Statistics for Astronomers: Lecture 20, 2021.01.26

Prof. Sundar Srinivasan

IRyA/UNAM

Review

Sampling techniques. Inverse-transform sampling. Rejection sampling. Importance sampling.

Monte Carlo. Quadrature as an expectation value.

We use Monte Carlo methods in order to either sample from a distribution or compute an expectation value of a function over a distribution.

Simple MC: $\mathbb{E}[f(X)] \approx \frac{1}{N}\sum_{i=1}^{N} f(X_i)$, where $X_i \sim p_X(x)$.

Problem: $p_X(x)$ may be too complicated (esp. multidimensional), and/or difficult to sample from.
Solution: rejection sampling, importance sampling – sample from a proposal distribution instead of the target distribution.

Problem: "curse of high dimensionality" – the proposal needs to be as close as possible to the target; as d increases, the discrepancy increases exponentially.
Solution: Markov Chain Monte Carlo (MCMC); explore multidimensional parameter space by sampling ("travelling") along regions/zones of high probability.

Regression

References

Hogg et al. (2010)

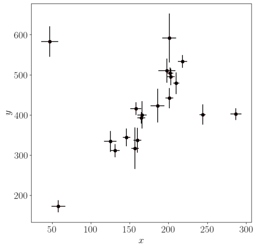
Feigelsen & Babu.

Motivation

We have observations of two random variables X and Y, possibly with uncertainties.

We want to investigate whether they are related, and quantify this relationship (possibly in terms of parameters).

The measurement/intrinsic uncertainties in the data translate to uncertainties in the parameter estimates.



Data from Table 1 of Hogg et al. (2010)

We want point and interval estimates for the parameters to quantify the X-Y relationship.

Once we know the relationship, we can predict future values of Y for given values of X. Given these data, what is the prediction for Y when $X = 105$?

Terminology and general procedure

X variable(s): predictor, regressor, feature, independent†.

Y: outcome, response, target, dependent. Discrete: “classification”; continuous: “regression”.

† Independent variable fallacy (Hogg et al. 2010): pick the one with lower uncertainties.

Terminology and general procedure

X variable(s): predictor, regressor, feature, independent[†].

Y: outcome, response, target, dependent. Discrete: “classification”; continuous: “regression”.

[†] Independent variable fallacy (Hogg et al. 2010): pick the one with lower uncertainties.

Regression function: $Y(x) = \mathbb{E}[Y|X = x]$ Regression model: $Y = f(X) + \epsilon$; $\mathbb{E}[\epsilon] = 0$.

Terminology and general procedure

X variable(s): predictor, regressor, feature, independent[†].

Y: outcome, response, target, dependent. Discrete: “classification”; continuous: “regression”.

[†] Independent variable fallacy (Hogg et al. 2010): pick the one with lower uncertainties.

Regression function: $Y(x) = \mathbb{E}[Y|X = x]$ Regression model: $Y = f(X) + \epsilon$; $\mathbb{E}[\epsilon] = 0$.

Randomness: ϵ – combination of measurement error and intrinsic variation. Typically ignore one w.r.t. the other.
Y random even if X isn't, because of ϵ.

Terminology and general procedure

X variable(s): predictor, regressor, feature, independent[†].

Y: outcome, response, target, dependent. Discrete: "classification"; continuous: "regression".

[†] Independent variable fallacy (Hogg et al. 2010): pick the one with lower uncertainties.

Regression function: $Y(x) = \mathbb{E}[Y|X = x]$ Regression model: $Y = f(X) + \epsilon$; $\mathbb{E}[\epsilon] = 0$.

Randomness: ϵ – combination of measurement error and intrinsic variation. Typically ignore one w.r.t. the other.
Y random even if X isn't, because of ϵ.

ϵ_i associated with y_i drawn from distribution with identical/differing variances: homoskedastic/heteroskedastic uncertainties.

Typically, astronomical measurements are heteroskedastic. Example: magnitudes of stars of a large range of masses.

Terminology and general procedure

X variable(s): predictor, regressor, feature, independent†.

Y: outcome, response, target, dependent. Discrete: "classification"; continuous: "regression".

†Independent variable fallacy (Hogg et al. 2010): pick the one with lower uncertainties.

