Part III Astrostatistics, Example Sheet 2 Solutions
Example Class: Friday, 02 Mar 2017, 2:00pm, MR5

1 Linear Regression with $(x, y)-$measurement error and intrinsic dispersion: Quasar X-ray Spectral Slopes vs. Eddington Ratios

In class we examined the problem of linear regression of the quasar X-ray spectral index vs. bolometric luminosity in the presence of measurement error in both quantities and intrinsic dispersion. (Regression is also described in Feigelson & Babu, Chapter 7, Ivezic et al., Chapter 8, and Kelly et al. 2007, The Astrophysical Journal, 665, 1506). Consider the probabilistic generative model described in class:

$$\xi_i \sim N(\mu, \tau^2) \tag{1}$$

$$\eta_i|\xi_i \sim N(\alpha + \beta\xi_i, \sigma^2) \tag{2}$$

$$x_i|\xi_i \sim N(\xi_i, \sigma^2_{x,i}) \tag{3}$$

$$y_i|\eta_i \sim N(\eta_i, \sigma^2_{y,i}) \tag{4}$$

The astronomer measures values $\mathcal{D} = \{x_i, y_i\}$ with known measurement error variances $\{\sigma^2_{x,i}, \sigma^2_{y,i}\}$, for $i = 1, \dots, N$ quasars.

1. Write down the joint distribution $P(x_i, y_i, \xi_i, \eta_i|\alpha, \beta, \sigma^2, \mu, \tau^2)$ for a single quasar.

   Solution:

$$\begin{aligned}
P(x_i, y_i, \xi_i, \eta_i|\, \alpha, \beta, \sigma^2, \mu, \tau^2) &= P(x_i, y_i|\xi_i, \eta_i)P(\xi_i, \eta_i|\alpha, \beta, \sigma^2, \mu, \tau) \\
&= P(y_i|\eta_i)P(x_i|\xi_i)P(\eta_i|\xi_i; \alpha, \beta, \sigma^2)P(\xi_i|\mu, \tau^2) \\
&= N(y_i|\, \eta_i, \sigma^2_{y,i})N(x_i|\xi_i, \sigma^2_{x,i})N(\eta_i|\alpha + \beta\xi_i, \sigma^2)N(\xi_i|\mu, \tau^2)
\end{aligned}$$

   which comes from expanding the joint into conditionals and marginals and using the modeling assumptions, Eq. 1-4.

2. Derive the observed data likelihood function for all the quasars:

$$L(\alpha, \beta, \sigma^2, \mu, \tau^2) = \prod_{s=1}^{N} P(x_i, y_i|\alpha, \beta, \sigma^2, \mu, \tau^2). \tag{5}$$

   Show all steps and maximally simplify.

   Solution: We could marginalise out the latent coordinates (ξ_i, η_i) analytically:

$$P(x_i, y_i|\alpha, \beta, \sigma^2, \mu, \sigma^2, \tau^2) = \int\int d\eta_i\, d\xi_i P(x_i, y_i, \xi_i, \eta_i|\, \alpha, \beta, \sigma^2, \mu, \tau^2),$$

but that would be tedious and not very insightful. Instead, we can utilise the properties of multivariate Gaussian random vectors. We use the fact that a marginal distribution of multivariate Gaussian vector $\boldsymbol{V}$,

$$\boldsymbol{V} \sim N(\boldsymbol{V}_0, \boldsymbol{\Sigma}_V)$$

and a conditional distribution of $\boldsymbol{U}|\boldsymbol{V}$:

$$\boldsymbol{U}|\boldsymbol{V} \sim N(\boldsymbol{U}_0 + \boldsymbol{X}\boldsymbol{V}, \boldsymbol{\Sigma}_{U|V})$$

for some matrix $\boldsymbol{X}$ of the appropriate dimensionality, are equivalent to a joint distribution that is also multivariate Gaussian

$$\begin{pmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{U}_0 + \boldsymbol{X}\boldsymbol{V}_0 \\ \boldsymbol{V}_0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{X}\boldsymbol{\Sigma}_V\boldsymbol{X}^T + \boldsymbol{\Sigma}_{U|V} & \boldsymbol{X}\boldsymbol{\Sigma}_V \\ \boldsymbol{\Sigma}_V\boldsymbol{X}^T & \boldsymbol{\Sigma}_V \end{pmatrix} \right).$$

Then marginally, $\boldsymbol{U} \sim N(\boldsymbol{U}_0 + \boldsymbol{X}\boldsymbol{V}_0, \boldsymbol{X}\boldsymbol{\Sigma}_V\boldsymbol{X}^T + \boldsymbol{\Sigma}_{U|V})$ is also multivariate Gaussian.

