Performance of Big Match for Varying Sample Sizes

Rachael Caelie (Rocky) Aikens

1/28/2019

This rmarkdown is meant for testing the performance of the current implementation of big_match on sample data.

Import Data

This sample cohort data was provided by Justin Lee at the Quantitative Sciences Unit. It contains ~900,000 observations of 112 variables.

```
dat <- read_sas("../sample_data/justincohort_june2017.sas7bdat")

# dimensions: ~900,000 x 112
dim(dat)
```

```
## [1] 893498    112
```

We include only hospitalizations with `totalct` > 1 and `arteryCt < 3`. A patient is considered to have recieved treatment if `arteryCt` is greater than 1. The outcome of this analysis is mortality.

```
# filter and add treatment column
dat <- filter(dat, totalct > 1 & arteryCt < 3) %>%
  mutate(treat = ifelse(arteryCt > 1, 1, 0))

# dimensions: ~900,000 x 112
dim(dat)
```

```
## [1] 833657    113
```

```
dat <- filter(dat, hosp_state != "Virgin Islands")
```

User Time

```
run_big_match <- function(n_samples){
  n_dat <- sample_n(dat, n_samples, replace = FALSE)

  # stratify
  t1 <- proc.time()
  a.strat <- auto_stratify(data = n_dat, treat = "treat",
                           outcome = "dead",
                           prog_formula = dead ~ totalct + AMI_7 + COPD_7 +
                             ISCHEMICHEART_7 + STROKE_TIA_7 + ATRIAL_FIB_7 + CHRONICKIDNEY_7 + DIABETES_7
                             ALZH_DEMEN_7 + Male,
                           size = 2500)
  print("Stratification complete.")
  strat_time <- proc.time() - t1

  # big match by strata in series
  t2 <- proc.time()
  big_match(a.strat, propensity_formula = treat ~ totalct + hosp_state +
            AMI_7 + COPD_7 + ISCHEMICHEART_7 + STROKE_TIA_7 + ATRIAL_FIB_7 +
            CHRONICKIDNEY_7 + DIABETES_7 + ALZH_DEMEN_7 + Male + race)
  strata_match_time <- proc.time() - t2

  # big match without strata
  t3 <- proc.time()
  big_match_nstrat(a.strat, propensity_formula = treat ~ totalct + hosp_state +
                   AMI_7 + COPD_7 + ISCHEMICHEART_7 + STROKE_TIA_7 + ATRIAL_FIB_7 +
                   CHRONICKIDNEY_7 + DIABETES_7 + ALZH_DEMEN_7 + Male + race)
  full_match_time = proc.time() - t3

  # big match with multidplyr
  t4 <- proc.time()
  big_match_multidplyr(a.strat, propensity_formula = treat ~ totalct + hosp_state +
                   AMI_7 + COPD_7 + ISCHEMICHEART_7 + STROKE_TIA_7 + ATRIAL_FIB_7 +
                   CHRONICKIDNEY_7 + DIABETES_7 + ALZH_DEMEN_7 + Male + race)
  multi_time <- proc.time() - t4
  return(data.frame(rbind(strat_time, strata_match_time, full_match_time, multi_time)))
}
```

```
options("optmatch_max_problem_size" = Inf)
n_samples <- c(5000, 10000, 15000, 20000, 25000, 30000, 35000)
time_list <- lapply(n_samples, run_big_match)
```

