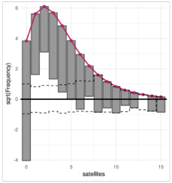
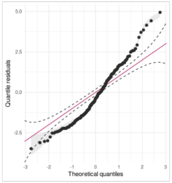
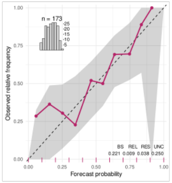
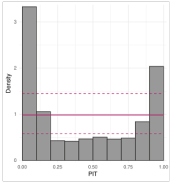
universität innsbruck









Visualizing Goodness of Fit of Probabilistic Regression Models

Achim Zeileis

https://topmodels.R-Forge.R-project.org/

Probabilistic regression models

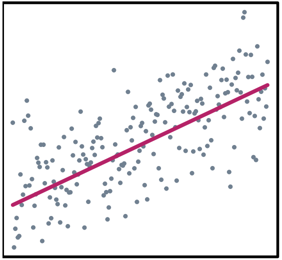
Classical approach: Model conditional expectation $\mathrm{E}(y_i|\boldsymbol{x}_i) = \mu_i$ ($i = 1, \ldots n$).

Regression model: $\mu_i \; = \; r(\boldsymbol{x}_i)$

Probabilistic regression models

Classical approach: Model conditional expectation $\text{E}(y_i|\boldsymbol{x}_i) = \mu_i$ $(i = 1, \ldots n)$.

Regression model: $\mu_i \;=\; r(\boldsymbol{x}_i)$

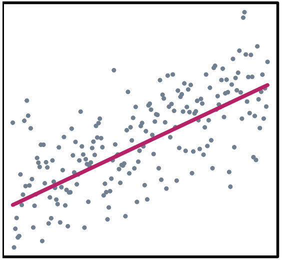


LM, GLM

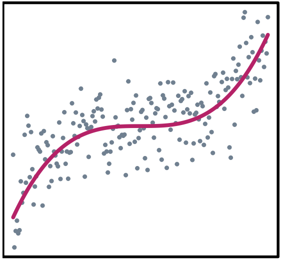
Probabilistic regression models

Classical approach: Model conditional expectation $\mathrm{E}(y_i|\boldsymbol{x}_i) = \mu_i$ ($i = 1, \dots n$).

Regression model: $\mu_i \; = \; r(\boldsymbol{x}_i)$



LM, GLM

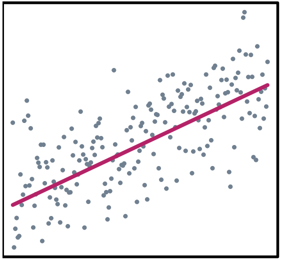


GAM

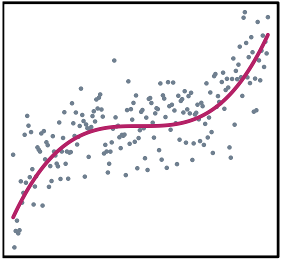
Probabilistic regression models

Classical approach: Model conditional expectation $\text{E}(y_i|\boldsymbol{x}_i) = \mu_i$ ($i = 1, \dots n$).

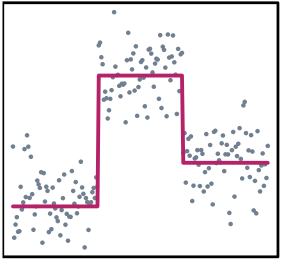
Regression model: $\mu_i \; = \; r(\boldsymbol{x}_i)$



LM, GLM



GAM

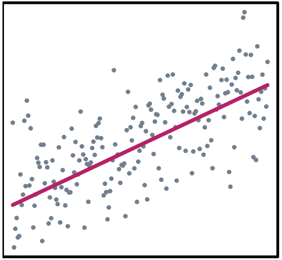


Regression tree

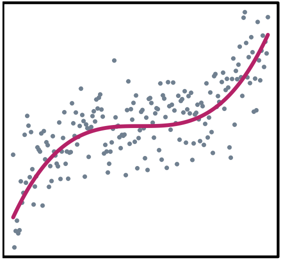
Probabilistic regression models

Classical approach: Model conditional expectation $\text{E}(y_i|\boldsymbol{x}_i) = \mu_i$ ($i = 1, \dots n$).

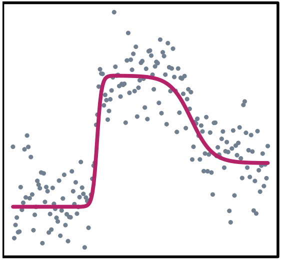
Regression model: $\mu_i \ = \ r(\boldsymbol{x}_i)$



LM, GLM



GAM



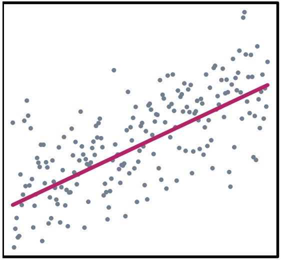
Random forest

Probabilistic regression models

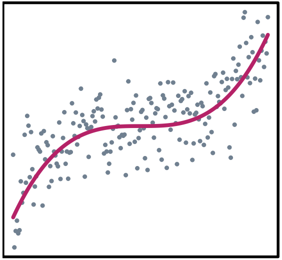
Classical approach: Model conditional expectation $\mathrm{E}(y_i|\boldsymbol{x}_i) = \mu_i$ ($i = 1, \ldots n$).

Regression model: $\mu_i \;=\; r(\boldsymbol{x}_i)$

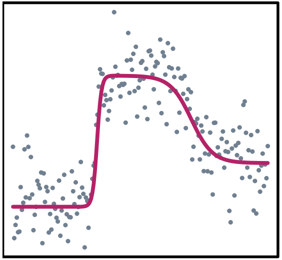
Often: Full conditional probability distribution is of interest.



LM, GLM



GAM



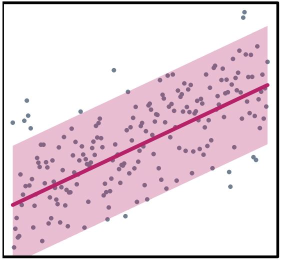
Random forest

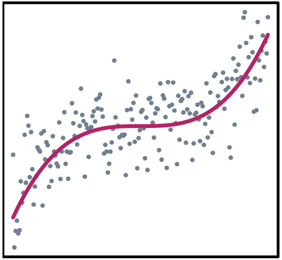
Probabilistic regression models

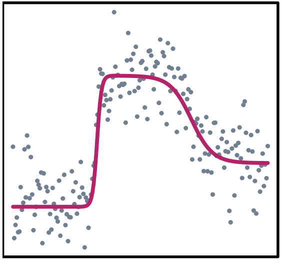
Classical approach: Model conditional expectation $\text{E}(y_i|\boldsymbol{x}_i) = \mu_i$ $(i = 1, \dots n)$.

