86. **Regression with asymmetric uncertainty,** (035018 exam, 21.10.2015) (p.310) Two measurements, $\overline{y}_1$ and $\overline{y}_2$, have been made at successive time steps. It is suspected that $\overline{y}_1$ is an over estimate of the true value, y_1, by as much as s_1 or more. Furthermore, it is suspected that $\overline{y}_2$ deviates from the true value, y_2, either above or below, by s_2 or more. We will represent this uncertainty with this info-gap model:

$$\mathcal{U}(h) = \left\{ y_1, y_2 : \ 0 \leq \frac{\overline{y}_1 - y_1}{s_1} \leq h, \ \left| \frac{y_2 - \overline{y}_2}{s_2} \right| \leq h \right\}, \quad h \geq 0 \tag{395}$$

We wish to choose a linear regression of the form:

$$y_i^{\mathrm{r}} = ci + b, \ i = 1, 2, \ldots \tag{396}$$

The squared error, with respect to measurements $y = (y_1, y_2)$, of a regression with coefficients $q = (c, b)$ is:

$$S(q, y) = \sum_{i=1}^{2} (y_i - y_i^{\mathrm{r}})^2 \tag{397}$$

If we knew that the measurements $\overline{y} = (\overline{y}_1, \overline{y}_2)$ were reliable we might choose the coefficients q to minimize $S(q, \overline{y})$. However, given the uncertainty in the measurements, and especially the asymmetric uncertainty in y_1, we wish to choose the coefficients q so that the regression reflects the measurements, $\overline{y}$, as well as the information about the uncertainty in these measurements. That is, we want the squared error to be small for a wide range of realizations of y as represented by the info-gap model. Consequently our performance requirement is:

$$S(q, y) \leq S_{\mathrm{c}} \tag{398}$$

(a) Let $\widetilde{q} = (\widetilde{c}, \widetilde{b})$ denote the regression coefficients for which the regression **intersects** both measurements, $\overline{y}$. Derive the robustness function for these coefficients.

(b) Let $q = (c, b)$ denote the regression coefficients for a regression that **intersects** measurement $\overline{y}_2$ and that falls **below** the value of $\overline{y}_1$. Derive the inverse of the robustness function for these coefficients.

(c) Let $q = (c, b)$ denote the regression coefficients for a regression that **intersects** measurement $\overline{y}_2$ and that falls **above** the value of $\overline{y}_1$. Derive the inverse of the robustness function for these coefficients.

(d) If y_i^{r} is the correct regression then a prediction, x, that is based on the regression, is normal with mean 0 and standard deviation 1.5. We have observed 6 statistically independent values of x, all evaluated with the same y_i^{r} and for data from the same situation (so if this y_i^{r} is correct for one then it is correct for all). These observed x values fall in 4 different ranges:

$$\begin{array}{lll} n_1 = 1 & \text{value in the range} & [-\infty, \ -2) \\ n_2 = 3 & \text{values in the range} & [-2, \ 0) \\ n_3 = 2 & \text{values in the range} & [0, \ 2) \\ n_4 = 0 & \text{values in the range} & [2, \ \infty) \end{array}$$

Consider the following two hypotheses:

$$H_0 : \quad y_i^{\mathrm{r}} \text{ is the correct regression} \tag{399}$$

$$H_1 : \quad y_i^{\mathrm{r}} \text{ is not the correct regression} \tag{400}$$

Given the observations, do you accept H_0 at 0.05 level of significance?

(e) We now know that y is distributed uniformly in the interval $[x,\ \overline{y}]$. Find the value, denoted $y_\alpha(x)$, such that:

$$\text{Prob}[y \leq y_\alpha(x)] = \alpha \tag{401}$$

(f) Continue part 86e but consider uncertainty in the value of x. We think the correct value is $\widetilde{x}$, but this could err by as much as w or more. Represent this uncertainty with this info-gap model:

$$\mathcal{U}(h) = \left\{ x : \left| \frac{x - \widetilde{x}}{w} \right| \leq h \right\}, \quad h \geq 0 \tag{402}$$

We require that $y_\alpha(\widetilde{x})$ over-estimate the true value, $y_\alpha(x)$, by no more than ε:

$$y_\alpha(\widetilde{x}) - y_\alpha(x) \leq \varepsilon \tag{403}$$

Derive an explicit algebraic expression for the robustness function.

Solution for problem 86: Regression with asymmetric uncertainty (035018 exam, 21.10.2015) (p.100).

