Partial identification in linear models: Regression with interval-valued data

Georg Schollmeyer Thomas Augustin

Situation

Situation: Simple Linear model $Y = \beta_0 + \beta_1 X + \varepsilon$ with interval-censored outcomes.

- Y unobserved, only $\underline{\mathrm{Y}}$ and $\overline{\mathrm{Y}}$ observed and all we know is $Y \in [\underline{\mathrm{Y}}, \overline{\mathrm{Y}}]$.
- model partially identified, there are many parameters (β_0, β_1) leading to the same distribution of the observable random variables $(X, \underline{\mathrm{Y}}, \overline{\mathrm{Y}})$.

- *Object of interest: The true parameter vector* $\beta = (\beta_0, \beta_1)$.

- *Object of interest: The true parameter vector* (β_0, β_1).
- *Problem: Because the model is only partially identified, there do not exist consistent classical point estimators for* (β_0, β_1).
- *'Solution': Try to estimate some (preferably sharp) set that contains the true parameter.*
- *The set of all parameters* (β_0, β_1) *that are compatible with the model assumptions and (the distribution of) the observable random variables is called* **identified set**.
- *Aim Here: Try to estimate the identified set.*
- *Different understandings of the linear model (***descriptive** *vs* **structural***) lead to different identification regions and estimators.*

Descriptive vs structural models (cf. Freedman, 1987)

Here:

- Descriptive linear model in the sense of the best linear predictor (blp) under squared loss: Find (estimate) that prediction function f linear in X that minimizes expected squared loss.
- Structural linear model in the sense that the conditional expectation of Y given x is assumed to be truly a linear function in x: Find (estimate) the intercept and slope of this truly linear relationship.

Different approaches

1. Moment inequality approach (Chernozhukov, Hong & Tamer, ECMA, 2007)
2. Cautious data completion approach (Beresteanu, Molchanov, Molinari, ECMA, 2011, Černỳ, Rada, Meas. Sci. Rev., 2011)
3. Approach based on the minimization of a set-domained loss function (Schollmeyer, Augustin, IJAR, 2015)

Moment inequality approach (Chernozhukov, Hong & Tamer, ECMA, 2007)

1) Approach based on a criterion function $Q : \mathbb{R}^2 \longrightarrow \mathbb{R}_{\geq 0}$ that characterizes the identified set (marrow region, **MR**) as

$$MR = \{(\beta_0, \beta_1) \in \mathbb{R}^2 \mid Q((\beta_0, \beta_1)) = 0\}.$$

Then MR can be estimated via an empirical analogue Q_n of Q as

$$\hat{MR} = \{(\beta_0, \beta_1) \mid Q_n((\beta_0, \beta_1)) = 0\}$$

or as

$$\hat{MR} = \underset{(\beta_0, \beta_1)}{\operatorname{argmin}}\, Q_n((\beta_0, \beta_1))$$

or as

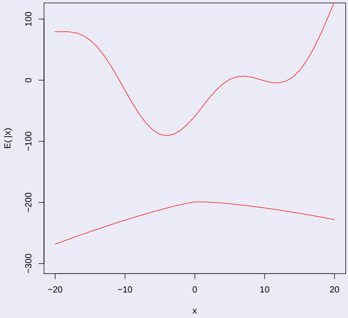
$$\hat{MR} = \{(\beta_0, \beta_1) \in \mathbb{R}^2 \mid Q_n((\beta_0, \beta_1)) \leq c\}$$

for some 'appropriately' choosen value c.

