Applied Bayesian Data Analysis — Chapter 17

Kim Albertsson

CERN and Luleå University of Technology

kim.albertsson@ltu.se

December 9, 2019

Chapter 17

Metric predicted Variable with One Metric Predictor

A.k.a. Univariate Regression

Recap of where we are

So far: Modelled a central tendency with added noise.

$$y = \beta_0 + \varepsilon \implies y \sim \mathcal{N}(\beta_0 + \mu_\varepsilon, \sigma_\varepsilon) \sim \mathcal{N}(\mu, \sigma)$$

In hierarchical models: Noise was *structured*, i.e. we changed the form of μ_ε and σ_ε.

Now: Expand structure in central tendency estimation

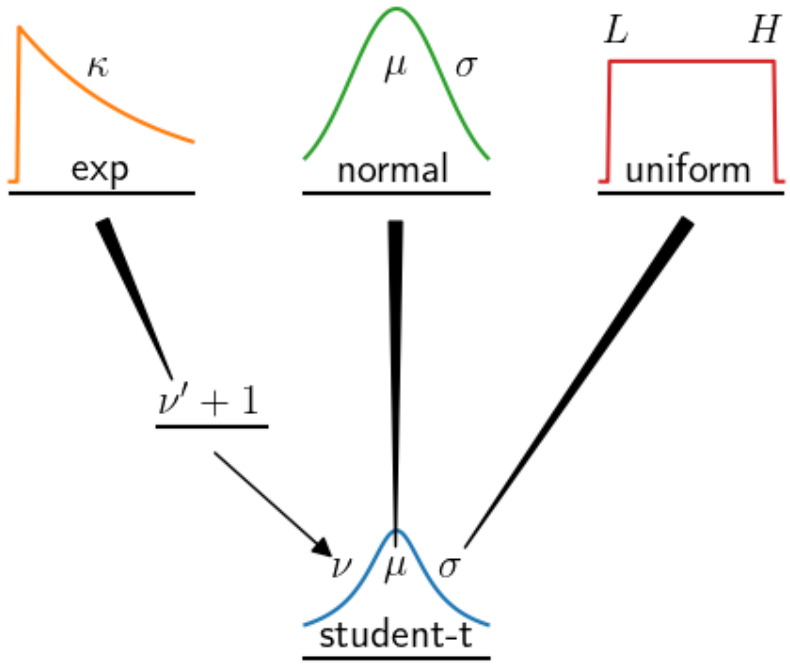
$$y = \beta_0 + \beta_1 x + \varepsilon \implies y \sim \mathcal{N}(\beta_0 + \beta_1 x + \mu_\varepsilon, \sigma_\varepsilon) \sim \mathcal{N}(\mu = c_0 + \beta_1 x, \sigma)$$

Note: We can use regression on any parameter e.g. σ, however, the model $y = \beta_0 + \beta_1 x + \varepsilon$ is a very useful one, hence we focus on this first.

Recap of where we are (in pictures!)

The first kind of model we considered was:

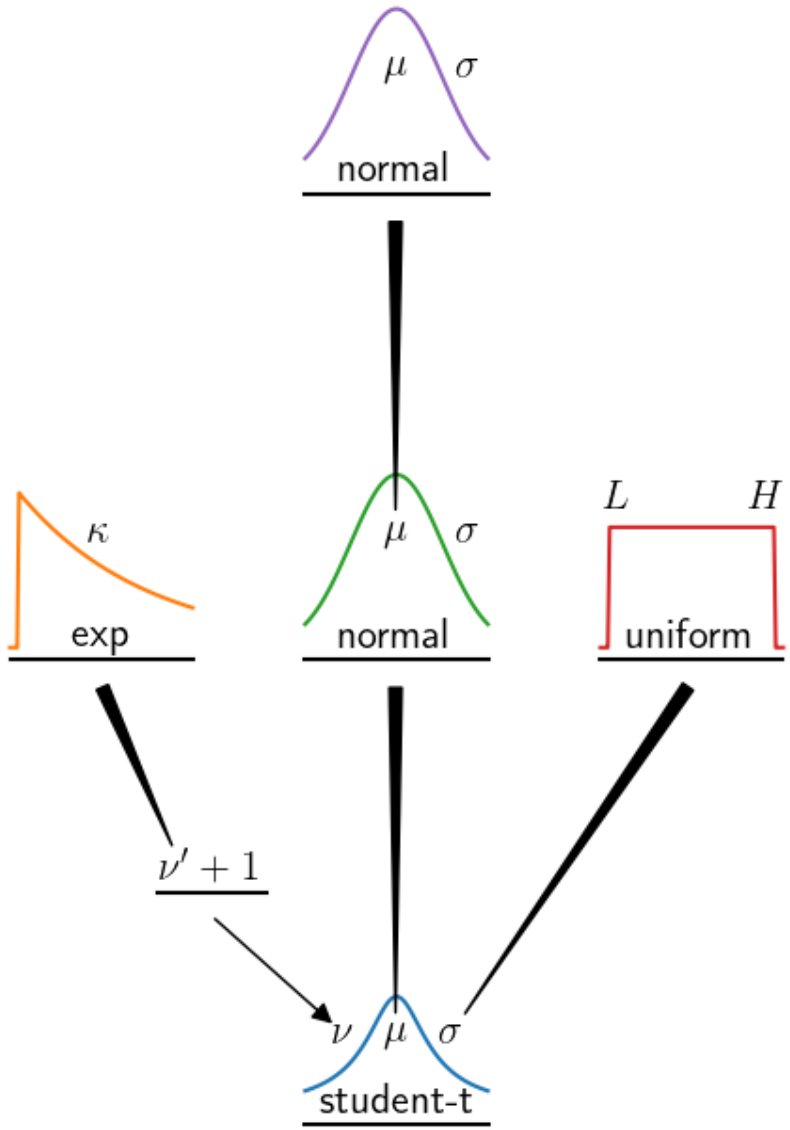
$$\begin{aligned} \text{model: } & y \sim \text{pdf}(\theta) \\ \text{prior: } & \theta \sim \text{pdf}(\varphi) \\ & \varphi = \Phi \end{aligned}$$



