ANOVA, ANCOVA and Time Trends Modeling: Solving Statistical Problems Using Interval Analysis

Sergei Zhilin

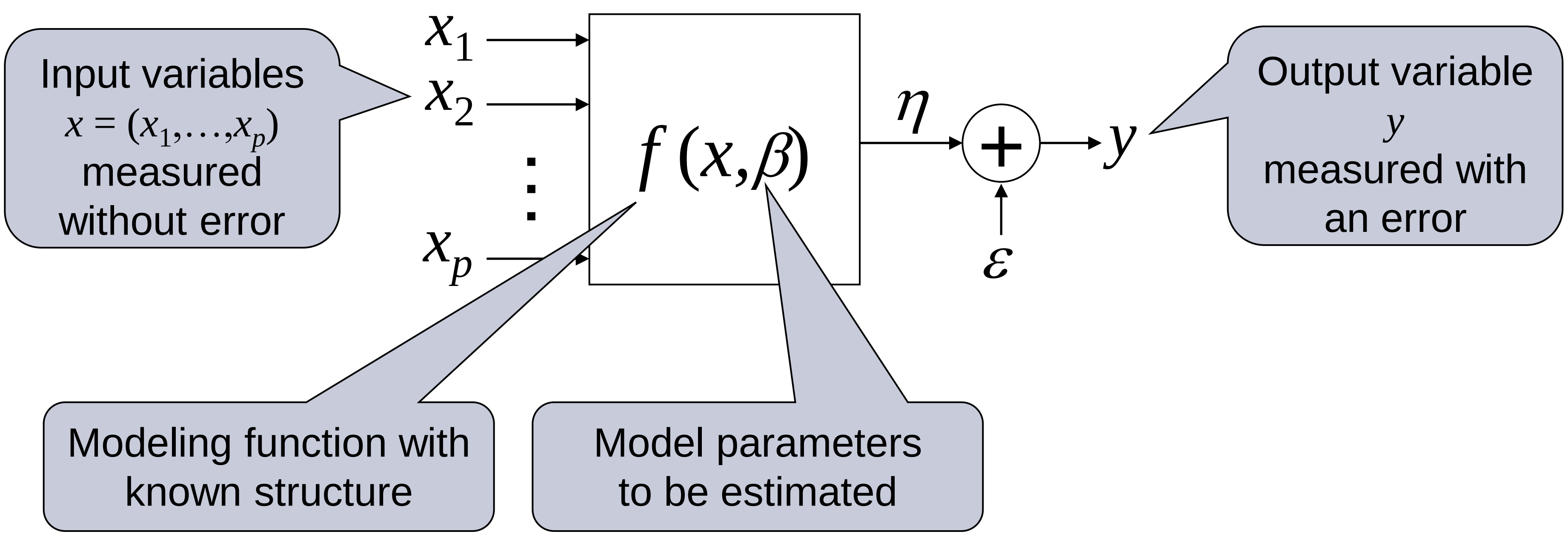
Altai State University, Barnaul, Russia

Outline

- Linear regression under interval error
- ANOVA and ANCOVA using interval regression
- Time Trends Modeling using interval regression
- Conclusions

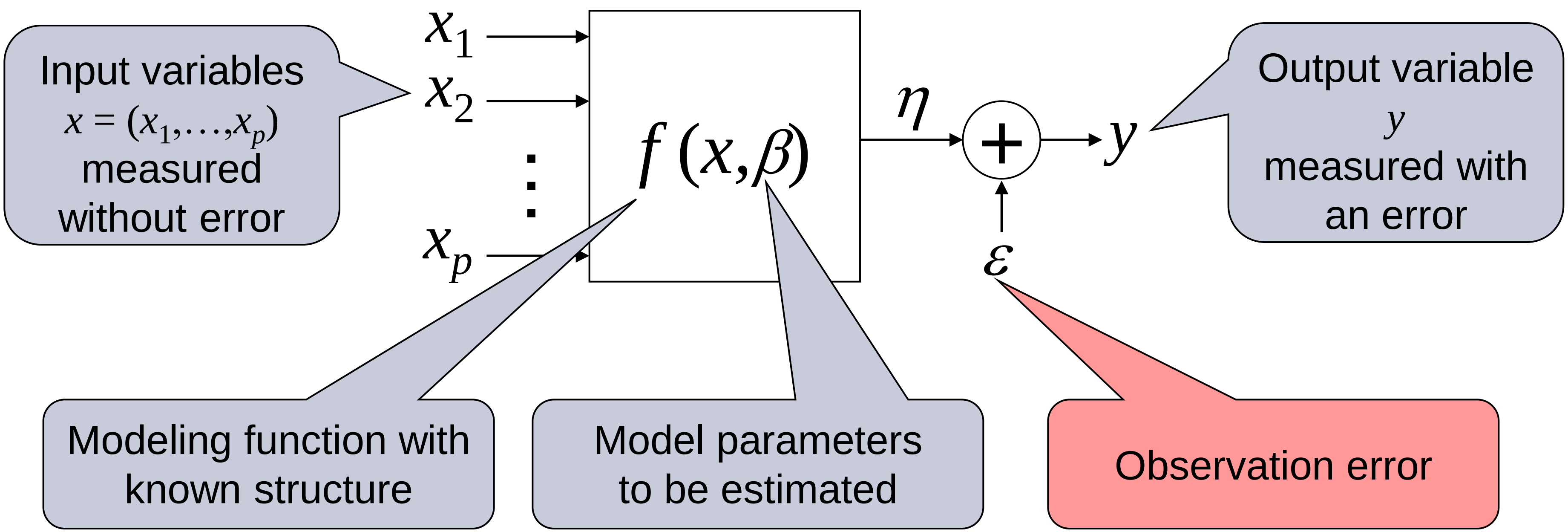
Linear Regression under Interval Error

- Black box approach



Linear Regression under Interval Error

- Black box approach



Linear Regression under Interval Error

- Classical statistical approach often assumes that the measurement error is Gaussian

