Please scan this code to register your presence on Ilias for contact-tracing purposes. Your Ilias identity will be used. – Only do this if you are physically present!

Data Literacy
Lecture 06: Linear Regression

Jakob Macke

Eberhard Karls Universität Tübingen
Faculty of Science
Department of Computer Science
Machine Learning in Science

November 29, 2021

Please register here if you want to participate in the course evaluation.

Plan for today

Linear regression: why it matters

Linear regression: The ‘Machine Learning’ view

Linear regression revisited: Estimation in linear Gaussian models

Linear regression: The Bayesian view

Summary

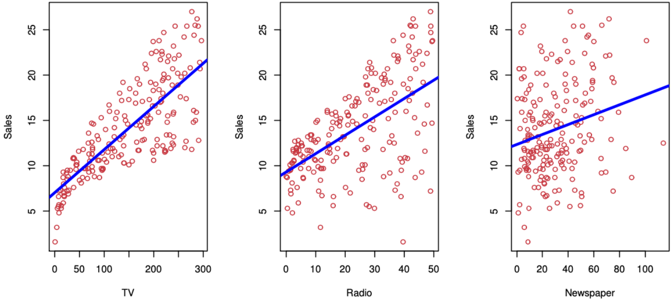
Regression

- *Regression analysis* aims to estimate the relationship between an *independent variable* x and a *dependent variable* t
- Often, we say we want to predict t from x.
- We want to find a function f with parameters ω such that

$$t \approx f(x, \omega), \tag{1}$$

  where x is (typically) an M-dimensional vector and t a scalar.
- Jargon:
  x is the *predictor*, *covariate*, *explanatory variable*, or *feature*.
  t is the *target*, *outcome*, or *response*.
- Often, we dont want to just *predict* t, but also *interpret* the relationship between x and t.

Example: Predicting sales from advertising



Source: James, Witten, Hastie, Tibshirani 2017

- Is there a relationship between advertising budget and sales?
- How strong is this relationship?
- Advertising in which media are most strongly associated with sales?
- Is the relationship linear?
- Are there synergies (or redundancies) among the different media?

Many other examples...

- How will Covid-Case Counts develop in the near future?
  (predict future counts from past counts)
- What vote-share will parties get at the next election?
  (predict votes from surveys, economic data)
- How will neural activity in the brain change when a particular image is shown (and which parts of the image are most important)?
- How well will students do in an exam?
- ...

Linear regression: The parent of all regression models

- Given data data $D = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_N, t_N)\}$, we seek a linear model

$$\begin{aligned} t_n &= \omega_1 x_n^{(1)} + \ldots + \omega_N x_n^{(N)} + \varepsilon_n && (2) \\ &= \omega^\top x_n + \varepsilon_n, && (3) \end{aligned}$$

  where the ε_n are error terms (or 'residuals').
- This is sometimes written in matrix form,

$$\mathbf{t} = X\omega + \varepsilon, \qquad (4)$$

  where $\mathbf{t}$ is an $N \times 1$ vector, X is the $N \times M$ matrix of covariates (the *design matrix*), and ε is a vector of residuals.

Why is linear regression so important?

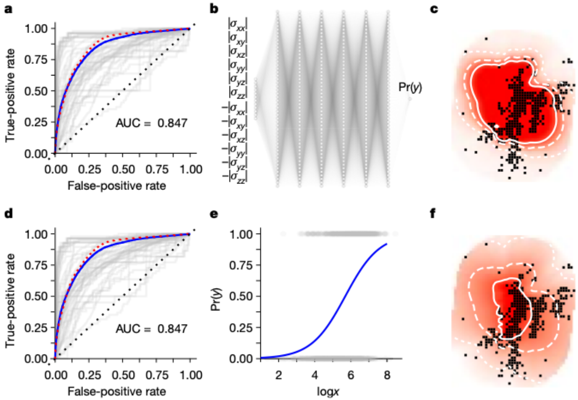
- The simplest regression model: (Almost) always try it first!
- Statistical properties of linear regression models (e.g. hypothesis tests) extremely well studied.
- Linear models are are easier to interpret (but even that is not always trivial, e.g. if covariates are correlated).
- It provides a building block for more complex, nonlinear models.

