Regression with Imprecise Data:
A Robust Approach

Marco Cattaneo and Andrea Wiencierz
Department of Statistics, LMU Munich

July 25, 2011
ISIPTA ’11, Innsbruck, Austria
Consider data on two variables, $X$ and $Y$. 

![Scatter plot of X vs Y]
Regression Analysis

- Consider data on two variables, $X$ and $Y$.
- The aim is to investigate the relationship between $X$ and $Y$. 
The relationship between $X$ and $Y$ is described by:

$$Y = f(X) = a + bX,$$

$a, b \in \mathbb{R}$. 

For which $a$ and $b$ does the function $f$ best fit the data?
The relationship between $X$ and $Y$ is described by:

$$Y = f(X) = a + bX,$$

$a, b \in \mathbb{R}$.

For which $a$ and $b$ does the function $f$ best fit the data?
For a given $f$, the (absolute) residuals are defined by 
$R_{f,i} := |Y_i - f(X_i)|$. 
Introduction

Linear Regression 2

- For a given $f$, the (absolute) residuals are defined by $R_{f,i} := |Y_i - f(X_i)|$.
- Ordinary Least Squares: $f_{OLS}$ minimizes the mean of $R_{f,i}^2$. 
Linear Regression 2

- For a given $f$, the (absolute) residuals are defined by $R_{f,i} := |Y_i - f(X_i)|$.
- Ordinary Least Squares: $f_{OLS}$ minimizes the mean of $R_{f,i}^2$.
- Least Median of Squares: $f_{LMS}$ minimizes the median of $R_{f,i}^2$. 

![Graph showing linear regression with residuals indicated]
Imprecise Data

- Observation spaces of $X$ and $Y$ are partitioned into disjoint intervals.

![Graph](image.png)
Imprecise Data

- Observation spaces of $X$ and $Y$ are partitioned into disjoint intervals.
- Rectangular data: $[X_i, \overline{X}_i) \times [Y_i, \overline{Y}_i)$. 

![Graph showing rectangular data distribution](image)
Imprecise Data

- Observation spaces of $X$ and $Y$ are partitioned into disjoint intervals.
- Rectangular data: $[X_i, \overline{X}_i) \times [Y_i, \overline{Y}_i)$.
- How to draw a line that reflects the relationship between $X$ and $Y$?

![Graph showing rectangular data intervals and a line reflecting the relationship between X and Y]
Imprecise Data

- Observation spaces of $X$ and $Y$ are partitioned into disjoint intervals.
- Rectangular data: $[X_i, \bar{X}_i) \times [Y_i, \bar{Y}_i)$.
- How to draw a line that reflects the relationship between $X$ and $Y$?
- Common simple method: OLS based on interval midpoints.
Imprecise Data

- Observation spaces of $X$ and $Y$ are partitioned into disjoint intervals.
- Rectangular data: $[X_i, \overline{X_i}] \times [Y_i, \overline{Y_i}]$.
- How to draw a line that reflects the relationship between $X$ and $Y$?
- Common simple method: OLS based on interval midpoints.
- Further approaches: e.g. Domingues et al. (2010) or Ferson et al. (2007).
A Robust Approach to Regression with Imprecise Data

- Theoretical framework: likelihood-based decisions (Cattaneo, 2007).
A Robust Approach to Regression with Imprecise Data

- Theoretical framework: likelihood-based decisions (Cattaneo, 2007).
- We assume that the variables have precise values, which are imprecisely observed:

\[ V_i := (X_i, Y_i) \quad \text{and} \quad V_i^* := [X_i, \overline{X}_i) \times [Y_i, \overline{Y}_i), \quad i = 1, \ldots, n. \]
A Robust Approach to Regression with Imprecise Data

- Theoretical framework: likelihood-based decisions (Cattaneo, 2007).
- We assume that the variables have precise values, which are imprecisely observed:

\[ V_i := (X_i, Y_i) \quad \text{and} \quad V_i^* := [X_i, \bar{X}_i] \times [Y_i, \bar{Y}_i], \quad i = 1, \ldots, n. \]

- Nonparametric probability model:

\[ \mathcal{P} := \{ P : (V_i, V_i^*), i = 1, \ldots, n, \text{i.i.d.} \land P(V_i \in V_i^*) \geq 1 - \varepsilon \}, \]

for some \( \varepsilon \in [0, 1] \).
A Robust Approach to Regression with Imprecise Data

- Theoretical framework: likelihood-based decisions (Cattaneo, 2007).
- We assume that the variables have precise values, which are imprecisely observed:
  \[ V_i := (X_i, Y_i) \text{ and } V_i^* := [X_i, \bar{X}_i] \times [Y_i, \bar{Y}_i], \quad i = 1, \ldots, n. \]
- Nonparametric probability model:
  \[ \mathcal{P} := \{ P : (V_i, V_i^*), i = 1, \ldots, n, \text{i.i.d.} \land P(V_i \in V_i^*) \geq 1 - \varepsilon \}, \]
  for some \( \varepsilon \in [0, 1] \).
- Given \( V_1^*, \ldots, V_n^* \), we reduce \( \mathcal{P} \) via the likelihood function to the set
  \[ \mathcal{P}_{>\beta} := \{ P \in \mathcal{P} : lik(P) > \beta \}, \quad \text{for some (chosen) } \beta \in (0, 1). \]
A Robust Approach to Regression with Imprecise Data

