Transformation Models

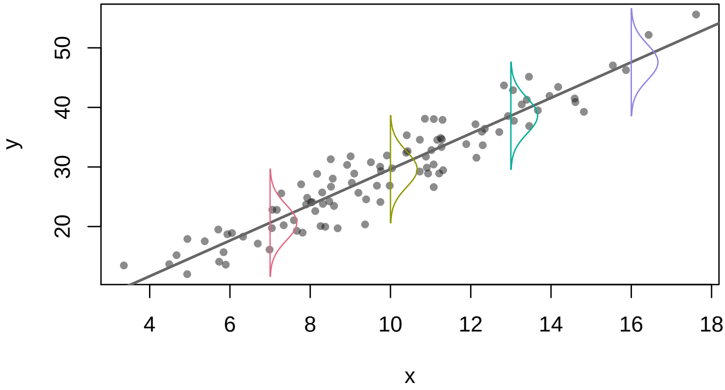
An Introduction

Lucas Kook
University of Zürich

October 10, 2019

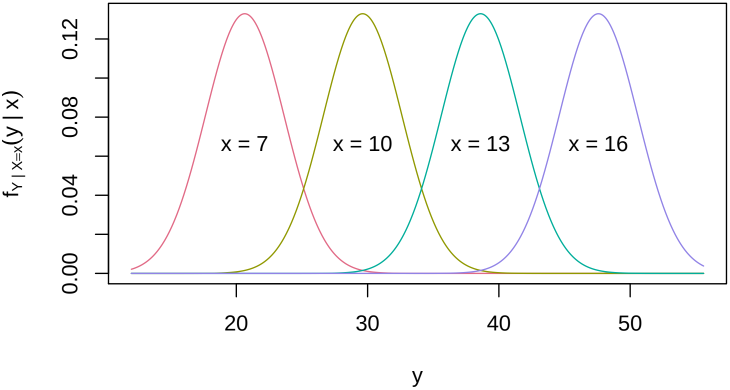
Regression in a classical sense

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \; \varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$$



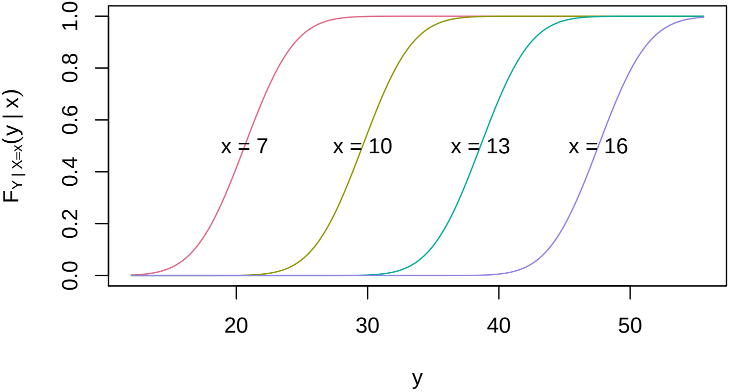
Regression in a classical sense

A different perspective: $f_{Y|X=x}(y|x) = \phi\left(\frac{y-\alpha-\beta x}{\sigma}\right)$



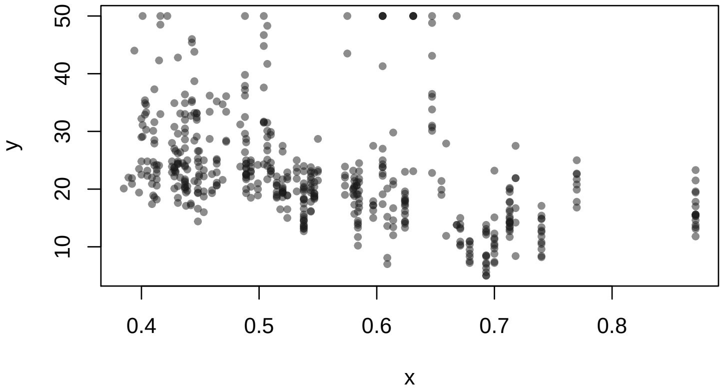
Regression in a classical sense

Yet another scale: $\mathbb{P}(Y \leq y|x) = F_{Y|X=x}(y|x) = \Phi\left(\frac{y-\alpha-\beta x}{\sigma}\right)$



Where classical regression breaks down

So how does one tackle a problem like this?