- In many real-life applications the error is rather interval than Gaussian

- “Interval” means “unknown but bounded”:
  - $\varepsilon \in [-\bar{\varepsilon}, \bar{\varepsilon}]$, where $\bar{\varepsilon}$ is upper bound of error
  - There are no other assumptions about the error

Linear Regression under Interval Error

- The structure of the modeling function $f(x, \beta)$ is assumed fixed

- Each row (x_j, y_j) of the measurements table constrains possible values of the parameter β with the set

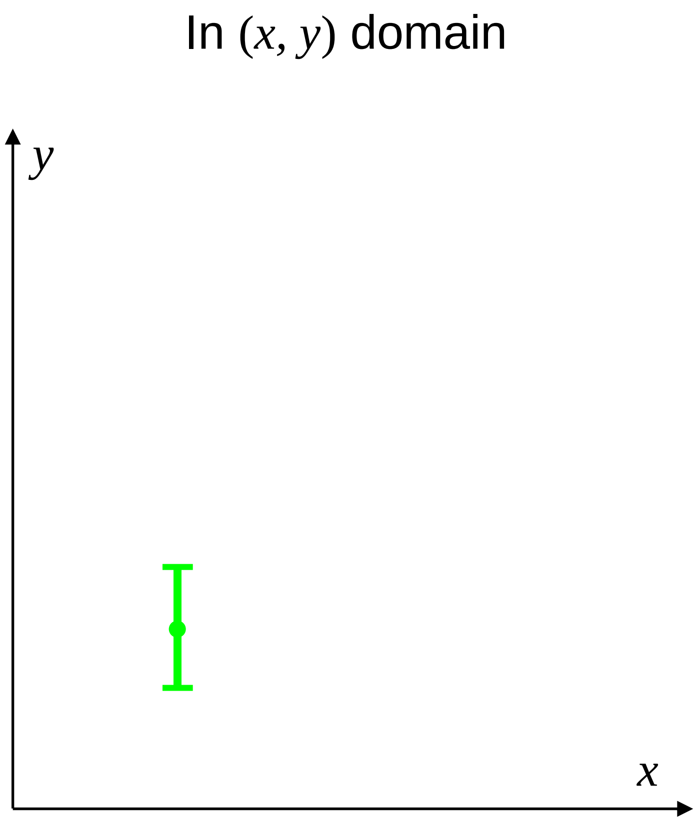
$$B_j = \left\{ \beta \;\middle|\; y_j - \bar{\varepsilon} \le f(x_j, \beta) \le y_j + \bar{\varepsilon} \right\}, \quad j = 1, \ldots, n.$$

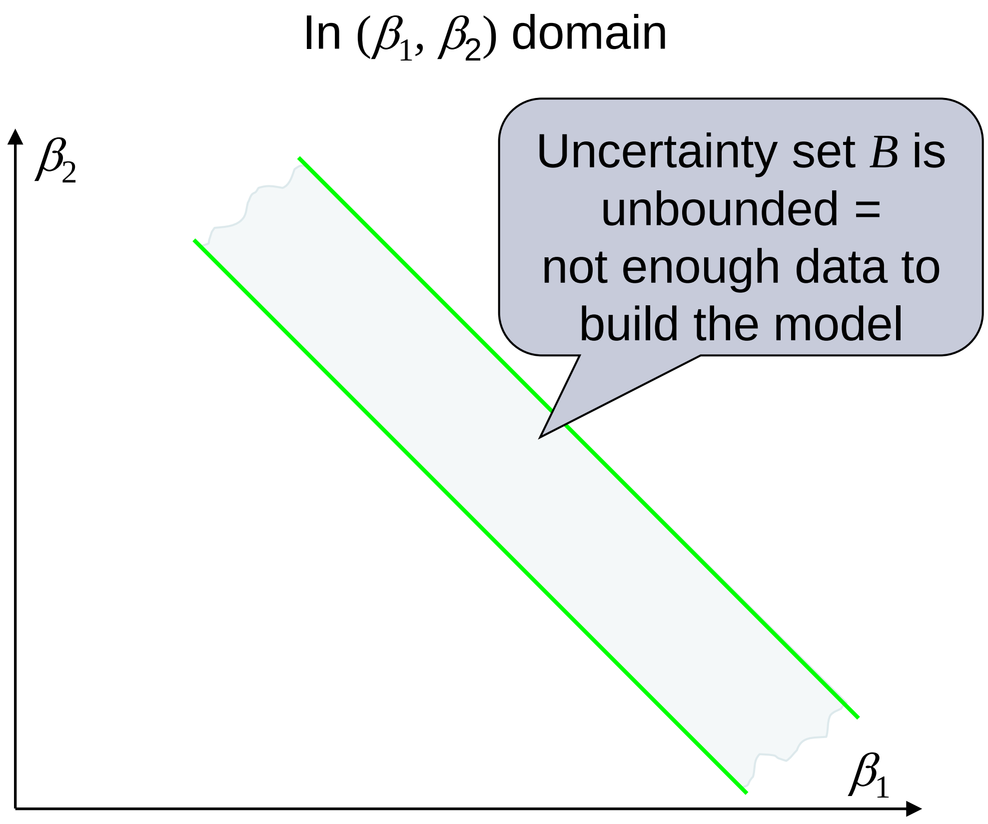
- Values of the parameter β consistent with all constraints form the uncertainty set