- Theoretical framework: likelihood-based decisions (Cattaneo, 2007).
- We assume that the variables have precise values, which are imprecisely observed:
  \[ V_i := (X_i, Y_i) \text{ and } V_i^* := \left[ X_i, X_i \right] \times \left[ Y_i, Y_i \right], \quad i = 1, \ldots, n. \]
- Nonparametric probability model:
  \[ \mathcal{P} := \{ P : (V_i, V_i^*), i = 1, \ldots, n, \text{i.i.d.} \land P(V_i \in V_i^*) \geq 1 - \varepsilon \}, \]
  for some \( \varepsilon \in [0, 1] \).
- Given \( V_1^*, \ldots, V_n^* \), we reduce \( \mathcal{P} \) via the likelihood function to the set
  \[ \mathcal{P}_{>\beta} := \{ P \in \mathcal{P} : \text{lik}(P) > \beta \}, \text{ for some (chosen) } \beta \in (0, 1). \]
- The set \( \mathcal{P}_{>\beta} \) determines interval-valued estimates of the median of the (absolute) residuals \( R_{f,i} \) for the regression functions \( f \).
• For each regression function $f$, we have an interval-valued evaluation, which is a confidence interval for the median of $R_{f,i}$. 
Regression as a Decision Problem

- For each regression function $f$, we have an interval-valued evaluation, which is a confidence interval for the median of $R_{f,i}$.
- Interval dominance leads to a set of optimal regression functions.
Regression as a Decision Problem

- For each regression function $f$, we have an interval-valued evaluation, which is a confidence interval for the median of $R_{f,i}$.
- Interval dominance leads to a set of optimal regression functions.
- $(\Gamma-)\text{minimax}$ leads to one optimal regression function.
Result of Regression 1

- Regression analysis of the imprecise dataset shown before.
Result of Regression 1

- Regression analysis of the imprecise dataset shown before.
- We considered linear regression functions, \( f(X) = a + bX \).
Result of Regression 1

- Regression analysis of the imprecise dataset shown before.
- We considered linear regression functions, $f(X) = a + bX$.
- Calculations are based on a grid search.
Result of Regression 1

- Regression analysis of the imprecise dataset shown before.
- We considered linear regression functions, \( f(X) = a + b X \).
- Calculations are based on a grid search.
- The imprecision of the result mainly reflects the imprecision of the data.
We performed the same analysis on the dataset with imprecise observations of $X$, but precise data of $Y$. The result of the regression analysis is much more precise.
Result of Regression 2

- We performed the same analysis on the dataset with imprecise observations of $X$, but precise data of $Y$.
- The result of the regression analysis is much more precise.
Summary and Outlook

- We introduced a likelihood-based imprecise regression approach.
We introduced a likelihood-based imprecise regression approach. The approach is very general and the regression method covers many different settings:

- We can consider all kinds of imprecise data, not only disjoint intervals.
- The imprecise data can be wrong with a certain probability (\( \varepsilon > 0 \)).
- It is possible to consider more than one explanatory variable.
- There can be imprecision in dependent and explanatory variables at the same time.
- We can consider arbitrary regression functions, not only linear ones.
- Instead of the median we can use any other quantile.

The presented regression method yields very robust results.
Summary and Outlook

- We introduced a likelihood-based imprecise regression approach.
- The approach is very general and the regression method covers many different settings:
  - We can consider all kinds of imprecise data, not only disjoint intervals.
Summary and Outlook

- We introduced a likelihood-based imprecise regression approach.
- The approach is very general and the regression method covers many different settings:
  - We can consider all kinds of imprecise data, not only disjoint intervals.
  - The imprecise data can be wrong with a certain probability ($\varepsilon > 0$).
Summary and Outlook

- We introduced a likelihood-based imprecise regression approach.
- The approach is very general and the regression method covers many different settings:
  - We can consider all kinds of imprecise data, not only disjoint intervals.
  - The imprecise data can be wrong with a certain probability ($\varepsilon > 0$).
  - It is possible to consider more than one explanatory variable.
Concluding Remarks

Summary and Outlook

- We introduced a likelihood-based imprecise regression approach.
- The approach is very general and the regression method covers many different settings:
  - We can consider all kinds of imprecise data, not only disjoint intervals.
  - The imprecise data can be wrong with a certain probability ($\varepsilon > 0$).
  - It is possible to consider more than one explanatory variable.
  - There can be imprecision in dependent and explanatory variables at the same time.
Summary and Outlook

- We introduced a likelihood-based imprecise regression approach.
- The approach is very general and the regression method covers many different settings:
  - We can consider all kinds of imprecise data, not only disjoint intervals.
  - The imprecise data can be wrong with a certain probability ($\varepsilon > 0$).
  - It is possible to consider more than one explanatory variable.
  - There can be imprecision in dependent and explanatory variables at the same time.
  - We can consider arbitrary regression functions, not only linear ones.
Summary and Outlook

- We introduced a likelihood-based imprecise regression approach.
- The approach is very general and the regression method covers many different settings:
  - We can consider all kinds of imprecise data, not only disjoint intervals.
  - The imprecise data can be wrong with a certain probability ($\epsilon > 0$).
  - It is possible to consider more than one explanatory variable.
  - There can be imprecision in dependent and explanatory variables at the same time.
  - We can consider arbitrary regression functions, not only linear ones.
  - Instead of the median we can use any other quantile.
Summary and Outlook

- We introduced a likelihood-based imprecise regression approach.
- The approach is very general and the regression method covers many different settings:
  - We can consider all kinds of imprecise data, not only disjoint intervals.
  - The imprecise data can be wrong with a certain probability ($\varepsilon > 0$).
  - It is possible to consider more than one explanatory variable.
  - There can be imprecision in dependent and explanatory variables at the same time.
  - We can consider arbitrary regression functions, not only linear ones.
  - Instead of the median we can use any other quantile.
- The presented regression method yields very robust results.
