Uncertainty Quantification

How bad (or good) is your model, really?

C. Johnstone

Air Force Operations Research Symposium

April 14, 2022

All materials used in this tutorial can be found here.

Uncertainty Quantification

- Machine learning algorithms emphasize point prediction (and classification) over inference
- A point prediction alone does not give information about the uncertainty of said prediction (or observation)

Interval Prediction

- Given some set of observations $D_n = \{(x_i, y_i)\}_{i=1}^n$, we want to know something about some future response y_{n+1}
- In a regression setting, we might want to generate a point prediction for y_{n+1}, say, $\hat{y}_{n+1}$
- In order to attach probability to prediction, we could also generate an interval $C_{1-\alpha}$ such that,

$$P(y_{n+1} \in C_{1-\alpha}(x)) \geq 1 - \alpha$$

Prediction Intervals with Linear Regression

Suppose we now have observations $\{(x_i, y_i)\}_{i=1}^{n}$, where x_i is an covariate vector of length d and

$$y_i = x_i'\beta + \epsilon_i,$$

where β is the vector of true parameters and ϵ_i is a $N(0, \sigma^2)$ error term associated with y_i.

Prediction Intervals with Linear Regression

For some new observation x_{n+1}, a $100(1-\alpha)\%$ prediction interval for y_{n+1} is,

$$\hat{y}_{n+1} \pm z_{\alpha/2}\hat{\sigma}\sqrt{1 + x'_{n+1}(X'X)^{-}x_{n+1}},$$

where $\hat{y}_{n+1}$ is the least-squares estimate $x'_{n+1}\hat{\beta}$ and X is a $n \times p$ matrix where row $i = x_i$

Prediction Intervals with Smoothing Splines

When a prediction is generated with a smoothing spline [1], i.e.,

$$\hat{y}_{n+1} = \sum_{i=1}^{m} \hat{\beta}_i g_i(x_{n+1}),$$

where $g_i(\cdot)$ is the i-th basis function, $\hat{\beta} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_m\right)'$ is the minimizer of

$$||y - G\beta||_2^2 + \lambda \beta' \Omega \beta,$$

$g(x) = \left(g_1(x), \ldots, g_m(x)\right)'$, $G = \{g(x_i)\}_{i=1}^{n}$, λ is the smoothing parameter and Ω a penalty matrix, then a $100(1-\alpha)\%$ prediction interval is of the form

$$\hat{y}_{n+1} \pm z_{\alpha/2} \hat{\sigma} \sqrt{1 + g(x_{n+1})'(G'G + \lambda\Omega)^{-} g(x_{n+1})}.$$

---

[1] https://www.stat.cmu.edu/ ryantibs/advmethods/notes/smoothspline.pdf
We still assume normality in this case.

Question of Interest #1

Can we eliminate some of the assumptions and still get valid prediction intervals?

Conformal Prediction Intervals

Conformal prediction[3] allows us to generate finite sample valid (conservative) prediction sets using **any** prediction method by repeatedly testing

$$
\begin{aligned}
H_0 &: y_{n+1} = y_c \\
H_a &: y_{n+1} \neq y_c,
\end{aligned}
$$

where y_c is some candidate value for y_{n+1}.

[3] Vovk et al. (2005)

Conformal Prediction Intervals

Utilizes *conformity scores* and *typicalness functions* to determine intervals, e.g.,

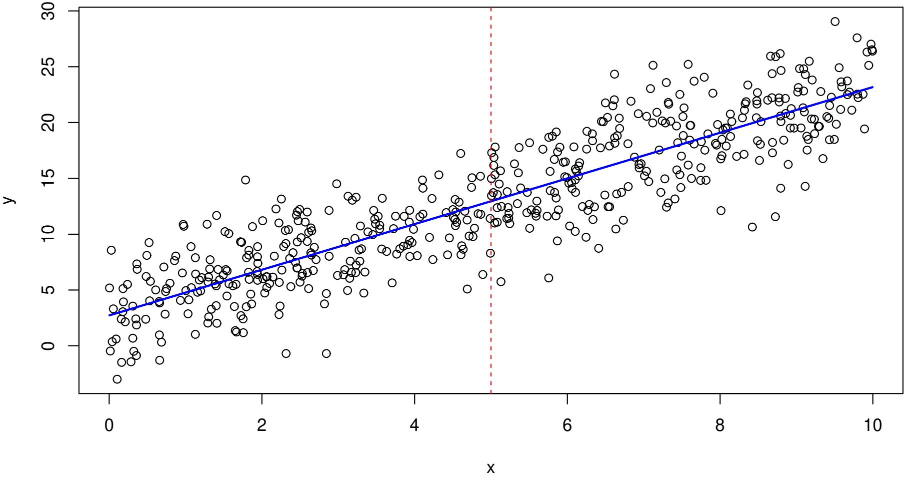
$$R_i(y_c) \equiv r_i(y_c) = \big|y_i - \hat{y}_i(y_c)\big|,$$

