Lab 1

Resampling

Shaina Trevino, Akhila Nekkanti, Jonathan Pedroza

Assigned 10/14/20, Due 10/21/20

Contents

Read in the `train.csv` data.

1. Initial Split

Split the data into a training set and a testing set as two named objects. Produce the `class` type for the initial split object and the training and test sets.

```
## <Analysis/Assess/Total>
## <28414/9471/37885>
```

```
## [1] 0.7500066
```

```
## [1] "rsplit"   "mc_split"
```

2. Use code to show the proportion of the `train.csv` data that went to each of the training and test sets.

```
## [1] 0.7500066
```

```
## [1] 0.2499934
```

3. *k*-fold cross-validation

Use 10-fold cross-validation to resample the training data.

4. Use `{purrr}` to add the following columns to your *k*-fold CV object:

- *analysis_n* = the *n* of the analysis set for each fold
- *assessment_n* = the *n* of the assessment set for each fold
- *analysis_p* = the proportion of the analysis set for each fold
- *assessment_p* = the proportion of the assessment set for each fold
- *sped_p* = the proportion of students receiving special education services (`sp_ed_fg`) in the analysis and assessment sets for each fold

5. Please demonstrate that that there are no common values in the `id` columns of the `assessment` data between `Fold01` & `Fold02`, and `Fold09` & `Fold10` (of your 10-fold cross-validation object).

```
## [1] 0
```

```
## [1] 0
```

6. Try to answer these next questions without running similar code on real data.

For the following code `vfold_cv(fictional_train, v = 20)`:

- What is the proportion in the analysis set for each fold?
- What is the proportion in the assessment set for each fold?

7. Use Monte Carlo CV to resample the training data with 20 resamples and .30 of each resample reserved for the assessment sets.

```
## [1] 0.700007
```

```
## [1] 0.299993
```

8. Please demonstrate that that there are common values in the `id` columns of the `assessment` data between `Resample 8` & `Resample 12`, and `Resample 2` & `Resample 20`in your MC CV object.

```
## [1] 2538
```

```
## [1] 2511
```

9. You plan on doing bootstrap resampling with a training set with $n = 500$.

- What is the sample size of an analysis set for a given bootstrap resample? Answer: The analysis set in a bootstrap sample is always the same as the total n, so the analysis set has a sample size of 500.
- What is the sample size of an assessment set for a given bootstrap resample? Answer: The assessment set in a bootstrap resample has an n of 180
- If each row was selected only once for an analysis set:
  - what would be the size of the analysis set? Answer: On average, 63.21% of the original sample ends up in a bootstrap sample, so n = 316.05
  - and what would be the size of the assessment set? Answer: The assessment set has an n of 183.95

```
## [1] 500
```

```
## [1] 180
```

```
## [1] 316.05
```

```
## [1] 183.95
```