MATTERS ARISING

https://doi.org/10.1038/s41586-019-1582-8

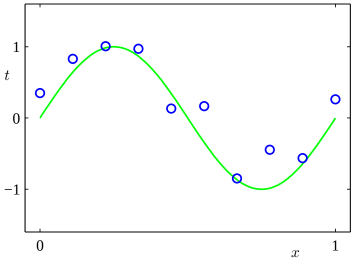
One neuron versus deep learning in aftershock prediction

Arnaud Mignan[1,2,3]* & Marco Broccardo[2,4]*



Linear regression: The ‘Machine Learning’ view

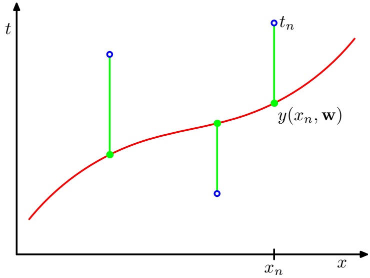
Using linear regression for nonlinear prediction problems



Source: Bishop PRML, Fig. 1.2.

$$
\begin{aligned}
y(x, \omega) &= \omega_0 + \omega_1 x + \omega_2 x^2 + \cdots + \omega_M x^M && (5)\\
&= \sum_{i=0}^{M} \omega_i x^i && (6)\\
&:= \omega^\top \phi(x) \qquad\qquad := \omega^\top z && (7)
\end{aligned}
$$

How to find ω? One idea: Minimize sum of squared errors



Source: Bishop PRML, Fig. 1.3.

$$E(\omega) = \frac{1}{2} \sum_{n=1}^{N} (y(x_n, \omega) - t_n)^2 \tag{8}$$

Minimum squared error solution for linear regression

$$
\begin{aligned}
E(\omega) &= \frac{1}{2}\sum_{n=1}^{N}(\omega^\top z_n - t_n)^2 && (9)\\
&= \frac{1}{2}\sum_{n=1}^{N}(\omega^\top z_n - t_n)(\omega^\top z_n - t_n)^\top && (10)\\
&= \frac{1}{2}\sum_{n=1}^{N}(\omega^\top z_n - t_n)(z_n^\top \omega - t_n) && (11)\\
&= \frac{1}{2}\sum_{n=1}^{N}\omega^\top z_n z_n^\top w - 2t_n\omega^\top z_n + t_n^2 && (12)\\
&= \omega^\top\left(\frac{1}{2}\sum_n z_n z_n^\top\right)\omega - \omega^\top\left(\sum_n t_n z_n\right) + \sum t_n^2 && (13)\\
&= \omega^\top \qquad A \qquad \omega + \omega^\top \qquad b \qquad + c && (14)
\end{aligned}
$$

Minimum squared error solution for linear regression

$$E(\omega) = \omega^\top A\omega + \omega^\top b + c \tag{15}$$

$$\Rightarrow \nabla_\omega E(\omega) = 2\omega^\top A + b^\top \tag{16}$$

$$= \omega^\top \left(\sum_n z_n z_n^\top\right) - \sum_n t_n z_n^\top \tag{17}$$

Setting the gradient to 0, we get the ω which minimizes the sum of squared errors,

$$\omega_{MSE} = \left(\sum_n z_n z_n^\top\right)^{-1} \sum_n t_n z_n \tag{18}$$

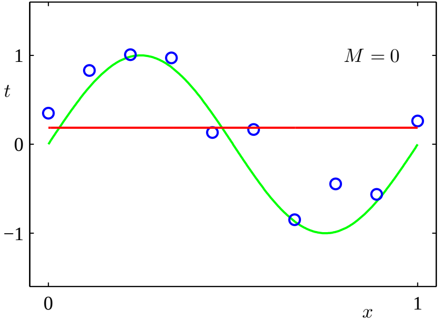
Note: If the z's have 0 mean, then

$$\mathrm{Cov}(z) = \frac{1}{N}\sum_n z_n z_n^\top \tag{19}$$

$$\mathrm{Cov}(z,t) = \frac{1}{N}\sum_n t_n z_n \tag{20}$$

Ok, but which *M* should we use?

