Resampling Methods

AU STAT-427/627

Emil Hvitfeldt

2021-10-04

Motivation

We are already familiar with train-test splits

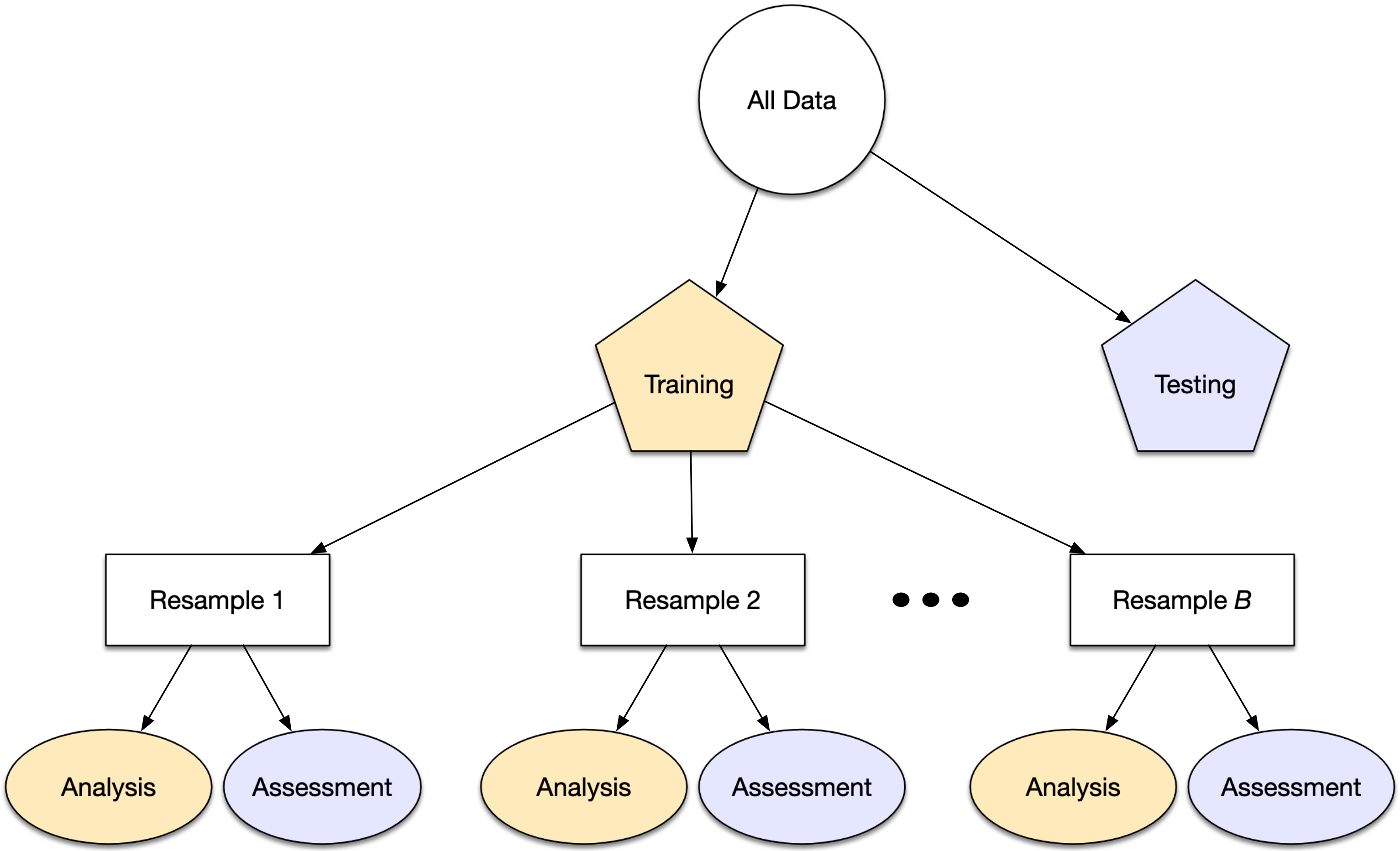
The main downside to train-test splits so far is that we can only use them once

This means we effectively can't make any decisions about the models we are using

Resampling

Resampling estimates of performance can generalize to new data

Resampling Workflow



Resampling Workflow

The resampling is only conducted on the training set

We are still keeping the test set. The test set is not involved.

For each iteration of resampling, the data are partitioned into two subsamples:

- The model is fitted with the analysis set
- The model is evaluated with the assessment set

Resampling Workflow

We have effectively created many train-test split out of our training data set.

The challange here now becomes how we are creating these resample sets

Resampling Workflow

Suppose we generate 10 different resamples

This means that we will be:

- Fitting 10 different models
- Perform predictions 10 times
- Produce 10 sets of performance statistics

The final estimate of the performance of the model will be the average of these 10 models

Resampling Workflow

If the resampling is done in an appropriate way then this average has very good generalization properties

Leave-One-Out Cross-Validation

- 1 observation is used as the assessment set
- The remaining observations make up the analysis set

Notes:

We are fitting the model on $n - 1$ observations and a prediction $\hat{y}_1$ is made on the assessment set using the value x_1

Leave-One-Out Cross-Validation

Since (x_1, y_1) is not used in the fitting process, then $MSE_1 = (y_1 - \hat{y}_1)^2$ provides an approximately unbiased estimate for the test error.

While this estimate is approximately unbiased, it is quite poor since it is highly variable

Leave-One-Out Cross-Validation

We can repeat this for

- $MSE_2 = (y_2 - \hat{y}_2)^2$
- $MSE_3 = (y_3 - \hat{y}_3)^2$
- ...
-

$MSE_n = (y_n - \hat{y}_n)^2$

to get n estimates of the test error

Leave-One-Out Cross-Validation

The LOOCV estimate of the test MSE is

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

Leave-One-Out Cross-Validation

Pros

The LOOCV estimate of the test MSE is going to have a low bias

There is no randomness in the LOOCV estimate

Cons

You need a lot of computational power even for modest data sets

(Some models don't need to be repeatedly refit)

K-Fold Cross-Validation

Could we think of a compromise between fitting 1 model and n models?

