

Package ‘moderndive’

May 13, 2022

Type Package

Title Tidyverse-Friendly Introductory Linear Regression

Version 0.5.4

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

Description Datasets and wrapper functions for tidyverse-friendly introductory linear regression, used in “Statistical Inference via Data Science: A ModernDive into R and the Tidyverse” available at <<https://moderndive.com/>>.

Depends R (>= 3.4.0)

License GPL-3

Encoding UTF-8

LazyData true

URL <https://moderndive.github.io/moderndive/>,
<https://github.com/moderndive/moderndive/>

BugReports <https://github.com/moderndive/moderndive/issues>

Imports magrittr, dplyr, ggplot2, tibble, janitor, broom (>= 0.4.3),
formula.tools, stringr, knitr, infer, rlang (>= 0.2.0), glue

RoxygenNote 7.1.2

Suggests testthat, covr, rmarkdown, vdiff, openintro, patchwork,
viridis, readr, nycflights13

VignetteBuilder knitr

NeedsCompilation no

Author Albert Y. Kim [aut, cre] (<<https://orcid.org/0000-0001-7824-306X>>),
Chester Ismay [aut] (<<https://orcid.org/0000-0003-2820-2547>>),
Andrew Bray [ctb] (<<https://orcid.org/0000-0002-4037-7414>>),
Delaney Moran [ctb],
Evgeni Chasnovski [ctb] (<<https://orcid.org/0000-0002-1617-4019>>),
Will Hopper [ctb] (<<https://orcid.org/0000-0002-7848-1946>>),
Benjamin S. Baumer [ctb] (<<https://orcid.org/0000-0002-3279-0516>>),
Marium Tapal [ctb] (<<https://orcid.org/0000-0001-5093-6462>>),
Wayne Ndlovu [ctb],

Catherine Peppers [ctb],
 Annah Mutaya [ctb],
 Anushree Goswami [ctb],
 Ziyue Yang [ctb] (<<https://orcid.org/0000-0002-9299-8327>>),
 Clara Li [ctb] (<<https://orcid.org/0000-0003-2456-0849>>),
 Caroline McKenna [ctb],
 Catherine Park [ctb] (<<https://orcid.org/0000-0002-8273-9620>>),
 Abbie Benfield [ctb],
 Georgia Gans [ctb],
 Kacey Jean-Jacques [ctb],
 Swaha Bhattacharya [ctb],
 Vivian Almaraz [ctb],
 Elle Jo Whalen [ctb],
 Jacqueline Chen [ctb],
 Michelle Flesaker [ctb],
 Irene Foster [ctb],
 Aushanae Haller [ctb],
 Benjamin Bruncati [ctb] (<<https://orcid.org/0000-0001-8545-5984>>),
 Quinn White [ctb] (<<https://orcid.org/0000-0001-5399-0237>>),
 Tianshu Zhang [ctb] (<<https://orcid.org/0000-0002-3004-4472>>),
 Katelyn Diaz [ctb] (<<https://orcid.org/0000-0001-6108-1682>>),
 Rose Porta [ctb],
 Renee Wu [ctb],
 Arris Moise [ctb],
 Kate Phan [ctb],
 Grace Hartley [ctb],
 Silas Weden [ctb],
 Emma Vejcik [ctb] (<<https://orcid.org/0000-0001-5093-6462>>),
 Nikki Schuldt [ctb],
 Tess Goldmann [ctb],
 Hongtong Lin [ctb],
 Alejandra Munoz [ctb],
 Elina Gordon-Halpern [ctb],
 Haley Schmidt [ctb] (<<https://orcid.org/0000-0002-6184-2266>>)

Repository CRAN

Date/Publication 2022-05-13 06:00:02 UTC

R topics documented:

alaska_flights	3
amazon_books	4
avocados	5
babies	6
bowl	7
bowl_samples	8
bowl_sample_1	8
coffee_ratings	9

DD_vs_SB	10
early_january_weather	11
evals	12
ev_charging	13
geom_categorical_model	14
geom_parallel_slopes	16
get_correlation	18
get_regression_points	19
get_regression_summaries	20
get_regression_table	21
gg_parallel_slopes	22
house_prices	23
ipf_lifts	24
mario_kart_auction	25
mass_traffic_2020	26
MA_schools	27
ma_traffic_2020_vs_2019	28
moderndive	28
movies_sample	29
mythbusters_yawn	30
orig_pennies_sample	30
pennies	31
pennies_resamples	31
pennies_sample	32
promotions	32
promotions_shuffled	33
saratoga_houses	33
tactile_prop_red	34
Index	35

alaska_flights	<i>Alaska flights data</i>
----------------	----------------------------

Description

On-time data for all Alaska Airlines flights that departed NYC (i.e. JFK, LGA or EWR) in 2013. This is a subset of the `flights` data frame from `nycflights13`.

