ECEN 689 Materials Informatics

Regression

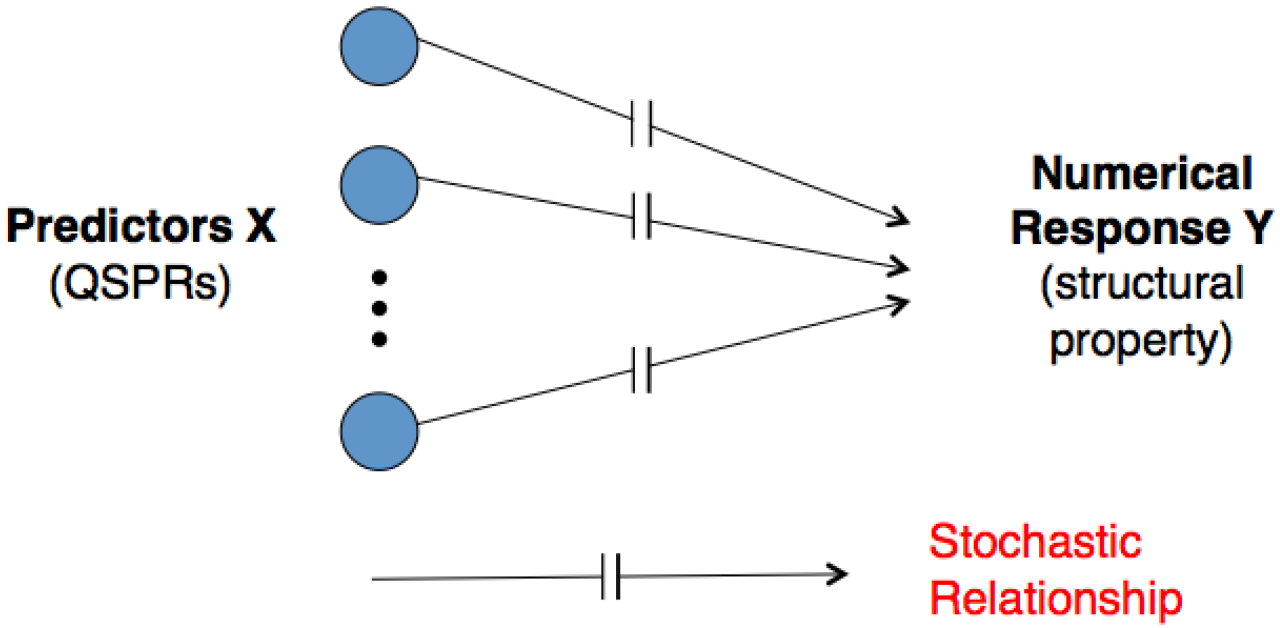
Ulisses Braga-Neto

ECE Department

Texas A&M University

Regression Problem

- We have a predictor vector $\mathbf{X} \in R^d$ (also called the *independent variable*) and a response $Y \in R$ (also called the *dependent variable*).
- Unlike in classification, the response Y is *numeric* and *nonfinite* (typically continuous-valued). For example, a real-valued energy measurement.



Regression Problem - II

- The stochastic relationship between $\mathbf{X}$ and Y is described by the joint probability density $p(\mathbf{x}, y)$.
- In some regression contexts, $\mathbf{X}$ is *not* random. Of great importance then is the *predictive density* $p(y\,|\,\mathbf{x})$.
- Regardless of whether $\mathbf{X}$ is random or not, one can always write the response Y at point $\mathbf{x}$ as:

$$Y = f(\mathbf{x}) + \varepsilon\,,$$

where ε is a zero-mean error term.

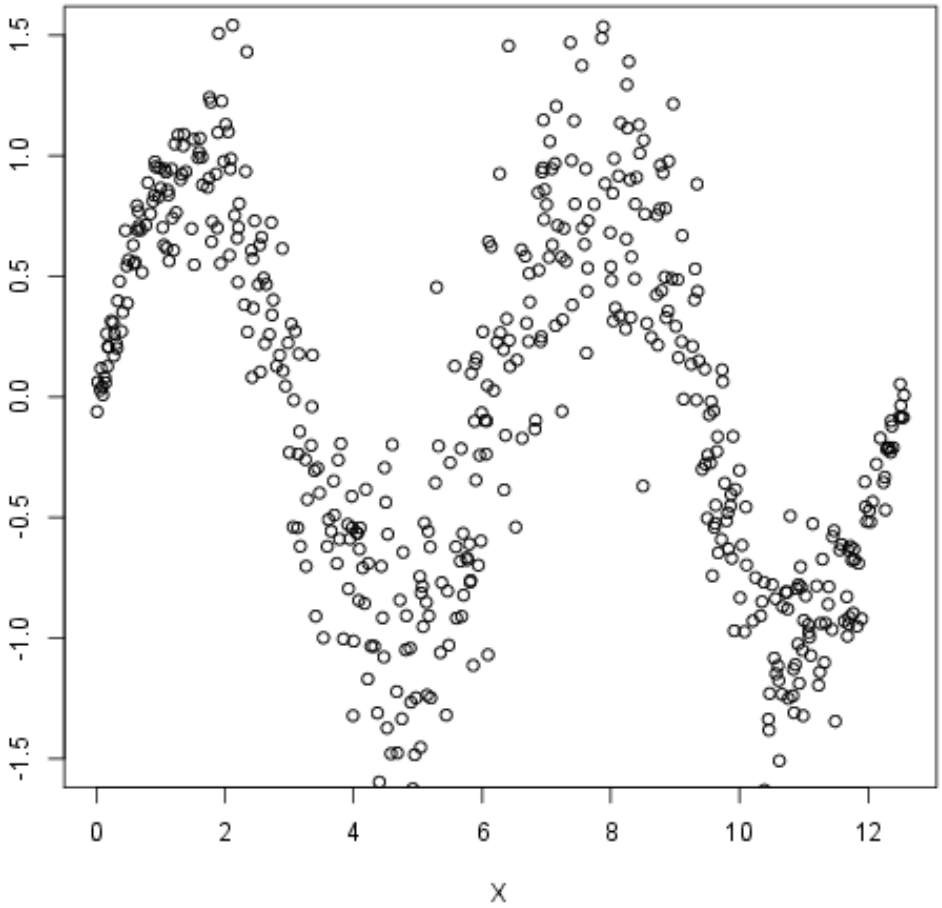
Proof: Let $f(\mathbf{x}) = E[Y \mid \mathbf{X} = \mathbf{x}]$, and define $\varepsilon = Y - E[Y \mid \mathbf{X} = \mathbf{x}]$. Then $Y = f(\mathbf{x}) + \varepsilon$ with

$$\begin{aligned} E[\varepsilon \mid \mathbf{X} = \mathbf{x}] &= E[Y - E[Y \mid \mathbf{X} = \mathbf{x}] \mid \mathbf{X} = \mathbf{x}] \\ &= E[Y \mid \mathbf{X} = \mathbf{x}] - E[Y \mid \mathbf{X} = \mathbf{x}] = 0\,. \end{aligned}$$

Q.E.D.

