Project Log: Multiple Outcomes

Contents

Stream of consciousness 2

Possible resampling algorithms 3

Other online resources 7

⋆ CV: Permutation tests vs. bootstrap tests 7

⋆ CV: Center first or resample first? 7

CV: Cases when naive bootstrap fails 7

CV: Cases when naive bootstrap fails 7

CV: Applying bootstrap to arbitrary smooth function of the data 8

Notes on past literature 9

Davison & Hinkley text 9

Good (2005) book chapter 9

Fox chapter on bootstrapping regressions 9

Hall & Wilson "Two guidelines" 9

⋆ Troendle 2004 "Slow convergence" 10

Westfall & Young textbook 10

Westfall & Troendle (2008) 11

Bretz, Hothorn, & Westfall (2008)'s R paper 11

Bickel 11

Chernick textbook 12

New reference 12

Q & A 13

Open questions 13

Resolved questions 13

Simulation notes 14

Summary of previous two 14
2017-8-1 (generate under alternative) 14
2017-7-25 (generate under null) 14
Summary so far 15
2017-7-24 “Sherlock sim” 15
2017-7-24 “Test smaller sim” 15
2017-7-22 15
2017-7-21 16
Summary so far 16
Summary so far 16

Stream of consciousness

Since we are bootstrapping and then fitting parametric model (OLS), I think we should use Fox’s or Westfall’s residual resampling approach (and indeed, this seemed to work in the last simulation). Seems like the basic approach automatically violates OLS assumptions, meaning our inference will be wrong and the probability of rejecting the test could be non-nominal. Our application of bootstrapping is unusual in that we need to resample in a way that is compatible with parametric assumptions of the test we’re going to apply for each outcome.

Metric #1: 95% CI for number of hits above expected value that we see in observed data (resample under original distribution, using first half of my theory)

Metric #2: 95% CI under the null for number of hits (resample under null, using second half of my theory)

What about other link functions or hypothesis tests not based on regression? How would we do residual resampling? Davison & Hinkley talk about other GLMs. I think we’d end up with either some form of residual or Y itself for each outcome, and then we’d jointly resample these things as planned.

Also, there might be models for which you can’t resample as planned. For example, with multiple regression, we

Possible resampling algorithms

Note that all of these methods assume $Var(Y^*|X^* = 0) = Var(Y^*|X^* = 1)$; in the resampled data, this is the average of the original within-stratum variances. I think this is fine since we are *not* avoiding parametric inference for OLS (unlike in other bootstrapping applications).

Basic "fix X, then resample Y"

1. Fix X and resample only Y: $(Y_i^*, X_i^*) = (\text{sample}(Y_i), X_i)$. So $Y_i^* \amalg X_i^*$.
2. Regress: $Y_i^* = \beta_0 + \beta X_i^* + \epsilon$.
3. Test $H_0 : \beta = 0$.

Has multimodality problem. No guarantee that residuals are normal, so OLS inference will be wrong anytime the original data are generated under H_A.

Figure 1: Original data.