Usage

```
alaska_flights
```

Format

A data frame of 714 rows representing Alaska Airlines flights and 19 variables

year, month, day Date of departure.

dep_time, arr_time Actual departure and arrival times (format HHMM or HMM), local tz.

sched_dep_time, sched_arr_time Scheduled departure and arrival times (format HHMM or HMM), local tz.

dep_delay, arr_delay Departure and arrival delays, in minutes. Negative times represent early departures/arrivals.

carrier Two letter carrier abbreviation. See [nycflights13::airlines](#) to get name.

flight Flight number.

tailnum Plane tail number. See [nycflights13::planes](#) for additional metadata.

origin, dest Origin and destination. See [nycflights13::airports](#) for additional metadata.

air_time Amount of time spent in the air, in minutes.

distance Distance between airports, in miles.

hour, minute Time of scheduled departure broken into hour and minutes.

time_hour Scheduled date and hour of the flight as a POSIXct date. Along with origin, can be used to join flights data to [nycflights13::weather](#) data.

Source

RITA, Bureau of transportation statistics

See Also

[nycflights13::flights](#).

amazon_books

Sample of Amazon books

Description

A random sample of 325 books from Amazon.com.

Usage

amazon_books

Format

A data frame of 325 rows representing books listed on Amazon and 13 variables.

title Book title

author Author who wrote book

list_price recommended retail price of book

amazon_price lowest price of book shown on Amazon

hard_paper book is either hardcover or paperback

num_pages number of pages in book

publisher Company that issues the book for sale

pub_year Year the book was published

isbn_10 10-character ISBN number

height, width, thick, weight_oz height, width, weight and thickness of the book

Source

The Data and Story Library (DASL) <https://dasl.datadescription.com/datafile/amazon-books>

avocados

Avocado Prices by US Region

Description

Gathered from <https://docs.google.com/spreadsheets/d/1cNuj9V-9Xe8fqV3DQRhvsXJhER3zTk01dSsQ1Q0j96g/edit#gid=1419070688>

Usage

avocados

Format

A data frame of 54 regions over 3 years of weekly results

date Week of Data Recording

average_price Average Price of Avocado

total_volume Total Amount of Avocados

small_hass_sold Amount of Small Haas Avocados Sold

large_hass_sold Amount of Large Haas Avocados Sold

xlarge_hass_sold Amount of Extra Large Haas Avocados Sold

total_bags Total Amount of Bags of Avocados

small_bags Total Amount of Bags of Small Haas Avocados

large_bags Total Amount of Bags of Large Haas Avocados
x_large_bags Total Amount of Bags of Extra Large Haas Avocados
type Type of Sale
year Year of Sale
region Region Where Sale Took Place

babies

*Data on maternal smoking and infant health***Description**

Data on maternal smoking and infant health

Usage

babies

Format

A data frame of 1236 rows of individual mothers.

id Identification number**plurality** Marked 5 for single fetus, otherwise number of fetuses**outcome** Marked 1 for live birth that survived at least 28 days**date** Birth date where 1096 is January 1st, 1961**birthday** Birth date in mm-dd-yyyy format**gestation** Length of gestation in days, marked 999 if unknown**sex** Infant's sex, where 1 is male, 2 is female, and 9 is unknown**wt** Birth weight in ounces, marked 999 if unknown**parity** Total number of previous pregnancies including fetal deaths and stillbirths, marked 99 if unknown**race** Mother's race where 0-5 is white, 6 is Mexican, 7 is Black, 8 is Asian, 9 is mixed, and 99 is unknown**age** Mother's age in years at termination of pregnancy, 99=unknown**ed** Mother's education 0= less than 8th grade, 1 = 8th -12th grade - did not graduate, 2= HS graduate--no other schooling , 3= HS+trade, 4=HS+some college 5= College graduate, 6&7 Trade school HS unclear, 9=unknown**ht** Mother's height in inches to the last completed inch, 99=unknown**wt_1** Mother prepregnancy wt in pounds, 999=unknown**drace** Father's race, coding same as mother's race**dage** Father's age, coding same as mother's age

- ded** Father's education, coding same as mother's education
- dht** Father's height, coding same as for mother's height
- dwt** Father's weight coding same as for mother's weight
- marital** 0= legally separated, 1=married, 2= divorced, 3=widowed, 5=never married
- inc** Family yearly income in \$2500 increments 0 = under 2500, 1=2500-4999, ..., 8= 12,500-14,999, 9=15000+, 98=unknown, 99=not asked
- smoke** Does mother smoke? 0=never, 1= smokes now, 2=until current pregnancy, 3=once did, not now, 9=unknown
- time** If mother quit, how long ago? 0=never smoked, 1=still smokes, 2=during current preg, 3=within 1 yr, 4= 1 to 2 years ago, 5= 2 to 3 yr ago, 6= 3 to 4 yrs ago, 7=5 to 9yrs ago, 8=10+yrs ago, 9=quit and don't know, 98=unknown, 99=not asked
- number** Number of cigs smoked per day for past and current smokers 0=never, 1=1-4, 2=5-9, 3=10-14, 4=15-19, 5=20-29, 6=30-39, 7=40-60, 8=60+, 9=smoke but don't know, 98=unknown, 99=not asked