```
## [1] "Constructing a model set via subsampling."
## [1] "Fitting prognostic model: dead ~ totalct + AMI_7 + COPD_7 + ISCHEMICHEART_7 + STROKE_TIA_7 +
## [1] "Generating strata assignments based on prognostic score."
## [1] "Completing strata diagnostics."
## [1] "Stratification complete."
## treat ~ totalct + hosp_state + AMI_7 + COPD_7 + ISCHEMICHEART_7 +
##     STROKE_TIA_7 + ATRIAL_FIB_7 + CHRONICKIDNEY_7 + DIABETES_7 +
##     ALZH_DEMEN_7 + Male + race + strata(stratum)
## <environment: 0x7fcac1af5c78>
## treat ~ totalct + hosp_state + AMI_7 + COPD_7 + ISCHEMICHEART_7 +
##     STROKE_TIA_7 + ATRIAL_FIB_7 + CHRONICKIDNEY_7 + DIABETES_7 +
##     ALZH_DEMEN_7 + Male + race
## <environment: 0x7fcac74914d0>

## Warning in value[[3L]](cond): Error gathering complete data. If the data
## has missing cases, imputation will not be performed. (Sometimes this can
## be fixed by supplying a `data` argument when fitting the model that's to
## be passed to `scores`. Alternatively, just take care of (impute) NAs before
## you fit that model.)

## Initialising 12 core cluster.

## Warning: group_indices_.grouped_df ignores extra arguments

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## [1] "Constructing a model set via subsampling."
## [1] "Fitting prognostic model: dead ~ totalct + AMI_7 + COPD_7 + ISCHEMICHEART_7 + STROKE_TIA_7 +
## [1] "Generating strata assignments based on prognostic score."
## [1] "Completing strata diagnostics."
## [1] "Stratification complete."
## treat ~ totalct + hosp_state + AMI_7 + COPD_7 + ISCHEMICHEART_7 +
##     STROKE_TIA_7 + ATRIAL_FIB_7 + CHRONICKIDNEY_7 + DIABETES_7 +
##     ALZH_DEMEN_7 + Male + race + strata(stratum)
## <environment: 0x7fcac4bca620>
## treat ~ totalct + hosp_state + AMI_7 + COPD_7 + ISCHEMICHEART_7 +
##     STROKE_TIA_7 + ATRIAL_FIB_7 + CHRONICKIDNEY_7 + DIABETES_7 +
##     ALZH_DEMEN_7 + Male + race
## <environment: 0x7fcac816e900>

## Warning in value[[3L]](cond): Error gathering complete data. If the data
## has missing cases, imputation will not be performed. (Sometimes this can
## be fixed by supplying a `data` argument when fitting the model that's to
## be passed to `scores`. Alternatively, just take care of (impute) NAs before
## you fit that model.)

## Initialising 12 core cluster.

## Warning: group_indices_.grouped_df ignores extra arguments

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## [1] "Constructing a model set via subsampling."
## [1] "Fitting prognostic model: dead ~ totalct + AMI_7 + COPD_7 + ISCHEMICHEART_7 + STROKE_TIA_7 +
## [1] "Generating strata assignments based on prognostic score."
## [1] "Completing strata diagnostics."
## [1] "Stratification complete."
## treat ~ totalct + hosp_state + AMI_7 + COPD_7 + ISCHEMICHEART_7 +
##     STROKE_TIA_7 + ATRIAL_FIB_7 + CHRONICKIDNEY_7 + DIABETES_7 +
##     ALZH_DEMEN_7 + Male + race + strata(stratum)
## <environment: 0x7fcac7d0ef58>
## treat ~ totalct + hosp_state + AMI_7 + COPD_7 + ISCHEMICHEART_7 +
##     STROKE_TIA_7 + ATRIAL_FIB_7 + CHRONICKIDNEY_7 + DIABETES_7 +
##     ALZH_DEMEN_7 + Male + race
## <environment: 0x7fcac3c610d0>
```

```
## Warning in value[[3L]](cond): Error gathering complete data. If the data
## has missing cases, imputation will not be performed. (Sometimes this can
## be fixed by supplying a `data` argument when fitting the model that's to
## be passed to `scores`. Alternatively, just take care of (impute) NAs before
## you fit that model.)
```

