Package ‘moderndive’

July 6, 2018

Type Package

Title Tidyverse-Friendly Introductory Linear Regression

Version 0.2.0

Maintainer Albert Y. Kim <albert.ys.kim@gmail.com>

Description Datasets and wrapper functions for tidyverse-friendly introductory linear regression, used in ModernDive: An Introduction to Statistical and Data Sciences via R available at <http://moderndive.com/> and DataCamp's Modeling with Data in the Tidyverse available at <https://www.datacamp.com/courses/modeling-with-data-in-the-tidyverse>.

Depends R (>= 3.2.4)

License MIT + file LICENSE

Encoding UTF-8

LazyData true

URL https://github.com/ModernDive/moderndive_package

BugReports https://github.com/ModernDive/moderndive_package/issues

Imports magrittr, dplyr, tibble, janitor, broom (>= 0.4.3),
formula.tools, stringr, knitr, assertive, infer, rlang

RoxygenNote 6.0.1

Suggests testthat, covr, ggplot2

NeedsCompilation no

Author Albert Y. Kim [cre],
Chester Ismay [aut],
Andrew Bray [ctb]

Repository CRAN

Date/Publication 2018-07-06 15:40:03 UTC

R topics documented:

`bowl` *A sampling bowl of red and white balls*

Description

A sampling bowl used as the population in a simulated sampling exercise. Also known as the urn sampling framework `https://en.wikipedia.org/wiki/Urn_problem`.

Usage

```
bowl
```

Format

A data frame 2400 rows representing different balls in the bowl, of which 900 are red and 1500 are white.

ball_ID ID variable used to denote all balls. Note this value is not marked on the balls themselves

color color of ball: red or white

Examples

```
library(dplyr)
library(ggplot2)

# Take 10 different samples of size n = 50 balls from bowl
bowl_samples_simulated <- bowl %>%
  rep_sample_n(50, reps = 10)

# Compute 10 different p_hats (prop red) based on 10 different samples of
# size n = 50
p_hats <- bowl_samples_simulated %>%
  group_by(replicate, color) %>%
  summarize(count = n()) %>%
  mutate(proportion = count/50) %>%
  filter(color == "red")

# Plot sampling distribution
ggplot(p_hats, aes(x = proportion)) +
  geom_histogram(binwidth = 0.05) +
  labs(x = expression(hat(p)), y = "Number of samples",
  title = "Sampling distribution of p_hat based 10 samples of size n = 50")
```

`bowl_samples` *Sampling from a tub of balls*

Description

Counting the number of red balls in 10 samples of size n = 50 balls from https://github.com/moderndive/moderndive/blob/master/data-raw/sampling_bowl.jpeg

Usage

```
bowl_samples
```

Format

A data frame 10 rows representing different groups of students' samples of size n = 50 and 5 variables

group Group name

red Number of red balls sampled

white Number of white balls sampled

green Number of green balls sampled

n Total number of balls samples

Examples

```
library(dplyr)
library(ggplot2)

# Compute proportion red
bowl_samples <- bowl_samples %>%
  mutate(prop_red = red / n)

# Plot sampling distributions
ggplot(bowl_samples, aes(x = prop_red)) +
  geom_histogram(binwidth = 0.05) +
  labs(x = expression(hat(p)), y = "Number of samples",
  title = "Sampling distribution of p_hat based 10 samples of size n = 50")
```

evals *Teaching evaluations at the UT Austin*

Description

The data are gathered from end of semester student evaluations for a large sample of professors from the University of Texas at Austin. In addition, six students rate the professors' physical appearance. The result is a data frame where each row contains a different course and each column has information on either the course or the professor https://www.openintro.org/stat/data/?data=evals

Usage

```
evals
```

Format

A data frame with 463 observations on the following 13 variables.

ID Identification variable used to distinguish rows.

score Average professor evaluation score: (1) very unsatisfactory - (5) excellent.

age Age of professor.

bty_avg Average beauty rating of professor.

gender Gender of professor: female, male.

ethnicity Ethnicity of professor: not minority, minority.

language Language of school where professor received education: English or non-English.

rank Rank of professor: teaching, tenure track, tenured.

pic_outfit Outfit of professor in picture: not formal, formal.

pic_color Color of professor's picture: color, black & white.

cls_did_eval Number of students in class who completed evaluation.

cls_students Total number of students in class.

cls_level Class level: lower, upper.