K-Fold Cross Validation has an answer:

Randomly divide the observations into k groups (or folds) or approximately equal size

K-Fold Cross-Validation

Randomly divide the observations into k groups (or folds) or approximately equal size

- 1 fold is used as the assessment set
- The remaining folds make up the analysis set

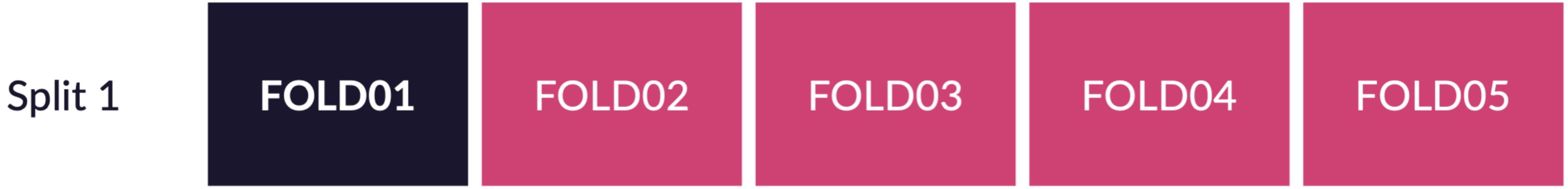
Everything else happens as before.