```
## Initialising 12 core cluster.
```

```
## Warning: group_indices_.grouped_df ignores extra arguments
```

```
## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character
```

```
## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector
```

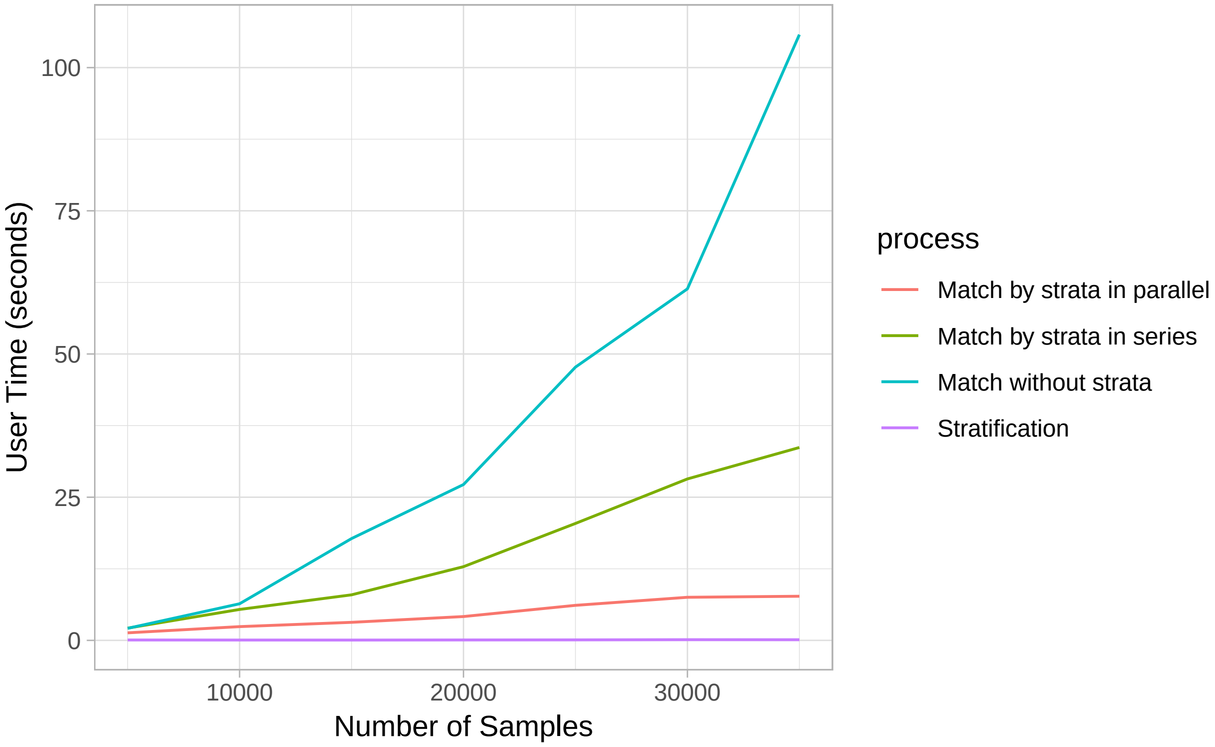
```
## [1] "Constructing a model set via subsampling."
## [1] "Fitting prognostic model: dead ~ totalct + AMI_7 + COPD_7 + ISCHEMICHEART_7 + STROKE_TIA_7 +
## [1] "Generating strata assignments based on prognostic score."
## [1] "Completing strata diagnostics."
## [1] "Stratification complete."
## treat ~ totalct + hosp_state + AMI_7 + COPD_7 + ISCHEMICHEART_7 +
##     STROKE_TIA_7 + ATRIAL_FIB_7 + CHRONICKIDNEY_7 + DIABETES_7 +
##     ALZH_DEMEN_7 + Male + race + strata(stratum)
## <environment: 0x7fcac7c351f8>
## treat ~ totalct + hosp_state + AMI_7 + COPD_7 + ISCHEMICHEART_7 +
##     STROKE_TIA_7 + ATRIAL_FIB_7 + CHRONICKIDNEY_7 + DIABETES_7 +
##     ALZH_DEMEN_7 + Male + race
## <environment: 0x7fcac3bc2e68>
## Warning in value[[3L]](cond): Error gathering complete data. If the data
## has missing cases, imputation will not be performed. (Sometimes this can
## be fixed by supplying a `data` argument when fitting the model that's to
## be passed to `scores`. Alternatively, just take care of (impute) NAs before
## you fit that model.)

## Initialising 12 core cluster.

## Warning: group_indices_.grouped_df ignores extra arguments

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## [1] "Constructing a model set via subsampling."