Regression model: $\mu_i \ = \ r(\boldsymbol{x}_i)$

Often: Full conditional probability distribution is of interest.


Normal (G)LM w/ constant variance


GAM

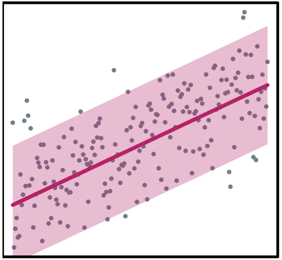

Random forest

Probabilistic regression models

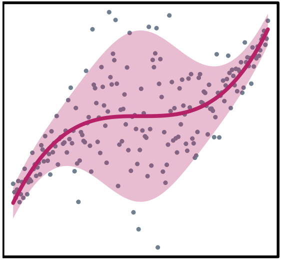
Classical approach: Model conditional expectation $\text{E}(y_i | \boldsymbol{x}_i) = \mu_i$ $(i = 1, \dots n)$.

Regression model: $\mu_i \ = \ r(\boldsymbol{x}_i)$

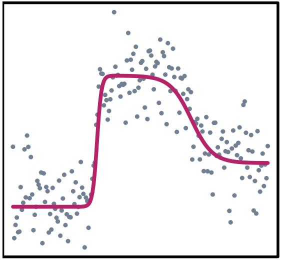
Often: Full conditional probability distribution is of interest.



Normal (G)LM w/ constant variance



GAMLSS



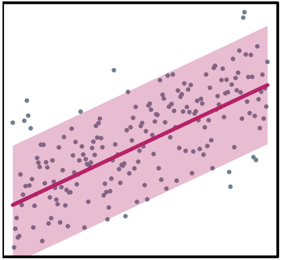
Random forest

Probabilistic regression models

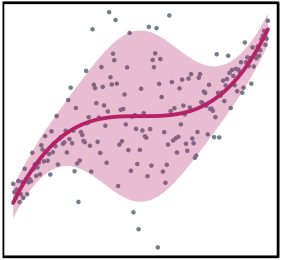
Classical approach: Model conditional expectation $\text{E}(y_i | \boldsymbol{x}_i) = \mu_i$ ($i = 1, \dots n$).

Regression model: $\mu_i \; = \; r(\boldsymbol{x}_i)$

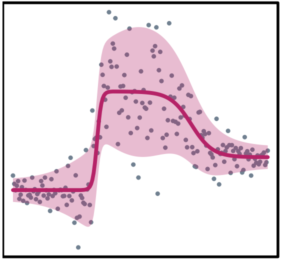
Often: Full conditional probability distribution is of interest.



Normal (G)LM w/ constant variance



GAMLSS



Distributional forest

Probabilistic regression models

Formally: Fit distribution with cumulative distribution function $F(y_i|\boldsymbol{\theta}_i)$ and parameter vector $\boldsymbol{\theta}_i$ for each observation y_i.

Probabilistic regression models

Formally: Fit distribution with cumulative distribution function $F(y_i|\boldsymbol{\theta}_i)$ and parameter vector $\boldsymbol{\theta}_i$ for each observation y_i.

Forecasting: $\hat{\theta}_i = \hat{\boldsymbol{r}}(\boldsymbol{x}_i)$.

- Model fit typically yields distribution parameters.
- Implies all other aspects of the distribution $F(\cdot|\boldsymbol{\theta}_i)$.
- Thus: Moments, quantiles, probabilities, ...

Illustration: Goals in the 2018 FIFA World Cup

Response: Goals scored by the two teams in all 64 matches.

Covariates: Basic match information and prediction of team (log-)abilities (based on bookmakers odds).

```
R> data("FIFA2018", package = "distributions3")
R> tail(FIFA2018, 2)
    goals team match  type    stage logability difference
127     4  FRA    64 Final knockout     0.8866      0.629
128     2  CRO    64 Final knockout     0.2576     -0.629
```

Model: Poisson GLM with mean λ_i using log link.

Illustration: Goals in the 2018 FIFA World Cup

In R:

```
R> m <- glm(goals ~ difference, data = FIFA2018, family = poisson)
```

Forecasting: In-sample for simplicity.

```
R> tail(procast(m), 2)
                              distribution
127 Poisson distribution (lambda = 1.6044)
128 Poisson distribution (lambda = 0.9538)
```

Illustration: Goals in the 2018 FIFA World Cup

In R:

```
R> m <- glm(goals ~ difference, data = FIFA2018, family = poisson)
```

Forecasting: In-sample for simplicity.

```
R> tail(procast(m), 2)
                              distribution
127 Poisson distribution (lambda = 1.6044)
128 Poisson distribution (lambda = 0.9538)
```

Implies:

- Probabilities for match results (assuming independence of goals).
- Corresponding probabilities for win/draw/lose.

Illustration: Goals in the 2018 FIFA World Cup

Example: Probabilities for final France-Croatia.

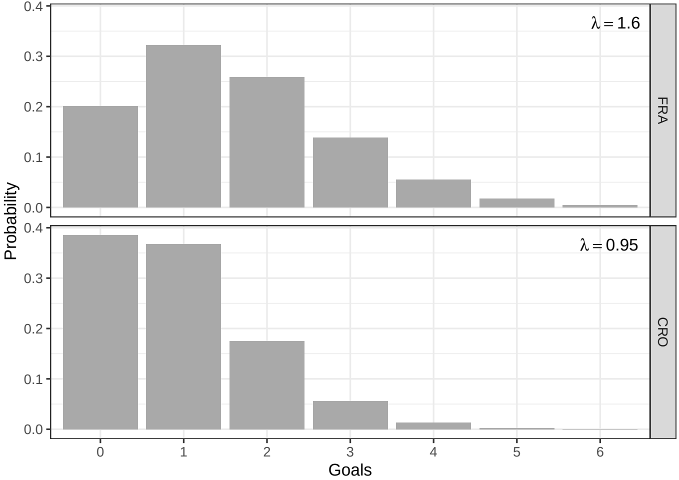
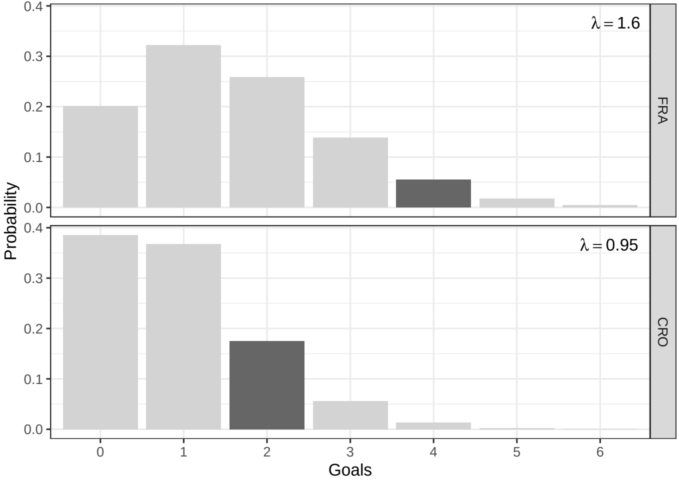


