Statistics for Astronomers: Lecture 22, 2019.05.21

Prof. Sundar Srinivasan

IRyA/UNAM

Recall: Linear regression

"Linear" parametric regression: linear in parameters, not necessary in the regressor.
Regression function: $Y(x) = \mathbb{E}[Y|X = x]$
Regression model: $Y = f(X) + \epsilon$; $\mathbb{E}[\epsilon] = 0$. (Even if X isn't random, Y is, because of ϵ.)
ϵ can be a combination of measurement error and intrinsic variation. Typically one is much smaller and can be ignored compared to the other.

General procedure:

1. Dependent variable decision
2. Model choice(s)
3. Method of parameter estimation/choice of goodness-of-fit ("objective function")
   Method of moments, least-squares, MLE, Bayesian inference.
   – estimate best values for parameters and their variances; confidence intervals.
4. Occam's Razor model validation/selection.
5. Prediction/classification can then be performed – for a given set of x values, compute y along with variances and thus confidence intervals for these y.

If the uncertainties ϵ_i associated with the dependent variable values y_i are drawn from a distribution with the same variance, the uncertainties are said to be homoskedastic. If they are drawn from distributions with differing variances, they are heteroskedastic.

Recall: Ordinary least-squares (OLS) estimation

The objective function for homoskedastic uncertainties is the residual sum of squares:
$RSS = \sum_{i=1}^{N} \left(y_i - y_{\mathrm{mod},i}\right)^2$. Since $y_i - y_{\mathrm{mod},i} \sim \mathcal{N}(0, \sigma^2)$, $RSS \sim \chi^2(N-2)$.

Minimising RSS w.r.t. the slope and intercept, we get $\hat{m} = \frac{S_{xy}}{S_{xx}}$ and $\hat{b} = \bar{y} - \hat{m}\bar{x}$.

Variances:
$Var(\hat{m}) = \frac{\sigma^2}{S_{xx}}$, $Var(\hat{b}) = \sigma^2 \left(\frac{1}{N} + \frac{\bar{x}^2}{S_{xx}}\right)$

Recall: OLS in matrix notation

In general, the regression relation becomes $\mathbf{y} = \mathbf{A}\mathbf{x}$.

Linear case: $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix}_{N\times 1}$ $\mathbf{A} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_N \end{bmatrix}_{N\times 2}$ $\mathbf{x} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}_{2\times 1}$ (θ_1 = intercept, θ_2 = slope)

The general covariance matrix $\mathbf{\Sigma}$ for heteroskedasstic uncertainties is such that $\Sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$, where ρ_{ij} is the correlation coefficient between σ_i and σ_j.

For uncorrelated uncertainties, $\mathbf{\Sigma}$ is diagonal: $\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \cdots & & & \\ 0 & 0 & \cdots & \sigma_N^2 \end{bmatrix}_{N\times N}$

$RSS \propto (\mathbf{y} - \mathbf{A}\mathbf{x})^T \mathbf{\Sigma}^{-1} (\mathbf{y} - \mathbf{A}\mathbf{x})$, where $\mathbf{\Sigma}^{-1} = \frac{1}{\sigma^2}\mathbb{I}$ (homoskedastic uncorrelated uncertainties).

If RSS is minimized w.r.t. $\mathbf{x}$, we get the matrix product version of the results obtained in the previous slide: $\hat{\mathbf{x}} = (\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{y} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}$.

$(\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A})^{-1}$ is the covariance matrix for the parameters.

Incorporating detection limits

see Appendix in Sawicki 2012 PASP 124, 1208

A faint source may be undetected during an observation. The detection limit is the upper limit to the true flux of a faint source.

Assuming that the background noise is normally distributed, the detection limit usually specified in terms of the background noise. *E.g.*, "the 3-σ upper limit is 0.5 μJy" means that the probability that the true flux of the faint source is below 0.5 μJy is about 98.75%.