We now get fewer performance metrics, BUT they are each less variable

Art by Alison Hill

Art by Alison Hill

Art by Alison Hill



Art by Alison Hill

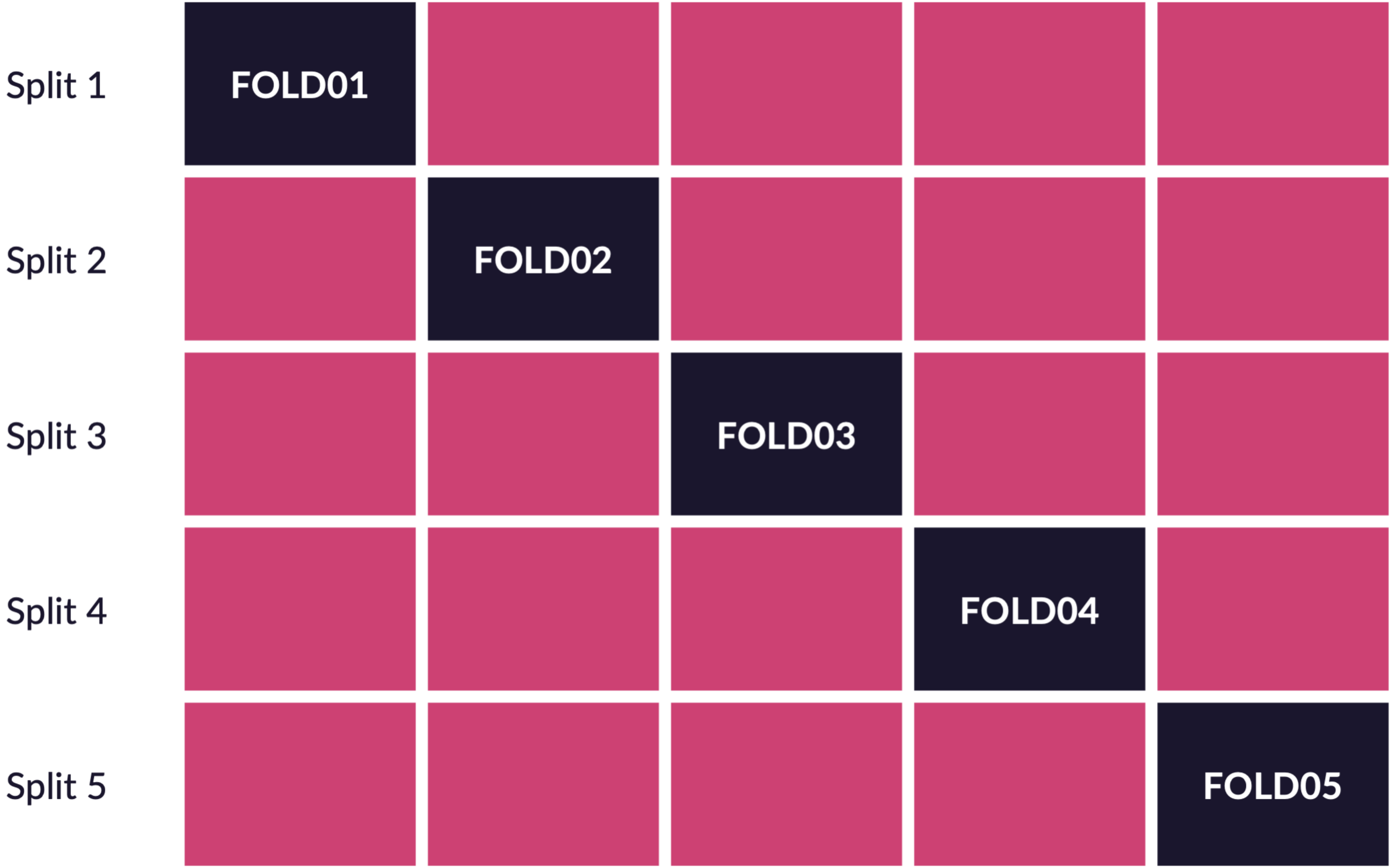
Art by Alison Hill

Art by Alison Hill

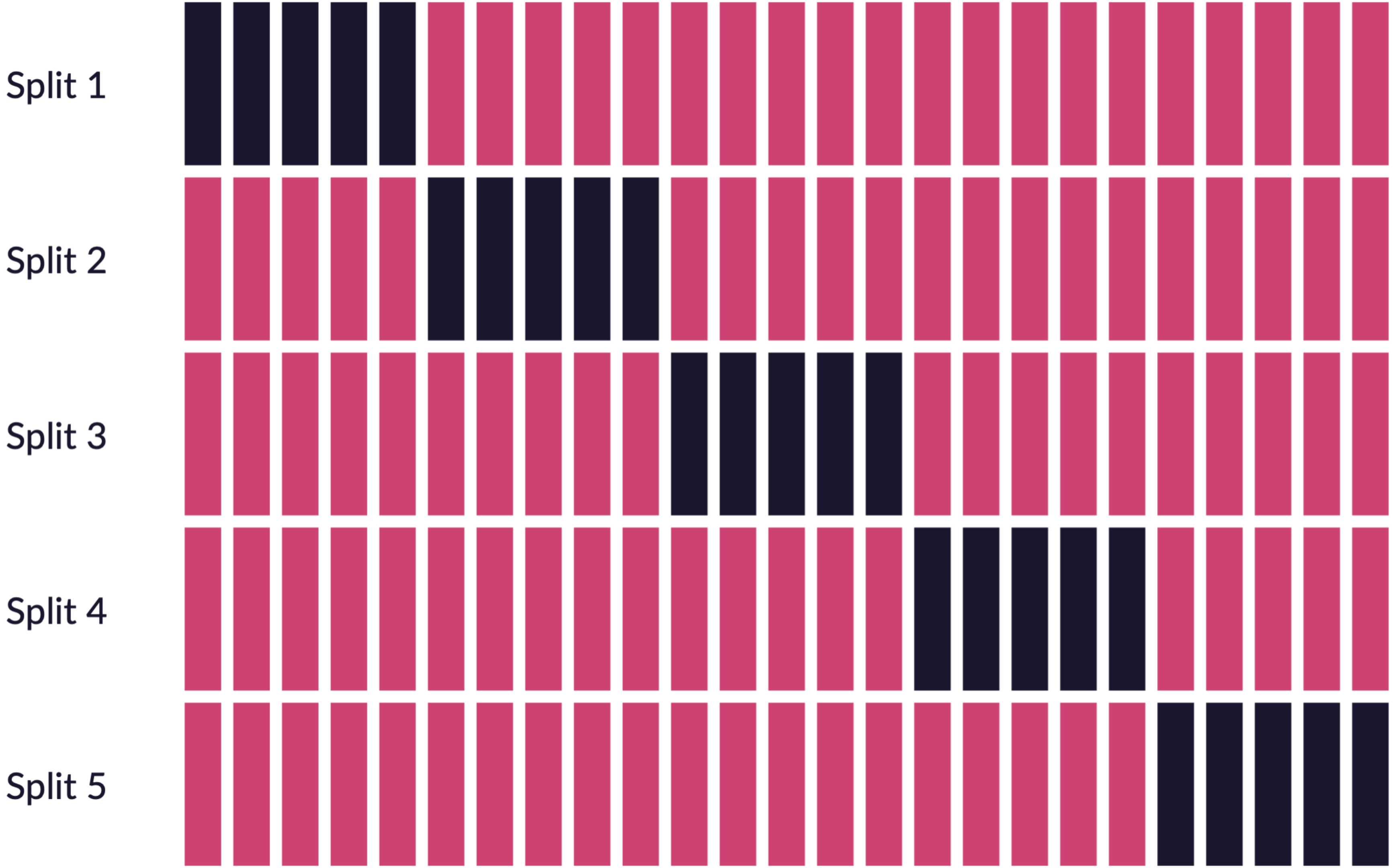
Split 1	**FOLD01**	FOLD02	FOLD03	FOLD04	FOLD05
Split 2	FOLD01	**FOLD02**	FOLD03	FOLD04	FOLD05
Split 3	FOLD01	FOLD02	**FOLD03**	FOLD04	FOLD05
Split 4	FOLD01	FOLD02	FOLD03	**FOLD04**	FOLD05

Art by Alison Hill

Art by Alison Hill



Art by Alison Hill



Art by Alison Hill

Cross validation

When we perform cross-validation our goal might be to determine how well a given model is expected to perform on new data

Other times we are using cross-validation to find the best model/hyperparameters

Bias-Variance tradeoff of LOOCV and k-fold Cross-Validation

LOOCV has a lower bias than k-fold CV

However, since the mean of many highly correlated quantities has higher variance than the mean of many not correlated quantities, we have that LOOCV has a higher variance than k-fold CV

Rsample

We are back with `rsample`

```r
library(rsample)

mtcars
```

```
##                      mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4           21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710          22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout   18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant             18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360          14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D           24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230            22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280            19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C           17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
```

Rsample

We can use the `vfold_cv()` function on a data.frame to create a `vfold_cv` object

```
mtcars_folds <- vfold_cv(mtcars, v = 4)
mtcars_folds
```

```
## #  4-fold cross-validation
## # A tibble: 4 × 2
##   splits         id
##   <list>         <chr>
## 1 <split [24/8]> Fold1
## 2 <split [24/8]> Fold2
## 3 <split [24/8]> Fold3
## 4 <split [24/8]> Fold4
```

Rsample

An under the hood, we have 4 analysis/assessment splits similar to `initial_split()`

```
mtcars_folds <- vfold_cv(mtcars, v = 4)
mtcars_folds$splits
```

```
## [[1]]
## <Analysis/Assess/Total>
## <24/8/32>
##
## [[2]]
## <Analysis/Assess/Total>
## <24/8/32>
##
## [[3]]
## <Analysis/Assess/Total>
## <24/8/32>
##
## [[4]]
## <Analysis/Assess/Total>
## <24/8/32>
```

Using resamples in action

We start by creating a linear regression specification

```
library(parsnip)
linear_spec <- linear_reg() %>%
  set_mode("regression") %>%
  set_engine("lm")
```

Workflows

A simple package that helps us formulate more about what happens to our model.

Main functions are `workflow()`, `add_model()`, `add_formula()` or `add_variables()` (we will see `add_recipe()` later in the course)

```
library(workflows)

linear_wf <- workflow() %>%
  add_model(linear_spec) %>%
  add_formula(mpg ~ disp + hp + wt)
```

Workflows

This allows up to combine the model with what variables it should expect

```r
library(workflows)

linear_wf <- workflow() %>%
  add_model(linear_spec) %>%
  add_formula(mpg ~ disp + hp + wt)
linear_wf
```

```
## ══ Workflow ════════════════════════════════════════
## Preprocessor: Formula
## Model: linear_reg()
##
## ── Preprocessor ────────────────────────────────────
## mpg ~ disp + hp + wt
##
## ── Model ───────────────────────────────────────────
## Linear Regression Model Specification (regressio
##
## Computational engine: lm
```

`add_variables()` allows for a different way of specifying the response and predictors in our model

Workflows

```r
library(workflows)

linear_wf <- workflow() %>%
  add_model(linear_spec) %>%
  add_variables(outcomes = mpg,
                predictors = c(disp, hp, wt))
linear_wf
```

```
## ══ Workflow ════════════════════════════════════════
## Preprocessor: Variables
## Model: linear_reg()
##
## ── Preprocessor ────────────────────────────────────
## Outcomes: mpg
## Predictors: c(disp, hp, wt)
##
## ── Model ───────────────────────────────────────────
## Linear Regression Model Specification (regressio
##
## Computational engine: lm
```

Workflows

You can use a `workflow` just like a parsnip object and fit it directly

```
fit(linear_wf, data = mtcars)
```

```
## ══ Workflow [trained] ══════════════════════════════════════════════════════════
## Preprocessor: Variables
## Model: linear_reg()
##
## ── Preprocessor ──────────────────────────────────────────────────────────────
## Outcomes: mpg
## Predictors: c(disp, hp, wt)
##
## ── Model ─────────────────────────────────────────────────────────────────────
##
## Call:
## stats::lm(formula = ..y ~ ., data = data)
##
## Coefficients:
## (Intercept)         disp           hp           wt
##   37.105505    -0.000937    -0.031157    -3.800891
```

Tune

We introduce the **tune** package. This package helps us fit many models in a controlled manner in the tidymodels framework. It relies heavily on parsnip and rsample

