# Imprecise regression analysis

Lev Utkin

Munich, September 2009

# Standard regression models

- Suppose that we have two variables $Y$ and $X$ with $Y$ being a dependent variable and $X$ being predictor variable, related to $Y$ according to the relation $Y = f(X, \mathbf{d}) + \epsilon$.
- The simplest case: the linear model $Y = bX + c + \epsilon$. Here $b$ and $c$ are parameters ($\mathbf{d} = (b, c)$) and $\epsilon$ is the random errors or the noise having zero mean and the unknown variance $\sigma^2$.
- A linear regression model fits a linear function to a set of data points. When the variable $X$ takes $n$ specific values $x_1, ..., x_n$, the variables $Y$ and $\epsilon$ take specific values $y_i$ and $\epsilon_i$, respectively, $i = 1, ..., n$, we get

$$y_i = bx_i + c + \epsilon_i, \ \ i = 1, ..., n.$$

# The first idea: maximum of the likelihood function over the set of CDFs (continuous case)

Denote $z_i = y_i - f(x_i, \mathbf{d})$ and $\mathbf{Z} = (z_1, ..., z_n)$.

1. Every r.v. $Z_i$ or $\epsilon_i$ is governed by an unknown CDF belonging to a set $\mathcal{M}_i(\mathbf{d})$ depending on a vector of parameters $\mathbf{d}$ and defined by **lower and upper CDFs**:

$$\underline{F}(z \mid \mathbf{d}) = \inf_{F(z) \in \mathcal{M}(\mathbf{d})} F(z), \ \overline{F}(z \mid \mathbf{d}) = \sup_{F(z) \in \mathcal{M}(\mathbf{d})} F(z).$$

2. The likelihood function $L(\mathbf{Z} \mid \mathbf{d}, F)$ is maximized over all distributions $F$ from $\mathcal{M}_i(\mathbf{d})$ and the resulting "modified" likelihood function depends on $\mathbf{d}$:

$$L(\mathbf{Z} \mid \mathbf{d}) = \max_{F \in \mathcal{M}(\mathbf{d})} L(\mathbf{Z} \mid \mathbf{d}, F).$$

## Proposition

*If random variables $Z_1, ..., Z_n$ are independent and continuous, then there holds*

$$\max_{\mathcal{M}} L(\mathbf{z}) = \prod_{i=1}^{n} \left\{ \overline{F}(z_i) - \underline{F}(z_i) \right\} \delta(z_i).$$

*Here $\delta(z)$ is Dirac function which has unit area concentrated in the immediate vicinity of point $z$.*

# Defining the lower and upper CDFs

**The second question:**

*How to define the functions* $\underline{F}(z \mid \mathbf{d})$ *and* $\overline{F}(z \mid \mathbf{d})$ *or the set* $\mathcal{M}(\mathbf{d})$*?*

**The first answer:**

*By using the imprecise Bayesian models!*

**The second answer:**

*By using the method of moments!*

**The third answer:**

*By using confidence intervals on the mean and variance!*

# The Bayesian normal model

Let $\epsilon \sim N(0, \sigma^2)$. Denote $\lambda = 1/\sigma^2$. The conjugate prior distribution for $\lambda$ is $\mathrm{Gamma}\left(\lambda | a, b\right)$.
The predictive density function is

$$p(x|a,b) = \frac{1}{\sqrt{\pi}} \frac{b^a}{\Gamma\left(a\right)} \int_{-\infty}^{\infty} \lambda^{a-1/2} \exp\left(-\left[\frac{x^2+2b}{2}\right]\lambda\right) \mathrm{d}\lambda.$$

If we denote

$$b^* = \frac{x^2+2b}{2},\ \ a^* = a + 1/2,$$

then

$$p(x|a,b) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(a^*\right)}{\Gamma\left(a\right)} \frac{b^a}{\left(b^*\right)^{a^*}}.$$

# Imprecise Bayesian normal model (1)

Replace $a$ and $b$ by parameters $s$ and $\gamma$ such that $a = (s+2)/2$ and $b = s\gamma/2$ (Quaeghebeur and de Cooman 2005).
The predictive density is

$$p(x|s,\gamma) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{s+3}{2}\right)}{\Gamma\left(\frac{s+2}{2}\right)} \frac{(s\gamma)^{\frac{s+2}{2}}}{(x^2 + s\gamma)^{\frac{s+3}{2}}}$$

Denote the parameters of the posterior distribution after having $k$ observations $s_k$ and $\gamma_k$:

$$s_k = s + k,\ \ \gamma_k = \frac{s\gamma + \tau_k}{s + k},\quad \tau_k = \sum_{j=1}^{k} z_j^2.$$