M=0

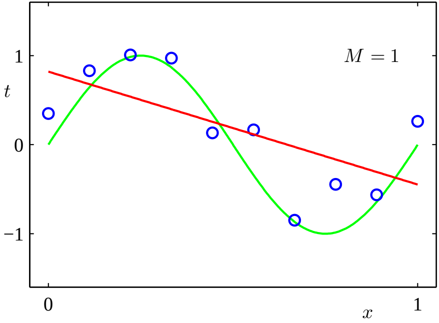


Source: Bishop PRML Fig. 1.4.

$$y = \omega_o \tag{21}$$

Ok, but which *M* should we use?

M=1

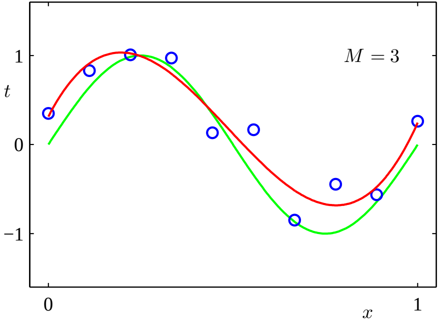


Source: Bishop PRML Fig. 1.4.

$$y = \omega_o + \omega_1 x \tag{22}$$

Ok, but which *M* should we use?

M=3

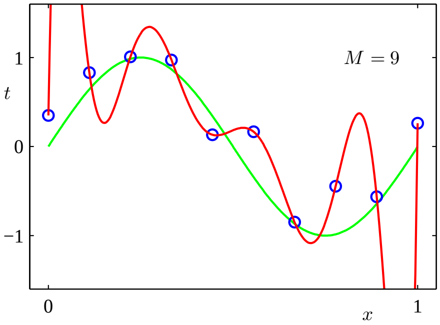


Source: Bishop PRML Fig. 1.4.

$$y = \omega_o + \omega_1 x + \omega_2 x^2 + \omega_3 x^3 \tag{23}$$

Ok, but which *M* should we use?

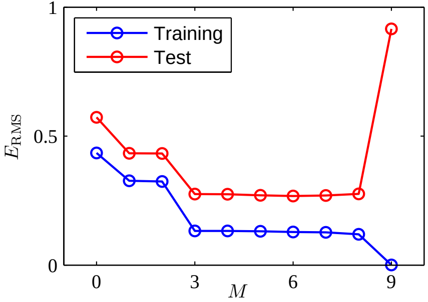
M=9



Source: Bishop PRML Fig. 1.4.

$$y = \omega_o + \omega_1 x + \omega_2 x^2 + \omega_3 x^3 \ldots + \omega_9 x^9 \tag{24}$$

Evaluating generalization performance: Use a separate test-set!



Source: Bishop PRML Fig. 1.5.

Note: The ‘optimal’ model will depend on *N*, i.e. the size of the data-set: With more data, we might favour more complex models.

Can we do better? Regularization

- Key idea: Penalize large parameters
- Add penalty on norm of ω to the loss function:

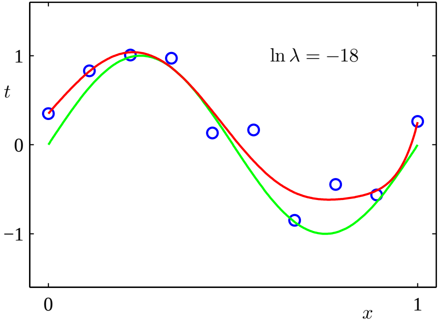
$$\tilde{E}(\omega|\lambda) = \frac{1}{2}\sum_{n=1}^{N}(y(x_n,\omega) - t_n)^2 + \frac{\lambda}{2}\|\omega\|^2 \tag{25}$$

- Solution:

$$\omega_{reg} = \left(\sum_{n=1}^{N} z_n z_n^\top + \lambda \mathbf{I}_M\right)^{-1} \sum_{n=1}^{N} z_n t_n \tag{26}$$

Can we do better? Regularization

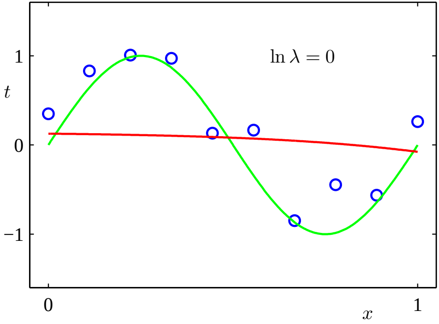
$\log \lambda = -18$



Source: Bishop PRML Fig. 1.7.