Recap of where we are (in pictures!)

We then considered hierarchical models, these provide *structured noise*:

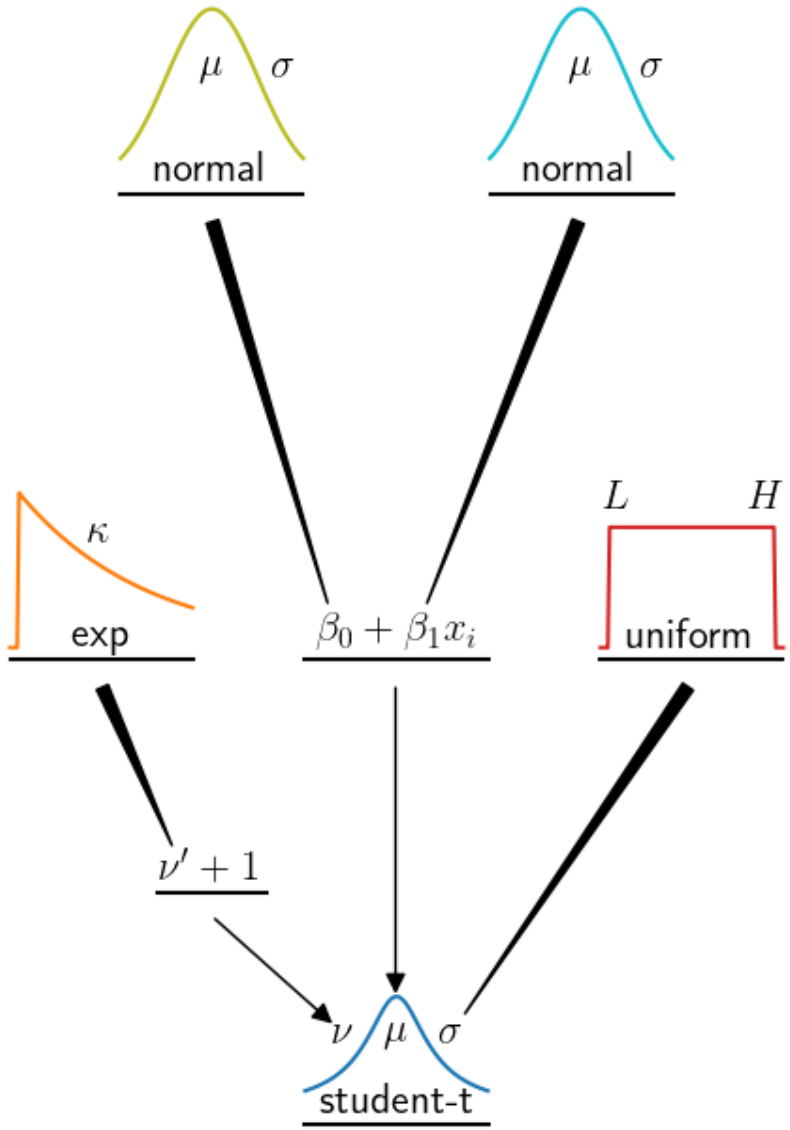
$$\begin{aligned} \text{model: } & y \sim \text{pdf}(\theta) \\ \text{prior: } & \theta \sim \text{pdf}(\varphi) \\ & \varphi \sim \text{pdf}(\lambda) \\ & \lambda = \Lambda \end{aligned}$$



Recap of where we are (in pictures!)

