Documentation for wyoung

December 2024

wyoung is a Stata command designed to control the family-wise error rate when performing multiple hypothesis tests. This document outlines the algorithm employed by **wyoung** and presents simulation results that demonstrate its effectiveness across various settings. To install the command and access its help file, type **ssc install wyoung, replace** at the Stata prompt. The latest development version is available for download from Github: https://github.com/reifjulian/wyoung

After installation, type **help wyoung** at the Stata prompt to view examples and learn the syntax. Companion Stata code for the simulations described below is available on Github: https://reifjulian.github.io/wyoung/documentation/simulations/wyoung_simulations.do

wyoung was originally developed for use in the Illinois Workplace Wellness Study. Please cite the command as Jones, Molitor and Reif (2019):

Jones, Damon, David Molitor, and Julian Reif. "What Do Workplace Wellness Programs Do? Evidence from the Illinois Workplace Wellness Study." *Quarterly Journal of Economics*, November 2019, 134(4): 1747–1791.

1 Methodology

Multiple hypotheses arise when there are multiple outcomes, subgroups, or independent parameters of interest. Consider testing $K > 1$ distinct null hypotheses. The family-wise error rate (FWER) is the probability of rejecting at least one true null hypothesis—commonly referred to as making "false discovery"—within this "family" of K hypotheses. A procedure is said to provide *strong control* of the FWER if it maintains the error rate at or below a specified level regardless of how many of the K hypotheses are true. In contrast, *weak control* of the FWER applies only under the assumption that all K hypotheses are true, i.e., when the complete null hypothesis holds.

wyoung controls the FWER using the free step-down resampling method of Westfall and Young (1993) (Algorithm 2.8, pp. 66–67). This method leverages resampling techniques, such as bootstrapping (sampling with replacement) or permutation (shuffling), to adjust the standard p-values obtained from model estimation. A detailed description of the algorithm is provided below.

1.1 Bootstrapping

The bootstrapping procedure involves the following steps:

1. Estimate $\{\widehat{\beta}_1, \widehat{\beta}_2, ..., \widehat{\beta}_K\}$. Calculate the conventional, unadjusted p-values $\{p_1, p_2, ..., p_K\}$ for individual tests of the null hypotheses $\widehat{\beta}_k = 0$.[1] Without loss of generality, assume the p-values are indexed such that $p_1 \leq p_2 \leq ... \leq p_K$.

2. Draw with replacement from the dataset to create a bootstrap sample.

   (a) Compute the bootstrap estimates $\{\widehat{\beta}^*_{i1}, \widehat{\beta}^*_{i2}, ..., \widehat{\beta}^*_{iK}\}$. Calculate the conventional, unadjusted p-values $\{p^*_{i1}, p^*_{i2}, ..., p^*_{iK}\}$ for individual tests of the null hypotheses $\widehat{\beta}^*_{ik} = \widehat{\beta}_k$. The k index here corresponds to the ranking computed in step 1. It will not generally be the case that $p^*_{i1} \leq p^*_{i2} \leq ... \leq p^*_{iK}$.

   (b) Enforce monotonicity with respect to the original ordering in step 1 by computing the successive minima:

$$\begin{aligned} q^*_{iK} &= p^*_{iK} \\ q^*_{i,K-1} &= \min(q^*_{iK}, p^*_{i,K-1}) \\ q^*_{i,K-2} &= \min(q^*_{i,K-1}, p^*_{i,K-2}) \\ &\vdots \\ q^*_{i1} &= \min(q^*_{i2}, p^*_{i1}) \end{aligned}$$

3. Repeat step 2 N times. For each bootstrap sample i and hypothesis k, define the indicator $COUNT_{ik} = 1$ if $q^*_{ik} \leq p_k$ and 0 otherwise.[2]

[1] Alternative hypotheses are also possible, including combinations of coefficients (see Section 2.3).

[2] To compute "single-step" p-values instead of "step-down" p-values, define the indicator $COUNT_{ik} = 1$ if $\min\{p^*_{i1}, p^*_{i2}, ..., p^*_{iK}\} < p_k$ and 0 otherwise. Resampling-based single-step methods often control family-wise type III (sign) error rates. Whether their step-down counterparts also control type III error rates is unknown (Westfall and Young, 1993, p. 51).