Can we do better? Regularization

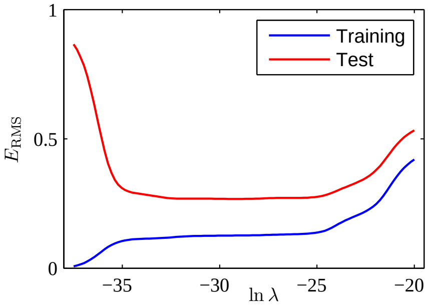
$\log \lambda = 0$



Origin: Bishop PRML Fig. 1.7.

Can we do better? Regularization

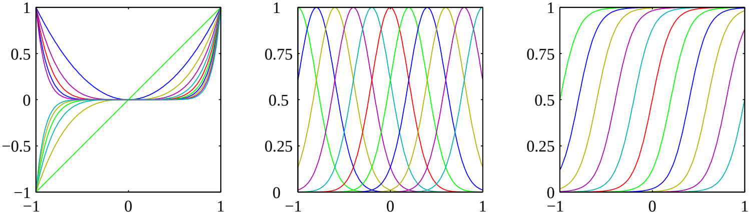
Cross-validation:



Source: Bishop PRML Fig. 1.8.

Linear regression with well-chosen basis functions ('features') can be a powerful modelling approach!.

- Basis functions $\phi(x)$ can model nonlinear relationships with $y(\omega, \mathbf{x}) = \omega^\top \phi(x)$.
- Polynomial regression: $\phi(x) = (1, x, x^2, x^3)$
- 'Gaussian bumps': $\phi_i(x) = \exp\left(-frac12(x - s_i)^2/\sigma_i^2\right)$
- Sigmoids $\phi_i(x) = 1/(1 + \exp(-x - s_i))$



Source: Bishop PRML Figures 3.1a-c

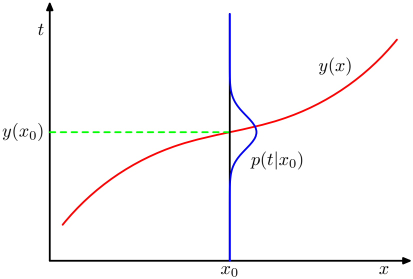
Linear regression revisited: Estimation in linear Gaussian models

Linear regression as MLE in a Gaussian model

- Suppose that we have data $D = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_N, t_N)\}$
- We model the data by $t_n \approx y(x, \omega) + \varepsilon$, where ε is additive Gaussian noise.
- We assume that noise is independent, identically distributed and Gaussian:

$$\varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{27}$$

$$t|\mathbf{x}, \omega, \sigma^2 \sim \mathcal{N}(y(\mathbf{x}, \omega), \sigma^2) \tag{28}$$



Source: Bishop PRML Figure 1.28

Linear regression as MLE in a Gaussian model

- We consider a linear model $y(\mathbf{x}, \omega) = \omega^\top \mathbf{x}$
- We want to infer ω by maximizing the log-likelihood, which is

$$\log P(D|\omega) = C - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y(\mathbf{x}_n, \omega) - t_n)^2 \tag{29}$$

- Thus, maximizing the likelihood for linear regression is the same as minimizing the sum of squared errors, and we get the MLE

$$\omega_{MLE} = \left(\sum_n \mathbf{x}_n \mathbf{x}_n^\top \right)^{-1} \sum_n \mathbf{x}_n t_n \tag{30}$$

Linear regression as MLE in a Gaussian model

$$y = \beta^\top x + \varepsilon_n \tag{31}$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d} \tag{32}$$

- The first entry of x is almost always a '1', and the corresponding parameter β_o is called the 'offset'.
- There is extensive, classical statistical theory for linear Gaussian models.
- Hypothesis tests: e.g. can one reject the null-hypothesis of all $\beta' s$ being 0?
- Confidence intervals: How well can we constrain each 'effect' β?
- Interpretation of β's: What does it 'mean' that a particular value of β is big?
- Are the assumptions of the model valid? Is the relationship linear? Are the errors independent, Gaussian, and all have the same variance?