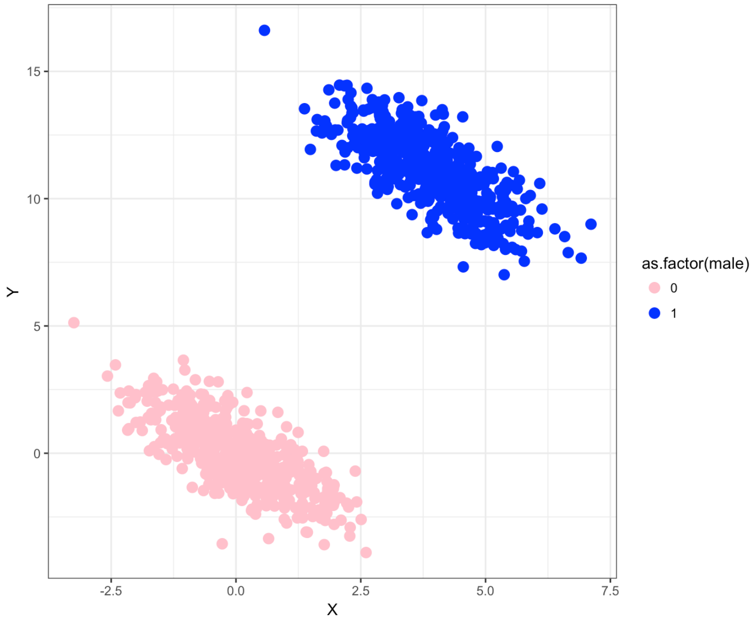
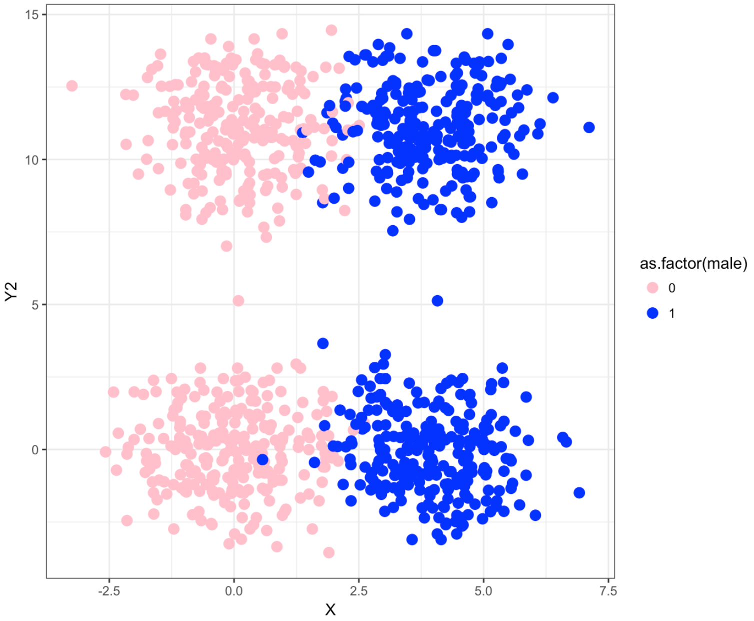


Figure 2: Resampled data under basic algorithm. Loses association between sex and Y, so regression residuals are bimodal.



Westfall's "center, then resample Y"

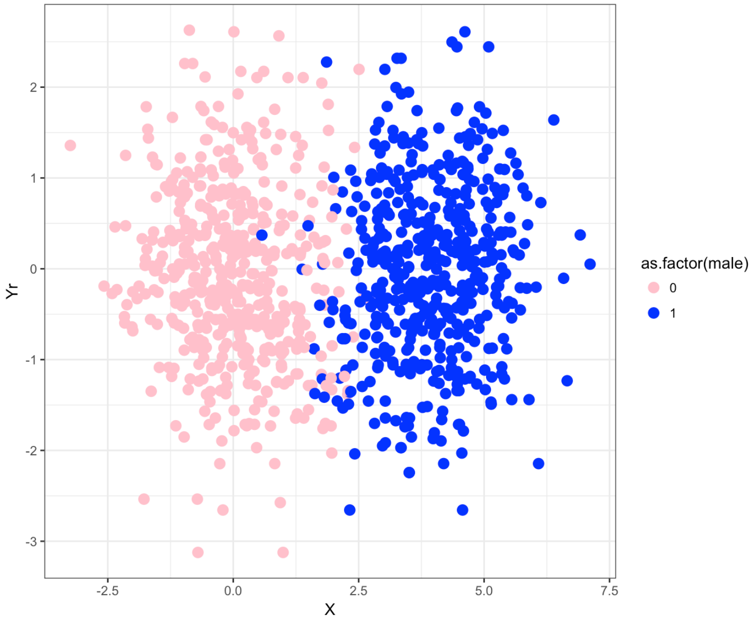
1. Fix X and set $Y_i^* = \text{sample}\left(Y - \widehat{Y}\right)$. So $Y_i^* \perp\!\!\!\perp X_i^*$.

2. Regress: $Y_i^* = \beta_0 + \beta X_i^* + \epsilon$.

3. Test $H_0 : \beta = 0$.

Avoids multimodality problem, but could still have non-normal residual distribution, as in the above. So OLS inference could be wrong, but only if original data already violate homoskedasticity. Works well in my simulations.

Figure 3: Resampled data under Westfall's algorithm. Residuals are normal.



Fox/Davison's residual resampling

See Davison Algorithm 6.1 in "Linear Regression" chapter.