$$B = \bigcap_{j=1}^{n} B_j$$

Linear Regression under Interval Error

- Fitting data with the model $y = \beta_1 + \beta_2 x$

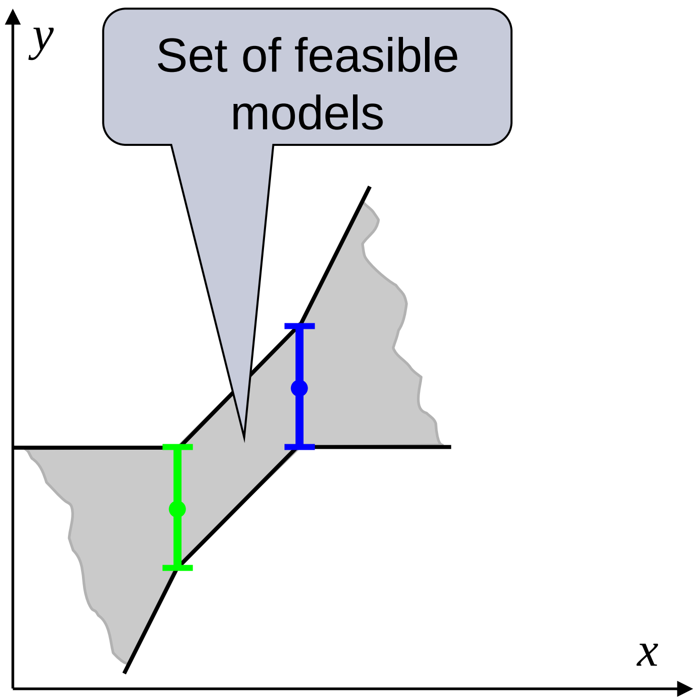




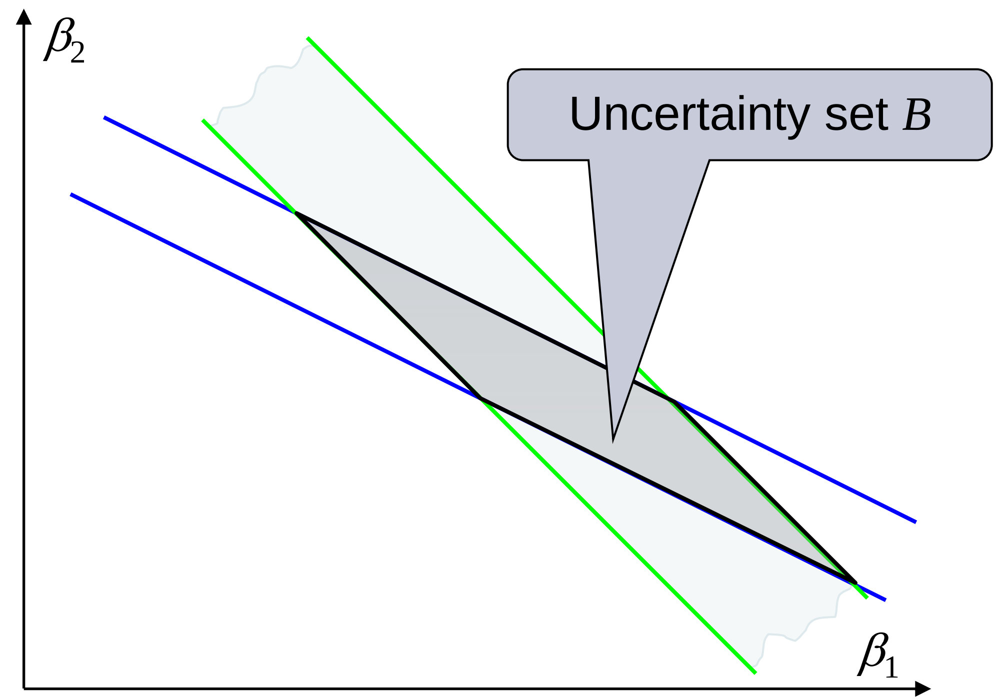
Linear Regression under Interval Error

- Fitting data with the model $y = \beta_1 + \beta_2 x$

In (x, y) domain



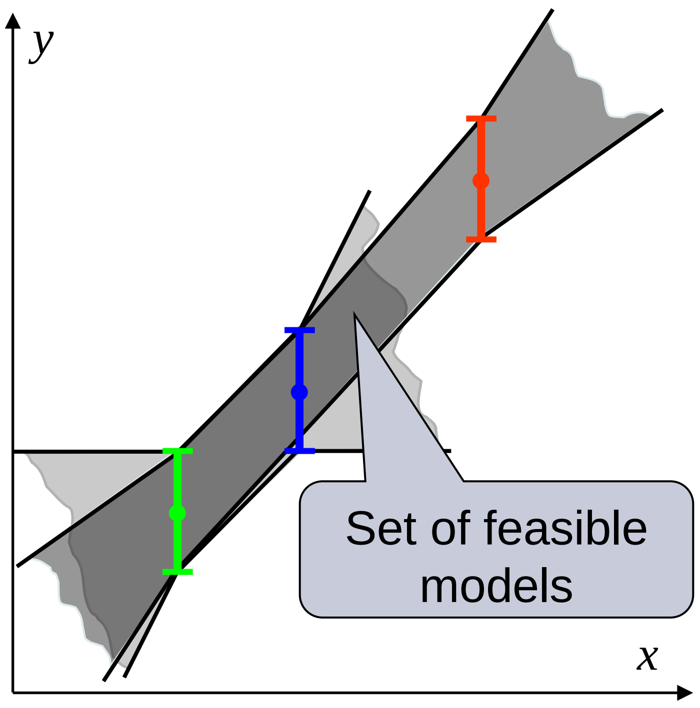
In (β_1, β_2) domain



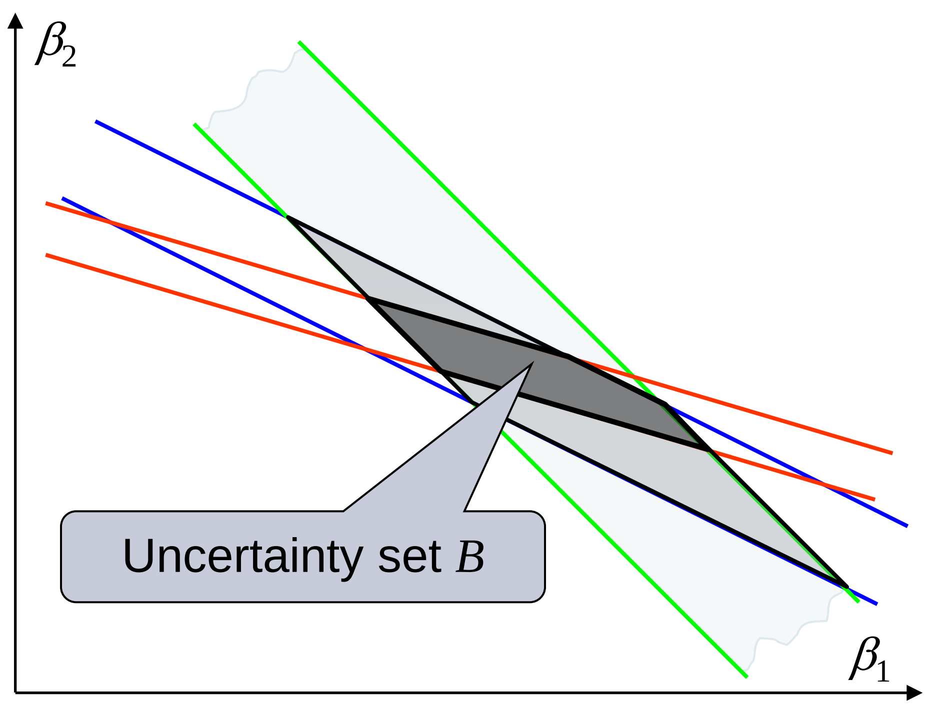