Identification regions for the simple linear model

a) Marrow Region (model understood as structural model):

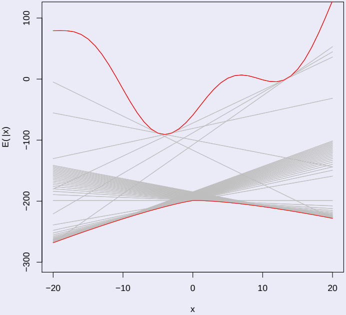
$$MR(\underline{\mathrm{Y}}, \overline{\mathrm{Y}}) = \{\beta \mid \mathbb{E}(\underline{\mathrm{Y}} \mid X) \leq \beta_0 + \beta_1 X \leq \mathbb{E}(\overline{\mathrm{Y}} \mid X)\}$$



Identification regions for the simple linear model

a) Marrow Region:

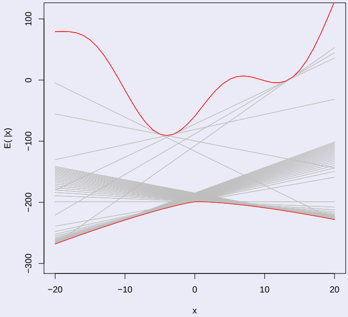
$$MR(\underline{Y}, \overline{Y}) = \{\beta \mid \mathbb{E}(\underline{Y} \mid X) \leq \beta_0 + \beta_1 X \leq \mathbb{E}(\overline{Y} \mid X)\}$$



Identification regions for the simple linear model

a) Marrow Region: $MR(\underline{Y}, \overline{Y}) = \{(\beta_0, \beta_1) \mid Q(\beta_0, \beta_1) = 0\}$ *with*

$$Q(\beta_0, \beta_1) = \int (\mathbb{E}(\underline{Y} \mid X) - (\beta_0 + \beta_1 X))_+^2 + (\mathbb{E}(\overline{Y} \mid X) - (\beta_0 + \beta_1 X))_-^2 \, d\mathbb{P}(x)$$



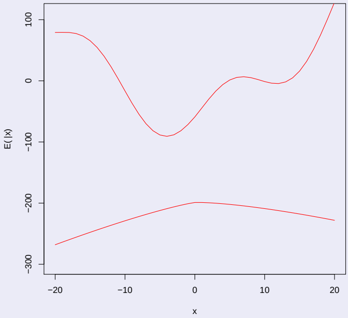
Cautious data completion approach (Beresteanu, Molchanov, Molinari, ECMA, 2011, Černỳ, Rada, Meas. Sci. Rev., 2011)

2) **Cautious data completion**: collect all best linear predictors for all random variables (X, Z) compatible with the observable random variables $(X, \underline{Y}, \overline{Y})$.

Identification regions for the simple linear model

b) Collection Region (model understood as descriptive model, collection of all blp's):

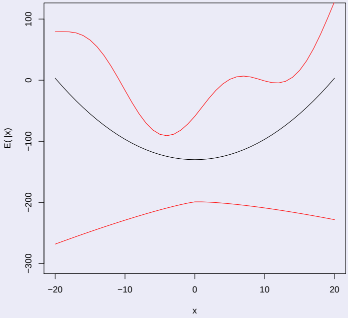
$$CR(\underline{Y}, \overline{Y}) = \{\operatorname{argmin} \mathbb{E}((\beta_0 + \beta_1 X - Z)^2) \mid Z \in [\underline{Y}, \overline{Y}]\}$$



Identification regions for the simple linear model

b) Collection Region:

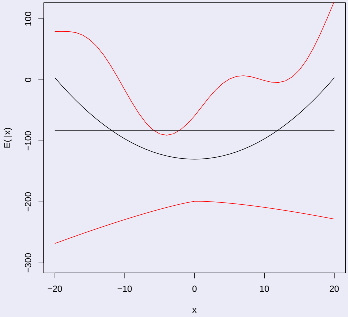
$$CR(\underline{Y}, \overline{Y}) = \{\operatorname{argmin} \mathbb{E}((\beta_0 + \beta_1 X - Z)^2) \mid Z \in [\underline{Y}, \overline{Y}]\}$$



Identification regions for the simple linear model

b) Collection Region:

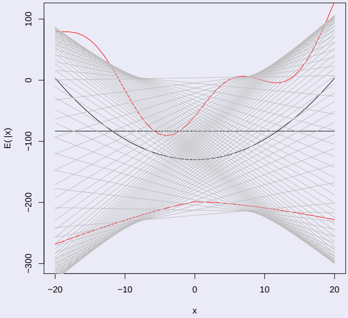
$$CR(\underline{Y}, \overline{Y}) = \{\operatorname{argmin} \mathbb{E}((\beta_0 + \beta_1 X - Z)^2) \mid Z \in [\underline{Y}, \overline{Y}]\}$$



Identification regions for the simple linear model

b) Collection Region:

$$CR(\underline{Y}, \overline{Y}) = \{\operatorname{argmin} \mathbb{E}((\beta_0 + \beta_1 X - Z)^2) \mid Z \in [\underline{Y}, \overline{Y}]\}$$



Approach based on the minimization of a set-domained loss function (Schollmeyer, Augustin, IJAR, 2015)

3) Approach based on minimizing a **set**-domained loss function: Find that **set** Γ for which the predicted boundaries

$$\sup_{\beta \in \Gamma} \beta_0 + \beta_1 x$$

and

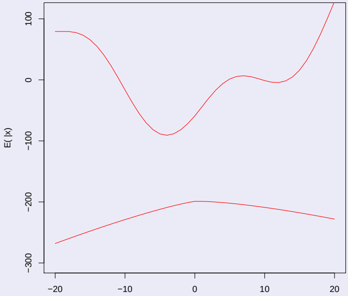
$$\inf_{\beta \in \Gamma} \beta_0 + \beta_1 x$$

are close to the observable boundaries $\overline{\mathrm{Y}}$ and $\underline{\mathrm{Y}}$ in terms of minimal expected loss.

Set-loss Region: Looking at the identified set rigorously as a set.

Set-loss Region (model understood as descriptive model, in a sense "best set-valued predictor"):

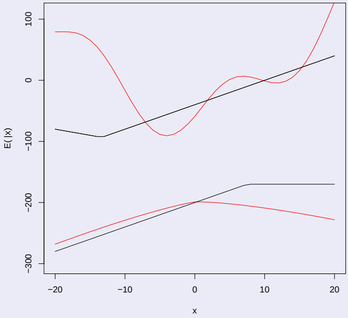
$$\begin{aligned} L_S(\underline{\mathrm{Y}}, \overline{\mathrm{Y}}, \Gamma) &= \int \left(\mathbb{E}(\overline{\mathrm{Y}} \mid x) - \sup_{\beta \in \Gamma} [\beta_0 + \beta_1 x] \right)^2 + \left(\mathbb{E}(\underline{\mathrm{Y}} \mid x) - \inf_{\beta \in \Gamma} [\beta_0 + \beta_1 x] \right)^2 d\mathbb{P}(x) \\ SR(\underline{\mathrm{Y}}, \overline{\mathrm{Y}}) &= \bigcup \operatorname*{argmin}_{\Gamma \subseteq \mathbb{R}^2} L_S(\underline{\mathrm{Y}}, \overline{\mathrm{Y}}, \Gamma) \end{aligned}$$



Set-loss Region

Set-loss Region:

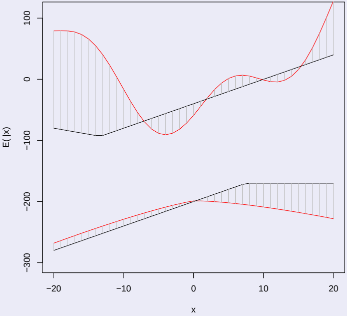
$$
\begin{aligned}
L_S(\underline{Y}, \overline{Y}, \Gamma) &= \int \left(\mathbb{E}(\overline{Y} \mid x) - \sup_{\beta \in \Gamma} [\beta_0 + \beta_1 x]\right)^2 + \left(\mathbb{E}(\underline{Y} \mid x) - \inf_{\beta \in \Gamma} [\beta_0 + \beta_1 x]\right)^2 d\mathbb{P}(x) \\
SR(\underline{Y}, \overline{Y}) &= \bigcup_{\Gamma \subseteq \mathbb{R}^2} \operatorname{argmin} L_S(\underline{Y}, \overline{Y}, \Gamma)
\end{aligned}
$$