Now:

$$
\begin{aligned}
\text{model:}\quad & y \sim \text{pdf}(\theta) \\
\text{prior:}\quad & \theta = \beta_0 + \beta_1 x \\
& \beta_0 \sim \text{pdf}(\varphi_0) \\
& \beta_1 \sim \text{pdf}(\varphi_1) \\
& \varphi_0 = \Phi_0 \\
& \varphi_1 = \Phi_1
\end{aligned}
$$



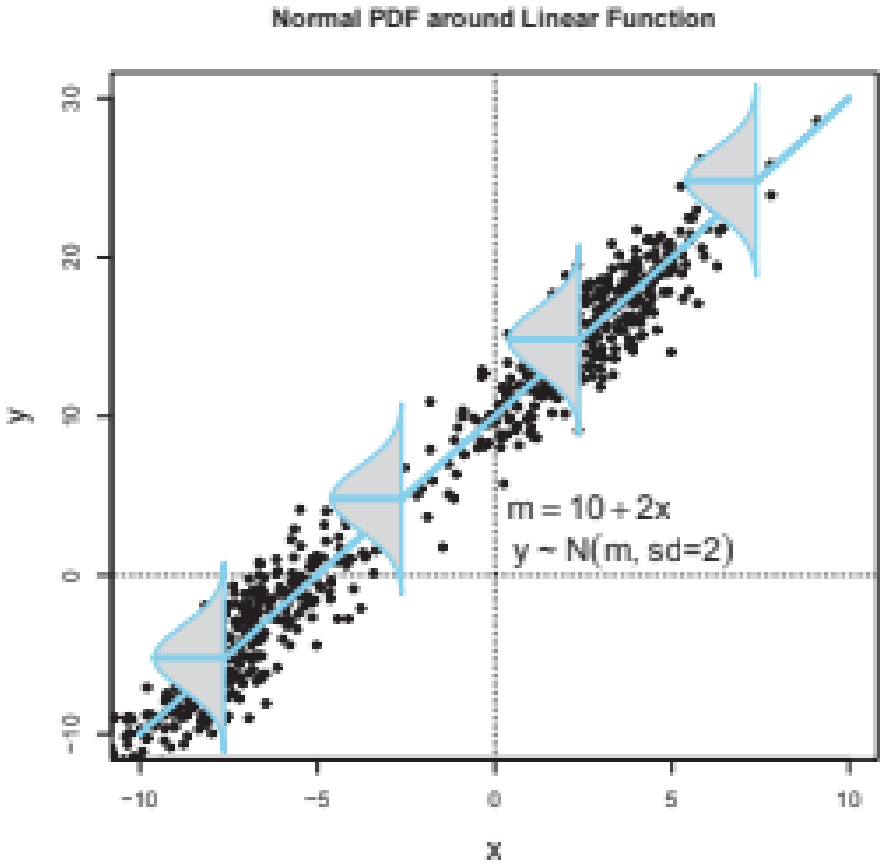
Simple Linear Regression

Model: $y = \beta_0 + \beta_1 x + \varepsilon$

Note: Given a particular distribution $x \sim p(X = x)$ the distribution for y will be a smeared version of this (in this case).

Homogeneity of variance: Noise magnitude is independent on scale of input. (Note: Measurement imprecision often increases with scale.)

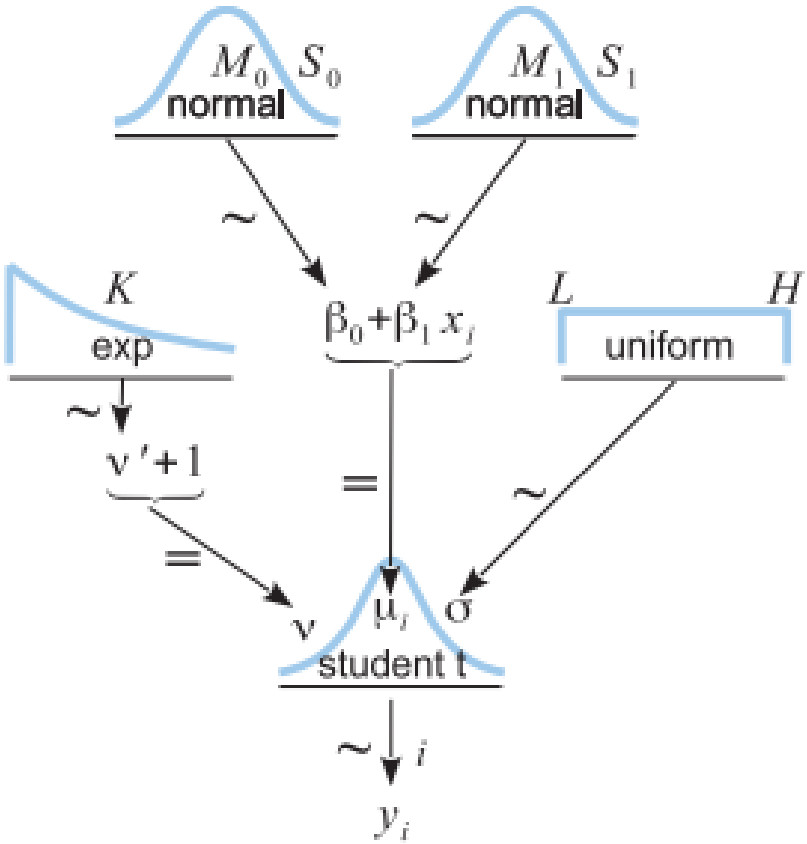
Homogeneity: From *homogeneous* + suffix *-ity*