(86a) The definition of the robustness function, for any regression coefficients q, is:

$$\widehat{h}(S_c, q) = \max\left\{h: \left(\max_{y\in\mathcal{U}(h)} S(q,y)\right) \le S_c\right\} \tag{1965}$$

Let $m(h)$ denote the inner maximum, which is the inverse of $\widehat{h}(S_c)$. In the present case we are considering coefficients $\widetilde{q} = (\widetilde{c}, \widetilde{b})$, for which $y_i^r = \overline{y}_i$, $i = 1, 2$. Thus $m(h)$ occurs for:

$$y_1 = \overline{y} - s_1 h, \quad y_2 = \overline{y}_2 + s_2 h \tag{1966}$$

Hence:

$$m(h) = [(\overline{y}_1 - s_1 h) - \overline{y}_1]^2 + [(\overline{y}_2 + s_2 h) - \overline{y}_2]^2 = (s_1^2 + s_2^2)h^2 \le S_c \implies \boxed{\widehat{h}(S_c, \widetilde{q}) = \sqrt{\frac{S_c}{s_1^2 + s_2^2}}} \tag{1967}$$

or zero if this is negative. Note that this expression for the robustness does not depend explicitly on $\widetilde{q}$ or on the measurements $\overline{y}_1$ and $\overline{y}_2$.

The robustness curve in eq.(1967) is shown by the solid blue curve in fig. 110 on p. 311.

(86b) In this case, $y_1^r < \overline{y}_1$ and $y_2^r = \overline{y}_2$. Thus $m(h)$ occurs for:

$$y_2 = \overline{y}_2 + s_2 h \tag{1968}$$

and for two different values of y_1 depending on the value of h. When h is small we obtain a maximum of $(y_1 - y_1^r)^2$ for $y_1 = \overline{y}_1$. However, when h is large enough we obtain a maximum of $(y_1 - y_1^r)^2$ for $y_1 = \overline{y}_1 - s_1 h$.

We first develop a concise solution, and then extend that solution explicitly.

Concise solution. Define the two possible realizations of $m(h)$:

$$m_1(h) = (s_2 h)^2 + (\overline{y}_1 - y_1^r)^2 \quad (\text{for } y_1 = \overline{y}) \tag{1969}$$

$$m_2(h) = (s_2 h)^2 + (\overline{y}_1 - s_1 h - y_1^r)^2 \quad (\text{for } y_1 = \overline{y}_1 - s_1 h) \tag{1970}$$

The inverse of the robustness function is the greater of these two:

$$m(h) = \max[m_1(h),\ m_2(h)] \tag{1971}$$

Detailed solution. To find the transition between these two cases we compare their values of $(y_1 - y_1^r)^2$. First note that, from eq.(396) on p.100, $y_1^r = c + b$. So:

$$(\underbrace{\overline{y}_1 - c - b}_{>0})^2 = [\underbrace{c + b - (\overline{y}_1 - s_1 h)}_{>0 \text{ for large } h}]^2 \iff \overline{y}_1 - c - b = c + b - \overline{y}_1 + s_1 h \iff h = \frac{2\overline{y}_1 - 2(c+b)}{s_1} \tag{1972}$$

Hence we conclude that $m(h)$ is obtained with the following choice of y_1:

$$y_1 = \begin{cases} \overline{y}_1 & \text{if } h \le \dfrac{2(\overline{y}_1 - y_1^r)}{s_1} \\ \overline{y}_1 - s_1 h & \text{else} \end{cases} \tag{1973}$$

Using eqs.(1968) and (1973) we find:

$$\boxed{m(h) = \begin{cases} (\overline{y}_1 - y_1^r)^2 + (s_2 h)^2 & \text{if } h \le \dfrac{2(\overline{y}_1 - y_1^r)}{s_1} \\ (\overline{y}_1 - s_1 h - y_1^r)^2 + (s_2 h)^2 & \text{else} \end{cases}} \tag{1974}$$

The robustness curve whose inverse appears in eq.(1974) is shown by the dashed red curve in fig. 110 on p. 311. Note the kink corresponding to the transition from the upper to the lower line in eq.(1974). This robustness curve (dashed red) crosses the robustness curve from eq.(1967) (solid blue) as we now explain. We first note that this makes sense because the sub-optimal regression (falling below $\overline{y}_1$) exploits the asymmetric uncertainty. Now for an analytical demonstration of crossing robustness curves.

Let $m_a(h)$ denote the inverse robustness function in eq.(1967) from part 86a, and let $m_b(h)$ denote the inverse robustness function in eq.(1974) from part 86b. We note that:

$$m_a(0) = 0 < (\overline{y}_1 - y_1^{\mathrm{r}})^2 = m_b(0) \tag{1975}$$

However, now we compare $m_a(h)$ with $m_b(h)$ for very large h:

$$m_a(h) = (s_1 h)^2 + (s_2 h)^2 \quad ? \quad [s_1 h - (\overline{y}_1 - y_1^{\mathrm{r}})]^2 + (s_2 h)^2 = m_b(h) \tag{1976}$$

$$\Longrightarrow \quad (s_1 h)^2 \quad ? \quad [s_1 h - (\overline{y}_1 - y_1^{\mathrm{r}})]^2 \tag{1977}$$

$$\Longrightarrow \quad \text{`?'} \quad \equiv \quad \text{`}>\text{'} \tag{1978}$$

$$\Longrightarrow \quad m_a(h) \quad > \quad m_b(h) \quad \text{for large } h \tag{1979}$$

Conbining eqs.(1975) and (1979) we see that $m_a(h)$ and $m_b(h)$ cross. Thus the corresponding robustness curves cross.