Source

Data on maternal smoking and infant health from <https://www.stat.berkeley.edu/~statlabs/labs.html>

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

bowl_samples	<i>Sampling from a bowl of balls</i>
--------------	--------------------------------------

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

See Also

[bowl\(\)](#)

bowl_sample_1	<i>Tactile sample of size 50 from a bowl of balls</i>
---------------	---

Description

A single tactile sample of size $n = 50$ balls from https://github.com/moderndive/moderndive/blob/master/data-raw/sampling_bowl.jpeg

Usage

```
bowl_sample_1
```

Format

A data frame of 50 rows representing different balls and 1 variable.

color Color of ball sampled

See Also

[bowl\(\)](#)

coffee_ratings *Data from the Coffee Quality Institute's review pages in January 2018*

Description

1,340 digitized reviews on coffee samples from <https://database.coffeeinstitute.org/>.

Usage

coffee_ratings

Format

A data frame of 1,340 rows representing each sample of coffee.

total_cup_points Number of points in final rating (scale of 0-100)

species Species of coffee bean plant (Arabica or Robusta)

owner Owner of coffee plant farm

country_of_origin Coffee bean's country of origin

farm_name Name of coffee plant farm

lot_number Lot number for tested coffee beans

mill Name of coffee bean's processing facility

ico_number International Coffee Organization number

company Name of coffee bean's company

altitude Altitude at which coffee plants were grown

region Region where coffee plants were grown

producer Name of coffee bean roaster

number_of_bags Number of tested bags

bag_weight Tested bag weight

in_country_partner Partner for the country

harvest_year Year the coffee beans were harvested

grading_date Day the coffee beans were graded

owner_1 Owner of the coffee beans

variety Variety of the coffee beans

processing_method Method used for processing the coffee beans

aroma Coffee aroma rating

flavor Coffee flavor rating

aftertaste Coffee aftertaste rating

acidity Coffee acidity rating

body Coffee body rating

balance Coffee balance rating
uniformity Coffee uniformity rating
clean_cup Cup cleanliness rating
sweetness Coffee sweetness rating
cupper_points Cupper Points, an overall rating for the coffee
moisture Coffee moisture content
category_one_defects Number of category one defects for the coffee beans
quakers Number of coffee beans that don't dark brown when roasted
color Color of the coffee beans
category_two_defects Number of category two defects for the coffee beans
expiration Expiration date of the coffee beans
certification_body Entity/Institute that certified the coffee beans
certification_address Body address of certification for coffee beans
certification_contact Certification contact for coffee beans
unit_of_measurement Unit of measurement for altitude
altitude_low_meters Lower altitude level coffee beans grow at
altitude_high_meters Higher altitude level coffee beans grow at
altitude_mean_meters Average altitude level coffee beans grow at

Source

Coffee Quality Institute. Access cleaned data available at <https://github.com/jldbc/coffee-quality-database>

DD_vs_SB

Dunkin Donuts vs Starbucks

Description

Number of Dunkin Donuts & Starbucks, median income, and population in 1024 census tracts in eastern Massachusetts in 2016.

Usage

DD_vs_SB

Format

A data frame of 1024 rows representing census tracts and 6 variables

county County where census tract is located. Either Bristol, Essex, Middlesex, Norfolk, Plymouth, or Suffolk county

FIPS Federal Information Processing Standards code identifying census tract

median_income Median income of census tract

population Population of census tract

shop_type Coffee shop type: Dunkin Donuts or Starbucks

shops Number of shops

Source

US Census Bureau. Code used to scrape data available at <https://github.com/DelaneyMoran/FinalProject>

early_january_weather *Early January hourly weather data*

Description

Hourly meteorological data for LGA, JFK and EWR for the month of January 2013. This is a subset of the weather data frame from `nycflights13`.

Usage

```
early_january_weather
```

Format

A data frame of 358 rows representing hourly measurements and 15 variables

origin Weather station. Named origin to facilitate merging with `nycflights13::flights` data.

year, month, day, hour Time of recording.

temp, dewp Temperature and dewpoint in F.

humid Relative humidity.

wind_dir, wind_speed, wind_gust Wind direction (in degrees), speed and gust speed (in mph).

precip Precipitation, in inches.

pressure Sea level pressure in millibars.

visib Visibility in miles.

time_hour Date and hour of the recording as a POSIXct date.

Source

ASOS download from Iowa Environmental Mesonet, <https://mesonet.agron.iastate.edu/request/download.phtml>.

See Also

[nycflights13::weather](#).

evals

Teaching evaluations at the UT Austin

Description

The data are gathered from end of semester student evaluations for a sample of 463 courses taught by 94 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/data/index.php?data=evals>

Usage

evals

Format

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

ID Identification variable for course.

prof_ID Identification variable for professor. Many professors are included more than once in this dataset.

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 (collected as a binary variable at the time of the study): 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).

See Also

The data in `evals` is a slight modification of `openintro::evals()`.

ev_charging

*Electric vehicle charging sessions for a workplace charging program***Description**

This dataset consists of information on 3,395 electric vehicle charging sessions across locations for a workplace charging program. The data contains information on multiple charging sessions from 85 electric vehicle drivers across 25 workplace locations, which are located at facilities of various types.

