Package 'OLSengine'

September 18, 2026

Type Package

Title Transparent and Assisted Linear Modeling Engine

Version 1.2.0

Description Unified estimation, diagnostics, and reporting for ordinary least squares (OLS) regression, ANOVA/t-tests, logistic regression, panel data (fixed/random effects with Hausman test), instrumental variables (2SLS with weak instrument diagnostics), and difference-in-differences. Designed for applied researchers in social sciences with integrated ``Methodological Customs'' that audit assumptions and provide literature references. All methods implemented in pure base R without external dependencies beyond stats and graphics packages.

License MIT + file LICENSE

Encoding UTF-8

Depends R (>= 3.5.0)

Imports stats, graphics

Suggests knitr, rmarkdown

VignetteBuilder knitr

URL https://github.com/msoto-perez/OLSengine

BugReports https://github.com/msoto-perez/OLSengine/issues

RoxygenNote 7.3.3

LazyData true

NeedsCompilation no

Author Manuel Soto-Pérez [aut, cre] (ORCID: <https://orcid.org/0000-0002-2298-2047>)

Maintainer Manuel Soto-Pérez <msoto@up.edu.mx>

Repository CRAN

Date/Publication 2026-09-18 15:32:10 UTC

Contents

`academic_salaries` *Academic Salaries Dataset for U.S. College Professors*

Description

Real data on 9-month academic salaries for assistant professors, associate professors, and full professors at a U.S. college. This dataset is provided for educational purposes to demonstrate regression modeling, ANOVA, and logistic regression with `paper_engine`.

Usage

```
academic_salaries
```

Format

A data frame with 397 observations and 7 variables:

rank Factor with 3 levels: `"AsstProf"` (Assistant Professor), `"AssocProf"` (Associate Professor), `"Prof"` (Full Professor). Represents academic rank.

discipline Factor with 2 levels: `"A"` (theoretical departments, e.g., mathematics, physics) and `"B"` (applied departments, e.g., engineering, business). Represents academic discipline category.

years_since_phd Numeric. Number of years since the faculty member earned their PhD.

years_service Numeric. Number of years the faculty member has served at this institution.

sex Factor with 2 levels: `"Female"` and `"Male"`.

salary Numeric. Nine-month academic salary in U.S. dollars (2008-09 academic year).

high_earner Integer. Binary indicator (0 = No, 1 = Yes) marking faculty in the top 33% of salaries. Created for logistic regression demonstrations.

Details

This dataset enables demonstration of OLSengine's three core methods:

- **OLS Regression**: Modeling salary as a function of rank, discipline, experience, and sex to assess wage determinants and potential gender disparities.
- **ANOVA**: Comparing mean salaries across academic ranks or disciplines.
- **Logistic Regression**: Predicting the probability of being a high earner based on experience, rank, and discipline.

The data were collected in the 2008-09 academic year and reflect institutional salary structures at that time. Gender wage gap research in academia remains an active area of inquiry (Ginther & Kahn, 2021).

Source

This dataset is adapted from the `Salaries` dataset in the **carData** package (Fox & Weisberg, 2019), which was originally compiled for the textbook *An R Companion to Applied Regression* (Fox & Weisberg, 2011). The original data source is a U.S. college during the 2008-09 academic year.

Licensed under GPL (>= 2), consistent with the **carData** package license.

References

Fox, J., & Weisberg, S. (2011). *An R Companion to Applied Regression* (2nd ed.). Thousand Oaks, CA: Sage. https://www.john-fox.ca/Companion/

Fox, J., & Weisberg, S. (2019). *carData: Companion to Applied Regression Data Sets*. R package version 3.0-3. https://CRAN.R-project.org/package=carData

Ginther, D. K., & Kahn, S. (2021). Women in academic science: A changing landscape. *Psychological Science in the Public Interest, 22*(1), 3-65.

Examples

```
# Load the dataset
data(academic_salaries)

# Explore structure
str(academic_salaries)
summary(academic_salaries)

# OLS: Modeling salary determinants
ols_model <- paper_engine(
  salary ~ rank + discipline + years_since_phd + sex,
  data = academic_salaries,
  model = "ols",
  robust = "auto"
)
print(ols_model$tables$Table2_OLS_Estimation)
print(ols_model$messages)

# ANOVA: Salary differences across academic ranks
anova_model <- paper_engine(
  salary ~ rank,
  data = academic_salaries,
  model = "anova"
)
print(anova_model$tables$Descriptive_Means)

# Logit: Predicting high earner status
logit_model <- paper_engine(
  high_earner ~ years_since_phd + rank + discipline,
  data = academic_salaries,
  model = "logit"
)
print(logit_model$tables$Table2_Logit_Estimation)
```

```
# Visualization
plot_engine(ols_model)
```

---

paper_engine *Transparent and Assisted Linear Modeling Engine*

Description

Estimates OLS regression, ANOVA/t-tests, binary logistic regression, panel data models, instrumental variables, or difference-in-differences using pure base R matrix algebra. Automatically audits statistical assumptions through an integrated methodological customs layer and returns publication-ready APA-formatted tables. Designed for applied researchers and early-career academics who need a single, transparent workflow from estimation to reporting.