Robust Linear Regression (I)

We can often rely on noise being approx. normal. However, specific processes (e.g. transcription errors) can significantly increase probability of outliers.

For Bayesian we can (easily) use other distributions. E.g. Student's t-distribution with heavy tails to resist outliers (w.r.t. the normal distribution).

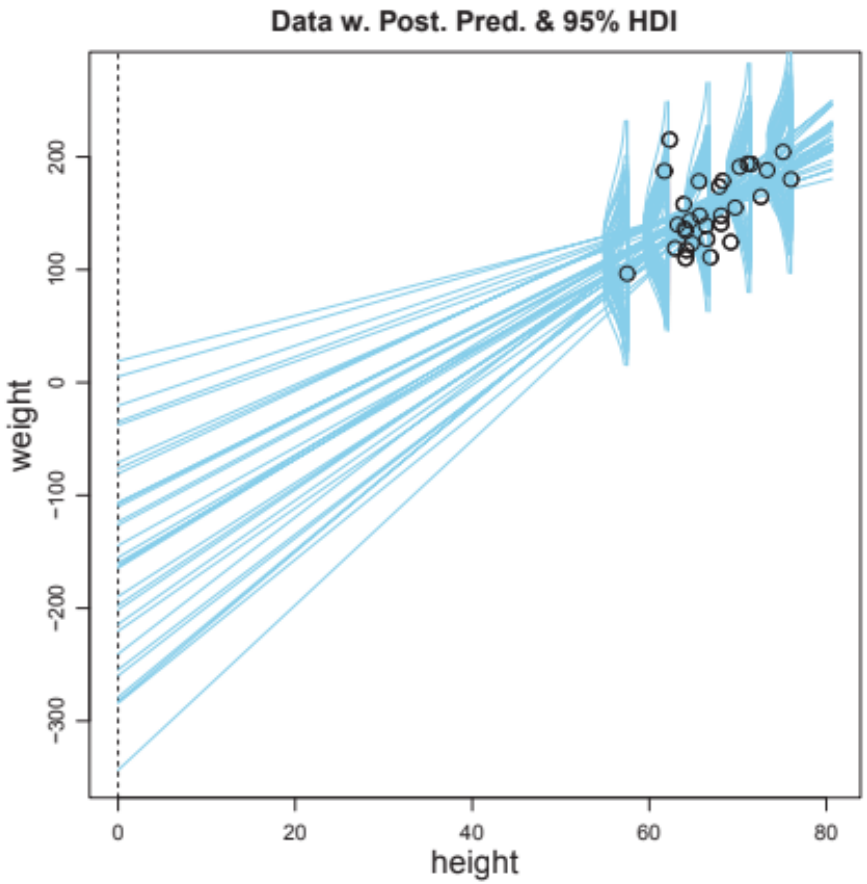


Standardisation (I)

Use *standardisation* to massasge data to better fit MCMC generators. Normalises data to be zero-centered and have unit variance. (Weak version of decorrelation, but computationally friendly?)

Standardisation can be defined for linear, quadratic, etc. trends.