Usage

ev_charging

Format

A data frame of 3,395 rows on 24 variables, where each row is an electric vehicle charging session.

session_id Unique identifier specifying the electric vehicle charging session

kwh_total Total energy used at the charging session, in kilowatt hours (kWh)

dollars Quantity of money paid for the charging session in U.S. dollars

created Date and time recorded at the beginning of the charging session

ended Date and time recorded at the end of the charging session

start_time Hour of the day when the charging session began (1 through 24)

end_time Hour of the day when the charging session ended (1 through 24)

charge_time_hrs Length of the charging session in hours

weekday First three characters of the name of the weekday when the charging session occurred

platform Digital platform the driver used to record the session (android, ios, web)

distance Distance from the charging location to the driver's home, expressed in miles NA if the driver did not report their address

user_id Unique identifier for each driver

station_id Unique identifier for each charging station

location_id Unique identifier for each location owned by the company where charging stations were located

manager_vehicle Binary variable that is 1 when the vehicle is a type commonly used by managers of the firm and 0 otherwise

facility_type Categorical variable that represents the facility type:

- 1 = manufacturing
- 2 = office
- 3 = research and development
- 4 = other

mon, tues, wed, thurs, fri, sat, sun Binary variables; 1 if the charging session took place on that day, 0 otherwise

reported_zip Binary variable; 1 if the driver did report their zip code, 0 if they did not

Source

Harvard Dataverse doi: [10.7910/DVN/NFPQLW](https://doi.org/10.7910/DVN/NFPQLW). Note data is released under a CC0: Public Domain license.

geom_categorical_model

Regression model with one categorical explanatory/predictor variable

Description

`geom_categorical_model()` fits a regression model using the categorical x axis as the explanatory variable, and visualizes the model's fitted values as piecewise horizontal line segments. Confidence interval bands can be included in the visualization of the model. Like `geom_parallel_slopes()`, this function has the same nature as `geom_smooth()` from the `ggplot2` package, but provides functionality that `geom_smooth()` currently doesn't have. When using a categorical predictor variable, the intercept corresponds to the mean for the baseline group, while coefficients for the non-baseline groups are offsets from this baseline. Thus in the visualization the baseline for comparison group's median is marked with a solid line, whereas all offset groups' medians are marked with dashed lines.

Usage

```
geom_categorical_model(
  mapping = NULL,
  data = NULL,
  position = "identity",
  ...,
  se = TRUE,
  level = 0.95,
  na.rm = FALSE,
  show.legend = NA,
  inherit.aes = TRUE
)
```

Arguments

mapping	Set of aesthetic mappings created by <code>aes()</code> or <code>aes_()</code> . If specified and <code>inherit.aes = TRUE</code> (the default), it is combined with the default mapping at the top level of the plot. You must supply mapping if there is no plot mapping.
data	The data to be displayed in this layer. There are three options: If <code>NULL</code> , the default, the data is inherited from the plot data as specified in the call to <code>ggplot()</code> . A <code>data.frame</code> , or other object, will override the plot data. All objects will be fortified to produce a data frame. See <code>fortify()</code> for which variables will be created.

	A function will be called with a single argument, the plot data. The return value must be a <code>data.frame</code> , and will be used as the layer data. A function can be created from a formula (e.g. <code>~ head(.x, 10)</code>).
<code>position</code>	Position adjustment, either as a string, or the result of a call to a position adjustment function.
<code>...</code>	Other arguments passed on to <code>layer()</code> . These are often aesthetics, used to set an aesthetic to a fixed value, like <code>colour = "red"</code> or <code>size = 3</code> . They may also be parameters to the paired geom/stat.
<code>se</code>	Display confidence interval around model lines? TRUE by default.
<code>level</code>	Level of confidence interval to use (0.95 by default).
<code>na.rm</code>	If FALSE, the default, missing values are removed with a warning. If TRUE, missing values are silently removed.
<code>show.legend</code>	logical. Should this layer be included in the legends? NA, the default, includes if any aesthetics are mapped. FALSE never includes, and TRUE always includes. It can also be a named logical vector to finely select the aesthetics to display.
<code>inherit.aes</code>	If FALSE, overrides the default aesthetics, rather than combining with them. This is most useful for helper functions that define both data and aesthetics and shouldn't inherit behaviour from the default plot specification, e.g. <code>borders()</code> .

