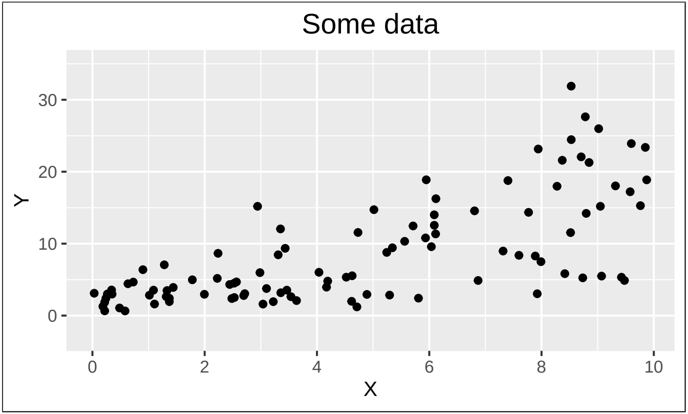
FH AACHEN
UNIVERSITY OF APPLIED SCIENCES

Bootstrap-based goodness-of-fit test for parametric families of conditional distributions

March 13, 2025

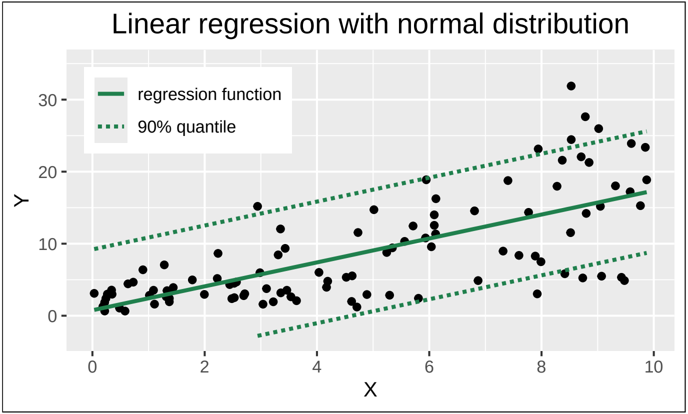
Gitte Kremling

Motivating Example

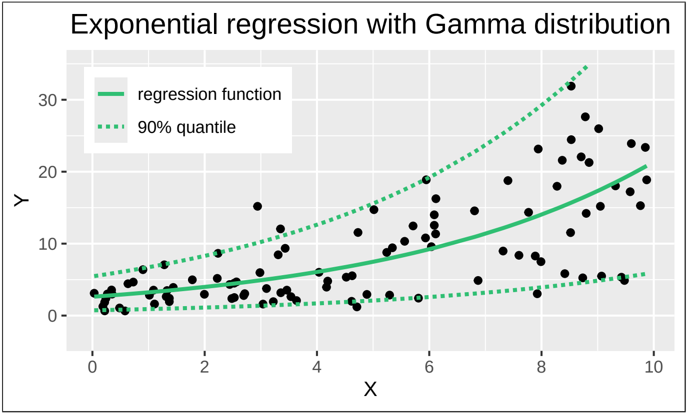


How does Y depend on X?

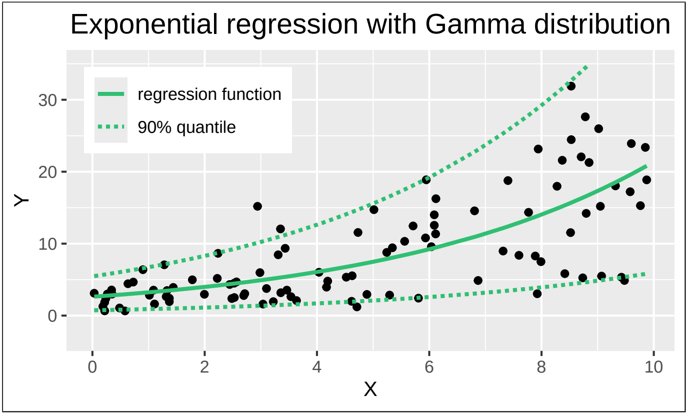
Motivating Example



Motivating Example



Motivating Example



Which one appropriately models the given data?