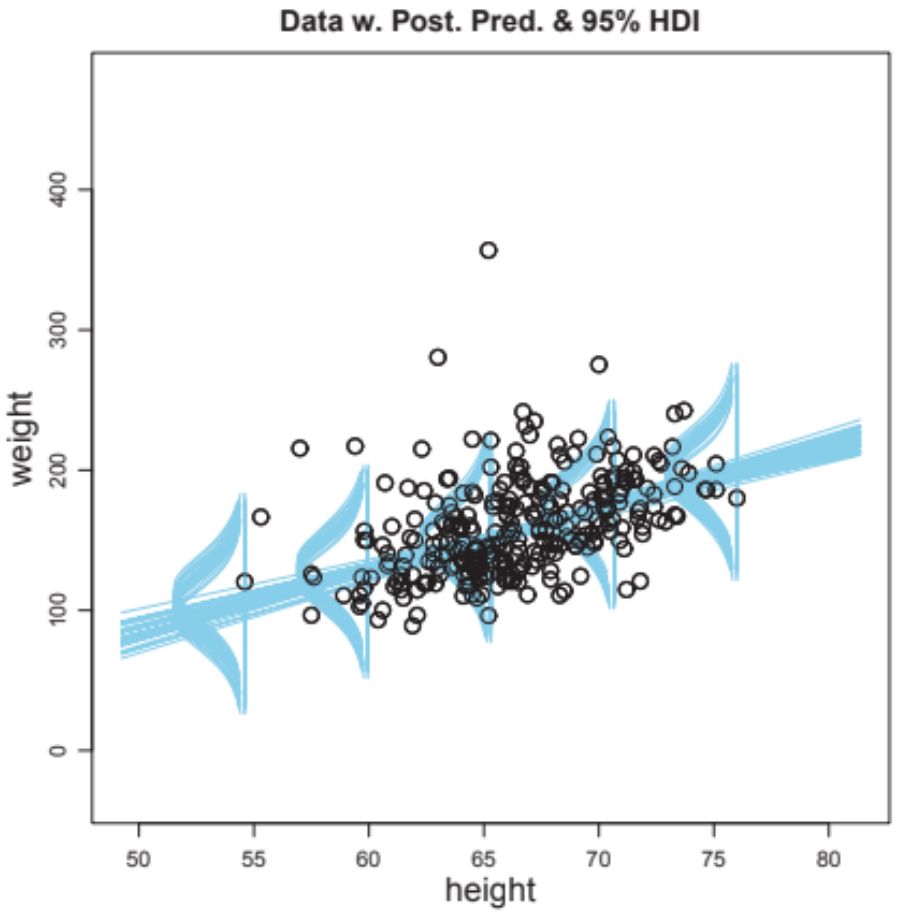
Note: Use only traning data to define your transformation (and inverse).



Standardisation (II)

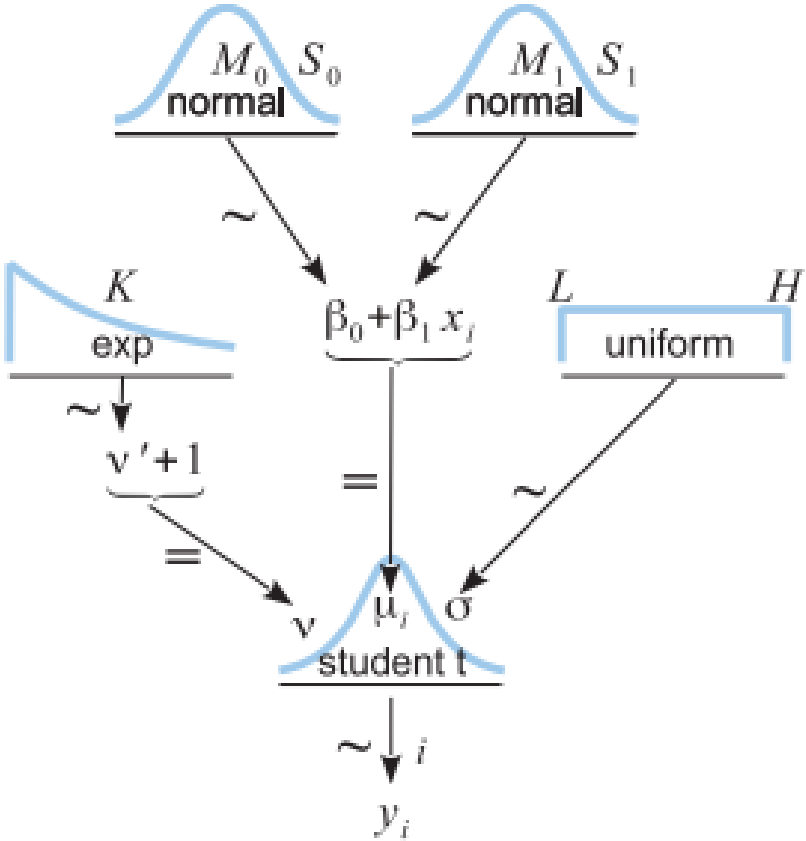
When using standardised data there is less correlation between slope and offset. (Shown here by the crop.)

Additionally, more data moves the estimation farther from the prior. (Not shown: normality paramter smaller for this case, i.e. less normal.)