Tune

We can use `fit_resamples()` to fit the workflow we created within each resample

```
library(tune)

linear_fold_fits <- fit_resamples(
    linear_wf,
    resamples = mtcars_folds
)
```

Tune

The results of this resampling come as a data.frame

```
linear_fold_fits
```

```
## # Resampling results
## # 4-fold cross-validation
## # A tibble: 4 × 4
##   splits         id    .metrics         .notes
##   <list>         <chr> <list>           <list>
## 1 <split [24/8]> Fold1 <tibble [2 × 4]> <tibble [0 × 1]>
## 2 <split [24/8]> Fold2 <tibble [2 × 4]> <tibble [0 × 1]>
## 3 <split [24/8]> Fold3 <tibble [2 × 4]> <tibble [0 × 1]>
## 4 <split [24/8]> Fold4 <tibble [2 × 4]> <tibble [0 × 1]>
```

Tune

`collect_metrics()` can be used to extract the CV estimate

```
library(tune)

linear_fold_fits <- fit_resamples(
    linear_wf,
    resamples = mtcars_folds
)

collect_metrics(linear_fold_fits)
```

```
## # A tibble: 2 × 6
##   .metric .estimator  mean     n std_err .config
##   <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 rmse    standard   2.97      4  0.384  Preprocessor1_Model1
## 2 rsq     standard   0.842     4  0.0616 Preprocessor1_Model1
```

Tune

Setting `summarize = FALSE` in `collect_metrics()` Allows us the see the individual performance metrics for each fold

```
collect_metrics(linear_fold_fits, summarize = FALSE)
```

```
## # A tibble: 8 × 5
##   id    .metric .estimator .estimate .config
##   <chr> <chr>   <chr>          <dbl> <chr>
## 1 Fold1 rmse    standard       2.93  Preprocessor1_Model1
## 2 Fold1 rsq     standard       0.898 Preprocessor1_Model1
## 3 Fold2 rmse    standard       4.06  Preprocessor1_Model1
## 4 Fold2 rsq     standard       0.659 Preprocessor1_Model1
## 5 Fold3 rmse    standard       2.29  Preprocessor1_Model1
## 6 Fold3 rsq     standard       0.885 Preprocessor1_Model1
## 7 Fold4 rmse    standard       2.61  Preprocessor1_Model1
## 8 Fold4 rsq     standard       0.926 Preprocessor1_Model1
```

Tune

There are some settings we can set with `control_resamples()`.

One of the handiest ones (IMO) is `verbose = TRUE`

```
library(tune)

linear_fold_fits <- fit_resamples(
  linear_wf,
  resamples = mtcars_folds,
  control = control_resamples(verbose = TRUE)
)
```

```
i Fold1: preprocessor 1/1
✓ Fold1: preprocessor 1/1
i Fold1: preprocessor 1/1, model 1/1
✓ Fold1: preprocessor 1/1, model 1/1
i Fold1: preprocessor 1/1, model 1/1 (predictions)
i Fold2: preprocessor 1/1
✓ Fold2: preprocessor 1/1
i Fold2: preprocessor 1/1, model 1/1
✓ Fold2: preprocessor 1/1, model 1/1
i Fold2: preprocessor 1/1, model 1/1 (predictions)
i Fold3: preprocessor 1/1
✓ Fold3: preprocessor 1/1
i Fold3: preprocessor 1/1, model 1/1
✓ Fold3: preprocessor 1/1, model 1/1
i Fold3: preprocessor 1/1, model 1/1 (predictions)
i Fold4: preprocessor 1/1
✓ Fold4: preprocessor 1/1
i Fold4: preprocessor 1/1, model 1/1
✓ Fold4: preprocessor 1/1, model 1/1
i Fold4: preprocessor 1/1, model 1/1 (predictions)
i Fold5: preprocessor 1/1
✓ Fold5: preprocessor 1/1
i Fold5: preprocessor 1/1, model 1/1
✓ Fold5: preprocessor 1/1, model 1/1
i Fold5: preprocessor 1/1, model 1/1 (predictions)
```

Tune

We can also directly specify the metrics that are calculated within each resample

```
library(tune)

linear_fold_fits <- fit_resamples(
    linear_wf,
    resamples = mtcars_folds,
    metrics = metric_set(rmse, rsq, mase)
)

collect_metrics(linear_fold_fits)
```

```
## # A tibble: 3 × 6
##   .metric .estimator  mean     n std_err .config
##   <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 mase    standard   0.459     4  0.213  Preprocessor1_Model1
## 2 rmse    standard   2.97      4  0.384  Preprocessor1_Model1
## 3 rsq     standard   0.842     4  0.0616 Preprocessor1_Model1
```

Bootstrapping

Last week we looked at a couple of different Cross-Validation methods

- Leave-One-Out Cross-Validation (LOOCV)
- K-fold Cross-Validation

Bootstrapping

This week we will look at Bootstrapping

This is a technique that uses resampling with replacement to estimate the uncertainty with a given estimator or statistical learning method

It is a powerful and general statistical tool and can be used with most estimators/methods

Bootstrapping VS Cross-Validation

- Cross-Validation: provide estimates of the test error.
- Bootstrap: provides the standard error of the estimates.

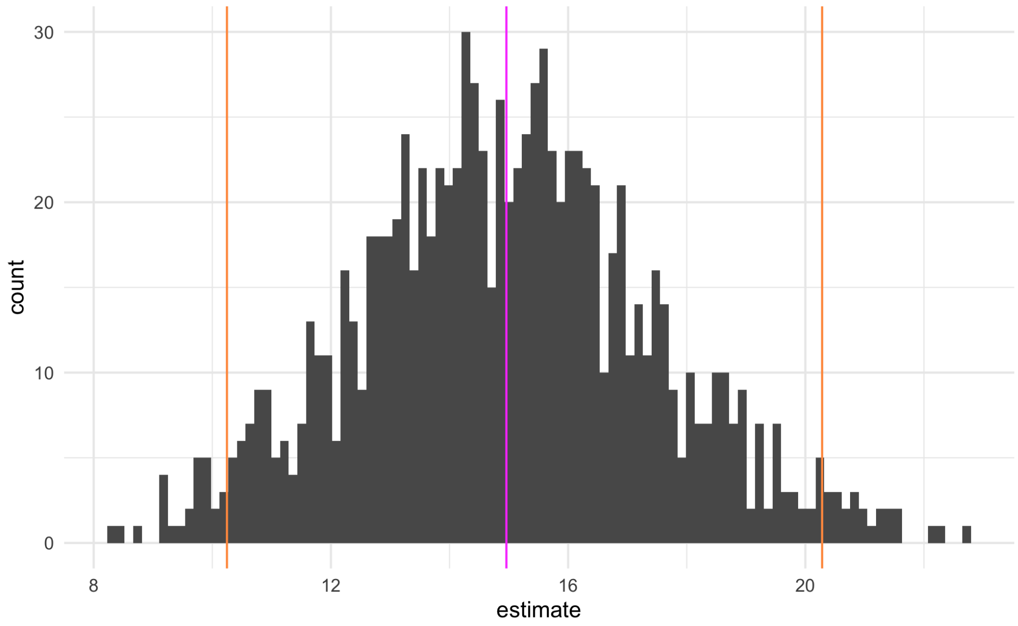
Motivation

Suppose We have an estimate we want to find out how variable it is.