A note on classical regression

Xiao-Li Meng
@XiaoLiMeng1

The term “regression” reflects statisticians’ modesty, and perhaps also our regrets? We should not have started statistical modeling with regression, for it confuses probabilistic model fitting with deterministic line/curve fitting, building wrong intuitions for generations.

7:48 am · 1 Oct 2019 · Twitter for iPhone

Source: twitter.com/XiaoLiMeng1

Perspectives on regression

Linear models:

$$\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}$$

Generalized linear models:

$$g(\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})) = \boldsymbol{x}^\top \boldsymbol{\beta}$$

Transformation models:

$$F_Y(y|\boldsymbol{x}) = F_Z(h_Y(y|\boldsymbol{x}))$$

Transformation models

$$F_Y(y|\boldsymbol{x}) = F_Z(h_Y(y|\boldsymbol{x}))$$

F_Y (Complex) conditional distribution of the response

F_Z (Simple) error distribution

h_Y (Flexible) transformation function

Motivation: Regression

Everything is in the conditional distribution function!

$$\mathbb{P}(Y \leq y | \boldsymbol{X} = \boldsymbol{x}) = F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y|\boldsymbol{x})$$

Q1: How do changes in $\boldsymbol{x}$ propagate to y?

Q2: How can we estimate $\hat{F}_{Y|\boldsymbol{X}=\boldsymbol{x}}$ from data?

Q3: Why model on the scale of the cdf?

The linear model as a transformation model

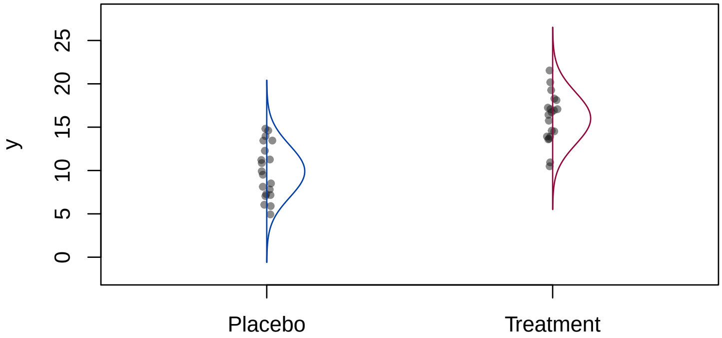
Starting from $Y|\boldsymbol{x} \overset{\text{i.i.d.}}{\sim} \mathrm{N}\left(\alpha + \boldsymbol{x}^\top \boldsymbol{\beta}, \sigma^2\right)$ we have

$$\mathbb{P}(Y \leq y|\boldsymbol{X} = \boldsymbol{x}) = \Phi\left(\frac{y - \alpha - \boldsymbol{x}^\top \boldsymbol{\beta}}{\sigma}\right).$$

Identify

$$\begin{aligned} F_Z &= \Phi \\ h_Y(y|\boldsymbol{x}) &= y/\sigma - \alpha/\sigma - \boldsymbol{x}^\top \boldsymbol{\beta}/\sigma \\ &= \vartheta_1 + \vartheta_2 y - \boldsymbol{x}^\top \tilde{\boldsymbol{\beta}} \end{aligned}$$

Example: Two group comparison



Example: Two group comparison

Continuous response Y and one binary treatment indicator $x \in \{0, 1\}$:

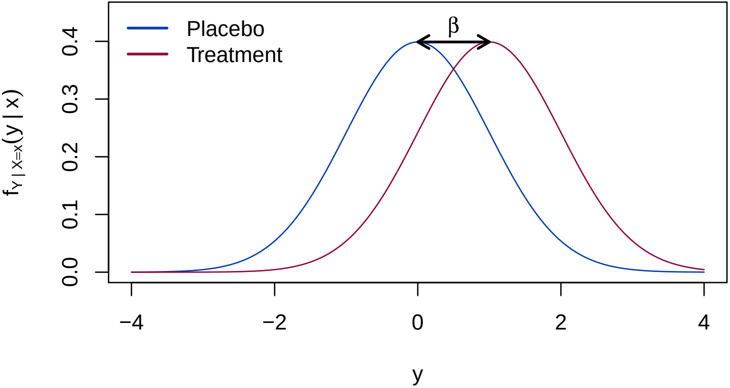
$$
\begin{aligned}
F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y|x=0) &= F_Z(h(y)) \\
h(y) &= F_Z^{-1}\left(F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y|x=0)\right) \\
Z &= h(y) \text{ is the transformed r.v.}
\end{aligned}
$$

