FH AACHEN
UNIVERSITY OF APPLIED SCIENCES

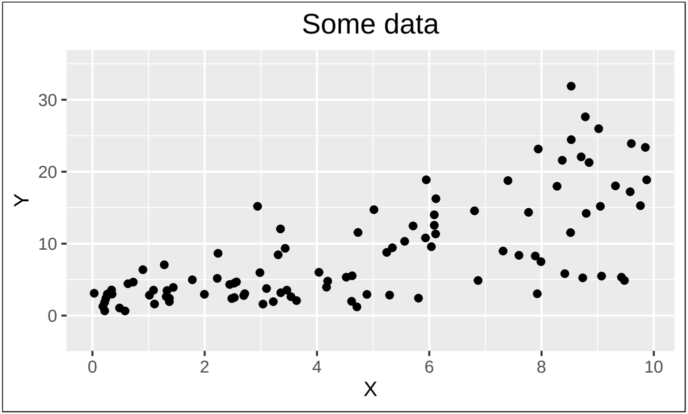
Bootstrap-based goodness-of-fit test for parametric regression based on conditional distribution families

August 7, 2024

Gitte Kremling, Gerhard Dikta, Richard Stockbridge

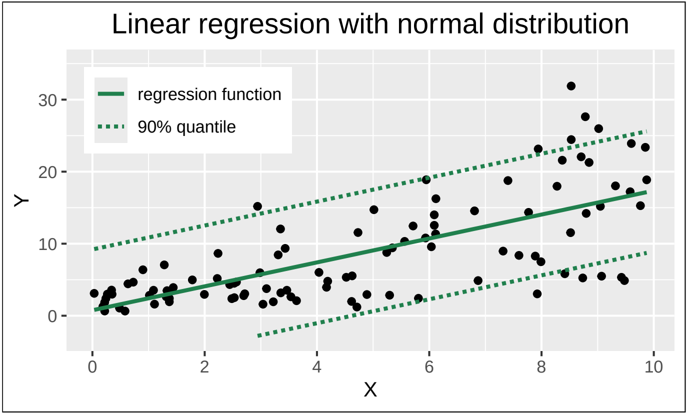
UNIVERSITY of WISCONSIN
UWMILWAUKEE

Motivating Example

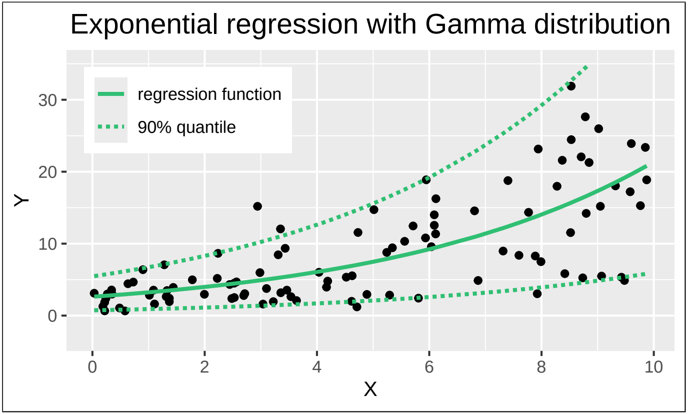


How does Y depend on X?

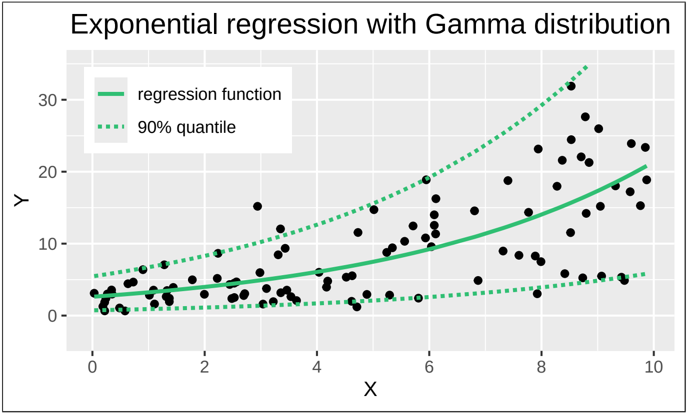
Motivating Example



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 $Y_i \in \mathbb{R}$

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

Method: Test goodness of fit for different parametric families

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

Goodness-of-fit test - Previous work

- Stute (1997):

  Test for a parametric family of regression functions $m(x) = \mathbb{E}[Y|X = x]$ based on empirical process of X weighted by the corresponding residuals (MEP)

Goodness-of-fit test - Previous work

- Stute (1997):

  Test for a parametric family of regression functions $m(x) = \mathbb{E}[Y|X = x]$ based on empirical process of X weighted by the corresponding residuals (MEP)