Checking model assumptions in linear regresssion

- Are the residuals Gaussian? Plot histogram of residuals, check for Gaussianity (either by visual inspection or by running a statistical test). In particular, check for any 'outliers', i.e. values which seem overly small or big– these will likely have a very big influence on the estimated fit!
- Are the residuals uncorrelated? For data which is 'ordered', e.g. time-series data, calculate the correlation between adjacent residuals, to check whether they are uncorrelated.
- Do the residuals have a constant variance? (Jargon: constant variance = 'homeoscedastic', non-constant variance = 'heteroscedastic'. Plot the fitted values (y) against the residuals, and check whether there is a correlation.

Linear regression: The Bayesian view

Maximum-a-posteriori (MAP) solution for linear regression:

- We use a multivariate Gaussian as prior for ω:

$$\omega_i \sim \mathcal{N}(0, \alpha^{-1}) \tag{33}$$

$$p(\omega|\alpha) = \left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\left(-\frac{\alpha}{2}\omega^\top \omega\right) \tag{34}$$

- Finding the maximum-a-posteriori of ω:

$$\omega_{MAP} = \left(\sum_{n=1}^{N} x_n x_n^\top + \frac{\alpha}{\gamma}\mathbf{I}_M\right)^{-1} \left(\sum_{n=1} x_n t_n\right), \tag{35}$$

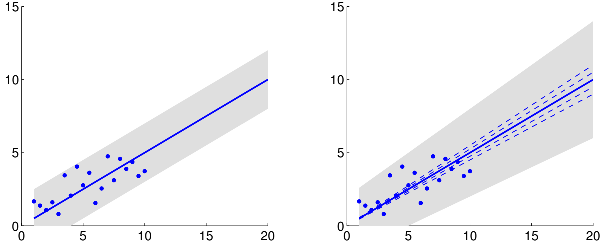
where $\gamma = 1/\sigma^2$.

In a linear Gaussian model, we can calculate the full posterior distribution in closed form (*)

$$p(\omega|D,\alpha,\gamma) = \mathcal{N}(\omega|\mu_{\text{post}},\Sigma_{\text{post}}), \text{ where} \tag{36}$$

$$\Sigma_{\text{post}}^{-1} = \alpha\mathbf{I} + \gamma\sum_n x_n x_n^\top \text{ and} \tag{37}$$

$$\mu_{post} = \gamma\Sigma_{post}\sum_n x_n t_n. \tag{38}$$



(*) Assuming that we fix α and γ– if we do not know them have to use MCMC or variational approximations.

Summary: Linear regression

- (Almost) always start with a linear regression!
- On nonlinear problems, use basis functions.
- ‘Machine learning view’: Minimize MSE, use regularization/cross-validation to control model complexity.
- ‘Statistical view’: Linear regression = Linear Gaussian model
- Can be used for hypothesis testing (e.g. ‘which, if any, parameters are significantly away from zero’).
- Care is needed when interpreting parameters!
- Bayesian linear regression: Model uncertainty over parameters. Closed form solutions (if we know prior and noise variances).
- Next week: Regression is just conditional density estimation.

Reading

- ‘Machine learning’ view of Linear Regression: Bishop PRML (“Pattern Recognition and Machine Learning” by Christopher Bishop, 2006) 3.1.1, 3.1.2, 3.1,4
- Statistical View of Linear Regression: “An introduction to statistical learning” by James, Witten, Hastie, Tibshirani, 2017, Chapter 3.
- Bayesian linear regression: Bishop PRML 3.3.1-3.3.2, 3.4-3.6 (advanced)