## [1] "Fitting prognostic model: dead ~ totalct + AMI_7 + COPD_7 + ISCHEMICHEART_7 + STROKE_TIA_7 +
## [1] "Generating strata assignments based on prognostic score."
## [1] "Completing strata diagnostics."
## [1] "Stratification complete."
## treat ~ totalct + hosp_state + AMI_7 + COPD_7 + ISCHEMICHEART_7 +
##     STROKE_TIA_7 + ATRIAL_FIB_7 + CHRONICKIDNEY_7 + DIABETES_7 +
##     ALZH_DEMEN_7 + Male + race + strata(stratum)
## <environment: 0x7fcac78ffa10>
## treat ~ totalct + hosp_state + AMI_7 + COPD_7 + ISCHEMICHEART_7 +
##     STROKE_TIA_7 + ATRIAL_FIB_7 + CHRONICKIDNEY_7 + DIABETES_7 +
##     ALZH_DEMEN_7 + Male + race
## <environment: 0x7fcac85f7af0>

## Warning in value[[3L]](cond): Error gathering complete data. If the data
## has missing cases, imputation will not be performed. (Sometimes this can
## be fixed by supplying a `data` argument when fitting the model that's to
## be passed to `scores`. Alternatively, just take care of (impute) NAs before
## you fit that model.)

## Initialising 12 core cluster.
## Warning: group_indices_.grouped_df ignores extra arguments

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## [1] "Constructing a model set via subsampling."
## [1] "Fitting prognostic model: dead ~ totalct + AMI_7 + COPD_7 + ISCHEMICHEART_7 + STROKE_TIA_7 +
## [1] "Generating strata assignments based on prognostic score."
## [1] "Completing strata diagnostics."
## [1] "Stratification complete."
## treat ~ totalct + hosp_state + AMI_7 + COPD_7 + ISCHEMICHEART_7 +
##     STROKE_TIA_7 + ATRIAL_FIB_7 + CHRONICKIDNEY_7 + DIABETES_7 +
##     ALZH_DEMEN_7 + Male + race + strata(stratum)
## <environment: 0x7fcac73fe548>
## treat ~ totalct + hosp_state + AMI_7 + COPD_7 + ISCHEMICHEART_7 +
##     STROKE_TIA_7 + ATRIAL_FIB_7 + CHRONICKIDNEY_7 + DIABETES_7 +
##     ALZH_DEMEN_7 + Male + race
## <environment: 0x7fcac8328628>

## Warning in value[[3L]](cond): Error gathering complete data. If the data
## has missing cases, imputation will not be performed. (Sometimes this can
## be fixed by supplying a `data` argument when fitting the model that's to
## be passed to `scores`. Alternatively, just take care of (impute) NAs before
## you fit that model.)

## Initialising 12 core cluster.

## Warning: group_indices_.grouped_df ignores extra arguments
## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## [1] "Constructing a model set via subsampling."
## [1] "Fitting prognostic model: dead ~ totalct + AMI_7 + COPD_7 + ISCHEMICHEART_7 + STROKE_TIA_7 +
## [1] "Generating strata assignments based on prognostic score."
## [1] "Completing strata diagnostics."
## [1] "Stratification complete."
## treat ~ totalct + hosp_state + AMI_7 + COPD_7 + ISCHEMICHEART_7 +
##     STROKE_TIA_7 + ATRIAL_FIB_7 + CHRONICKIDNEY_7 + DIABETES_7 +
##     ALZH_DEMEN_7 + Male + race + strata(stratum)
## <environment: 0x7fcac7c72890>
## treat ~ totalct + hosp_state + AMI_7 + COPD_7 + ISCHEMICHEART_7 +
##     STROKE_TIA_7 + ATRIAL_FIB_7 + CHRONICKIDNEY_7 + DIABETES_7 +
##     ALZH_DEMEN_7 + Male + race
## <environment: 0x7fcac3c48790>

## Warning in value[[3L]](cond): Error gathering complete data. If the data
## has missing cases, imputation will not be performed. (Sometimes this can
## be fixed by supplying a `data` argument when fitting the model that's to
## be passed to `scores`. Alternatively, just take care of (impute) NAs before
## you fit that model.)