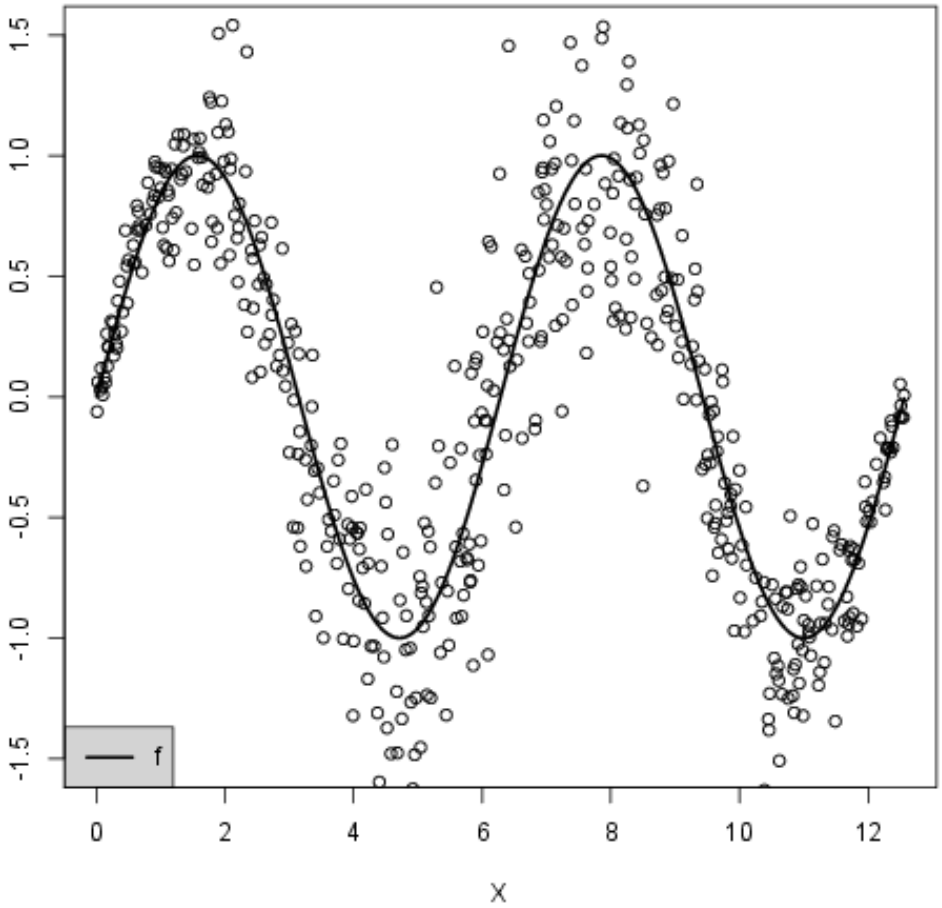
Graphical Representation

Example of noisy data, regression function, error bands.
Error is zero-mean, Gaussian, Std$(\varepsilon) = 0.05 + 0.02\pi x - 0.01x^2$



Graphical Representation

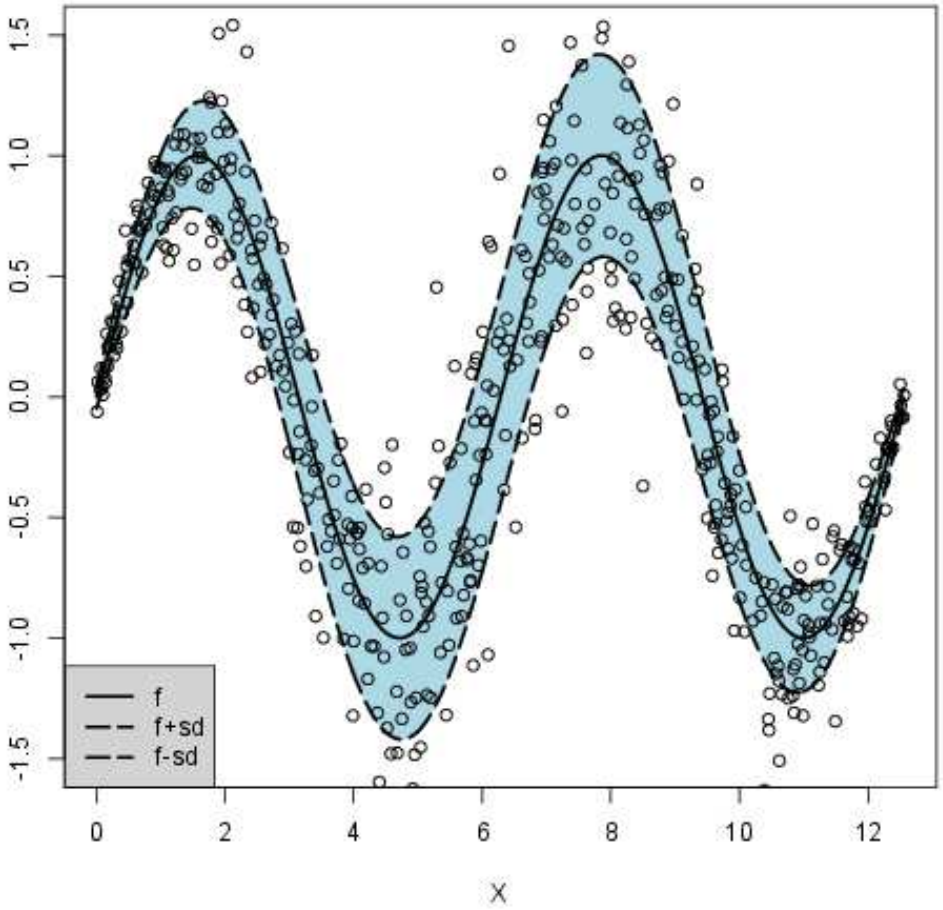
Example of noisy data, regression function, error bands.
Error is zero-mean, Gaussian, Std$(\varepsilon) = 0.05 + 0.02\pi x - 0.01x^2$



$$f(x) = \sin(x)$$

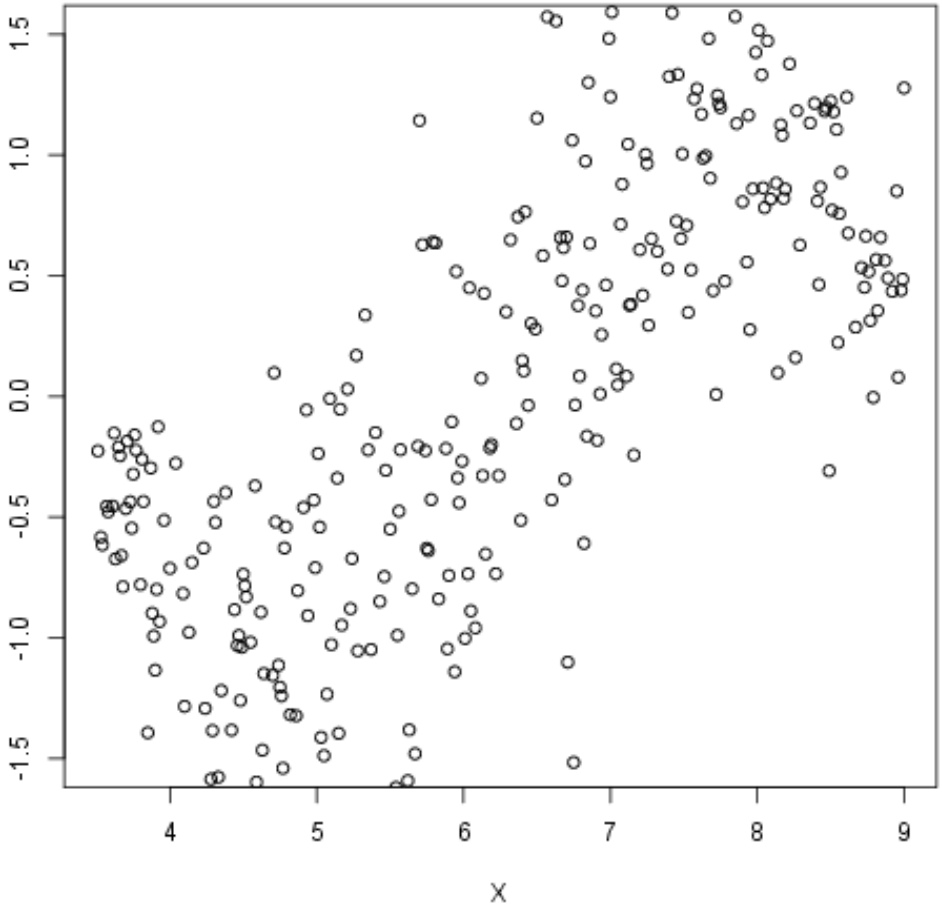
Graphical Representation

Example of noisy data, regression function, error bands.
Error is zero-mean, Gaussian, Std$(\varepsilon) = 0.05 + 0.02\pi x - 0.01x^2$