See Also

[geom_parallel_slopes\(\)](#)

Examples

```
library(dplyr)
library(ggplot2)

p <- ggplot(mpg, aes(x = drv, y = hwy)) +
  geom_point() +
  geom_categorical_model()
p

# In the above visualization, the solid line corresponds to the mean of 19.2
# for the baseline group "4", whereas the dashed lines correspond to the
# means of 28.19 and 21.02 for the non-baseline groups "f" and "r" respectively.
# In the corresponding regression table however the coefficients for "f" and "r"
# are presented as offsets from the mean for "4":
model <- lm(hwy ~ drv, data = mpg)
get_regression_table(model)

# You can use different colors for each categorical level
p %+% aes(color = drv)

# But mapping the color aesthetic doesn't change the model that is fit
p %+% aes(color = class)
```

geom_parallel_slopes *Parallel slopes regression model*

Description

geom_parallel_slopes() fits parallel slopes model and adds its line output(s) to a ggplot object. Basically, it fits a unified model with intercepts varying between groups (which should be supplied as standard ggplot2 grouping aesthetics: group, color, fill, etc.). This function has the same nature as geom_smooth() from ggplot2 package, but provides functionality that geom_smooth() currently doesn't have.

Usage

```
geom_parallel_slopes(  
  mapping = NULL,  
  data = NULL,  
  position = "identity",  
  ...,  
  se = TRUE,  
  formula = y ~ x,  
  n = 100,  
  fullrange = FALSE,  
  level = 0.95,  
  na.rm = FALSE,  
  show.legend = NA,  
  inherit.aes = TRUE  
)
```

Arguments

mapping	Set of aesthetic mappings created by <code>aes()</code> or <code>aes_()</code> . If specified and <code>inherit.aes = TRUE</code> (the default), it is combined with the default mapping at the top level of the plot. You must supply mapping if there is no plot mapping.
data	The data to be displayed in this layer. There are three options: If NULL, the default, the data is inherited from the plot data as specified in the call to <code>ggplot()</code> . A <code>data.frame</code> , or other object, will override the plot data. All objects will be fortified to produce a data frame. See <code>fortify()</code> for which variables will be created. A function will be called with a single argument, the plot data. The return value must be a <code>data.frame</code> , and will be used as the layer data. A function can be created from a formula (e.g. <code>~ head(.x, 10)</code>).
position	Position adjustment, either as a string, or the result of a call to a position adjustment function.

...	Other arguments passed on to <code>layer()</code> . These are often aesthetics, used to set an aesthetic to a fixed value, like <code>colour = "red"</code> or <code>size = 3</code> . They may also be parameters to the paired geom/stat.
<code>se</code>	Display confidence interval around model lines? TRUE by default.
<code>formula</code>	Formula to use per group in parallel slopes model. Basic linear $y \sim x$ by default.
<code>n</code>	Number of points per group at which to evaluate model.
<code>fullrange</code>	Should the fit span the full range of the plot, or just the data?
<code>level</code>	Level of confidence interval to use (0.95 by default).
<code>na.rm</code>	If FALSE, the default, missing values are removed with a warning. If TRUE, missing values are silently removed.
<code>show.legend</code>	logical. Should this layer be included in the legends? NA, the default, includes if any aesthetics are mapped. FALSE never includes, and TRUE always includes. It can also be a named logical vector to finely select the aesthetics to display.
<code>inherit.aes</code>	If FALSE, overrides the default aesthetics, rather than combining with them. This is most useful for helper functions that define both data and aesthetics and shouldn't inherit behaviour from the default plot specification, e.g. <code>borders()</code> .

See Also

[geom_categorical_model\(\)](#)

Examples

```
library(dplyr)
library(ggplot2)

ggplot(evals, aes(x = age, y = score, color = ethnicity)) +
  geom_point() +
  geom_parallel_slopes(se = FALSE)

# Basic usage
ggplot(evals, aes(x = age, y = score, color = ethnicity)) +
  geom_point() +
  geom_parallel_slopes()
ggplot(evals, aes(x = age, y = score, color = ethnicity)) +
  geom_point() +
  geom_parallel_slopes(se = FALSE)

# Supply custom aesthetics
ggplot(evals, aes(x = age, y = score, color = ethnicity)) +
  geom_point() +
  geom_parallel_slopes(se = FALSE, size = 4)

# Fit non-linear model
example_df <- house_prices %>%
  slice(1:1000) %>%
  mutate(
    log10_price = log10(price),
    log10_size = log10(sqft_living)
```

```

)
ggplot(example_df, aes(x = log10_size, y = log10_price, color = condition)) +
  geom_point(alpha = 0.1) +
  geom_parallel_slopes(formula = y ~ poly(x, 2))

# Different grouping
ggplot(example_df, aes(x = log10_size, y = log10_price)) +
  geom_point(alpha = 0.1) +
  geom_parallel_slopes(aes(fill = condition))

```

get_correlation	<i>Get correlation value in a tidy way</i>
-----------------	--

Description

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

Usage

```
get_correlation(data, formula, na.rm = FALSE, ...)
```

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
na.rm	a logical value indicating whether NA values should be stripped before the computation proceeds.
...	further arguments passed to <code>stats::cor()</code>

Value

A 1x1 data frame storing the correlation value

Examples

```

library(moderndiver)

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

# Group by one variable:
library(dplyr)
mtcars %>%
  group_by(am) %>%
  get_correlation(formula = mpg ~ cyl)

```

```
# Group by two variables:
mtcars %>%
  group_by(am, gear) %>%
  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,
  ID = NULL
)
```

Arguments

<code>model</code>	an <code>lm()</code> model object
<code>digits</code>	number of digits precision in output table
<code>print</code>	If TRUE, return in print format suitable for R Markdown
<code>newdata</code>	A new data frame of points/observations to apply <code>model</code> to obtain new fitted values and/or predicted values \hat{y} . Note the format of <code>newdata</code> must match the format of the original data used to fit <code>model</code> .
<code>ID</code>	A string indicating which variable in either the original data used to fit <code>model</code> or <code>newdata</code> should be used as an identification variable to distinguish the observational units in each row. This variable will be the left-most variable in the output data frame. If <code>ID</code> is unspecified, a column <code>ID</code> with values 1 through the number of rows is returned as the identification variable.