Linear Regression under Interval Error

- Fitting data with the model $y = \beta_1 + \beta_2 x$

In (x, y) domain



In (β_1, β_2) domain



Linear Regression under Interval Error

- Problems stated with respect to uncertainty set B
  - Prediction of the response value for fixed values of input variables
    - Interval estimates of y

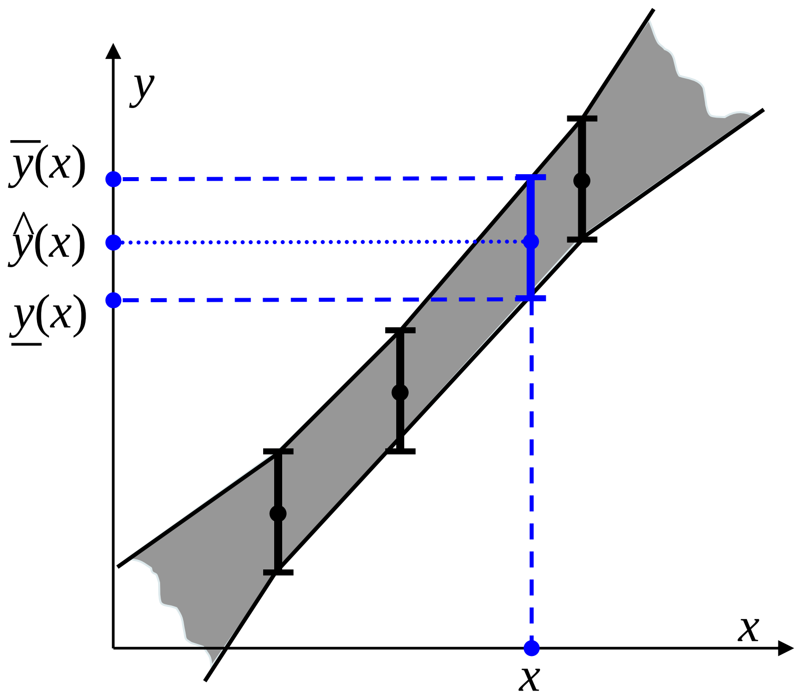
$$\boldsymbol{y}(x) = \left[\underline{y}(x), \overline{y}(x)\right]:$$