Now assume

$$
F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y|x=1) = F_Z(h(y) - \beta)
$$

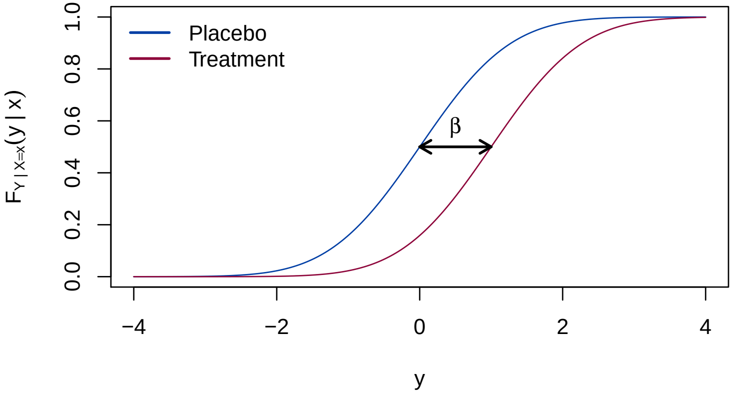
Example: Two group comparison

$$f_{Y|\boldsymbol{X}=\boldsymbol{x}}(y|x) = \phi(h(y) - \tilde{\beta}x)h'(y)$$



Example: Two group comparison

$$F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y|x) = \Phi(h(y) - \tilde{\beta}x)$$



Example: Two group comparison

Now $F_Z = \Phi$ determines the interpretational scale of $\tilde{\beta}$:

$$\mathbb{E}(h(y) \mid x = 1) - \mathbb{E}(h(y) \mid x = 0) = \tilde{\beta}$$

Since

$$(h(y)|x = 0) \sim \mathrm{N}(0, 1) \text{ and}$$
$$(h(y)|x = 1) \sim \mathrm{N}(\tilde{\beta}, 1)$$

Example: Two group comparison

Now $F_Z = \Phi$ determines the interpretational scale of $\tilde{\beta}$:

$$\mathbb{E}(h(y) \mid x = 1) - \mathbb{E}(h(y) \mid x = 0) = \tilde{\beta}$$

Since

$$(h(y)|x = 0) \sim \mathrm{N}(0, 1) \text{ and}$$
$$(h(y)|x = 1) \sim \mathrm{N}(\tilde{\beta}, 1)$$

Bonus: $\mathbb{E}(Y|x = 1) - \mathbb{E}(Y|x = 0) = \beta$ if $h(y)$ affine

Example: Two group comparison

```
set.seed(24101968)
n <- 20; beta <- 2
x <- rep(c(0, 1), each = 10)
y <- 10 + x * beta + rnorm(n, sd = 0.5)
coef(m0 <- stats::lm(y ~ x))

## (Intercept)           x
##        9.76        2.44

coef(m1 <- tram::Lm(y ~ x), with_baseline = TRUE)

## (Intercept)           y           x
##      -21.34        2.19        5.34
```

Q: How do we arrive at the same coefficients?

Example: Two group comparison

Since $\tilde{\alpha} = -\alpha/\sigma$ and $\tilde{\beta} = \beta/\sigma$

```
coef(m1, with_baseline = TRUE)[-2] /
        coef(m1, with_baseline = TRUE)[2] * c(-1, 1)



## (Intercept)           x
##        9.76        2.44
```

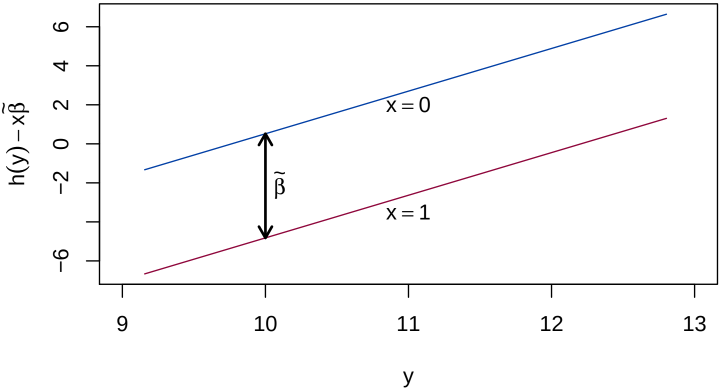
But why favor `Lm` over `lm`?