Set-loss Region

Set-loss Region:

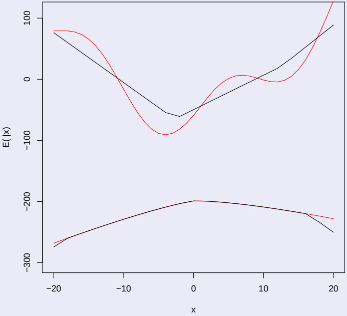
$$
\begin{aligned}
L_S(\underline{Y}, \overline{Y}, \Gamma) &= \int \left(\mathbb{E}(\overline{Y} \mid x) - \sup_{\beta \in \Gamma} [\beta_0 + \beta_1 x] \right)^2 + \left(\mathbb{E}(\underline{Y} \mid x) - \inf_{\beta \in \Gamma} [\beta_0 + \beta_1 x] \right)^2 d\mathbb{P}(x) \\
SR(\underline{Y}, \overline{Y}) &= \bigcup_{\Gamma \subseteq \mathbb{R}^2} \operatorname{argmin} L_S(\underline{Y}, \overline{Y}, \Gamma)
\end{aligned}
$$



Set-loss Region

Set-loss Region:

$$
\begin{aligned}
L_S(\underline{\mathrm{Y}}, \overline{\mathrm{Y}}, \Gamma) &= \int \left(\mathbb{E}(\overline{\mathrm{Y}} \mid x) - \sup_{\beta \in \Gamma} [\beta_0 + \beta_1 x] \right)^2 + \left(\mathbb{E}(\underline{\mathrm{Y}} \mid x) - \inf_{\beta \in \Gamma} [\beta_0 + \beta_1 x] \right)^2 d\mathbb{P}(x) \\
SR(\underline{\mathrm{Y}}, \overline{\mathrm{Y}}) &= \bigcup \operatorname*{argmin}_{\Gamma \subseteq \mathbb{R}^2} L_S(\underline{\mathrm{Y}}, \overline{\mathrm{Y}}, \Gamma)
\end{aligned}
$$



Problems

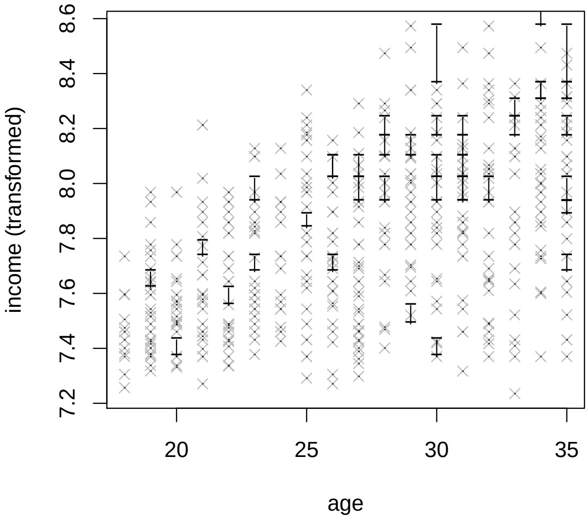
- The sharp marrow region is **not** continuously dependent on the conditional expectations $\mathbb{E}(Y \mid X)$.
- Collection Region: Is the collection of descriptive models still a reasonable descriptive model itself?
  Not in every case.
- Set-loss region: Is it a reasonable, not too 'big' set that contains the true parameter in the case of a correctly specified linear (structural) model?
  It always contains the true parameter if the model is correctly specified, but sometimes it is bigger as e.g., the collection region that also contains the true parameter.