$$\underline{y}(x) = \min_{\beta \in B} \beta^T x,$$

$$\overline{y}(x) = \max_{\beta \in B} \beta^T x,$$

    - Point estimates of y

$$\hat{y}(x) = \tfrac{1}{2}\left(\underline{y}(x) + \overline{y}(x)\right)$$



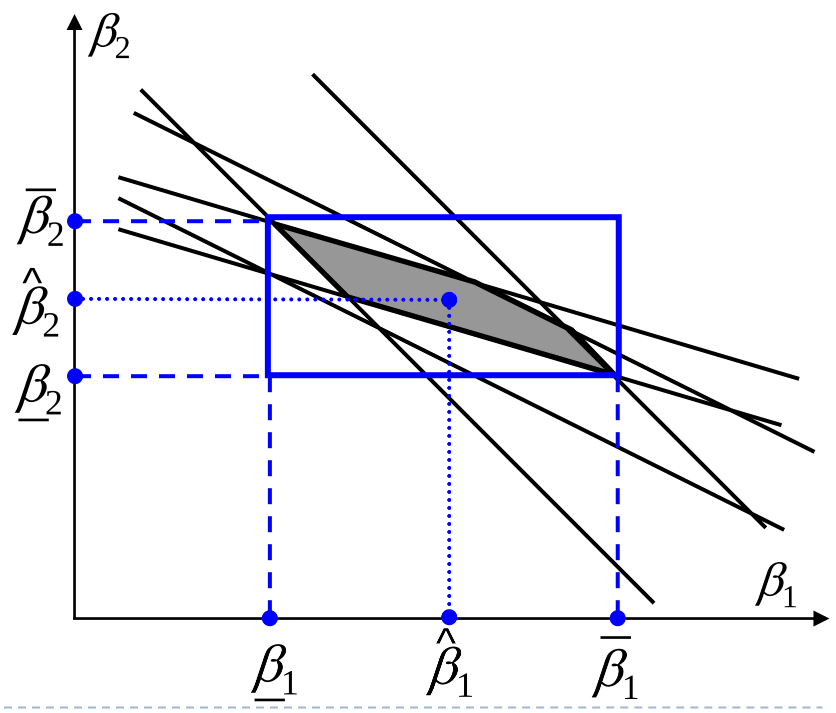
Linear Regression under Interval Error

- Problems stated with respect to uncertainty set B
  - Model parameters estimation
    - Interval estimates of β

$$\square B = \left(\left[\underline{\beta}_1, \overline{\beta}_1 \right], \ldots, \left[\underline{\beta}_p, \overline{\beta}_p \right] \right):$$

$$\underline{\beta}_i = \min_{\beta \in B} \beta_i, \quad \overline{\beta}_i = \max_{\beta \in B} \beta_i,$$

$$i = 1, \ldots, p.$$



    - Point estimates of β

$$\hat{\beta} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p \right):$$

$$\hat{\beta}_i = \tfrac{1}{2} \left(\underline{\beta}_i + \overline{\beta}_i \right), \quad i = 1, \ldots, p.$$

Minimal feasible error bound

- Shrink error bound until $B = \emptyset$.
  - $\varepsilon_{\min}$ is a minimal feasible error bound
  - $B_{\varepsilon_{\min}}$ is uncertainty set corresponding to $\varepsilon_{\min}$
- Take the center of $\square B_{\varepsilon_{\min}}$ as a point estimate of β
- If data are inconsistent ($B = \emptyset$) for certain model structure and $\bar{\varepsilon}$ value
  - Expand error bound until $B \neq \emptyset$
  - Analyze $\varepsilon_{\min}$ and boundary samples to detect outliers or to correct model structure

Linear Regression under Interval Error

- Testing hypothesis

$$C_{i_1}\beta_{i_1} + C_{i_2}\beta_{i_2} + \ldots + C_{i_k}\beta_{i_k} = C$$

  is equivalent to checking

$$B \cup \left\{ C_{i_1}\beta_{i_1} + C_{i_2}\beta_{i_2} + \ldots + C_{i_k}\beta_{i_k} = C \right\} = \varnothing$$

- Testing hypothesis about significance of β_i is equivalent to checking

$$0 \in \left[\underline{\beta_i}, \overline{\beta_i} \right]$$

Fitting experimental data under unknown-but-bounded error

- Years and Authors
  - 1962 L.V. Kantorovich
  - 1970 S.I. Spivak et al.
  - 1982 G. Belforte, M. Milanese et al.
  - 1983 N.M. Oskorbin et al.
  - 1986 J.P. Norton
  - 1987 S.I. Kumkov et al.
  - 1987 E. Walter, H. Piet-Lahanier
  - 1989 A.P. Voshchinin et al.
  - 1993 P.L. Combettes
  - 2000 O.E. Rodionova, A.L. Pomerantsev
  - 2003 A.A. Podruzhko, A.S. Podruzhko

Analysis of variance (ANOVA)

- The purpose of ANOVA is to test for significant differences between means in different groups of observations
- ANOVA model can be expressed as a regression model with categorical predictors
  - ANCOVA model mixes categorical and quantitative predictors
- Then the essence of the problem becomes to test significance of regression coefficients
- Some statisticians propose to replace hypothesis testing with the providing of confidence intervals