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(dplyr)
library(tibble)

# Convert rownames to column
mtcars <- mtcars %>%
  rownames_to_column(var = "automobile")

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

# Get information on all points in regression:
get_regression_points(mpg_model, ID = "automobile")

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

# 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, ID = "automobile")
```

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 <code>lm()</code> 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(moderndiver)

# 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,
  conf.level = 0.95,
  digits = 3,
  print = FALSE,
  default_categorical_levels = FALSE
)
```

Arguments

<code>model</code>	an <code>lm()</code> model object
<code>conf.level</code>	The confidence level to use for the confidence interval if <code>conf.int = TRUE</code> . Must be strictly greater than 0 and less than 1. Defaults to 0.95, which corresponds to a 95 percent confidence interval.
<code>digits</code>	number of digits precision in output table
<code>print</code>	If <code>TRUE</code> , return in print format suitable for R Markdown
<code>default_categorical_levels</code>	If <code>TRUE</code> , do not change the non-baseline categorical variables in the term column. Otherwise non-baseline categorical variables will be displayed in the format "categorical_variable_name: level_name"

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`; the confidence levels default to 95%

See Also

[tidy\(\)](#), [get_regression_points\(\)](#), [get_regression_summaries\(\)](#)

Examples

```
library(moderndiver)

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

# Get regression table:
get_regression_table(mpg_model)

# Vary confidence level of confidence intervals
get_regression_table(mpg_model, conf.level = 0.99)
```

gg_parallel_slopes *Plot parallel slopes model*

Description

NOTE: This function is deprecated; please use [geom_parallel_slopes\(\)](#) instead. Output a visualization of linear regression when you have one numerical and one categorical explanatory/predictor variable: a separate colored regression line for each level of the categorical variable

Usage

```
gg_parallel_slopes(y, num_x, cat_x, data, alpha = 1)
```

Arguments

y	Character string of outcome variable in data
num_x	Character string of numerical explanatory/predictor variable in data
cat_x	Character string of categorical explanatory/predictor variable in data
data	an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in data, the variables are taken from environment (formula), typically the environment from which <code>lm</code> is called.
alpha	Transparency of points

Value

A `ggplot2::ggplot()` object.

See Also

[geom_parallel_slopes\(\)](#)

Examples

```
## Not run:
library(ggplot2)
library(dplyr)
library(moderndiver)

# log10() transformations
house_prices <- house_prices %>%
  mutate(
    log10_price = log10(price),
    log10_size = log10(sqft_living)
  )

# Output parallel slopes model plot:
gg_parallel_slopes(
  y = "log10_price", num_x = "log10_size", cat_x = "condition",
  data = house_prices, alpha = 0.1
) +
  labs(
    x = "log10 square feet living space", y = "log10 price in USD",
    title = "House prices in Seattle: Parallel slopes model"
  )

# Compare with interaction model plot:
ggplot(house_prices, aes(x = log10_size, y = log10_price, col = condition)) +
  geom_point(alpha = 0.1) +
  geom_smooth(method = "lm", se = FALSE, size = 1) +
  labs(
    x = "log10 square feet living space", y = "log10 price in USD",
    title = "House prices in Seattle: Interaction model"
  )

## End(Not run)
```

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.

ipf_lifts

International Power Lifting Results A subset of international powerlifting results.

Description

International Power Lifting Results A subset of international powerlifting results.

Usage

ipf_lifts

Format

A data frame with 41,152 entries, one entry for individual lifter

name Individual lifter name

sex Binary sex (M/F)

event The type of competition that the lifter entered

equipment The equipment category under which the lifts were performed

age The age of the lifter on the start date of the meet

age_class The age class in which the lifter falls

division division of competition

bodyweight_kg The recorded bodyweight of the lifter at the time of competition, to two decimal places

weight_class_kg The weight class in which the lifter competed, to two decimal places

best3squat_kg Maximum of the first three successful attempts for the lift

best3bench_kg Maximum of the first three successful attempts for the lift

best3deadlift_kg Maximum of the first three successful attempts for the lift

place The recorded place of the lifter in the given division at the end of the meet

date Date of the event

federation The federation that hosted the meet

meet_name The name of the meet

Source

This data is a subset of the open dataset [Open Powerlifting](#)

mario_kart_auction *Data from Mario Kart Ebay auctions*

Description

Ebay auction data for the Nintendo Wii game Mario Kart.