We can apply this to Eqs. 1 & 2 to get the joint distribution of (η_i, ξ_i),

$$\begin{pmatrix} \eta_i \\ \xi_i \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha + \beta\mu \\ \mu \end{pmatrix}, \begin{pmatrix} \beta^2\tau^2 + \sigma^2 & \beta\tau^2 \\ \beta\tau^2 & \tau^2 \end{pmatrix} \right).$$

We also note that Eqs 3 & 4 can be described jointly as a multivariate Gaussian vector:

$$\begin{pmatrix} y_i \\ x_i \end{pmatrix} \Big| \begin{pmatrix} \eta_i \\ \xi_i \end{pmatrix} \sim N\left(\begin{pmatrix} \eta_i \\ \xi_i \end{pmatrix}, \begin{pmatrix} \sigma_{y,i}^2 & 0 \\ 0 & \sigma_{x,i}^2 \end{pmatrix} \right)$$

We can apply the same lemma again (with $\boldsymbol{X} = I$) to obtain the marginal density of $\boldsymbol{U} = (y_i, x_i)^T$, integrating out the latent variables $\boldsymbol{V} = (\eta_i, \xi)^T$:

$$\begin{pmatrix} y_i \\ x_i \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha + \beta\mu \\ \mu \end{pmatrix}, \begin{pmatrix} \beta^2\tau^2 + \sigma^2 + \sigma_{y,i}^2 & \beta\tau^2 \\ \beta\tau^2 & \tau^2 + \sigma_{x,i}^2 \end{pmatrix} \right) \equiv N(\boldsymbol{\zeta}, \boldsymbol{V}_i).$$

Hence the joint sampling distribution for $\boldsymbol{z}_i \equiv (y_i, x_i)^T$ is $P(y_i, x_i|\alpha, \beta, \sigma^2, \mu, \tau^2) = N(\boldsymbol{z}_i|\,\boldsymbol{\zeta}, \boldsymbol{V}_i)$, so $\boldsymbol{z}_i$ is also a multivariate Gaussian vector. The likelihood function of the parameters is then

$$L(\alpha, \beta, \sigma^2, \mu, \tau^2) = \prod_{i=1}^{N} N(\boldsymbol{z}_i|\,\boldsymbol{\zeta}, \boldsymbol{V}_i),$$

assuming the data from each quasar are independently sampled.

3. Write a code to find the maximum likelihood estimate, if given $\{x_i, y_i\}$ and their known measurement variances for $i = 1 \dots N$ quasars. (If you were unable to complete steps 1 & 2, use Eqs 19 - 23 of Kelly et al. 2007 with K= 1, $\gamma_1 = \pi_1 = 1$). Test your code on simulated data you generate from the model with known true parameter values. Find an approximate 95% confidence interval for each parameter using the observed Fisher information. (Use a generic optimisation library or toolbox to numerically minimise a given function, e.g. scipy.optimize in Python, fmincon in Matlab, or optim in R, or equivalent).

   Solution: see code.

4. Using the dataset provided online ("example_sheet2_prb1_data.txt"), find the maximum likelihood estimates (MLEs) of the parameters $\alpha, \beta, \sigma^2, \mu, \tau^2$, and their uncertainties.

   Solution: see code.

5. Suppose the distribution of the latent (true) independent variables $\{\xi_i\}$ is assumed to be "non-informative" or flat. Take the limit $\tau \to \infty$ of Eq. 5 to derive $L_{\tau\to\infty}(\alpha, \beta, \sigma^2)$.

   Solution: Let $\boldsymbol{\theta} = (\alpha, \beta, \sigma^2)$ be the parameters of the regression model, and $\boldsymbol{\phi} = (\mu, \tau^2)$ be the parameters of the population of the latent independent variable. Noting that the joint distribution of $P(y_i, x_i|\boldsymbol{\theta}, \boldsymbol{\phi})$ is Gaussian, we can decompose it into the product of the conditional $P(y_i|x_i, \boldsymbol{\theta}, \boldsymbol{\phi})$ and the marginal $P(x_i|\boldsymbol{\theta}, \boldsymbol{\phi})$. We can read from the solution to part 2, the marginal

$$P(x_i|\boldsymbol{\theta}, \boldsymbol{\phi}) = N(x_i|\, \mu, \tau^2 + \sigma_{x,i}^2).$$