## Initialising 12 core cluster.
## Warning: group_indices_.grouped_df ignores extra arguments

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector
```

```
bind_rows(time_list, .id = "column_label")
```

```
##    column_label user.self sys.self elapsed user.child sys.child
## 1             1     0.071    0.004   0.075      0.000     0.000
## 2             1     2.139    0.486   2.627      0.000     0.000
## 3             1     2.104    0.261   2.366      0.000     0.000
## 4             1     1.320    0.236   8.160      0.013     0.016
## 5             2     0.062    0.001   0.063      0.000     0.000
## 6             2     5.405    1.100   6.521      0.000     0.000
## 7             2     6.393    1.187   7.584      0.000     0.000
## 8             2     2.403    0.356   9.857      0.012     0.014
## 9             3     0.059    0.001   0.060      0.000     0.000
## 10            3     7.953    1.918   9.935      0.000     0.000
## 11            3    17.779    2.856  20.645      0.000     0.000
## 12            3     3.152    0.628  12.315      0.012     0.015
## 13            4     0.078    0.013   0.091      0.000     0.000
## 14            4    12.873    3.072  16.066      0.000     0.000
## 15            4    27.209    4.968  32.196      0.000     0.000
## 16            4     4.163    0.673  15.314      0.012     0.015
## 17            5     0.092    0.027   0.129      0.000     0.000
## 18            5    20.416    4.437  25.047      0.000     0.000
## 19            5    47.712   10.002  58.185      0.000     0.000
## 20            5     6.119    1.058  20.228      0.013     0.016
## 21            6     0.121    0.056   0.184      0.000     0.000
## 22            6    28.191    6.515  34.815      0.000     0.000
## 23            6    61.369   12.626  74.028      0.000     0.000
## 24            6     7.525    1.761  21.713      0.012     0.015
## 25            7     0.115    0.023   0.139      0.000     0.000
## 26            7    33.677    8.645  42.948      0.000     0.000
## 27            7   105.760   23.688 132.200      0.000     0.000
## 28            7     7.709    1.547  23.983      0.013     0.015
```

```
trials <- length(n_samples)
time_df <- bind_rows(time_list, .id = "column_label") %>%
  mutate(n_samples = rep(n_samples, each = 4)) %>%
  mutate(process = rep(c("Stratification", "Match by strata in series", "Match without strata", "Match
  select(c(process, n_samples, user.self, sys.self, elapsed))
```

```
a <- ggplot(time_df, aes(x = n_samples, y = user.self, group = process, color = process)) +
  geom_line() +
  labs(x = "Number of Samples", y = "User Time (seconds)")
a
```



```
n_dat <- sample_n(dat, 30000, replace = FALSE)
a.strat <- auto_stratify(data = n_dat, treat = "treat",
                       outcome = "dead",
                       prog_formula = dead ~ totalct  + AMI_7 + COPD_7 +
                         ISCHEMICHEART_7 + STROKE_TIA_7 + ATRIAL_FIB_7 + CHRONICKIDNEY_7 + DIABETES_7 +
                         ALZH_DEMEN_7 + Male,
                       size = 2500)
```

```
## [1] "Constructing a model set via subsampling."
## [1] "Fitting prognostic model: dead ~ totalct + AMI_7 + COPD_7 + ISCHEMICHEART_7 + STROKE_TIA_7 +
## [1] "Generating strata assignments based on prognostic score."
## [1] "Completing strata diagnostics."
```

```
print("Stratification complete.")
```

```
## [1] "Stratification complete."
```

```
t1 <- proc.time()
mymatch <- big_match_multidplyr(a.strat, propensity_formula = treat ~ totalct + hosp_state +
                   AMI_7 + COPD_7 + ISCHEMICHEART_7 + STROKE_TIA_7 + ATRIAL_FIB_7 +
                   CHRONICKIDNEY_7 + DIABETES_7 + ALZH_DEMEN_7 + Male + race)
```

```
## Initialising 12 core cluster.
```

```
## Warning: group_indices_.grouped_df ignores extra arguments
```

```
## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character
```

```
## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector
```

```
multi_time <- proc.time() - t1
multi_time
```

```
##    user  system elapsed
##   3.654   0.513  13.288
```