Illustration: Goals in the 2018 FIFA World Cup

Example: Probabilities for final France-Croatia. Result 4-2.



Goodness of fit

Idea:

- Use visualizations instead of just summing up scores.
- Gain more insights graphically.
- Reveal different types of model misspecification.

Goodness of fit

Idea:

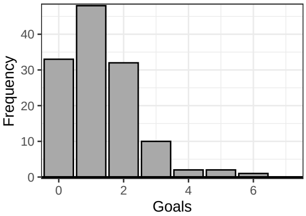
- Use visualizations instead of just summing up scores.
- Gain more insights graphically.
- Reveal different types of model misspecification.

Questions: Graphics are not new but novel unifying view.

- What are useful elements of such graphics?
- What are relative (dis)advantages?

Goodness of fit

Ideas: Illustrated for FIFA Poisson model.

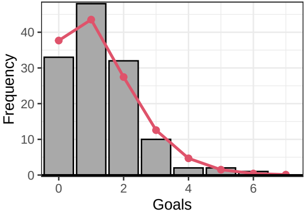


Marginal calibration:

- Observed frequencies.

Goodness of fit

Ideas: Illustrated for FIFA Poisson model.

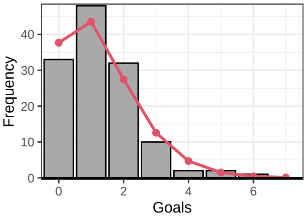


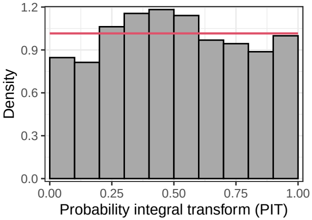
Marginal calibration:

- Observed frequencies.
- Compare: Expected.

Goodness of fit

Ideas: Illustrated for FIFA Poisson model.





Marginal calibration:

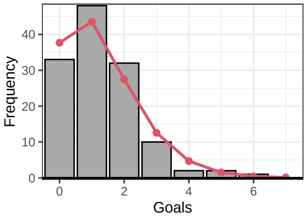
- Observed frequencies.
- Compare: Expected.

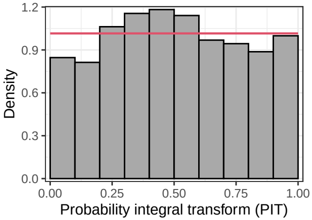
Probabilistic calibration:

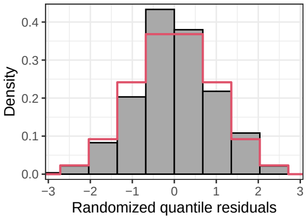
- Probability integral $u_i = F(y_i \mid \hat{\theta}_i)$.
- Compare: Uniform.

Goodness of fit

Ideas: Illustrated for FIFA Poisson model.







Marginal calibration:

- Observed frequencies.
- Compare: Expected.

Probabilistic calibration:

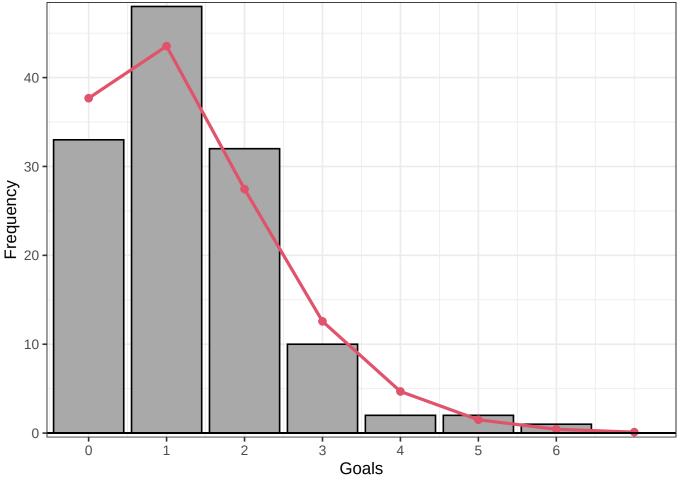
- Probability integral $u_i = F(y_i \mid \hat{\theta}_i)$.
- Compare: Uniform.

Probabilistic calibration:

- Quantile residuals $\Phi^{-1}(u_i)$.
- Compare: Normal

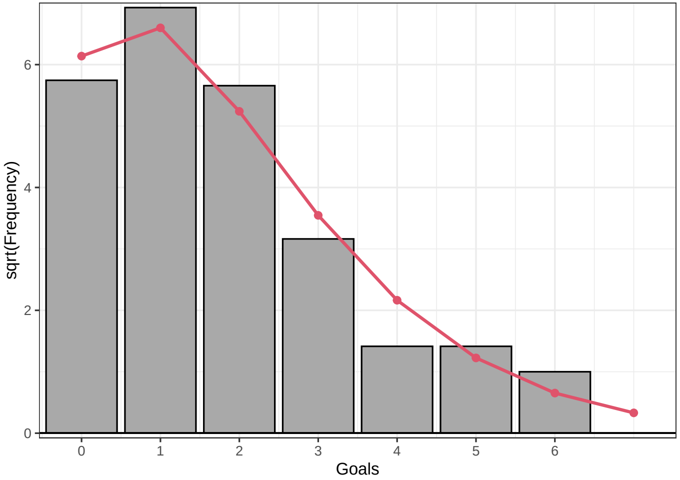
Goodness of fit: Marginal calibration

Observed vs. expected frequencies: Standing, with reference line.