   From the properties of the multivariate Gaussian, the conditional density derived from the solution to part 2 is:

$$P(y_i|\, x_i; \boldsymbol{\theta}, \boldsymbol{\phi}) = N(y_i|\, \mathbb{E}[y_i|\, x_i; \boldsymbol{\theta}, \boldsymbol{\phi}], \mathbf{Var}[y_i|\, x_i; \boldsymbol{\theta}, \boldsymbol{\phi}])$$

   in which the conditional expectation is:

$$\begin{aligned}\mathbb{E}[y_i|\, x_i; \boldsymbol{\theta}, \boldsymbol{\phi}] &= \alpha + \beta\mu + \frac{\beta\tau^2}{\tau^2 + \sigma_{x,i}^2}(x_i - \mu) \\ &= \alpha + \frac{\beta\tau^2}{\tau^2 + \sigma_{x,i}^2} x_i + \frac{\beta\sigma_{x,i}^2}{\tau^2 + \sigma_{x,i}^2}\mu\end{aligned}$$

$$\mathbf{Var}[y_i|\, x_i; \boldsymbol{\theta}, \boldsymbol{\phi}] = \beta^2\tau^2 + \sigma^2 + \sigma_{y,i}^2 - \frac{(\beta\tau^2)^2}{\tau^2 + \sigma_{x,i}^2}$$

   In the limit $\tau \to \infty$, $\mathbb{E}[y_i|\, x_i; \boldsymbol{\theta}, \boldsymbol{\phi}] \to \alpha + \beta x_i$, $\mathbf{Var}[y_i|\, x_i; \boldsymbol{\theta}, \boldsymbol{\phi}] \to \sigma^2 + \sigma_{y,i}^2 + \beta^2\sigma_{x,i}^2$. So,

$$L(\alpha, \beta, \sigma^2, \mu, \tau^2) \to \prod_{i=1}^{N} N(y_i|\, \alpha + \beta x_i, \sigma^2 + \sigma_{y,i}^2 + \beta^2\sigma_{x,i}^2) \times \prod_{i=1}^{N} N(x_i|\, \mu, \tau^2)$$

   Note that only the first product depends on the regression parameters α, β, σ^2, so we can take

$$L_{\tau\to\infty}(\alpha, \beta, \sigma^2) \propto \prod_{i=1}^{N} N(y_i|\, \alpha + \beta x_i, \sigma^2 + \sigma_{y,i}^2 + \beta^2\sigma_{x,i}^2)$$

   while the MLE for μ is independently $\hat{\mu} = \bar{x}$ from the second product.

6. Compare your MLE for β using Eq 5 against what you get using ordinary least squares (OLS), minimum χ^2, FITEXY modified χ^2 methods, and the MLE with $L_{\tau\to\infty}(\alpha, \beta, \sigma^2)$.

   (a) Ordinary Least Squares minimises the residual sum of squares (RSS) with respect to the parameters:

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \alpha - \beta\, x_i)^2\,. \tag{6}$$

Solution: This can be solved by direct numerical minimisation of the above. However, the solution can be written in closed form:

$$\textbf{RSS} = (\boldsymbol{y} - \boldsymbol{X}b)^T(\boldsymbol{y} - \boldsymbol{X}b)$$

where $\boldsymbol{y}$ has elements y_i, $\boldsymbol{X}$ is a matrix with the ith row being $X_i = (1, x_i)$ and parameter vector $b = (\alpha, \beta)^T$. This ordinary least squares problem has a linear algebra solution:

$$\hat{b} = \arg\min_b \textbf{RSS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

(b) Minimum χ^2 or Weighted Least Squares minimises the following with respect to the parameters:

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - \alpha - \beta\, x_i)^2}{\sigma_{y,i}^2}. \tag{7}$$

What minimum value do you find for the reduced $\chi^2_\nu = \chi^2/(N-2)$?
Solution: This can be solved by direct numerical minimisation of the above. However, the solution can be written in closed form:

$$\chi^2 = (\boldsymbol{y} - \boldsymbol{X}b)^T\boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{X}b)$$

where $\boldsymbol{y}$ has elements y_i, $\boldsymbol{X}$ is a matrix with the ith row being $X_i = (1, x_i)$ and parameter vector $b = (\alpha, \beta)^T$. The weight matrix is diagonal with elements $(W^{-1})_{ii} = \sigma_{y,i}^{-2}$. This weighted least squares problem has a linear algebra solution:

$$\hat{b} = \arg\min_b \chi^2 = (\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{Y}$$

(c) The FITEXY methods (Press et al. *Numerical Recipes in C*) minimise an “effective” χ^2 statistic that takes in account x-measurement errors

$$\chi^2_{EXY} = \sum_{i=1}^{N} \frac{(y_i - \alpha - \beta\, x_i)^2}{\sigma_{y,i}^2 + \beta^2\sigma_{x,i}^2}. \tag{8}$$