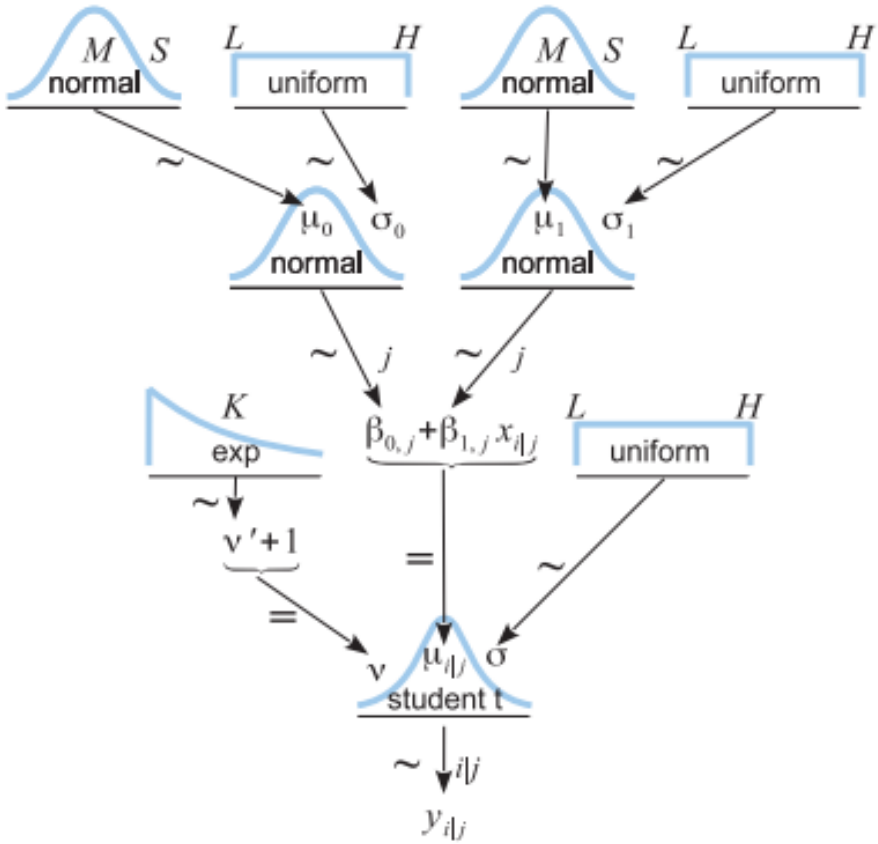


Structured Noise (I)

Old model



New model

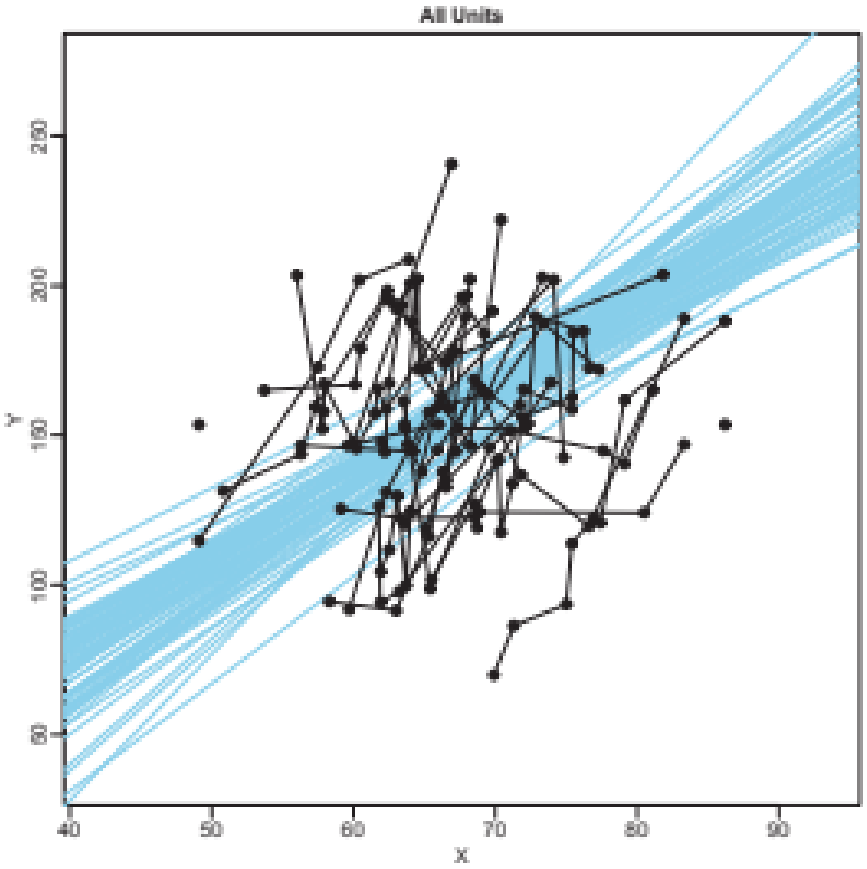


Structured Noise (II)

Providing an hierarchical model can significantly change the modelling.