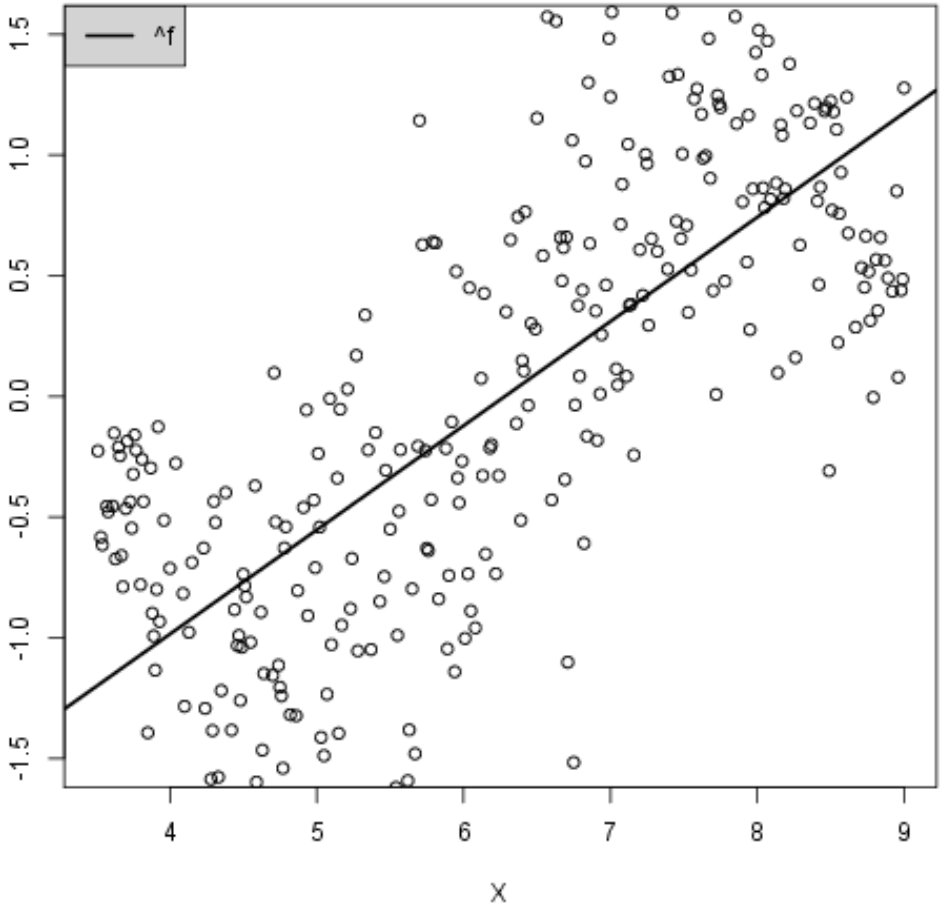
Regression Estimation

In practice, the function f is not known, and an estimate $\hat{f}$ must be estimated from the data.



Regression Estimation

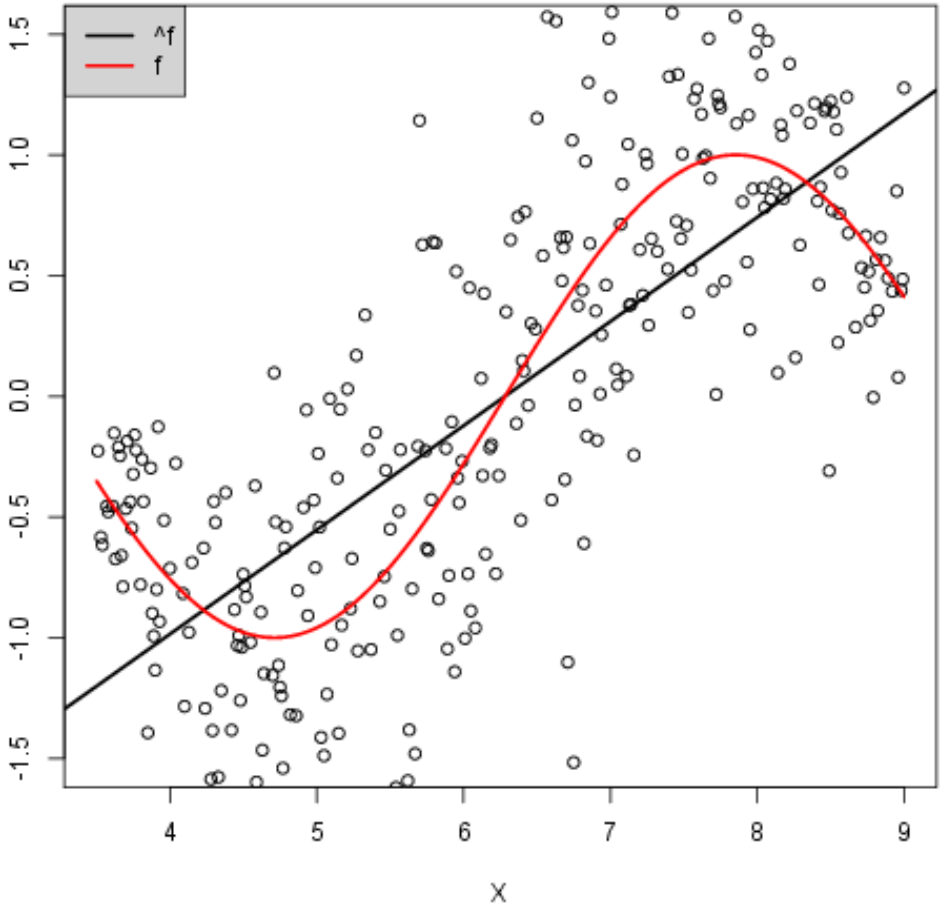
In practice, the function f is not known, and an estimate $\hat{f}$ must be estimated from the data.



Linear regression.

Regression Estimation

In practice, the function f is not known, and an estimate $\hat{f}$ must be estimated from the data.



The true regression is nonlinear. This is the best possible regression, but it is not very predictive itself.

Regression Error

- For a given regression estimate $\hat{f}$, the *conditional regression error* at a point $\mathbf{x}$ is given by:

$$L[\hat{f}](\mathbf{x}) = \int \ell(y, \hat{f}(\mathbf{x}))p(y \mid \mathbf{x})dy\,, \quad \mathbf{x} \in R^d$$

where $\ell : R \times R \to R$ is an appropriate *loss function*:

  - Quadratic loss: $\ell(y, \hat{f}(\mathbf{x})) = (y - \hat{f}(\mathbf{x}))^2$
  - Absolute loss: $\ell(y, \hat{f}(\mathbf{x})) = |y - \hat{f}(\mathbf{x})|$
  - Minkowski loss: $\ell(y, \hat{f}(\mathbf{x})) = |y - \hat{f}(\mathbf{x})|$, $q > 0$.

Optimal Regression

- Assuming that $\mathbf{X}$ is a random variable, then the *regression error* is the average error over R^d:

$$L[\hat{f}] = \int \left(\int \ell(y, \hat{f}(\mathbf{x})) p(y \mid \mathbf{x}) dy \right) p(\mathbf{x}) dx$$

$$= \int \ell(y, \hat{f}(\mathbf{x})) p(\mathbf{x}, y) dx dy = E[\ell(Y, \hat{f}(\mathbf{X}))] .$$

- The optimal regression *for a given loss function* is

$$f^* = \arg\min_{\hat{f} \in F} L[\hat{f}] = \arg\min_{\hat{f} \in F} \int \ell(y, \hat{f}(\mathbf{x})) p(\mathbf{x}, y) dx dy$$

where F is the class of admissible regression functions (e.g., all measurable functions).

Optimal Regression

- It can be shown that the optimal regression function for the quadratic loss is:

$$f^*(\mathbf{x}) \;=\; E[Y \mid \mathbf{X} = \mathbf{x}]\,, \;\; \mathbf{x} \in R^d$$

  i.e., the conditional mean.

- It can be shown likewise that the optimal regression function for the absolute loss is the *conditional median*, whereas the optimal regression function for the Minkowski loss with $q \to 0$ is the *conditional mode*.

- We will focus from this point on on the quadratic loss, in which case the regression error

$$L[\hat{f}] \;=\; E[\ell(Y, \hat{f}(\mathbf{X}))] \;=\; E[(Y - \hat{f}(\mathbf{X}))^2]$$

  is called the *mean square error* (MSE).

Reducible and Irreducible Error

- Using the decomposition $Y = f^*(\mathbf{X}) + \varepsilon$, where $f^*(\mathbf{x}) = E[Y \mid \mathbf{X} = \mathbf{x}]$ and ε is zero-mean, we can write

$$\begin{aligned} L[\hat{f}](\mathbf{x}) &= E[(f^*(X) + \varepsilon - \hat{f}(\mathbf{X}))^2 \mid \mathbf{X} = \mathbf{x}] \\ &= (f^*(\mathbf{x}) - \hat{f}(\mathbf{x}))^2 + E[\varepsilon^2 \mid \mathbf{X} = \mathbf{x}] \\ &= (f^*(\mathbf{x}) - \hat{f}(\mathbf{x}))^2 + \text{Var}(\varepsilon \mid \mathbf{X} = \mathbf{x}) \\ &= \text{reducible part} \quad + \quad \text{irreducible part} \end{aligned}$$

- The reducible error $(f^*(\mathbf{x}) - \hat{f}(\mathbf{x}))^2$ can be made small by good algorithm design, but the irreducible part $\text{Var}(\varepsilon \mid \mathbf{X} = \mathbf{x})$ is intrinsic to the problem.

Homoskedasticity

- If ε is independent of $\mathbf{X}$, then

$$\text{Var}(\varepsilon \mid \mathbf{X} = \mathbf{x}) = \text{Var}(\varepsilon) = \sigma^2$$

and the problem is much simpler. This is called the *homoskedastic case*. Otherwise, we have the *heteroskedastic case*.

