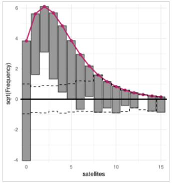
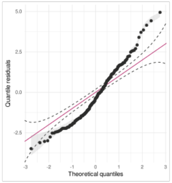
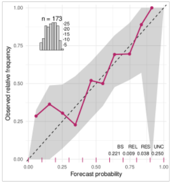
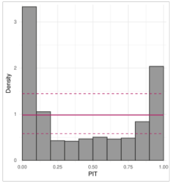
universität
innsbruck









A Toolbox for Probabilistic Regression Models

Forecasts, Visualizations, Scoring Rules, and Software Infrastructure

Achim Zeileis

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

Probabilistic regression models

Classical approach: Model conditional expectation $\text{E}(y_i|\boldsymbol{x}_i) = \mu_i$ of a response y_i given explanatory variables $\boldsymbol{x}_i$ for $i = 1, \ldots n$.

Regression model:

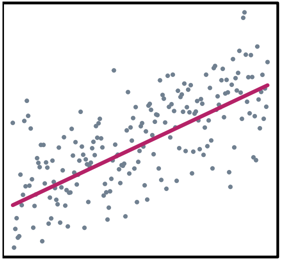
$$\mu_i \ = \ r(\boldsymbol{x}_i)$$

Probabilistic regression models

Classical approach: Model conditional expectation $\mathsf{E}(y_i|\boldsymbol{x}_i) = \mu_i$ of a response y_i given explanatory variables $\boldsymbol{x}_i$ for $i = 1, \dots n$.

Regression model: Linear model.

$$\mu_i \ = \ r(\boldsymbol{x}_i) \ = \ \beta_0 + \beta_1 \cdot x_{i,1} + \dots + \beta_k \cdot x_{i,k}$$



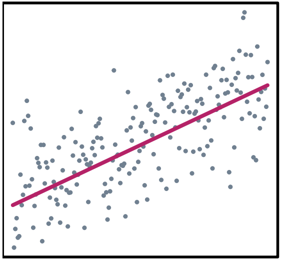
LM, GLM

Probabilistic regression models

Classical approach: Model conditional expectation $\text{E}(y_i|\boldsymbol{x}_i) = \mu_i$ of a response y_i given explanatory variables $\boldsymbol{x}_i$ for $i = 1, \dots n$.

Regression model: Generalized linear model with link function $g(\cdot)$.

$\mu_i \;=\; r(\boldsymbol{x}_i) \;=\; g^{-1}(\beta_0 + \beta_1 \cdot x_{i,1} + \dots + \beta_k \cdot x_{i,k})$



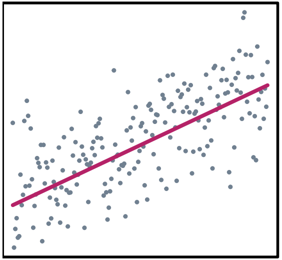
LM, GLM

Probabilistic regression models

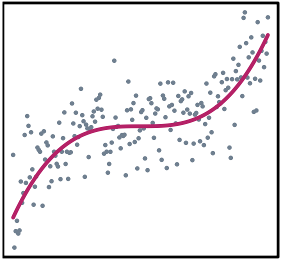
Classical approach: Model conditional expectation $\mathrm{E}(y_i|\boldsymbol{x}_i) = \mu_i$ of a response y_i given explanatory variables $\boldsymbol{x}_i$ for $i = 1, \dots n$.

Regression model: Generalized additive model with link function $g(\cdot)$.

$$\mu_i \;=\; r(\boldsymbol{x}_i) \;=\; g^{-1}(\beta_0 + s(x_{i,1}) + \cdots + s(x_{i,k}))$$



LM, GLM



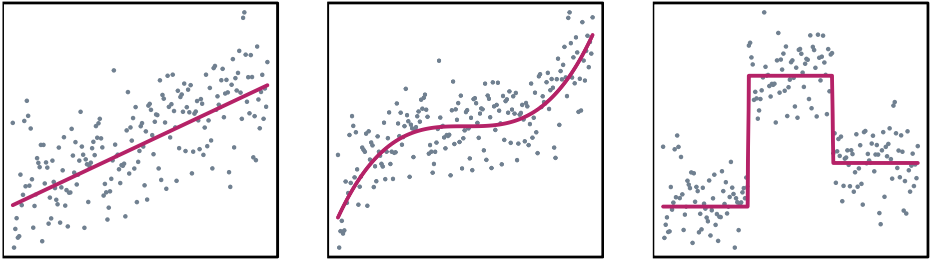
GAM

Probabilistic regression models

Classical approach: Model conditional expectation $E(y_i|\boldsymbol{x}_i) = \mu_i$ of a response y_i given explanatory variables $\boldsymbol{x}_i$ for $i = 1, \dots n$.

Regression model: Algorithmic, machine learning, nonparametric, . . .

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



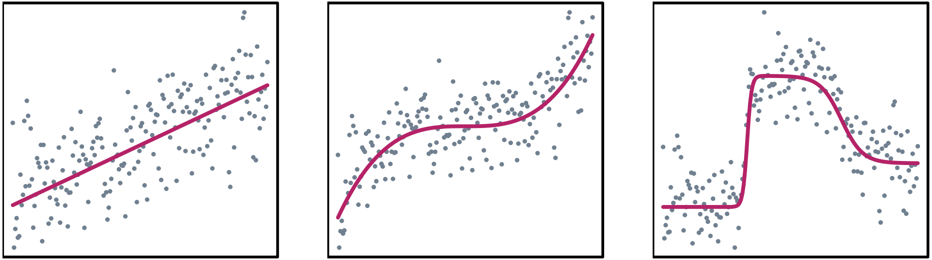
LM, GLM   GAM   Regression tree

Probabilistic regression models

Classical approach: Model conditional expectation $\text{E}(y_i|\boldsymbol{x}_i) = \mu_i$ of a response y_i given explanatory variables $\boldsymbol{x}_i$ for $i = 1, \dots n$.

Regression model: Algorithmic, machine learning, nonparametric, ...

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



LM, GLM — GAM — Random forest

Probabilistic regression models

Often: Further assumptions are made beyond the mean specification, especially for estimation and inference.

- Constant variance for least squares.
- Higher moments may co-vary with expectation μ_i, e.g., in exponential family (Poisson, binomial, ...)
- Full distribution for maximum likelihood or Bayesian MCMC, etc.

Probabilistic regression models

Often: Further assumptions are made beyond the mean specification, especially for estimation and inference.

- Constant variance for least squares.
- Higher moments may co-vary with expectation μ_i, e.g., in exponential family (Poisson, binomial, …)
- Full distribution for maximum likelihood or Bayesian MCMC, etc.

But typically: Focus is on conditional means.

- *Forecasting:* $\hat{\mu}_i = \hat{r}(\boldsymbol{x}_i)$.
- *Scores:* $(y_i - \hat{\mu}_i)^2$ or $|y_i - \hat{\mu}_i|$.
- *Inference:* Robustness/adjustments under misspecification.

Probabilistic regression models

However: Mean forecasts are often of limited interest.