Problem

Data: i.i.d. sample of covariates $X_i \in \mathbb{R}^p$ and output variables $Y_i \in \mathbb{R}$

Aim: Find a good model for the conditional distribution $Y|X \sim F$

Method: Test goodness-of-fit for different parametric families

$$H_0 : F \in \mathcal{F} = \{(x, y) \mapsto F_\vartheta(y|x) \mid \vartheta \in \Theta\} \quad \text{vs.} \quad H_1 : F \notin \mathcal{F}$$

Goodness-of-fit test - Existing work

- Andrews (1997):

  Based on the difference between a non- and semi-parametric estimate of $F_{X,Y}$

Goodness-of-fit test - Existing work

- Andrews (1997):

  Based on the difference between a non- and semi-parametric estimate of $F_{X,Y}$

- Bierens and Wang (2012):

  Based on the difference between a non- and semi-parametric estimate of $\varphi_{X,Y}$

Goodness-of-fit test - Existing work

- Andrews (1997):

  Based on the difference between a non- and semi-parametric estimate of $F_{X,Y}$

- Bierens and Wang (2012):

  Based on the difference between a non- and semi-parametric estimate of $\varphi_{X,Y}$

- Stute and Zhu (2002) / Dikta and Scheer (2021):

  Based on the difference between a non- and semi-parametric estimate of $\mathbb{E}[\mathbb{1}_{\{\beta^T X \leq t\}} Y]$

  (for parametric GLMs only; tests for the regression function)

Goodness-of-fit test - Existing work

- Andrews (1997):

  Based on the difference between a non- and semi-parametric estimate of $F_{X,Y}$

- Bierens and Wang (2012):

  Based on the difference between a non- and semi-parametric estimate of $\varphi_{X,Y}$

- Stute and Zhu (2002) / Dikta and Scheer (2021):

  Based on the difference between a non- and semi-parametric estimate of $\mathbb{E}[\mathbb{1}_{\{\beta^T X \leq t\}} Y]$

  (for parametric GLMs only; tests for the regression function)

- Other approaches using kernel estimators

Goodness-of-fit test - New approach

- Based on the difference between a non-parametric and semi-parametric estimate of F_Y

Goodness-of-fit test - New approach

- Based on the difference between a non-parametric and semi-parametric estimate of F_Y
- Non-parametric fit: empirical distribution function (ecdf)

$$\hat{F}_{Y,n}(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{Y_i \leq t\}}$$

Goodness-of-fit test - New approach

- Based on the difference between a non-parametric and semi-parametric estimate of F_Y
- Non-parametric fit: empirical distribution function (ecdf)

$$\hat{F}_{Y,n}(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Y_i \leq t\}}$$

- Semi-parametric fit, using MLE $\hat{\vartheta}_n$ and ecdf $\hat{H}_n$ of $\{X_i\}_{i=1}^n$:

$$\hat{F}_{Y,\hat{\vartheta}_n}(t) := \int F_{\hat{\vartheta}_n}(t|x)\hat{H}_n(dx) = \frac{1}{n}\sum_{i=1}^{n} F_{\hat{\vartheta}_n}(t|X_i)$$

Goodness-of-fit test - New approach

- Conditional empirical process with estimated parameters:

$$\tilde{\alpha}_n(t) = \sqrt{n}\left(\hat{F}_{Y,n}(t) - \hat{F}_{Y,\hat{\vartheta}_n}(t)\right)$$
$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{1}_{\{Y_i \leq t\}} - F_{\hat{\vartheta}_n}(t|X_i)$$

- Kolmogorov-Smirnov type distance $\|\tilde{\alpha}_n\|_\infty = \sup_t |\tilde{\alpha}_n(t)|$ should be small under H_0

Goodness-of-fit test - New approach

- Conditional empirical process with estimated parameters:

$$\begin{aligned}\tilde{\alpha}_n(t) &= \sqrt{n}\left(\hat{F}_{Y,n}(t) - \hat{F}_{Y,\hat{\vartheta}_n}(t)\right) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbb{1}_{\{Y_i \le t\}} - F_{\hat{\vartheta}_n}(t|X_i)\end{aligned}$$

- Kolmogorov-Smirnov type distance $\|\tilde{\alpha}_n\|_\infty = \sup_t |\tilde{\alpha}_n(t)|$ should be small under H_0... **But how small is small enough?**

Goodness-of-fit test - New approach

- Conditional empirical process with estimated parameters:

$$\tilde{\alpha}_n(t) = \sqrt{n}\left(\hat{F}_{Y,n}(t) - \hat{F}_{Y,\hat{\vartheta}_n}(t)\right)$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbb{1}_{\{Y_i \le t\}} - F_{\hat{\vartheta}_n}(t|X_i)$$

- Kolmogorov-Smirnov type distance $\|\tilde{\alpha}_n\|_\infty = \sup_t |\tilde{\alpha}_n(t)|$ should be small under H_0... **But how small is small enough?**

Find distribution of $\|\tilde{\alpha}_n\|_\infty$ to decide when H_0 should be rejected.