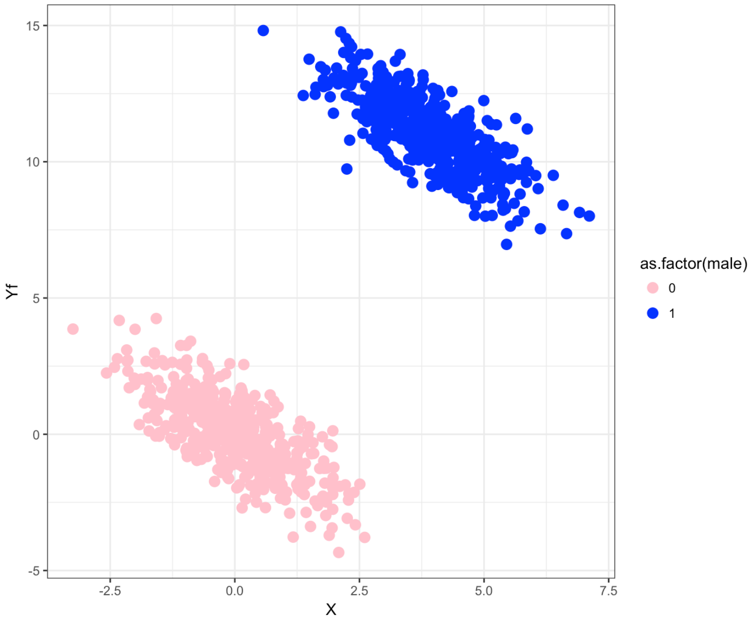
1. Fix X and set $Y_i^* = \widehat{Y}_i + \text{sample}\left(Y - \widehat{Y}\right)$. So $Y_i^* \not\!\perp\!\!\!\perp X_i^*$.

2. Regress: $Y_i^* = \beta_0 + \beta X_i + \epsilon$.

3. Test $H_0 : \beta = \widehat{\beta}$ (i.e., the fitted coefficient from original sample).

Avoids multimodality problem residuals will be normal as long as they are normal within strata of X, so OLS inference should be correct. $Var(Y^*|X^* = 0) = Var(Y^*|X^* = 0)$ and is an average of the original conditional variances.

Figure 4: Resampled data under Fox's algorithm. Residuals are normal.



Other online resources

$\star$ CV: Permutation tests vs. bootstrap tests

What we are doing is similar to a permutation test because we're not centering the observations. In this answer, Snow says "permutation tests test a specific null hypothesis of exchangeability, i.e. that only the random sampling/randomization explains the difference seen". Also, permutation tests are more powerful.

"Resample Y while fixing X" strategy is "similar in spirit" to permutation tests and assumes that the distribution of Y in each group of X is identical. In contrast, "mean-center, then resample within each group of X" does not make this assumption.

https://stats.stackexchange.com/questions/20217/bootstrap-vs-permutation-hypotheis-testing

$\star$ CV: Center first or resample first?

https://stats.stackexchange.com/questions/136661/using-bootstrap-under-h0-to-perform-a-test-for-the-difference-of-two-r
187630

Goes over difference between "resample, then center" (OP's first bootstrap) and "center, then resample" (OP's second bootstrap).

Accepted answer says that first approach is actually testing whether the distribution of x and y are identical.
I DON'T REALLY GET THIS.

CV: Cases when naive bootstrap fails

https://stats.stackexchange.com/questions/9664/what-are-examples-where-a-naive-bootstrap-fails/
9722#9722

CV: Cases when naive bootstrap fails

https://stats.stackexchange.com/questions/11210/assumptions-regarding-bootstrap-estimates-of-uncertainty?
noredirect=1&lq=1

CV: Applying bootstrap to arbitrary smooth function of the data

https://stats.stackexchange.com/questions/246632/smoothness-of-a-statistic-for-bootstrapping

Notes on past literature

Davison & Hinkley text

- Linear regression chapter has great explanation of case resampling (Tyler's approach) vs. residual resampling (Fox, sort of Westfall).
- Pg 12 in "Linear Regression" chapter: Says that with case resampling, "the null hypothesis being tested is stronger than just zero slope"

Good (2005) book chapter

- Clear explanation of difference in assumptions between nonparametric bootstrap (equal parameters) and permutation test (exchangeability).
- I think the "resample, then center" strategy is closer to a permutation test since it forces equal distributions, while the "center, then resample" or "just resample" strategies are more like bootstrapping.