What minimum value do you find for the reduced $\chi^2_{EXY,\nu} = \chi^2_{EXY}/(N-2)$?
Solution: see code

(d) The maximum likelihood solution assuming a non-informative distribution on the $\{\xi_i\}$ is obtained by minimising:

$$-\log L_{\tau\to\infty}(\alpha, \beta, \sigma^2) = \lim_{\tau\to\infty} -\log L(\alpha, \beta, \sigma^2, \mu, \tau^2). \tag{9}$$

Solution: see code

7. State and employ appropriate non-informative priors on the parameters $\alpha, \beta, \sigma^2, \mu, \tau^2$ defined in Eqs. 1 - 5. Construct and implement a MCMC algorithm to sample from the posterior probability density:

$$P(\alpha, \beta, \sigma^2, \mu, \tau^2 |\, \mathcal{D}) \propto L(\alpha, \beta, \sigma^2, \mu, \tau^2) \times P(\alpha, \beta, \sigma^2, \mu, \tau^2) \tag{10}$$

Run 4 independent chains to diagnose convergence using the Gelman-Rubin ratio. Remove “burn-in” and use the combined chains to compute the marginal distributions of the parameters, and compare against the point estimates you previously obtained with the other methods. **Solution: see code**

2 Importance Sampling for Bayesian Estimates of the Milky Way Mass using Angular Momentum Measurements

Look up the paper Patel et al. 2017, "Orbits of massive satellite galaxies II. Bayesian estimates of the Milky Way and Andromeda masses using high-precision astrometry and cosmological simulations." *Monthly Notices of the Royal Astronomical Society*, 468, 3428. Use the measurements in Table 1 and the online data from the Illustris simulation to estimate the Milky Way mass using angular momentum j and the rotational velocity $v_{\max}$ of the Large Magellanic Cloud (LMC). In this context, the Milky Way is the central (host) galaxy of the system, and the LMC is a "satellite" galaxy.

1. Let $\boldsymbol{x} = (v_{\max}, j)$ be the latent parameters, and let $\boldsymbol{d} = (v_{\max}^{\text{obs}}, j^{\text{obs}})$ be their measured values, with uncertainties shown in Table 1. Write down the likelihood function $P(\boldsymbol{d}|\,\boldsymbol{x})$, assuming Gaussian measurement errors.

   Solution:
$$P(\boldsymbol{d}|\boldsymbol{x}) = N(v_{\mathbf{max}}^{\mathbf{obs}}|\, v_{\mathbf{max}}, \sigma_v^2) \times N(j^{\mathbf{obs}}|\, j, \sigma_j^2)$$

   where σ_v and σ_k are the standard deviations of measurement errors associated with $v_{\mathbf{max}}^{\mathbf{obs}}$ and $j^{\mathbf{obs}}$, listed in Table 1.

2. The Illustris simulation implicitly encodes a joint distribution between these latent dynamical parameters of satellites and the $\log_{10}$ masses of central (or host) galaxies, $P(\boldsymbol{x}, \log_{10} M)$. Assuming this exists, write down an expression for the normalised posterior probability density of the Milky Way $\log_{10}$ mass.

   Solution: The joint posterior density of the latent parameters $\boldsymbol{x}$ and $m \equiv \log_{10} M$ is given by Bayes' Theorem:
$$P(\boldsymbol{x}, m|\boldsymbol{d}) = \frac{P(\boldsymbol{d}|\boldsymbol{x}, m)P(\boldsymbol{x}, m)}{\int P(\boldsymbol{d}|\boldsymbol{x}, m)P(\boldsymbol{x}, m)\, d\boldsymbol{x}\, dm}$$

   but since $\boldsymbol{d}$ is conditionally independent of m, given $\boldsymbol{x}$, $P(\boldsymbol{d}|\boldsymbol{x}, m) = P(\boldsymbol{d}|\boldsymbol{x})$. Then marginalising over the latent variables $\boldsymbol{x}$, we have
$$P(m|\boldsymbol{d}) = \frac{\int P(\boldsymbol{d}|\,\boldsymbol{x})P(\boldsymbol{x}, m)\, d\boldsymbol{x}}{\int P(\boldsymbol{d}|\,\boldsymbol{x})P(\boldsymbol{x}, m)\, d\boldsymbol{x}\, dm}$$

3. Write down an expression for the posterior mean estimate of the $\log_{10}$ MW mass in terms of integrals involving the likelihood and prior.