Data example: Allbus 2014 (Only for illustration)

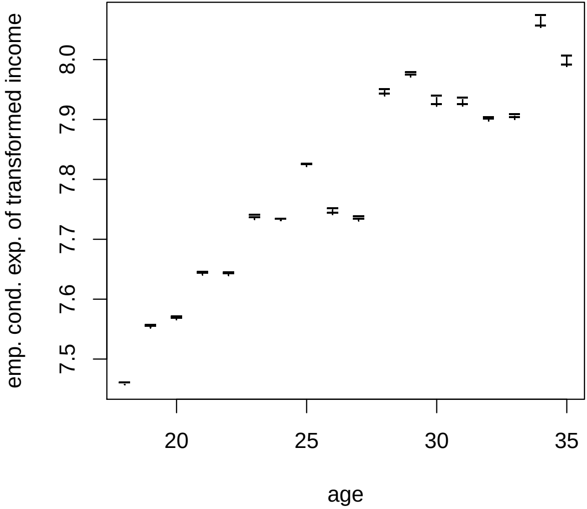
Example (Allbus 2014): age and (transformed) income of people from former West Germany aged from 18 to 35.

- $x:$ age in years
- $y = log($ monthly net income in Euro $- 1000)$: transformed income
- $n = 559$ people (for the sake of simplicity, non-response was ignored here)
- 40 persons (7%) gave only categorized answers for income.