Linear models with categorical predictors

- General model

$$y = \sum_{i=0}^{p} \beta_i x_i \qquad x_i \in X \subseteq R, \quad i = 0,\dots,q-1, \quad \text{– quantitative variables}$$

$$x_i \in \{x_{i1},\dots,x_{iL_i}\}, \quad i = q,\dots,p \quad \text{– categorical variables}$$

- Coding values of categorical variables
  - Model with dummy variables

$$y = \sum_{i=0}^{q-1} \beta_i x_i + \sum_{i=q}^{p} \sum_{k=1}^{L_i-1} \delta_{ik} d_{ik}$$

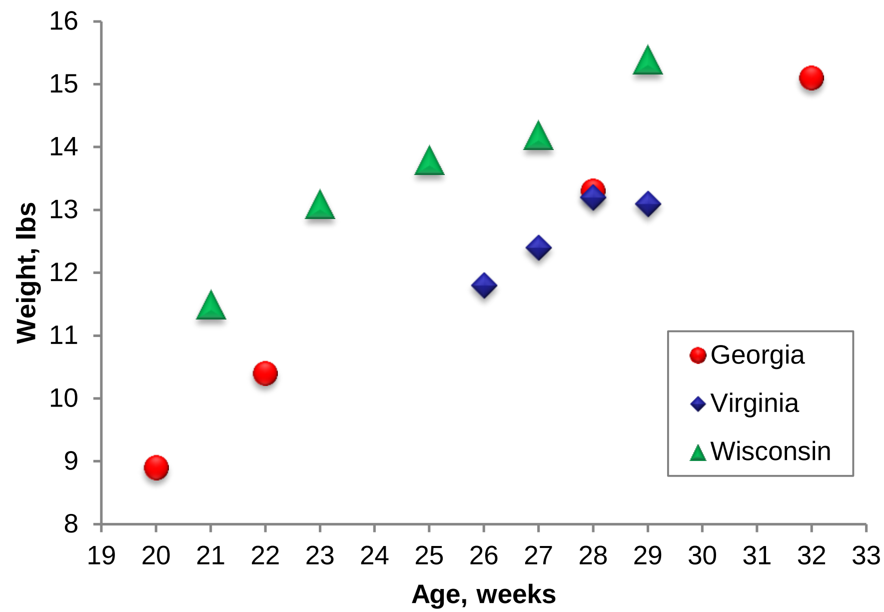
Dummy variables

Levels of cat. variable

x_i	d_{i1}	d_{i2}	d_{i3}		$d_{i(L_i-1)}$
x_{i1}	0	0	0	…	0
x_{i2}	1	0	0	…	0
x_{i3}	0	1	0	…	0
x_{i4}	0	0	1	…	0
:	:	:	:	…	:
x_{iL_i}	0	0	0	…	1

Turkey dataset

#	Weight, lbs	Age, weeks	Origin
1	13.3	28	Georgia
2	8.9	20	Georgia
3	15.1	32	Georgia
4	10.4	22	Georgia
5	13.1	29	Virginia
6	12.4	27	Virginia
7	13.2	28	Virginia
8	11.8	26	Virginia
9	11.5	21	Wisconsin
10	14.2	27	Wisconsin
11	15.4	29	Wisconsin
12	13.1	23	Wisconsin
13	13.8	25	Wisconsin

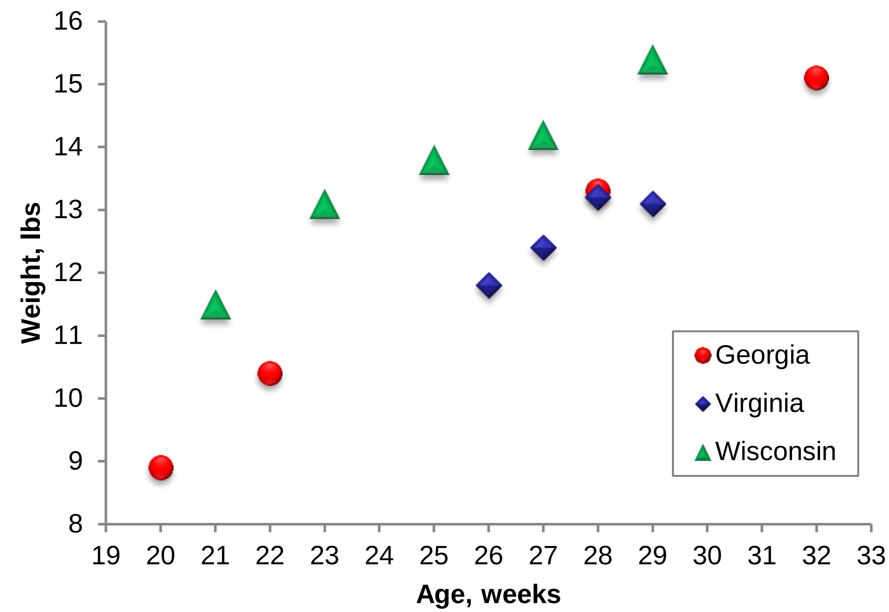


$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$\varepsilon_{\min} = 1.4$$

Turkey dataset

#	Weight, lbs	Age, weeks	Origin	d_1	d_2
1	13.3	28	Georgia	1	0
2	8.9	20	Georgia	1	0
3	15.1	32	Georgia	1	0
4	10.4	22	Georgia	1	0
5	13.1	29	Virginia	0	1
6	12.4	27	Virginia	0	1
7	13.2	28	Virginia	0	1
8	11.8	26	Virginia	0	1
9	11.5	21	Wisconsin	0	0
10	14.2	27	Wisconsin	0	0
11	15.4	29	Wisconsin	0	0
12	13.1	23	Wisconsin	0	0
13	13.8	25	Wisconsin	0	0