   Solution:
$$\mathbb{E}[m|\boldsymbol{d}] = \int m\, P(m|\,\boldsymbol{d})\, dm = \frac{\int m\, P(\boldsymbol{d}|\,\boldsymbol{x})P(\boldsymbol{x}, m)\, d\boldsymbol{x}\, dm}{\int P(\boldsymbol{d}|\,\boldsymbol{x})P(\boldsymbol{x}, m)\, d\boldsymbol{x}\, dm}$$

4. Using an arbitrary importance sampling distribution $Q(\boldsymbol{x}, \log_{10} M)$ from which we can easily draw samples, rewrite this expression in terms of expectations with respect to Q. **Solution: We can multiply the top and bottom integrands by $1 = Q(\boldsymbol{x}, m)/Q(\boldsymbol{x}, m)$.**
$$\mathbb{E}[m|\boldsymbol{d}] = \frac{\int m\, P(\boldsymbol{d}|\,\boldsymbol{x})\, \frac{P(\boldsymbol{x},m)}{Q(\boldsymbol{x},m)}\, Q(\boldsymbol{x}, m)\, d\boldsymbol{x}\, dm}{\int P(\boldsymbol{d}|\,\boldsymbol{x})\, \frac{P(\boldsymbol{x},m)}{Q(\boldsymbol{x},m)}\, Q(\boldsymbol{x}, m)\, d\boldsymbol{x}\, dm}$$

5. Rewrite this expression now assuming now that the importance sampling distribution is the same as the prior $Q(\boldsymbol{x}, \log_{10} M) = P(\boldsymbol{x}, \log_{10} M)$. Approximate this expression with weighted sums over the prior samples, suitable for the Monte Carlo method, and derive the importance weights.

   Solution: In general, if we have K samples from the importance sampling function, $\boldsymbol{\theta}_i \sim Q(\boldsymbol{\theta})$, then we can approximate expectations with respect to $P(\boldsymbol{\theta})$ using the Monte Carlo sum:

$$\mathbb{E}[f(\boldsymbol{\theta}] = \int f(\boldsymbol{\theta})\, P(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \int f(\boldsymbol{\theta}) \frac{P(\boldsymbol{\theta})}{Q(\boldsymbol{\theta})} Q(\boldsymbol{\theta})\, d\boldsymbol{\theta} \approx \frac{1}{K} \sum_{i=1}^{K} f(\boldsymbol{\theta}_i) w_i$$

   in which the importance weights are $w_i = P(\boldsymbol{\theta}_i)/Q(\boldsymbol{\theta}_i)$ evaluated over the samples $\boldsymbol{\theta}_i$. Applying this we have:

$$\mathbb{E}[m|\boldsymbol{d}] \approx \frac{\frac{1}{K} \sum_{i=1}^{K} m_i\, P(\boldsymbol{d}|\, \boldsymbol{x}_i) \frac{P(\boldsymbol{x}_i, m)}{Q(\boldsymbol{x}_i, m_i)}}{\frac{1}{K} \sum_{i=1}^{K} P(\boldsymbol{d}|\, \boldsymbol{x}_i) \frac{P(\boldsymbol{x}_i, m)}{Q(\boldsymbol{x}_i, m_i)}}$$

   However, since we are using the Illustris simulation both as the prior and importance sampling distribution $P(\boldsymbol{x}, m) = Q(\boldsymbol{x}, m)$, and this simplifies to:

$$\mathbb{E}[m|\boldsymbol{d}] \approx \frac{\sum_{i=1}^{K} m_i\, P(\boldsymbol{d}|\, \boldsymbol{x}_i)}{\sum_{i=1}^{K} P(\boldsymbol{d}|\, \boldsymbol{x}_i)} = \sum_{i=1}^{K} m_i\, w_i$$

   where the importance weights are

$$w_i = \frac{P(\boldsymbol{d}|\, \boldsymbol{x}_i)}{\sum_{i=1}^{K} P(\boldsymbol{d}|\, \boldsymbol{x}_i)}$$

   which are already normalised ($\sum_{i=1}^{K} w_i = 1$).

6. Use the Illustris host-satellite data in the online file "Patel17b_Illustris_Data_KM.txt" as samples from the prior. Use the columns labelled "MVIR", "SATVMAX" and "SATJMAG". Compute the importance weights, and estimate the posterior mean and standard deviation of the $\log_{10}$ MW mass, given the LMC data $\boldsymbol{d}$. Also compute an effective sample size using Eq. B2 in the paper, and compare against the number of samples from the prior.
   Solution: see code

7. Using the bandwidth Eq. B1, create a weighted KDE representation of the posterior distribution $P(\log_{10} M|\boldsymbol{d})$. Plot it over a KDE representation of the prior $P(\log_{10} M)$.
   Solution: see code