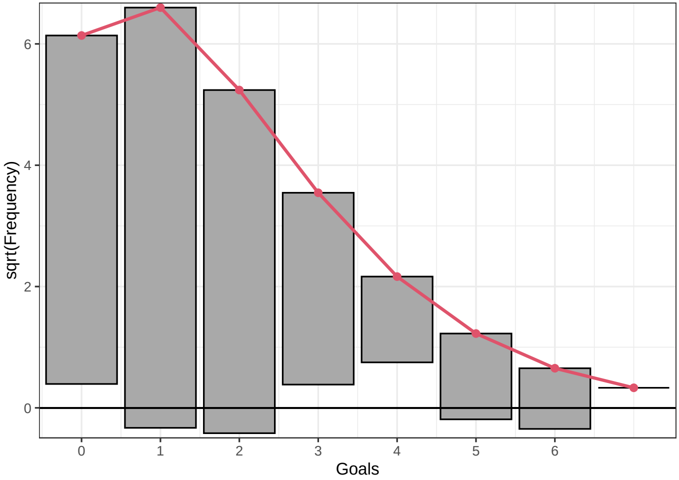
Goodness of fit: Marginal calibration

$\sqrt{\textbf{Observed}}$ **vs.** $\sqrt{\textbf{expected}}$ **frequencies:** Standing, with reference line.



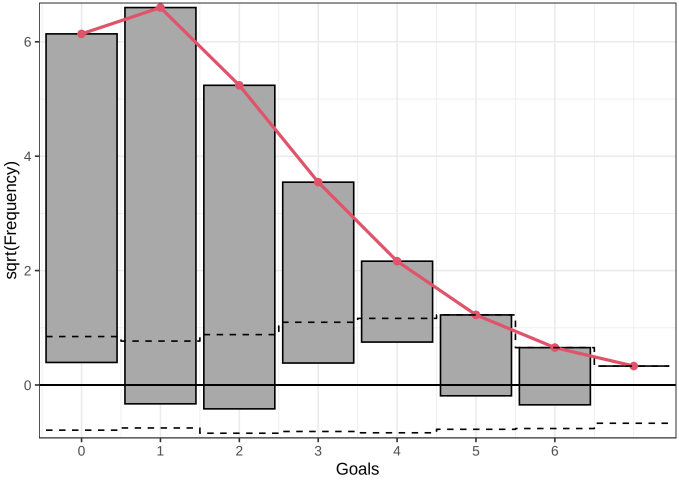
Goodness of fit: Marginal calibration

$\sqrt{\textbf{Observed}}$ **vs.** $\sqrt{\textbf{expected}}$ **frequencies:** Hanging.



Goodness of fit: Marginal calibration

$\sqrt{\textbf{Observed}}$ **vs.** $\sqrt{\textbf{expected}}$ **frequencies:** Hanging, with confidence interval.



Goodness of fit: Marginal calibration

Rootogram:

- Frequencies on raw or square-root scale.
- Hanging, standing, or suspended styled rootograms.

Goodness of fit: Marginal calibration

Rootogram:

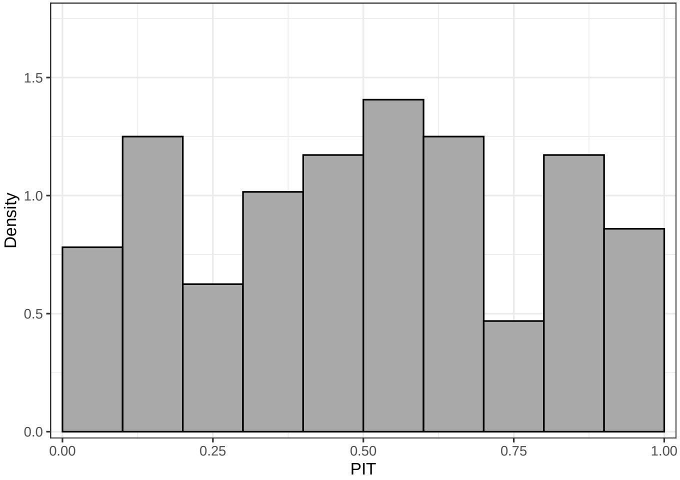
- Frequencies on raw or square-root scale.
- Hanging, standing, or suspended styled rootograms.

Overall:

- *Advantage:* Scale of observations is natural, direct interpretation.
- *Disadvantage:* Needs to be compared with a combination of distributions.

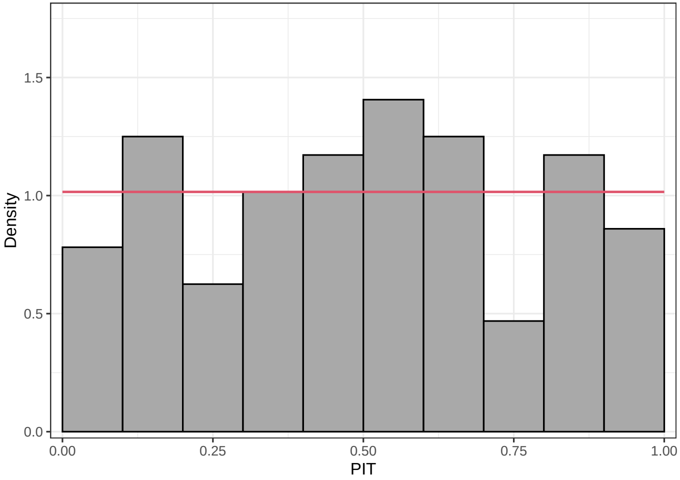
Goodness of fit: Probabilistic calibration

PIT: Randomization 1a.



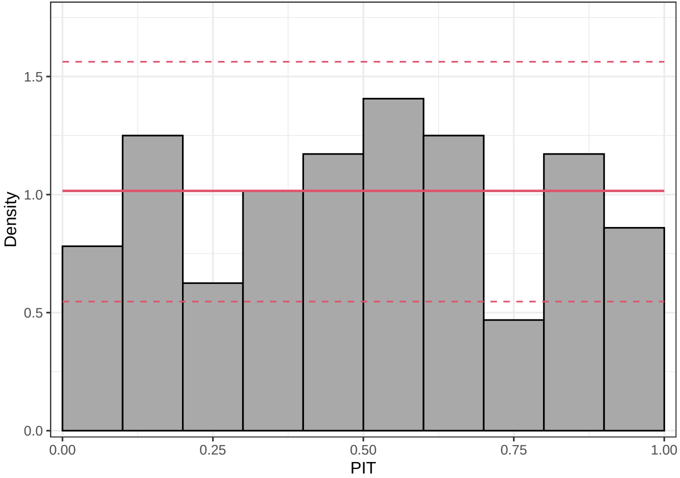
Goodness of fit: Probabilistic calibration

PIT: Randomization 1a, with reference line.



