Summarizer README

1 Summarise variables/factors by a categorical variable

`summary.factorlist()` is a simple wrapper used to summarise any number of variables by a single categorical variable. This is usually "Table 1" of a study report.

```
library(summarizer)
library(dplyr)
library(stringr)
library(kableExtra)

# Load example dataset, modified version of survival::colon
data(colon_s)

# Table 1 - Patient demographics ----
explanatory = c("age", "age.factor", "sex.factor", "obstruct.factor")
dependent = "perfor.factor"
colon_s %>%
  summary.factorlist(dependent, explanatory, p=T) %>%
    kable(., booktabs = TRUE)
```

	label	levels	No	Yes	pvalue
1	Age (years)	Mean (SD)	59.8 (11.9)	58.4 (13.3)	0.578
2	Age	<40 years	68 (97.1)	2 (2.9)	1.000
3		40-59 years	334 (97.1)	10 (2.9)	
4		60+ years	500 (97.1)	15 (2.9)	
7	Sex	Female	432 (97.1)	13 (2.9)	0.979
8		Male	470 (97.1)	14 (2.9)	
5	Obstruction	No	715 (97.7)	17 (2.3)	0.018
6		Yes	166 (94.3)	10 (5.7)	

`summary.factorlist()` is also commonly used to summarise any number of variables by an *outcome variable* (say dead yes/no).

```
# Table 2 - 5 yr mortality ----
explanatory = c("age.factor", "sex.factor", "obstruct.factor", "perfor.factor")
dependent = 'mort_5yr'
colon_s %>%
  summary.factorlist(dependent, explanatory) %>%
    kable(., booktabs = TRUE)
```

	label	levels	Alive	Died
1	Age	<40 years	31 (46.3)	36 (53.7)
2		40-59 years	208 (61.4)	131 (38.6)
3		60+ years	272 (53.4)	237 (46.6)
8	Sex	Female	243 (55.6)	194 (44.4)
9		Male	268 (56.1)	210 (43.9)
4	Obstruction	No	408 (56.7)	312 (43.3)
5		Yes	89 (51.1)	85 (48.9)
6	Perforation	No	497 (56.0)	391 (44.0)
7		Yes	14 (51.9)	13 (48.1)

2 Summarise regression model results in final table format

The second main feature is the ability to create final tables for logistic `glm()`, hierarchical logistic `lme4::glmer()` and Cox proprotional hazard `survival::coxph()` regression models.

The `summarizer()` "all-in-one" function takes a single dependent variable with a vector of explanatory variable names (continuous or categorical variables) to produce a final table for publication including summary statistics, univariable and multivariable regression analyses. The first columns are those produced by `summary.factorist()`.

2.1 glm

`glm(depdendent ~ explanatory, family="binomial")`

```r
explanatory = c("age.factor", "sex.factor", "obstruct.factor", "perfor.factor")
dependent = 'mort_5yr'
colon_s %>%
  summarizer(dependent, explanatory) %>%
    kable(., booktabs = TRUE)
```

	label	levels	Alive	Died	OR (univariable)	OR (multivariable)
1	Age	<40 years	31 (6.1)	36 (8.9)	-	-
2		40-59 years	208 (40.7)	131 (32.4)	0.54 (0.32-0.92, p=0.023)	0.57 (0.34-0.98, p=0.041)
3		60+ years	272 (53.2)	237 (58.7)	0.75 (0.45-1.25, p=0.270)	0.81 (0.48-1.36, p=0.426)
8	Sex	Female	243 (47.6)	194 (48.0)	-	-
9		Male	268 (52.4)	210 (52.0)	0.98 (0.76-1.27, p=0.889)	0.98 (0.75-1.28, p=0.902)
4	Obstruction	No	408 (82.1)	312 (78.6)	-	-
5		Yes	89 (17.9)	85 (21.4)	1.25 (0.90-1.74, p=0.189)	1.25 (0.90-1.76, p=0.186)
6	Perforation	No	497 (97.3)	391 (96.8)	-	-
7		Yes	14 (2.7)	13 (3.2)	1.18 (0.54-2.55, p=0.672)	1.12 (0.51-2.44, p=0.770)

2.2 multi-level

Where a multivariable model contains a subset of the variables specified in the full univariable set, this can be specified.

```
explanatory = c("age.factor", "sex.factor", "obstruct.factor", "perfor.factor")
explanatory.multi = c("age.factor", "obstruct.factor")
dependent = 'mort_5yr'
colon_s %>%
  summarizer(dependent, explanatory, explanatory.multi) %>%
    kable(., booktabs = TRUE)
```

	label	levels	Alive	Died	OR (univariable)	OR (multivariable)
1	Age	<40 years	31 (6.1)	36 (8.9)	-	-
2		40-59 years	208 (40.7)	131 (32.4)	0.54 (0.32-0.92, p=0.023)	0.57 (0.34-0.98, p=0.041)
3		60+ years	272 (53.2)	237 (58.7)	0.75 (0.45-1.25, p=0.270)	0.81 (0.48-1.36, p=0.424)
8	Sex	Female	243 (47.6)	194 (48.0)	-	-
9		Male	268 (52.4)	210 (52.0)	0.98 (0.76-1.27, p=0.889)	-
4	Obstruction	No	408 (82.1)	312 (78.6)	-	-
5		Yes	89 (17.9)	85 (21.4)	1.25 (0.90-1.74, p=0.189)	1.26 (0.90-1.76, p=0.176)
6	Perforation	No	497 (97.3)	391 (96.8)	-	-
7		Yes	14 (2.7)	13 (3.2)	1.18 (0.54-2.55, p=0.672)	-

2.3 Random effects.

```
lme4::glmer(dependent ~ explanatory + (1 | random_effect), family="binomial")
```

```
explanatory = c("age.factor", "sex.factor", "obstruct.factor", "perfor.factor")
explanatory.multi = c("age.factor", "obstruct.factor")
random.effect = "hospital"
dependent = 'mort_5yr'
colon_s %>%
  summarizer(dependent, explanatory, explanatory.multi, random.effect) %>%
    kable(., booktabs = TRUE)
```

	label	levels	Alive	Died	OR (univariable)	OR (multilevel)
1	Age	<40 years	31 (6.1)	36 (8.9)	-	-
2		40-59 years	208 (40.7)	131 (32.4)	0.54 (0.32-0.92, p=0.023)	0.73 (0.38-1.40, p=0.342)
3		60+ years	272 (53.2)	237 (58.7)	0.75 (0.45-1.25, p=0.270)	1.01 (0.53-1.90, p=0.984)
8	Sex	Female	243 (47.6)	194 (48.0)	-	-
9		Male	268 (52.4)	210 (52.0)	0.98 (0.76-1.27, p=0.889)	-
4	Obstruction	No	408 (82.1)	312 (78.6)	-	-
5		Yes	89 (17.9)	85 (21.4)	1.25 (0.90-1.74, p=0.189)	1.24 (0.83-1.85, p=0.292)
6	Perforation	No	497 (97.3)	391 (96.8)	-	-
7		Yes	14 (2.7)	13 (3.2)	1.18 (0.54-2.55, p=0.672)	-

2.4 with metrics

`metrics=TRUE` provides common model metrics.

```
colon_s %>%
  summarizer(dependent, explanatory, explanatory.multi,  metrics=TRUE) %>%
    kable(., booktabs = TRUE)
```

	label	levels	Alive	Died	OR (univariable)	OR (multivariable)
1	Age	<40 years	31 (6.1)	36 (8.9)	-	-
2		40-59 years	208 (40.7)	131 (32.4)	0.54 (0.32-0.92, p=0.023)	0.57 (0.34-0.98, p=0.041)
3		60+ years	272 (53.2)	237 (58.7)	0.75 (0.45-1.25, p=0.270)	0.81 (0.48-1.36, p=0.424)
8	Sex	Female	243 (47.6)	194 (48.0)	-	-
9		Male	268 (52.4)	210 (52.0)	0.98 (0.76-1.27, p=0.889)	-
4	Obstruction	No	408 (82.1)	312 (78.6)	-	-
5		Yes	89 (17.9)	85 (21.4)	1.25 (0.90-1.74, p=0.189)	1.26 (0.90-1.76, p=0.176)
6	Perforation	No	497 (97.3)	391 (96.8)	-	-
7		Yes	14 (2.7)	13 (3.2)	1.18 (0.54-2.55, p=0.672)	-

x
Number in dataframe = 929, Number in model = 894, Missing = 35, AIC = 1226.8, C-statistic = 0.555

2.5 Cox proportional hazards

```
survival::coxph(dependent ~ explanatory)
```

```
explanatory = c("age.factor", "sex.factor", "obstruct.factor", "perfor.factor")
dependent = "Surv(time, status)"

colon_s %>%
    summarizer(dependent, explanatory) %>%
    kable(., booktabs = TRUE)
```

	label	levels	HR (univariable)	HR (multivariable)
1	Age	<40 years	-	-
2		40-59 years	0.76 (0.53-1.09, p=0.132)	0.79 (0.55-1.13, p=0.196)
3		60+ years	0.93 (0.66-1.31, p=0.668)	0.98 (0.69-1.40, p=0.926)
8	Sex	Female	-	-
9		Male	1.01 (0.84-1.22, p=0.888)	1.02 (0.85-1.23, p=0.812)
4	Obstruction	No	-	-
5		Yes	1.29 (1.03-1.62, p=0.028)	1.30 (1.03-1.64, p=0.026)
6	Perforation	No	-	-
7		Yes	1.17 (0.70-1.95, p=0.556)	1.08 (0.64-1.81, p=0.785)

3 Subsets

Rather than going all-in-one, any number of subset models can be manually added on to a `summary.factorlist()` table using `summarizer.merge()`. This is particularly useful when models take a long-time to run or are complicated.

3.1 glm

Note requirement for `glm.id=TRUE`. `fit2df` is a subfunction extracting most common models to a dataframe.

```
explanatory = c("age.factor", "sex.factor", "obstruct.factor", "perfor.factor")
explanatory.multi = c("age.factor", "obstruct.factor")
```

```
random.effect = "hospital"
dependent = 'mort_5yr'

# Separate tables
colon_s %>%
  summary.factorlist(dependent, explanatory, glm.id=TRUE) -> example.summary

colon_s %>%
  glmuni(dependent, explanatory) %>%
  fit2df(estimate.suffix=" (univariable)") -> example.univariable

colon_s %>%
  glmmulti(dependent, explanatory) %>%
  fit2df(estimate.suffix=" (multivariable)") -> example.multivariable


colon_s %>%
  glmmixed(dependent, explanatory, random.effect) %>%
  fit2df(estimate.suffix=" (multilevel") -> example.multilevel

# Pipe together
example.summary %>%
  summarizer.merge(example.univariable) %>%
  summarizer.merge(example.multivariable) %>%
  summarizer.merge(example.multilevel) %>%
  select(-c(glm.id, index)) -> example.final
example.final %>%
    kable(., booktabs = TRUE) %>%
    kable_styling(latex_options = "scale_down")
```

	label	levels	Alive	Died	OR (univariable)	OR (multivariable)	OR (multilevel
1	Age	<40 years	31 (46.3)	36 (53.7)	-	-	-
2		40-59 years	208 (61.4)	131 (38.6)	0.54 (0.32-0.92, p=0.023)	0.57 (0.34-0.98, p=0.041)	0.75 (0.39-1.44, p=0.382)
3		60+ years	272 (53.4)	237 (46.6)	0.75 (0.45-1.25, p=0.270)	0.81 (0.48-1.36, p=0.426)	1.03 (0.55-1.96, p=0.916)
8	Sex	Female	243 (55.6)	194 (44.4)	-	-	-
9		Male	268 (56.1)	210 (43.9)	0.98 (0.76-1.27, p=0.889)	0.98 (0.75-1.28, p=0.902)	0.80 (0.58-1.11, p=0.180)
4	Obstruction	No	408 (56.7)	312 (43.3)	-	-	-
5		Yes	89 (51.1)	85 (48.9)	1.25 (0.90-1.74, p=0.189)	1.25 (0.90-1.76, p=0.186)	1.23 (0.82-1.83, p=0.320)
6	Perforation	No	497 (56.0)	391 (44.0)	-	-	-
7		Yes	14 (51.9)	13 (48.1)	1.18 (0.54-2.55, p=0.672)	1.12 (0.51-2.44, p=0.770)	1.03 (0.43-2.51, p=0.940)

3.2 Cox Proportional Hazards example with separate tables merged together.

```
explanatory = c("age.factor", "sex.factor", "obstruct.factor", "perfor.factor")
explanatory.multi = c("age.factor", "obstruct.factor")
dependent = "Surv(time, status)"

# Separate tables
colon_s %>%
    summary.factorlist(dependent, explanatory, glm.id=TRUE) -> example2.summary
```

```
## Warning in summary.factorlist(., dependent, explanatory, glm.id = TRUE):
```

```
## Dependent variable is a survival object
```

```
colon_s %>%
    coxphuni(dependent, explanatory) %>%
    fit2df(estimate.suffix=" (univariable)") -> example2.univariable

colon_s %>%
  coxphmulti(dependent, explanatory.multi) %>%
  fit2df(estimate.suffix=" (multivariable)") -> example2.multivariable

# Pipe together
example2.summary %>%
    summarizer.merge(example2.univariable) %>%
    summarizer.merge(example2.multivariable) %>%
    select(-c(glm.id, index)) -> example2.final
example2.final %>%
    kable(., booktabs = TRUE)
```

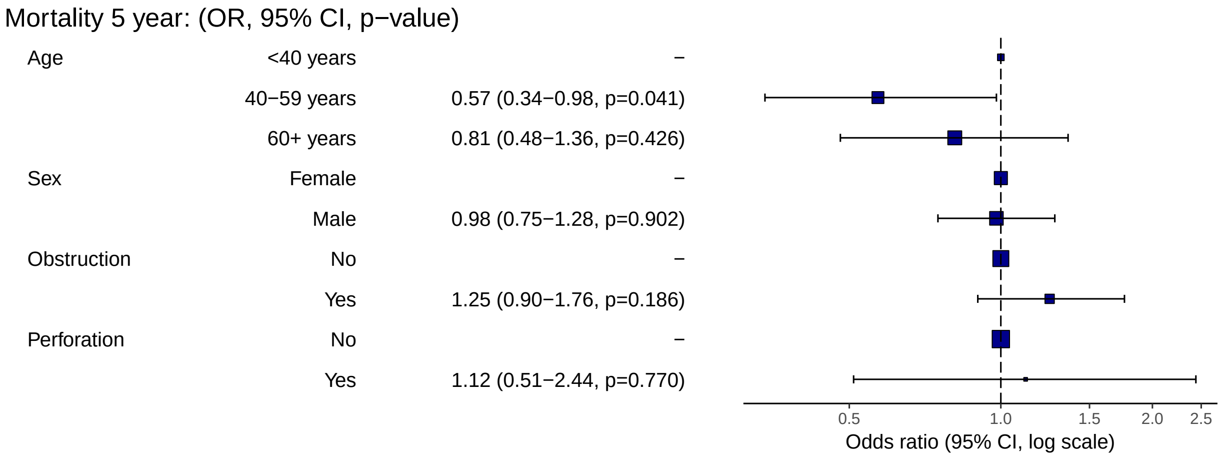
	label	levels	all	HR (univariable)	HR (multivariable)
1	Age	<40 years	70 (100.0)	-	-
2		40-59 years	344 (100.0)	0.76 (0.53-1.09, p=0.132)	0.79 (0.55-1.14, p=0.203)
3		60+ years	515 (100.0)	0.93 (0.66-1.31, p=0.668)	0.99 (0.70-1.40, p=0.943)
8	Sex	Female	445 (100.0)	-	-
9		Male	484 (100.0)	1.01 (0.84-1.22, p=0.888)	-
4	Obstruction	No	732 (100.0)	-	-
5		Yes	176 (100.0)	1.29 (1.03-1.62, p=0.028)	1.31 (1.04-1.64, p=0.022)
6	Perforation	No	902 (100.0)	-	-
7		Yes	27 (100.0)	1.17 (0.70-1.95, p=0.556)	-

4 Summarise regression model results in plot

Models can be summarized with odds ratio/hazard ratio plots using `or.plot` or `hr.plot` (hr.plot not fully tested).

```
# OR plot
explanatory = c("age.factor", "sex.factor", "obstruct.factor", "perfor.factor")
dependent = 'mort_5yr'
colon_s %>%
  or.plot(dependent, explanatory)
```

```
## Warning: Removed 4 rows containing missing values (geom_errorbarh).
```

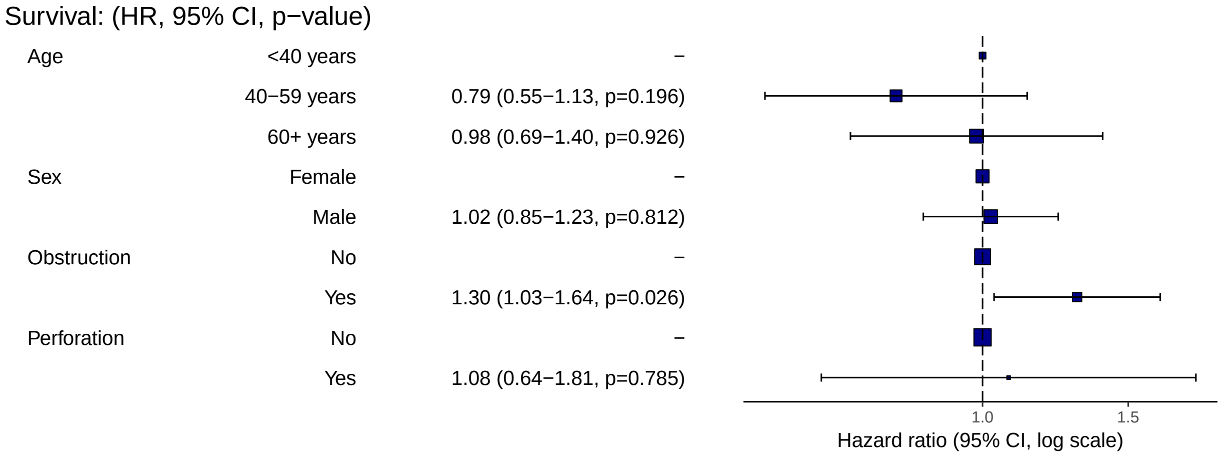


```r
# Previously fitted models (`glmmulti`) can be provided directly to `glmfit`

# HR plot (not fully tested)
explanatory = c("age.factor", "sex.factor", "obstruct.factor", "perfor.factor")
dependent = "Surv(time, status)"
colon_s %>%
  hr.plot(dependent, explanatory, dependent_label = "Survival")
```

```
## Warning in summary.factorlist(df, dependent, explanatory, glm.id = TRUE):
## Dependent variable is a survival object

## Warning in summary.factorlist(df, dependent, explanatory, glm.id = TRUE):
## Removed 4 rows containing missing values (geom_errorbarh).
```



```r
# Previously fitted models (`coxphmulti`) can be provided directly using `coxfit`
```

Our own particular `Rstan` models are supported and will be documented in the future. Broadly, if you are running (hierarchical) logistic regression models in Stan with coefficients specified as a vector labelled `beta`, then `fit2df()` will work directly on the `stanfit` object in a similar manner to if it was a `glm` or `glmerMod` object.

5 Notes

Use `Hmisc::label()` to assign labels to variables for tables and plots.

```r
label(colon_s$age.factor) = "Age (years)"
```

Export dataframe tables directly or to R Markdown using `knitr::kable()`.

Note wrapper `summary.missing()` can be useful. Wraps `mice::md.pattern`.

```r
colon_s %>%
  summary.missing(dependent, explanatory) %>%
    kable(., booktabs = TRUE)
```

	sex.factor	perfor.factor	age.factor	obstruct.factor	
908	1	1	1	1	0
21	1	1	1	0	1
	0	0	0	21	21