Regression function: $Y(x) = \mathbb{E}[Y|X=x]$ Regression model: $Y = f(X) + \epsilon$; $\mathbb{E}[\epsilon] = 0$.

Randomness: ϵ – combination of measurement error and intrinsic variation. Typically ignore one w.r.t. the other.
Y random even if X isn't, because of ϵ.

ϵ_i associated with y_i drawn from distribution with identical/differing variances: homoskedastic/heteroskedastic uncertainties.

Typically, astronomical measurements are heteroskedastic. Example: magnitudes of stars of a large range of masses.

Procedure: (1) Dependent variable decision (2) Model choice(s) (3) Method of parameter estimation (choice of objective function/goodness-of-fit). (4) Model validation/selection (Occam's Razor, odds ratios, information criteria).

Terminology and general procedure

X variable(s): predictor, regressor, feature, independent†.

Y: outcome, response, target, dependent. Discrete: "classification"; continuous: "regression".

†Independent variable fallacy (Hogg et al. 2010): pick the one with lower uncertainties.

Regression function: $Y(x) = \mathbb{E}[Y|X = x]$ Regression model: $Y = f(X) + \epsilon;\ \mathbb{E}[\epsilon] = 0$.

Randomness: ϵ – combination of measurement error and intrinsic variation. Typically ignore one w.r.t. the other.
Y random even if X isn't, because of ϵ.

ϵ_i associated with y_i drawn from distribution with identical/differing variances: homoskedastic/heteroskedastic uncertainties.

Typically, astronomical measurements are heteroskedastic. Example: magnitudes of stars of a large range of masses.

Procedure: (1) Dependent variable decision (2) Model choice(s) (3) Method of parameter estimation (choice of objective function/goodness-of-fit). (4) Model validation/selection (Occam's Razor, odds ratios, information criteria).

Regression can be nonparametric (*e.g.*, ML interpolation) or parametric (*e.g.*, χ^2 fitting, MLE).

"Linear" parametric regression: linear in parameters, not necessary in the regressor.

Linear: $Y = mX + c$, $Y = \alpha\sqrt{X^2 + 1}$, $Y = \text{constant}$. Nonlinear: $Y = \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left[-\dfrac{1}{2}\left(\dfrac{X-\mu}{\sigma}\right)^2\right]$, $Y = \dfrac{A}{X+B}$.

Objective function/goodness-of-fit

We want our model predictions for Y to be as close as possible to the observations for Y.

Typically, this means we want to minimise some function $\mathcal{L}(Y_{\rm mod}, Y_{\rm obs})$.

$\mathcal{L}$ is variously called the loss function, objective function, or goodness-of-fit.

Objective function/goodness-of-fit

We want our model predictions for Y to be as close as possible to the observations for Y.

Typically, this means we want to minimise some function $\mathcal{L}(Y_{\rm mod}, Y_{\rm obs})$.

$\mathcal{L}$ is variously called the loss function, objective function, or goodness-of-fit.

Example:

Model predictions $y_{{\rm mod},i}$ for N observations y_i without uncertainties. We want to minimise the residue $y_i - y_{{\rm mod},i}$.

Objective function/goodness-of-fit

We want our model predictions for Y to be as close as possible to the observations for Y.

Typically, this means we want to minimise some function $\mathcal{L}(Y_{\rm mod}, Y_{\rm obs})$.

$\mathcal{L}$ is variously called the loss function, objective function, or goodness-of-fit.

Example:

Model predictions $y_{\mathrm{mod},i}$ for N observations y_i without uncertainties. We want to minimise the residue $y_i - y_{\mathrm{mod},i}$.

If magnitude irrelevant: minimise $\mathcal{L}$ = sum-of-squares of residues, $\sum_{i=1}^{N}(y_i - y_{\mathrm{mod},i})^2$.

Objective function/goodness-of-fit

We want our model predictions for Y to be as close as possible to the observations for Y.

Typically, this means we want to minimise some function $\mathcal{L}(Y_{\mathrm{mod}}, Y_{\mathrm{obs}})$.

$\mathcal{L}$ is variously called the loss function, objective function, or goodness-of-fit.

Example:

Model predictions $y_{\mathrm{mod},i}$ for N observations y_i without uncertainties. We want to minimise the residue $y_i - y_{\mathrm{mod},i}$.

