Regression

Eric Feigelson

2nd East Asian Workshops on Astrostatistics
Nanjing & Guiyang
July 2018

Regression vs. density estimation

Density estimation is nonparametric: no functional form for the shape of the distribution, or relationship between the variables, is assumed. It is usually applied to 1-3-dimensional problems.

Regression differs in two respects:

- It addresses problems where one seeks to understand the dependency of a *pre-specified response variable* Y on one (or more) independent variables X (or **X**).

  The science question should determine the response variable

- It addresses problems of modeling where the *functional form of the relationship between the variables is pre-specified*. The function has parameters, and the goal of the regression is to find the `best' parameter values that `fit' the data.

  Astronomers perform regressions with heuristic functions (e.g. power laws) and with functions from astrophysical theory

Classical regression model:

$$E[Y|X] = f(X, \theta) + \epsilon$$

``The expectation (mean) of the dependent (response) variable Y for a given value of the independent variable X (or vector of variables **X**) is equal to a specified function *f*, which depends on both X and a vector of parameters θ, plus a random error (scatter)."

The `error' ε is commonly assumed to be a normal (Gaussian) i.i.d. random variable with zero mean, $\varepsilon = N(0,\sigma^2)$. Note that all of the randomness is in this error term; the functional relationship is deterministic with a known mathematical form.

Warning

Astronomers may be using classical regression too often, perhaps due to its familiarity compared to other (e.g. nonparametric) statistical methods.

- If there is no basis for choosing a functional form (e.g. an astrophysical theory), then nonparametric density estimation may be more appropriate than regression using a heuristic function.

- If there is no basis for choosing the dependency relationship (i.e. that Y depends on X, rather than X on Y or both on some hidden variables), then a form of regression that treats the variables symmetrically should be used (e.g. OLS bisector, orthogonal regression, Principal Component Analysis).

The error term ε

There may be different causes of the scatter:

- It could be intrinsic to the underlying population (`equation error’). This is called a `structural regression model’.

- It may arise from an imperfect measurement process (`measurement error’) and the true Y exactly satisfy Y=f(X).

  This is called a `functional regression model’.

- Or both intrinsic and measurement errors may be present.

Astronomers encounter all of these situations

Parameter estimation & model selection

Once a mathematical model is chosen, and a dataset is provided, then the `best fit' parameters are estimated by one (or more) of the techniques discussed in MSMA Chpt. 3:

- Method of moments
- Ordinary least squares (OLS, L_2)
- Least absolute deviation (L_1)
- Maximum likelihood estimation (MLE)
- Bayesian inference

Seek balance between model complexity and parsimony (Occam's Razor):

- Does the ΛCDM model have a w-dot term?
- Are three or four planets orbiting the star?
- Is the star cluster an isothermal sphere or ellipsoid?

Choice of model form and complexity is called `model selection'. Methods include: χ^2_ν, BIC, AIC, …

The final model should be validated against the dataset (or other datasets) using goodness-of-fit test (e.g. Anderson-Darling test with bootstrap resamples for significance levels) and residual analysis.

Important!
In statistical parlance, `linear' means `linear in the parameters β_i', not `linear in the variable X'.

Examples of linear regression functions:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

1st order polynomial

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

high order polynomial

$$Y = \beta_0 e^{-X} + \epsilon$$

exponential decay

$$Y = \beta_0 + \beta_1 \cos X + \beta_2 \sin X + \epsilon$$

periodic sinusoid with fixed phase

Examples of non-linear regression functions:

$$Y = \left(\frac{X}{\beta_0}\right)^{-\beta_1} + \epsilon$$ power law (Pareto)

$$Y = \frac{\beta_0}{1+(X/\beta_1)^2} + \epsilon$$ isothermal sphere

$$Y = \beta_0 + \beta_1 \cos(X+\beta_2) + \beta_3 \sin(X+\beta_2) + \epsilon$$ sinusoid with arbitrary phase

$$Y = \begin{cases} \beta_0 + \beta_1 X & \text{for } X < x_o \\ \beta_2 + \beta_3 X & \text{for } X > x_o \end{cases}$$ segmented linear

Assumptions of ordinary least squares regression

- The model is correctly specified (i.e. the population truly follows the specified relationship)
- The errors have (conditional) mean zero: $E[\varepsilon|X] = E[\varepsilon] = 0$
- The errors are homoscedastic, $E[\varepsilon_i^2|X] = \sigma^2$, and uncorrelated, $E[\varepsilon_i\varepsilon_j] = 0$ ($i \neq j$)
- For some purposes, assume the errors are normally distributed, $\varepsilon|X \sim N(0,\sigma^2)$
- For some purposes, assume the data are i.i.d., (x_i,y_i) are independent from (x_j,y_j) but share the same distribution
- For multivariate covariates $\mathbf{X}=(X_1, X_2,\ldots, X_p)$, some additional assumptions:
  - $X_i \ldots X_p$ are linearly independent
  - The matrix $E[X_i,X_i']$ is positive-definite

OLS gives the maximum likelihood estimator for regression when $\varepsilon|X \sim N(0,\sigma^2)$

A very common astronomical regression procedure

Dataset of the form: $(X_i, \sigma_{X,i}, Y_i, \sigma_{Y,i})$
(bivariate data with heteroscedastic measurement errors with known variances)

Linear (or other) regression model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

Best fit parameters from minimizing the function: $$S_{r,wt} = \sum_{i=1}^{n} \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{\sigma_{Y,i}^2}$$

The distributions of the parameters are estimated from tables of the χ^2 distribution (e.g. $\Delta\chi^2 = 2.71$ around best-fit parameter values for 90% confidence interval for 1 degree of freedom; *Numerical Recipes*, Fig 15.6.4)

This procedure is often called `minimum chi-square regression' or `chi-square fitting' because a random variable that is the sum of squared normal random variables follows a χ^2 distribution. If all of the variance in Y is attributable to the known measurement errors $\sigma_{Y,i}$ and these errors are normally distributed, then the model is valid.

However, from a statistical viewpoint ….

… this is a non-standard procedure! Pearson's (1900) chi-square statistic was developed for a very specific problem: hypothesis testing for the multinomial experiment producing a contingency table of counts in a *pre-determined* number k of categories.

$$X^2_P(\theta_p) = \frac{\sum_{i=1}^{k} [O_i - M_i(\theta_p)]^2}{M_i(\theta_p)}$$

where O_i are the observed counts, and M_i are the model counts dependent on the p model parameters θ. The weights (denominator) are completely different than in the astronomers' procedure.

A better approach uses a more complicated likelihood that includes the measurement errors & model error, and proceeds with MLE or Bayesian inference. See important article by Brandon C. Kelly, ApJ 2007.

Concluding remarks

Regression is very widely used in astronomy, and often in a reasonable fashion. It is crucial for understanding situations where the data are explained by astrophysical models.

But poor practice does occur:

- Overuse of heuristic models
- Ill-defined response variable
- Improper used of `minimum chi-squared' method
- Inadequate model selection
- Inadequate residual analysis
- Overuse of Bayesian inference with uninformative priors