Introduction to Regression

Antanas Verikas
antanas.verikas@hh.se

IDE, Halmstad University

2013

Example

Regression aims at finding a function that fits the observations.

Figure: Consumer prise index in Sweden.

Observations:
(x,y) pairs

(1949, 100)
(1950, 117)
...
(1996, 1462)
(1997, 1469)
(1998, 1467)
(1999, 1474)

Linear fit

The linear fit is not so good.

y	$\widehat{y}$
100	-215
117	-184
...	...
1467	1314
1474	1345

Figure: Consumer prise index in Sweden, linear fit.

Example

Apply a transformation.

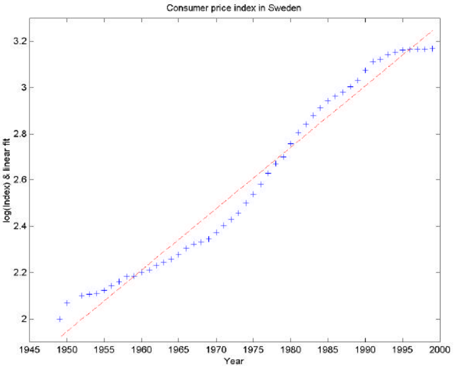


Figure: Consumer prise index in Sweden.

Take logarithm of y and fit a straight line.

Linear fit

Transform y back to the original. The fit is better.

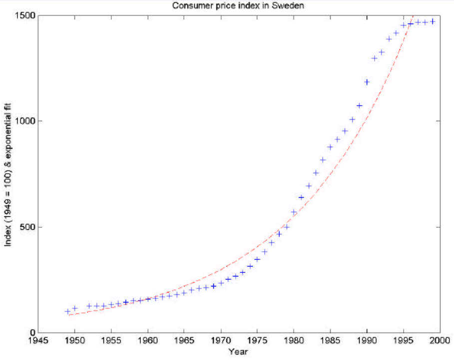


Figure: Consumer prise index in Sweden.

y	$\widehat{y}$
100	83
117	88
...	...
1467	1660
1474	1765

Regression task

Construct a model of a process, using examples of the process.

Input: $\mathbf{x}$ (possibly a vector)

Output: $y = g(\mathbf{x})$ (generated by the process)

Examples: Pairs of input and output $\{y(n), \mathbf{x}(n)\}$

Our model: $\widehat{y} = f(\mathbf{x})$

The function f is our estimate of the true function g

Data and assumptions

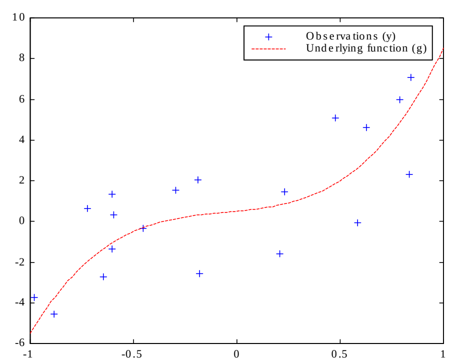
$$
\begin{aligned}
\texttt{Data set } \mathbf{Z} &= \{\mathbf{x}(n), y(n)\}_{n=1,\dots,N} \\
y(n) &= g[(\mathbf{x}(n)] + \varepsilon(n)
\end{aligned}
$$

$\mathbf{x}(n)$	Observed input
$y(n)$	Observed output
$\mathrm{g}[(\mathbf{x}(\mathrm{n})]$	True underlying function
$\varepsilon(n)$	i.i.d noise process with zero mean

Example

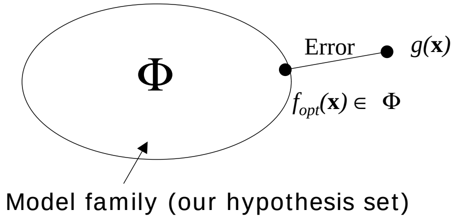
Underlying function: $g(x) = 0.5 + x + x^2 + 6x^3$