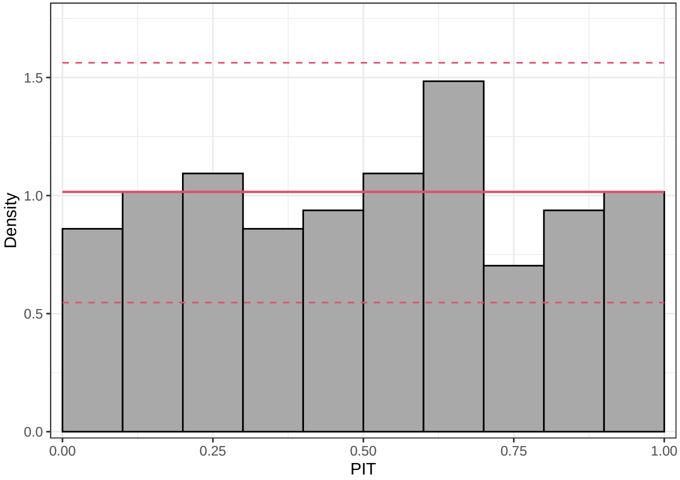
Goodness of fit: Probabilistic calibration

PIT: Randomization 1a, with reference line and confidence interval.



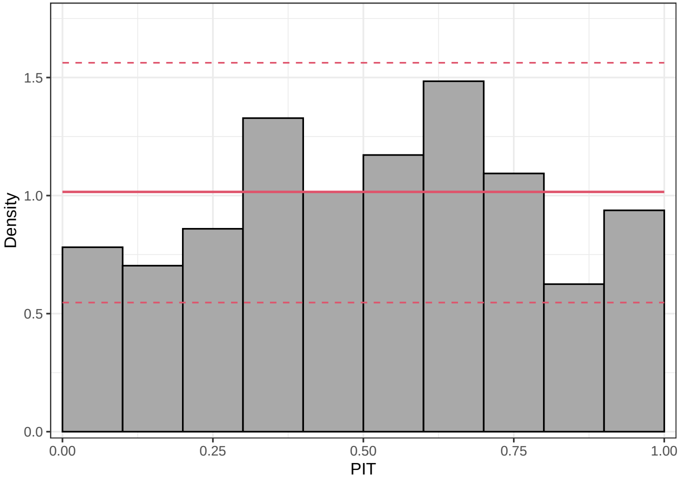
Goodness of fit: Probabilistic calibration

PIT: Randomization 1b.



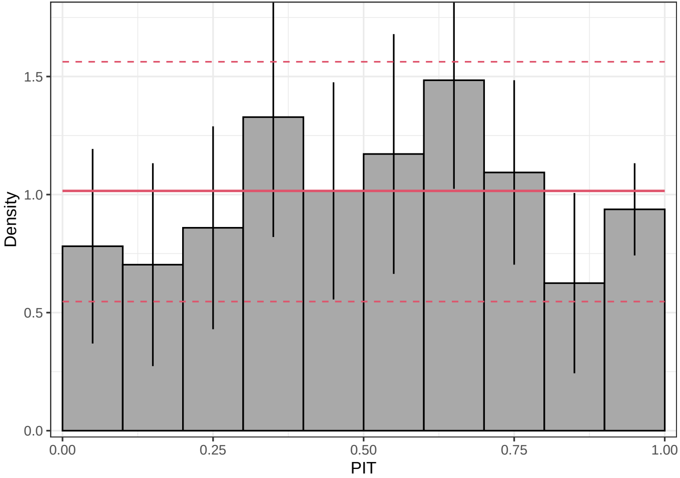
Goodness of fit: Probabilistic calibration

PIT: Randomization 1c.



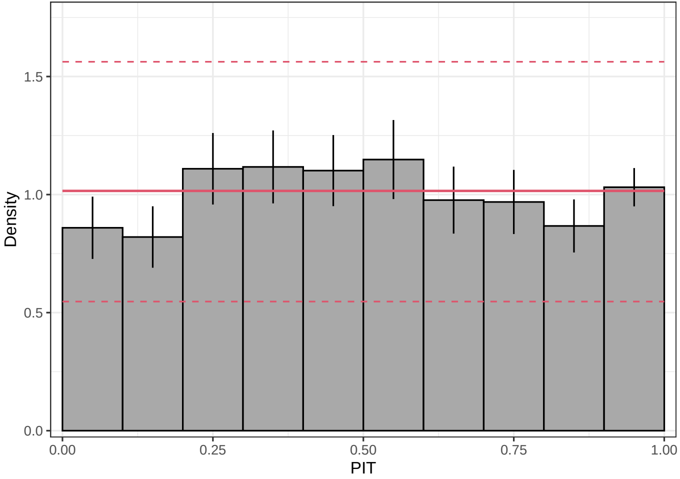
Goodness of fit: Probabilistic calibration

PIT: Randomization 1c, with simulation intervals.



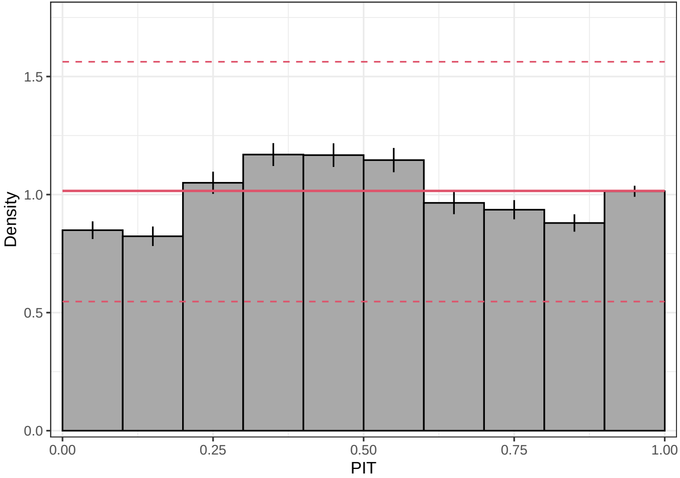
Goodness of fit: Probabilistic calibration

PIT: 10 random draws.



