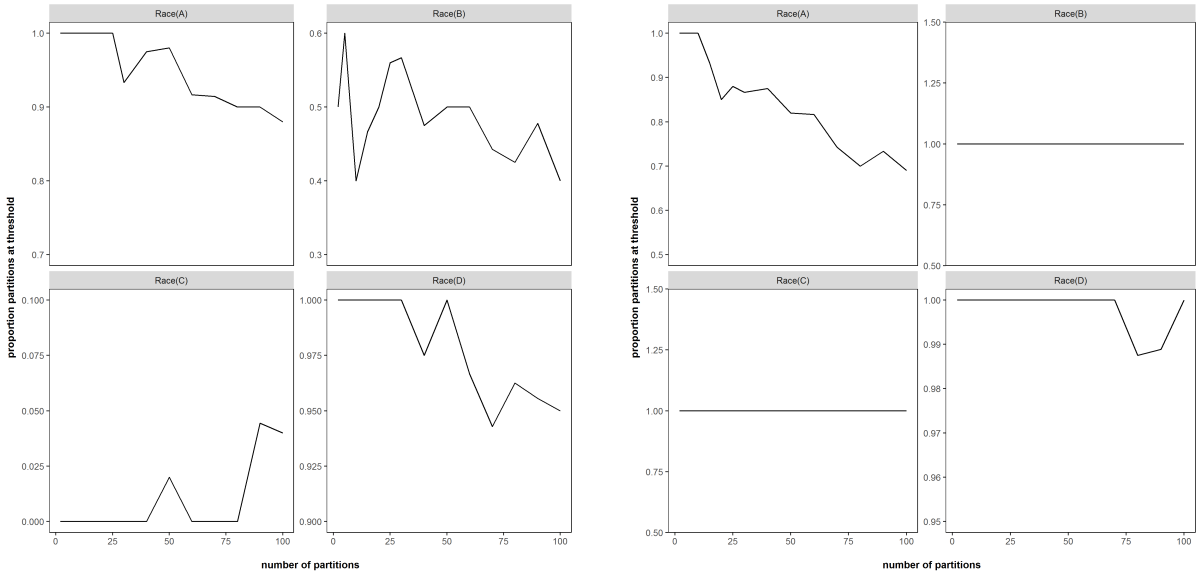
[1] For purposes of this discussion, we ignore the privacy preserving statistical noise that is added to partition counts prior to reporting.

[2] It is reasonable to expect each employee to represent a restricted region for various predictors, but random assignment of a large number of employees to a partition is expected to maintain an unbiased representation of the population within the partition.



(a) Female (b) Male

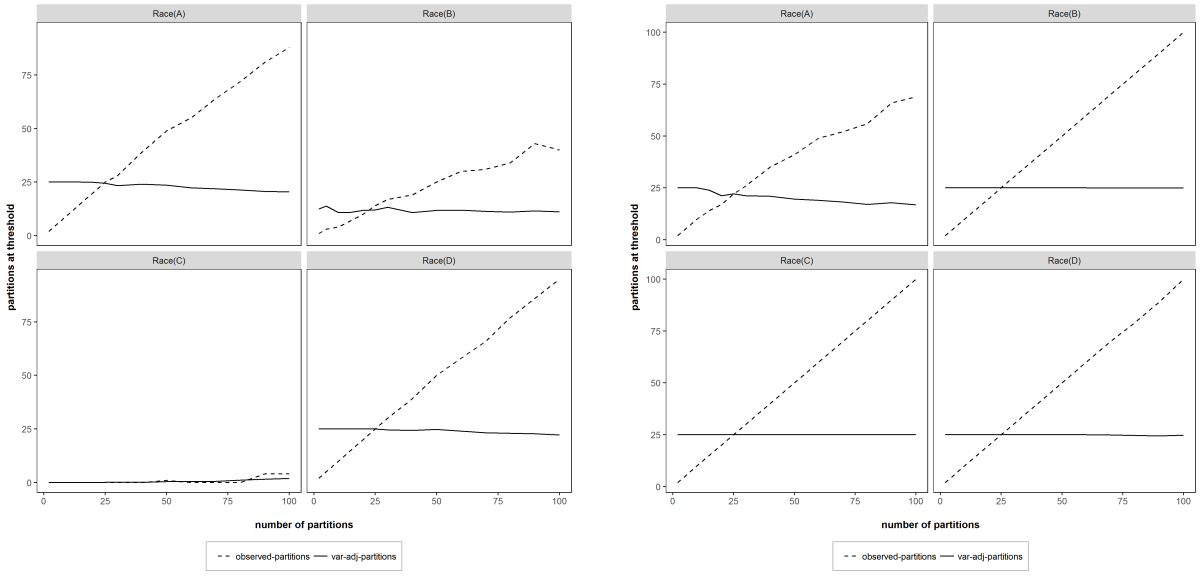
Figure 1: Partition parameter estimate distributions for race disparity fixed effects model. Number of partitions on x-axis. Dashed line is threshold. Solid line, given as reference, is disparity estimate from entire synthetic data set.



