

Package: quartets (via r-universe)

September 7, 2024

Type Package

Title Datasets to Help Teach Statistics

Version 0.1.3

Description In the spirit of Anscombe's quartet, this package includes datasets that demonstrate the importance of visualizing your data, the importance of not relying on statistical summary measures alone, and why additional assumptions about the data generating mechanism are needed when estimating causal effects. The package includes ``Anscombe's Quartet'' (Anscombe 1973) <doi:10.1080/00031305.1973.10478966>, D'Agostino McGowan & Barrett (2023) ``Causal Quartet'' <doi:10.1080/26939169.2023.2276446>, ``Datasaurus Dozen'' (Matejka & Fitzmaurice 2017), ``Interaction Triptych'' (Rohrer & Arslan 2021) <doi:10.1177/25152459211007368>, ``Rashomon Quartet'' (Biecek et al. 2023) <doi:10.48550/arXiv.2302.13356>, and Gelman ``Variation and Heterogeneity Causal Quartets'' (Gelman et al. 2023) <doi:10.48550/arXiv.2302.12878>.

License MIT + file LICENSE

URL <https://github.com/r-causal/quartets>,
<https://r-causal.github.io/quartets/>

BugReports <https://github.com/r-causal/quartets/issues>

Encoding UTF-8

LazyData true

Depends R (>= 2.10)

RoxygenNote 7.2.3

Roxygen list(markdown = TRUE)

Suggests knitr, rmarkdown, tidyverse

VignetteBuilder knitr

Repository <https://r-causal.r-universe.dev>

RemoteUrl <https://github.com/r-causal/quartets>

RemoteRef HEAD

RemoteSha b812b77f9d1bb7544aca5ea9e5ec0c9f1cd7bc07

Contents

anscombe_leverage	2
anscombe_linear	3
anscombe_nonlinear	4
anscombe_outlier	5
anscombe_quartet	6
causal_collider	7
causal_collider_time	7
causal_confounding	10
causal_mediator	10
causal_m_bias	11
causal_quartet	11
datasaurus_dozen	12
heterogeneous_causal_quartet	13
interaction_triptych	14
rashomon_quartet	15
variation_causal_quartet	16
Index	18

anscombe_leverage	<i>Anscombe's Quartet High Leverage Data</i>
-------------------	--

Description

This dataset contains 11 observations generated by Francis Anscombe to demonstrate that statistical summary measures alone cannot capture the full relationship between two variables (here, x and y). Anscombe emphasized the importance of visualizing data prior to calculating summary statistics.

Usage

anscombe_leverage

Format

A dataframe with 11 rows and 2 variables:

- x: the x-variable
- y: the y-variable

Details

This Dataset has a no relationship between x and y with a single high leverage point. Additionally, the following statistical summaries hold:

- mean of x: 9
- variance of x: 11

- mean of y: 7.5
- variance of y: 4.125
- correlation between x and y: 0.816
- linear regression between x and y: $y = 3 + 0.5x$
- R^2 for the regression: 0.67

References

Anscombe, F. J. (1973). "Graphs in Statistical Analysis". *American Statistician*. 27 (1): 17–21. doi:10.1080/00031305.1973.10478966. JSTOR 2682899.

anscombe_linear *Anscombe's Quartet Linear Data*

Description

This dataset contains 11 observations generated by Francis Anscombe to demonstrate that statistical summary measures alone cannot capture the full relationship between two variables (here, x and y). Anscombe emphasized the importance of visualizing data prior to calculating summary statistics.

Usage

anscombe_linear

Format

A dataframe with 11 rows and 2 variables:

- x: the x-variable
- y: the y-variable

Details

This Dataset has a linear relationship between x and y
Additionally, the following statistical summaries hold:

- mean of x: 9
- variance of x: 11
- mean of y: 7.5
- variance of y: 4.125
- correlation between x and y: 0.816
- linear regression between x and y: $y = 3 + 0.5x$
- R^2 for the regression: 0.67

References

Anscombe, F. J. (1973). "Graphs in Statistical Analysis". *American Statistician*. 27 (1): 17–21. doi:10.1080/00031305.1973.10478966. JSTOR 2682899.

anscombe_nonlinear *Anscombe's Quartet Nonlinear Data*

Description

This dataset contains 11 observations generated by Francis Anscombe to demonstrate that statistical summary measures alone cannot capture the full relationship between two variables (here, x and y). Anscombe emphasized the importance of visualizing data prior to calculating summary statistics.