Fox chapter on bootstrapping regressions

- In section on bootstrap hypothesis testing, uses this residual-resampling algorithm:
  1. Fix X and set $Y_i^* = \widehat{Y}_i + \text{sample}\left(Y - \widehat{Y}\right)$. So $Y_i^* \not\perp X_i^*$.
  2. Regress: $Y_i^* = \beta_0 + \beta X_i + \epsilon$.
  3. Test $H_0 : \beta = \widehat{\beta}$ (i.e., the fitted coefficient from original sample).

Hall & Wilson "Two guidelines"

- "Sample, then center" algorithm (here, I am extrapolating what they present to the regression setting):
  1. Resample entire vector $(Y_i^*, X_i^*) = \text{sample}\left((Y_i, X_i)\right)$. So $Y_i^* \not\perp X_i^*$.
  2. Regress: $Y_i^* = \beta_0 + \beta X_i^* + \epsilon$.
  3. Test $H_0 : \beta = \widehat{\beta}$ (i.e., the fitted coefficient from original sample).
- The point of using asymptotically pivotal statistics (centered and scaled) is that then the asymptotic distribution of the statistic does not depend on any unknowns.

- When you can't estimate variance of estimator in order to create pivot, suggests using accelerated bias correction bootstrap.

⋆ Troendle 2004 "Slow convergence"

- Talks about multiple 2-sample t-tests
- Two approaches (see page 3 for very clear overview):
  1. Permute a subject's outcome variables while assigning group labels (with *uncentered* variables)
  2. Bootstrap by resampling a subject's entire row (with *centered* outcome variables)
- "Center, then sample" bootstrapping algorithm (here, I am extrapolating what they present to the regression setting):
  1. Center outcome by setting $\widetilde{Y}_i = Y_i - \widehat{Y}$
  2. Resample entire vector: $(Y_i^*, X_i^*) = \text{sample}\Big((\widetilde{Y}_i, X_i)\Big)$. So $Y_i^* \amalg X_i^*$.
  3. Regress: $Y_i^* = \beta_0 + \beta X_i^* + \epsilon$.
  4. Test $H_0 : \beta = 0$.

Westfall & Young textbook

- I think their algorithm is closest to the one Tyler initially suggested (fix X while resampling Y), except that they first center the outcomes (i.e., resample the residuals).
- "Center, then resample" strategy
  1. Center outcome by setting $\widetilde{Y}_i = Y_i - \widehat{Y}$
  2. Fix X and resample only Y: $(Y_i^*, X_i^*) = \Big(\text{sample}(\widetilde{Y}_i), X_i\Big)$. So $Y_i^* \amalg X_i^*$.
  3. Regress: $Y_i^* = \beta_0 + \beta X_i^* + \epsilon$.
  4. Test $H_0 : \beta = 0$.
- 106-109: More on OLS
- Page 79: "Resampling from the residuals, rather than the actual data values, is important. If the actual data are pooled and resampled, then the underlying residual distribution will be estimated

using a multi-modal empirical distribution function when the means differ. Such resampling violates the first principle of Hall & Wilson (1991): ′...even if the data might be drawn in a way that fails to satisfy H_0, resampling should be done in a way that reflects H_0′." I DON'T GET WHY UNCENTERED APPROACH VIOLATES H0!

- Page 39-41: Analyzes $P(P^* < \alpha)$ in bootstrap iterates
- Page 92-93: Re-randomization
- Bootstraps of pivotal statistics converge faster (Section 2.2.2)

Westfall & Troendle (2008)

- Setting: You have multiple outcomes measured on multiple categorical groups and are interested in full exchangeability on the outcome across the groups (not just different means, for example).
- Shows that "permutation methods are distribution-free under an appropriate exchangeability assumption and FWER follows mathematically, regardless of sample size"
- Implemented in SAS' PROC MULTTEST
- According to SAS' documentation, this(?) paper shows that: "when subset pivotality holds, the joint distribution of p-values under the subset is identical to that under the complete null"

Bretz, Hothorn, & Westfall (2008)'s R paper

- They are fixing the Y and permuting the Xs (in their application, these are just the group labels), so are keeping the correlation structure of the Ys as we are (pg. 130)
- Bootstrapping vs. permutation pros and cons (pg. 140)"

Bickel

- Classic paper that goes over basic asymptotics of bootstrap
- Uses SLLN

Chernick textbook

- Great table comparing different CI methods with their assumptions (Ch 3, page 8)

New reference

- asdf

Q & A

Open questions

Resolved questions

Q: Why does Blakesley/Westfall's minP bootstrapping work since it is an extreme order statistic? A: Probably because minP is not an order statistic of the data themselves.