If magnitude irrelevant: minimise $\mathcal{L} =$ sum-of-squares of residues, $\sum_{i=1}^{N}(y_i - y_{\mathrm{mod},i})^2$.

If we have uncertainties, we want $y_i - y_{\mathrm{mod},i}$ small compared to uncertainty $\sigma_{y,i}$.

Objective function/goodness-of-fit

We want our model predictions for Y to be as close as possible to the observations for Y.

Typically, this means we want to minimise some function $\mathcal{L}(Y_{\mathrm{mod}}, Y_{\mathrm{obs}})$.

$\mathcal{L}$ is variously called the loss function, objective function, or goodness-of-fit.

Example:

Model predictions $y_{\mathrm{mod},i}$ for N observations y_i without uncertainties. We want to minimise the residue $y_i - y_{\mathrm{mod},i}$.

If magnitude irrelevant: minimise $\mathcal{L}$ = sum-of-squares of residues, $\sum_{i=1}^{N} (y_i - y_{\mathrm{mod},i})^2$.

If we have uncertainties, we want $y_i - y_{\mathrm{mod},i}$ small compared to uncertainty $\sigma_{y,i}$.

$\implies \mathcal{L}$ = weighted sum-of-squares of residues, $\sum_{i=1}^{N} \frac{(y_i - y_{\mathrm{mod},i})^2}{\sigma^2_{y,i}} = (\mathbf{y} - \mathbf{y}_{\mathrm{mod}})^T \cdot \mathbf{\Sigma}^{-1} \cdot (\mathbf{y} - \mathbf{y}_{\mathrm{mod}})$ in matrix form,

with $N \times 1$ column vectors $\mathbf{y}$ and $\mathbf{y}_{\mathrm{mod}}$ and $\mathbf{\Sigma}$ the $N \times N$ covariance matrix.

$\mathbf{\Sigma}$ stores information about correlations in the uncertainties.

If errors are homoskedastic, $\mathbf{\Sigma}^{-1} = \frac{1}{\sigma^2} \mathbb{I}_{N \times N}$

Objective function/goodness-of-fit

We want our model predictions for Y to be as close as possible to the observations for Y.

Typically, this means we want to minimise some function $\mathcal{L}(Y_{\mathrm{mod}}, Y_{\mathrm{obs}})$.

$\mathcal{L}$ is variously called the loss function, objective function, or goodness-of-fit.

Example:

Model predictions $y_{\mathrm{mod},i}$ for N observations y_i without uncertainties. We want to minimise the residue $y_i - y_{\mathrm{mod},i}$.

If magnitude irrelevant: minimise $\mathcal{L}$ = sum-of-squares of residues, $\sum_{i=1}^{N}(y_i - y_{\mathrm{mod},i})^2$.

If we have uncertainties, we want $y_i - y_{\mathrm{mod},i}$ small compared to uncertainty $\sigma_{y,i}$.

$\implies \mathcal{L}$ = weighted sum-of-squares of residues, $\sum_{i=1}^{N}\frac{(y_i - y_{\mathrm{mod},i})^2}{\sigma_{y,i}^2} = (\mathbf{y} - \mathbf{y}_{\mathrm{mod}})^T \cdot \mathbf{\Sigma}^{-1} \cdot (\mathbf{y} - \mathbf{y}_{\mathrm{mod}})$ in matrix form,

with $N \times 1$ column vectors $\mathbf{y}$ and $\mathbf{y}_{\mathrm{mod}}$ and $\mathbf{\Sigma}$ the $N \times N$ covariance matrix.

$\mathbf{\Sigma}$ stores information about correlations in the uncertainties.

If errors are homoskedastic, $\mathbf{\Sigma}^{-1} = \frac{1}{\sigma^2}\,\mathbb{I}_{N\times N}$

$\mathbf{y}_{\mathrm{mod}} \propto$ parameters $\boldsymbol{\theta} \implies$, estimate by minimising $\mathcal{L}(\boldsymbol{\theta})$.

This is done by setting the derivative w.r.t. each parameter to zero (similar to MLE!).