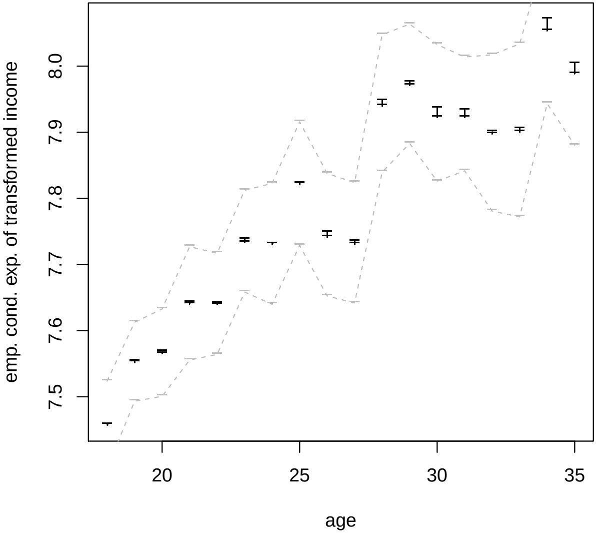
Data example: Allbus 2014



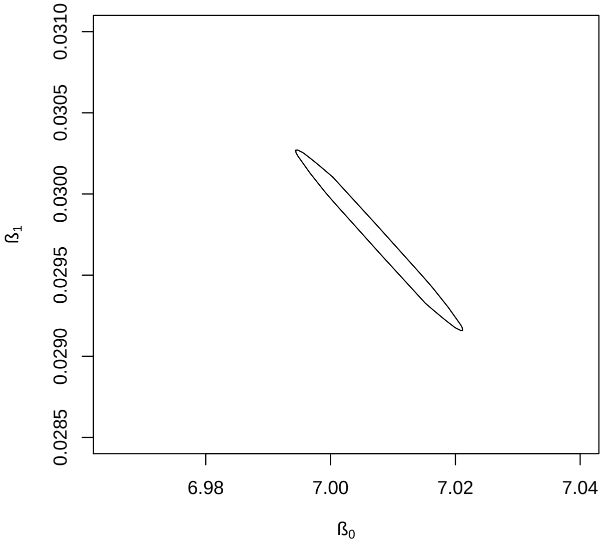
Data example: Allbus 2014



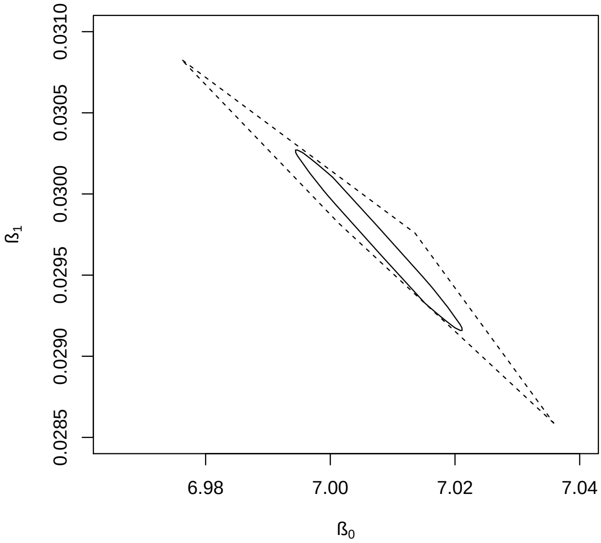
Data example: Allbus 2014



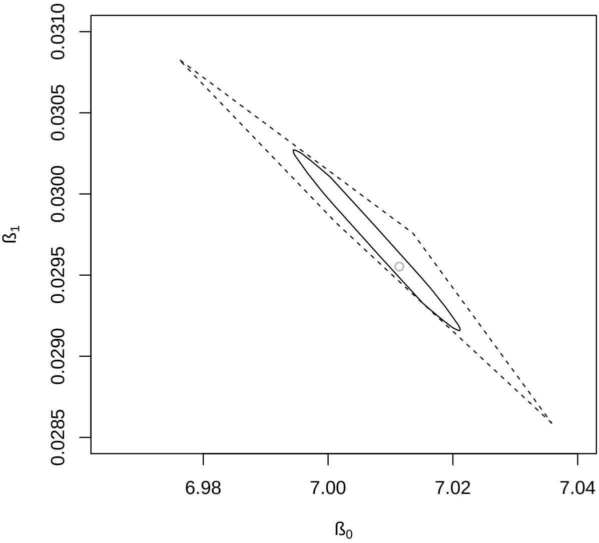
Data example: Allbus 2014



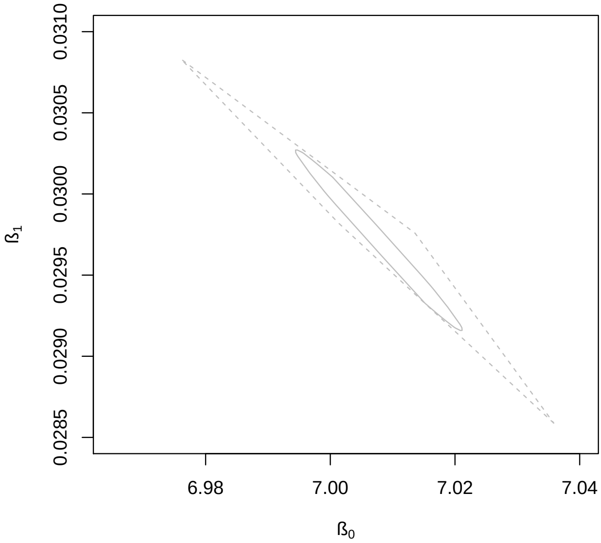
Data example: Allbus 2014