- *Football:* Average goals of team A vs. team B.
- *Precipitation:* Average amount of precipitation today.

Probabilistic regression models

However: Mean forecasts are often of limited interest.

- *Football:* Average goals of team A vs. team B.
- *Precipitation:* Average amount of precipitation today.

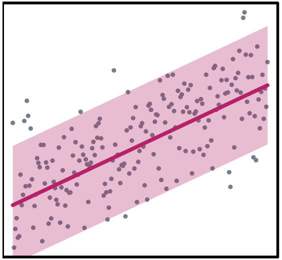
Instead: Full distribution of interest.

- *Football:* Probability for $0, 1, \dots$ goals, implying win/draw/lose probability.
- *Precipitation:* Probability of no/moderate/extreme precipitation.

Probabilistic regression models

Models:

- Classical models under full assumptions.

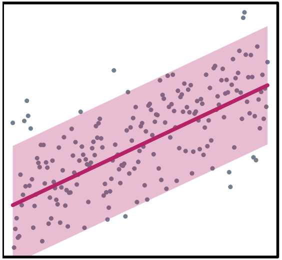


Normal (G)LM w/ constant variance

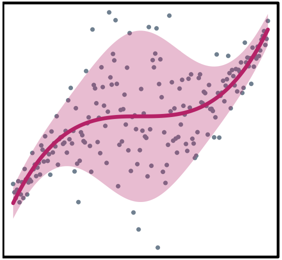
Probabilistic regression models

Models:

- Classical models under full assumptions.
- Generalized additive models for location, scale, and shape.



Normal (G)LM w/ constant variance

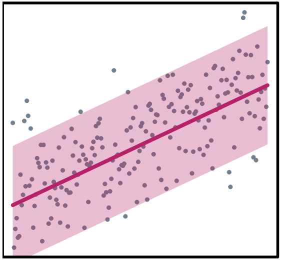


GAMLSS

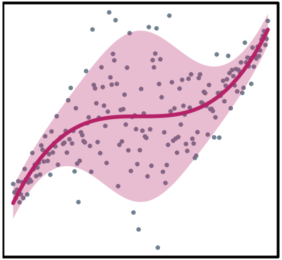
Probabilistic regression models

Models:

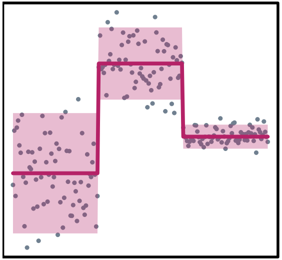
- Classical models under full assumptions.
- Generalized additive models for location, scale, and shape.
- Other distributional regression (Bayesian, trees, forests, neural nets, . . .).



Normal (G)LM w/ constant variance



GAMLSS

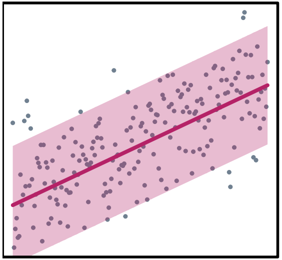


Distributional tree

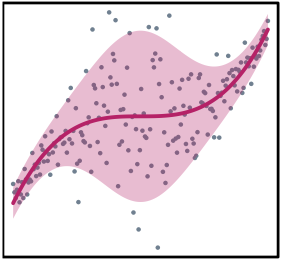
Probabilistic regression models

Models:

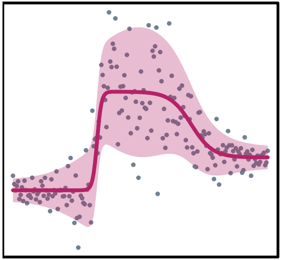
- Classical models under full assumptions.
- Generalized additive models for location, scale, and shape.
- Other distributional regression (Bayesian, trees, forests, neural nets, . . .).



Normal (G)LM w/ constant variance



GAMLSS



Distributional forest

Probabilistic regression models

Formally: Fit full probability distribution for each observation y_i.

Often: Assume parametric response distribution with parameter vector $\boldsymbol{\theta}_i$.

Cumulative distribution function: $F(y_i|\boldsymbol{\theta}_i)$.

Probability density function: $f(y_i|\boldsymbol{\theta}_i)$.

Probabilistic regression models

Formally: Fit full probability distribution for each observation y_i.

Often: Assume parametric response distribution with parameter vector $\boldsymbol{\theta}_i$.

Cumulative distribution function: $F(y_i | \boldsymbol{\theta}_i)$.

Probability density function: $f(y_i | \boldsymbol{\theta}_i)$.

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> head(FIFA2018)
  goals team match type stage logability difference
1     5  RUS     1    A group     0.1531     0.8638
2     0  KSA     1    A group    -0.7108    -0.8638
3     0  EGY     2    A group    -0.2066    -0.4438
4     1  URU     2    A group     0.2372     0.4438
5     3  RUS     3    A group     0.1531     0.3597
6     1  EGY     3    A group    -0.2066    -0.3597
```

Illustration: Goals in the 2018 FIFA World Cup

Model: Poisson GLM with log link.

Regression: Number of goals per team explained by ability difference.

$\log(\hat{\lambda}_i) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \mathsf{difference}_i$

Illustration: Goals in the 2018 FIFA World Cup

Model: Poisson GLM with log link.

Regression: Number of goals per team explained by ability difference.

$\log(\hat{\lambda}_i) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{difference}_i$

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

```
z test of coefficients:

            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.2127     0.0813    2.62   0.0088 **
difference    0.4134     0.1058    3.91  9.3e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Illustration: Goals in the 2018 FIFA World Cup

Forecasting: In-sample for simplicity.

```
R> head(procast(m))
                          distribution
1 Poisson distribution (lambda = 1.7680)
2 Poisson distribution (lambda = 0.8655)
3 Poisson distribution (lambda = 1.0297)
4 Poisson distribution (lambda = 1.4862)
5 Poisson distribution (lambda = 1.4354)
6 Poisson distribution (lambda = 1.0661)
```

Illustration: Goals in the 2018 FIFA World Cup

Forecasting: In-sample for simplicity.

```
R> head(procast(m))
                            distribution
1 Poisson distribution (lambda = 1.7680)
2 Poisson distribution (lambda = 0.8655)
3 Poisson distribution (lambda = 1.0297)
4 Poisson distribution (lambda = 1.4862)
5 Poisson distribution (lambda = 1.4354)
6 Poisson distribution (lambda = 1.0661)
```

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.