4. For each hypothesis $k = 1, 2, ..., K$, calculate the fraction of successive minima that were lower than the original p-value:

$$r_k = \frac{1}{N} \sum_{i=1}^{N} COUNT_{ik}$$

5. Enforce monotonicity using successive maximization to calculate the adjusted p-value:

$$\begin{aligned} p_1^{adj} &= r_1 \\ p_2^{adj} &= \max(r_1, r_2) \\ &\vdots \\ p_K^{adj} &= \max(r_{K-1}, r_K) \end{aligned}$$

This resampling algorithm provides strong control of the FWER under the condition of subset pivotality, a multivariate generalization of pivotality.[3] Subset pivotality requires that the joint distribution of any subvector of p-values remains unaffected by the truth or falsehood of hypotheses corresponding to p-values not included in the subvector. This condition is satisfied in many settings, including significance testing for coefficients in a general multivariate regression model with possibly non-normal or heteroskedastic errors (Westfall and Young, 1993, pp. 122–123). If subset pivotality does not hold, the procedure provides only weak control of the FWER. In this case, the adjustment is valid solely for the smallest p-value. One notable example of subset pivotality failure arises in tests involving overlapping subgroups. Consider a scenario where hypothesis 1 tests whether an effect exists for the entire group, and hypothesis 2 tests whether the effect exists specifically for women. If hypothesis 2 is true, hypothesis 1 must also be true, creating a dependency between the hypotheses that violates the subset pivotality condition.[4]

It is possible for this algorithm to produce adjusted p-values that are smaller than unadjusted p-values. For instance, in the extreme case where only one bootstrap sample is used ($N = 1$ in steps 3 and 4), all adjusted p-values are either zero or one. Those equal to zero will, of course, be smaller than the unadjusted values. To avoid this issue, we recommend using a large number of bootstraps. Westfall and Young (1993) recommend at least 10,000 bootstrap draws. If adjusted p-values consistently lie below the unadjusted p-values, even when the number of bootstraps is large, this may indicate model misspecification. For example, in simulations with clustered errors (described below), we found that adjusted p-values were often smaller than unadjusted values if a cluster bootstrap was not employed.

1.2 Permutation

The permutation procedure follows the bootstrapping approach described above.[5] However, it includes one key modification in step 2(a), where a sharp null hypothesis is tested:

[3] The sampling distribution of a pivotal statistic does not depend on the specific distribution that generated the data; the t-statistic is a common example.

[4] In principle, it is possible for an effect in one subgroup to perfectly offset the effect in other subgroups, resulting in no overall effect. However, if treatment effects are heterogeneous and drawn from a continuous distribution, the probability of such an exact cancellation is zero.

[5] Thanks to Adam Sacarny for helping implement the permutation procedure.

2. Permute (shuffle) the data under the null hypothesis.

   (a) Compute the shuffled estimates $\{\widehat{\beta}^*_{i1}, \widehat{\beta}^*_{i2}, ..., \widehat{\beta}^*_{iK}\}$. Calculate the conventional, unadjusted p-values $\{p^*_{i1}, p^*_{i2}, ..., p^*_{iK}\}$ for individual tests of the sharp null hypotheses $\widehat{\beta}^*_{ik} = 0$. The k index here corresponds to the ranking computed in step 1. It will not generally be the case that $p^*_{i1} \leq p^*_{i2} \leq ... \leq p^*_{iK}$.

Permutation breaks the link between the shuffled variable and the outcome, producing a sharp null hypothesis that assumes an exact treatment effect of zero for all observations. In **wyoung**, permuting a variable by default also severs its association with all other covariates. However, users can optionally specify that multiple variables be permuted jointly if the analysis requires preserving their relationship.

As with bootstrapping, this permutation algorithm requires subset pivotality in order to provide strong control of the FWER. Westfall and Young (1993) note that this condition must hold exactly for permutation analyses "provided that only one test per [outcome] variable is performed" (p. 115).

