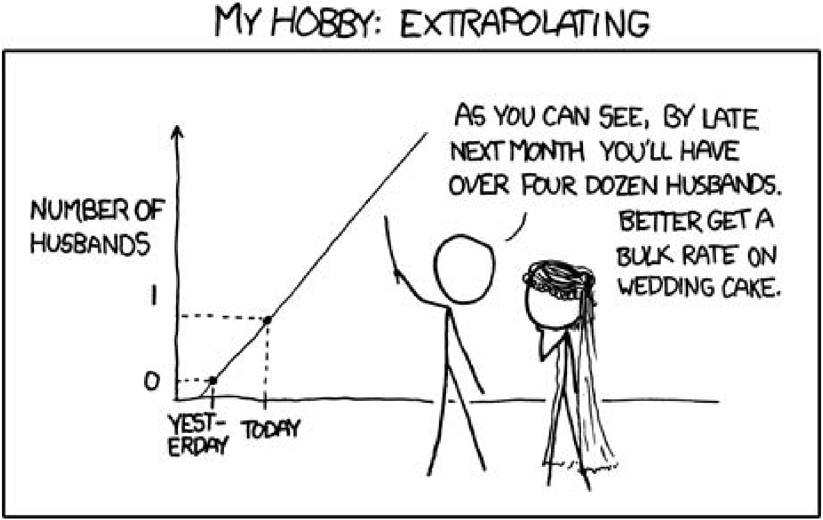

MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS.
BETTER GET A BULK RATE ON WEDDING CAKE.

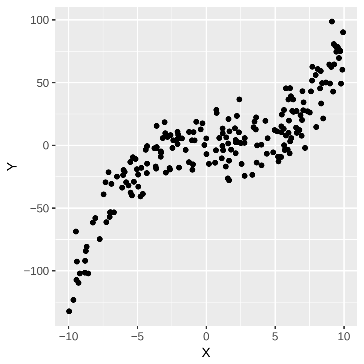
Linear regression

Erik Štrumbelj

2019

Example (Simple bivariate distribution)

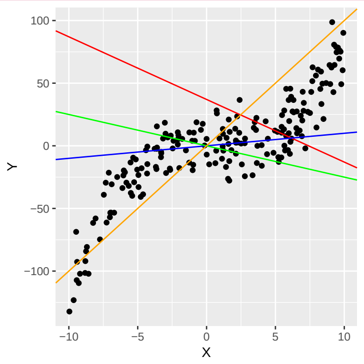
200 samples of (x,y) pairs:



We wish to model the relationship between x and y, with the purpose of predicting y for future observations of x.

Example (Modelling with a line)

There are infinitely many lines:



No line gives a perfect fit. Which one is the best?

Simple linear model

Model:

$$y_i = \beta_2 x_i + \beta_1 + \epsilon_i, \qquad \epsilon \sim_{\text{iid}} \mathcal{N}(0, \sigma^2), i = 1..n.$$

Or, equivalently:

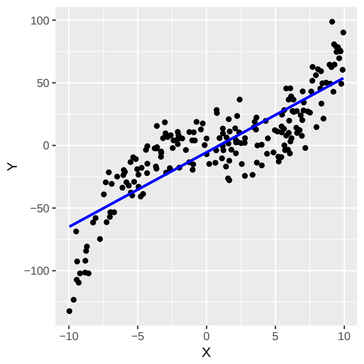
$$y_i \sim \mathcal{N}(\beta_2 x_i + \beta_1, \sigma^2).$$

Selecting the line that maximizes the likelihood:

$$p(y|x, \beta_1, \beta_2, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \beta_2 x_i - \beta_1)^2}{2\sigma^2}}$$

Example (Simple linear model)

Maximum likelihood estimates: $\beta_2 = 5.96, \beta_1 = -5.49, \sigma = 21.65$:



Linear regression

Simple generalization to k:

$$y_i = \beta_k x_{i,k} + ... + \beta_2 x_{i,2} + \beta_1 x_{i,1} + \epsilon_i, \qquad \epsilon \sim_{\text{iid}} \mathcal{N}(0, \sigma^2), i = 1..n.$$

Or, equivalently:

$y_i \sim \mathcal{N}(\beta^T x_i, \sigma^2)$ or $y \sim \mathcal{N}_n(X\beta, \sigma^2 \mathbb{I})$.

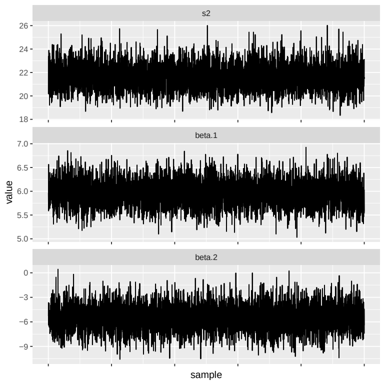
The likelihood

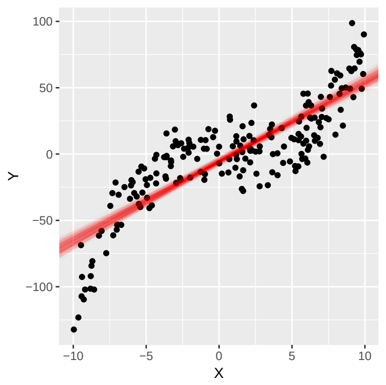
$$p(y|X, \beta, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{\sum_{i=1}^n (y_i - \beta^T x_i)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{SSR(\beta)}{2\sigma^2}}.$$

is maximized by $\beta^* = (X^T X)^{-1} X^T y$ (Ordinar Least Squares).