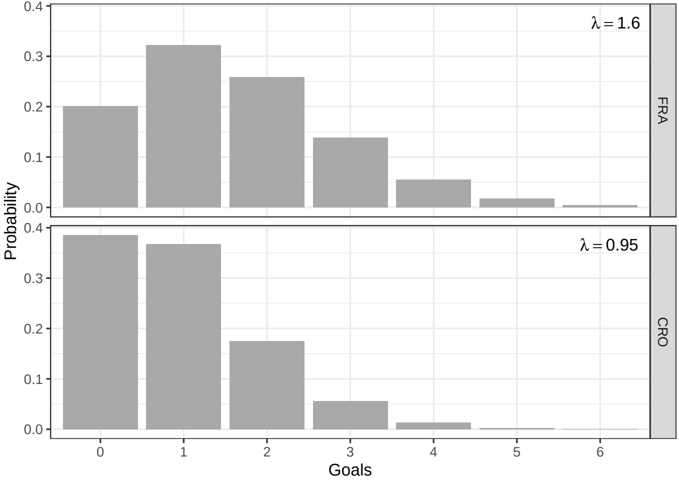


Illustration: Goals in the 2018 FIFA World Cup

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

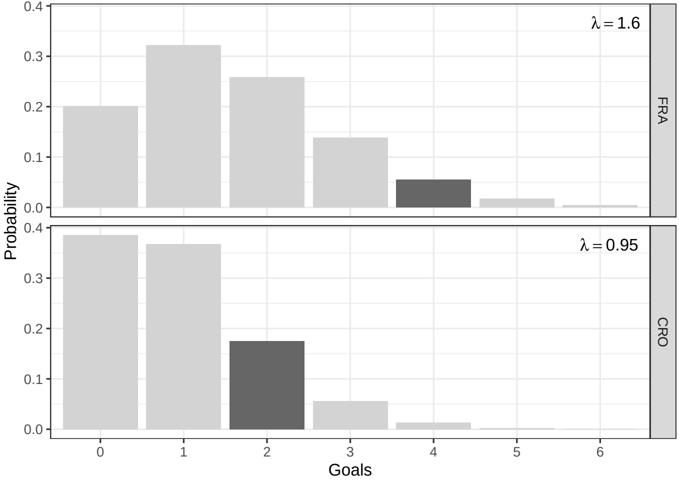


Illustration: Goals in the 2018 FIFA World Cup

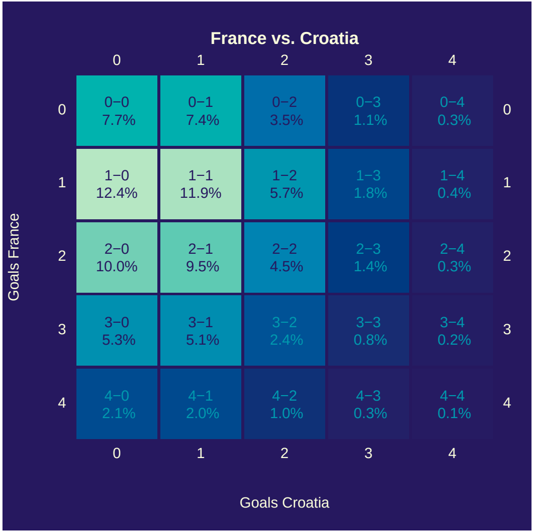


Illustration: Goals in the 2018 FIFA World Cup

Possible extensions:

- More observations: Fit on previous World Cups, forecast out-of-sample.
- More covariates: Previous matches, team structure, economic indicators.
- More flexible models: GAM, random forests, boosting, . . .
- More flexible distributions: Bivariate, overdispersion, zero inflation.

Illustration: Goals in the 2018 FIFA World Cup

Possible extensions:

- More observations: Fit on previous World Cups, forecast out-of-sample.
- More covariates: Previous matches, team structure, economic indicators.
- More flexible models: GAM, random forests, boosting, ...
- More flexible distributions: Bivariate, overdispersion, zero inflation.

Here: Focus on goodness-of-fit assessment.

In particular: Graphical assessment of model calibration.

Goodness of fit: Scoring rules

Log-score: Log-likelihood; basis for information criteria and classical inference.

$\log f(y_i \mid \hat{\theta}_i)$

Goodness of fit: Scoring rules

Log-score: Log-likelihood; basis for information criteria and classical inference.

$$\log f(y_i \mid \hat{\theta}_i)$$

(Continuous) ranked probability score: Bounded alternative to log-score.

$$\int (F(z \mid \hat{\theta}_i) - 1(z \geq y_i))^2 \mathrm{d}z$$

Goodness of fit: Residuals

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

- Uniformly distributed if model correctly specified.
- Uniquely defined for continuous distributions.
- Otherwise consider uniform draw between $F(y_i - 1 \mid \hat{\theta}_i)$ and $F(y_i \mid \hat{\theta}_i)$.

Goodness of fit: Residuals

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

- Uniformly distributed if model correctly specified.
- Uniquely defined for continuous distributions.
- Otherwise consider uniform draw between $F(y_i - 1 \mid \hat{\theta}_i)$ and $F(y_i \mid \hat{\theta}_i)$.

(Randomized) quantile residuals: $\Phi^{-1}(u_i)$.

- Map to normal scale (from uniform).
- More similar to residuals in classical linear regression.
- More emphasis on deviations in the tails of the distribution.

Goodness of fit: Graphical assessment

Ideas:

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

Goodness of fit: Graphical assessment

Ideas:

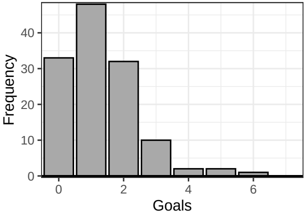
- Use visualizations instead of just summing up scores.
- Gain more insights graphically.
- Reveal different types of model misspecification.

Questions: Graphics are not new but novel unifying view.

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

Goodness of fit: Graphical assessment

Ideas: Illustrated for FIFA Poisson model.

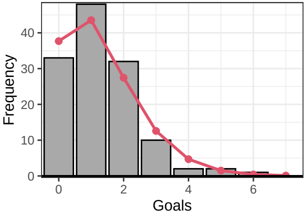


Marginal calibration:

- Observed frequencies.

Goodness of fit: Graphical assessment

Ideas: Illustrated for FIFA Poisson model.

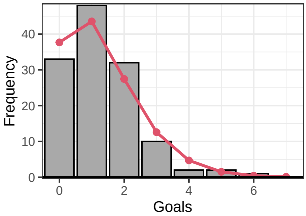


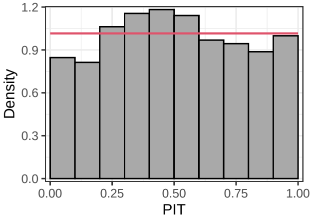
Marginal calibration:

- Observed frequencies.
- Compare: Expected.

Goodness of fit: Graphical assessment

Ideas: Illustrated for FIFA Poisson model.





Marginal calibration:

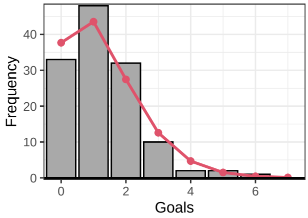
- Observed frequencies.
- Compare: Expected.

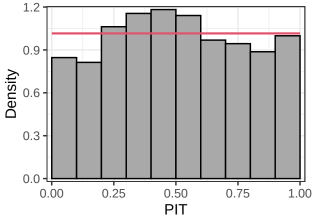
Probabilistic calibration:

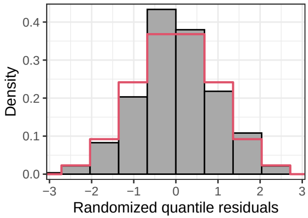
- Probability integral transform.
- Compare: Uniform.