# Imprecise Bayesian normal model (2)

The predictive CDF is

$$F(z|s,\gamma) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{s+k+3}{2}\right)(s\gamma+\tau_k)^{\frac{s+k+2}{2}}}{\Gamma\left(\frac{s+k+2}{2}\right)} \times \int_{-\infty}^{z} \left(s\gamma+\tau_k+x^2\right)^{-\frac{s+k+3}{2}} \mathrm{d}x.$$

Properties of the predictive CDF:

1. if $\gamma_1 \geq \gamma_2$ and $z < 0$, then $F(z|s,\gamma_1) \geq F(z|s,\gamma_2)$,
2. if $\gamma_1 \geq \gamma_2$ and $z > 0$, then $F(z|s,\gamma_1) \leq F(z|s,\gamma_2)$,
3. if $\gamma_1 \geq \gamma_2$ and $z = 0$, then $F(z|s,\gamma_1) = F(z|s,\gamma_2) = 0.5$.

# Imprecise Bayesian normal model (3)

We take $\inf \gamma \to 0$ and $\sup \gamma = \overline{\gamma} = \max\left(z_1^2, ..., z_m^2\right)$. (ad-hoc rule).
Then the lower bound $\underline{F}^{(s)}(z)$ for the set $\mathcal{M}(\mathbf{d})$

$$\underline{F}^{(s)}(z) = \begin{cases} F(z|s, 0), & z < 0 \\ F(z|s, \overline{\gamma}), & z \geq 0 \end{cases}.$$

The upper bound $\overline{F}^{(s)}(z)$ for the set $\mathcal{M}(\mathbf{d})$

$$\overline{F}^{(s)}(z) = \begin{cases} F(z|s, \overline{\gamma}), & z < 0 \\ F(z|s, 0), & z \geq 0 \end{cases}.$$

## The imprecise Bayesian regression model

The logarithm of the likelihood function is

$$\ln L^{(s)}(\mathbf{Z} \mid \mathbf{d}) = \sum_{i=1}^{m} \ln\left(\overline{F}^{(s)}(y_i - \mathbf{x}_i\mathbf{d}) - \underline{F}^{(s)}(y_i - \mathbf{x}_i\mathbf{d})\right).$$

After some transformations:

$$\ln L^{(s)}(\mathbf{Z} \mid \mathbf{d}) = \sum_{i=1}^{m} \ln\left(H^{(s)}((y_i - \mathbf{x}_i\mathbf{d})^2 \,|0) - H^{(s)}((y_i - \mathbf{x}_i\mathbf{d})^2 \,|\overline{\gamma})\right).$$

In particular, if $\overline{\gamma} \longrightarrow \infty$, we get

$$\ln L^{(s)}(\mathbf{Z} \mid \mathbf{d}) = \sum_{j=1}^{m} \ln \int_0^{(y_j - \mathbf{x}_j\mathbf{d})^2} \frac{y^{-\frac{1}{2}} \tau_k^{\frac{s+k+2}{2}}}{(\tau_k + y)^{\frac{s+k+3}{2}}} \mathrm{d}y.$$

# The imprecise Bayesian regression model (simplification)

The logarithmic likelihood function can be simplified by taking some terms of a power series. By taking two terms of the power series of the function

$$h(y)=\left(\frac{y^{-\frac{1}{2}}\tau_k^{\frac{s+k+2}{2}}}{(\tau_k+y)^{\frac{s+k+3}{2}}}-\frac{y^{-\frac{1}{2}}\left(s\overline{\gamma}+\tau_k\right)^{\frac{s+k+2}{2}}}{(s\overline{\gamma}+\tau_k+y)^{\frac{s+k+3}{2}}}\right),$$

we get

$$\ln L^{(s)}(\mathbf{Z}\mid\mathbf{d})=\sum_{j=1}^{m}\ln\left(2\left(z_j^{1/2}-w_j^{1/2}\right)-\frac{(s+k+3)}{3}\left(z_j^{3/2}-w_j^{3/2}\right)\right)$$

$$z_j=\frac{(y_j-\mathbf{x}_j\mathbf{d})^2}{\tau_k},\quad w_j=\frac{(y_j-\mathbf{x}_j\mathbf{d})^2}{s\overline{\gamma}+\tau_k}.$$

# The imprecise Bayesian regression model

The logarithm of the likelihood function is

$$\ln L^{(s)}(\mathbf{Z} \mid \mathbf{d}) = \sum_{i=1}^{m} \ln \left( \overline{F}^{(s)}(y_i - \mathbf{x}_i\mathbf{d}) - \underline{F}^{(s)}(y_i - \mathbf{x}_i\mathbf{d}) \right).$$