Usage

```
paper_engine(
  formula,
  data,
  model = "ols",
  robust = FALSE,
  non_parametric = FALSE,
  paired = FALSE,
  entity_id = NULL,
  time_id = NULL,
  method = "auto",
  instruments = NULL,
  treatment_var = NULL,
  time_var = NULL,
  treatment_level = NULL,
  post_level = NULL,
  digits = 2
)
```

Arguments

formula	A formula object specifying the model (e.g., y ~ x1 + x2).
data	A data frame containing all variables referenced in formula.
model	A character string indicating the estimation engine. One of "ols" (default), "anova", "logit", "panel", "iv", or "did".
robust	Logical or "auto". Controls heteroskedasticity-robust standard errors (HC3) for OLS models. If TRUE, HC3 SEs are always applied. If "auto", they are applied only when the Breusch-Pagan test detects heteroskedasticity ($p < .05$). Default is FALSE.

non_parametric	Logical or "auto". Controls non-parametric fallback for ANOVA/t-test models. If TRUE, Kruskal-Wallis or Wilcoxon tests are used. If "auto", transition occurs when Shapiro-Wilk detects non-normality ($p < .05$). Default is FALSE.
paired	Logical. If TRUE, assumes paired/dependent samples for ANOVA/t-test models (pre-post designs). Default is FALSE.
entity_id	Character string. Name of the entity/individual identifier variable for panel data models. Required when model = "panel".
time_id	Character string. Name of the time period identifier variable for panel data models. Required when model = "panel".
method	Character string for panel data. One of "auto" (default, uses Hausman test to select between FE and RE), "fe" (Fixed Effects), or "re" (Random Effects). Only used when model = "panel".
instruments	A formula specifying instrumental variables for IV models (e.g., ~ z1 + z2). Required when model = "iv". Instruments must satisfy relevance (correlated with endogenous X) and exogeneity (uncorrelated with error term).
treatment_var	Character string. Name of the treatment group variable for DiD models. Required when model = "did".
time_var	Character string. Name of the time period variable (pre/post) for DiD models. Required when model = "did".
treatment_level	Character string. Which level of treatment_var represents the treated group. If NULL, the second level is used.
post_level	Character string. Which level of time_var represents the post-treatment period. If NULL, the second level is used.
digits	Integer. Number of decimal places in output tables. Default is 2.

Value

An object of class basic_model, which is a list containing:

tables A list of formatted data frames with estimation results.

diagnostics A list of raw diagnostic statistics (p-values, fit indices).

messages A character vector of methodological guidance messages from the customs layer.

method A character string indicating the engine used ("ols", "anova", "logit", "panel", "iv", or "did").

data The cleaned data frame used for estimation (after listwise deletion).

Examples

```
# OLS example
set.seed(42)
df <- data.frame(y = rnorm(100), x1 = rnorm(100), x2 = rnorm(100))
result <- paper_engine(y ~ x1 + x2, data = df, model = "ols")
print(result$tables)
print(result$messages)
```

```
# ANOVA example
df2 <- data.frame(score = c(rnorm(30, 5), rnorm(30, 7)),
                  group = rep(c("A", "B"), each = 30))
result2 <- paper_engine(score ~ group, data = df2, model = "anova")
print(result2$tables)

# Logit example
df3 <- data.frame(y = rbinom(100, 1, 0.5), x = rnorm(100))
result3 <- paper_engine(y ~ x, data = df3, model = "logit")
print(result3$tables)
```

`plot_engine` *Generate Publication-Ready Plots for Basic Models*

Description

Produces minimalist APA-style plots from a `basic_model` object returned by `paper_engine`. The plot type is selected automatically based on the estimation method: a forest plot of coefficients with 95 and a logistic probability curve for logistic regression.

Usage

```
plot_engine(model_object, y_label = NULL, x_label = NULL)
```

Arguments

`model_object`	An object of class `basic_model` generated by `paper_engine`.
`y_label`	A character string for the Y-axis label. If `NULL` (default), a label is generated automatically from the model type.
`x_label`	A character string for the X-axis label. If `NULL` (default), a label is generated automatically from the model type.

Value

A base R plot rendered in the active graphics device. The function is called for its side effect (the plot) and returns `NULL` invisibly.

Examples

```
set.seed(42)
df <- data.frame(y = rnorm(100), x1 = rnorm(100), x2 = rnorm(100))
result <- paper_engine(y ~ x1 + x2, data = df, model = "ols")
plot_engine(result, y_label = "Outcome", x_label = "Predictors")
```

Index