Goodness of fit: Graphical assessment

Ideas: Illustrated for FIFA Poisson model.







Marginal calibration:

- Observed frequencies.
- Compare: Expected.

Probabilistic calibration:

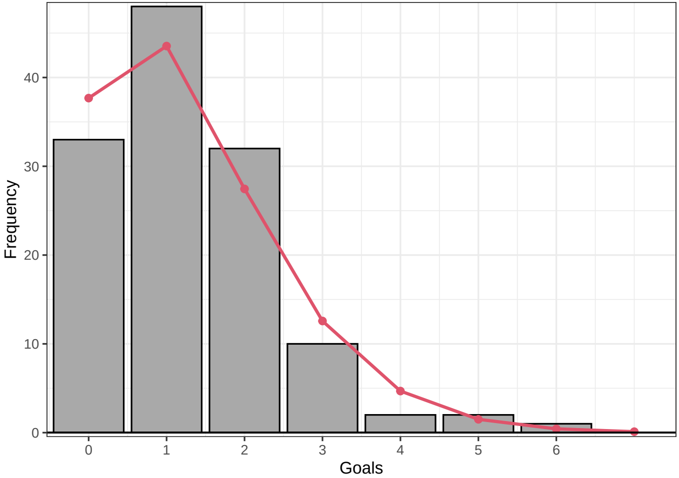
- Probability integral transform.
- Compare: Uniform.

Probabilistic calibration:

- (Randomized) quantile residuals.
- Compare: Normal

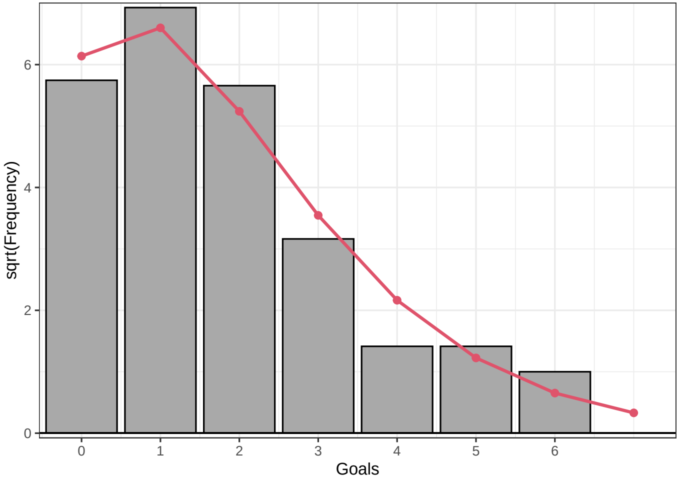
Goodness of fit: Marginal calibration

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



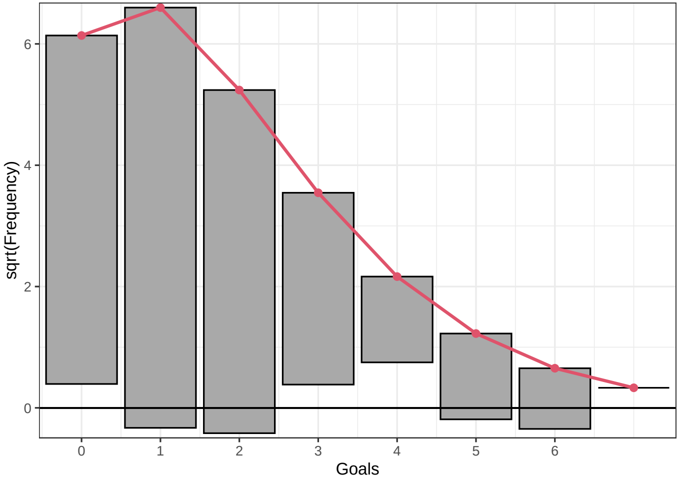
Goodness of fit: Marginal calibration

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



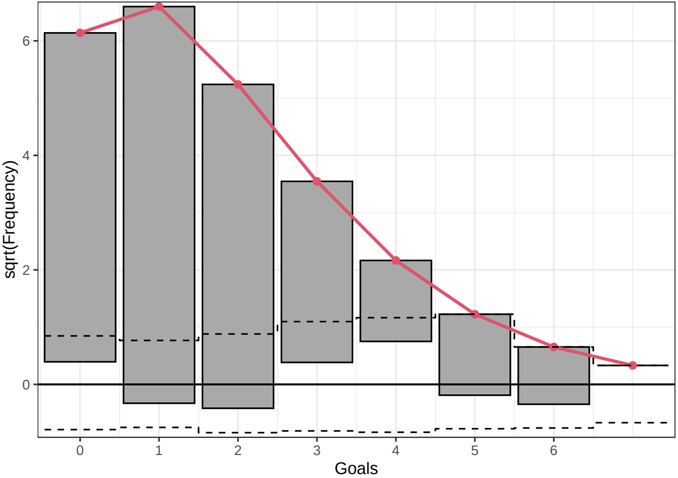
Goodness of fit: Marginal calibration

$\sqrt{\textbf{Observed}}$ **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:

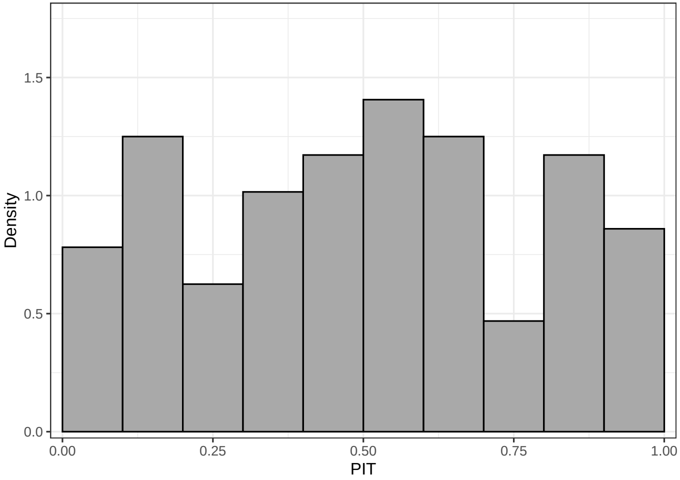
- Frequencies on raw or square-root scale.
- Hanging, standing, or suspended styled rootograms.

Overall:

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

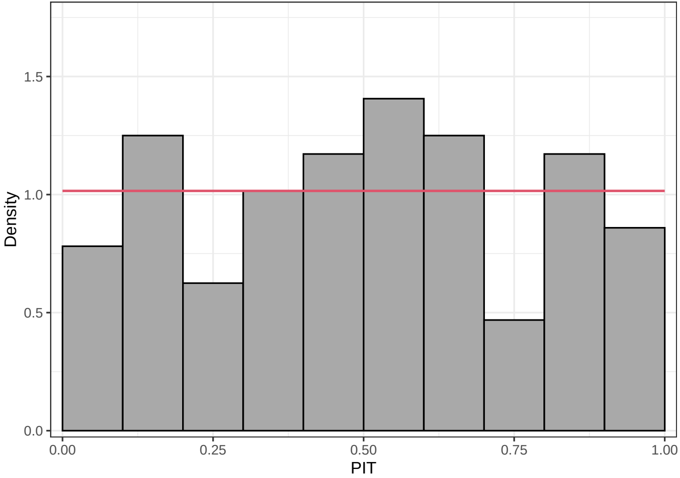
Goodness of fit: Probabilistic calibration

PIT: Randomization 1a.