(86c) In this case, $y_1^{\mathrm{r}} > \overline{y}_1$ and $y_2^{\mathrm{r}} = \overline{y}_2$. Thus $m(h)$ occurs for:

$$y_1 = \overline{y}_1 - s_1 h \quad \text{and} \quad y_2 = \overline{y}_2 + s_2 h \tag{1980}$$

Thus:

$$m(h) = [y_1^{\mathrm{r}} - (\overline{y}_1 - s_1 h)]^2 + [y_2^{\mathrm{r}} - (\overline{y}_2 + s_1 h)]^2 = (y_1^{\mathrm{r}} - \overline{y}_1 + s_1 h)^2 + (s_2 h)^2 \tag{1981}$$

It seems that this inverse robustness function does not cross the inverse robustness function in eq.(1967). This makes sense because the sub-optimal regression (falling above $\overline{y}_1$) contradicts the asymmetric uncertainty.

The robustness curve in eq.(1981) is shown by the dot-dashed yellow curve in fig. 110 on p. 311.

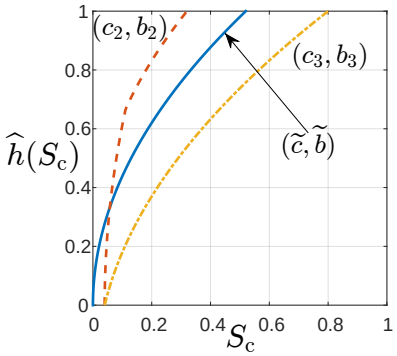

Figure 110: Robustness curves for problems 86a–86c. $\overline{y}_1 = 2$, $\overline{y}_2 = 1.5$, $s_1 = 0.6$, $s_2 = 0.4$, $(c_2, b_2) = (-0.3, 2.1)$, $(c_3, b_3) = (-0.7, 2.9)$. Calculated with prob86_001.m.

(86d) We will use a χ^2 test. Let $F(x)$ denote the cumulative probability distribution for $\mathcal{N}(0, 1.5^2)$. Table 5 contains values of $F(x)$.

x	$F(x)$
5.0000	0.9996
4.0000	0.9962
3.0000	0.9772
2.0000	0.9088
1.0000	0.7475
0	0.5000
−1.0000	0.2525
−2.0000	0.0912
−3.0000	0.0228
−4.0000	0.0038
−5.0000	0.0004

Table 5: Cumulative probability distribution for problem 86d.

Let P_i denote the probability of x falling in the ith bin:

$$P_1 = F(-2) - F(-4) = 0.0912 - 0 = \mathbf{0.0912} \quad (1982)$$

$$P_2 = F(0) - F(-2) = 0.5 - 0.0912 = \mathbf{0.4088} \quad (1983)$$

$$P_3 = F(2) - F(0) = 0.9088 - 0.5 = \mathbf{0.4088} \quad (1984)$$

$$P_4 = F(4) - F(2) = 1 - 0.9088 = \mathbf{0.0912} \quad (1985)$$

The χ^2 statistic is:

$$\chi^2 = \sum_{i=1}^{4} \frac{(n_i - NP_i)^2}{NP_i} = 1.1276 \quad (1986)$$

There are $K = 4 - 1 = 3$ degrees of freedom. We reject H_0 at 0.05 level of significance if:

$$\text{Prob}(\chi^2_{(3)} \geq 1.1276) \leq 0.05 \quad (1987)$$

From a χ^2 table we find:

$$\text{Prob}(\chi^2_{(3)} \geq 1.1276) > 0.1 \quad (1988)$$

Thus we **do not reject** H_0 at 0.05 level of significance.

(86e) For the uniform distribution of y on the interval $[x,\ \overline{y}]$ we find:

$$\text{Prob}[y \leq y_\alpha(x)] = \alpha \iff \frac{y_\alpha(x) - x}{\overline{y} - x} = \alpha \iff \boxed{y_\alpha(x) = (\overline{y} - x)\alpha + x = \alpha\overline{y} + (1-\alpha)x} \quad (1989)$$

(86f) The definition of the robustness function is:

$$\widehat{h}(\varepsilon) = \max\left\{h:\ \left(\max_{x \in \mathcal{U}(h)} [y_\alpha(\widetilde{x}) - y_\alpha(x)]\right) \leq \varepsilon\right\} \quad (1990)$$

Let $m(h)$ denote the inner maximum, which is the inverse of the robustness function, $\widehat{h}(\varepsilon)$. This maximum occurs for $x = \widetilde{x} - wh$, so:

$$m(h) = \alpha\overline{y} + (1-\alpha)\widetilde{x} - [\alpha\overline{y} + (1-\alpha)(\widetilde{x} - wh)] = (1-\alpha)wh \leq \varepsilon \longrightarrow \boxed{\widehat{h}(\varepsilon) = \frac{\varepsilon}{(1-\alpha)w}} \quad (1991)$$