- `lm()` estimates $\hat{\sigma}^2$ and $\hat{\beta}$ separately via REML
- `lm()` cannot deal with any form of censoring

Example: Two group comparison

Affine baseline transformations are very restrictive!

$$h_Y(y|\boldsymbol{x}) = \vartheta_1 + \vartheta_2 y - \boldsymbol{x}^\top \tilde{\boldsymbol{\beta}}$$



Beyond the linear model: Box-Cox type models

Allow $h(y)$ to be more flexible, e.g. a basis expansion

$$h(y; \boldsymbol{\vartheta}) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}$$

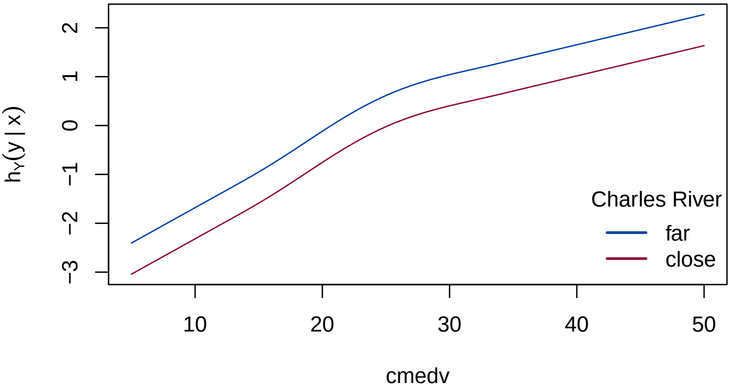
```
m2 <- BoxCox(cmedv ~ chas, order = 6, data = BostonHousing2,
             extrapolate = TRUE)
coef(m2, with_baseline = TRUE)
```

```
## Bs1(cmedv) Bs2(cmedv) Bs3(cmedv) Bs4(cmedv) Bs5(cmedv) Bs6(cmedv)
##     -1.262     -0.737     -0.213      0.873      0.873      1.097
## Bs7(cmedv)      chas1
##      1.322      0.638
```

$$F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y|\boldsymbol{x}) = F_Z\left(h(y) - \boldsymbol{x}^\top \boldsymbol{\beta}\right)$$
$$= \Phi\left(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta} - \boldsymbol{x}^\top \boldsymbol{\beta}\right)$$

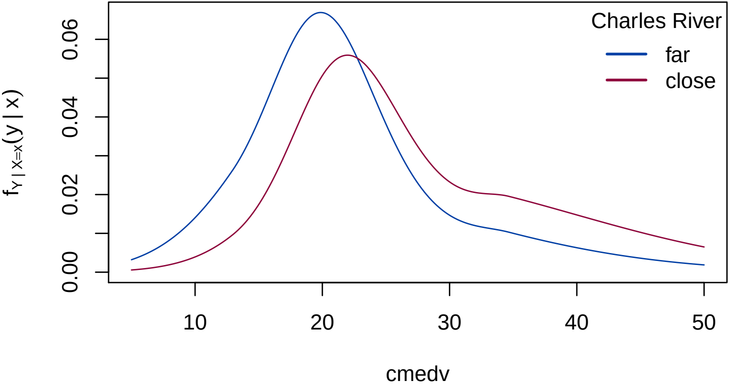
Beyond linear baseline transformations

$$h_Y(y|\text{chas}) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta} - \beta \cdot \text{chas}$$



Beyond linear baseline transformations

$$f_Y(y|\text{chas}) = \phi\left(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta} - \beta \cdot \text{chas}\right) \boldsymbol{a}'(y)^\top \boldsymbol{\vartheta}$$



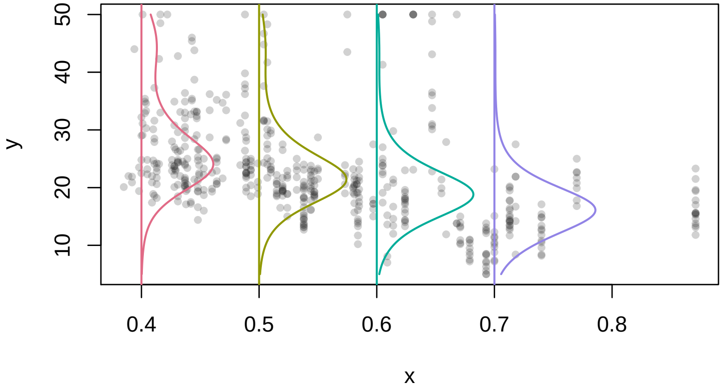
Beyond $F_Z(z) = \Phi(z)$

Interpretational scale of β changes with F_Z:

$\Phi(z)$	β difference in expectation
$F_{\text{SL}}(z) = \text{expit}(z)$	β log odds ratio
$F_{\text{MEV}}(z) = 1 - \exp(-\exp(z))$	β log hazard ratio
$F_{\text{Gumbel}}(z) = \exp(-\exp(-z))$	β log Lehmann alternative

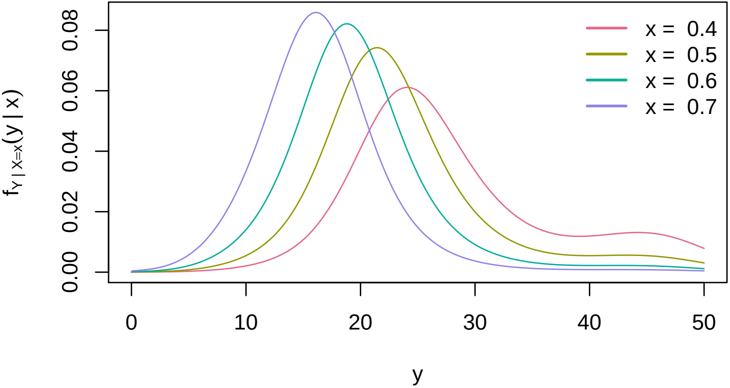
Back to the beginning

$$F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y|x) = F_{\text{SL}}\left(h(y) + \beta x\right)$$



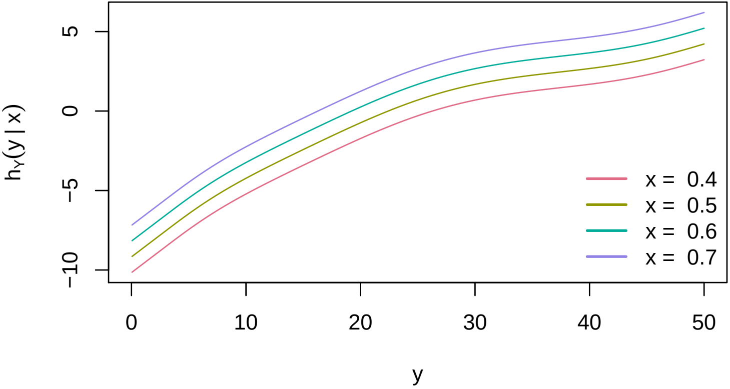
Back to the beginning

$$F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y|x) = f_{\text{SL}}(h(y) + \beta x)\, h'(y)$$

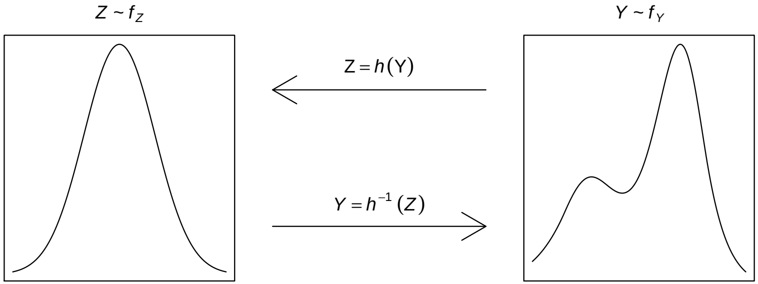


Back to the beginning

$$h_Y(y|x) = h(y) + \beta x$$



Connection to Flow-based methods



Outlook: Beyond stratified linear transformation models

- Conditional transformation models {`mlt`} (TH)
- Transformation mixed models {`tramm`} (BT)
- Count-transformation models {`cotram`} (SS)
- Regularized transformation models {`tramnet`} (LK)
- Transformation trees and random forests {`trtf`} (TH)
- Transformation boosting machines {`tbm`} (TH)
- Multivariate transformation models (LB)