The hypothesis tests performed in step 2(a) (and 1(a)) rely on standard normal-theory approximations rather than permutation. While permutation could theoretically be used to compute these tests, doing so would scale computational complexity with order N^2 rather than N, making the algorithm impractical for many applications. To assess whether this simplification compromises validity, we recommend comparing unadjusted p-values derived from permutation with those based on normal-theory approximations. Significant discrepancies between the two indicate that the adjusted p-values may be unreliable.

2 Simulations

We conducted simulations to evaluate the effectiveness and statistical power of the resampling algorithm described in Section 1.1. Let μ be a ten-dimensional zero vector $(0, 0, ..., 0)'$. Let I be a 10×10 identity matrix. Let Σ be a 10×10 covariance matrix where all off-diagonal elements are equal to 0.9. The data-generating process for each simulation scenario is described below:

1. Normal i.i.d. errors (ten outcomes)

   $e \sim \mathcal{N}(\mu, I)$

   $Y = e$

2. Normal i.i.d. errors (one outcome, ten subgroups)

   $e \sim \mathcal{N}(0, 1)$

   $Y = e$

3. Correlated errors (ten outcomes)

   $X \sim \mathcal{N}(\mu, I)$

   $e \sim \mathcal{N}(\mu, \Sigma)$

   $Y = 0.2X + e$

4. Lognormal, mean-zero i.i.d. errors (ten outcomes)[6]
   $e \sim \exp[\mathcal{N}(\mu, I)] - \sqrt{\exp[1]}$
   $Y = e$

We simulated 2,000 datasets for each of these four data-generating processes. In each of these 2,000 simulations, we estimated a series of 10 regressions:

$$Y_i = \alpha + \beta_i X_i + \varepsilon_i, i = 1...10.$$

The sample size for each regression was 100. The regressor $X_i \sim N(0, 1)$ in scenarios 1, 2, and 3. In scenario 4, the regressor is just a constant equal to one (α is omitted). There are ten null hypotheses that correspond to these ten regressions: $\beta_i = 0, i = 1, ..., 10$. These ten null hypotheses are all true in scenarios 1, 2, and 4; the hypotheses are all false in scenario 3 (correlated errors).

Table 1 compares the effectiveness of the Westfall-Young resampling algorithm to other well-known multiple inference adjustment methods.[7] Each column in the table reports how often at least one null hypothesis was rejected using each adjustment method. When outcomes are independent and normally distributed, the probability that at least one of the ten hypotheses is statistically significant is equal to $1 - (1 - 0.05)^{10} = 0.401$. This calculation accords well with the simulation: the first row of column (1) reports that at least one of the ten hypotheses was rejected at $\alpha = 0.05$ in 39.8 percent of the 2,000 simulations when no adjustment was performed. By contrast, the Bonferroni-Holm, Sidak-Holm, and Westfall-Young adjustments reject at least one null hypothesis only about 4 percent of the time, thus achieving a FWER of less than 5 percent.

In column (2), the ten hypotheses arise from examining multiple subgroups rather than multiple outcome variables. Failing to adjust the p-values again results in a high rejection rate of nearly 40 percent. The Bonferroni-Holm, Sidak-Holm, and Westfall-Young adjustment methods, however, all achieve rejection rates of around 5 percent.

A key limitation of the Bonferroni-Holm and Sidak-Holm adjustment methods is their reliance on the assumption of independence among outcomes, which can lead to overly conservative results when outcomes are correlated. This limitation is evident in column (3), which reports rejection rates for a scenario where the ten null hypotheses are all false. Under these conditions, the Bonferroni-Holm and Sidak-Holm methods reject at least one hypothesis in only about 35 percent of simulations. By contrast, the Westfall-Young resampling algorithm, which accounts for correlations among outcomes, achieves a rejection rate in excess of 50 percent, demonstrating its superior performance in this context.

Although traditional adjustment methods such as Bonferroni-Holm and Sidak-Holm are generally thought to be conservative, Westfall and Young (1993) emphasize that these traditional methods can actually over-reject when the data-generating process is nonnormal.

[6] The mean of the standard lognormal distribution is $\sqrt{\exp[1]}$.