Example (Bayesian linear regression)

Gibbs sampling ($m = 5000$; $\beta_0 = 0, \Sigma_0 = 2500\mathbb{I}, a_0 = 1,\ b_0 = 20$):



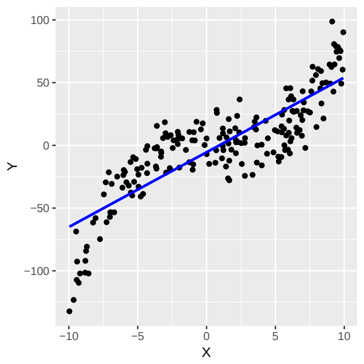


All $m_{eff} > 4000$, $se_{MCMC} < 0.05$.

Bayesian posterior means: $\beta_1 = 5.96, \beta_0 = -5.49, \sigma = 21.63$

Maximum likelihood estimates: $\beta_1 = 5.96, \beta_0 = -5.49, \sigma = 21.65$

Example (Simple linear regression)

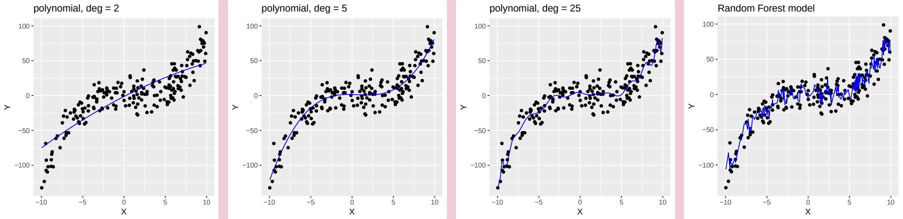


Any suggestions for a more appropriate model?

Example (Non-linearity via input space transformation)

This is still linear regression:

$$y_i = \beta_k x_i^k + \dots + \beta_2 x_i^2 + \beta_1 x_i + \beta_0 + \epsilon_i$$



Which model is the best?

Model selection

The goal is to select a model that will best generalize to new (unseen) data from the problem.

Two key components:

- Error (quality) criterion.
- Error estimation procedure.

Error criteria

In general, the log-score (log-likelihood) is a good choice:

$$LOG = \log\left(\overbrace{\prod_{i=1}^{n} p_{model}(y_i|y)}^{\text{likelihood}}\right) = \sum_{i=1}^{n} \log p_{model}(y_i|y).$$

Often, especially in machine learning, Mean Squared Error (MSE) is used:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_{\text{model}} - y_i)^2.$$

Other: accuracy, precision, recall, rank probability score, Brier score,...

Procedure for error estimation

Estimating error on the dataset used for fitting the model will result in biased (optimistic) estimates.

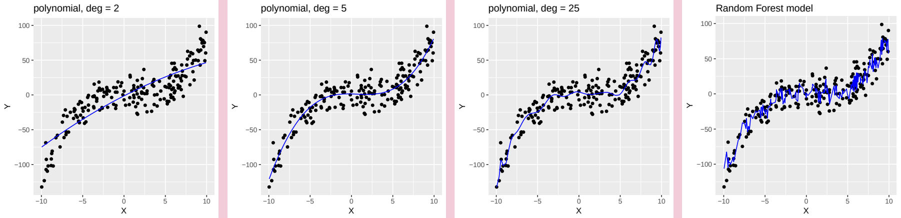
Alternative 1 (separate evaluation dataset): Reserve a random subset for evaluation and use the rest for fitting the model.

Alternative 2 (cross-validation):

- Partition the data into k (approximately) equal parts: y_1, y_2, ..., y_k.
- Fit k models $M_i, i = 1..k$, each without the i-th partition.
- Estimate error with $\overline{err} \approx \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} err_{M_i}(y_{i,j})$.

Special case $k = n$: leave-one-out cross-validation (LOOCV).

Example (Model evaluation)



In-sample and out-of-sample estimated (LOOCV) MSE:

Model	in-sample	out-of-sample
linear	464.1	475.8
2nd	450.0	468.7
3rd	200.1	208.6
5th	199.8	213.1
25th	180.5	12859.4
RF	299.1	296.9

Overfitting

Fitting to the noise in the data and not the data generating process.

Consequence: incorrect inference, poor generalization and predictions for new/unseen data.

A proper procedure of estimating the models error will detect overfitting and there exist several approaches to preventing overfitting:

- feature selection,
- early stopping,
- tree pruning,
- **regularization**,
- ...

Example (Bear weight)

AGE	SEX	HEAD_L	HEAD_W	NECK	LENGTH	CHEST	**WEIGHT**
19	1	11	5.5	16	53	26	80
55	1	16.5	9	28	67.5	45	344
81	1	15.5	8	31	72	54	416
115	1	17	10	31.5	72	49	348
104	2	15.5	6.5	22	62	35	166
100	2	13	7	21	70	41	220
56	1	15	7.5	26.5	73.5	41	262
...							

Random sample of 54 bears.

Task:

- Predict bear weight from other variables: .