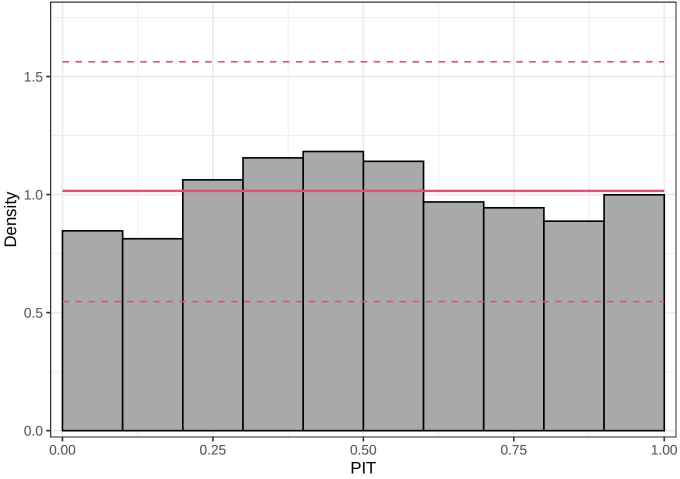
Goodness of fit: Probabilistic calibration

PIT: 100 random draws.



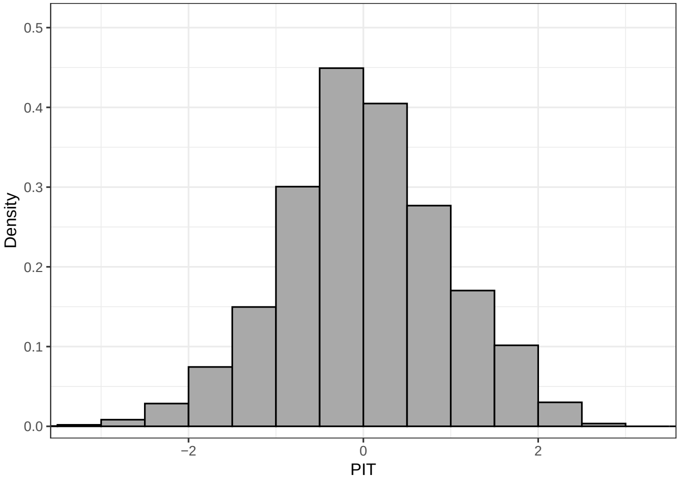
Goodness of fit: Probabilistic calibration

PIT: Expected.



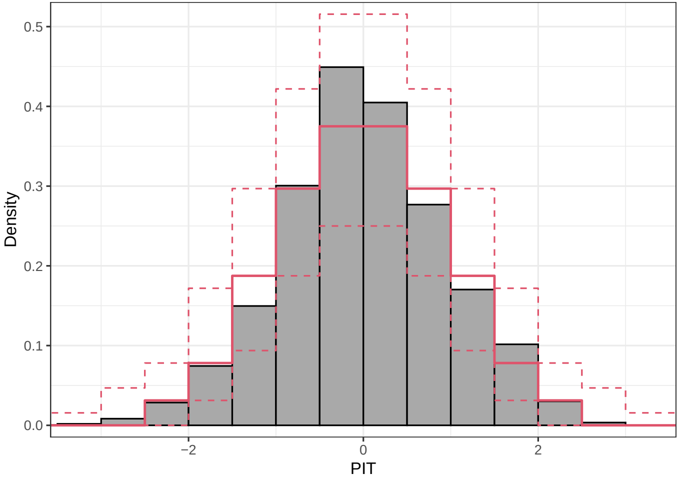
Goodness of fit: Probabilistic calibration

Randomized quantile residuals: Expected.



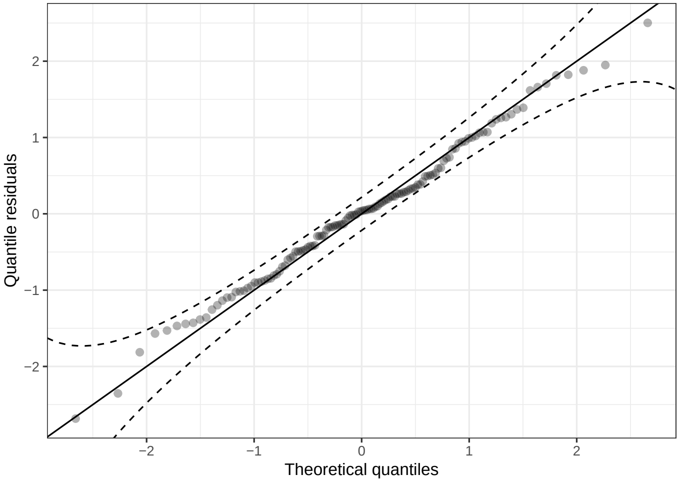
Goodness of fit: Probabilistic calibration

Randomized quantile residuals: Expected, with reference.



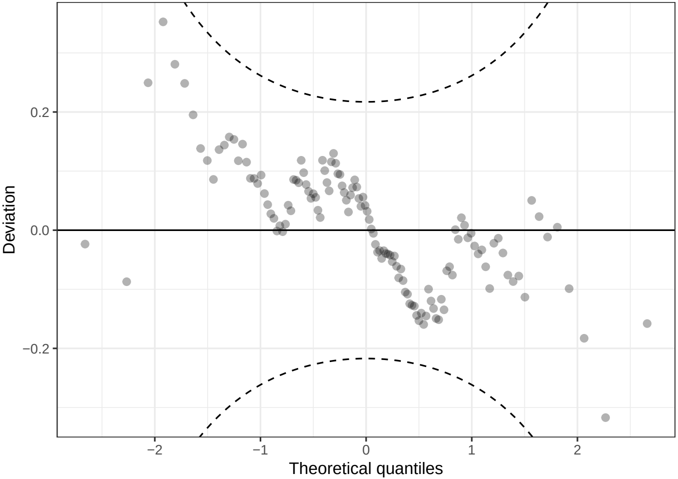
Goodness of fit: Probabilistic calibration

Observed vs. expected quantiles: Q-Q plot.



Goodness of fit: Probabilistic calibration

Observed vs. expected quantiles: Detrended Q-Q plot (worm plot).



Goodness of fit: Probabilistic calibration

PIT histogram:

- Probability scale or transformed to normal scale.
- Randomized or expected for discrete distributions.

Goodness of fit: Probabilistic calibration

PIT histogram:

- Probability scale or transformed to normal scale.
- Randomized or expected for discrete distributions.

Q-Q residuals plot:

- Normal or uniform scale.
- Detrended Q-Q plot (worm plot).