- Notice that the MSE of the optimal regression function (i.e., the *optimal MSE*) in the homoskedastic case is just

$$L[f^*] = E[(Y - f^*(\mathbf{X}))^2] = E[\varepsilon^2] = \sigma^2$$

This is a lower bound on the performance of any regression algorithm.

The Residual Sum of Squares

- In practice, the true MSE $L[\hat{f}]$ requires distributional kwnoledge and is not know. It has therefore to be estimated from the data.
- A straightforward to do this is to test the regression function on the training data.
- Given the training data $\{(\mathbf{x}_1, y1), \ldots, (\mathbf{x}_n, y_n)\}$ used to derive $\hat{f}$, we define

$$\mathrm{RSS} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2$$

  which is called the *residual sum of squares* (RSS).
- The RSS is the analog of the aparent (resubstitution) error in classification. It also tends to be optimistically biased, and more so for more "flexible" algorithms.

The R^2 Statistic

- The RSS is not normalized. A very popular alternative is the R^2 statistic, which is always between 0 and 1.
- If there is no predictor, it is natural to estimate Y with the mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. The RSS of this predictor is

$$\mathrm{TSS} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2$$

  where TSS stands for *total sum of squares*.
- The R^2 statistic is the relative improvement in RSS by having predictor X over having no predictors:

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

- This is also called the *coefficient of determination.*

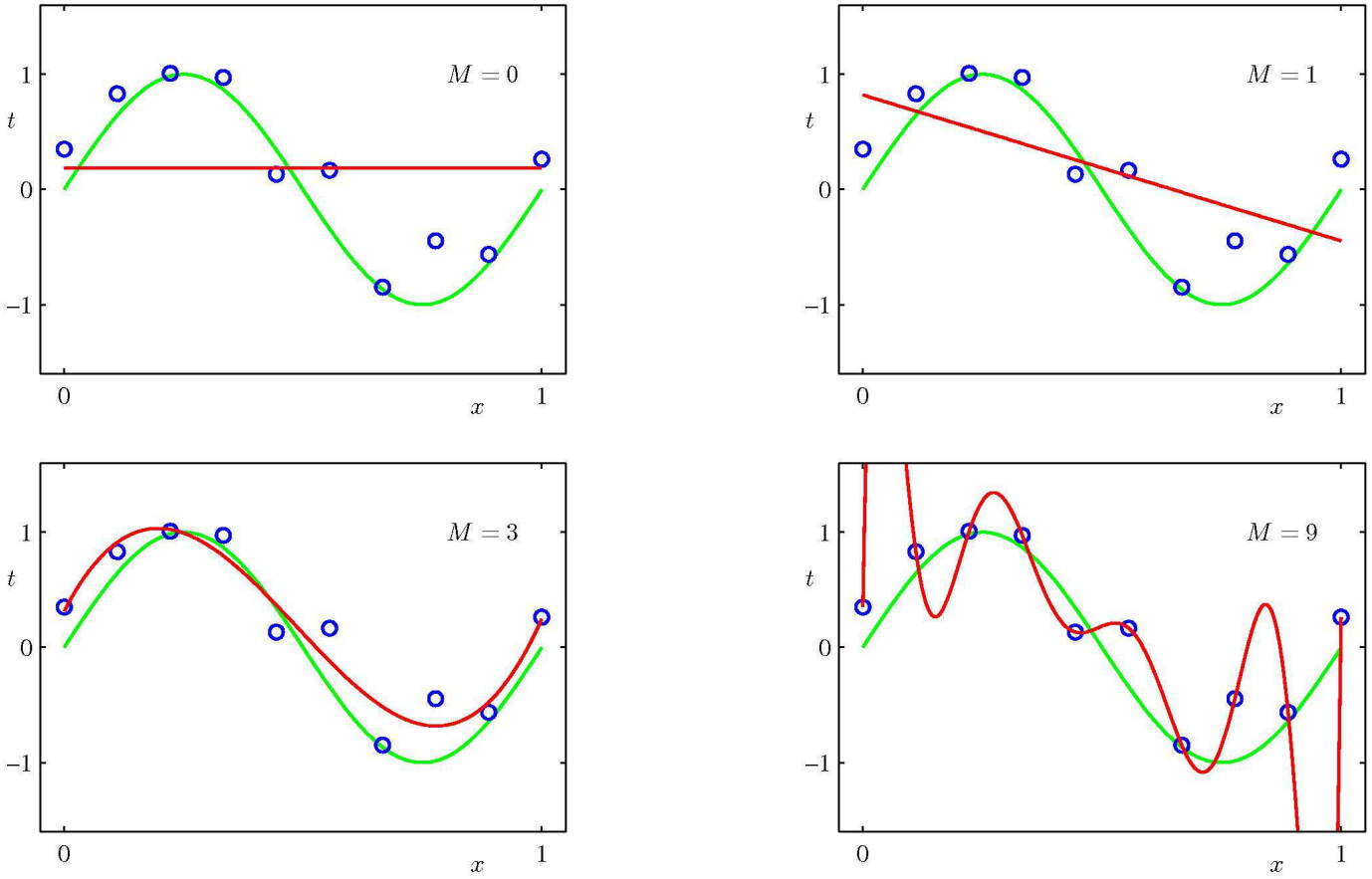
Bias-Variance Decomposition

- Let us consider now the estimation problem in more detail. Since the data is random, the regression estimator $\hat{f}$ is also random and therefore so is the MSE.
- The *expected MSE* for a regression algorithm, which is independent of the data, is just $E[L[\hat{f}]] = E[(Y - \hat{f}(\mathbf{X}))^2]$ (where $\hat{f}$ is no longer fixed, but changes with the data).
- In the homoskedastic case, after some algebra, we get:

$$\begin{aligned} E[L[\hat{f}]] &= E[(Y - \hat{f}(\mathbf{X}))^2] = E[(f^*(\mathbf{X}) + \varepsilon - \hat{f}(\mathbf{X}))^2] \\ &= E[(f^*(\mathbf{X}) - \hat{f}(\mathbf{X}))^2] + \mathrm{Var}(\varepsilon) \\ &= E[(f^*(\mathbf{X}) - \hat{f}(\mathbf{X}))]^2 + \mathrm{Var}(f^*(\mathbf{X}) - \hat{f}(\mathbf{X})) + \sigma^2 \\ &= \mathrm{Bias}(\hat{f}) + \mathrm{Variance}(\hat{f}) + \sigma^2 \end{aligned}$$

Bias-Variance Decomposition

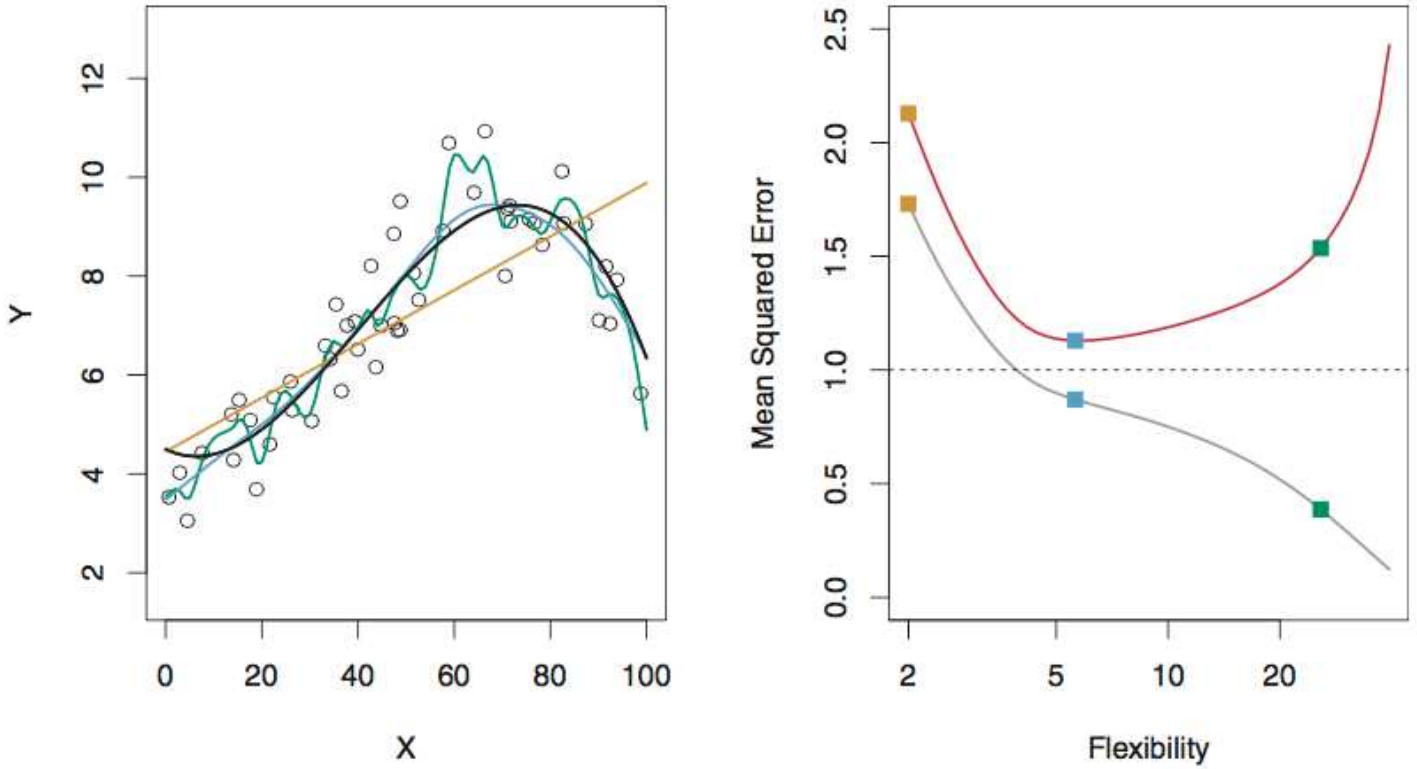
- Example: Bishop.



