universität
innsbruck

Graphical Model Assessment of Probabilistic Forecasts

Moritz N. Lang, Reto Stauffer, Achim Zeileis

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

Motivation

Probabilistic regression models:

- Modelling the entire probability distribution rather than just the expectation.
- Various model classes and types.

Motivation

Probabilistic regression models:

- Modelling the entire probability distribution rather than just the expectation.
- Various model classes and types.

Goodness of fit:

- Scoring rules for evaluating the predictive performance, e.g., using the log-score or the (continuous) ranked probability score.
- Visualizations especially suitable for identifying possible misspecifications.

Motivation

Probabilistic regression models:

- Modelling the entire probability distribution rather than just the expectation.
- Various model classes and types.

Goodness of fit:

- Scoring rules for evaluating the predictive performance, e.g., using the log-score or the (continuous) ranked probability score.
- Visualizations especially suitable for identifying possible misspecifications.

$\Rightarrow$ **What are useful elements of such graphics?**

Motivation

Probabilistic regression models:

- Modelling the entire probability distribution rather than just the expectation.
- Various model classes and types.

Goodness of fit:

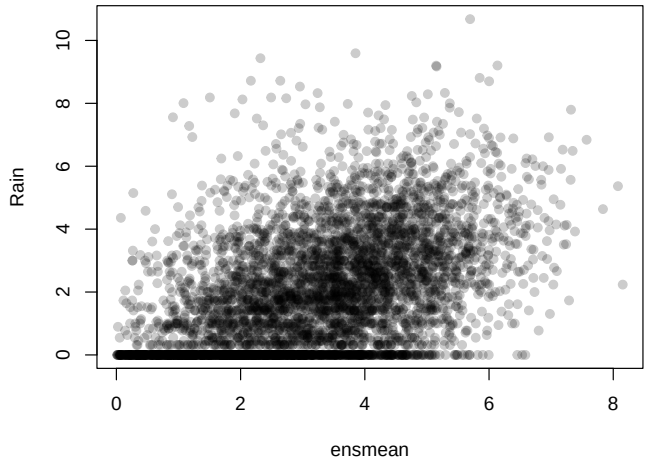
- Scoring rules for evaluating the predictive performance, e.g., using the log-score or the (continuous) ranked probability score.
- Visualizations especially suitable for identifying possible misspecifications.

⇒ **What are useful elements of such graphics?**

⇒ **What are relative (dis)advantages?**

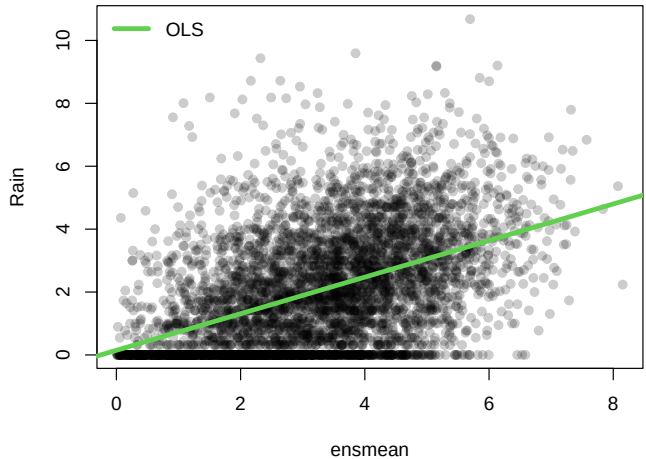
Probabilistic precipitation forecasting

Observed vs. ensmean:



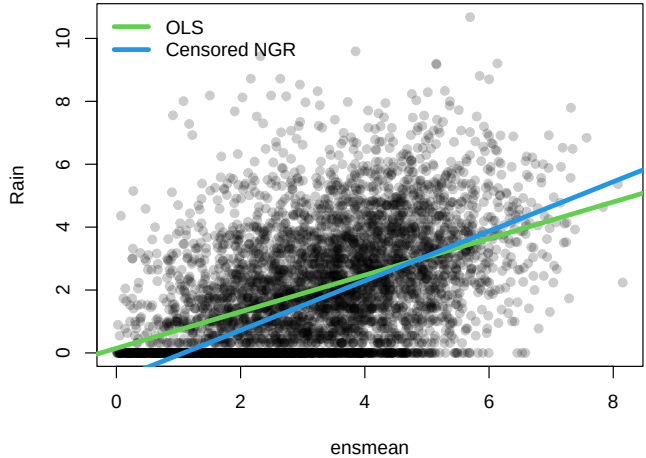
Probabilistic precipitation forecasting

Observed vs. ensmean:



Probabilistic precipitation forecasting

Observed vs. ensmean:

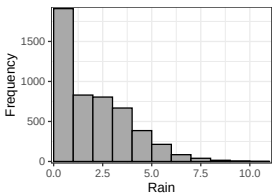


Graphical assessment

However: Is the model calibrated?

Graphical assessment

However: Is the model calibrated?

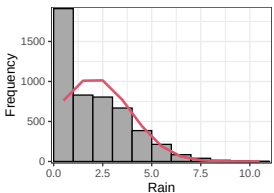


Marginal calibration:

- Observed frequencies.

Graphical assessment

However: Is the model calibrated?

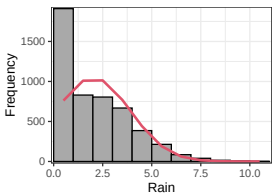


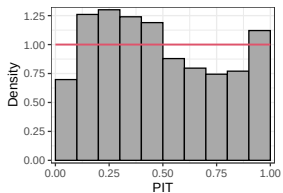
Marginal calibration:

- Observed frequencies.
- Compare: Expected.

Graphical assessment

However: Is the model calibrated?





Marginal calibration:

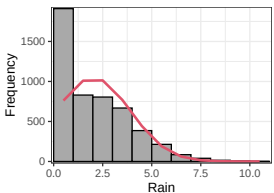
- Observed frequencies.
- Compare: Expected.

Probabilistic calibration:

- PIT residuals: $u_i = F(y_i | \hat{\theta}_i)$.
- Compare: Uniform.

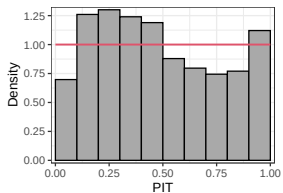
Graphical assessment

However: Is the model calibrated?



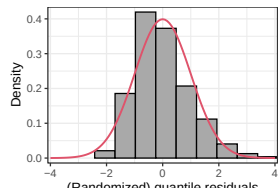
Marginal calibration:

- Observed frequencies.
- Compare: Expected.



Probabilistic calibration:

- PIT residuals: $u_i = F(y_i|\,\hat{\theta}_i)$.
- Compare: Uniform.

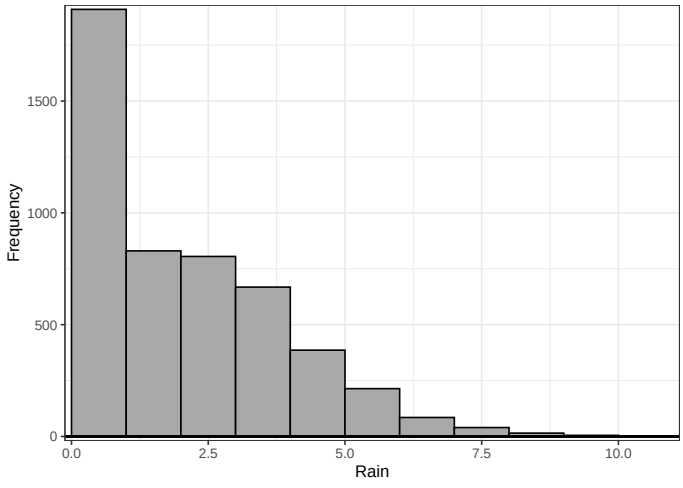


Probabilistic calibration:

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

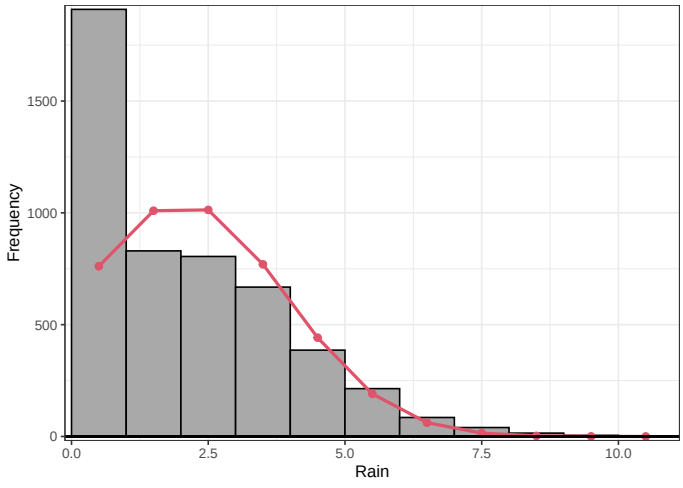
Marginal calibration

Frequencies: Observed



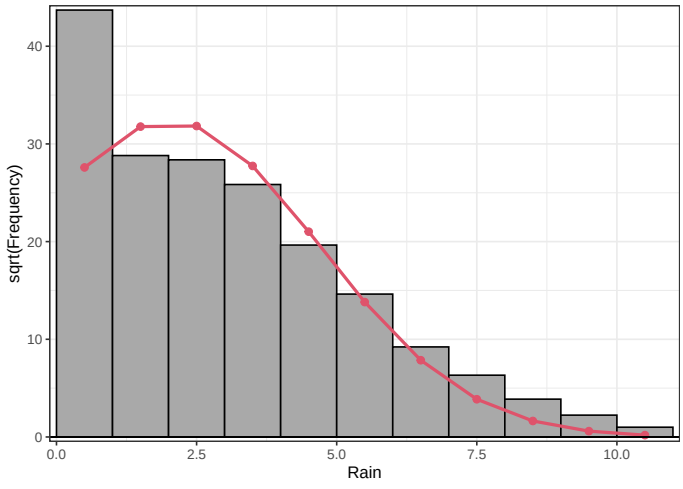
Marginal calibration

Frequencies: Observed vs. expected



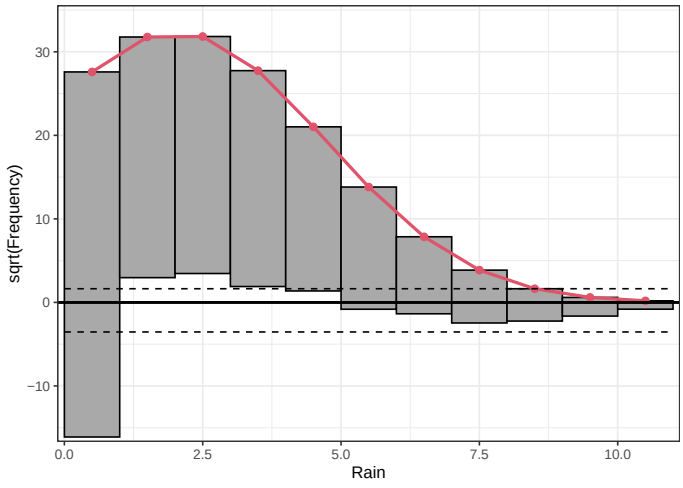
Marginal calibration

Frequencies: $\sqrt{\text{Observed}}$ vs. $\sqrt{\text{expected}}$



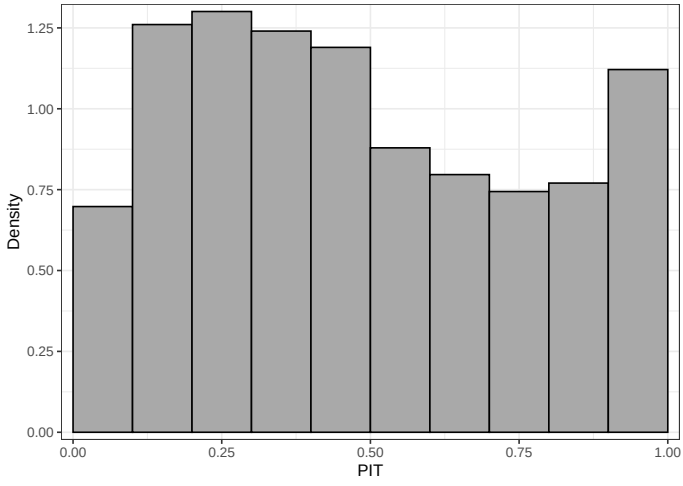
Marginal calibration

Frequencies: $\sqrt{\text{Observed}}$ vs. $\sqrt{\text{expected}}$



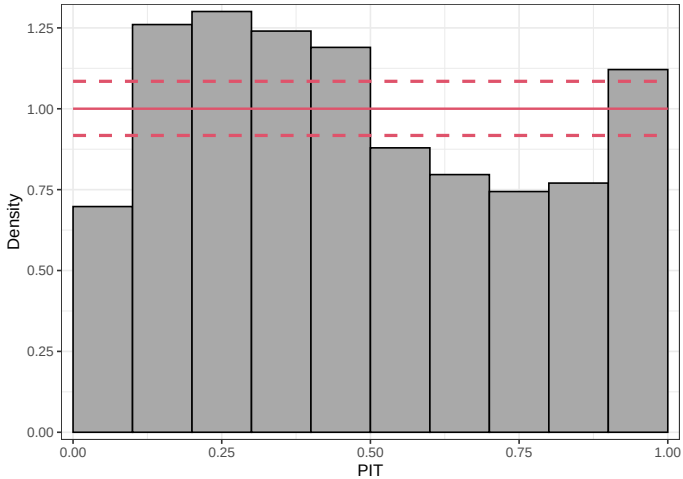
Probabilistic calibration

PIT residuals:



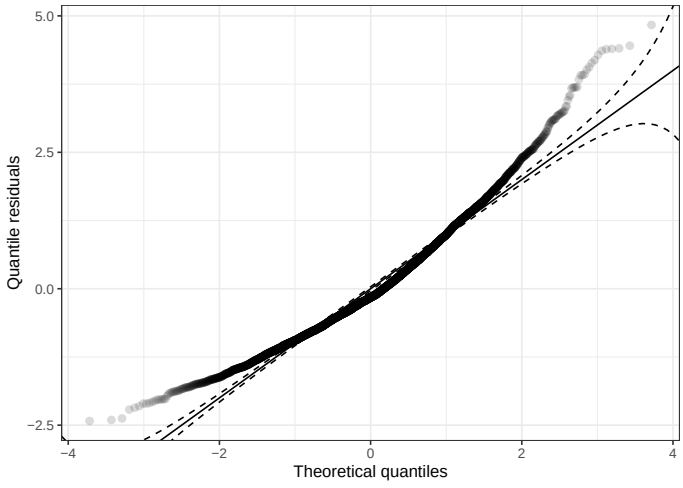
Probabilistic calibration

PIT residuals:



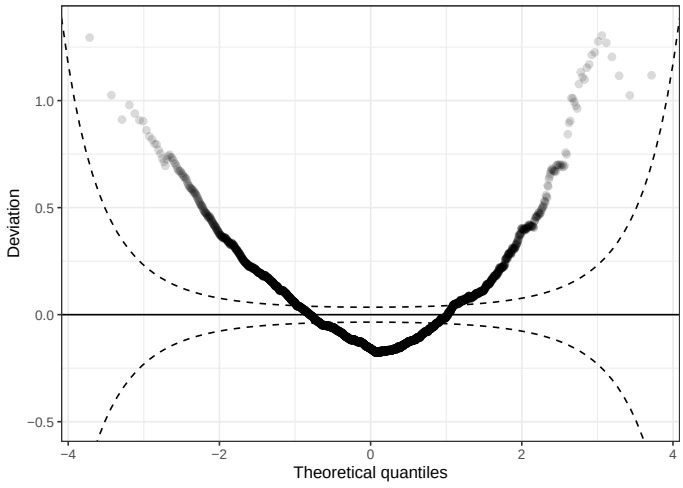
Probabilistic calibration

Quantile residuals: Observed vs. expected



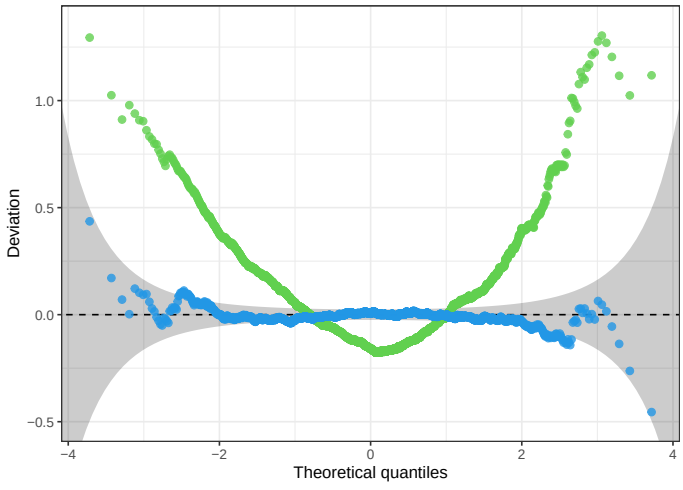
Probabilistic calibration

Quantile residuals: Deviations



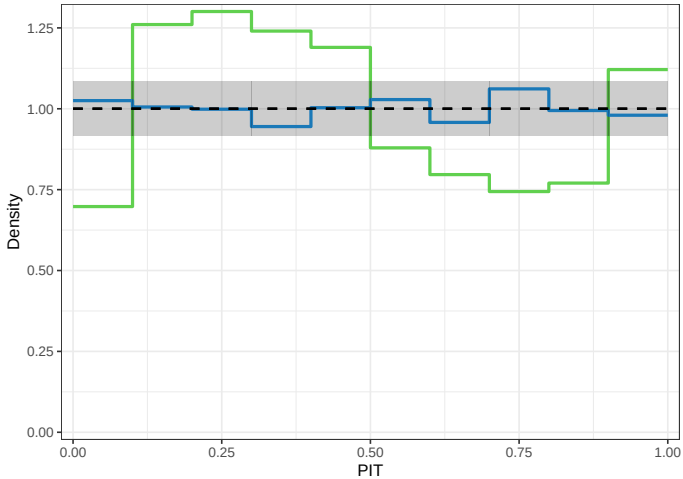
Model comparison

Quantile residuals: Deviations



Model comparison

PIT residuals:



Summary

Graphical assessments: Various possibilities suggested in different parts of the literature.

- Rootogram.
- Probability integral transform (PIT) histogram.
- (Randomized) quantile-quantile residuals plot.
- Detrended Q-Q residuals plot or worm plot.
- Reliability diagram at prespecified thresholds.

Summary

Graphical assessments: Various possibilities suggested in different parts of the literature.

- Rootogram.
- Probability integral transform (PIT) histogram.
- (Randomized) quantile-quantile residuals plot.
- Detrended Q-Q residuals plot or worm plot.
- Reliability diagram at prespecified thresholds.

topmodels: Unifying toolbox for graphical model assessment.

- available on R-Forge at `https://topmodels.R-Forge.R-project.org/`

References

Lang MN, Zeileis A *et al.* (2021). “topmodels: Infrastructure for Inference and Forecasting in Probabilistic Models.” *R package version 0.2-0*. https://topmodels.R-Forge.R-project.org/

Dunn PK, Smyth GK (1996). “Randomized Quantile Residuals.” *Journal of Computational and Graphical Statistics*, **5**(3), 236–244. doi:10.2307/1390802

Gneiting T, Balabdaoui F, Raftery AE (2007) “Probabilistic Forecasts, Calibration and Sharpness.” *Journal of the Royal Statistical Society: Series B (Methodological)*, **69**(2), 243–268. doi:10.1111/j.1467-9868.2007.00587.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

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

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

moritz.lang@uibk.ac.at MoritzNLang