Goodness of fit: Probabilistic calibration

PIT histogram:

- Probability scale or transformed to normal scale.
- Randomized or expected for discrete distributions.

Q-Q residuals plot:

- Normal or uniform scale.
- Detrended Q-Q plot (worm plot).

Overall:

- *Advantage:* Comparison with only one distribution (uniform or normal).
- *Disadvantages:* Scale is not so natural. May require randomization.

Illustration: Loss aversion in adolescents

Experiment: Behaviour of adolescents (mostly 11–19).

- *Setup:* Nine rounds of a lottery with positive expectation.
- *Response:* Proportion of invested points across all rounds.
- *Covariates:* Arrangement (single vs. team), gender, age.

Illustration: Loss aversion in adolescents

Experiment: Behaviour of adolescents (mostly 11–19).

- *Setup:* Nine rounds of a lottery with positive expectation.
- *Response:* Proportion of invested points across all rounds.
- *Covariates:* Arrangement (single vs. team), gender, age.

Models:

- Ordinary least squares, interpreted as homoscedastic Gaussian model.
- Extended-support beta mixture regression (with point masses for 0 and 1).

Illustration: Loss aversion in adolescents

Experiment: Behaviour of adolescents (mostly 11–19).

- *Setup:* Nine rounds of a lottery with positive expectation.
- *Response:* Proportion of invested points across all rounds.
- *Covariates:* Arrangement (single vs. team), gender, age.

Models:

- Ordinary least squares, interpreted as homoscedastic Gaussian model.
- Extended-support beta mixture regression (with point masses for 0 and 1).

Goodness of fit: Similar fitted means but rather different distributions.

Illustration: Loss aversion in adolescents

Rootogram:

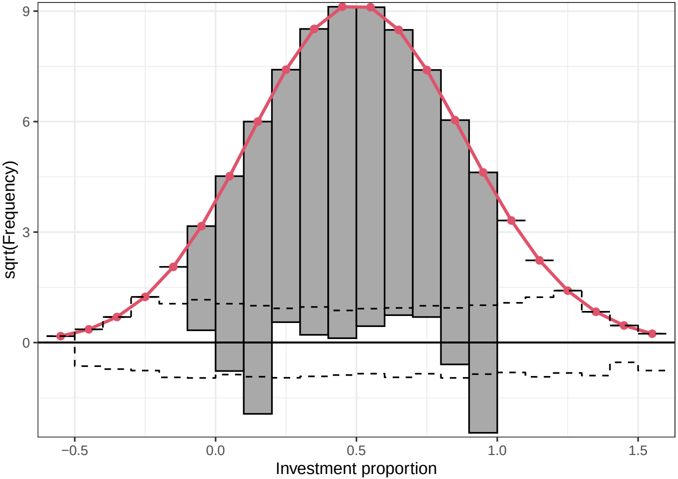


Illustration: Loss aversion in adolescents

Rootogram:

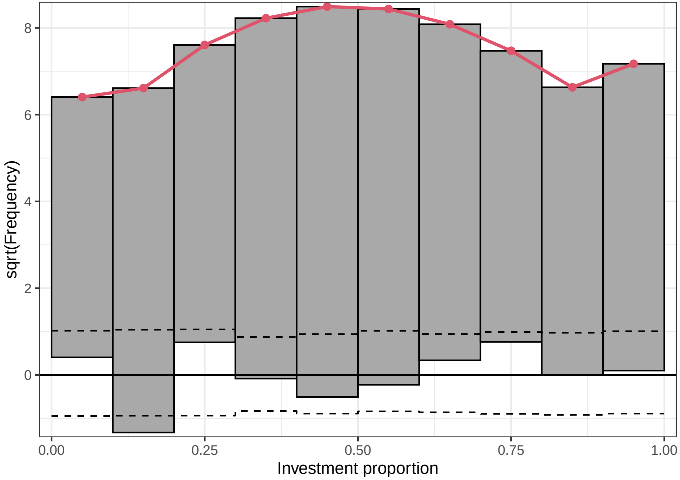


Illustration: Loss aversion in adolescents

PIT histogram:

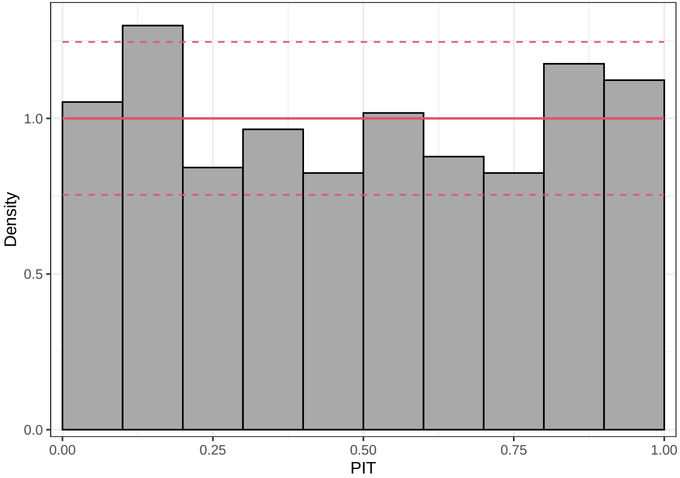


Illustration: Loss aversion in adolescents

PIT histogram:

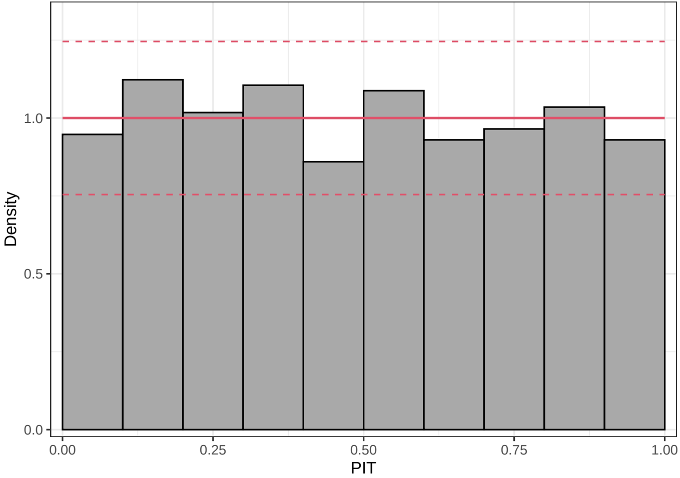


Illustration: Loss aversion in adolescents

PIT histogram:

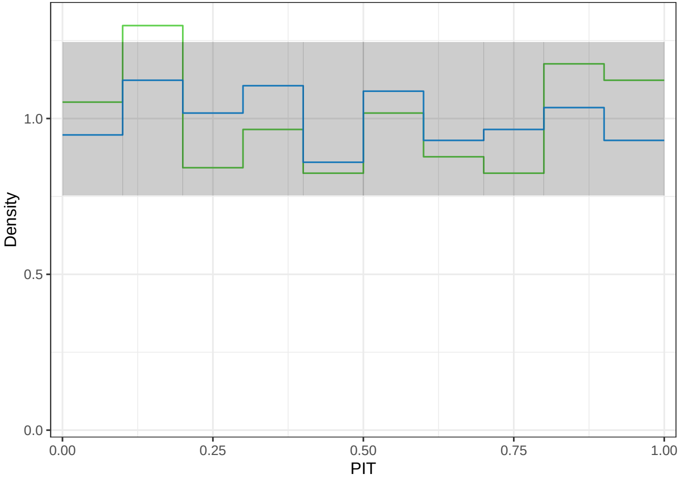


Illustration: Loss aversion in adolescents

Q-Q residual plot:

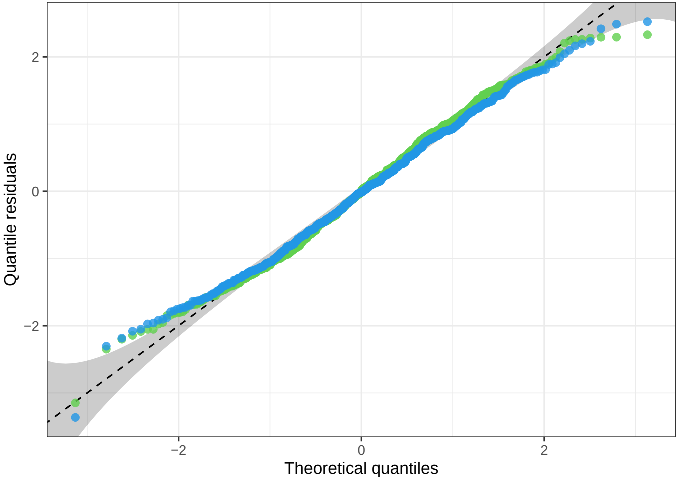
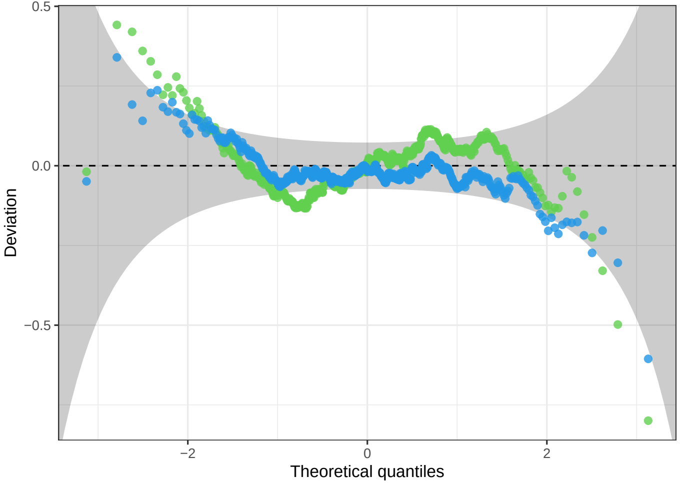


Illustration: Loss aversion in adolescents

Q-Q residual plot: Detrended.



Software: topmodels

R package: *topmodels*. Forecasting and assessment of probabilistic models.

Not yet on CRAN: `https://topmodels.R-Forge.R-project.org/`

Visualizations:

`rootogram()`	Rootograms of observed and fitted frequencies
`pithist()`	PIT histograms
`qqrplot()`	Q-Q plots for quantile residuals
`wormplot()`	Worm plots for quantile residuals
`reliagram()`	(Extended) reliability diagrams

Software: topmodels

Numeric quantities:

`procast()`	Probabilistic forecasts (probabilities, quantiles, etc.)
`proscore()`	Evaluate scoring rules for procasts
`pitresiduals()`	Probability integral transform (PIT) residuals
`qresiduals()`	(Randomized) quantile residuals

Software: topmodels

Numeric quantities:

`procast()`	Probabilistic forecasts (probabilities, quantiles, etc.)
`proscore()`	Evaluate scoring rules for procasts
`pitresiduals()`	Probability integral transform (PIT) residuals
`qresiduals()`	(Randomized) quantile residuals

Object orientation:

- Work with distribution objects (vectorized) from *distributions3*.
- Model classes like `lm`, `glm`, `gamlss`, `bamlss`, `hurdle`, `zeroinfl`, ...
- New model classes can be easily added if distribution can be extracted.

References

Lang MN, Zeileis A, Stauffer R, *et al.* (2023). "topmodels: Infrastructure for Inference and Forecasting in Probabilistic Models." *R package version 0.3-0*. `https://topmodels.R-Forge.R-project.org/`

Hayes A, Moller-Trane R, Jordan D, Northrop P, Lang MN, Zeileis A, *et al.* (2022). "distributions3: Probability Distributions as S3 Objects." *R package version 0.2.1*. `https://alexpghayes.github.io/distributions3/`

Czado C, Gneiting T, Held L (2009). "Predictive Model Assessment for Count Data." *Biometrics*, **65**(4), 1254–1261. `doi:10.1111/j.1541-0420.2009.01191.x`

Kleiber C, Zeileis A (2016). "Visualizing Count Data Regressions Using Rootograms." *The American Statistician*, **70**(3), 296–303. `doi:10.1080/00031305.2016.1173590`

Zeileis A, Leitner C, Hornik K (2018) "Probabilistic Forecasts for the 2018 FIFA World Cup Based on the Bookmaker Consensus Model." Working Paper 2018-09. Working Papers in Economics; Statistics, Research Platform Empirical; Experimental Economics, Universität Innsbruck. `https://EconPapers.RePEc.org/RePEc:inn:wpaper:2018-09`.

Glätzle-Rützler D, Sutter M, Zeileis A (2015). "No Myopic Loss Aversion in Adolescents? An Experimental Note." *Journal of Economic Behavior & Organization*, **111**, 169–176. `doi:10.1016/j.jebo.2014.12.021`

Contact

Mastodon: @zeileis@fosstodon.org

X/Twitter: @AchimZeileis

Web: https://www.zeileis.org/