- Regression with a polynomial of order M. Bias decreases and variance increases with increasing M.

Bias-Variance Decomposition

- Example: James.



- Models with decreasing bias but increasing variance. True MSE (red) vs. RSS (gray). The true MSE combines bias and variance and displays a peaking phenomenon.

Linear Regression Model

- The most common form of parametric regression model is the linear model:

$$Y = a_0 + a_1 X_1 + \cdots + a_d X_d + \varepsilon \,,$$

where $X_1, \ldots, X_d$ are the predictors, ε an error term, and a_0 and $a = (a_1, \ldots, a_d)$ are the parameters.

- It can be generalized to a linear *basis function* model:

$$\begin{aligned} Y &= \theta_0 \phi_0(\mathbf{X}) + \theta_1 \phi_1(\mathbf{X}) + \cdots + \theta_k \phi_k(\mathbf{X}) + \varepsilon \\ &= \mathbf{\Phi}(\mathbf{X})^T \boldsymbol{\theta} + \varepsilon \,, \end{aligned}$$

where $\mathbf{X} = (X_1, \ldots, X_d)$ is the predictor vector, $\phi_i : R^d \to R$ are the basis functions, $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_k)$ is the *parameter vector* and $\mathbf{\Phi} = (\phi_1, \ldots, \phi_k)$.

Linear Regression Model

- For example, in the standard linear model, $\phi_0(\mathbf{X}) = 1$, $\phi_1(\mathbf{X}) = X_1, \ldots, \phi(\mathbf{X}) = X_d$, and $\theta_i = a_i$, for $i = 1, \ldots, d$.
- The key point is that the model be linear *in the parameters*.
- Another example is the (univariate) *polynomial regression* model:

$$Y = a_0 + a_1 X + a_2 X^2 + \cdots + a_k X^k + \epsilon \,,$$

  Here the basis functions are $\phi_0(X) = 1$, $\phi_1(X) = X$, $\phi_2(X) = X^2, \ldots, \phi_k(X) = X^k$, and $\theta_i = a_i$, for $i = 1, \ldots, k$.
- How to estimate the parameters from training data?

The Least-Squares Method

"The most probable value of the unknown quantities will be that in which *the sum of the squares of the differences* between the actually observed and the computed values multiplied by numbers that measure the degree of precision *is a minimum.*"

Karl Friedrich Gauss, *Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium*, 1809.

Matrix-Based Formulation

- Given the training data $S_n = \{(\mathbf{X}_1, Y_1), \ldots, \mathbf{X}^d\}$, write one equation for each data point:

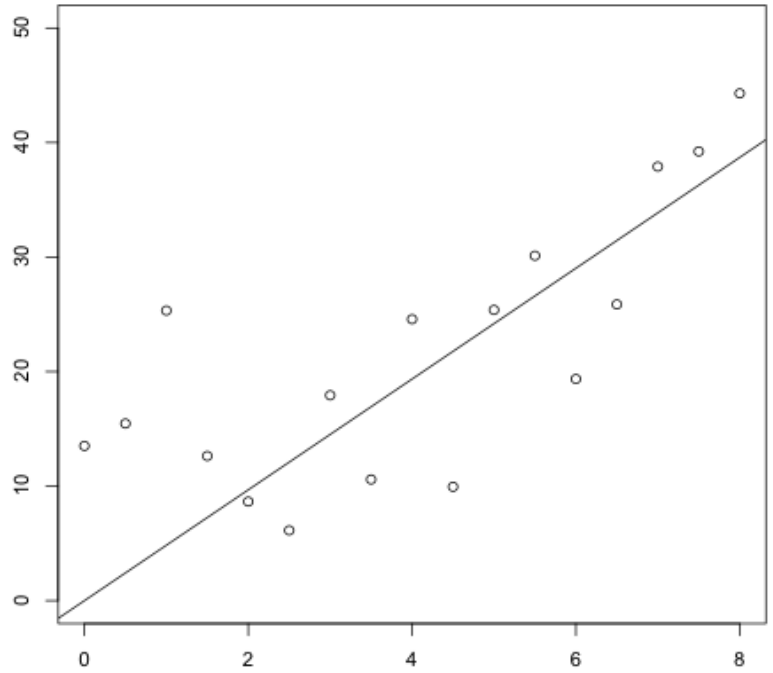
$$\begin{aligned} Y_1 &= \theta_0\phi_0(\mathbf{X}_1) + \theta_1\phi_1(\mathbf{X}_1) + \cdots + \theta_k\phi_k(\mathbf{X}_1) + \epsilon_1 \\ &\vdots \\ Y_n &= \theta_0\phi_0(\mathbf{X}_n) + \theta_1\phi_1(\mathbf{X}_n) + \cdots + \theta_k\phi_k(\mathbf{X}_n) + \epsilon_n \end{aligned}$$

where $n > k$. In other words, $\mathbf{Y}_{n\times 1} = H_{n\times k}\,\boldsymbol{\theta}_{k\times 1} + \boldsymbol{\epsilon}_{n\times 1}$ where $\mathbf{Y} = (Y_1, \ldots, Y_n)$, $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_k)$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ and

$$H = \begin{bmatrix} \phi_0(\mathbf{X}_1) & \cdots & \phi_k(\mathbf{X}_1) \\ \vdots & \ddots & \vdots \\ \phi_0(\mathbf{X}_n) & \cdots & \phi_k(\mathbf{X}_n) \end{bmatrix}$$

Example

Example: Univariate Linear Regression passing through the origin (only parameter is the slope).

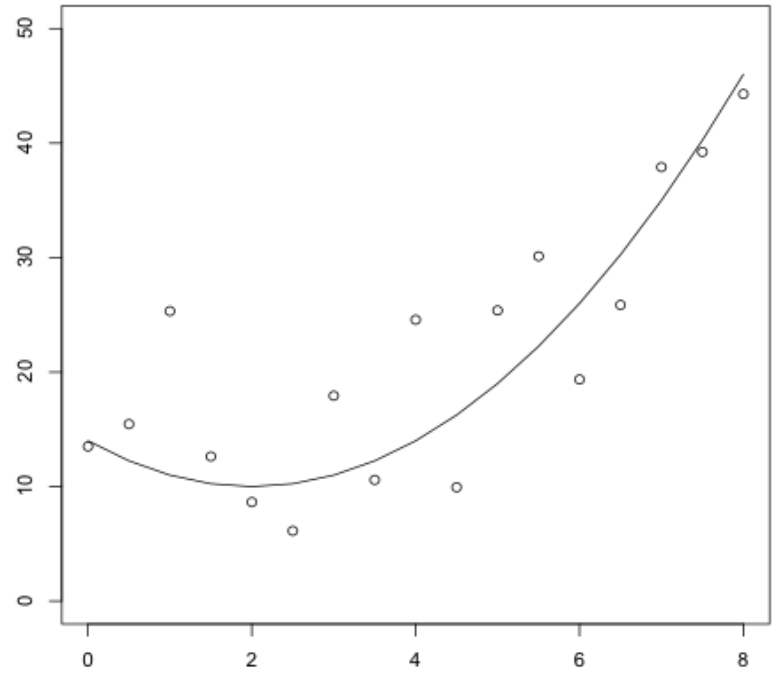


Model:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_n \end{bmatrix} \theta + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_n \end{bmatrix}$$

Example

Example: Univariate Polynomial Regression of Order 2.