where $\hat{y}_i(y_c)$ is the prediction for y_i trained on an augmented data set $\{(x_1, y_1), \dots, (x_n, y_n), (x_{n+1}, y_c)\}$.

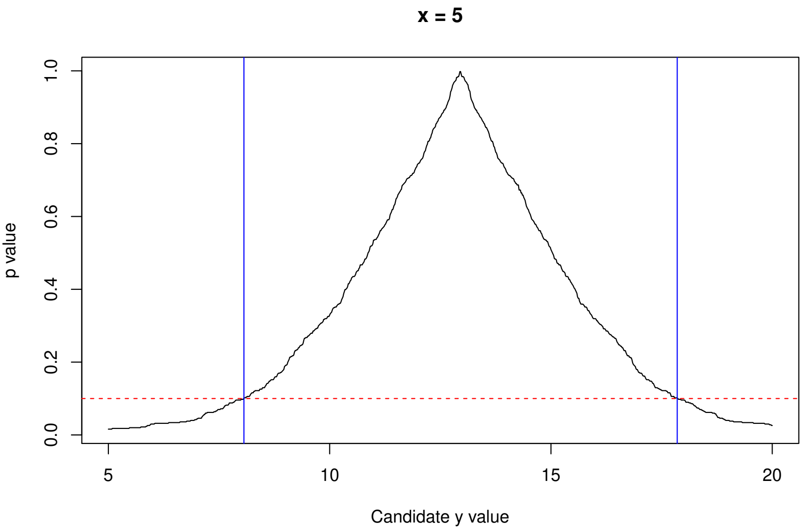
$$\pi(y_c) = \frac{1}{n+1} + \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{I}\{r_i(y_c) \le r_{n+1}(y_c)\}$$

$$C_{1-\alpha}^{conf}(x) = \{y_c \in \mathbb{R} \ : \ (n+1)\pi(y_c) \le \lceil (1-\alpha)(n+1) \rceil\}.$$

Conformal Prediction Intervals



Conformal Prediction Intervals



Split-Conformal Prediction Intervals

- Requires fewer computations, i.e., fewer model retrains
- Same guarantees as conformal prediction[4]

[4] Lei et al. (2018)

Split-Conformal Prediction Intervals

- Partition data set $\{(x_i, y_i)\}_{i=1}^{n}$ into $\mathcal{I}_1$ and $\mathcal{I}_2$
- Train predictor with observations in $\mathcal{I}_1$
- Generate conformity scores for observations in $\mathcal{I}_2$ using predictor trained with $\mathcal{I}_1$
- Prediction interval constructed using conformity scores associated with $\mathcal{I}_2$ rather than an augmented data set

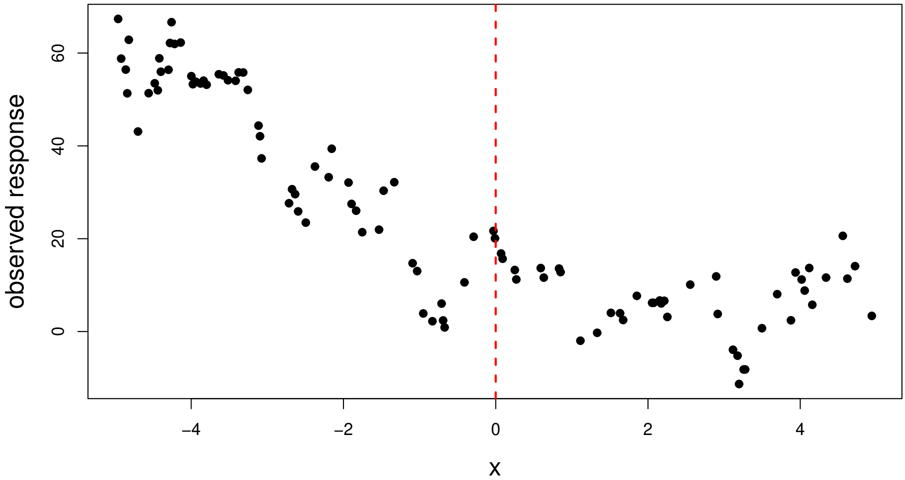
What are some potential drawbacks of the split-conformal approach?