Goodness-of-fit test - Asymptotic distribution

Theorem

Under H_0 and some regularity conditions, $\tilde{\alpha}_n$ converges weakly to a centered Gaussian process $\tilde{\alpha}_\infty$ with known covariance function which is dependent on the true distribution functions of X and Y.

Goodness-of-fit test - Asymptotic distribution

Theorem

Under H_0 and some regularity conditions, $\tilde{\alpha}_n$ converges weakly to a centered Gaussian process $\tilde{\alpha}_\infty$ with known covariance function which is dependent on the true distribution functions of X and Y.

Proof sketch:

- Splitting as in Durbin (1973):

$$\tilde{\alpha}_n(t) = \sqrt{n}\Big(\hat{F}_{Y,n}(t) - \hat{F}_{Y,\vartheta_0}(t)\Big) + \sqrt{n}\left(\hat{F}_{Y,\vartheta_0}(t) - \hat{F}_{Y,\hat{\vartheta}_n}(t)\right)$$

- Apply Kosorok (2008), Theorem 7.17:
  convergence of fidis and tightness $\Rightarrow$ weak convergence

Goodness-of-fit test - Asymptotic distribution

Theorem

Under H_0 and some regularity conditions, $\tilde{\alpha}_n$ converges weakly to a centered Gaussian process $\tilde{\alpha}_\infty$ with known covariance function which is ***dependent on the true distribution functions of X and Y****.*

Use bootstrap to approximate the distribution of $\|\tilde{\alpha}_n\|_\infty$.

Goodness-of-fit test - Bootstrap

Aim: Estimate the distribution of $\|\tilde{\alpha}_n\|_\infty$ under H_0

Method:

- Resample from the given data in a way that H_0 is fulfilled:
  $X_i^* = X_i$, $Y_i^* \sim F_{\hat{\vartheta}_n}(\,\cdot\,|X_i^*)$
- Compute the test statistic $\|\tilde{\alpha}_n^*\|_\infty$ for this new sample
- Repeat these steps many times and use the ecdf of the resulting test statistics as an estimate

Usage: p-value is approximated by the percentage of $\|\tilde{\alpha}_n^*\|_\infty$ that are greater than or equal to $\|\tilde{\alpha}_n\|_\infty$

Goodness-of-fit test - Asymptotic correctness

FH AACHEN
UNIVERSITY OF APPLIED SCIENCES

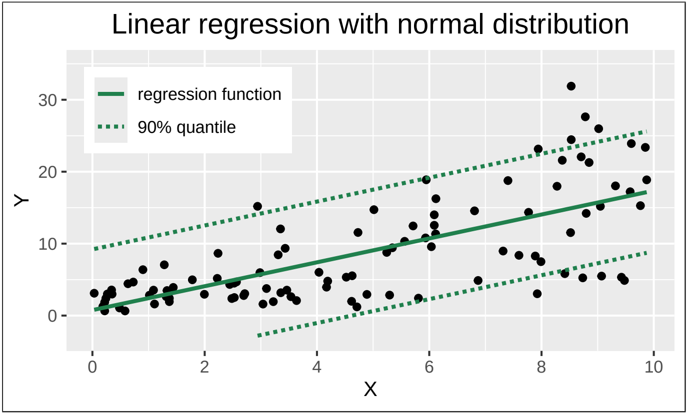
Theorem

Under H_0 and some regularity conditions, $\tilde{\alpha}_n$ converges weakly to a centered Gaussian process $\tilde{\alpha}_\infty$ with known covariance function which is dependent on the true distribution functions of X and Y.