We could collect data n times and calculate the estimates.

We then have a distribution of and can see how well it is doing

1000 realizations
pink line is the mean
orange lines 95% percent quantiles



Motivation

The Problem

We are not always able to conduct multiple data collections at will

Sometimes for resource issues or time-sensitive data

We need the different samples to come from the same underlying distribution

Motivation

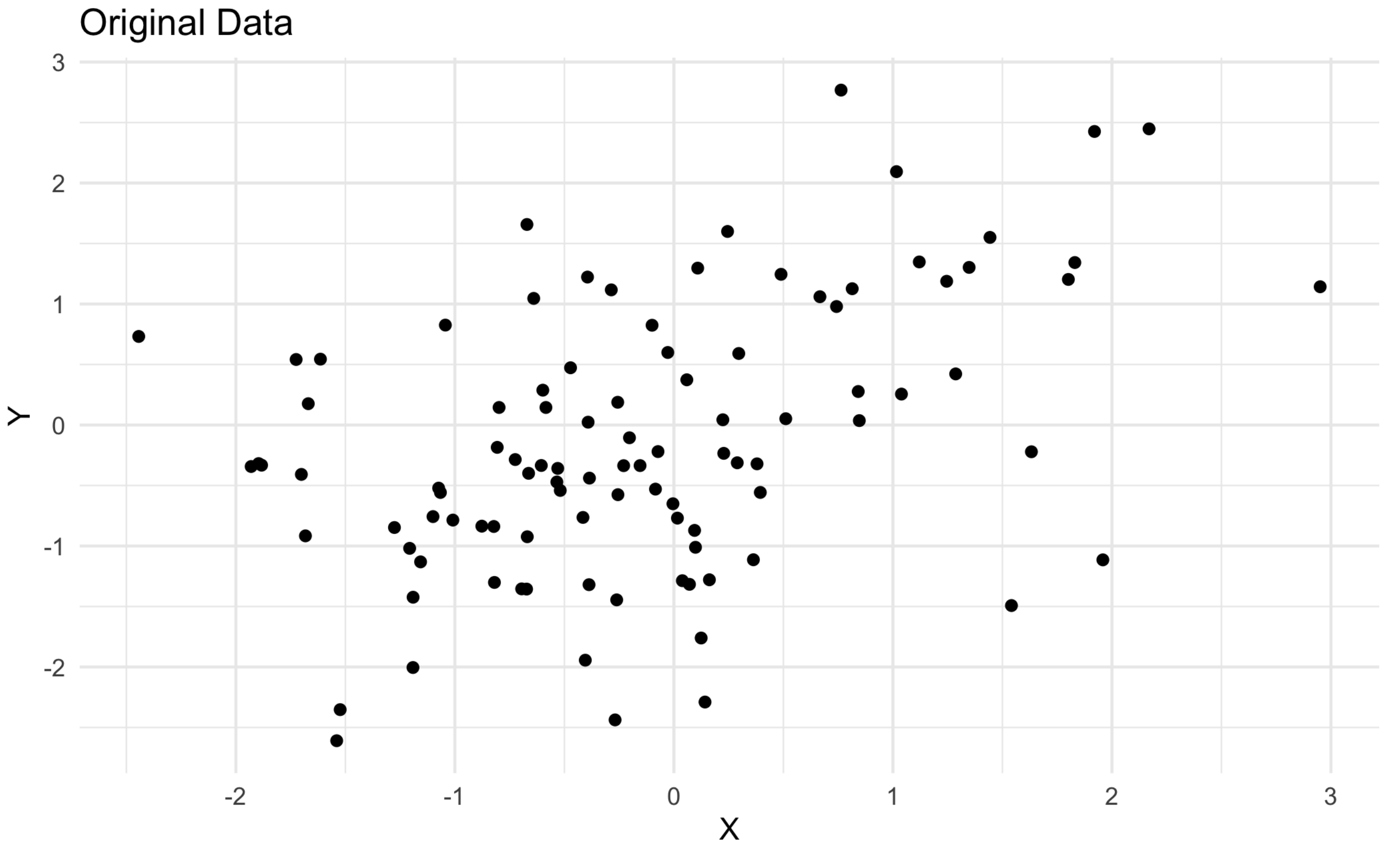
The Solution

We take our one data set and resample the rows with replacement. This allows us to get new data sets that approximate the original data set

If the original data set is close to the underlying true distribution then the resampled data sets are also approximations of the true underlying distribution

Example

From "An Introduction to Statistical Learning"