(a) Female (b) Male

Figure 2: Proportion partitions at threshold for race disparity model. General pattern of decreasing proportion at threshold with increasing partition count.

To compensate for potential bias from selecting a number of partitions, the k-fold variance relationship can be used to estimate a *normalized* parameter variance estimate for each partition. Say that, for privacy concerns, the minimum allowable number of partitions is M_0 and the partition count chosen is $M > M_0$. Then variance estimates normalized to M_0 are the observed partition variances multiplied by M/M_0. Alternatively, the number of observations, n_{M_0} and n_M can be used, in which case partition variances are multiplied by n_{M_0}/n_M. Using the normalized variance estimates and assuming normally distributed parameter estimates, the proportion of expected distribution at threshold for each partition can be computed. Accumulating the proportions gives an estimate of the number of partitions at threshold. This is demonstrated in figure 3, for M_0=25, using the same partitions as were used to compose figure 1. Dashed lines indicate actual partitions at threshold, while solid lines indicate expected number of partitions at threshold, given normally distributed parameter estimates with variance normalized to a partition count of M_0=25. The relatively constant number of expected partitions at threshold across actual partition counts (x-axis) implies lack of bias.



(a) Female (b) Male

Figure 3: Estimated partitions at threshold using normally distributed parameter estimates with variance scaled to 25 partitions. Solid line is estimated partitions at threshold using variance estimate. Dashed line is actual.

Choice of partition count also affects the number of partitions with singular design matrices, designs with one or more linearly dependent columns. When this occurs, some combination of model parameters are confounded and cannot be independently estimated. It is seen from table 1 that the number of singular designs increases with partition count or, alternatively, increases with decreasing partition observation count. This seems to violate the assumption of population characteristic representation by random assignment of employees to partitions. Yet, a plausible explanation may be that the probability of selecting a group of employees with some identical career aspect (*year* and *agency* or *agency* and *occupation*) in all observations is greater with small partitions than with large ones. When a singular design is encountered, one linearly dependent column is randomly selected and omitted from the model, then the reduced model is fit to the partition data. Dependencies generally occur in *bureau* and *occupation* columns and do not affect *race-pay-disparity* estimates. Elimination of a fixed effect level effectively combines, or confounds it, with the associated reference level and alters the strict notion of controlling fixed effects. An interesting property of table 1 is that males tend to exhibit a greater number of singular partitions than do females. Since female and male employee counts are approximately equal, this might indicate greater diversity of career fixed effect level combinations (*bureau*, *occupation*, *year*) for females.

Table 1: Partitions with nonsingular design matrices. Estimates computed after removal of all but one dependent column. No dependencies detected in race columns.

n-Partitions	n-Singular-Female	n-Singular-Male
2	0	0
5	0	0
10	0	0
15	0	0
20	0	1
25	0	1
30	1	0
40	2	4
50	3	4
60	2	16
70	7	30
80	3	35
90	14	38
100	18	42

CONCLUSION

Computed threshold verification measures are sensitive to choice of partition count, but sensitivity can be mitigated by normalizing parameter estimate standard errors to those of a reference partition count. Verification measure results must be interpreted with respect to the number of partitions used since a result of k partitions at threshold could be due either to the magnitude of parameter estimates in k partitions or to high variance parameter estimates due to partition size. Once a standard partition count is chosen, a question arises as to what, aside from being required for statistical purposes, the standard means. Knowing that a minimum number of partitions is required for privacy concerns, while a maximum is required for variance concerns, perhaps a central partition count can be specified that satisfies both requirements such that verification results can be said to be with respect to optimal privacy and variance properties of analyses done with the data set. A complication arises with sub-setting the data, since a smaller data set will likely have an optimal partition count that is different than that of a larger data set. Further, different subsets (*occupation categories*, *years*, etc.) could have different subset sizes, requiring different optimal partition counts, which adds complexity to interpretation of results.