Model:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 & X_1^2 \\ 1 & X_2 & X_2^2 \\ \dots & \dots & \dots \\ 1 & X_n & X_n^2 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_n \end{bmatrix}$$

Least-Squares Linear Regression

- Let us write again the model:

$$\mathbf{Y}_{n\times 1} = H_{n\times k}\,\boldsymbol{\theta}_{k\times 1} + \boldsymbol{\epsilon}_{n\times 1}$$

where $n > k$ and $\mathrm{Rank}(H) = n$.

- Let $\hat{\mathbf{Y}} = H\hat{\boldsymbol{\theta}}$. Gauss prescribes minimizing the sum of squares

$$||\mathbf{Y} - \hat{\mathbf{Y}}||^2 = (\mathbf{Y} - H\hat{\boldsymbol{\theta}})^T(\mathbf{Y} - H\hat{\boldsymbol{\theta}})$$

- The solution is easily found to be

$$\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = H^L\mathbf{Y} = (H^TH)^{-1}H^T\mathbf{Y}$$

where $H^L = (H^TH)^{-1}H^T$ is the "left-inverse" of full-rank matrix H.

Univariate Linear Regression Formulas

- (Without intercept) Model:

$$\begin{bmatrix} Y_1 \\ \cdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_1 \\ \cdots \\ X_n \end{bmatrix} \theta + \begin{bmatrix} \epsilon_1 \\ \cdots \\ \epsilon_n \end{bmatrix}$$

- Least-Squares Solution

$$\hat{\theta}_{\mathrm{LS}} = (H^T H)^{-1} H^T \mathbf{Y}$$

$$= \left([X_1 \ \cdots \ X_n] \begin{bmatrix} X_1 \\ \cdots \\ X_n \end{bmatrix} \right)^{-1} [X_1 \ \cdots \ X_n] \begin{bmatrix} Y_1 \\ \cdots \\ Y_n \end{bmatrix}$$

$$= \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} = \frac{R_{XY}}{R_{XX}}$$

Univariate Linear Regression Formulas

- (With intercept) Model:

$$\begin{bmatrix} Y_1 \\ \dots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ \dots & \dots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \dots \\ \epsilon_n \end{bmatrix}$$

- Least-Squares Solution (prove it):

$$\hat{\theta}_{0,\mathrm{LS}} = \overline{Y} - \hat{\theta}_{1,\mathrm{LS}}\,\overline{X}$$

$$\hat{\theta}_{1,\mathrm{LS}} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{S_{XY}}{S_{XX}}$$

where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

Gauss-Markov Theorem

- The previous results are purely deterministic. If $\boldsymbol{\epsilon}$ is considered a random vector, then so is $\mathbf{Y}$, and thus $\hat{\boldsymbol{\theta}}$.
- One can now talk about estimator bias and variance.
- **Gauss-Markov Theorem.** (homoskedastic case) If $E[N] = 0$ (zero-mean noise) and $E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \sigma^2 I_n$, then

$$\hat{\boldsymbol{\theta}}_{\text{BLUE}} = (H^T H)^{-1} H^T \mathbf{Y}$$

is unbiased ($E[\hat{\boldsymbol{\theta}}_{\text{BLUE}}] = \boldsymbol{\theta}$), and its variance is minimum among all linear estimators $\hat{\boldsymbol{\theta}} = \mathbf{B}\mathbf{Y}$, in the sense that

$$\text{Var}\left(\hat{\boldsymbol{\theta}}_{\text{BLUE},i}\right) = \sigma^2 \left[(H^T H)^{-1}\right]_{ii}$$

is minimum for each $i = 1, \ldots, n$.

Gaussian Noise Case

- If we can further assume the noise ϵ to be Gaussian, then we can show that the least-square solution is the maximum-likelihood solution to the model.
- Let $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$ (homoskedastic case). Then

$$\mathbf{Y}_{n\times 1} = H_{n\times k}\,\boldsymbol{\theta}_{k\times 1} + \boldsymbol{\epsilon}_{n\times 1} = \hat{\mathbf{Y}}_{n\times 1} + \boldsymbol{\epsilon}_{n\times 1} \sim \mathcal{N}(\hat{\mathbf{Y}}_{n\times 1}, \sigma^2 I_n)$$

- The likelihood function for this model is:

$$L(\boldsymbol{\theta}) = p(\mathbf{Y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(Y_i - \hat{Y}_i)^2}{2\sigma^2}\right)$$

Gaussian Noise Case - II

- The log-likelihood can be thus be written as

$$\log L(\boldsymbol{\theta}) \;=\; \text{const} \;-\; \sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{2\sigma^2}$$

- Therefore, maximizing the likelihood is equivalent to minimizing the sum of squares, and the MLE estimator and the least-squares estimator are the same.

- As an MLE estimator, the least-squares estimator is asymptotically unbiased, consistent, asymptotically efficient, and asymptotically normal.

Gaussian Noise Case - III

- Maximizing the log-likelihood with respect to σ produces the MLE estimator of the variance

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

The MLE of σ^2 is therefore the RSS!

- However, a straightforward calculation (try it) shows that, for a linear model with k parameters,

$$E[\hat{\sigma}^2] = \frac{n-k}{n}\sigma^2$$

so that the MLE is biased (though it is *asymptoticallly* unbiased, as a MLE). For this reason, one prefers:

$$\hat{\sigma}^2_{\text{unbiased}} = \frac{n}{n-k}\hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Penalized Least Squares

- In some cases, it is desirable to introduce some constraint on the coefficients of the regression in order to avoid overfitting. This is called penalized least squares or *ridge regression*.
- This also has the effect of producing more stable numerical solutions.
- One replace the least squares criterion $||\mathbf{Y} - H\boldsymbol{\theta}||^2$ by

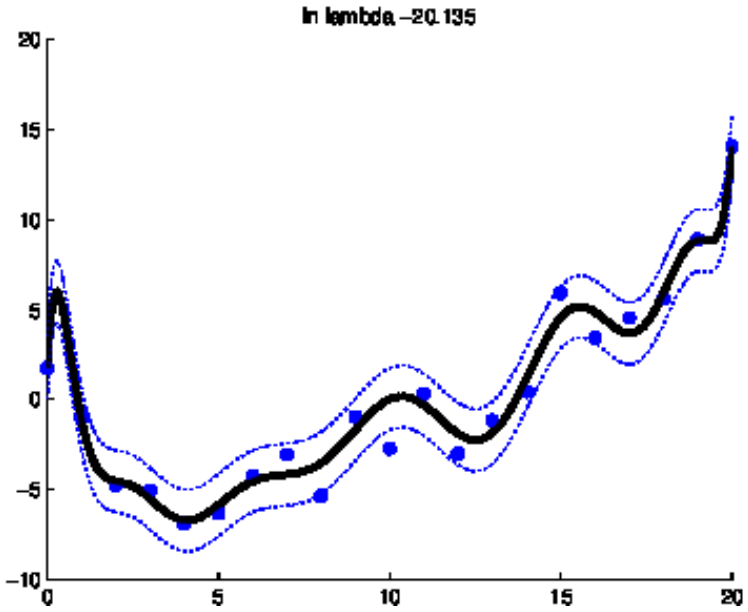
$$||\mathbf{Y} - H\hat{\boldsymbol{\theta}}||^2 + \lambda||\hat{\boldsymbol{\theta}}||^2 = (\mathbf{Y} - H\hat{\boldsymbol{\theta}})^T(\mathbf{Y} - H\hat{\boldsymbol{\theta}}) + \lambda\hat{\boldsymbol{\theta}}^T\hat{\boldsymbol{\theta}}$$

- Minimizing as before, we get the solution (prove it):

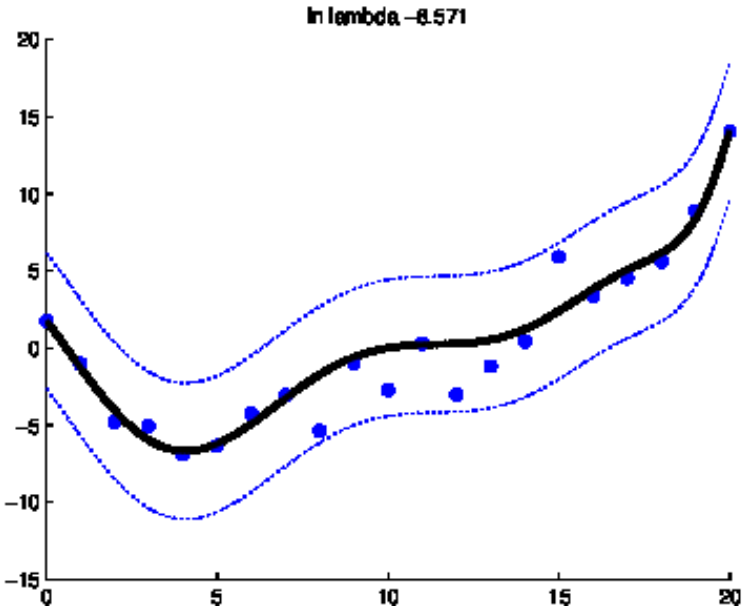
$$\hat{\boldsymbol{\theta}}_{\mathrm{PLS}} = (H^T H + \lambda I_k)^{-1} H^T \mathbf{Y}$$

Example

- Polynomial regression (order 14) to $n = 21$ data points (Murphy)



Small λ



Large λ

Variable Selection

- There are three basic ways to perform variable selection in regression.
  - Wrapper search (similar to classification).
    - Exhaustive
    - Sequential forward/backward search
    - Floating Search
  - Statistical testing of each coefficients for the hypothesis that the coefficient is zero (and discarding those that are not significant).
  - Shrinking the coefficients towards zero to generate *sparse* solutions.