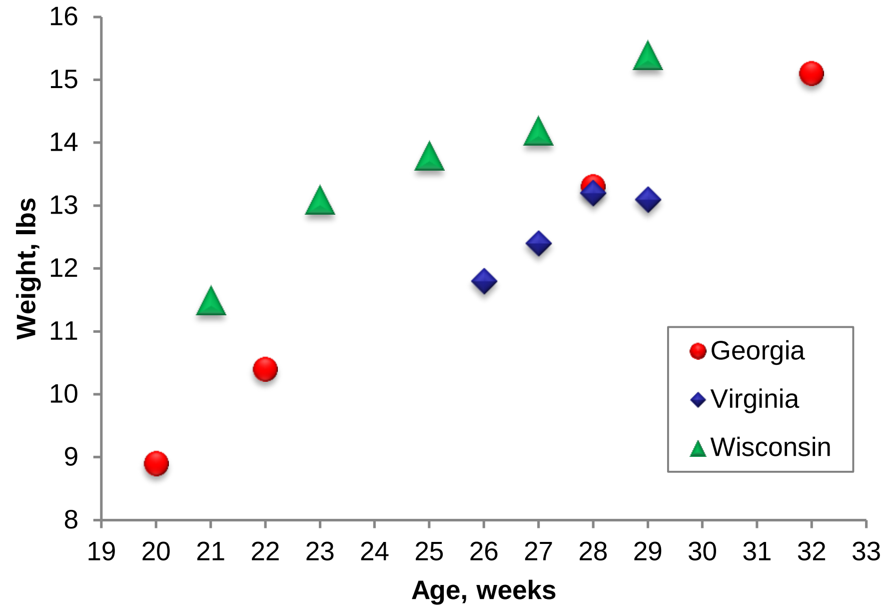


$$y = \beta_0 + \beta_1 x + \delta_1 d_1 + \delta_2 d_2 + \varepsilon$$

$$\varepsilon_{\min} = 0.37$$

Turkey dataset

#	Weight, lbs	Age, weeks	Origin	d_1	d_2
1	13.3	28	Georgia	1	0
2	8.9	20	Georgia	1	0
3	15.1	32	Georgia	1	0
4	10.4	22	Georgia	1	0
5	13.1	29	Virginia	0	1
6	12.4	27	Virginia	0	1
7	13.2	28	Virginia	0	1
8	11.8	26	Virginia	0	1
9	11.5	21	Wisconsin	0	0
10	14.2	27	Wisconsin	0	0
11	15.4	29	Wisconsin	0	0
12	13.1	23	Wisconsin	0	0
13	13.8	25	Wisconsin	0	0



$$y = \beta_0 + \beta_1 x + \delta_1 d_1 + \delta_2 d_2 + \varepsilon$$

$$\bar{\varepsilon} = 1$$

$$\hat{\beta}_0 \in [-3.31,\ 5.15] \quad \hat{\beta}_1 \in [0.35,\ 0.67] \quad \hat{\delta}_1 \in [-3.38, -0.65] \quad \hat{\delta}_2 \in [-3.60, -0.45]$$

Turkey dataset

- General model

$$y = \beta_0 + \beta_1 x + \delta_1 d_{\mathbf{1}} + \delta_2 d_{\mathbf{2}} + \varepsilon \qquad \bar{\varepsilon} = 1$$

$$\hat{\beta}_0 \in [-3.31,\ 5.15] \qquad \hat{\beta}_1 \in [0.35,\ 0.67]$$

$$\hat{\delta}_1 \in [-3.38, -0.65] \qquad \hat{\delta}_2 \in [-3.60, -0.45]$$

- Individual models
  - For Georgia $\hat{\beta}_0 \in [-6.69,\ 5.15]$ $\hat{\beta}_1 \in [0.35,\ 0.67]$
  - For Virginia $\hat{\beta}_0 \in [-6.91,\ 4.7]$ $\hat{\beta}_1 \in [0.35,\ 0.67]$
  - For Wisconsin $\hat{\beta}_0 \in [-3.31,\ 5.15]$ $\hat{\beta}_1 \in [0.35,\ 0.67]$