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

  Test for a family of parametric generalized linear models (GLMs) based on a slightly modified MEP

Goodness-of-fit test - Previous work

- Stute (1997):

  Test for a parametric family of regression functions $m(x) = \mathbb{E}[Y|X = x]$ based on empirical process of X weighted by the corresponding residuals (MEP)

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

  Test for a family of parametric generalized linear models (GLMs) based on a slightly modified MEP

- Andrews (1997):

  Test for a family of conditional distribution functions based on difference between non- and semi-parametric fit of $F_{X,Y}$

Goodness-of-fit test - Previous work

- Stute (1997):

  Test for a parametric family of regression functions $m(x) = \mathbb{E}[Y|X = x]$ based on empirical process of X weighted by the corresponding residuals (MEP)

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

  Test for a family of parametric generalized linear models (GLMs) based on a slightly modified MEP

- Andrews (1997):

  Test for a family of conditional distribution functions based on difference between non- and semi-parametric fit of $F_{X,Y}$

- Other approaches using kernel estimators

Goodness-of-fit test - New approach

- Difference between 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:

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

- 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?**

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 - Limit distribution

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.

Goodness-of-fit test - Limit 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

Goal: 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 $(Y_i^* \sim F_{\hat{\vartheta}_n}(\,\cdot\,|X_i))$
- 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

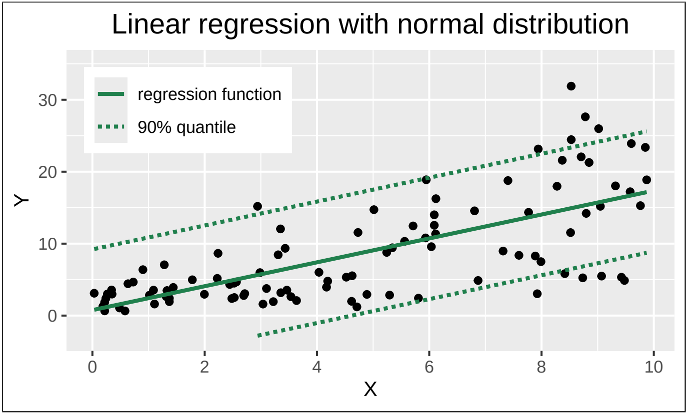
Theorem

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

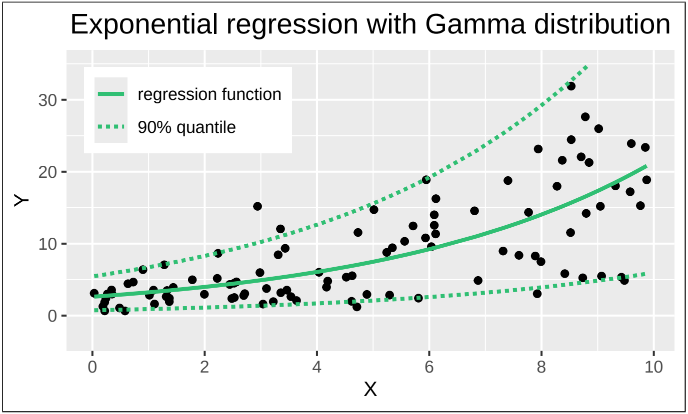
Under H_0 and some regularity conditions, $\tilde{\alpha}_n^*$ converges weakly to the same limit process $\tilde{\alpha}_\infty$.

Back to our motivating example



p-value $= 0$

Back to our motivating example



p-value $= 0.37$

Comparison to previous work

- Compared to Stute (1997):
  More specific because it includes the distribution not just regression function
- Compared to Stute and Zhu (2002) / Dikta and Scheer (2021):
  More general because it cannot only be applied for GLMs
- Compared to Andrews (1997):
  Better applicable to cases with high-dimensional covariates
- Compared to all of them:
  In some cases more sensitive to deviations from H_0

References I

- Andrews, Donald WK (1997). "A conditional Kolmogorov test". In: *Econometrica: Journal of the Econometric Society*, pp. 1097–1128.
- 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 (1997). "Nonparametric model checks for regression". In: *The Annals of Statistics*, pp. 613–641.
- Stute, Winfried and Li-Xing Zhu (2002). "Model checks for generalized linear models". In: *Scandinavian Journal of Statistics* 29.3, pp. 535–545.

Comparison to previous work

- Compared to Stute (1997):
  More specific because it includes the distribution not just regression function
- Compared to Stute and Zhu (2002) / Dikta and Scheer (2021):
  More general because it cannot only be applied for GLMs
- Compared to Andrews (1997):
  Better applicable to cases with high-dimensional covariates
- Compared to all of them:
  In some cases more sensitive to deviations from H_0

Thank you for your attention! Any questions? :)