Acknowledgements

Torsten Hothorn

Muriel Buri

Luisa Barbanti

Sandra Sigfried

Balint Tamasi

Beate Sick

Appendix

Basis Expansions

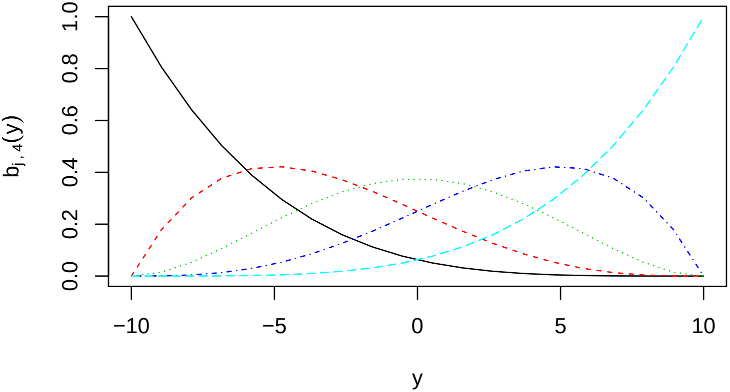
Trams are parametrized using Bernstein Polynomials.

$$b_{\nu,n}(y) = \binom{n}{\nu} y^{\nu}(1-y)^{n-\nu}, \quad \nu = 0, \dots, n$$
$$h_Y(y) = \boldsymbol{a}_{\text{Bs},p}(y)^\top \boldsymbol{\vartheta}$$

- Monotonicity constraint nicely translates into $\boldsymbol{D}^{(1)}\boldsymbol{\vartheta} \geq 0$
- Taking derivatives is easy, i.e. to compute $f_Y(y)$
- Direct connection to the Beta distribution
- Computational convenience

Basis Expansions

$$b_{\nu,n}(y) = \binom{n}{\nu} y^{\nu}(1-y)^{n-\nu}, \quad \nu = 0, \ldots, n$$



Interpretational scales induced by F_Z

F_Z	Interpretation of $\boldsymbol{x}^\top \boldsymbol{\beta}$
Φ	$\mathbb{E}(h_Y(Y) \mid \boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}$
F_{SL}	$\frac{F_{Y \mid \boldsymbol{X} = \boldsymbol{x}}(y \mid \boldsymbol{x})}{1 - F_{Y \mid \boldsymbol{X} = \boldsymbol{x}}(y \mid \boldsymbol{x})} = \frac{F_Y(y)}{1 - F_Y(y)} \exp(-\boldsymbol{x}^\top \boldsymbol{\beta})$
F_{MEV}	$1 - F_{Y \mid \boldsymbol{X} = \boldsymbol{x}}(y \mid \boldsymbol{x}) = (1 - F_Y(y))^{\exp(-\boldsymbol{x}^\top \boldsymbol{\beta})}$
F_{Gumbel}	$F_{Y \mid \boldsymbol{X} = \boldsymbol{x}}(y \mid \boldsymbol{x}) = F_Y(y)^{\exp(\boldsymbol{x}^\top \boldsymbol{\beta})}$

Beyond shift effects

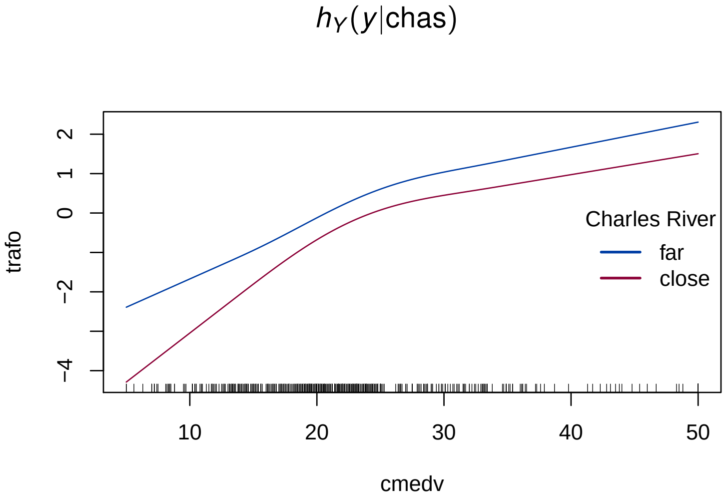
Stratum variables and response varying effects

$$h_Y(y|\boldsymbol{s}, \boldsymbol{x}) = h_Y(y|\boldsymbol{s}) - \boldsymbol{x}^\top \beta$$

```
m3 <- BoxCox(cmedv | chas ~ 1, order = 6, data = BostonHousing2, extra
```

- Binary stratum variable: Separate baseline trafos
- Continuous strata: response varying effect

Beyond shift effects



Beyond shift effects

$$f_Y(y|\text{chas}) = \phi\left(h_Y(y|\text{chas})\right) h'_Y(y|\text{chas})$$