After some transformations:

$$\ln L^{(s)}(\mathbf{Z} \mid \mathbf{d}) = \sum_{i=1}^{m} \ln \left( H^{(s)}((y_i - \mathbf{x}_i\mathbf{d})^2 \, |0) - H^{(s)}((y_i - \mathbf{x}_i\mathbf{d})^2 \, |\overline{\gamma}) \right).$$

In particular, if $\overline{\gamma} \longrightarrow \infty$, we get

$$\ln L^{(s)}(\mathbf{Z} \mid \mathbf{d}) = \sum_{j=1}^{m} \ln \int_0^{(y_j - \mathbf{x}_j\mathbf{d})^2} \frac{y^{-\frac{1}{2}} \tau_k^{\frac{s+k+2}{2}}}{(\tau_k + y)^{\frac{s+k+3}{2}}} \mathrm{d}y.$$

# The uniform distribution

$|Z_i|$ has the uniform distribution (UD)

$$p(z|\theta) = \begin{cases} 1/\theta, & 0 \leq z \leq \theta \\ 0, & \text{otherwise} \end{cases}.$$

The conjugate prior to the UD is the Pareto distribution $Pa(\theta|b, a)$. The predictive CDF by given $D = \max(z_1, ..., z_n)$ and $c = \max(b, D)$ is

$$F(z|\mathbf{Z}) = \begin{cases} \dfrac{z(a+n)}{(a+n+1)c}, & z \leq c \\ 1 - \dfrac{c^{a+n}}{z^{a+n}(a+n+1)}, & z > c \end{cases}.$$

# The imprecise Pareto-uniform model

The parameters $a$ and $b$ are replaced by $a = s + 1$ and $b = st$, $t \in [0, \infty)$.

The bounds for the predictive cumulative distribution function are

$$\underline{F}(x|\mathbf{X}) = 0,$$

$$\overline{F}(x|\mathbf{X}) = \begin{cases} \dfrac{x(s+n+1)}{(s+n+2)D}, & x \leq D \\ 1 - \dfrac{D^{s+n+1}}{x^{s+n+1}(s+n+2)}, & x > D \end{cases}.$$

# The imprecise Pareto-uniform regression model

The logarithm of the likelihood function is

$$\ln L^{(s)}(\mathbf{Z}|\ \mathbf{d}) = \sum_{i=1}^{n} \ln \left( \frac{|y_i - \mathbf{x}_i \mathbf{d}| \cdot (s+n+1)}{(s+n+2)D} \right)$$
$$= n \cdot \ln \left( \frac{(s+n+1)}{(s+n+2)D} \right) + \sum_{i=1}^{n} \ln \left( |y_i - \mathbf{x}_i \mathbf{d}| \right)$$

*The parameters* **d** *of the regression model do not depend on the caution parameter s of the imprecise model.*

# The standard method of moments

The *method of moments* is a technique for constructing estimators of the parameters that is based on matching the sample moments with the corresponding distribution moments.

Let $\mu_i(\theta)$ be the $i$-th moment of $Z$: $\mu_i(\theta) = \mathbb{E}_\theta(Z^i)$, $i = 1, .., k$.

Let $M_i(Z)$ be the $i$-th sample moment: $M_i(\mathbf{Z}) = \frac{1}{n}\sum_{j=1}^{n} Z_j^i$, $i = 1, .., k$.

To construct estimators $(W_1, ..., W_k)$ for parameters $(\theta_1, ..., \theta_k)$, we solve the set of equations

$$\mu_i(W_1, ..., W_k) = M_i(Z_1, ..., Z_n), \ \ i = 1, .., k$$

for $(W_1, ..., W_k)$ in terms of $(Z_1, ..., Z_n)$.

# The imprecise method of moments

By having $k$ moments, we can restrict a set of probability distributions (or pdfs) by the constraints:

$$\mathbb{E}(Z^i) = M_i(\mathbf{Z} \mid \mathbf{d}),\ \ i = 1, ..., k,$$

or

$$\sum\nolimits_{j=1}^{N} \pi(v_j) v_j^i = \frac{1}{n} \sum\nolimits_{j=1}^{n} z_j^i = \frac{1}{n} \sum\nolimits_{j=1}^{n} \left(y_j - f\left(\mathbf{x}_j, \mathbf{d}\right)\right)^i, \\ i = 1, ..., k.$$

Here $\pi \in \mathcal{M}$. In other words, the set of sample moments produces the set $\mathcal{M}$.