Ordinary least-squares (OLS) – linear model

Homoskedastic errors: $y_{\rm obs} = y_{{\rm mod},i} + \epsilon_i$, with $\mathbb{E}[\epsilon_i] = 0$, ${\rm Var}[\epsilon_i] = \sigma^2$.

Model linear in the regressor: $y_{{\rm mod},i} = mx_i + b$. $(m, b) =$ slope and intercept.

Ordinary least-squares (OLS) – linear model

Homoskedastic errors: $y_{\mathrm{obs}} = y_{\mathrm{mod},i} + \epsilon_i$, with $\mathbb{E}[\epsilon_i] = 0$, $\mathrm{Var}[\epsilon_i] = \sigma^2$.

Model linear in the regressor: $y_{\mathrm{mod},i} = mx_i + b$. (m, b) = slope and intercept.

$$\mathcal{L} = \sum_{i=1}^{N} \left(y_i - y_{\mathrm{mod},i}\right)^2 = \sum_{i=1}^{N} \left(y_i - mx_i - b\right)^2.$$

Optimisation:

Ordinary least-squares (OLS) – linear model

Homoskedastic errors: $y_{\rm obs} = y_{{\rm mod},i} + \epsilon_i$, with $\mathbb{E}[\epsilon_i] = 0$, $\mathrm{Var}[\epsilon_i] = \sigma^2$.

Model linear in the regressor: $y_{{\rm mod},i} = mx_i + b$. $(m, b) =$ slope and intercept.

$$\mathcal{L} = \sum_{i=1}^{N} \Big(y_i - y_{{\rm mod},i}\Big)^2 = \sum_{i=1}^{N} \Big(y_i - mx_i - b\Big)^2.$$

Optimisation:

$$\frac{\partial \mathcal{L}}{\partial m}\Big|_{(m,b)=(\hat{m},\hat{b})} \propto \sum_{i=1}^{N} \Big(y_i - \hat{m}x_i - \hat{b}\Big) \cdot x_i = 0 \Longrightarrow \hat{m} = \frac{\displaystyle\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\displaystyle\sum_{i=1}^{N}(x_i - \bar{x})^2} \equiv \frac{S_{xy}}{S_{xx}}.$$

$$\frac{\partial \mathcal{L}}{\partial b}\Big|_{(m,b)=(\hat{m},\hat{b})} \propto \sum_{i=1}^{N} \Big(y_i - \hat{m}x_i - \hat{b}\Big) = 0 \Longrightarrow \hat{b} = \bar{y} - \hat{m}\bar{x}.$$

Ordinary least-squares (OLS) – linear model

Homoskedastic errors: $y_{\mathrm{obs}} = y_{\mathrm{mod},i} + \epsilon_i$, with $\mathbb{E}[\epsilon_i] = 0$, $\mathrm{Var}[\epsilon_i] = \sigma^2$.

Model linear in the regressor: $y_{\mathrm{mod},i} = mx_i + b$. $(m, b) =$ slope and intercept.

$$\mathcal{L} = \sum_{i=1}^{N} \Big(y_i - y_{\mathrm{mod},i}\Big)^2 = \sum_{i=1}^{N} \Big(y_i - mx_i - b\Big)^2.$$

Optimisation:

$$\frac{\partial \mathcal{L}}{\partial m}\Big|_{(m,b)=(\hat{m},\hat{b})} \propto \sum_{i=1}^{N} \Big(y_i - \hat{m}x_i - \hat{b}\Big) \cdot x_i = 0 \Longrightarrow \hat{m} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \equiv \frac{S_{xy}}{S_{xx}}.$$

$$\frac{\partial \mathcal{L}}{\partial b}\Big|_{(m,b)=(\hat{m},\hat{b})} \propto \sum_{i=1}^{N} \Big(y_i - \hat{m}x_i - \hat{b}\Big) = 0 \Longrightarrow \hat{b} = \bar{y} - \hat{m}\bar{x}.$$

Gaussian errors: $\epsilon_i \sim \mathcal{N}(0, \sigma^2), \mathcal{L} \sim \chi^2_{N-2}$ (2 parameters estimated using data). "χ^2 fitting".

Ordinary least-squares (OLS) – linear model

Homoskedastic errors: $y_{\rm obs} = y_{{\rm mod},i} + \epsilon_i$, with $\mathbb{E}[\epsilon_i] = 0$, ${\rm Var}[\epsilon_i] = \sigma^2$.