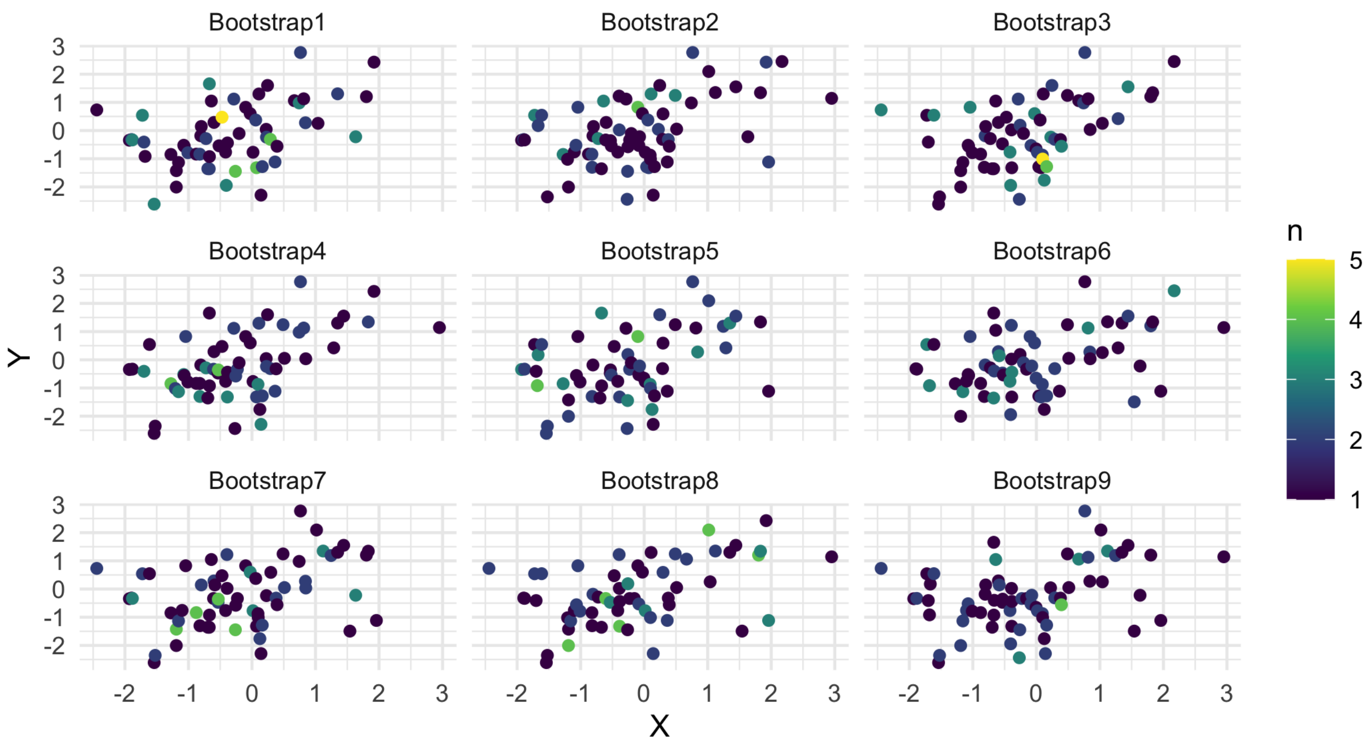


Example

Visualizing multiple bootstraps

9 reaelizations of Bootstrapping

Color indicates the number of times the observation is sampled



Example

We want to minimize

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

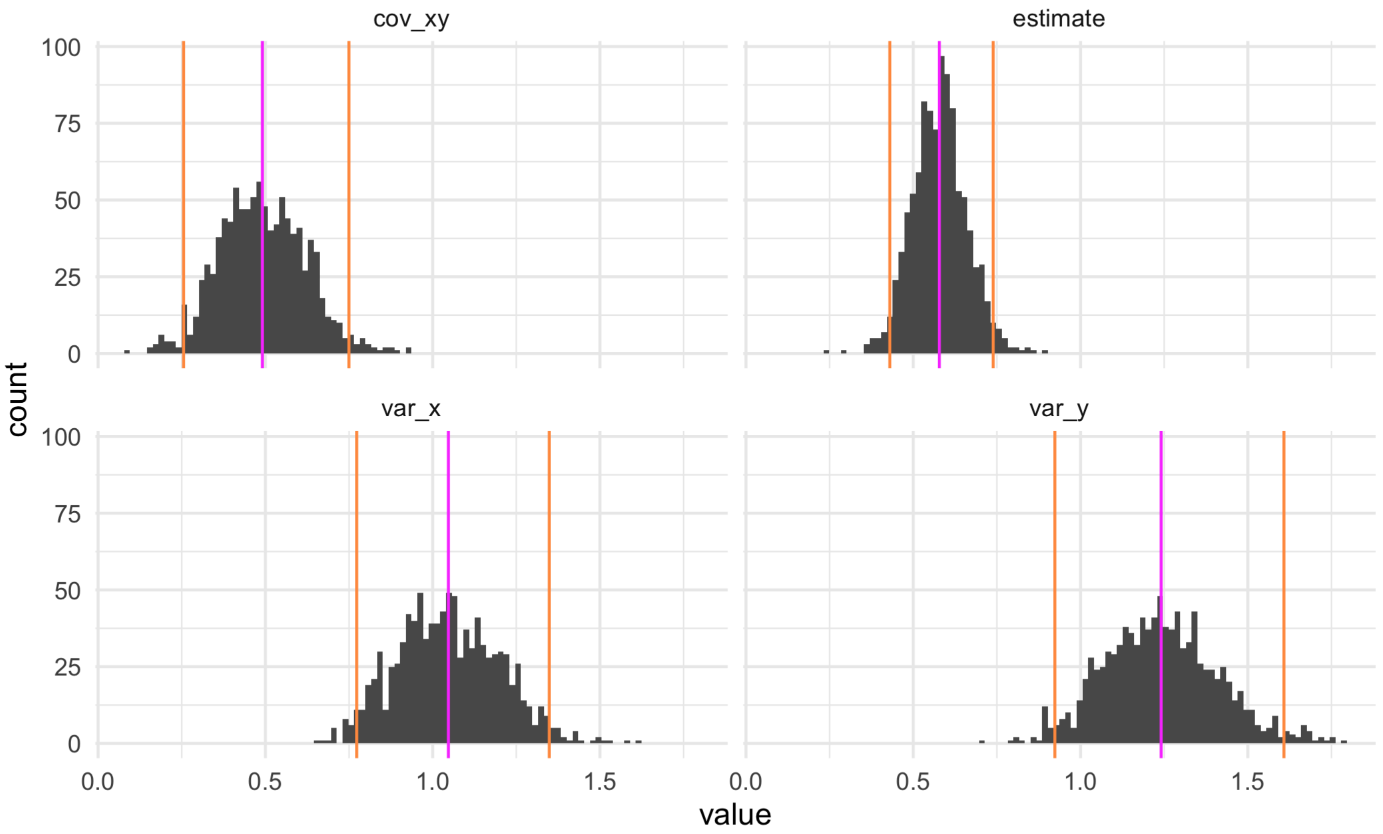
Where $\sigma_X^2 = \mathrm{Var}(X)$, $\sigma_Y^2 = \mathrm{Var}(Y)$, and $\sigma_{XY} = \mathrm{Cov}(X, Y)$

Bootstrapping results

```
## # A tibble: 1,000 × 5
##    id            var_x var_y cov_xy estimate
##    <chr>         <dbl> <dbl>  <dbl>    <dbl>
##  1 Bootstrap0001 1.04   1.33  0.583    0.618
##  2 Bootstrap0002 0.958  1.21  0.416    0.596
##  3 Bootstrap0003 0.950  1.44  0.479    0.671
##  4 Bootstrap0004 0.909  1.27  0.326    0.617
##  5 Bootstrap0005 1.05   1.24  0.413    0.563
##  6 Bootstrap0006 0.747  1.52  0.386    0.759
##  7 Bootstrap0007 0.899  1.33  0.488    0.673
##  8 Bootstrap0008 0.897  1.43  0.515    0.705
##  9 Bootstrap0009 1.21   1.29  0.531    0.527
## 10 Bootstrap0010 0.879  1.06  0.381    0.576
## # … with 990 more rows
```