Example: Without grouping, no trend can be seen. Grouping by individual makes clear that there is significant trend.

Note: Two individuals have only a single sample. Despite this, they have different predictions (the data point pull from group mean).



Extensions etc.

If there is significant mismodelling (check with posterior predicitve distribution), one can extend the model (also applies for initial modelling).

Consider:

- Trends: linear, quadratic, sinusodial, etc.
- Hierarchical models.
- Per-individual parameters (req. more data)
- Robust distributions (t-distribution in higher level paramters).
- Weighting of data (book exemplifies weighting of data credibility).

Note: Careful with extending to more paramters, this expands paramter space and can result in less precise estimations.

Recap of where we are

So far: Modelled a central tendency with added noise.

$$y = \beta_0 + \varepsilon \implies y \sim \mathcal{N}(\beta_0 + \mu_\varepsilon, \sigma_\varepsilon) \sim \mathcal{N}(\mu, \sigma)$$

Now: Expand structure in central tendency estimation

$$y = \beta_0 + \beta_1 x + \varepsilon \implies y \sim \mathcal{N}(\beta_0 + \beta_1 x + \mu_\varepsilon, \sigma_\varepsilon) \sim \mathcal{N}(\mu = c_0 + \beta_1 x, \sigma)$$

Improve model: Posterior predictive check. Model: trends, group-level parameters, robust distributions, weight data.