[7] The Bonferroni-Holm and Sidak-Holm (step-down) p-values are calculated as follows. Sort the K unadjusted p-values so that $p_1 \leq p_2 \leq ... \leq p_K$. The Bonferroni-Holm adjusted p-values are calculated as $\{p_1 K, \max[p_1, p_2(K-1)], ..., \max[p_{K-1}, p_K]\}$. The Sidak-Holm adjusted p-values are calculated as $\{1-(1-p_1)^K, \max[p_1, 1-(1-p_2)^{(K-1)}], ..., \max[p_{K-1}, p_K]\}$. If the calculation yields a value larger than one, then the adjusted p-value is set equal to one.

Column (4) illustrates this issue: the resampling method of Westfall-Young achieves a FWER below 6 percent, whereas the Bonferroni-Holm and Sidak-Holm methods incorrectly reject at least one null hypothesis in more than 20 percent of simulations, far exceeding the target threshold of 5 percent.

2.1 Clustered standard errors

Westfall and Young (1993) do not discuss methods for conducting multiple inference in regression models where observations are grouped into clusters and model errors exhibit within-cluster correlation. While clustered errors do not violate subset pivotality—a condition automatically satisfied in standard linear regression models—it is important to adapt the resampling procedure in step 2 to account for clustering. Specifically, the rsampling must be performed over entire clusters rather than individual observations. This adjustment can be accomplished by specifying the **cluster()** option in the **wyoung** command.

To illustrate the importance of resampling over clusters, we conducted an additional set of simulations. Let μ be a ten-dimensional zero vector $(0, 0, ..., 0)'$, and let I be a 10×10 identity matrix. The data-generating process for this simulation scenario is:

5. Serially correlated errors (ten outcomes)

   $i = 1...100$ clusters

   $t = 1...10$ time periods

   $\eta_i \sim \mathcal{N}(\mu, I)$

   $e_{it} \sim \mathcal{N}(\mu, I)$

   $Y_{it} = \eta_i + e_{it}$

We again simulated 2,000 datasets. In each simulation, we estimated the following ten regressions:

$$Y_{it} = \alpha + \beta_i D_{it} + \varepsilon_{it}, i = 1...10,$$

where the dummy variable $D_{it} = 1\{t > START_i\}$ and $START_i$ is a Poisson random variable with mean equal to five. These regressions were estimated under two different assumptions about the standard errors (homoskedastic or clustered) and with and without a bootstrap cluster. The results are presented in Table 2.

Comparing column (2) to column (1) in the first row of Table 2, we observe that clustering the standard errors results in a smaller FWER relative to assuming homoskedasticity. However, the rejection rate for the unadjusted value in column (2) still significantly exceeds 5 percent, as it does not account for the number of hypotheses being tested.[8]

The second and third rows of Table 2 show that the Bonferroni-Holm and Sidak-Holm corrections achieve an FWER of less than 5 percent when the standard errors are clustered. This result is expected, as the outcome variables in this simulation are independent.

The fourth row of Table 2 highlights the importance of properly accounting for clustered standard errors when implementing the Westfall-Young correction. Column (2) shows

[8]The unadjusted, Bonferroni-Holm, and Sidak-Holm values do not vary across columns (2) and (3) because these two columns differ only in their bootstrapping methodology, which affects only the Westfall-Young correction.

that (erroneously) employing a simple bootstrap that resamples over individual observations, rather than clusters, causes the Westfall-Young correction to perform worse than the unadjusted specification. However, column (3) shows that employing a cluster bootstrap restores the Westfall-Young correction's ability to control the FWER at 5 percent.

2.2 Multiple regressors

The simulations described above address scenarios involving multiple outcomes or multiple subgroups. Another common context for multiple hypotheses testing arises when there are multiple coefficients of interest within the same model. The Westfall-Young adjustment exhibits strong control of the FWER in this setting as well (Westfall and Young, 1993, p. 134).