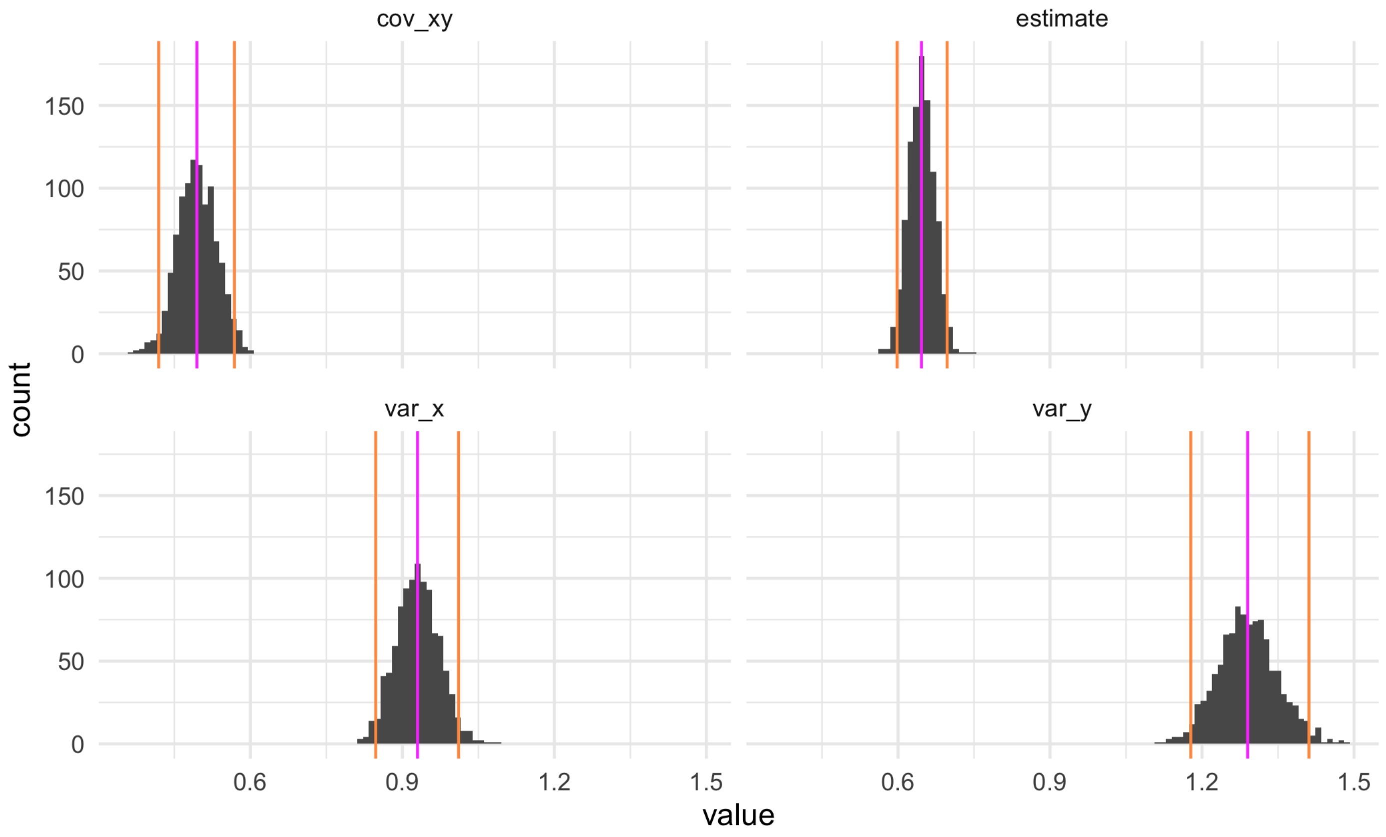
Bootstrapping results

With $n = 100$ in original data set



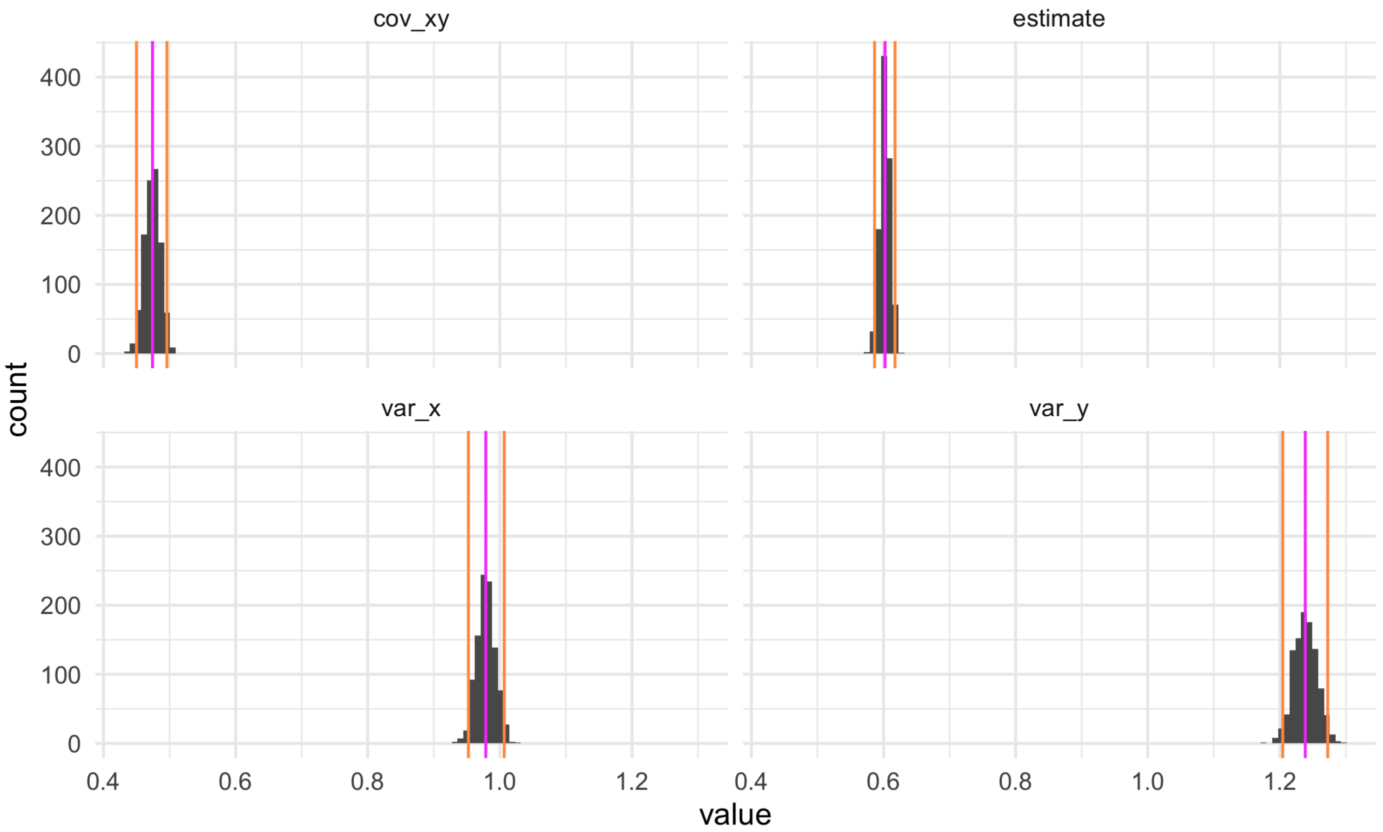
Bootstrapping results

With $n = 1000$ in original data set



Bootstrapping results

With $n = 10000$ in original data set



What size of bootstraps are we looking for?

We are using bootstrapping sizes to be the same size to get a comparative estimate of the variation

Rsample

We are back with `rsample` and the `mtcars` data set

```
library(rsample)

mtcars
```

```
##                      mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4           21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710          22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout   18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant             18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360          14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D           24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230            22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280            19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C           17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
```

Rsample

We can use the `bootstraps()` function on a data.frame to create a `bootstraps` object

```
mtcars_boots <- bootstraps(mtcars, times = 10
mtcars_boots
```

```
## # Bootstrap sampling
## # A tibble: 100 × 2
##    splits          id
##    <list>          <chr>
##  1 <split [32/12]> Bootstrap001
##  2 <split [32/11]> Bootstrap002
##  3 <split [32/12]> Bootstrap003
##  4 <split [32/9]>  Bootstrap004
##  5 <split [32/10]> Bootstrap005
##  6 <split [32/11]> Bootstrap006
##  7 <split [32/12]> Bootstrap007
##  8 <split [32/11]> Bootstrap008
##  9 <split [32/11]> Bootstrap009
## 10 <split [32/11]> Bootstrap010
## # … with 90 more rows
```

Rsample

An under the hood, we have 100 analysis/assesment splits similar to `initial_split()` and `vfold_cv()`

```
mtcars_boots <- bootstraps(mtcars, times = 10
mtcars_boots$splits
```

```
## [[1]]
## <Analysis/Assess/Total>
## <32/12/32>
##
## [[2]]
## <Analysis/Assess/Total>
## <32/12/32>
##
## [[3]]
## <Analysis/Assess/Total>
## <32/9/32>
##
## [[4]]
## <Analysis/Assess/Total>
## <32/14/32>
##
## [[5]]
```

Using resamples in action

We start by creating a linear regression specification and create a `workflow` object with `workflows()`

```
library(parsnip)
linear_spec <- linear_reg() %>%
  set_mode("regression") %>%
  set_engine("lm")

library(workflows)

linear_wf <- workflow() %>%
  add_model(linear_spec) %>%
  add_formula(mpg ~ disp + hp + wt)
```