Usage

anscombe_nonlinear

Format

A dataframe with 11 rows and 2 variables:

- x : the x -variable
- y : the y -variable

Details

This Dataset has a nonlinear relationship between x and y

Additionally, the following statistical summaries hold:

- mean of x : 9
- variance of x : 11
- mean of y : 7.5
- variance of y : 4.125
- correlation between x and y : 0.816
- linear regression between x and y : $y = 3 + 0.5x$
- R^2 for the regression: 0.67

References

Anscombe, F. J. (1973). "Graphs in Statistical Analysis". *American Statistician*. 27 (1): 17–21. doi:10.1080/00031305.1973.10478966. JSTOR 2682899.

anscombe_outlier *Anscombe's Quartet Outlier Data*

Description

This dataset contains 11 observations generated by Francis Anscombe to demonstrate that statistical summary measures alone cannot capture the full relationship between two variables (here, x and y). Anscombe emphasized the importance of visualizing data prior to calculating summary statistics.

Usage

anscombe_outlier

Format

A dataframe with 11 rows and 2 variables:

- x : the x -variable
- y : the y -variable

Details

This Dataset has a linear relationship between x and y with a single outlier

Additionally, the following statistical summaries hold:

- mean of x : 9
- variance of x : 11
- mean of y : 7.5
- variance of y : 4.125
- correlation between x and y : 0.816
- linear regression between x and y : $y = 3 + 0.5x$
- R^2 for the regression: 0.67

References

Anscombe, F. J. (1973). "Graphs in Statistical Analysis". *American Statistician*. 27 (1): 17–21. doi:10.1080/00031305.1973.10478966. JSTOR 2682899.

anscombe_quartet *Anscombe's Quartet Data*

Description

This dataset contains 44 observations, 11 observations from 4 datasets generated by Francis Anscombe to demonstrate that statistical summary measures alone cannot capture the full relationship between two variables (here, x and y). Anscombe emphasized the importance of visualizing data prior to calculating summary statistics.

Usage

anscombe_quartet

Format

A dataframe with 44 rows and 3 variables:

- dataset: the dataset the values come from
- x: the x-variable
- y: the y-variable

Details

- Dataset 1 has a linear relationship between x and y
- Dataset 2 has shows a nonlinear relationship between x and y
- Dataset 3 has a linear relationship between x and y with a single outlier
- Dataset 4 has shows no relationship between x and y with a single outlier that serves as a high-leverage point.

In each of the datasets the following statistical summaries hold:

- mean of x : 9
- variance of x : 11
- mean of y : 7.5
- variance of y : 4.125
- correlation between x and y : 0.816
- linear regression between x and y : $y = 3 + 0.5x$
- R^2 for the regression: 0.67

References

Anscombe, F. J. (1973). "Graphs in Statistical Analysis". *American Statistician*. 27 (1): 17–21. doi:10.1080/00031305.1973.10478966. JSTOR 2682899.

causal_collider	<i>Collider Data</i>
-----------------	----------------------

Description

This dataset contains 100 observations, generated under the following mechanism: $X \sim N(0, 1)$ (exposure) $Y \sim X + N(0,1)$ (outcome) $Z \sim 0.45X + 0.77Y + N(0, 1)$ (measured factor: collider)

Usage

causal_collider

Format

A dataframe with 100 rows and 3 variables:

- exposure: exposure
- outcome: outcome
- covariate: a known factor (collider)

References

Lucy D'Agostino McGowan, Travis Gerke & Malcolm Barrett (2023) Causal inference is not just a statistics problem, *Journal of Statistics and Data Science Education*, DOI: 10.1080/26939169.2023.2276446

causal_collider_time	<i>Time-varying Causal Quartet Data</i>
----------------------	---

Description

These datasets contains 100 observations, each generated under a different data generating mechanism:

- (1) A collider
- (2) A confounder
- (3) A mediator
- (4) M-bias

Usage

`causal_collider_time`

`causal_confounding_time`

`causal_mediator_time`

`causal_m_bias_time`

`causal_quartet_time`

Format

`causal_collider_time`: A dataframe with 100 rows and 7 variables:

- `covariate_baseline`: known factor measured at baseline
- `exposure_baseline`: exposure measured at baseline
- `outcome_baseline`: outcome measured at baseline
- `exposure_followup`: exposure measured at the followup visit (final time)
- `outcome_followup`: outcome measured at the followup visit (final time)
- `covariate_followup`: known factor measured at the followup visit (final time)

`causal_confounding_time`: A dataframe with 100 rows and 7 variables:

- `covariate_baseline`: known factor measured at baseline
- `exposure_baseline`: exposure measured at baseline
- `outcome_baseline`: outcome measured at baseline
- `exposure_followup`: exposure measured at the followup visit (final time)
- `outcome_followup`: outcome measured at the followup visit (final time)
- `covariate_followup`: known factor measured at the followup visit (final time)

`causal_mediator_time`: A dataframe with 100 rows and 7 variables:

- `covariate_baseline`: known factor measured at baseline
- `exposure_baseline`: exposure measured at baseline
- `outcome_baseline`: outcome measured at baseline
- `covariate_mid`: known factor measured at some mid-point
- `exposure_mid`: exposure measured at some mid-point
- `outcome_mid`: outcome measured at some mid-point
- `exposure_followup`: exposure measured at the followup visit (final time)
- `outcome_followup`: outcome measured at the followup visit (final time)
- `covariate_followup`: known factor measured at the followup visit (final time)

`causal_m_bias_time`: A dataframe with 100 rows and 9 variables:

- u1: unmeasured factor
- u2: unmeasured factor
- covariate_baseline: known factor measured at baseline
- exposure_baseline: exposure measured at baseline
- outcome_baseline: outcome measured at baseline
- exposure_followup: exposure measured at the followup visit (final time)
- outcome_followup: outcome measured at the followup visit (final time)
- covariate_followup: known factor measured at the followup visit (final time)

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 400 rows and 12 columns.

Details

There are two time points:

- baseline
- follow up

These datasets help demonstrate that a model that includes only pre-exposure covariates (that is, only adjusting for covariates measured at baseline), will be less prone to potential biases. Adjusting for only pre-exposure covariates "solves" the bias in datasets 1-3. It does not solve the data generated under the "M-bias" scenario, however this is more of a toy example, it has been shown many times that the assumptions needed for this M-bias to hold are often not ones we practically see in data analysis.

References

Lucy D'Agostino McGowan, Travis Gerke & Malcolm Barrett (2023) Causal inference is not just a statistics problem, *Journal of Statistics and Data Science Education*, DOI: 10.1080/26939169.2023.2276446

Examples

```
## incorrect model because covariate is post-treatment
lm(outcome_followup ~ exposure_baseline + covariate_followup,
  data = causal_collider_time)

## correct model because covariate is pre-treatment
## even though the true mechanism dictates that the covariate is a collider,
## because the pre-exposure variable is used, the collider bias does not
## occur.
lm(outcome_followup ~ exposure_baseline + covariate_baseline,
  data = causal_collider_time)
```

causal_confounding *Confounder Data*

Description

This dataset contains 100 observations, generated under the following mechanism: $Z \sim N(0, 1)$ (measured factor: confounder) $X \sim Z + N(0,1)$ (exposure) $Y \sim 0.5X + Z + N(0, 1)$ (outcome)

Usage

causal_confounding

Format

A dataframe with 100 rows and 3:

- covariate: a known factor (confounder)
- exposure: exposure
- outcome: outcome

References

Lucy D'Agostino McGowan, Travis Gerke & Malcolm Barrett (2023) Causal inference is not just a statistics problem, *Journal of Statistics and Data Science Education*, DOI: 10.1080/26939169.2023.2276446

causal_mediator *Mediator Data*

Description

This dataset contains 100 observations, generated under the following mechanism: $X \sim N(0, 1)$ (exposure) $Z \sim X + N(0,1)$ (measured factor: mediator) $Y \sim Z + N(0, 1)$ (outcome)

Usage

causal_mediator

Format

A dataframe with 100 rows and 3 variables:

- exposure: exposure
- covariate: a known factor (mediator)
- outcome: outcome

References

Lucy D'Agostino McGowan, Travis Gerke & Malcolm Barrett (2023) Causal inference is not just a statistics problem, Journal of Statistics and Data Science Education, DOI: 10.1080/26939169.2023.2276446

causal_m_bias

M-Bias Data

Description

This dataset contains 100 observations, generated under the following mechanism: $U1 \sim N(0, 1)$
 $U2 \sim N(0, 1)$ $Z \sim 8 U1 + U2 + N(0, 1)$ (measured factor) $X \sim U1 + N(0, 1)$ (exposure) $Y \sim X + U2 + N(0, 1)$ (outcome)

Usage

causal_m_bias

Format

A dataframe with 100 rows and 5 variables:

- u1: an unknown factor
- u2: an unknown factor
- covariate: a known factor
- exposure: exposure
- outcome: outcome

References

Lucy D'Agostino McGowan, Travis Gerke & Malcolm Barrett (2023) Causal inference is not just a statistics problem, Journal of Statistics and Data Science Education, DOI: 10.1080/26939169.2023.2276446

causal_quartet

Causal Quartet Data

Description

This dataset contains 400 observations, each generated under a different data generating mechanism:

- (1) A collider
- (2) A confounder
- (3) A mediator
- (4) M-bias

Usage

```
causal_quartet
```

Format

A dataframe with 400 rows and 6 variables:

- dataset: The data generating mechanism
- exposure: exposure
- outcome: outcome
- covariate: a known factor
- u1: an unknown factor
- u2: an unknown factor

References

Lucy D'Agostino McGowan, Travis Gerke & Malcolm Barrett (2023) Causal inference is not just a statistics problem, *Journal of Statistics and Data Science Education*, DOI: 10.1080/26939169.2023.2276446

datasaurus_dozen

Datasaurus Dozen Data

Description

A dataset containing 12 datasets that are equal in mean, variance, and Pearson's correlation but very different when visualized.

Usage

```
datasaurus_dozen
```

Format

A data frame with 1846 rows and 3 variables:

- dataset: the dataset the values come from
- x: the x-variable
- y: the y-variable

References

Davies R, Locke S, D'Agostino McGowan L (2022). *datasauRus: Datasets from the Datasaurus Dozen*. R package version 0.1.6, <https://CRAN.R-project.org/package=datasauRus>.

Matejka, J., & Fitzmaurice, G. (2017). Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. CHI 2017 Conference proceedings: ACM SIGCHI Conference on Human Factors in Computing Systems. Retrieved from <https://www.autodesk.com/research/publications/same-stats-different-graphs>

heterogeneous_causal_quartet

Gelman Heterogeneity Causal Quartet Data

Description

This dataset contains 88 observations, each generated under a different mechanism treatment heterogeneity with respect to some pre-exposure characteristic, z :

- (1) Linear interaction
- (2) No effect then steady increase
- (3) Plateau
- (4) Intermediate zone with large effects

Usage

heterogeneous_causal_quartet

Format

A dataframe with 88 rows and 5 variables:

- `dataset`: The data generating mechanism
- `exposure`: exposure
- `covariate`: a pre-exposure factor
- `outcome`: outcome
- `.causal_effect`: latent true causal effect

References

Gelman, A., Hullman, J., & Kennedy, L. (2023). Causal quartets: Different ways to attain the same average treatment effect. arXiv preprint arXiv:2302.12878.

Hullman J (2023). *causalQuartet: Create Causal Quartets for Interrogating Average Treatment Effects*. R package version 0.0.0.9000.

interaction_triptych *Interaction Triptych Data*

Description

This dataset contains 2,700 observations, generated under 3 different conditions

- (1) Ideal case
- (2) Floor effect, No latent interaction
- (3) Smaller correlation at larger slope

Usage

interaction_triptych

Format

A dataframe with 2700 rows and 5 variables:

- dataset: ideal, floor, or smaller correlation at larger slope
- moderator: a factor that potentially interacts with x, values: low, medium, or high
- x
- y

Details

In the ideal scenario, only the slopes differ by moderator level. In the "floor effect" scenario, there is an illusion of an interaction, even though only main effects were simulated. In the third scenario, the slopes increase with higher moderator values but the correlation decreases. Running only a linear model would not allow for appropriate differentiation between these effects.

In each case there is a potential moderator with "low" "medium" and "high" values.