Noise: $\varepsilon \sim N[0, 2]$

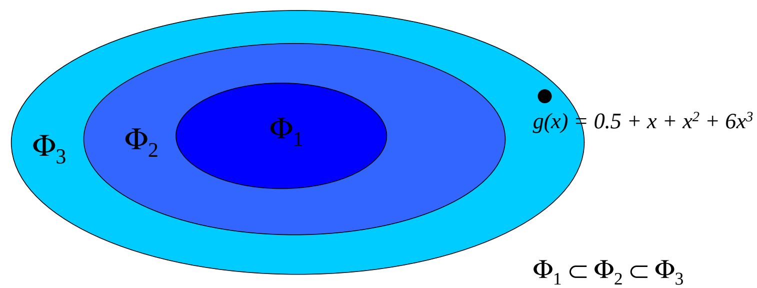


Idealized regression

Find appropriate model family Φ and $f(\mathbf{x}) \in \Phi$ with a minimum "distance" (error) to $g(\mathbf{x})$



Examples of model families



Linear	$\Phi_1 = \{a + bx\}$
Quadratic	$\Phi_2 = \{a + bx + cx^2\}$
Cubic	$\Phi_3 = \{a + bx + cx^2 + dx^3\}$

How to measure “distance”?

Q: What does the distance between functions f and g mean?

A: The difference between the functions f and g.

Q: How do we measure difference (error) between functions?

The summed squared error (SSE)

$$E = \mathrm{SSE} = \sum_{n=1}^{N} \{ f[\mathbf{x}(n), \mathbf{w}] - y(n) \}^2 \tag{1}$$

$\mathbf{w}$ = the parameters of the function f.

SSE assumes zero mean i.i.d noise

SSE $\Longleftrightarrow$ "Least squares" fit.

Negative log-likelihood

$$\begin{aligned} \texttt{Data set } \mathbf{Z} &= \{\mathbf{x}(n), y(n)\}_{n=1,\ldots,N} \\ y(n) &= g[(\mathbf{x}(n)] + \varepsilon(n) \end{aligned} \qquad (2)$$

$$E = -\ln L = -\ln\left[\prod_{n=1}^{N} p[\mathbf{z}(n)|\mathbf{w}]\right] \qquad (3)$$

It is common to assume normally distributed noise $\Longrightarrow$

$$p[\mathbf{z}(n)|\mathbf{w}] = p\{f[\mathbf{x}(n), \mathbf{w}] - y(n)\} \sim N[0, \sigma] \qquad (4)$$

This leads to $E \propto \texttt{SSE}$.

The Bayesian error measure (1)

- Why maximize the likelihood for the observations given the model parameters?
- Maximize the likelihood for the model parameters given the observations, instead.
- Bayes' theorem tells us how we should do.

The Bayesian error measure (2)

The probability for the model parameters, given the observations:

$$p(\mathbf{w}|\mathbf{Z}) = \frac{p(\mathbf{Z}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{Z})} = \frac{\mathcal{L}(\mathbf{Z}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{Z})} \tag{5}$$

where $p(\mathbf{w})$ is our "prior" for the model parameters $\mathbf{w}$.
More convenient to minimize the negative likelihood:

$$\begin{aligned} E = -\ln p(\mathbf{w}|\mathbf{Z}) &= -\ln \mathcal{L}(\mathbf{Z}|\mathbf{w}) - \ln p(\mathbf{w}) + \ln p(\mathbf{Z}) \\ \rightarrow &= -\ln \mathcal{L}(\mathbf{Z}|\mathbf{w}) - \ln p(\mathbf{w}) \end{aligned} \tag{6}$$

since the third term does not depend on the model parameters w.