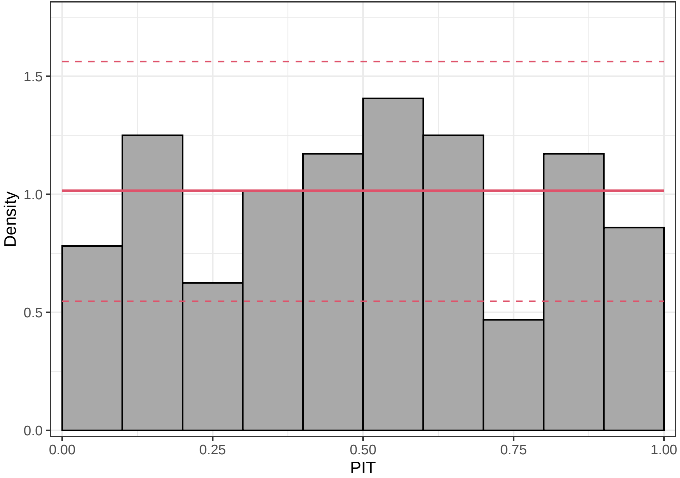
Goodness of fit: Probabilistic calibration

PIT: Randomization 1a, with reference line.



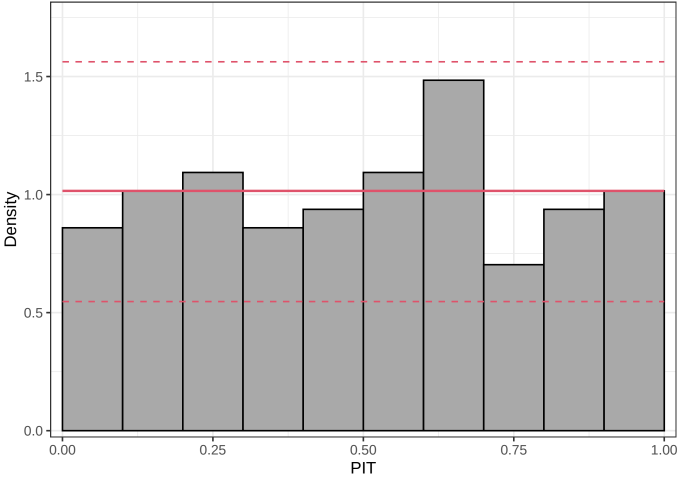
Goodness of fit: Probabilistic calibration

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



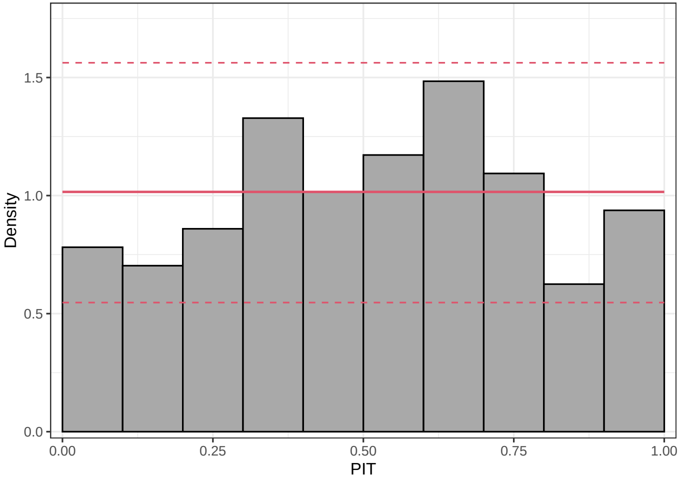
Goodness of fit: Probabilistic calibration

PIT: Randomization 1b.



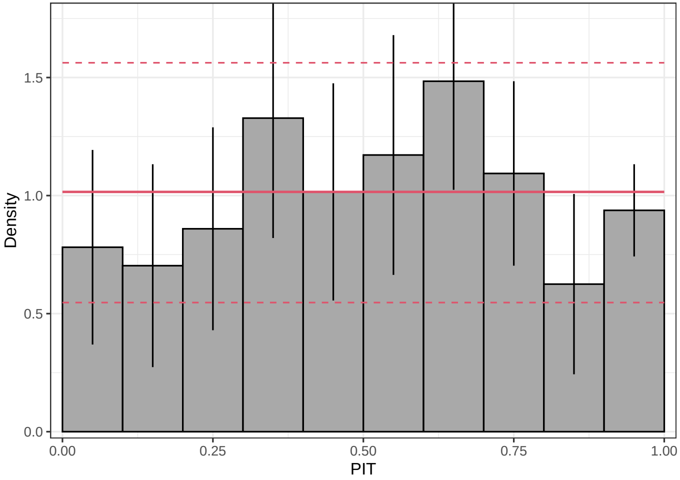
Goodness of fit: Probabilistic calibration

PIT: Randomization 1c.



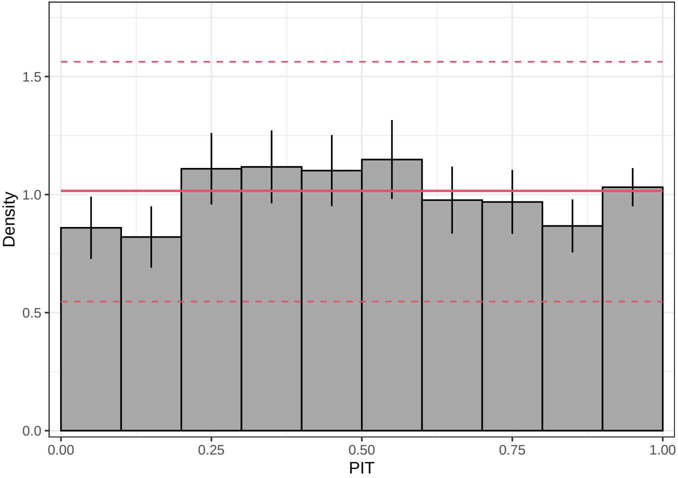
Goodness of fit: Probabilistic calibration

PIT: Randomization 1c, with simulation intervals.



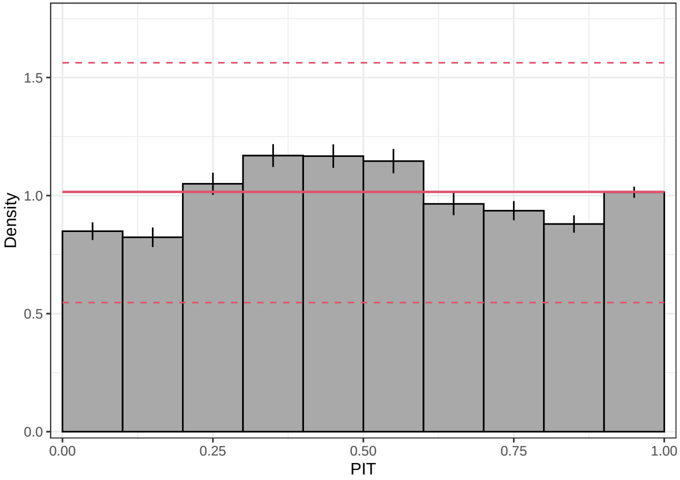
Goodness of fit: Probabilistic calibration

PIT: 10 random draws.



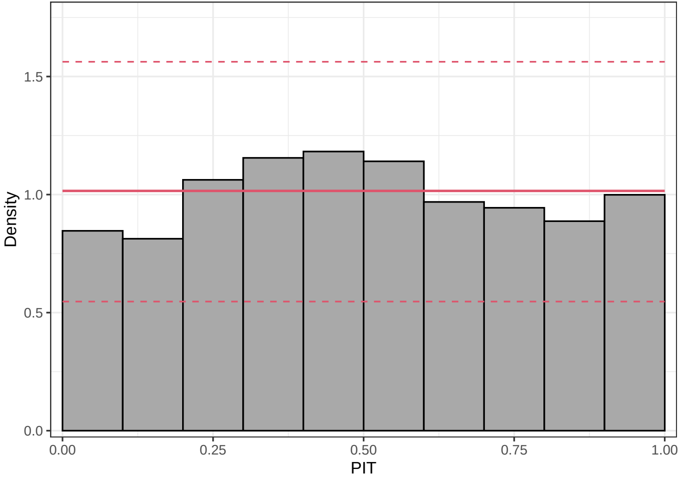
Goodness of fit: Probabilistic calibration

PIT: 100 random draws.



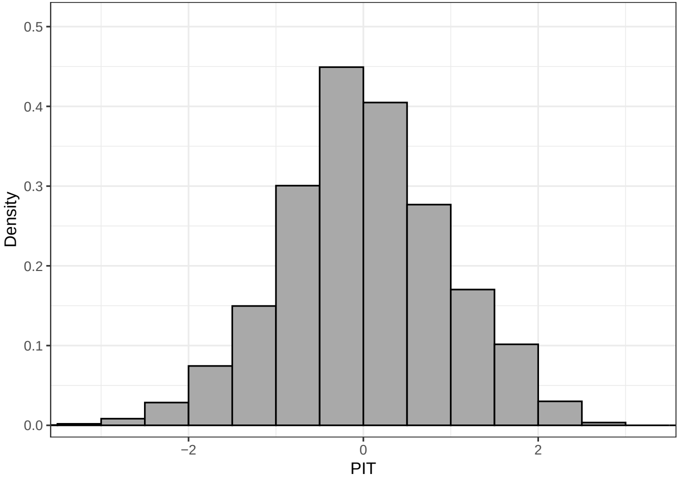
Goodness of fit: Probabilistic calibration

PIT: Expected.



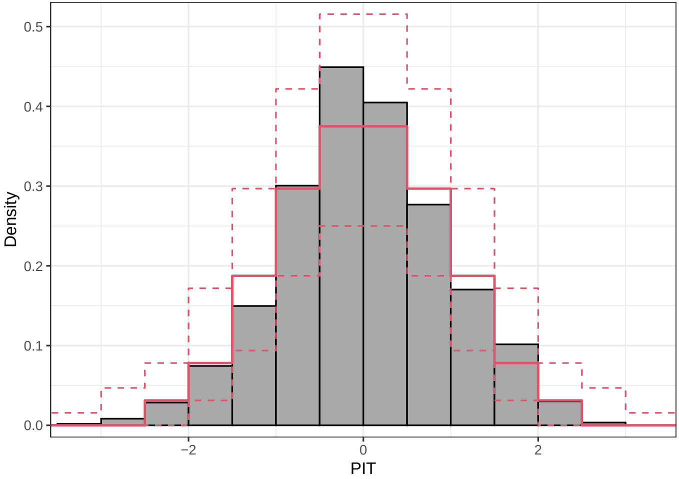
Goodness of fit: Probabilistic calibration

Randomized quantile residuals: Expected.



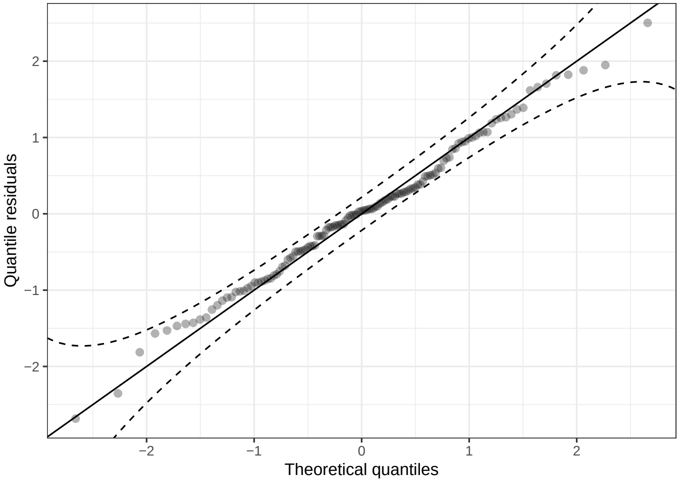
Goodness of fit: Probabilistic calibration

Randomized quantile residuals: Expected, with reference.



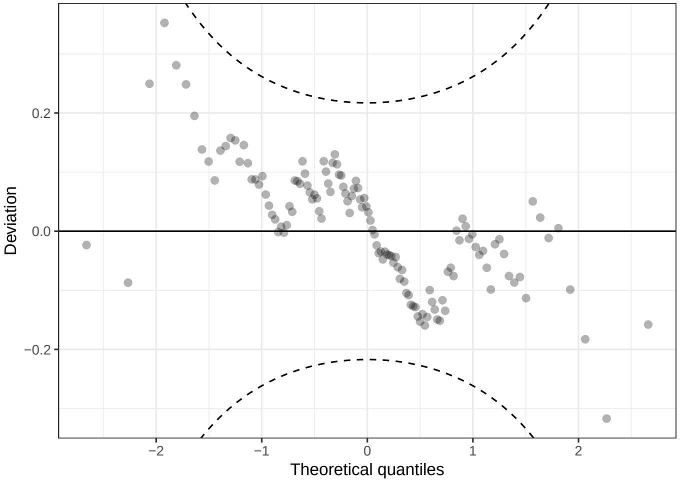
Goodness of fit: Probabilistic calibration

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: Precipitation in Innsbruck

Observation data:

- 3 day-accumulated precipitation amounts over 13 years (2000–2013).
- Observation station “Innsbruck” in Austria.

Illustration: Precipitation in Innsbruck

Observation data:

- 3 day-accumulated precipitation amounts over 13 years (2000–2013).
- Observation station “Innsbruck” in Austria.

Covariates:

- Ensemble mean and standard deviation of numerical precipitation forecasts.

Illustration: Precipitation in Innsbruck

Observation data:

- 3 day-accumulated precipitation amounts over 13 years (2000–2013).
- Observation station "Innsbruck" in Austria.

Covariates:

- Ensemble mean and standard deviation of numerical precipitation forecasts.

Model assumptions:

- Homoscedastic linear regression:
  $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{ensmean}_i, \quad \hat{\sigma} = \text{sd}(\epsilon)$
- Heteroscedastic censored regression with a logistic distribution assumption:
  $\text{y}_i \sim \text{Logistic}_0\big(\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{ensmean}_i, \, \hat{\sigma}_i = \exp(\hat{\gamma}_0 + \hat{\gamma}_1 \cdot \text{enssd}_i)\big)$

Illustration: Precipitation in Innsbruck

Data: Observations and numerical ensemble mean.

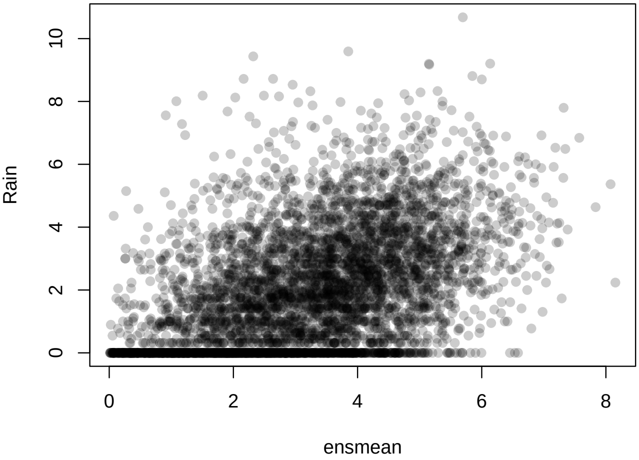


Illustration: Precipitation in Innsbruck

Data: Observations and numerical ensemble mean.

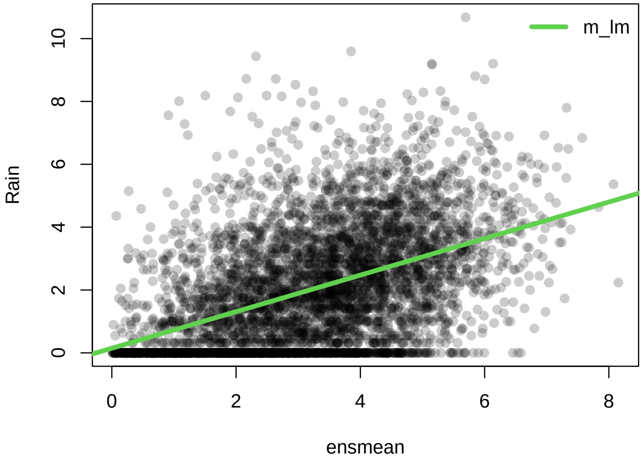


Illustration: Precipitation in Innsbruck

Data: Observations and numerical ensemble mean.

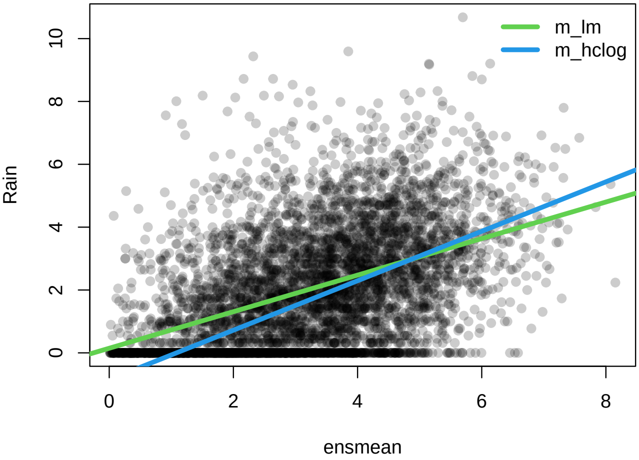


Illustration: Precipitation in Innsbruck

Rootogram:

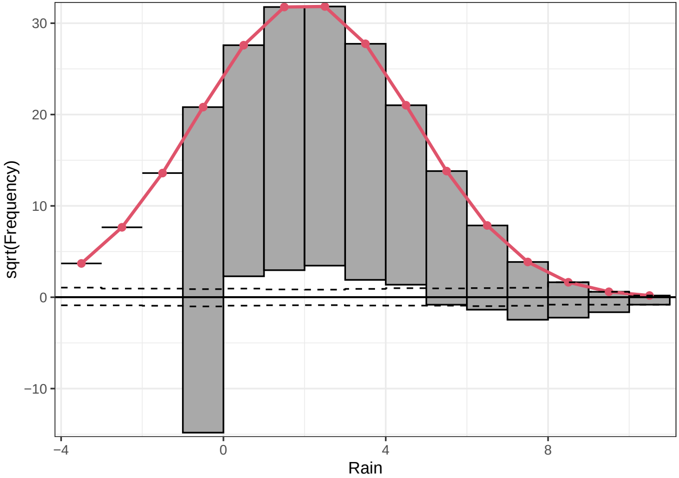


Illustration: Precipitation in Innsbruck

Rootogram:

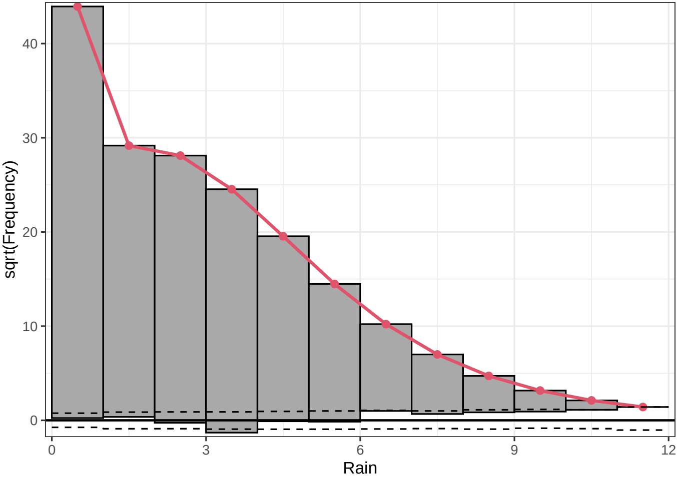


Illustration: Precipitation in Innsbruck

PIT histogram:

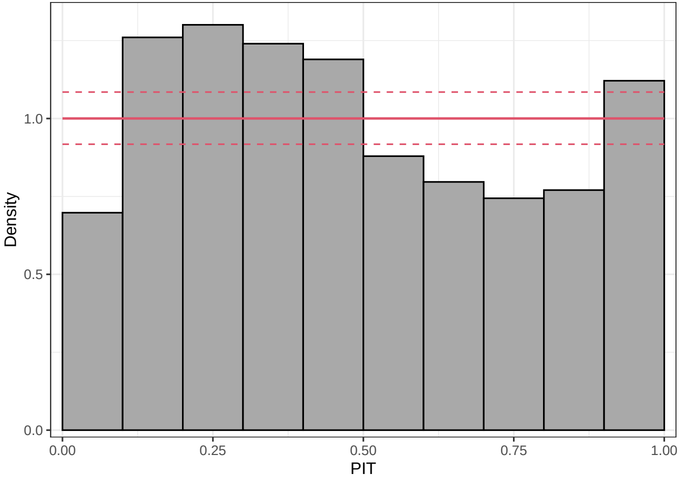


Illustration: Precipitation in Innsbruck

PIT histogram:

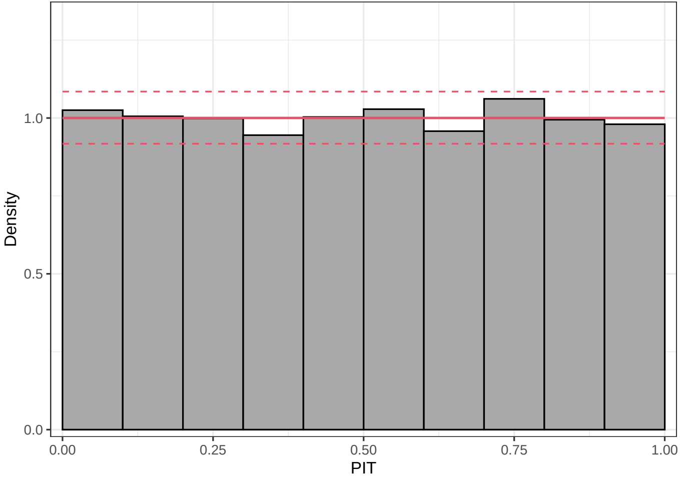


Illustration: Precipitation in Innsbruck

PIT histogram:

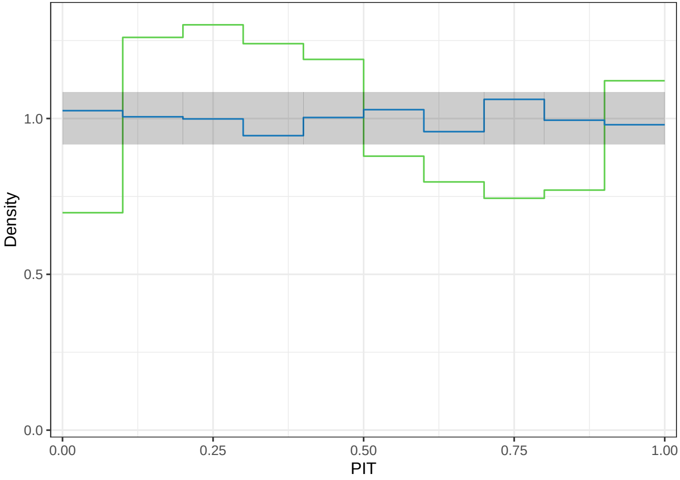


Illustration: Precipitation in Innsbruck

Q-Q residual plot:

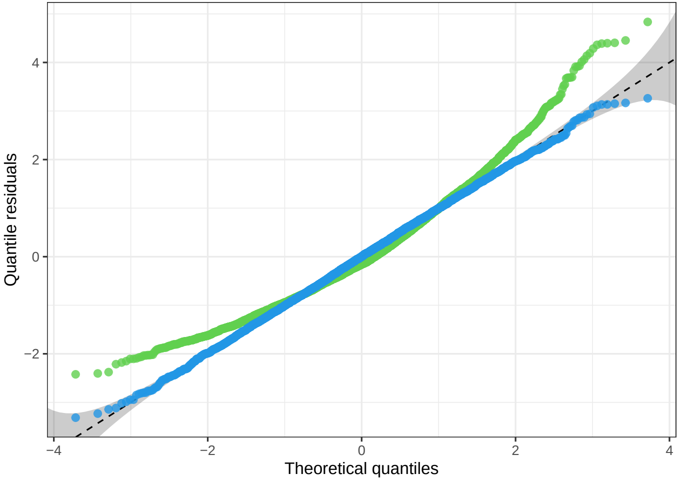
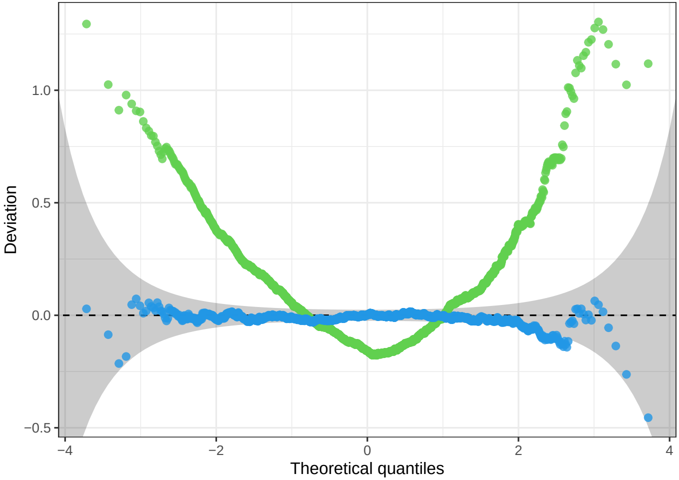


Illustration: Precipitation in Innsbruck

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.

Software: topmodels & distributions3

Probabilistic forecasts:

```
R> p <- procast(m)
R> head(p, 3)
                            distribution
1 Poisson distribution (lambda = 1.7680)
2 Poisson distribution (lambda = 0.8655)
3 Poisson distribution (lambda = 1.0297)
```

Software: topmodels & distributions3

Probabilistic forecasts:

```
R> p <- procast(m)
R> head(p, 3)
                              distribution
1 Poisson distribution (lambda = 1.7680)
2 Poisson distribution (lambda = 0.8655)
3 Poisson distribution (lambda = 1.0297)
```

For final:

```
R> p_final <- tail(p$distribution, 2)
R> pdf(p_final, 0:4)
       d_0    d_1    d_2     d_3     d_4
127 0.2010 0.3225 0.2587 0.13836 0.05550
128 0.3853 0.3675 0.1752 0.05572 0.01329
```

Software: topmodels & distributions3

Probabilistic forecasts:

```
R> p <- procast(m)
R> head(p, 3)
                          distribution
1 Poisson distribution (lambda = 1.7680)
2 Poisson distribution (lambda = 0.8655)
3 Poisson distribution (lambda = 1.0297)
```

For final:

```
R> p_final <- tail(p$distribution, 2)
R> pdf(p_final, 0:4)
       d_0    d_1    d_2     d_3     d_4
127 0.2010 0.3225 0.2587 0.13836 0.05550
128 0.3853 0.3675 0.1752 0.05572 0.01329
```

Scoring rules:

```
R> proscore(m, type = c("LogS", "CRPS", "MSE"), aggregate = TRUE)
    LogS  CRPS   MSE
1 -1.388 0.562 1.162
```

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`.

Messner JW, Mayr GJ, Zeileis A (2016). “Heteroscedastic Censored and Truncated Regression with crch.” *The R Journal.*, **8**(1), 173–181. `doi:10.32614/RJ-2016-012`

Contact

Mastodon: @zeileis@fosstodon.org

X/Twitter: @AchimZeileis

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