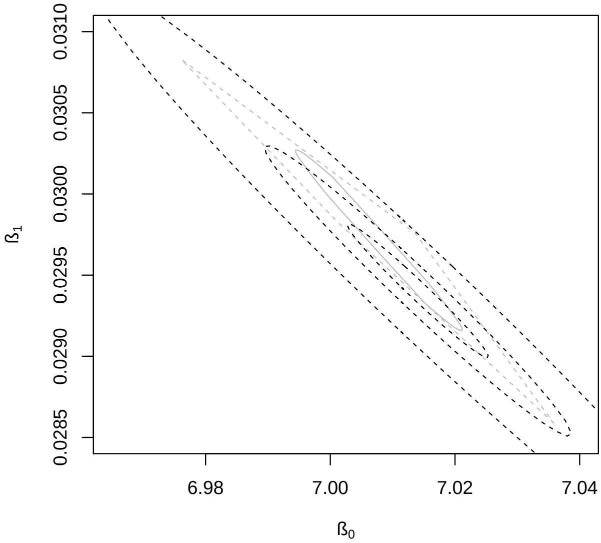
Data example: Allbus 2014



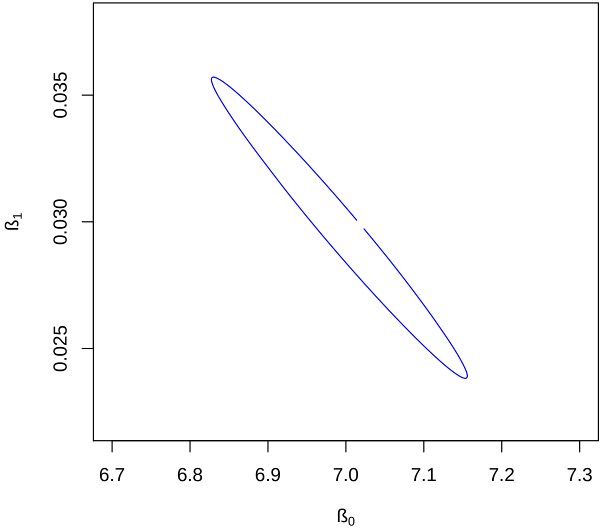
Data example: Allbus 2014



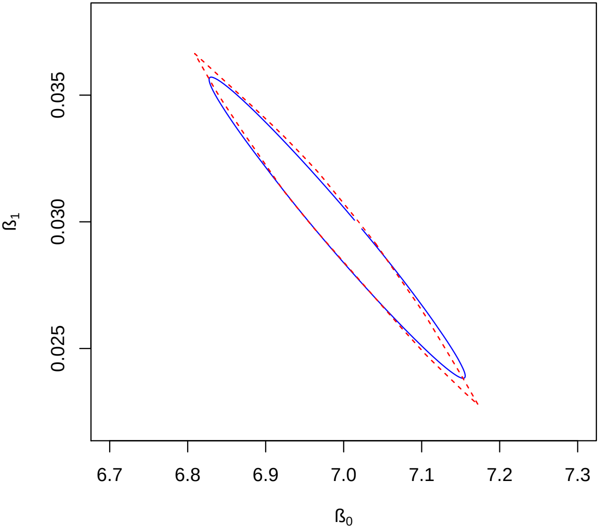
Data example: Allbus 2014



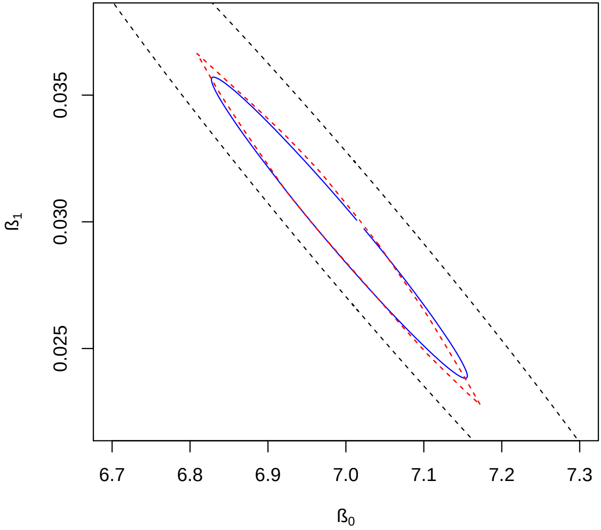
Data example: Allbus 2014



Data example: Allbus 2014



Data example: Allbus 2014



0.035
0.030
0.025
ß1
6.7
6.8
6.9
7.0
7.1
7.2
7.3
ß0