Model linear in the regressor: $y_{{\rm mod},i} = mx_i + b$. $(m, b) =$ slope and intercept.

$$\mathcal{L} = \sum_{i=1}^{N} \Big(y_i - y_{{\rm mod},i}\Big)^2 = \sum_{i=1}^{N} \Big(y_i - mx_i - b\Big)^2.$$

Optimisation:

$$\frac{\partial \mathcal{L}}{\partial m}\Big|_{(m,b)=(\hat{m},\hat{b})} \propto \sum_{i=1}^{N} \Big(y_i - \hat{m}x_i - \hat{b}\Big) \cdot x_i = 0 \implies \hat{m} = \frac{\sum\limits_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{N}(x_i - \bar{x})^2} \equiv \frac{S_{xy}}{S_{xx}}.$$

$$\frac{\partial \mathcal{L}}{\partial b}\Big|_{(m,b)=(\hat{m},\hat{b})} \propto \sum_{i=1}^{N} \Big(y_i - \hat{m}x_i - \hat{b}\Big) = 0 \implies \hat{b} = \bar{y} - \hat{m}\bar{x}.$$

Gaussian errors: $\epsilon_i \sim \mathcal{N}(0, \sigma^2), \mathcal{L} \sim \chi^2_{N-2}$ (2 parameters estimated using data). "χ^2 fitting".

Estimate for the variance in **y** is then $S^2 = \dfrac{\mathcal{L}}{N-2} \sim \sigma^2 \dfrac{\chi^2_{N-2}}{N-2}$. Reduced χ^2.

OLS linear model – parameter variances

Recall: x_i not random but y_i random because of the uncertainties ϵ_i, which have variance σ^2.

$$S_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}$$

OLS linear model – parameter variances

Recall: x_i not random but y_i random because of the uncertainties ϵ_i, which have variance σ^2.

$$S_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}$$

$$Var(S_{xy}) = \frac{1}{N^2}\sum_{i=1}^{N} x_i^2 \sigma^2 - \bar{x}^2 \frac{\sigma^2}{N} = \sigma^2 \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2\right) = \sigma^2 S_{xx}.$$

OLS linear model – parameter variances

Recall: x_i not random but y_i random because of the uncertainties ϵ_i, which have variance σ^2.

$$S_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}$$

$$Var(S_{xy}) = \frac{1}{N^2}\sum_{i=1}^{N} x_i^2 \sigma^2 - \bar{x}^2\frac{\sigma^2}{N} = \sigma^2\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2\right) = \sigma^2 S_{xx}.$$

$$\implies Var(\hat{m}) = Var\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}^2} Var(S_{xy}) = \frac{\sigma^2}{S_{xx}}$$

OLS linear model – parameter variances

Recall: x_i not random but y_i random because of the uncertainties ϵ_i, which have variance σ^2.

$$S_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}$$

$$Var(S_{xy}) = \frac{1}{N^2}\sum_{i=1}^{N} x_i^2 \sigma^2 - \bar{x}^2 \frac{\sigma^2}{N} = \sigma^2\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2\right) = \sigma^2 S_{xx}.$$

$$\implies Var(\hat{m}) = Var\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}^2} Var(S_{xy}) = \frac{\sigma^2}{S_{xx}}$$

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

OLS linear model – parameter variances

Recall: x_i not random but y_i random because of the uncertainties ϵ_i, which have variance σ^2.

$$S_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}$$

$$Var(S_{xy}) = \frac{1}{N^2}\sum_{i=1}^{N} x_i^2 \sigma^2 - \bar{x}^2 \frac{\sigma^2}{N} = \sigma^2 \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2\right) = \sigma^2 S_{xx}.$$

$$\implies Var(\hat{m}) = Var\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}^2} Var(S_{xy}) = \frac{\sigma^2}{S_{xx}}$$

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

For large N, best-fit parameter estimates normally distributed about their means with variances as above.

If σ unknown, must also be estimated from data: $\hat{\sigma} = S$. $\hat{m}, \hat{b}$ then t-distributed.