Q: If we go the confidence interval route, why not just treat the p-values as our data and resample directly from them to produce CI? A: Because there are only a fixed number of p-values (k), and this is not necessarily asymptotically large.

Q: Why does bootstrapping work with small samples since we need the ECDF in sample to go to the true CDF? Or does it not work in small samples? A: Correct. It does not necessarily work in small samples.

Simulation notes

Summary of previous two

- Plot1 (CIs)
  - Shows that the residual resampling is working correctly (since results are same regardless of whether we generate under H0 or under HA)
  - As expected, more variable when Y's are more correlated

2017-8-1 (generate under alternative)

- Same as 2017-7-25, but now generating under weak alternative (rho.XY = 0.03)
- Expect same exact results for CIs (since they only use results from resampling under H0), but higher rejection rates

2017-7-25 (generate under null)

- CIs are 95% two-sided ones for both values of alpha
- Hypothesis tests are one-sided and have sample alpha as the individual tests
- N=1,000 and B=2,000 (j=10 per simulation)
- Figure out why file 125 has extra rows (try running it again)
- **The apparently too-conservative behavior for the uncorrelated case is a benign artifact** of $\widehat{\theta}$'s discreteness.
  - It occurs because when the Y's are uncorrelated, the distribution of the number of rejections is less right-skewed, so it drops off more quickly in the tails. Since $\widehat{\theta}$ is discrete, we are forced to pick a quantile that doesn't have exactly 5% of the mass above it. By default, R inverts the ECDF, so uses the quantile with LESS than 5% above it. Because the distribution of $\widehat{\theta}$ drops off quickly around the chosen quantile, we end up with conservative performance (see artifact_plot1 and artifact_plot2, where I shade in red the proportion that are above the variable's own critical value).
  - Artifact might improve a little with more simulation reps, but only to a point.

  - For this case, we could benchmark against the truth since $\widehat{\theta}$ is binomial. Indeed, the empirical quantiles of both `n.rej` and the bootstrap estimates are exactly correct. :)

Summary so far

- Seems like N=1,000 and B=2,000 is enough for good asymptotics
- When data are generated under null, resampling the Y's or the residuals seems to works. But former theoretically should work only in more limited cases.

2017-7-24 "Sherlock sim"

- Huge: N=10,000, B=10,000
- Resampling under joint null (Y's rather than residuals)
- This seemed to work based on interim results (though I had trouble with existing stitching script, so stitched results in overall_stitched folder might be wrong)

2017-7-24 "Test smaller sim"

- N=1,000, B=2,000
- Switched to resampling residuals instead of Ys themselves for the first time
- Seems to work! Going to try on cluster to have larger simulation

2017-7-22

- N=100, B=1,000: seems too small based on the below
- Looked good at alpha=0.05 (4.4% rejections) but maybe too conservative for alpha=0.01? (3.6% rejections)
- Took about 1.5 hours locally

2017-7-21

- With N=5,000, B=2,000, n.sims = 250, looks good
- 5.5% rejection rate on joint null for alpha = 0.05
- 6% rejection rate for alpha = 0.01
- Took about 30 hours locally
- Seems good

Summary so far

- Resample under H0 by drawing Y's separately; reject by inverted CI using its percentiles ⇒ **WORKS** (see Sherlock sim)
- Resample from original with single Y1; look at variance of mean ⇒ **WORKS**
- Bootstrap from original sample with 100 independent std. normal; reject using z-test that's only a function of mean ⇒ **DOESN'T WORK**
  - The bootstrapped distribution of the absolute differences is too variable compared to true distribution
  - Hence rejects only 1.8% of time
- Conjecture: Hall & Wilson approach seems to rely on the idea that the standardized estimator is pivotal, i.e., that it comes from a location-scale family. This is why we can get away with only sampling from the original distribution. Since our estimator definitely isn't from a location-scale distribution, maybe that's why it does not work. Also, based on the different results between currently running one and very first one, I think B=500 is not enough, and B = 10,000 is enough. Not sure about intermediate ones.

Summary so far

- Seems like N=1,000 and B=2,000 is enough for good asymptotics
- When data are generated under null, resampling the Y's or the residuals seems to works. But former theoretically should work only in more limited cases.

For earlier simulation results, see old project log in Word.

References