To assess the effectiveness of the adjustment in this setting, we conducted additional simulations. Let μ be a ten-dimensional zero vector $(0, 0, ..., 0)'$, and let I be a 10×10 identity matrix. The data-generating process for this simulation scenario is:

6. Normal i.i.d. errors (ten outcomes, two regression coefficients)

   $D_1 \sim 1\{\mathcal{U}(0,1) > 0.5\}$

   $D_2 \sim 1\{\mathcal{U}(0,1) > 0.5\}$

   $e \sim \mathcal{N}(\mu, I)$

   $Y = e$

We simulated 2,000 datasets using this data-generating process. In each simulation, we estimated a series of 10 regressions:

$$Y_i = \alpha + \beta_{i1} D_1 + \beta_{i2} D_2 + \varepsilon_i, i = 1...10.$$

The sample size for each regression was 100. Across the 10 regressions, we tested 20 null hypotheses: $\beta_1 = 0$ and $\beta_2 = 0$. All null hypotheses are true by construction.

Column (1) of Table 3 shows that without any adjustment, the rejection rate exceeds 60 percent. However, applying a multiple-testing adjustment reduces the rejection rate to approximately 4 percent, successfully controlling the FWER below the target threshold of 5 percent.

2.3 Linear and nonlinear combinations

wyoung enables researchers to perform multiple inference when testing hypotheses about any linear or nonlinear combination of coefficients. To evaluate its effectiveness, we conducted simulations testing both linear and nonlinear restrictions involving two regression coefficients. Let μ be a ten-dimensional zero vector $(0, 0, ..., 0)'$. Let I be a 10×10 identity matrix. The data-generating process is:

7. Multiple restrictions (ten outcomes)

   $X_1 \sim \mathcal{N}(\mu, I)$

$X_2 \sim \mathcal{N}(\mu, I)$

$e \sim \mathcal{N}(\mu, I)$

$Y = 2X_1 + 0.5X_2 + e$

We simulated 2,000 datasets using this data-generating process. In each simulation, we estimated a series of 10 regressions:

$$Y_i = \alpha + \beta_{i1}X_{i1} + \beta_{i2}X_{i2} + \varepsilon_i, i = 1...10.$$

The sample size for each regression was 100. We separately tested the following two sets of 10 null hypotheses: (1) the linear restriction $\beta_{i1} - 4\beta_{i2} = 0$; and (2) the nonlinear restriction $\beta_{i1}\beta_{i2} - 1 = 0$. Both these null hypotheses are true by construction.

The results, reported in Columns (2) and (3) of Table 3, show that rejection rates exceed 40 percent when no adjustment is applied. By contrast, the rejection rates for adjusted p-values are approximately 5 percent for the linear restriction and 6 percent for the nonlinear restriction, demonstrating that **wyoung** effectively controls the FWER in both scenarios.

2.4 Permutation

The simulations described above employed the bootstrapping algorithm outlined in Section 1.1. Next, we compare the performance of this algorithm with the permutation algorithm presented in Section 1.2. Let μ denote a ten-dimensional zero vector $(0, 0, ..., 0)'$, and let I denote a 10×10 identity matrix. We begin with the following normal data-generating process:

8. Normal i.i.d. errors (ten outcomes)

   $e \sim \mathcal{N}(\mu, I)$

   $Y = e$

We analyze two distinct treatment assignment processes. The first is simple random assignment at the individual level, where each individual has an equal probability of 0.5 of being assigned to the treatment group. The second is stratified random assignment, in which the population is divided into 10 equally sized strata, and treatment is randomly assigned within each stratum.

Next, we consider clustered random assignment, governed by the following data-generating process:

9. Clustered random assignment (ten outcomes)

   $i = 1...100$ clusters

   $j = 1...10$ units per cluster

   $\eta_i \sim \mathcal{N}(\mu, I)$

   $e_{ij} \sim \mathcal{N}(\mu, I)$

   $Y_{ij} = \eta_i + e_{ij}$

Treatment is assigned at the cluster level using simple random assignment, with a 50 percent probability of being assigned to the treatment group.

Table 4 reports rejection rates for all three scenarios. Without any adjustment, the rejection rates are about 40 percent. Both the Bonferroni-Holm and Sidak-Holm methods effectively control the FWER at approximately 5 percent, as expected in this setting where outcomes are uncorrelated.The Westfall-Young correction achieves rejection rates between 4 and 6 percent when using bootstrapping and slightly tigher control, with rates between 4 and 5 percent, when using permutation. These results suggest that the choice between the two methods does not substantially affect performance in this setting.