Tune

We can use `fit_resamples()` to fit the workflow we created within each bootstrap

```
library(tune)

linear_fold_fits <- fit_resamples(
    linear_wf,
    resamples = mtcars_boots
)
```

Tune

The results of this resampling come as a data.frame

```
linear_fold_fits
```

```
## # Resampling results
## # Bootstrap sampling
## # A tibble: 100 × 4
##    splits          id           .metrics         .notes
##    <list>          <chr>        <list>           <list>
##  1 <split [32/12]> Bootstrap001 <tibble [2 × 4]> <tibble [0 × 1]>
##  2 <split [32/12]> Bootstrap002 <tibble [2 × 4]> <tibble [0 × 1]>
##  3 <split [32/9]>  Bootstrap003 <tibble [2 × 4]> <tibble [0 × 1]>
##  4 <split [32/14]> Bootstrap004 <tibble [2 × 4]> <tibble [0 × 1]>
##  5 <split [32/16]> Bootstrap005 <tibble [2 × 4]> <tibble [0 × 1]>
##  6 <split [32/13]> Bootstrap006 <tibble [2 × 4]> <tibble [0 × 1]>
##  7 <split [32/15]> Bootstrap007 <tibble [2 × 4]> <tibble [0 × 1]>
##  8 <split [32/12]> Bootstrap008 <tibble [2 × 4]> <tibble [0 × 1]>
##  9 <split [32/14]> Bootstrap009 <tibble [2 × 4]> <tibble [0 × 1]>
## 10 <split [32/11]> Bootstrap010 <tibble [2 × 4]> <tibble [0 × 1]>
## # … with 90 more rows
```

Tune

`collect_metrics()` can be used to extract the CV estimate

```r
library(tune)

collect_metrics(linear_fold_fits)
```

```
## # A tibble: 2 × 6
##   .metric .estimator  mean     n std_err .config
##   <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 rmse    standard   2.95    100 0.0633  Preprocessor1_Model1
## 2 rsq     standard   0.828   100 0.00670 Preprocessor1_Model1
```

Tune

Setting `summarize = FALSE` in `collect_metrics()` Allows us the see the individual performance metrics for each fold

```
collect_metrics(linear_fold_fits, summarize = FALSE)
```

```
## # A tibble: 200 × 5
##    id           .metric .estimator .estimate .config
##    <chr>        <chr>   <chr>          <dbl> <chr>
##  1 Bootstrap001 rmse    standard       2.78  Preprocessor1_Model1
##  2 Bootstrap001 rsq     standard       0.938 Preprocessor1_Model1
##  3 Bootstrap002 rmse    standard       3.53  Preprocessor1_Model1
##  4 Bootstrap002 rsq     standard       0.752 Preprocessor1_Model1
##  5 Bootstrap003 rmse    standard       2.49  Preprocessor1_Model1
##  6 Bootstrap003 rsq     standard       0.802 Preprocessor1_Model1
##  7 Bootstrap004 rmse    standard       2.52  Preprocessor1_Model1
##  8 Bootstrap004 rsq     standard       0.811 Preprocessor1_Model1
##  9 Bootstrap005 rmse    standard       2.98  Preprocessor1_Model1
## 10 Bootstrap005 rsq     standard       0.826 Preprocessor1_Model1
## # … with 190 more rows
```