The Bayesian error measure (3)

$$E = -\ln p(\mathbf{w}|\mathbf{Z}) \propto -\ln \mathcal{L}(\mathbf{Z}|\mathbf{w}) - \ln p(\mathbf{w}) \tag{7}$$

Allows including a prior belief, expressed in $p(\mathbf{w})$, about the function $f(\mathbf{x}, \mathbf{w})$.

An example is:

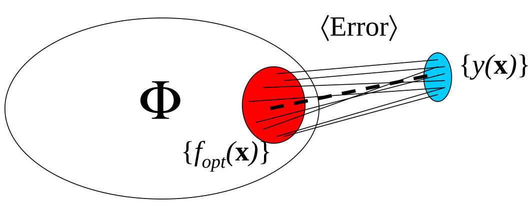
$$p(\mathbf{w}) \propto \exp(-\|\mathbf{w}\|^2 / 2\sigma_W^2) \tag{8}$$

The Bayesian error measure (4)

- The Bayesian error measure is more general than the ML error.
- The ML error is the special case of the Bayesian error with a uniform prior.
- The Bayesian error is very important to avoid over-fitting.

The real regression

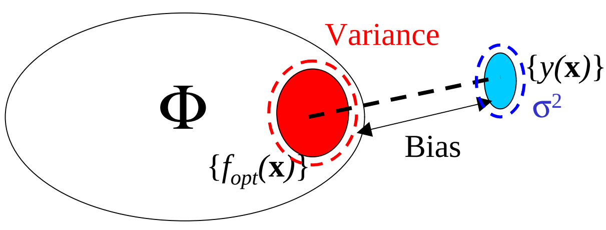
Find an appropriate model family Φ and minimize the **expected** distance to $y(\mathbf{x})$ ("generalization error")



Data is never noise free, and never available in infinite amounts, thus we get variation in data and model. The generalization error is a function of both the training data and the hypothesis selection method.

Model “bias” & model “variance”

$$\langle \texttt{Error} \rangle = (\texttt{Bias})^2 + (\texttt{Variance}) + \sigma_\varepsilon^2 \tag{9}$$



Example (1)

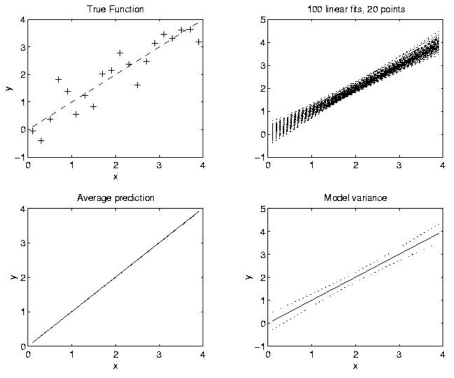


Figure: A linear function $g(x)$ fitted with a linear model $f(x)$, small variance.

Example (2)

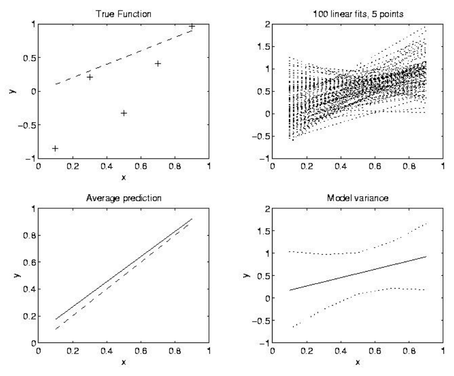


Figure: A linear function $g(x)$ fitted with a linear model $f(x)$, larger variance.

Example (1)

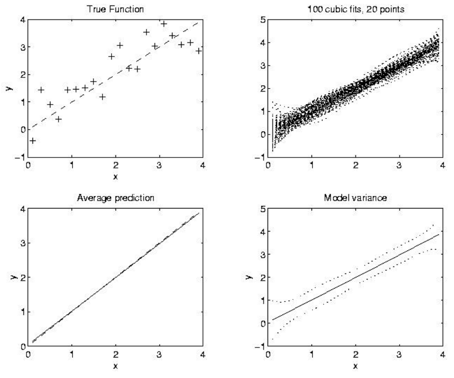


Figure: A linear function $g(x)$ fitted with a cubic model $f(x)$, small variance.