References

Rohrer, Julia M., and Ruben C. Arslan. "Precise answers to vague questions: Issues with interactions." *Advances in Methods and Practices in Psychological Science* 4.2 (2021): 25152459211007368.

rashomon_quartet	<i>Rashomon Quartet Data</i>
------------------	------------------------------

Description

This dataset contains 2,000 observations, 1,000 training observations and 1,000 testing observations. These were generated such that 4 modeling techniques (regression tree, linear model, neural network, random forest) will yield the same R^2 and RMSE but will fit the models very differently.

Usage

rashomon_quartet

rashomon_quartet_train

rashomon_quartet_test

Format

rashomon_quartet: A dataframe with 2000 rows and 5 variables:

- split: train, test
- x1
- x2
- x3
- y

rashomon_quartet_train: A dataframe with 1000 rows and 4 variables:

- x1
- x2
- x3
- y

rashomon_quartet_test: A dataframe with 1000 rows and 4 variables:

- x1
- x2
- x3
- y

Details

There are three explanatory variables x_1 , x_2 , x_3 and one outcome y generated as:

$$y = \sin((3x_1 + x_2)/5) + \varepsilon$$

where $\varepsilon \sim N(0, 1/3)$ and $[x_1, x_2, x_3] \sim N(0, \Sigma_{3 \times 3})$ and $\Sigma_{3 \times 3}$ has 1 on the diagonal and 0.9 elsewhere.

If fit using the following hyperparameters, each model will yield an R^2 of 0.73 and an RMSE of 0.354

- Regression tree: max depth: 3, min split: 250
- Linear model: all main effects
- Random Forest: mtry: 1, number of trees: 100
- Neural network: hidden neurons in each layer: 8, 4, threshold for partial derivatives of the error function as stopping criteria: 0.05

[rashomon_quartet_train](#) contains just the training data and [rashomon_quartet_test](#) contains only the test data.

References

P. Biecek, H. Baniecki, M. Krzyżiński, D. Cook. Performance is not enough: the story of Rashomon’s quartet. Preprint arXiv:2302.13356v2, 2023.

variation_causal_quartet

Gelman Variation Causal Quartet Data

Description

This dataset contains 88 observations, each generated under a different mechanism of variation of the treatment effect with respect to some pre-exposure characteristic, z :

- (1) Constant effect
- (2) Low variation
- (3) High variation
- (4) Occasional large effects

Usage

variation_causal_quartet

Format

A dataframe with 88 rows and 5 variables:

- dataset: The data generating mechanism
- exposure: exposure
- covariate: a pre-exposure factor
- outcome: outcome
- .causal_effect: Latent true causal effect

References

Gelman, A., Hullman, J., & Kennedy, L. (2023). Causal quartets: Different ways to attain the same average treatment effect. arXiv preprint arXiv:2302.12878.

Hullman J (2023). *causalQuartet: Create Causal Quartets for Interrogating Average Treatment Effects*. R package version 0.0.0.9000.

Index

* datasets

- anscombe_leverage, 2
- anscombe_linear, 3
- anscombe_nonlinear, 4
- anscombe_outlier, 5
- anscombe_quartet, 6
- causal_collider, 7
- causal_collider_time, 7
- causal_confounding, 10
- causal_m_bias, 11
- causal_mediator, 10
- causal_quartet, 11
- datasaurus_dozen, 12
- heterogeneous_causal_quartet, 13
- interaction_triptych, 14
- rashomon_quartet, 15
- variation_causal_quartet, 16

- anscombe_leverage, 2
- anscombe_linear, 3
- anscombe_nonlinear, 4
- anscombe_outlier, 5
- anscombe_quartet, 6

- causal_collider, 7
- causal_collider_time, 7
- causal_confounding, 10
- causal_confounding_time
 - (causal_collider_time), 7
- causal_m_bias, 11
- causal_m_bias_time
 - (causal_collider_time), 7
- causal_mediator, 10
- causal_mediator_time
 - (causal_collider_time), 7
- causal_quartet, 11
- causal_quartet_time
 - (causal_collider_time), 7

- datasaurus_dozen, 12

- heterogeneous_causal_quartet, 13

- interaction_triptych, 14

- rashomon_quartet, 15
- rashomon_quartet_test, 16
- rashomon_quartet_test
 - (rashomon_quartet), 15
- rashomon_quartet_train, 16
- rashomon_quartet_train
 - (rashomon_quartet), 15

- variation_causal_quartet, 16