References

Jones, D., D. Molitor, and J. Reif (2019). What do workplace wellness programs do? Evidence from the Illinois Workplace Wellness Study. *Quarterly Journal of Economics 134*(4), 1747–1791.

Westfall, P. H. and S. S. Young (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, Volume 279. John Wiley & Sons.

Table 1: Family-wise rejection proportions at $\alpha = 0.05$

	(1)	(2)	(3)	(4)
Adjustment method	Normal errors	Multiple subgroups	Correlated errors	Lognormal errors
Unadjusted	0.398	0.387	0.685	0.577
Bonferroni-Holm	0.040	0.047	0.344	0.234
Sidak-Holm	0.040	0.051	0.347	0.237
Westfall-Young	0.041	0.045	0.513	0.058
Num. observations	100	100	100	100
Num. hypotheses	10	10	10	10
Hypotheses are true	Y	Y	N	Y

Notes: Table reports the proportion of 2,000 simulations where at least one null hypothesis in a family of 10 hypotheses was rejected. In the simulations reported in columns (1), (2), and (4), all hypotheses are true, so lower rejection rates indicate better performance. In contrast, for the simulation reported in column (3), all hypotheses are false, so higher rejection rates indicate better performance. The Westfall-Young adjustment is applied using 1,000 bootstraps.

Table 2: Family-wise rejection proportions at $\alpha = 0.05$, when the data generating process is serially correlated

Adjustment method	(1)	(2)	(3)
Unadjusted	0.652	0.401	0.401
Bonferroni-Holm	0.187	0.049	0.049
Sidak-Holm	0.188	0.049	0.049
Westfall-Young	0.191	0.498	0.046
Num. observations	1,000	1,000	1,000
Num. hypotheses	10	10	10
Model std. errors	Homoskedastic	Clustered	Clustered
Cluster bootstrap	N	N	Y

Notes: Table reports the proportion of 2,000 simulations where at least one null hypothesis in a family of 10 hypotheses was rejected. The difference between columns (1) and (2) is the assumption about the standard errors (homoskedastic or clustered). The difference between columns (2) and (3) is the method of bootstrapping (resampling over individual observations versus clusters), which matters only for the Westfall-Young adjustment. All null hypotheses are true, so lower rejection rates indicate better performance. Each simulation generated 100 panels (clusters) with 10 time periods. The Westfall-Young adjustment is applied using 1,000 bootstraps.

Table 3: Family-wise rejection proportions at $\alpha = 0.05$, when testing hypotheses with multiple regressors or restrictions

	(1)	(2)	(3)
Adjustment method	Multiple regressors	Linear restriction	Nonlinear restriction
Unadjusted	0.634	0.440	0.435
Bonferroni-Holm	0.043	0.052	0.064
Sidak-Holm	0.045	0.052	0.066
Westfall-Young	0.041	0.051	0.062
Num. observations	100	100	100
Num. hypotheses	20	10	10

Notes: Table reports the proportion of 2,000 simulations where at least one null hypothesis in the family was rejected. All null hypotheses are true, so lower rejection rates indicate better performance. Section 2.2 describes the data-generating process used in column (1). Section 2.3 describes the data-generating process used in columns (2) and (3). The Westfall-Young adjustment is applied using 1,000 bootstraps.

Table 4: Family-wise rejection proportions at $\alpha = 0.05$, when treatment is randomized

	(1)	(2)	(3)
	Method of random assignment		
Adjustment method	Individual	Stratified	Clustered
Unadjusted	0.392	0.409	0.391
Bonferroni-Holm	0.051	0.045	0.045
Sidak-Holm	0.054	0.047	0.045
Westfall-Young (bootstrap)	0.053	0.064	0.043
Westfall-Young (permutation)	0.052	0.048	0.043
Num. observations	100	100	1,000
Num. hypotheses	10	10	10

Notes: Table reports the proportion of 2,000 simulations where at least one null hypothesis in the family was rejected. All null hypotheses are true, so lower rejection rates indicate better performance. In column (1), individuals are randomly assigned to treatment with a probability of 0.5. In column (2), assignment is stratified into 10 equally sized strata. In column (3), treatment is assigned at the cluster level, with 100 clusters of 10 observations each. The Westfall-Young adjustments are applied using 1,000 bootstraps/permutations.