# The imprecise method of moments and the likelihood function

## Proposition

*Suppose that the initial information about the i-th discrete random variable $Z_i$ produces a set $\mathcal{M}_i$ of probability distributions $\pi(z)$, $i = 1, ..., n$. If $Z_1, ..., Z_n$ are independent, then there holds*

$$\max_{\mathcal{M}_1,...,\mathcal{M}_n} \Pr\{Z_1 = z_1, ..., Z_n = z_n\} = \prod_{i=1}^{n} \overline{\pi}(z_i),$$

*where*

$$\overline{\pi}(z_i) = \max_{\mathcal{M}_i} \Pr\{Z_i = z_i\}.$$

# The imprecise regression model using the method of moments

The logarithmic likelihood function is

$$\max_{\mathcal{M}(\mathbf{d})} \ln L(\mathbf{Z} \mid \mathbf{d}) = \sum_{i=1}^{n} \ln \overline{\pi}(y_i - f(\mathbf{x}_i, \mathbf{d})).$$

Now the optimal vector $\mathbf{d}_0$ can be found from the following system of equations:

$$\frac{\partial \ln L(\mathbf{Z} \mid \mathbf{d})}{\partial d_i} = 0, \; i = 1, ..., m$$

$$\sum_{i=1}^{n} \frac{\partial \overline{\pi}(y_i - f(\mathbf{x}_i, \mathbf{d})) / \partial d_j}{\overline{\pi}(y_i - f(\mathbf{x}_i, \mathbf{d}))} = 0, \; j = 1, ..., m.$$

# A special case: two moments (modification of Chebyshev's inequality)

We take only first two moments $m_1 = M_1(\mathbf{Z})$ and $m_2 = M_2(\mathbf{Z})$.

**Proposition**

*Suppose that the first moment $m_1 = \mathbb{E}X$ and the second moment $m_2 = \mathbb{E}X^2$ of a continuous random variable $X$ defined on the sample space $\mathbb{R}$ are known. Then the upper probability of the event $t \leq X \leq t + \varepsilon$ is defined as*

$$\overline{P}(t, t+\varepsilon) = \begin{cases} 1, & t < m_1 < t + \varepsilon \\ \dfrac{m_2 - m_1^2}{(m_1 - t)^2 + m_2 - m_1^2}, & \text{otherwise} \end{cases}.$$

# A special case: two moments (2)

The logarithm of the likelihood function is

$$\max_{\mathcal{M}(\mathbf{d})} \ln L(\mathbf{Z} \mid \mathbf{d}) = \sum_{i=1}^{n} \ln \frac{m_2(\mathbf{d}) - m_1^2(\mathbf{d})}{(m_1(\mathbf{d}) - y_i - f(\mathbf{x}_i, \mathbf{d}))^2 + m_2(\mathbf{d}) - m_1^2(\mathbf{d})},$$

where

$$m_1(\mathbf{d}) = \frac{1}{n} \sum_{j=1}^{n} (y_j - f(\mathbf{x}_j, \mathbf{d})),$$

$$m_2(\mathbf{d}) = \frac{1}{n} \sum_{j=1}^{n} (y_j - f(\mathbf{x}_j, \mathbf{d}))^2.$$

# Confidence intervals on the mean and variance

Confidence intervals on the mean and variance:

$$[\underline{m}_1(\mathbf{d}), \overline{m}_1(\mathbf{d})] = \left[ m_1(\mathbf{d}) - \frac{t_{\alpha/2,N-1}\hat{\sigma}(\mathbf{d})}{\sqrt{N}},\ m_1(\mathbf{d}) + \frac{t_{\alpha/2,N-1}\hat{\sigma}(\mathbf{d})}{\sqrt{N}} \right],$$

$$[\underline{\sigma}^2(\mathbf{d}), \overline{\sigma}^2(\mathbf{d})] = \left[ \frac{(N-1)\hat{\sigma}^2(\mathbf{d})}{\chi^2{}_{\alpha/2,N-1}},\ \frac{(N-1)\hat{\sigma}^2(\mathbf{d})}{\chi^2{}_{1-\alpha/2,N-1}} \right],$$

$$\underline{F}(x \mid \mathbf{d}) = \min \left\{ \Phi\left( (x - \overline{m}_1(\mathbf{d})) / \overline{\sigma}(\mathbf{d}) \right), \Phi\left( (x - \overline{m}_1(\mathbf{d})) / \underline{\sigma}(\mathbf{d}) \right) \right\},$$
$$\overline{F}(x \mid \mathbf{d}) = \max \left\{ \Phi\left( (x - \underline{m}_1(\mathbf{d})) / \overline{\sigma}(\mathbf{d}) \right), \Phi\left( (x - \underline{m}_1(\mathbf{d})) / \underline{\sigma}(\mathbf{d}) \right) \right\}.$$

$$\hat{\sigma}^2(\mathbf{d}) = \frac{1}{n} \sum_{j=1}^{n} \left( y_j - f(\mathbf{x}_j, \mathbf{d}) \right)^2 - \left( \frac{1}{n} \sum_{j=1}^{n} \left( y_j - f(\mathbf{x}_j, \mathbf{d}) \right) \right)^2 .$$

# Questions

?