,

Taking into account time trends

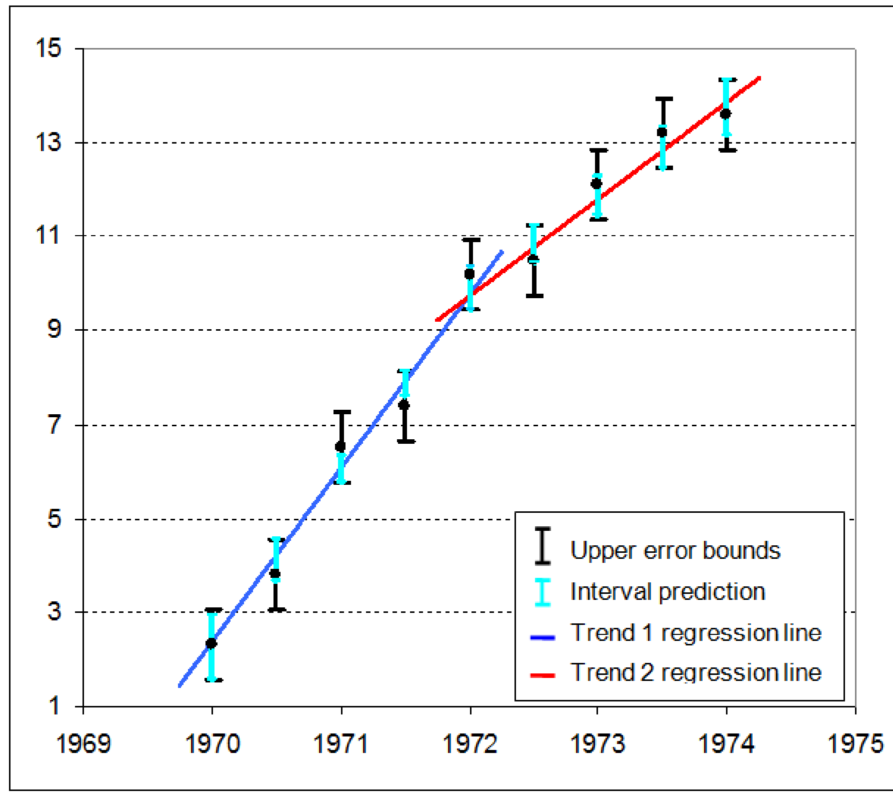
X – Time, years
Y – Response
d_i – Dummy variables
$\bar{\varepsilon} = 0.75$

$$y = \delta_0 + \delta_1 d_1 + \delta_2 d_2 + \varepsilon$$

δ_0 is the value of y in the common point

δ_1, δ_2 are angular coefficients of trend lines

X	Y	d_1	d_2
1970	2.3	– 4	0
1970½	3.8	– 3	0
1971	6.5	– 2	0
1971½	7.4	– 1	0
1972	10.2	0	0
1972½	10.5	0	1
1973	12.1	0	2
1973½	13.2	0	3
1974	13.6	0	4



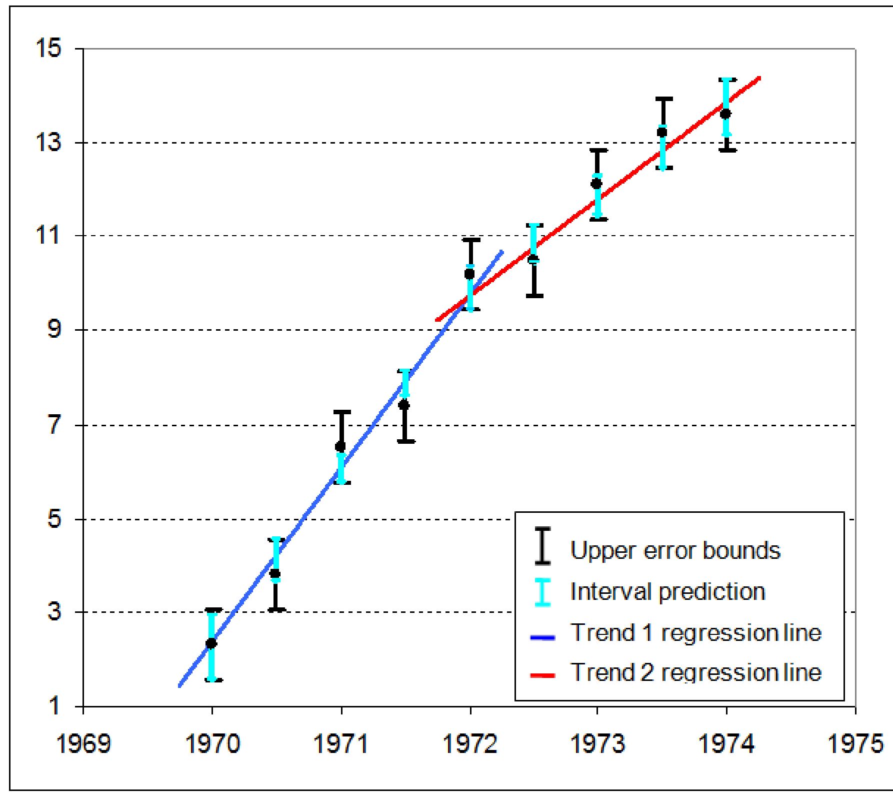
Taking into account time trends

X – Time, years
Y – Response
d_i – Dummy variables
$\bar{\varepsilon} = 0.75$

$$y = \delta_0 + \delta_1 d_1 + \delta_2 d_2 + \varepsilon$$

$$\hat{\delta}_0 \in [9.45,\ 10.35] \quad \hat{\delta}_1 \in [1.63,\ 2.20] \quad \hat{\delta}_2 \in [0.70,\ 1.23]$$

X	Y	d_1	d_2
1970	2.3	– 4	0
1970½	3.8	– 3	0
1971	6.5	– 2	0
1971½	7.4	– 1	0
1972	10.2	0	0
1972½	10.5	0	1
1973	12.1	0	2
1973½	13.2	0	3
1974	13.6	0	4



Conclusions

- Using linear regression under interval error with categorical predictors one can solve ANOVA/ANCOVA type problems and time trends modeling problems.

- Proposed approach provides flexible way to express and take into account *a priori* information (as supplemental constraints)

- Solving more complex variant of ANOVA problem and case studies is a challenge