Usage

mario_kart_auction

Format

A data frame of 143 auctions.

id Auction ID assigned by Ebay

duration Auction length in days

n_bids Number of bids

cond Game condition, either new or used

start_pr Price at the start of the auction

ship_pr Shipping price

total_pr Total price, equal to auction price plus shipping price

ship_sp Shipping speed or method

seller_rate Seller's rating on Ebay, equal to the number of positive ratings minus the number of negative ratings

stock_photo Whether the auction photo was a stock photo or not, pictures used in many options were considered stock photos

wheels Number of Wii wheels included in the auction

title The title of the auctions

Source

This data is from <https://www.openintro.org/data/index.php?data=mariokart>

mass_traffic_2020	<i>2020 road traffic volume and crash level date for 13 Massachusetts counties</i>
-------------------	--

Description

2020 road traffic volume and crash level date for 13 Massachusetts counties

Usage

mass_traffic_2020

Format

A data frame of 874 rows representing traffic data at the 874 sites

site_id Site id

county County in which the site is located

community Community in which the site is located

rural_urban Rural (R) or Urban (U)

- dir** Direction for traffic movement. Either 1-WAY, 2-WAY, EB (eastbound), RAMP or WB (westbound)
- functional_class** Classification of road. Either Arterial, Collector, Freeway & Expressway, Interstate or Local Road
- avg_speed** Average traffic speed
- total_volume** Number of vehicles recorded at each site in 2020
- crashes** Number of vehicle crashes at each site
- nonfatal_injuries** Number of non-fatal injuries for all recorded vehicle crashes
- fatal_injuries** Number of fatal injuries for all recorded vehicle crashes

 MA_schools

Massachusetts Public High Schools Data

Description

Data on Massachusetts public high schools in 2017

Usage

MA_schools

Format

A data frame of 332 rows representing Massachusetts high schools and 4 variables

school_name High school name.

average_sat_math Average SAT math score. Note 58 of the original 390 values of this variable were missing; these rows were dropped from consideration.

perc_disadvan Percent of the student body that are considered economically disadvantaged.

size Size of school enrollment; small 13-341 students, medium 342-541 students, large 542-4264 students.

Source

The original source of the data are Massachusetts Department of Education reports https://profiles.doe.mass.edu/state_report/, however the data was downloaded from Kaggle at <https://www.kaggle.com/ndalziel/massachusetts-public-schools-data>

ma_traffic_2020_vs_2019

Massachusetts 2020 vs. 2019 Traffic Data Comparison

Description

This dataset contains information about changes in speed, volume, and accidents of traffic between 2020 and 2019 by community and class of road in Massachusetts.

Usage

ma_traffic_2020_vs_2019

Format

A data frame of 264 rows each representing a different community in Massachusetts.

community City or Town

functional_class Class or group the road belongs to

change_in_speed Average estimated Speed (mph)

change_in_volume Average traffic

change_in_accidents Average number of accidents

Source

<https://massdot-impact-crashes-vhb.opendata.arcgis.com/datasets/MassDOT::2020-vehicle-level-crash-collisions>
explore <https://mhd.public.ms2soft.com/tcds/tsearch.asp?loc=Mhd&mod=>

moderndive

moderndive - Tidyverse-Friendly Introductory Linear Regression

Description

Datasets and wrapper functions for tidyverse-friendly introductory linear regression, used in "Statistical Inference via Data Science: A ModernDive into R and the tidyverse" available at <https://moderndive.com/>.

Examples

```
library(moderndiver)

# 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)

# Plotting parallel slopes models
library(ggplot2)
ggplot(evals, aes(x = age, y = score, color = ethnicity)) +
  geom_point() +
  geom_parallel_slopes(se = FALSE)
```

movies_sample

Random sample of 68 action and romance movies

Description

A random sample of 32 action movies and 36 romance movies from <https://www.imdb.com/> and their ratings.

Usage

```
movies_sample
```

Format

A data frame of 68 rows movies.

title Movie title

year Year released

rating IMDb rating out of 10 stars

genre Action or Romance

See Also

This data was sampled from the `movies` data frame in the `ggplot2movies` package.

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

orig_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

orig_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:443/app/index.html?dataid=301596>

See Also

`pennies()`

pennies	<i>A population of 800 pennies sampled in 2011</i>
---------	--

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:443/app/index.html?dataid=301596>

pennies_resamples	<i>Bootstrap resamples of a sample of 50 pennies</i>
-------------------	--

Description

35 bootstrap resamples with replacement of sample of 50 pennies contained in a 50 cent roll from Florence Bank on Friday February 1, 2019 in downtown Northampton, Massachusetts, USA <https://goo.gl/maps/AF88fpvVfm12>. The original sample of 50 pennies is available in [pennies_sample\(\)](#).