OLS linear model – parameter variances

Recall: x_i not random but y_i random because of the uncertainties ϵ_i, which have variance σ^2.

$$S_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}$$

$$Var(S_{xy}) = \frac{1}{N^2}\sum_{i=1}^{N} x_i^2 \sigma^2 - \bar{x}^2 \frac{\sigma^2}{N} = \sigma^2 \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2\right) = \sigma^2 S_{xx}.$$

$$\implies Var(\hat{m}) = Var\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}^2} Var(S_{xy}) = \frac{\sigma^2}{S_{xx}}$$

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

For large N, best-fit parameter estimates normally distributed about their means with variances as above.

If σ unknown, must also be estimated from data: $\hat{\sigma} = S$. $\hat{m}, \hat{b}$ then t-distributed.

OLS linear model – parameter variances

Recall: x_i not random but y_i random because of the uncertainties ϵ_i, which have variance σ^2.

$$S_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}$$

$$Var(S_{xy}) = \frac{1}{N^2}\sum_{i=1}^{N} x_i^2\sigma^2 - \bar{x}^2\frac{\sigma^2}{N} = \sigma^2\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2\right) = \sigma^2 S_{xx}.$$

$$\implies Var(\hat{m}) = Var\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}^2}Var(S_{xy}) = \frac{\sigma^2}{S_{xx}}$$

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

For large N, best-fit parameter estimates normally distributed about their means with variances as above.

If σ unknown, must also be estimated from data: $\hat{\sigma} = S$. $\hat{m}, \hat{b}$ then t-distributed.

Distribution of predicted value of y:

$y_{\mathrm{pred}}(x)$ is t-distributed around $mx + b$ with standard deviation $S\sqrt{1 + \frac{1}{N}\frac{(x-\bar{x})^2}{S_{xx}}}$.

OLS in matrix notation

For multivariate problems, it's much easier to work with matrices.

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

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

Covariance matrix $\Sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$, with ρ_{ij} the correlation coefficient between σ_i and σ_j.

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

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

If $\mathcal{L}$ 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.

To the Jupyter notebook!

Demonstration of Exercise 1 from Hogg et al. (2010)

1. ▸ Download this Jupyter notebook.
2. ▸ Navigate to Colaboratory.
3. Sign in
4. Click on "Upload" and upload the notebook you downloaded in step 1.

OLS – comparison to MLE

If the uncertainties ϵ are normally distributed and homoskedastic, the associated likelihood is

$$\mathscr{L}(m,b) = \prod_{i=1}^{N} \exp\left[-\frac{1}{2}\left(\frac{y_i - mx_i - b}{\sigma}\right)^2\right] \Rightarrow \ln \mathscr{L} = \mathrm{constant} - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{y_i - mx_i - b}{\sigma}\right)^2.$$

$$= \mathrm{constant} - \frac{1}{2}\,\mathcal{L}.$$

The objective function $\mathcal{L}$ is related to $\ln \mathscr{L}$, so the results from optimising $\mathcal{L}$ are equivalent to the maximum likelihood estimate for this problem.

For heteroskedastic uncertainties, we replace σ with N distinct values σ_i. The matrix product version of the log-likelihood is

$$\ln \mathscr{L} = \mathrm{constant} + (\mathbf{y} - \mathbf{Ax})^T \mathbf{\Sigma}^{-1} (\mathbf{y} - \mathbf{Ax}).$$

Robust regression

A robust statistic is one whose value isn't sensitive to outliers.
Examples: Median vs. mean, IQR or MADM vs. standard deviation.

Manual removal of outliers is neither objective nor reproducible.

"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.

Outlier rejection can be done using a robust technique. Many such techniques exist (see Feigelsen & Babu).

One example: Bayesian Outlier Rejection (Hogg et al. (2010); AstroML Sec. 8.9). Similar to assuming a Gaussian mixture model for the data.

Robust regression (Outlier rejection)

Core assumption: outliers are drawn from a different distribution than the "true data" values.

Robust regression (Outlier rejection)

Core assumption: outliers are drawn from a different distribution than the “true data” values.

Data model: Gaussian mixture of “true data” distribution and outlier distribution.

True data is such that residue $\epsilon_i \equiv y_{\mathrm{data},i} - y_{\mathrm{mod},i} \sim \mathcal{N}(0, \sigma_i^2)$.