Suppose our model predicts a flux of $F_{\rm mod}$ for this source. The probability that the source flux is drawn from this model is

$$P \propto \int_{-\infty}^{F_{\rm lim}} dF \ \exp\left[-\frac{1}{2}\left(\frac{F-F_{\rm mod}}{\sigma^2}\right)^2\right] = \sqrt{\frac{\pi}{2}}\sigma\left[1+{\rm erf}\left(\frac{F_{\rm lim}-F_{\rm mod}}{\sigma}\right)\right].$$

For multi-band photometry combining detections and non-detections,

$$\chi^2 = \sum_i \left(\frac{F_i - F_{{\rm mod},i}}{\sigma_i}\right)^2 + \sum_j \ln\left\{\sqrt{\frac{\pi}{2}}\sigma_j\left[1+{\rm erf}\left(\frac{F_{{\rm lim},j}-F_{{\rm mod},j}}{\sigma_j}\right)\right]\right\}.$$

χ^2 minimisation can now be performed numerically to obtain the best-fit model parameters.

Same technique can be applied to bright sources – the integration limits then go from $F_{\rm sat}$ to ∞.

Robust regression (Outlier rejection)

"Robust statistics provide strategies to reduce the influence of outliers when scientific knowledge of the identity of the discordant data points is not available." – Feigelsen & Babu.

In such a case, manual removal of outliers is neither objective nor reproducible.
Many robust regression techniques exist (see Feigelsen & Babu for a summary).

We will discuss one example among these, Bayesian Outlier Rejection (see Hogg et al. (2010) and Section 8.9 in the AstroML book). The method is similar to assuming a Gaussian mixture model for the data.

Assumptions:

– The uncertainties associated with "true data" are distributed according to $\mathcal{N}(0,\sigma^2)$.

– Outliers are generated from a Gaussian distribution with mean Y_b with variance V_b (substantially larger than σ).

– The probability that a given data point is an outlier in the resulting Gaussian mixture is P_b. Or, equivalently, we can flag each point according to whether or not we think it is an outlier. Each point then has an associated flag variable q_i ($q_i = 0$ if the point is "bad", 1 if "good").

We want to fit a straight line to the "good" points. In addition to m and b, we now have $N+3$ extra parameters. The q_i are nuisance parameters which we can marginalise over. BUT for a given point j we could also marginalise over all other parameters except q_j to see if it was flagged as a true data point or an outlier! This is the strength of the Bayesian method.

Bayesian outlier rejection: likelihood

With 'fg' and 'bg' referring to the true data ("foreground") and outliers ("background"),

$$\mathscr{L} = \prod_{i=1}^{N} p_{fg}(\mathrm{data}|m,b)^{q_i} \cdot p_{bg}(\mathrm{data}|Y_b, V_b)^{1-q_i}$$

$$= \prod_{i=1}^{N} \left\{ \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\left(\frac{y_i - mx_i - b}{\sigma_i} \right)^2 \right] \right\}^{q_i} \left\{ \frac{1}{\sqrt{2\pi(V_b + \sigma_i^2)}} \exp\left[-\frac{1}{2}\frac{(y_i - Y_b)^2}{V_b + \sigma_i^2} \right] \right\}^{1-q_i}$$

In terms of P_b, instead, we could also write

$$\mathscr{L} = \prod_{i=1}^{N} (1 - P_b) \cdot p_{fg}(\mathrm{data}|m,b) + P_b \cdot p_{bg}(\mathrm{data}|Y_b, V_b)$$

$$= \prod_{i=1}^{N} (1-P_b) \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\left(\frac{y_i - mx_i - b}{\sigma_i} \right)^2 \right] + P_b \frac{1}{\sqrt{2\pi(V_b + \sigma_i^2)}} \exp\left[-\frac{1}{2}\frac{(y_i - Y_b)^2}{V_b + \sigma_i^2} \right]$$

Bayesian outlier rejection: priors

For a certain combination $\{q_i\}$, the distribution is binomial: $p(\{q_i\}|P_b) = \prod_{i=1}^{N} (1 - P_b)^{q_i} P_b^{1-q_i}$.