Example (2)

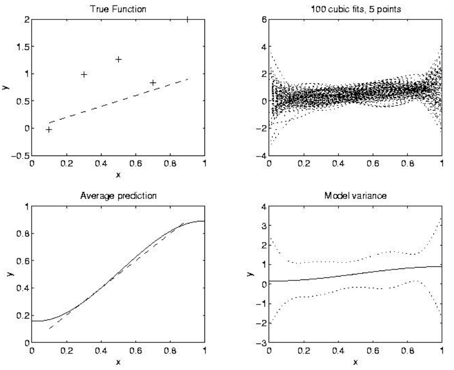


Figure: A linear function $g(x)$ fitted with a cubic model $f(x)$, larger variance.

Example (1)

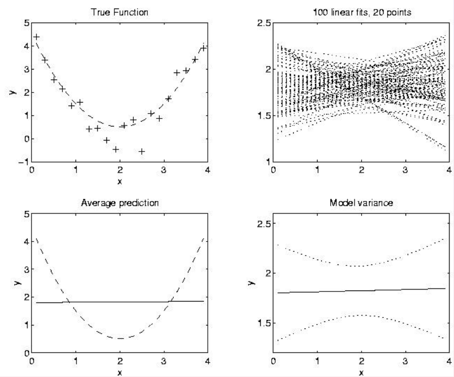


Figure: A quadratic function $g(x)$ fitted with a linear model $f(x)$, small variance.

Example (2)

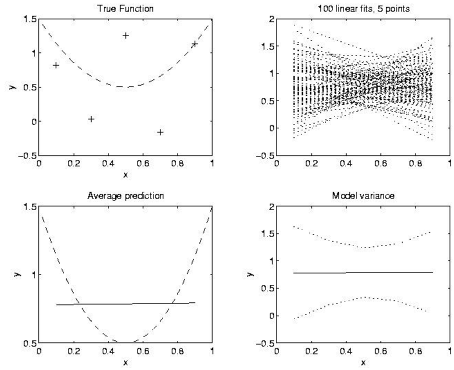


Figure: A quadratic function $g(x)$ fitted with a linear model $f(x)$, larger variance.

Model selection

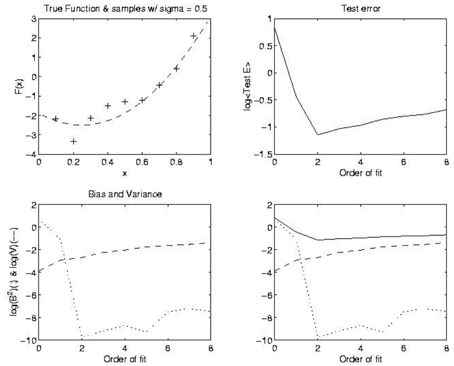


Figure: Model with the lowest generalization error is a bias versus variance trade-off.

Model complexity

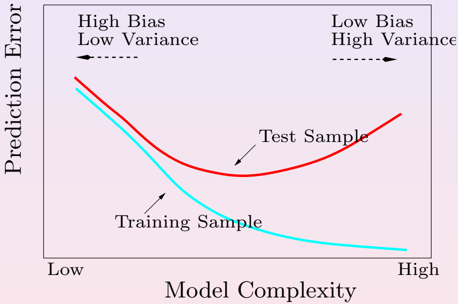


Figure: Model with the lowest generalization error is a bias versus variance trade-off.

Variable selection

More variables imply larger variance

For linear regression models:

$$\langle E_{\mathtt{Test}} \rangle = \langle E_{\mathtt{Train}} \rangle + \frac{\sigma_\varepsilon^2 (D+1)}{N} \tag{10}$$

$\Rightarrow$ A penalty is payed for each input.