Usage

```
pennies_resamples
```

Format

A data frame of 1750 rows representing 35 students' bootstrap resamples of size 50 and 3 variables

replicate ID variable of replicate/resample number.

name Name of student

year Year on resampled penny

See Also

[pennies_sample\(\)](#)

pennies_sample	<i>A sample of 50 pennies</i>
----------------	-------------------------------

Description

A sample of 50 pennies contained in a 50 cent roll from Florence Bank on Friday February 1, 2019 in downtown Northampton, Massachusetts, USA <https://goo.gl/maps/AF88fpvVfm12>.

Usage

pennies_sample

Format

A data frame of 50 rows representing 50 sampled pennies and 2 variables

ID Variable used to uniquely identify each penny.

year Year of minting.

Note

The original pennies_sample has been renamed [orig_pennies_sample\(\)](#) as of moderndive v0.3.0.

promotions	<i>Bank manager recommendations based on (binary) gender</i>
------------	--

Description

Data from a 1970's study on whether gender influences hiring recommendations. Originally used in OpenIntro.org.

Usage

promotions

Format

A data frame with 48 observations on the following 3 variables.

id Identification variable used to distinguish rows.

gender gender (collected as a binary variable at the time of the study): a factor with two levels male and female

decision a factor with two levels: promoted and not

Source

Rosen B and Jerdee T. 1974. Influence of sex role stereotypes on personnel decisions. *Journal of Applied Psychology* 59(1):9-14.

See Also

The data in `promotions` is a slight modification of `openintro::gender_discrimination()`.

`promotions_shuffled` *One permutation/shuffle of promotions*

Description

Shuffled/permutated data from a 1970's study on whether gender influences hiring recommendations.

Usage

```
promotions_shuffled
```

Format

A data frame with 48 observations on the following 3 variables.

id Identification variable used to distinguish rows.

gender shuffled/permutated (binary) gender: a factor with two levels male and female

decision a factor with two levels: promoted and not

See Also

`promotions()`.

`saratoga_houses` *House Prices and Properties in Saratoga, New York*

Description

Random sample of 1057 houses taken from full Saratoga Housing Data (De Veaux)

Usage

```
saratoga_houses
```

Format

A data frame with 1057 observations on the following 8 variables

price price (US dollars)

living_area Living Area (square feet)

bathrooms Number of Bathroom (half bathrooms have no shower or tub)

bedrooms Number of Bedrooms

fireplaces Number of Fireplaces

lot_size Size of Lot (acres)

age Age of House (years)

fireplace Whether the house has a Fireplace

Source

Gathered from <https://docs.google.com/spreadsheets/d/1AY5eECqNIggKpYF3kYzJQBIuuOdkic1FhbjAmY3Yc8E/edit#gid=622599674>

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

See Also

[bowl\(\)](#)

Index

* datasets

alaska_flights, 3
amazon_books, 4
avocados, 5
babies, 6
bowl, 7
bowl_sample_1, 8
bowl_samples, 8
coffee_ratings, 9
DD_vs_SB, 10
early_january_weather, 11
ev_charging, 13
evals, 12
house_prices, 23
ipf_lifts, 24
MA_schools, 27
ma_traffic_2020_vs_2019, 28
mario_kart_auction, 25
mass_traffic_2020, 26
movies_sample, 29
mythbusters_yawn, 30
orig_pennies_sample, 30
pennies, 31
pennies_resamples, 31
pennies_sample, 32
promotions, 32
promotions_shuffled, 33
saratoga_houses, 33
tactile_prop_red, 34

aes(), 14, 16
aes_(), 14, 16
alaska_flights, 3
amazon_books, 4
as.data.frame, 22
augment(), 19
avocados, 5

babies, 6
borders(), 15, 17

bowl, 7
bowl(), 8, 34
bowl_sample_1, 8
bowl_samples, 8

coffee_ratings, 9

DD_vs_SB, 10

early_january_weather, 11
ev_charging, 13
evals, 12

fortify(), 14, 16

geom_categorical_model, 14
geom_categorical_model(), 17
geom_parallel_slopes, 16
geom_parallel_slopes(), 14, 15, 22
get_correlation, 18
get_regression_points, 19
get_regression_points(), 21, 22
get_regression_summaries, 20
get_regression_summaries(), 19, 22
get_regression_table, 21
get_regression_table(), 19, 21
gg_parallel_slopes, 22
ggplot(), 14, 16
ggplot2::ggplot(), 22
glance(), 21

house_prices, 23

ipf_lifts, 24

layer(), 15, 17

MA_schools, 27
ma_traffic_2020_vs_2019, 28
mario_kart_auction, 25
mass_traffic_2020, 26

moderndive, 28
movies_sample, 29
mythbusters_yawn, 30

nycflights13::airlines, 4
nycflights13::airports, 4
nycflights13::flights, 4, 11
nycflights13::planes, 4
nycflights13::weather, 4, 11

openintro::evals(), 12
openintro::gender_discrimination(), 33
orig_pennies_sample, 30
orig_pennies_sample(), 32

pennies, 31
pennies(), 30
pennies_resamples, 31
pennies_sample, 32
pennies_sample(), 31
promotions, 32
promotions(), 33
promotions_shuffled, 33

saratoga_houses, 33
stats::cor(), 18

tactile_prop_red, 34
tidy(), 22