Criteria for Wrapper Search

- If one tries to perform wrapper selection by minimizing the RSS, or equivalently maximizing R^2, one ends up with an overfit model with all variables.
- For variable selection, one usually employs the adjusted R^2 statistic:

$$\text{Adjusted}R^2 \;=\; 1 - \frac{\text{RSS}/(\text{n} - \text{d} - 1)}{\text{TSS}(\text{n} - 1)}$$

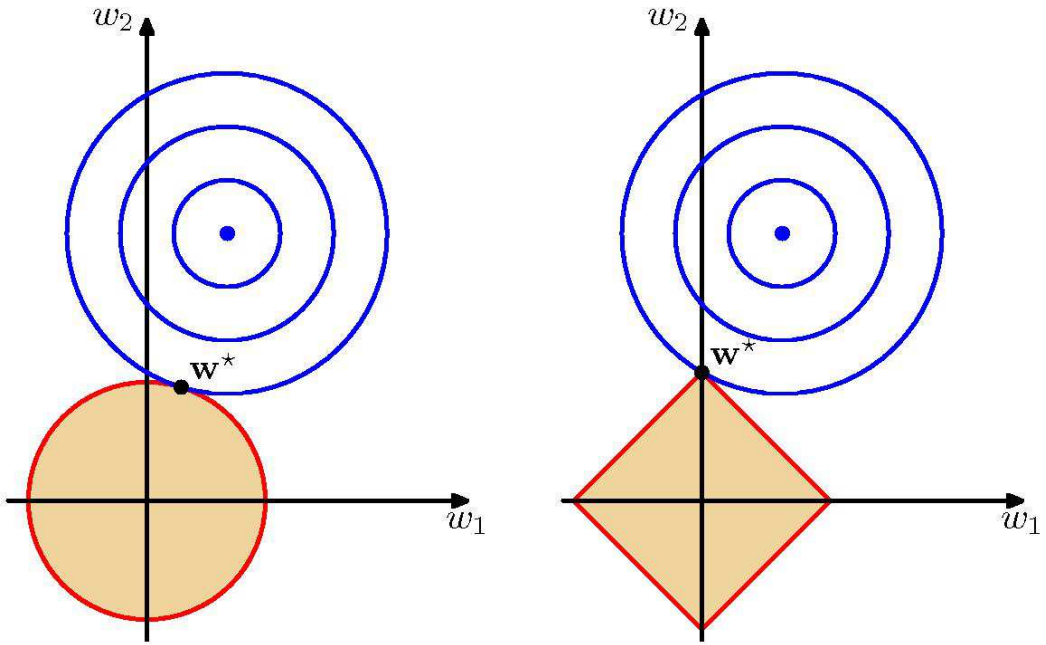
- Other criteria that are used: Mallows' C_p, AIC, BIC. For example, Mallows' C_p for the Gaussian model is

$$C_p \;=\; \text{RSS} + \frac{2d}{n}\hat{\sigma}^2 \,.$$

where $\hat{\sigma}^2$ is the variance estimator mentioned before.

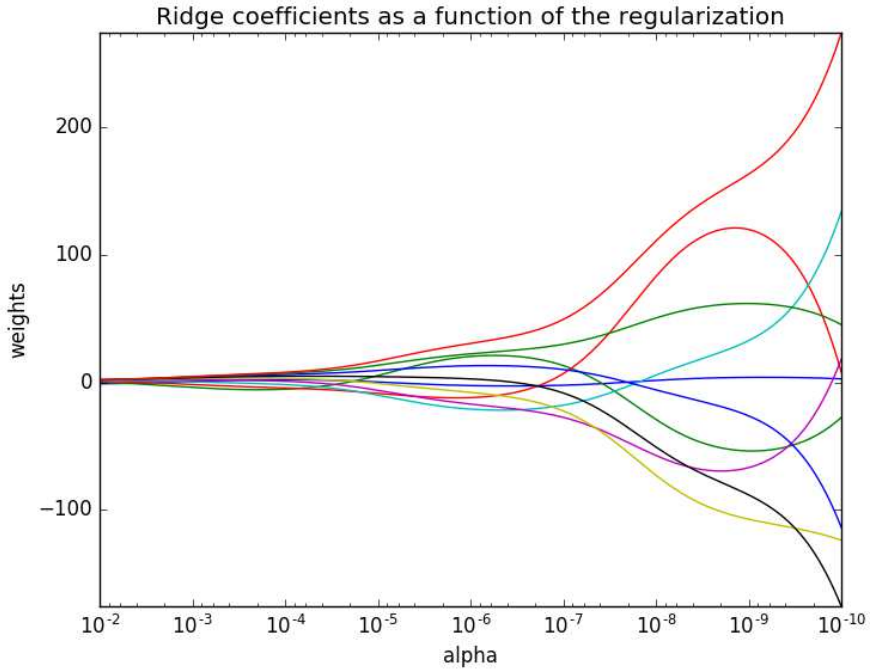
Shrinkage: Lasso and Elastic Net

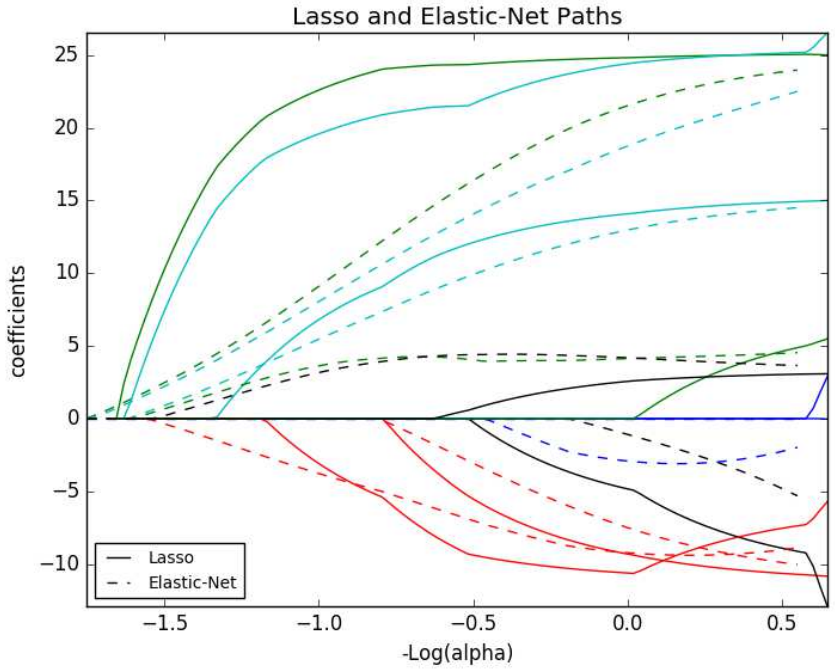
- Similar to ridge regression, but uses L_1 norm, and can drive coefficient values to zero.
  - Lasso: $||\mathbf{Y} - H\hat{\boldsymbol{\theta}}||^2 + \lambda||\hat{\boldsymbol{\theta}}||_1$.
  - Elastic Net: $||\mathbf{Y} - H\hat{\boldsymbol{\theta}}||^2 + \lambda_1||\hat{\boldsymbol{\theta}}||_1 + \lambda_2||\hat{\boldsymbol{\theta}}||_2^2$.



Example: Coefficient Paths

Coefficient paths as a function of regularization parameter.