Source

Çetinkaya-Rundel M, Morgan KL, Stangl D. 2013. Looking Good on Course Evaluations. CHANCE 26(2). http://chance.amstat.org/2013/04/looking-good/

Examples

```
library(dplyr)
glimpse(evals)
```

get_correlation *Get correlation value in a tidy way*

Description

Determine the Pearson correlation coefficient between two variables in a data frame using pipeable and formula-friendly syntax

Usage

```
get_correlation(data, formula)
```

Arguments

`data`	a data frame object
`formula`	a formula with the response variable name on the left and the explanatory variable name on the right

Value

A 1x1 data frame storing the correlation value

Examples

```
library(moderndive)

# Compute correlation between mpg and cyl:
mtcars %>%
   get_correlation(formula = mpg ~ cyl)
```

get_regression_points *Get regression points*

Description

Output information on each point/observation used in an `lm()` regression in "tidy" format. This function is a wrapper function for `broom::augment()` and renames the variables to have more intuitive names.

Usage

```
get_regression_points(model, digits = 3, print = FALSE, newdata = NULL)
```

Arguments

model	an lm() model object
digits	number of digits precision in output table
print	If TRUE, return in print format suitable for R Markdown
newdata	A new data frame of points/observations to apply model to obtain new fitted values and/or predicted values y-hat. Note the format of newdata must match the format of the original data used to fit model.

Value

A tibble-formatted regression table of outcome/response variable, all explanatory/predictor variables, the fitted/predicted value, and residual.

See Also

augment, get_regression_table, get_regression_summaries

Examples

```
library(moderndive)
library(dplyr)
library(tibble)

# Fit lm() regression:
mpg_model <- lm(mpg ~ cyl, data = mtcars)

# Get information on all points in regression:
get_regression_points(mpg_model)

# Create training and test set based on mtcars:
mtcars <- mtcars %>%
  rownames_to_column(var = "model")
training_set <- mtcars %>%
  sample_frac(0.5)
test_set <- mtcars %>%
  anti_join(training_set, by = "model")

# Fit model to training set:
mpg_model_train <- lm(mpg ~ cyl, data = training_set)

# Make predictions on test set:
get_regression_points(mpg_model_train, newdata = test_set)
```

get_regression_summaries

Get regression summary values

Description

Output scalar summary statistics for an `lm()` regression in "tidy" format. This function is a wrapper function for `broom::glance()`.

Usage

```
get_regression_summaries(model, digits = 3, print = FALSE)
```

Arguments

`model`	an `lm()` model object
`digits`	number of digits precision in output table
`print`	If TRUE, return in print format suitable for R Markdown

Value

A single-row tibble with regression summaries. Ex: `r_squared` and `mse`.

See Also

`glance`, `get_regression_table`, `get_regression_points`

Examples

```
library(moderndive)

# Fit lm() regression:
mpg_model <- lm(mpg ~ cyl, data = mtcars)

# Get regression summaries:
get_regression_summaries(mpg_model)
```

get_regression_table *Get regression table*

Description

Output regression table for an `lm()` regression in "tidy" format. This function is a wrapper function for `broom::tidy()` and includes confidence intervals in the output table by default.

Usage

```
get_regression_table(model, digits = 3, print = FALSE)
```

Arguments

`model`	an `lm()` model object
`digits`	number of digits precision in output table
`print`	If TRUE, return in print format suitable for R Markdown

Value

A tibble-formatted regression table along with lower and upper end points of all confidence intervals for all parameters `lower_ci` and `upper_ci`.

See Also

`tidy`, `get_regression_points`, `get_regression_summaries`

Examples

```
library(moderndive)

# Fit lm() regression:
mpg_model <- lm(mpg ~ cyl, data = mtcars)

# Get regression table:
get_regression_table(mpg_model)
```

house_prices *House Sales in King County, USA*

Description

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. This dataset was obtained from Kaggle.com https://www.kaggle.com/harlfoxem/housesalesprediction/data

Usage

```
house_prices
```

Format

A data frame with 21613 observations on the following 21 variables.

id a notation for a house

date Date house was sold

price Price is prediction target

bedrooms Number of Bedrooms/House

bathrooms Number of bathrooms/bedrooms

sqft_living square footage of the home

sqft_lot square footage of the lot

floors Total floors (levels) in house

waterfront House which has a view to a waterfront

view Has been viewed

condition How good the condition is (Overall)

grade overall grade given to the housing unit, based on King County grading system

sqft_above square footage of house apart from basement

sqft_basement square footage of the basement

yr_built Built Year

yr_renovated Year when house was renovated

zipcode zip code

lat Latitude coordinate

long Longitude coordinate

sqft_living15 Living room area in 2015 (implies– some renovations) This might or might not have affected the lotsize area

sqft_lot15 lotSize area in 2015 (implies– some renovations)

Source

Kaggle https://www.kaggle.com/harlfoxem/housesalesprediction. Note data is released under a CC0: Public Domain license.

Examples

```
library(dplyr)
library(ggplot2)

# Create variable log of house price
house_prices <- house_prices %>%
  mutate(log_price = log(price))

# Plot histogram of log of house price
ggplot(house_prices, aes(x = log_price)) +
  geom_histogram()
```

moderndive — *moderndive - Tidyverse-Friendly Introductory Linear Regression*

Description

Datasets and wrapper functions for tidyverse-friendly introductory linear regression, used in ModernDive: An Introduction to Statistical and Data Sciences via R available at http://moderndive.com/ and DataCamp's Modeling with Data in the Tidyverse available at https://www.datacamp.com/courses/modeling-with-data-in-the-tidyverse.

Examples

```
library(moderndive)

# Fit regression model:
mpg_model <- lm(mpg ~ hp, data = mtcars)

# Regression tables:
get_regression_table(mpg_model)

# Information on each point in a regression:
get_regression_points(mpg_model)

# Regression summaries
get_regression_summaries(mpg_model)
```

mythbusters_yawn *Data from Mythbusters' study on contagiousness of yawning*

Description

From a study on whether yawning is contagious https://www.imdb.com/title/tt0768479/. The data here was derived from the final proportions of yawns given in the show.

Usage

```
mythbusters_yawn
```

Format

A data frame of 50 rows representing each of the 50 participants in the study.

subj integer value corresponding to identifier variable of subject ID

group string of either "seed", participant was shown a yawner, or "control", participant was not shown a yawner

yawn string of either "yes", the participant yawned, or "no", the participant did not yawn

Examples

```
library(ggplot2)

# Plot both variables as a stacked proportional bar chart
ggplot(mythbusters_yawn, aes(x = group, fill = yawn)) +
  geom_bar(position = "fill") +
  labs(x = "", y = "Proportion",
  title = "Proportion of yawn and not yawn for each group")
```

pennies *A population of 800 pennies sampled in 2011*

Description

A dataset of 800 pennies to be treated as a sampling population. Data on these pennies were recorded in 2011.

Usage

```
pennies
```

Format

A data frame of 800 rows representing different pennies and 2 variables

year Year of minting

age_in_2011 Age in 2011

Source

StatCrunch https://www.statcrunch.com/app/index.php?dataid=301596

Examples

```
library(dplyr)
library(ggplot2)

# Take 25 different samples of size n = 50 pennies from population
many_samples <- pennies %>%
  rep_sample_n(size = 50, reps = 25)
many_samples

# Compute mean year of minting for each sample
sample_means <- many_samples %>%
  group_by(replicate) %>%
  summarize(mean_year = mean(year))

# Plot sampling distribution
ggplot(sample_means, aes(x = mean_year)) +
  geom_histogram(binwidth = 1, color = "white") +
  labs(x = expression(bar(x)), y = "Number of samples",
  title = "Sampling distribution of x_bar based 25 samples of size n = 50")
```

pennies_sample *A random sample of 40 pennies sampled from the* pennies *data frame*

Description

A dataset of 40 pennies to be treated as a random sample with pennies acting as the population. Data on these pennies were recorded in 2011.

Usage

```
pennies_sample
```

Format

A data frame of 40 rows representing 40 randomly sampled pennies from pennies and 2 variables

year Year of minting

age_in_2011 Age in 2011

Source

StatCrunch https://www.statcrunch.com/app/index.php?dataid=301596

See Also

pennies

Examples

```
library(dplyr)
library(ggplot2)

# Take 50 different resamples/bootstraps from the original sample
many_bootstraps <- pennies_sample %>%
  rep_sample_n(size = 40, replace = TRUE, reps = 50)
many_bootstraps

# Compute mean year of minting for each bootstrap sample
bootstrap_means <- many_bootstraps %>%
  group_by(replicate) %>%
  summarize(mean_year = mean(year))

# Plot sampling distribution
ggplot(bootstrap_means, aes(x = mean_year)) +
  geom_histogram(binwidth = 1, color = "white") +
  labs(x = expression(bar(x)), y = "Number of samples",
  title = "Bootstrap distribution of x_bar based 50 resamples of size n = 40")
```

tactile_prop_red *Tactile sampling from a tub of balls*

Description

Counting the number of red balls in 33 tactile samples of size n = 50 balls from https://github.com/moderndive/moderndive/blob/master/data-raw/sampling_bowl.jpeg

Usage

```
tactile_prop_red
```

Format

A data frame of 33 rows representing different groups of students' samples of size n = 50 and 4 variables

group Group members

replicate Replicate number

red_balls Number of red balls sampled out of 50

prop_red Proportion red balls out of 50

Examples

```
library(ggplot2)

# Plot sampling distributions
ggplot(tactile_prop_red, aes(x = prop_red)) +
  geom_histogram(binwidth = 0.025) +
  labs(x = expression(hat(p)), y = "Number of samples",
  title = "Sampling distribution of p_hat based 33 samples of size n = 50")
```

Index