Outliers are such that residue $\varepsilon_i \equiv y_{\mathrm{data},i} - y_{\mathrm{mod},i} \sim \mathcal{N}(Y_b, V_b)$.

Robust regression (Outlier rejection)

Core assumption: outliers are drawn from a different distribution than the "true data" values.

Data model: Gaussian mixture of "true data" distribution and outlier distribution.

True data is such that residue $\epsilon_i \equiv y_{\mathrm{data},i} - y_{\mathrm{mod},i} \sim \mathcal{N}(0, \sigma_i^2)$.

Outliers are such that residue $\varepsilon_i \equiv y_{\mathrm{data},i} - y_{\mathrm{mod},i} \sim \mathcal{N}(Y_b, V_b)$.

Probability that a given data point is an outlier $\equiv P_b$.

$$\Longrightarrow p(y_i \mid x_i, m, b, \sigma_i, P_b, V_b) = \frac{(1-P_b)}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\left(\frac{y_i - mx_i - b}{\sigma_i}\right)^2\right] + \frac{P_b}{\sqrt{2\pi(V_b + \sigma_i^2)}} \exp\left[-\frac{1}{2}\frac{(y_i - Y_b)^2}{V_b + \sigma_i^2}\right]$$

Robust regression (Outlier rejection)

Core assumption: outliers are drawn from a different distribution than the “true data” values.

Data model: Gaussian mixture of “true data” distribution and outlier distribution.

True data is such that residue $\epsilon_i \equiv y_{\mathrm{data},i} - y_{\mathrm{mod},i} \sim \mathcal{N}(0, \sigma_i^2)$.

Outliers are such that residue $\varepsilon_i \equiv y_{\mathrm{data},i} - y_{\mathrm{mod},i} \sim \mathcal{N}(Y_b, V_b)$.

Probability that a given data point is an outlier $\equiv P_b$.

$$\implies p(y_i \mid x_i, m, b, \sigma_i, P_b, V_b) = \frac{(1-P_b)}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\left(\frac{y_i - mx_i - b}{\sigma_i}\right)^2\right] + \frac{P_b}{\sqrt{2\pi(V_b + \sigma_i^2)}} \exp\left[-\frac{1}{2}\frac{(y_i - Y_b)^2}{V_b + \sigma_i^2}\right]$$

OR, equivalently, 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”).

Then, probability that a data point is “bad” $= P(q_i = 0) \equiv P_b = \mathrm{constant}$.

Robust regression (Outlier rejection)

Core assumption: outliers are drawn from a different distribution than the “true data” values.

Data model: Gaussian mixture of “true data” distribution and outlier distribution.

True data is such that residue $\epsilon_i \equiv y_{\mathrm{data},i} - y_{\mathrm{mod},i} \sim \mathcal{N}(0, \sigma_i^2)$.

Outliers are such that residue $\varepsilon_i \equiv y_{\mathrm{data},i} - y_{\mathrm{mod},i} \sim \mathcal{N}(Y_b, V_b)$.

Probability that a given data point is an outlier $\equiv P_b$.

$$\implies p(y_i \mid x_i, m, b, \sigma_i, P_b, V_b) = \frac{(1-P_b)}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\left(\frac{y_i - mx_i - b}{\sigma_i}\right)^2\right] + \frac{P_b}{\sqrt{2\pi(V_b+\sigma_i^2)}} \exp\left[-\frac{1}{2}\frac{(y_i - Y_b)^2}{V_b + \sigma_i^2}\right]$$

OR, equivalently, 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”).

Then, probability that a data point is “bad” $= P(q_i = 0) \equiv P_b = \mathrm{constant}$.

$$\implies p(y_i \mid x_i, m, b, \sigma_i, q_i, V_b) = \left\{\exp\left[-\frac{1}{2}\left(\frac{y_i - mx_i - b}{\sigma_i}\right)^2\right]\right\}^{q_i} \left\{\frac{1}{\sqrt{2\pi(V_b+\sigma_i^2)}} \exp\left[-\frac{1}{2}\frac{(y_i - Y_b)^2}{V_b + \sigma_i^2}\right]\right\}^{1-q_i}$$

Robust regression (Outlier rejection)

Core assumption: outliers are drawn from a different distribution than the "true data" values.