For P_b, Y_b, ("location" parameters) and V_b ("scale" parameter), we can use prior information or choose uninformative priors.

Bayesian outlier rejection: marginalisation

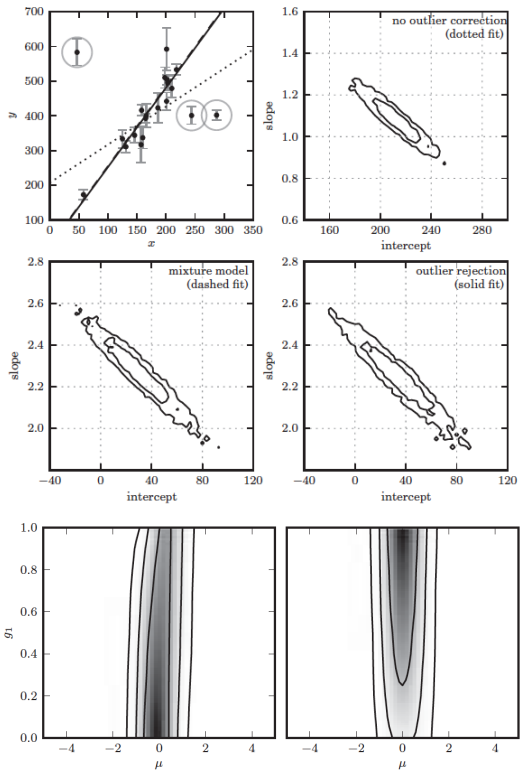
The posterior is $\propto$ likelihood $\times$ the priors.

We can marginalise this posterior over the nuisance parameters q_i to obtain the joint distribution of m and b.

Since the q_i are discrete (they take values 0 or 1), marginalising over them means summing over these possible values instead of integrations.

Once this is done, we also marginalise over P_b, V_b, and Y_b.

This is a multidimensional problem, perfect for MCMC. The implementation is part of the AstroML book (Section 8.9). Part of the homework this week.



source: AstroML book Sections 5.6.7 and 8.9

Parameter uncertainties

For the OLS setup, the parameter uncertainties were in the matrix

$$(\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A})^{-1} = \begin{bmatrix} \sigma_b^2 & \sigma_b\sigma_m \\ \sigma_m\sigma_b & \sigma_m^2 \end{bmatrix}.$$

For more complicated situations (which is most of the time):

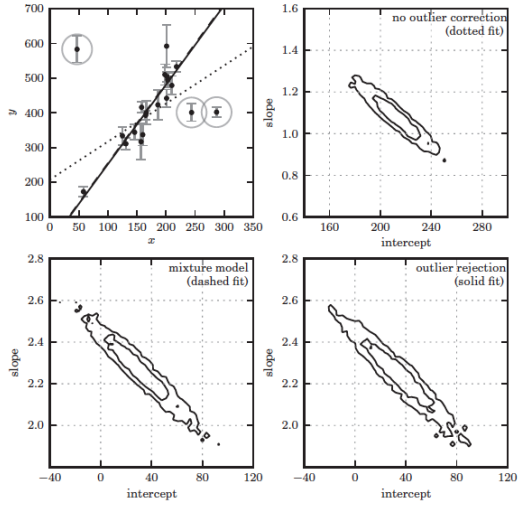
Frequentist version:
(1) generate the distributions for b and m using bootstrap/jackknife.

$$\sigma_m^2 = \frac{1}{B}\sum_{j=1}^{N}\Big(m_j - m\Big)^2$$

(m is the estimate using all the data, m_j is from partial samples).
(2) use these distributions to compute CIs for b and m.

Bayesian version:
(1) generate the posterior distribution of b and m.
(2) use these to compute the MAP values and CrIs.

In general, the off-diagonal terms will be non-zero (the parameters will be correlated).



from AstroML book Section 8.9

To do

Install AstroML if you haven't already!

Read through the first 4 sections of Hogg et al. 2010 for more details, and because you'll need that for the homework.