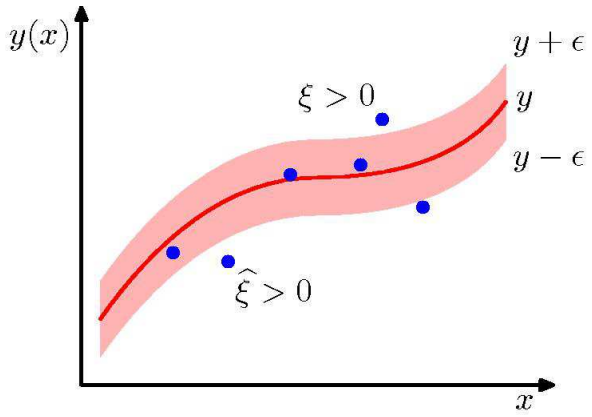




(source: http://scikit-learn.org/stable/modules/linear_model.html)

SVM Regression

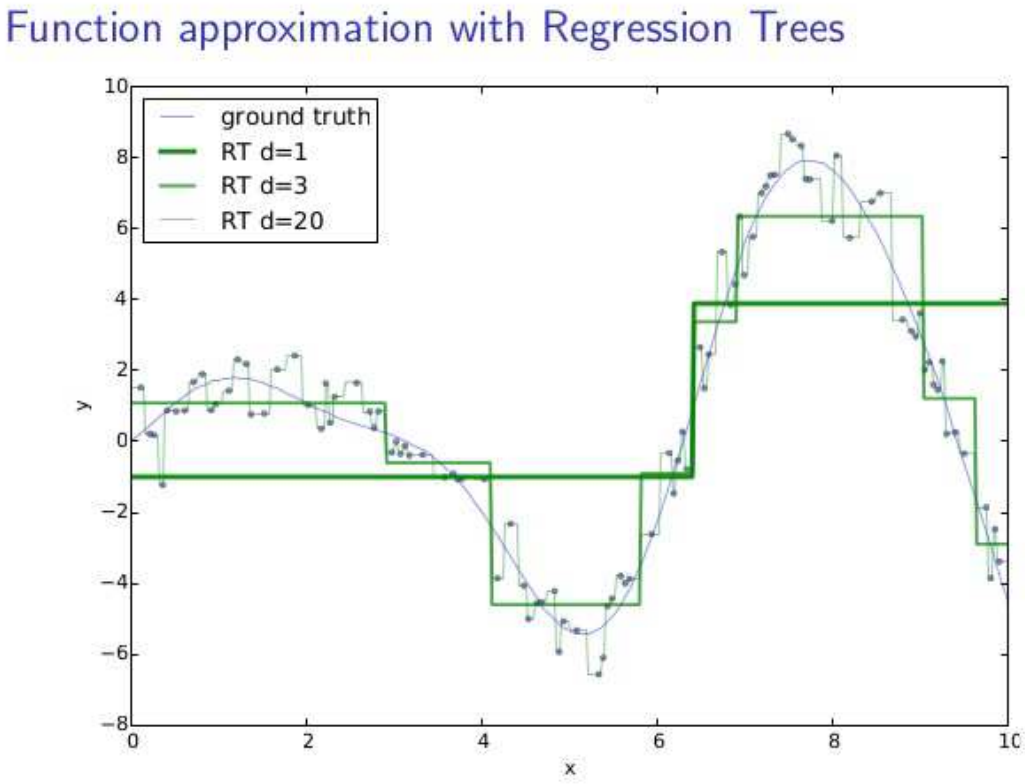
- Similar to SVM for classification. The same idea of margin reappears, but now the points are supposed to be *within* a margin of width ϵ.
- Slack variables are associated to outliers that break the margin criterion. Only support vectors (margin and outlier vectors) determine regression curve.
- Nonlinear version with mapping to high-dimensional space with kernels, as before.



(source: Bishop)

Regression Trees

- Similar to classification trees (CART).
- A regression is fit at each node and the node impurity is the RSS. The value at assigned to leaf nodes is the mean.



Logistic Regression

- Well-known in Statistics. It is based on the "logit" (i.e., log-odds) function

$$\text{logit}(p) \;=\; \ln\left(\frac{p}{1-p}\right)\,, \quad 0 < p < 1\,.$$

- Here $Y \sim \text{Bernoulli}(\pi(\mathbf{X}))$, i.e., $P(Y = 1 \mid \mathbf{X}) = \pi(\mathbf{X})$. This is modelled as a linear regression in logit space:

$$\text{logit}(\pi(\mathbf{X})) \;=\; \ln\left(\frac{\pi(\mathbf{X})}{1-\pi(\mathbf{X})}\right) \;=\; a_0 + a_1 X_1 + \cdots + a_d X_d\,,$$

so that

$$\pi(\mathbf{x}) \;=\; \frac{1}{1 + e^{-(a_0 + a_1 X_1 + \cdots + a_d X_d)}}$$

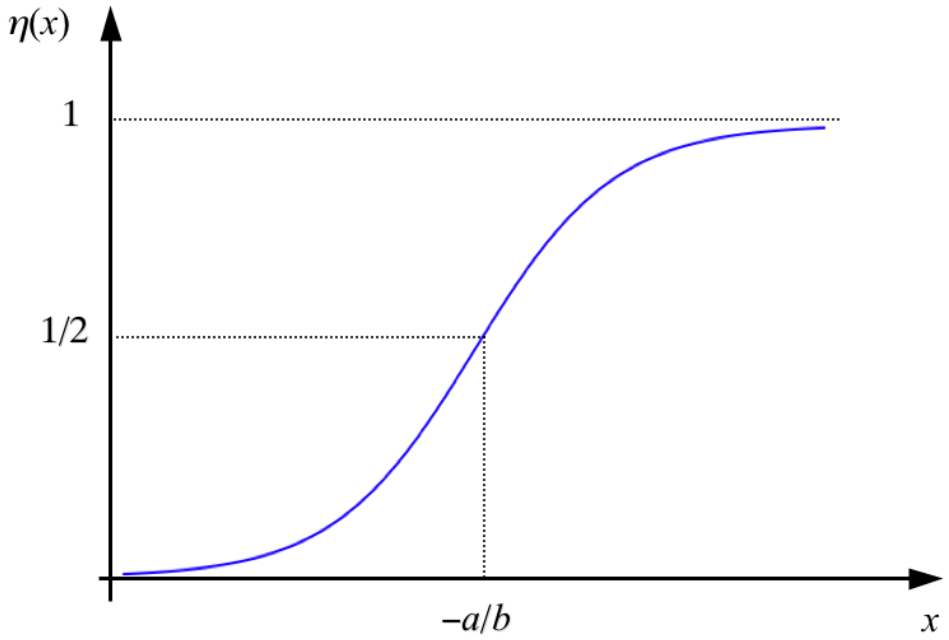
Logistic Regression - II

- The function

$$\eta(\mathbf{x}) = \frac{1}{1 + e^{-(a_0 + a_1 x_1 + \cdots + a_d x_d)}}$$

  is called the *logistic curve*.

- Univariate example: $\eta(x) = 1/(1 + e^{-(a+bx)})$.



Logistic Regression - III

- One estimates coefficients by maximum likelihood.
- One common way to fit the model is to estimate $a_0, a_1, \ldots, a_d$ by *maximum likelihood*.
- Given data $S_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$, the likelihood (discarding $P(\mathbf{X}_i)$) is given by

$$L \;=\; \Pi_{i=1}^{n} P(Y_i|\mathbf{X}_i) \;=\; \Pi_{i=1}^{n} \pi(\mathbf{X}_i)^{Y_i} (1 - \pi(\mathbf{X}_i))^{1-Y_i}$$

  The maximum of this function cannot be found in close form and is found by iterative methods.
- The role of the RSS is played by the *deviance*, which is the relative difference in likelihood between the full model and the current reduced model.

Example: Challenger Accident

- Source: Challenger Shuttle Accident Investigation.
- There were 23 launches before the one in question.
- Data on O-ring failure vs. temperature.

Flight	Temp	Damage	Flight	Temp	Damage	Flight	Temp	Damage
STS-1	66	NO	STS-9	70	NO	STS 51-B	75	NO*
STS-2	70	YES	STS 41-B	57	YES	STS 51-G	70	NO
STS-3	69	NO	STS 41-C	63	YES	STS 51-F	81	NO
STS-4	80	???	STS 41-D	70	YES	STS 51-I	76	NO
STS-5	68	NO	STS 41-G	78	NO	STS 51-J	79	NO
STS-6	67	NO	STS 51-A	67	NO	STS 61-A	75	YES
STS-7	72	NO	STS 51-C	53	YES	STS 61-B	76	NO
STS-8	73	NO	STS 51-D	67	NO	STS 61-C	58	YES

- The problem is to estimate the probability that the launch would have failed at the actual launch temperature of 31 F.

Example: Challenger Accident

- Fitted model:

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = 15.043 - 0.2322 \times \text{temp}$$

- Prediction at temp $= 31$:

$$\pi(x) = \frac{1}{1+e^{0.2322\times\text{temp}-15.043}} \Rightarrow \pi(31) \approx 1.$$