Theorem

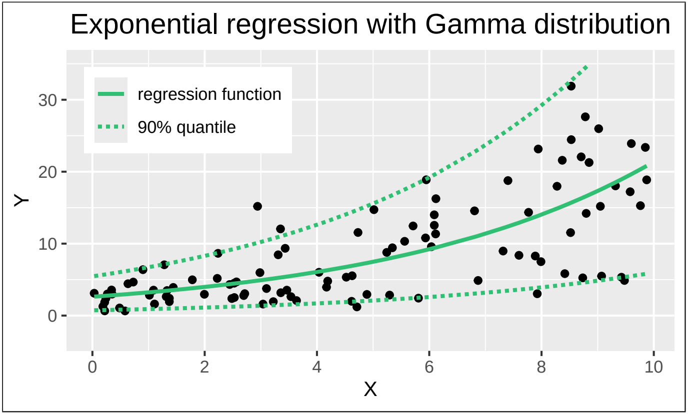
Under H_0 and some regularity conditions, $\tilde{\alpha}_n^*$ converges weakly to the same limit process $\tilde{\alpha}_\infty$.

Back to our motivating example



p-value $= 0$

Back to our motivating example



p-value $= 0.37$

Simulation study - Setup

$X \sim \mathcal{N}(0,1)$

$H_0 : (Y|X) \sim \mathcal{N}(\beta^T X, \sigma^2)$

(A) $Y = 1 + X + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0,1)$

(B) $Y = 1 + X + \varepsilon$ where ε follows a standard logistic distribution

(C) $Y = 1 + X + \varepsilon$ where $\varepsilon \sim t_5$

(D) $Y = 1 + X + X^2 + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0,1)$

(E) $Y = 1 + X + X\varepsilon$ where $\varepsilon \sim \mathcal{N}(0,1)$

Simulation study - Setup

$X \sim \mathcal{N}(0,1)$

$H_0 : (Y|X) \sim \mathcal{N}(\beta^T X, \sigma^2)$

$n = 200$ observations

$m = 500$ bootstrap iterations

$r = 1000$ simulation repetitions

Proportion of rejection for significance level $\alpha = 5\%$

(A) $Y = 1 + X + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0,1)$

(B) $Y = 1 + X + \varepsilon$ where ε follows a standard logistic distribution

(C) $Y = 1 + X + \varepsilon$ where $\varepsilon \sim t_5$

(D) $Y = 1 + X + X^2 + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0,1)$

(E) $Y = 1 + X + X\varepsilon$ where $\varepsilon \sim \mathcal{N}(0,1)$

Simulation study - Results

Proportion of rejection in percentage terms for $\alpha = 5\%$:

	(A)	(B)	(C)	(D)	(E)
New approach	**5.2**	**22.7**	**45.4**	**5.2**	**99.8**
Andrews (1997)	5.6	18.2	37.6	7.4	100.0
Bierens & Wang (2012)	4.6	9.4	19.7	5.5	99.8
Dikta & Scheer (2021)	5.9	6.0	4.8	13.9	16.2

(A) $Y = 1 + X + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0,1)$

(B) $Y = 1 + X + \varepsilon$ where ε follows a standard logistic distribution

(C) $Y = 1 + X + \varepsilon$ where $\varepsilon \sim t_5$

(D) $Y = 1 + X + X^2 + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0,1)$

(E) $Y = 1 + X + X\varepsilon$ where $\varepsilon \sim \mathcal{N}(0,1)$

Advantages over former methods

- Compared to Andrews (1997):
  Better applicable to cases with high-dimensional covariates

- Compared to Bierens and Wang (2012):
  More sensitive to deviations from H_0 (in our simulation study)

- Compared to Stute and Zhu (2002) / Dikta and Scheer (2021):
  More specific because it tests for the whole conditional distribution not just the regression function

- Additional advantage:
  Easily applicable due to *gofreg*-package in R

References

Andrews, Donald WK (1997). "A conditional Kolmogorov test". In: *Econometrica: Journal of the Econometric Society*, pp. 1097–1128.

Bierens, Herman J and Li Wang (2012). "Integrated conditional moment tests for parametric conditional distributions". In: *Econometric Theory* 28.2, pp. 328–362.

Dikta, Gerhard and Marsel Scheer (2021). *Bootstrap Methods. With Applications in R*. 1st ed. Springer International Publishing.

Durbin, James (1973). "Weak convergence of the sample distribution function when parameters are estimated". In: *The Annals of Statistics*, pp. 279–290.

Kosorok, Michael R (2008). *Introduction to empirical processes and semiparametric inference*. Vol. 61. Springer.

Stute, Winfried and Li-Xing Zhu (2002). "Model checks for generalized linear models". In: *Scandinavian Journal of Statistics* 29.3, pp. 535–545.

References

Preprint on arXiv:

R-package on CRAN:

Any questions? :)