Alternative Typicalness Function

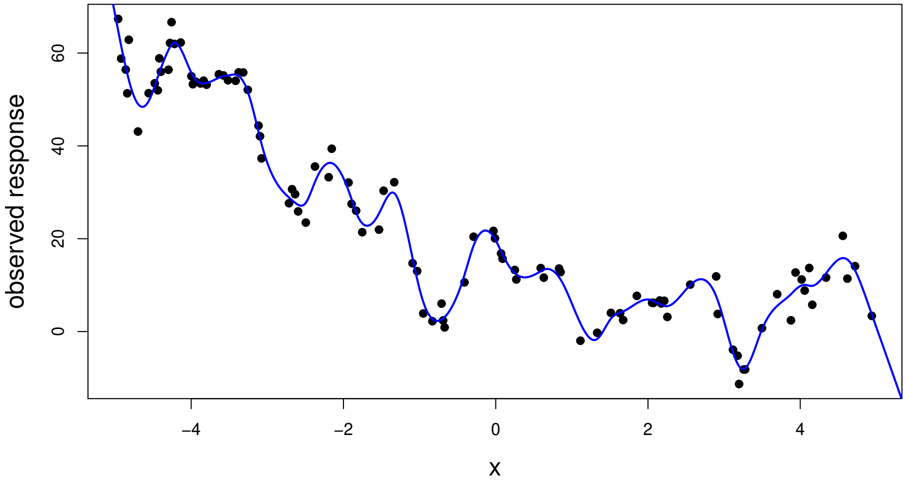
- We can use any *measureable* function as a typicalness function, e.g., distance measures, k-nearest neighbors.
- We can also use *kernel density* estimators as our typicalness function.

$$R_i(y_c) = \left[\frac{1}{n+1} \sum_{i=1}^{n+1} \frac{1}{h} K\left(\frac{r_{n+1}(y_c) - r_i(y_c)}{h} \right) \right]^{-1}$$

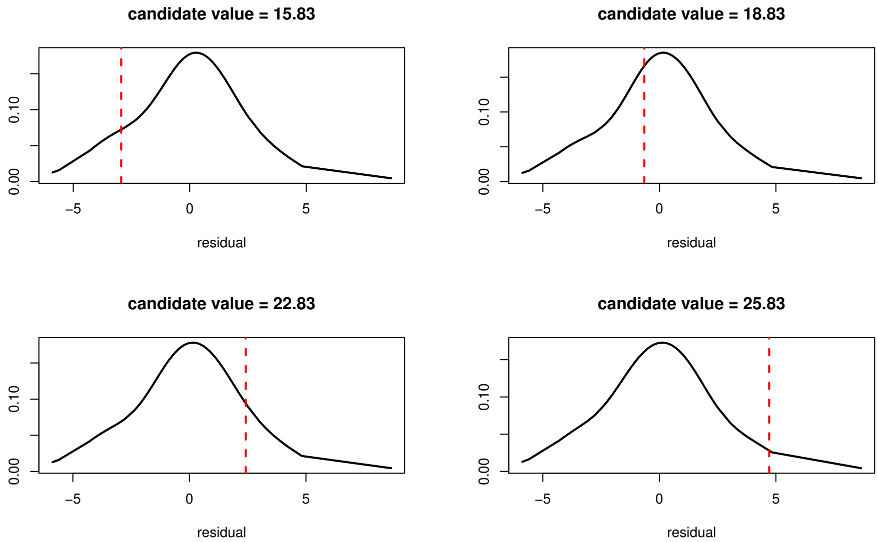
Data Example



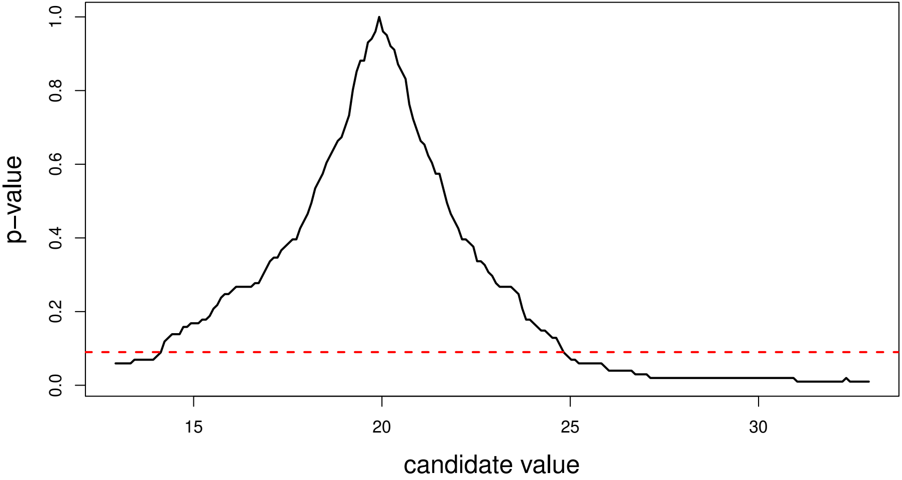
Data Example Smoothing Spline



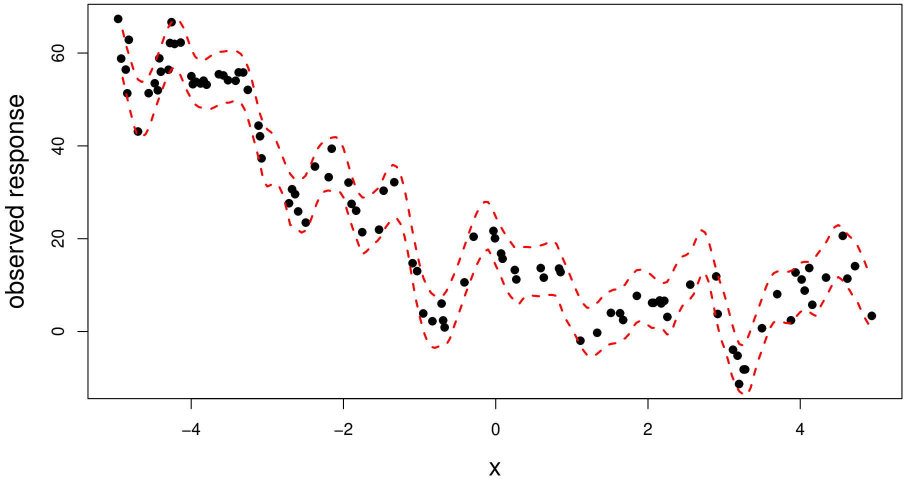
KDE for candidate values when $x = 0$



p-values for candidate values when $x = 0$



Prediction Intervals



Classification Example

Figure 1: Prediction sets for “fox squirrel” exemplar images (Angelopoulos et al., 2020)

Conformal Inference Classification Walkthrough

We will now walk through an example of conformal inference for classification using some data from the CogPilot Data Challenge.

Extensions of Conformal Inference

- Mondrian conformal inference (Boström and Johansson, 2020)
- Conformal inference under covariate shift (Tibshirani et al., 2019)
- Conformal uncertainty sets for robust optimization (Johnstone and Cox, 2021)
- Conformal inference with dependent data (Chernozhukov et al., 2018)
- Conformal inference for classification (Angelopoulos et al., 2020)
- Stable Conformal Prediction Sets (Ndiaye, 2021)

▸ awesome conformal prediction

References

Angelopoulos, A., Bates, S., Malik, J., and Jordan, M. I. (2020). Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*.

Boström, H. and Johansson, U. (2020). Mondrian conformal regressors. In *Conformal and Probabilistic Prediction and Applications*, pages 114–133. PMLR.

Chernozhukov, V., Wüthrich, K., and Yinchu, Z. (2018). Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference On Learning Theory*, pages 732–749. PMLR.

Johnstone, C. and Cox, B. (2021). Conformal uncertainty sets for robust optimization. In *Conformal and Probabilistic Prediction and Applications*, pages 72–90. PMLR.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Ndiaye, E. (2021). Stable conformal prediction sets. *arXiv preprint arXiv:2112.10224*.

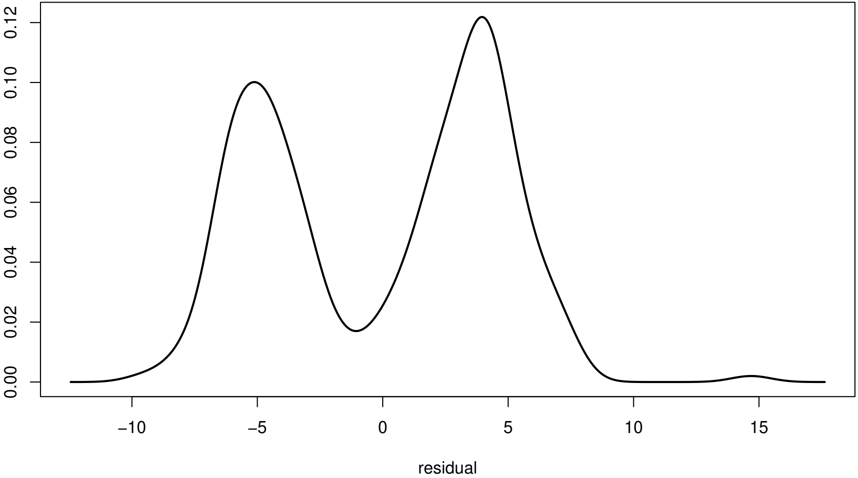
Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

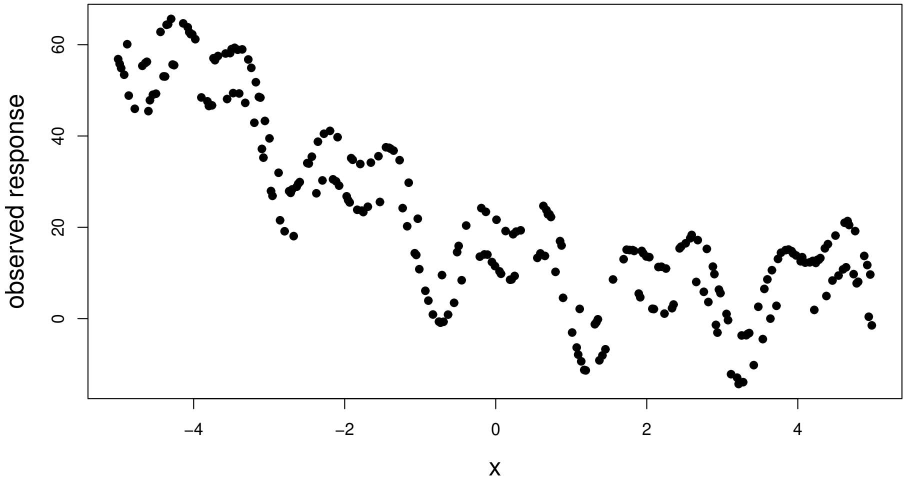
Question of Interest #2

What if we have some "funky" error distribution?

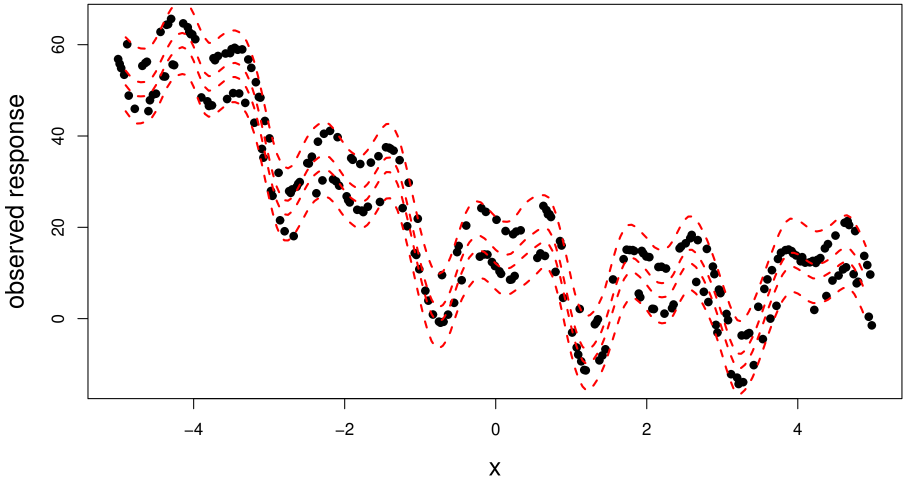
Funky Error KDE



Funky Data Example



Funky Prediction Intervals



Question of Interest #3

What are some situations where conformal inference performs "poorly"?

Conformal Inference Guarantees

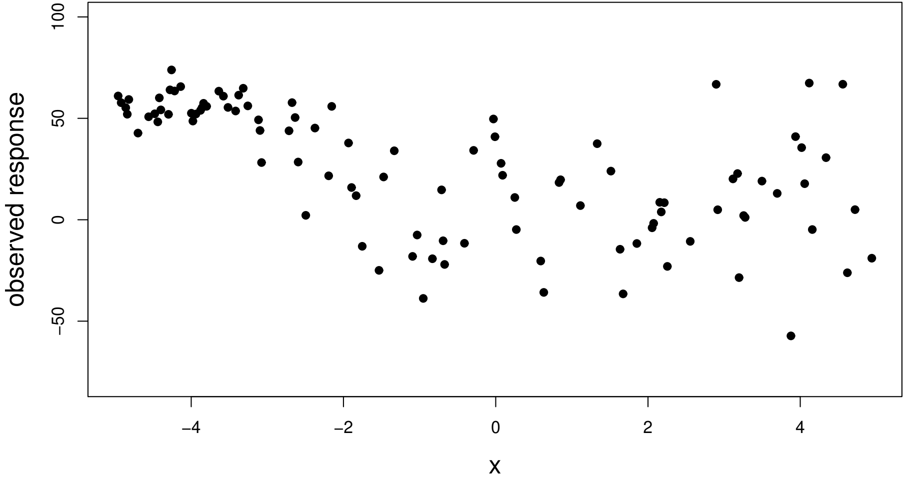
- Under *exchangeability*, prediction intervals generted with conformal inference guarantee *marginal* coverage, i.e.,

$$P(y \in C_{1-\alpha}(x)) \geq 1 - \alpha$$

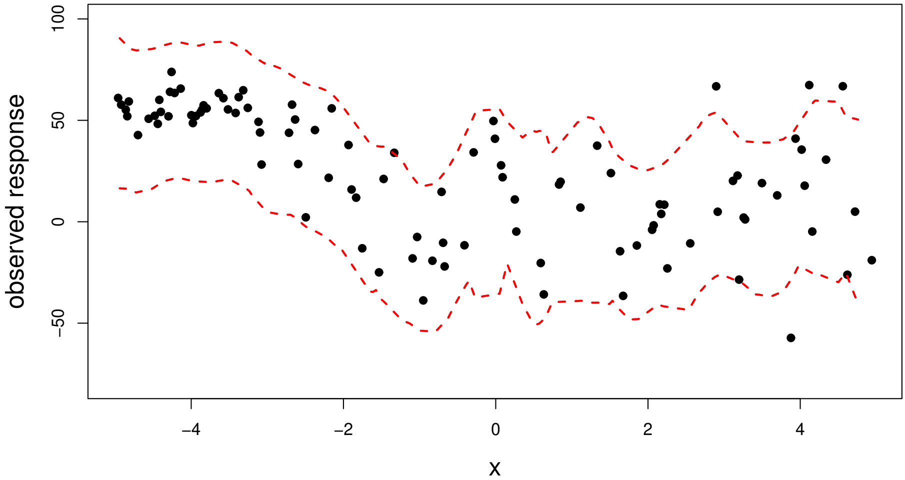
- It does not guarantee conditional coverage, i.e.,

$$P(y \in C_{1-\alpha}(x) | X = x) \geq 1 - \alpha$$

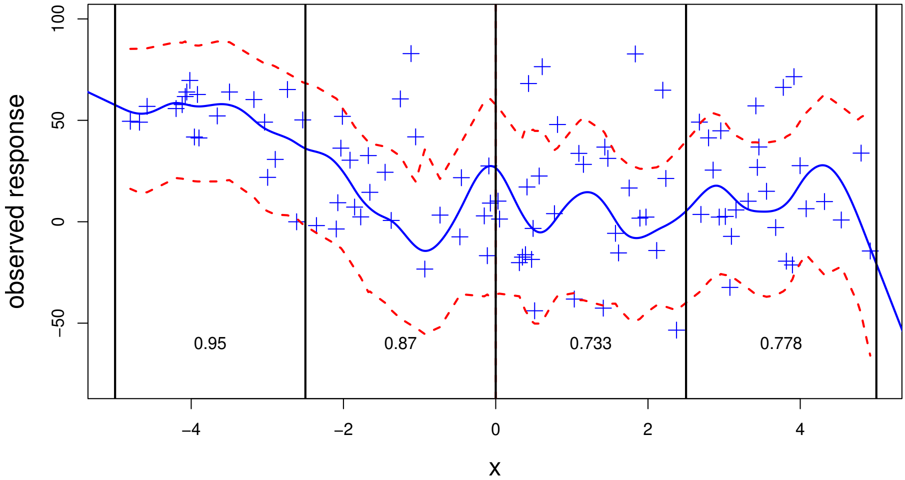
Non-constant Variance Data Example



Non-constant Variance Prediction Interval



Conditional Coverage on Test Data



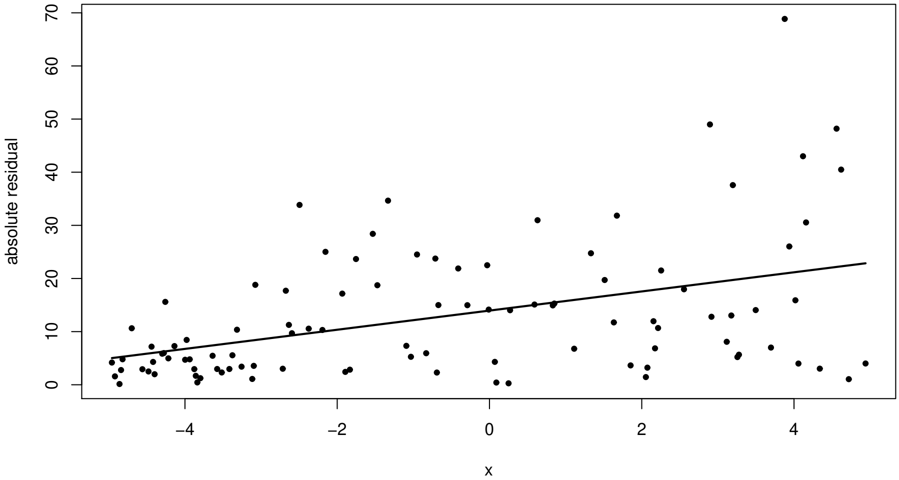
Locally-Weighted Conformal Inference

- We can get *better* conditional coverage if we localize our typicalness score based on some measure of spread at x.
- We can adjust our typicalness score to

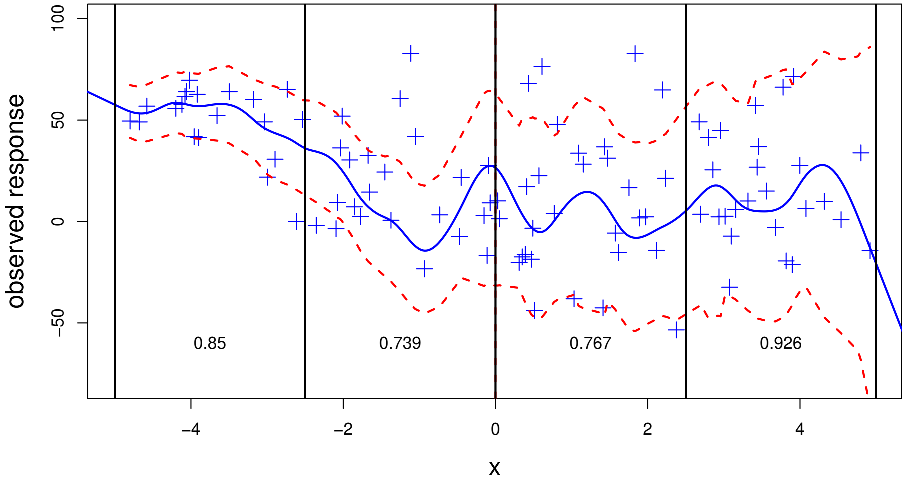
$$R_i = \frac{r_i(y_c)}{\hat{\rho}(x_i)}$$

where $\hat{\rho}(x)$ is the estimated *mean-absolute deviation* at x and $r_i(y_c)$ is the absolute residual.

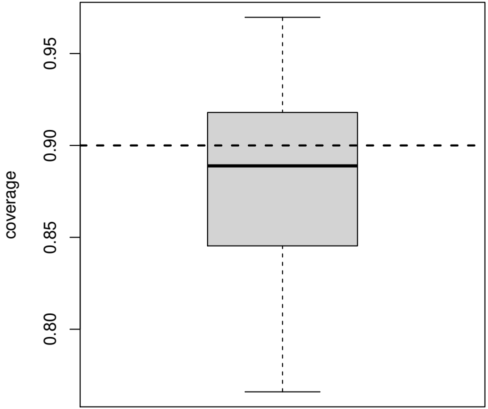
Estimating MAD



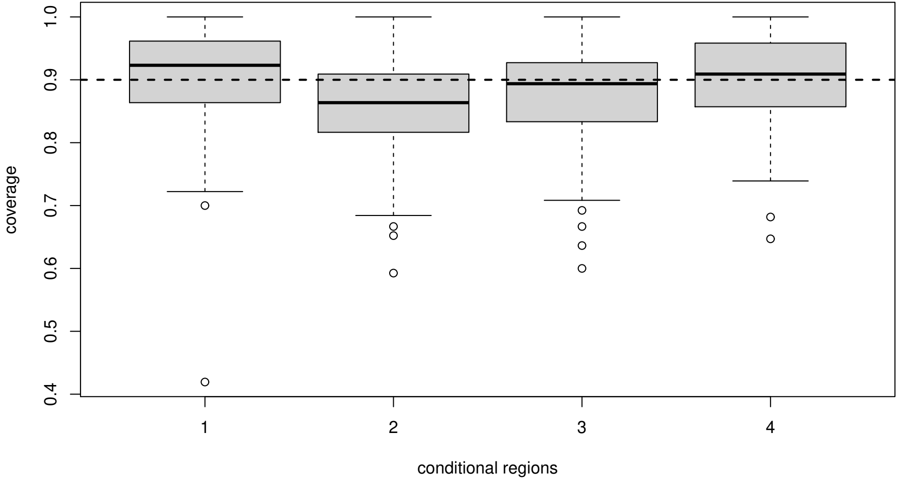
Conditional Coverage on Test Data - Local



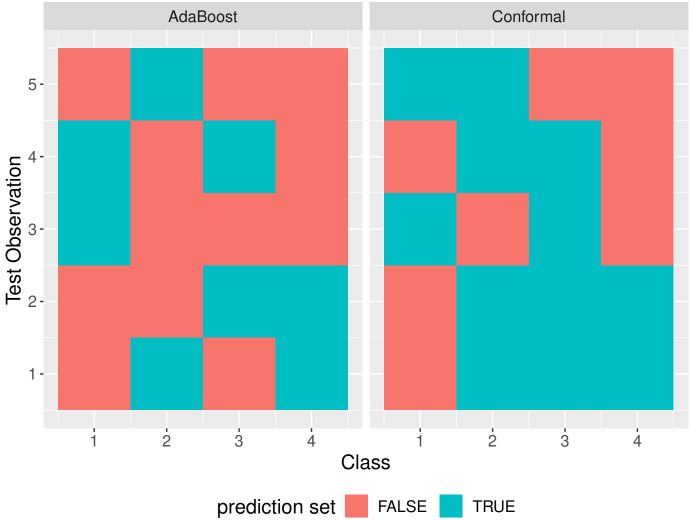
Marginal Coverage Simulation Results



Conditional Coverage Simulation Results



CogPilot Test Example



CogPilot Classification Calibration