Data model: Gaussian mixture of "true data" distribution and outlier distribution.

True data is such that residue $\epsilon_i \equiv y_{\mathrm{data},i} - y_{\mathrm{mod},i} \sim \mathscr{N}(0, \sigma_i^2)$.

Outliers are such that residue $\varepsilon_i \equiv y_{\mathrm{data},i} - y_{\mathrm{mod},i} \sim \mathscr{N}(Y_b, V_b)$.

Probability that a given data point is an outlier $\equiv P_b$.

$$\Longrightarrow p(y_i \mid x_i, m, b, \sigma_i, P_b, V_b) = \frac{(1-P_b)}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\left(\frac{y_i - mx_i - b}{\sigma_i}\right)^2\right] + \frac{P_b}{\sqrt{2\pi(V_b+\sigma_i^2)}} \exp\left[-\frac{1}{2}\frac{(y_i - Y_b)^2}{V_b+\sigma_i^2}\right]$$

OR, equivalently, 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").

Then, probability that a data point is "bad" $= P(q_i = 0) \equiv P_b = \mathrm{constant}$.

$$\Longrightarrow p(y_i \mid x_i, m, b, \sigma_i, q_i, V_b) = \left\{\exp\left[-\frac{1}{2}\left(\frac{y_i - mx_i - b}{\sigma_i}\right)^2\right]\right\}^{q_i} \left\{\frac{1}{\sqrt{2\pi(V_b+\sigma_i^2)}} \exp\left[-\frac{1}{2}\frac{(y_i - Y_b)^2}{V_b+\sigma_i^2}\right]\right\}^{1-q_i}$$

Total # parameters: $2 + (N + 3)$. The q_i are nuisance parameters, can marginalise over them.

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 and priors

With ‘fg’ and ‘bg’ referring to the true data (“foreground”) and outliers (“background”),

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

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

Bayesian outlier rejection: likelihood and priors

With ‘fg’ and ‘bg’ referring to the true data (“foreground”) and outliers (“background”),

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

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

In terms of P_b, instead,

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

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

Bayesian outlier rejection: likelihood and priors

With ‘fg’ and ‘bg’ referring to the true data (“foreground”) and outliers (“background”),

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

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

In terms of P_b, instead,

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

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

Joint prior on the $\{q_i\}$: $p(\{q_i\}|P_b) = \prod_{i=1}^{N} (1-P_b)^{q_i} P_b^{1-q_i}$.

For P_b, Y_b, (“locations”) and V_b (“scale”), we can use prior information or uninformative priors.

Bayesian outlier rejection: marginalisation

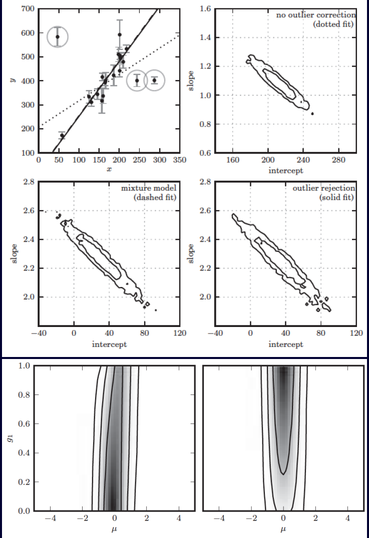
The posterior is $\propto$ likelihood $\times$ the priors.

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 (value $= 0$ 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).



source: AstroML book Sections 5.6.7 and 8.9

Parameter uncertainties

For the OLS setup, the parameter uncertainties were $(\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A})^{-1} = \begin{bmatrix} \sigma_b^2 & \sigma_b\sigma_m \\ \sigma_m\sigma_b & \sigma_m^2 \end{bmatrix}$.

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

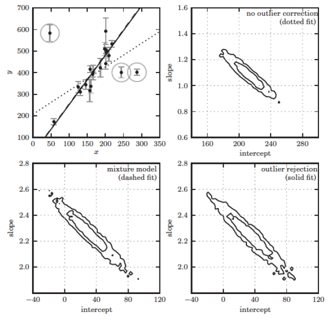
Frequentist version:
(1) generate the distributions for b and m using bootstrap.

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

(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.

Correlated parameters
→ Nonzero off-diagonal terms.



from AstroML book Section 8.9