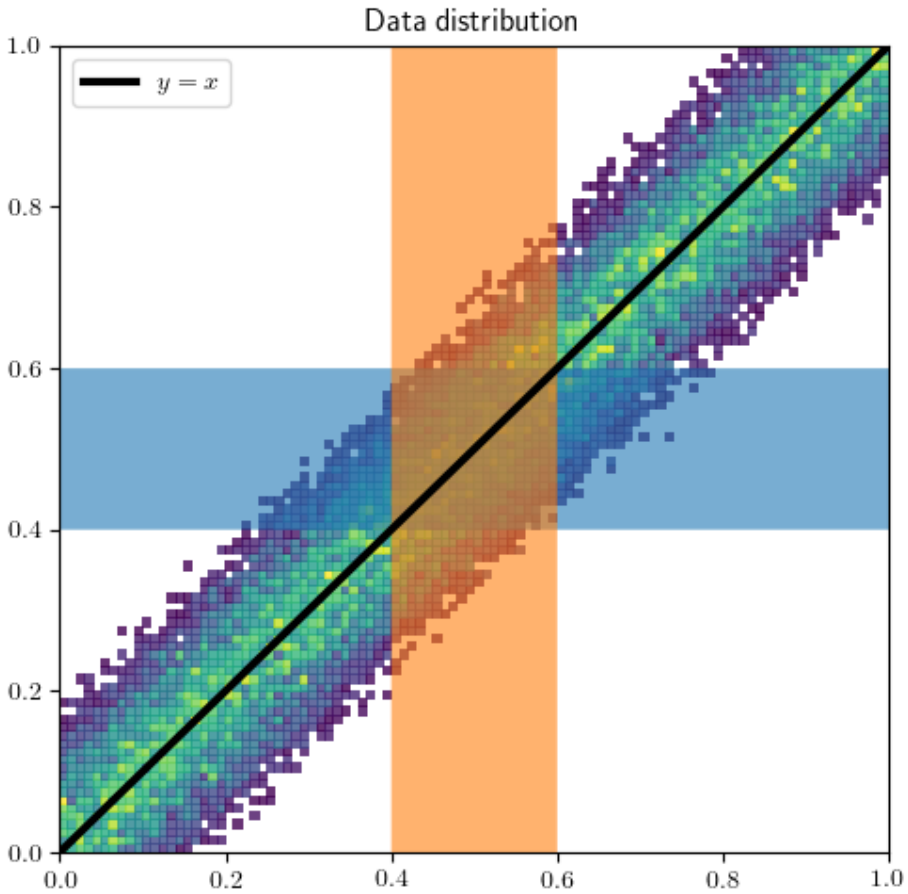
Chapter 17

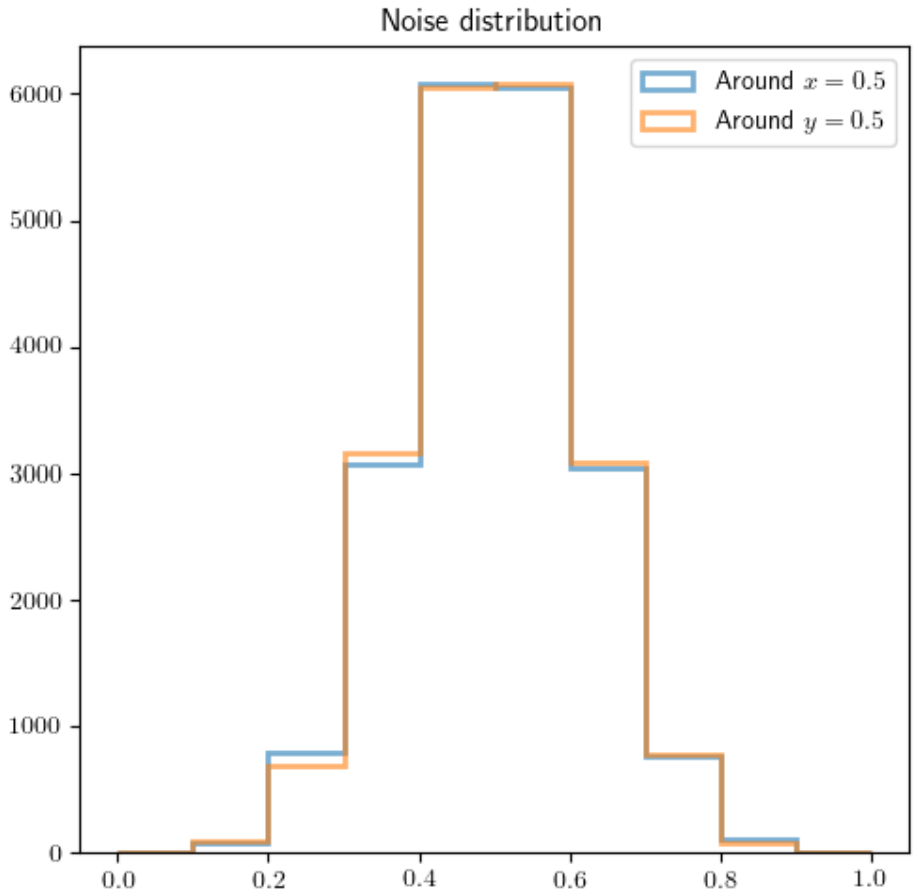
End

Appendix below

Aside: Why assume all noise lies in y?

Note: We assume all measurement noise lies in y. Possible since perfect meas. of y and imperfect meas. in x can be interpreted as the inverse (under additive normal noise atleast). Below shown for $x = y + \varepsilon$.





Aside: Link function

Note: Incomplete discussion.
Suppose $y \sim \mathcal{N}(\mu, \sigma)$**:** Modelling the pdf is non-polynomial.
Idea: We can apply a transformation to y to simplify regression.
Introduce *link function*:

$$f(y) = v = \text{lin}_\beta(x)$$
$$y = f^{-1}(v) = f^{-1}(\text{lin}_\beta(x))$$

If v is approximately polynomial regression will be simplified, e.g. apply non-linear least squares.
For logistic inverse link function f^{-1} the link function f becomes:

$$f^{-1}(v) = \text{logistic}(v)$$
$$f(y) = \log \frac{y}{1-y}$$

Aside: Link function

If one considers ν as the variable to do regression on, it makes sense that f is the "primary" direction.

$$x \overset{\text{lin}}{\rightarrow} \nu \overset{f^{-1}}{\rightarrow} y$$

$$x \overset{\text{lin}^{-1}}{\leftarrow} \nu \overset{f}{\leftarrow} y$$

Aside: Link function

Worked example for normal y and f^{-1} logistic:

$$
\begin{aligned}
f(y) &= c_0 \exp -\frac{(x-\mu)^2}{2\sigma^2} \\
&= \log\left(\frac{c_0 \exp -\frac{(x-\mu)^2}{2\sigma^2}}{1 - c_0 \exp -\frac{(x-\mu)^2}{2\sigma^2}}\right) \\
&= \log\left(c_0 \exp -\frac{(x-\mu)^2}{2\sigma^2}\right) - \log\left(1 - c_0 \exp -\frac{(x-\mu)^2}{2\sigma^2}\right) \\
&\approx \log\left(c_0 \exp -\frac{(x-\mu)^2}{2\sigma^2}\right) \\
&\approx c_1 - \frac{(x-\mu)^2}{2\sigma^2}
\end{aligned}
$$

The key in the approximation is that $c_0 \exp -\frac{(x-\mu)^2}{2\sigma^2}$ has a range of $(0, c_0)$ which is $\mathcal{O}(1)$ for reasonable choices of σ.

Aside: Link function

Worked example shown visually (and qualitatively), regression by hand: