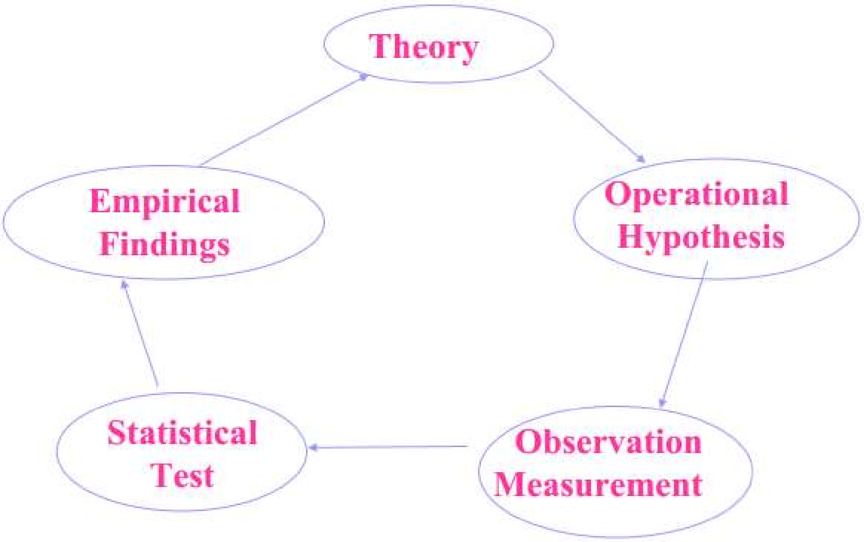
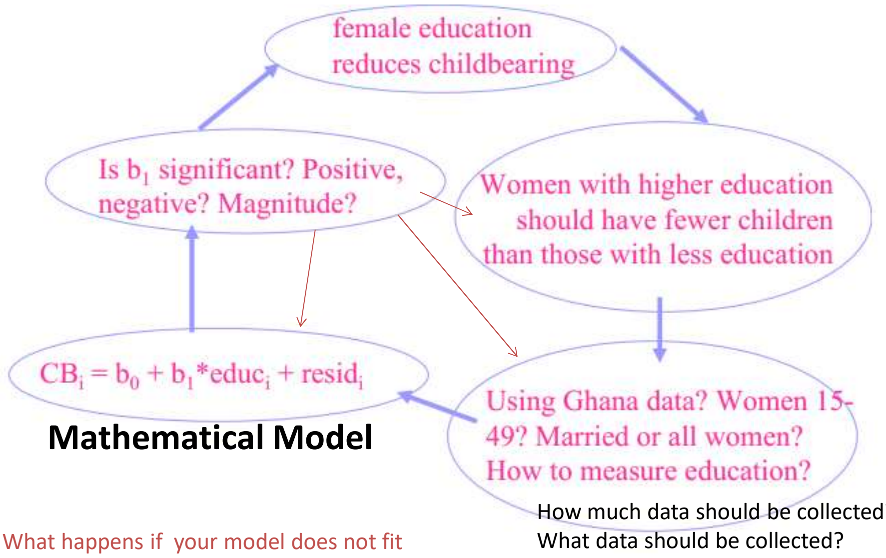
Data Modelling
and
Regression Techniques

M. Fatih Amasyalı

The traditional scientific approach



Example of a scientific approach



How much data should be collected?
What data should be collected?

What happens if your model does not fit your theory/hypothesis?

- A model is an underlying theory about how the world works. It includes:
  - Assumptions
  - Key players (independent variables)
  - Interactions between variables
  - Outcome set (dependent variables)
- CB=x1+educ*x2+resid
  - Assumptions?, variables?, interactions?

Regression Models

- Relationship between one dependent variable and explanatory variable(s)
- Use equation to set up relationship
  - Numerical Dependent (Response) Variable
  - 1 or More Numerical or Categorical Independent (Explanatory) Variables

Regression Modeling Steps

1. Hypothesize Deterministic Component
   - Estimate Unknown Parameters
2. Evaluate the fitted Model
3. Use Model for Prediction & Estimation

Specifying the deterministic component

1. Define the dependent variable and independent variable(s)

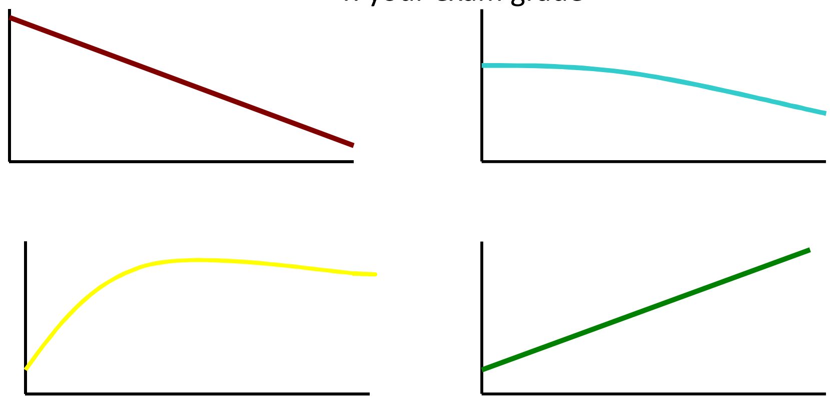
2. Hypothesize Nature of Relationship
   – Expected Effects (i.e., Coefficients' Signs)
   – Functional Form (Linear or Non-Linear)
   – Interactions

Model Specification Is Based on Theory

1. Previous Research
2. 'Common Sense'
3. Data (which variables, linear/non-linear etc.)

Thinking Challenge: Which Is More Logical?

X: How many hours did you study in the night before the exam?
Y: your exam grade



Types of Regression Models

The linear first order model $Y=\beta_0+ \beta_1X+\varepsilon$
The linear second order model $Y=\beta_0+ \beta_1X+ \beta_2X^2+ \varepsilon$
The linear n order model $Y=\beta_0+ \beta_1X+ \beta_2X^2+ \dots + \beta_nX^n+ \varepsilon$

ε is random error.

The word **linear** refers to the linearity of the parameters β_i .

The **order** (or **degree**) of the model refers to the highest power of the predictor variable X.

Types of Regression Models

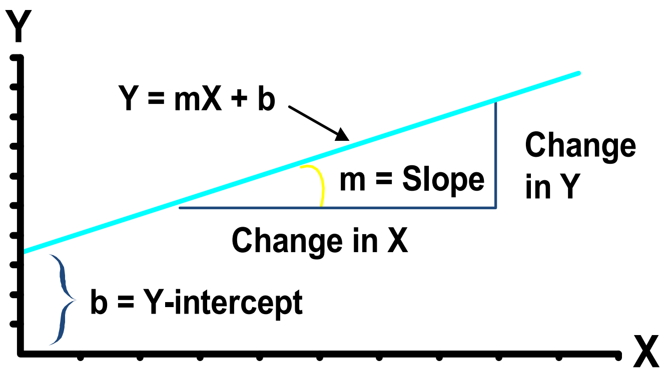
- If the parameters are linear related to the each other the model is linear. A non-linear first order model: $Y=(\beta_0 X) / (\beta_1+X)+\varepsilon$
- If X has d dimensions, a linear first order model: $Y=\beta_0+ \beta_1 X_1+ \beta_2 X_2+ \ldots + \beta_d X_d+ \varepsilon$

- A linear first order model $Y=\beta_0+ \beta_1 X+\varepsilon$
- To get the model, we need to estimate the parameters β_0 and β_1
- Thus, the estimate of our model is

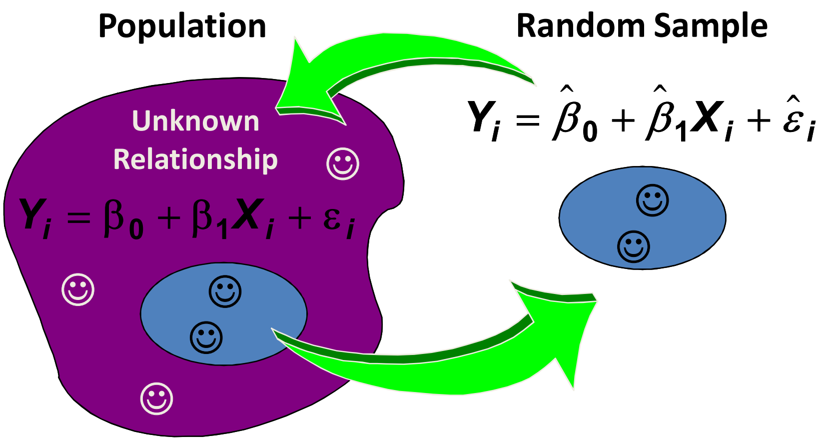
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- Y_{hat} denotes the predicted value of Y for some value of X, and β_{0hat} and β_{1hat} are the estimated parameters.

An Old Friend

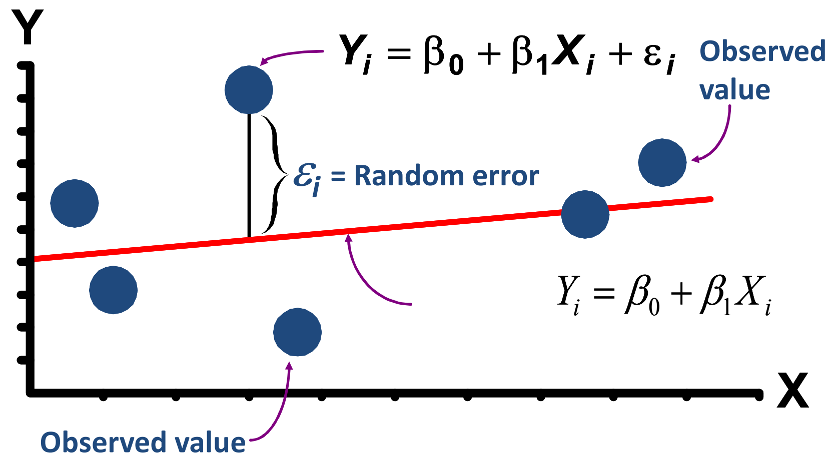


Population & Sample Regression Models

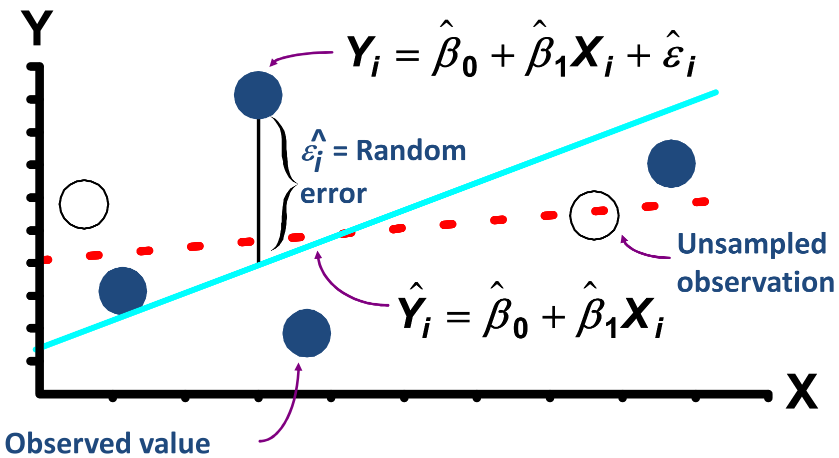


Mostly, you can not reach the all population.

Population Linear Regression Model



Sample Linear Regression Model



Estimating Parameters: Least Squares Method

Least Squares

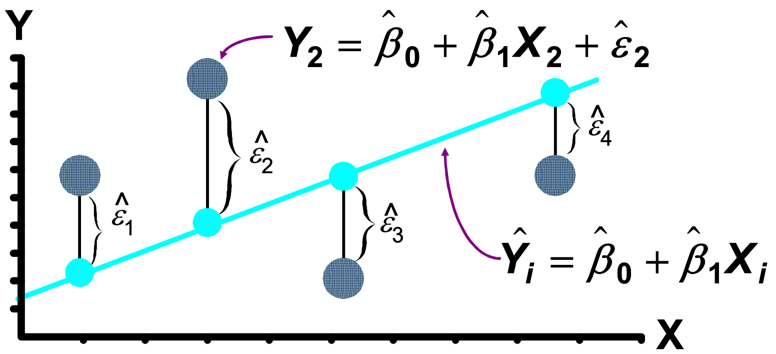
- 1. 'Best Fit' means difference between actual Y values & predicted Y values are a minimum. *But* positive differences off-set negative. So square errors!

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

- 2. LS Minimizes the Sum of the Squared Differences (errors) (SSE)
- Mean squared error (MSE)= $\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$

Least Squares Graphically

LS minimizes $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$



Mehmet Fatih AMASYALI Optimization Techniques Lecture Notes YILDIZ TECHNICAL UNIVERSITY COMPUTER ENG DEPT.

Interpretation of Coefficients

- 1. Slope ($\hat{\beta}_1$)
  - Estimated *Y* Changes by $\hat{\beta}_1$ for Each 1 Unit Increase in *X*
    - If $\hat{\beta}_1$ = 2, then *Y* Is Expected to Increase by 2 for Each 1 Unit Increase in *X*
- 2. Y-Intercept ($\hat{\beta}_0$)
  - Average Value of *Y* When *X* = 0
    - If $\hat{\beta}_0$ = 4, then Average *Y* Is Expected to Be 4 When *X* Is 0

Mehmet Fatih AMASYALI Optimization Techniques Lecture Notes YILDIZ TECHNICAL UNIVERSITY COMPUTER ENG DEPT.

Assume that our model is Y=β

- How can we estimate β using LS?
- Least Squares Minimize squared error

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta)^2$$

$$\frac{\partial \sum_{i=1}^{n} (y_i - \beta)^2}{\partial \beta} = 0$$

$$-2\sum_{i=1}^{n} (y_i - \beta) = 0$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \beta = 0$$

$$\sum y = n\beta$$

$$\beta = \frac{1}{n}\sum y$$

A new look

- A linear first order model (X has d dim.)
- $Y=\beta_0+ \beta_1 X_1+ \beta_2 X_2+ \ldots + \beta_d X_d+ \varepsilon$
- This can be written in matrix form as
- $Y_{n*1}=X_{n*(d+1)}\beta_{(d+1)*1}+ \varepsilon_{n*1}$
- n is the sample size

Example-1

- $Y_{n*1} = X_{n*(d+1)} \beta_{(d+1)*1} + \varepsilon_{n*1}$
- n =4, d=1

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1\, x_1 \\ 1\, x_2 \\ 1\, x_3 \\ 1\, x_4 \end{bmatrix} * \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

$$Y_i = 1 * \beta_0 + \beta_1 * X_i + \varepsilon_i$$

Example-2

- $Y_{n*1} = X_{n*(d+1)} \beta_{(d+1)*1} + \varepsilon_{n*1}$
- n =4, d=2

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1\, x_{11}\ x_{12} \\ 1\, x_{21}\ x_{22} \\ 1\, x_{31}\ x_{32} \\ 1\, x_{41}\ x_{42} \end{bmatrix} * \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

$$Y_i = 1 * \beta_0 + \beta_1 * X_{i1} + \beta_2 * X_{i2} + \varepsilon_i$$

Example-3

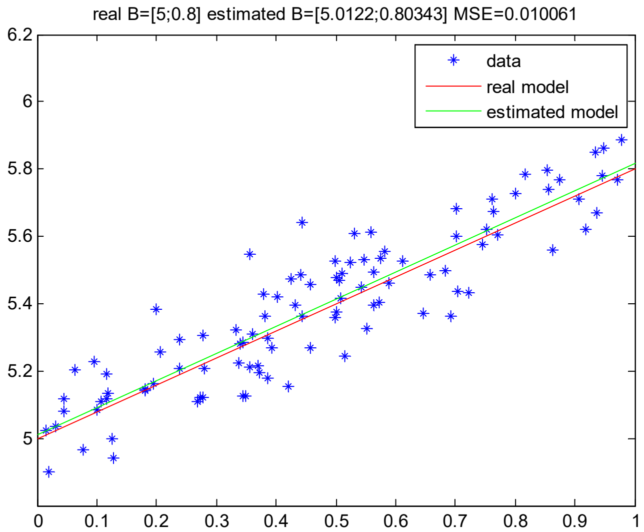
$$Y_i = 1*\beta_0 + \beta_1 * X_i + \beta_2 * X_i^2 + \varepsilon_i$$

- n =4, d=1, order=2

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \end{bmatrix} * \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

- All examples have the following form:
- $Y=X\beta$
- How can we estimate β?
- $X^{-1}Y=X^{-1}X\beta$ ($X^{-1}X=I$)
- $\beta=X^{-1}Y$ (OK, but what if X is not square matrix?)
- $X^TY=X^TX\beta$ (X^TX is always a square matrix)
- $(X^TX)^{-1}(X^TY)=(X^TX)^{-1}(X^TX)\beta$ $[(X^TX)^{-1}(X^TX)=I]$
- $\beta=(X^TX)^{-1}(X^TY)$

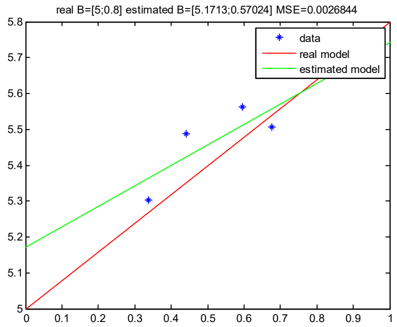
Lin_reg1.m

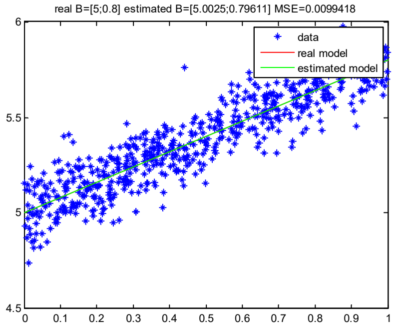


Sample size effect
lin_reg1.m

N=6

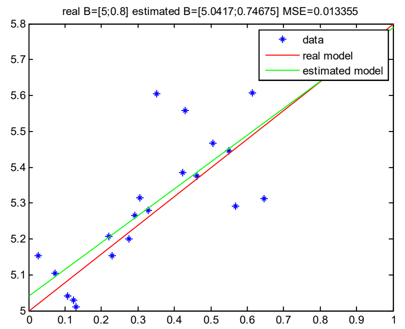
N=600



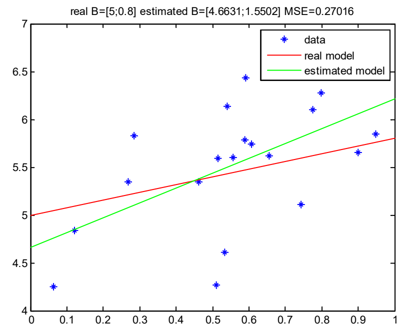


Error rate effect
lin_reg1.m

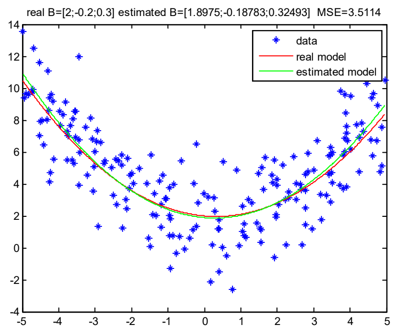
e=randn/10

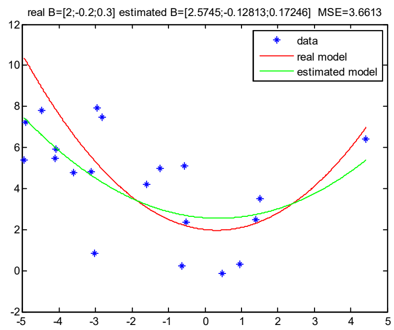


e=randn/2



Y=b0+b1*x+b2*x^2
Lin_reg2.m

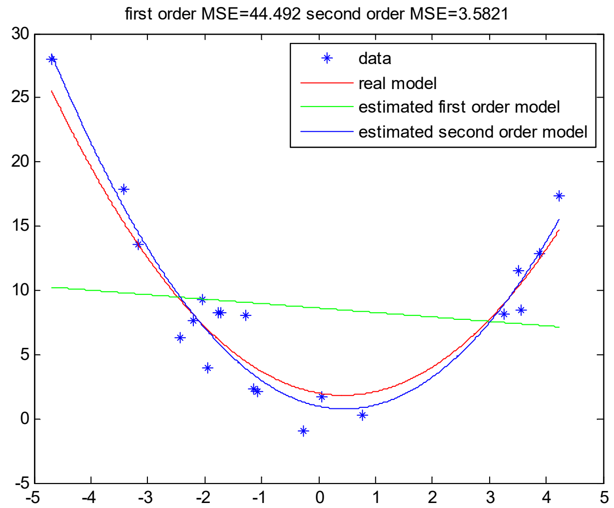




Mehmet Fatih AMASYALI Optimization Techniques Lecture Notes YILDIZ TECHNICAL UNIVERSITY COMPUTER ENG DEPT.

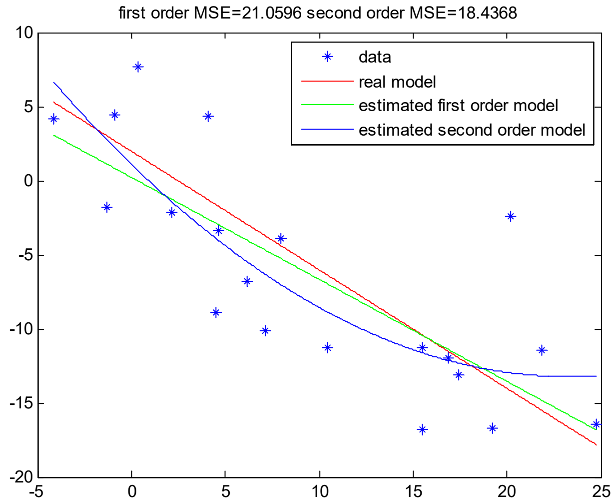
Real 2nd order, estimated 1st and 2nd order

Lin_reg3.m



Real 1st order, estimated 1st and 2nd order

Lin_reg4.m

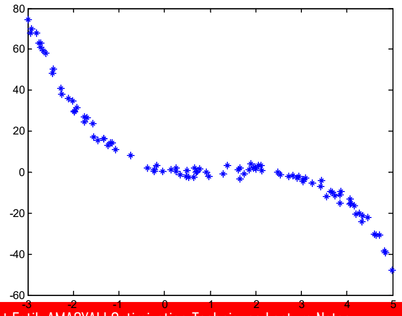


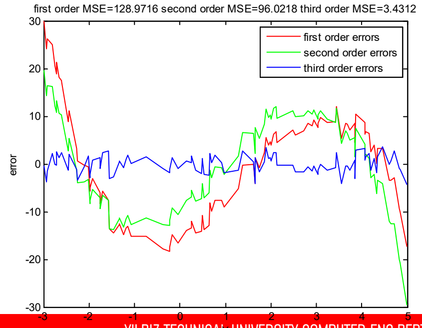
← What happens in these areas? →

% real model Y=B0+B1*X+B2*X^2+B3*X^3
% estimated model1 Y=B0+B1*X
% estimated model2 Y=B0+B1*X+B2*X^2
% estimated model3 Y=B0+B1*X+B2*X^2+B3*X^3

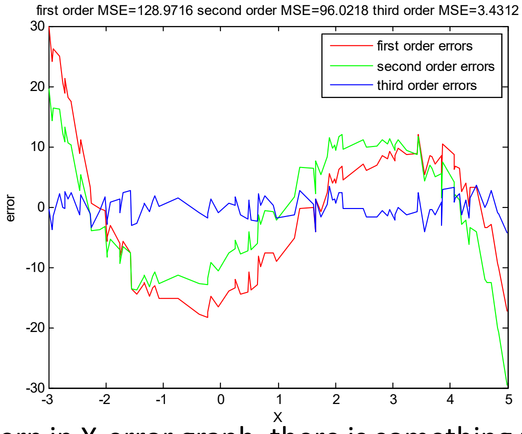
B = [1 -4 4 -1]’
fBhat = [16.3823 -9.4018]’
sBhat = [10.3682 -11.4712 1.1441]’
tBhat = [1.0170 -4.0199 3.9625 -0.9901]’

Lin_reg5.m





X-error graph

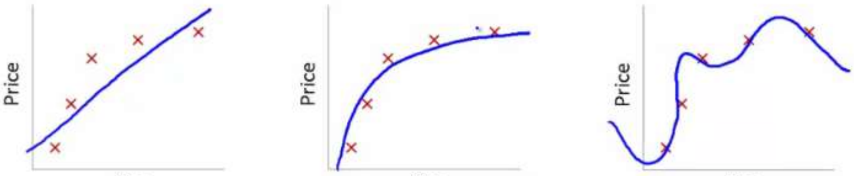


If there is a pattern in X-error graph, there is something unmodeled.

If the model degree is sufficient,
the errors have the same variances along X.

Overfitting

- **Overfitting** occurs when a statistical model describes random error or noise instead of the underlying relationship.

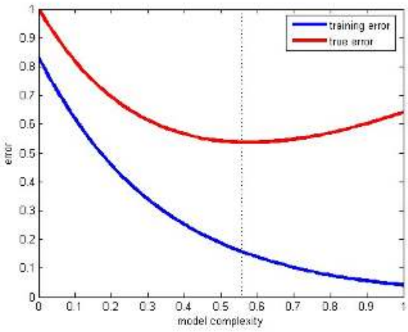


$\theta_0 + \theta_1 x$ | $\theta_0 + \theta_1 x + \theta_2 x^2$ | $\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

High bias (underfit) | "Just right" | High variance (overfit)

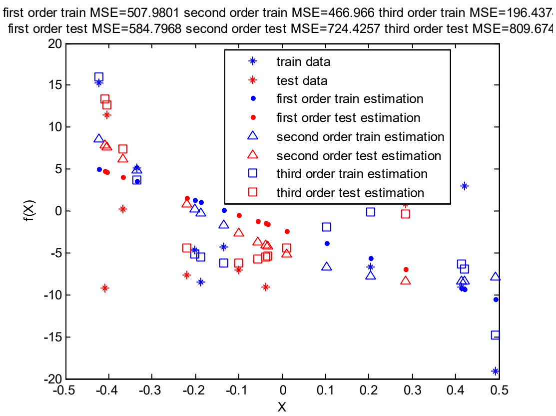
Model Validation

- Training set MSE is not reliable. WHY?
- Because, we can not determine the overfitting with training set MSE.
- Training set is used for parameter estimation.
- Test set is used for model validation.



Model Validation

- Training and test sets are seperated data samples.
- Lin_reg6.m



Y=X*β Linear System Construction

- Data= (X_{n*d}, Y_{n*1})
- n: number of data, d: dimension of data

- Model Y= β1 + β2*X
- $X_{n*2}=[1_{n*1}\ X_{n*1}]$ $\beta_{2*1}=[\beta1 ; \beta2]$

- Model Y= β1*X + β2*X^2
- $X_{n*2}=[X_{n*1}\ X^2_{n*1}]$ $\beta_{2*1}=[\beta1 ; \beta2]$

- Model Y= β1 + β2*X^2 + β3*X^3
- $X_{n*3}=[1_{n*1}\ X^2_{n*1}\ X^3_{n*1}]$ $\beta_{3*1}=[\beta1 ; \beta2 ; \beta3]$

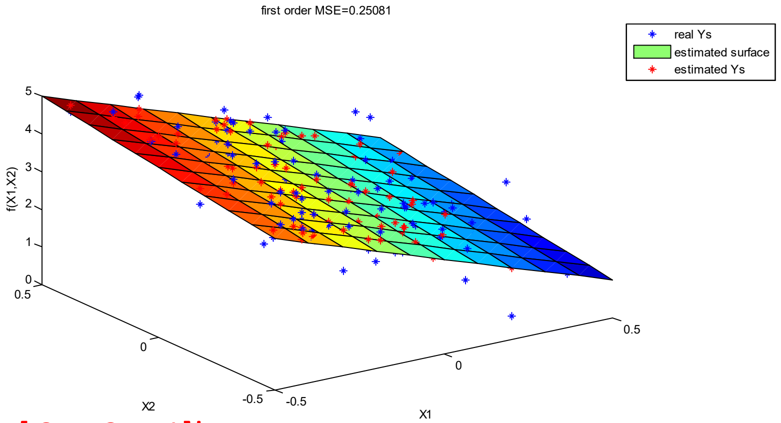
Y=X*β Linear System Construction

- Model Y= β1*X + β2*cos(X^2)
- X_{n*2}=[X_{n*1} cos(X^2)$_{n*1}$] $β_{2*1}$=[β1 ; β2]

- Model Y= β1*X1 + β2*cos(X2^2) + β3*sin(X1)
- X_{n*3}=[$X1_{n*1}$ cos($X2^2$)$_{n*1}$ sin(X1)$_{n*1}$] $β_{3*1}$=[β1 ; β2 ; β3]

- Model Y= β1 + β2*X1*X2*X3 + β3*X1
- X_{n*3}=[1_{n*1} X1*X2*$X3_{n*1}$ $X1_{n*1}$] $β_{3*1}$=[β1 ; β2 ; β3]

model Y=B0+B1*X1+B2*X2 multivariate first order

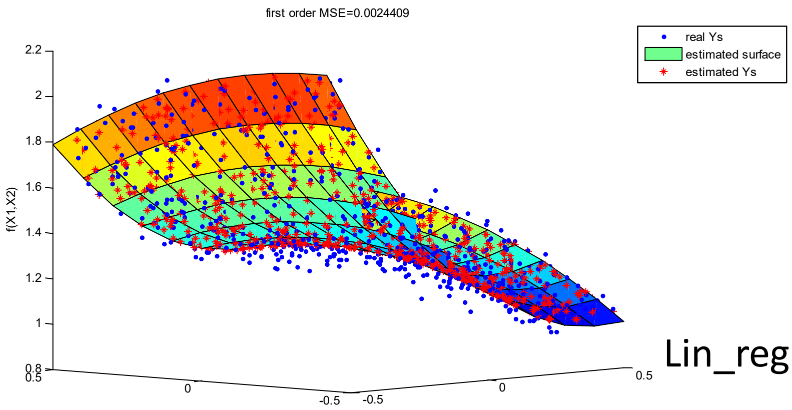


B = [3 -3 1]'

Bhat =[3.0216 -3.0439 1.0387]'

Lin_reg7.m

model
Y=B0+B1*X1+B2*X2+B3*X1^2+B4*X2^2+B5*X1*X2
multivariate second order



Lin_reg8.m

B= [1.3666 -0.2511 0.5365 -0.4343 1.4466 0.8925]'
Bhat = [1.3656 -0.2548 0.5247 -0.4571 1.4572 0.9054]'

What if we can not write linear equation system?

- (linear according to b)
- Y=b1*sin(X)
- Y=X1*b1*sin(X2)
- Y=b1*sin(X) +b2*cos(X)
- Y=b1*exp(X) +b2*cos(X)
- Y=b1*exp(X1) +b2*cos(X2)+b3*cos(X1)

- (non-linear according to b)
- Y=sin(b1*X)
- Y=b1+sin(b2*X)
- Y=exp(b1*X)
- Y=(b1*X)/(b2+X)
- Y=b1*(exp –((X–b2)^2 /(b3^2))

Non-linearity

- A parameter β of the function f appears nonlinearly if the derivative ∂f/∂β is a function of β.
- The model M(β, x) is nonlinear if at least one of the parameters in β appear nonlinearly.
- f(x)= β*sin(x), ∂f/∂β =sin(x), which is independent of β, so the model M(β,x) is linear.
- f(x)=sin(β*x), ∂f/∂β =x*cos(β*x), which is dependent of β, so the model M(β,x) is non-linear.

Non-linearity

- f(x)= β1*sin(x) + β2*cos(x) , ∂f/∂β1 =sin(x), which is independent of β1, ∂f/∂β2 =cos(x), which is independent of β2, so the model M(β,x) is linear.
- f(x)= β1+ cos(β2*x) , ∂f/∂β1 =1, which is independent of β1, but ∂f/∂β2 = -x*sin(b2*x), which is dependent of β2, so the model M(β,x) is non-linear.

What if we can not write linear equation system?

- There are two ways:
  - Transformations to achieve linearity
  - Nonlinear regression (iterative estimation)

Transformations to achieve linearity

- Some tips:
- ln(e)=1, ln(1)=0
- ln(x^r)=r*ln(x)
- ln(e^A)=A^ln(e)=A
- ln(A*B)=ln(A)+ln(B)
- ln(A/B)=ln(A)-ln(B)
- e^(A*B)=(e^A) ^B
- e^(A+B)=(e^A) *(e^B)
- e^(A-B)=(e^A) /(e^B)

Transformations to achieve linearity

- Original Y=b0*exp(b1*X)
- ln (Y)= ln(b0)+(b1*X)
- Z=ln(Y), b2=ln(b0)
- Z=b2+b1*X (linear)

- Original Y=exp(b0)*exp(b1*X)
- ln(Y)=b0+b1*X
- Z=ln(Y)
- Z=b0+b1*X (linear)

Transformations to achieve linearity

- Original Y=(b0+b1*X)^2
- sqrt(Y)= b0+b1*X
- Z=sqrt(Y),
- Z=b0+b1*X (linear)

- Original Y=1/(b0+b1*X)
- 1/Y=b0+b1*X
- Z=1/Y
- Z=b0+b1*X (linear)

Nonlinear regression (iterative estimation)

- Data={x_i , y_i } i=1..n (n= number of data points)
- y=f(β,x)
- x = n*d matrix
- y= n*1 matrix
- $r_i=y_i-f(\beta,x_i)$ r= residuals (n*1 matrix)
- β = parameters to be optimized
- $E(\beta)=\sum(r_i)$^2 i=1..n
- $\min_\beta E(\beta)$
- dE(β)/dβ = 0 (optimize E according to β)

Nonlinear regression (iterative estimation)

- dE(β)/dβ= 2*r*dr/dβ
- dr/dβ= n*d matrix = [$dr_1/d\beta_1$ $dr_1/d\beta_2$ … $dr_1/d\beta_d$
  $dr_2/d\beta_1$ $dr_2/d\beta_2$ … $dr_2/d\beta_d$
  …
  $dr_n/d\beta_1$ $dr_n/d\beta_2$ … $dr_n/d\beta_d$]
- dr/dβ is called Jacobian matrix (J)
- $\beta_{k+1}= \beta_k$-eps*dE(β)/dβ (Gradient descent)
- $\beta_{k+1}= \beta_k$-eps*J^T*r (Gradient descent)
- (d*1)=(d*1)-eps(d*n)*(n*1)

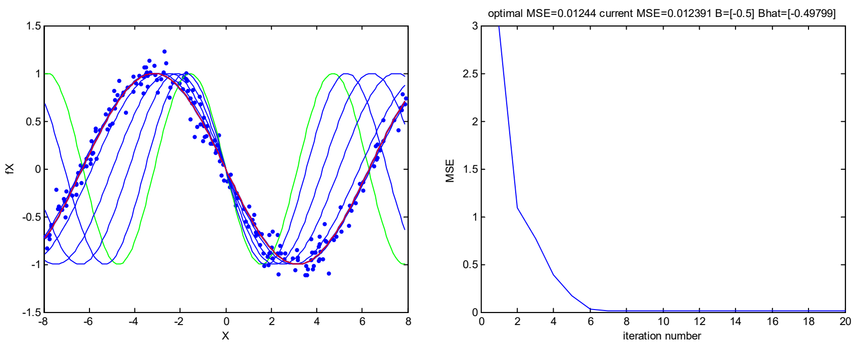
Mehmet Fatih AMASYALI Optimization Techniques Lecture Notes YILDIZ TECHNICAL UNIVERSITY COMPUTER ENG DEPT.

Nonlinear regression (iterative estimation)

- $\beta_{k+1} = \beta_k - eps*dE(\beta)/d\beta$ (Gradient descent)
- $\beta_{k+1} = \beta_k - (dE(\beta)/d\beta) / (ddE(\beta)/dd\beta)$ (Newton Raphson)
- $ddE(\beta)/dd\beta \approx J^T*J$
- $\beta_{k+1} = \beta_k - (J^T*r) / (J^T*J)$
- $\beta_{k+1} = \beta_k - inv(J^T*J)*(J^T*r)$
- $pinv(J)=inv(J^T*J)*J^T$
- $\beta_{k+1} = \beta_k - pinv(J)*r$ (Newton Raphson)
- $\beta_{k+1} = \beta_k - eps*J^T*r$ (Gradient descent)

Blue points: data
Green line: initial guess
Blue lines: iterative guesses
Red line: last guess

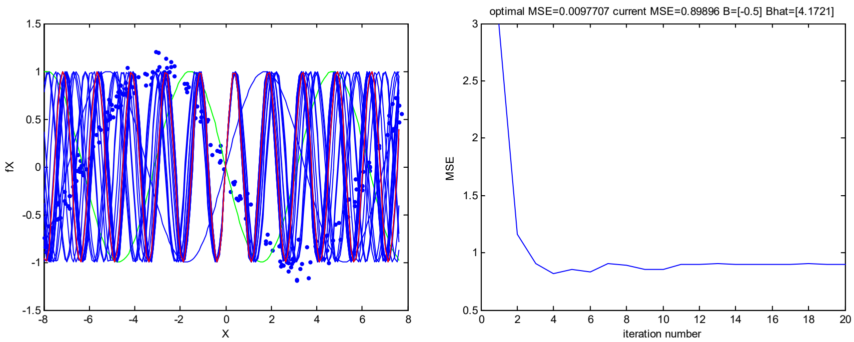
Y=sin(b1*X)



Gradient descent
Eps=0.0005
Real b=-0.5
initial b=-1

Non_lin_reg0.m

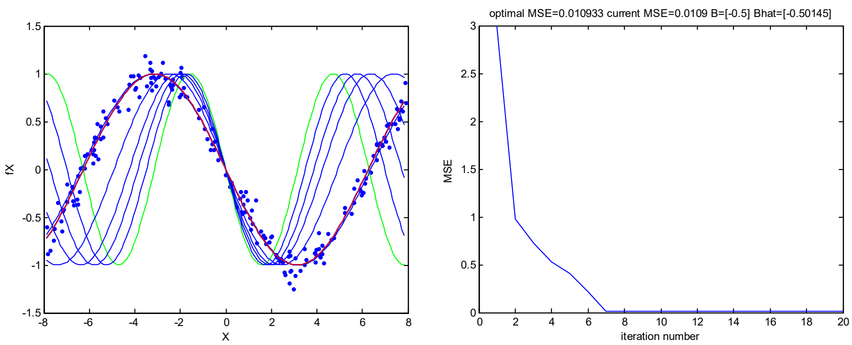
Y=sin(b1*X)



Gradient descent
Eps=0.01
Real b=-0.5
initial b=-1

Non_lin_reg0.m

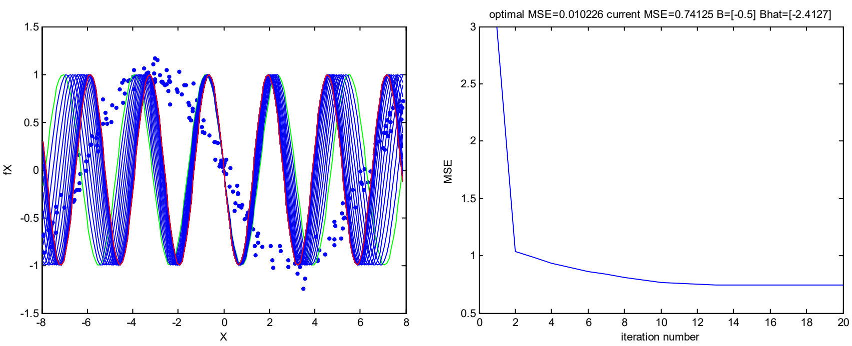
Y=sin(b1*X)



Newton raphson
Real b=-0.5
initial b=-1

Non_lin_reg0.m

Y=sin(b1*X)

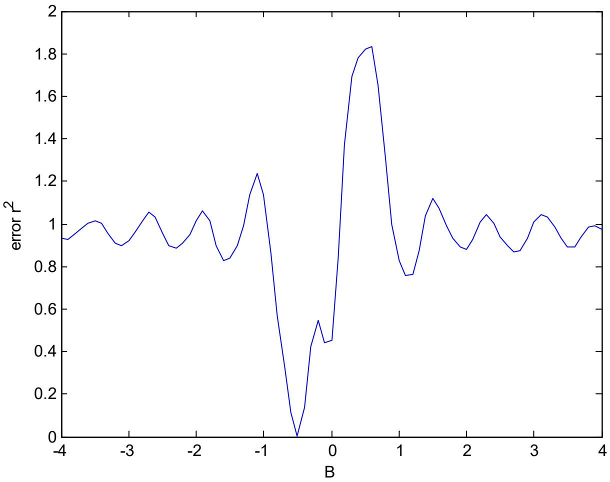


Newton raphson
Real b=-0.5
initial b=-2

Non_lin_reg0.m

Y=sin(b1*X)
b1 vs. error

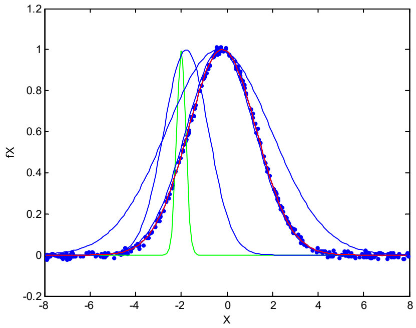
- B1=-0.5

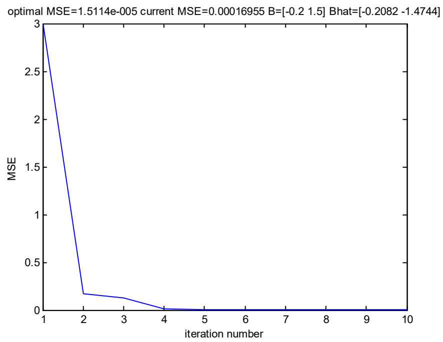


Non_lin_reg1.m

Blue points: data
Green line: initial guess
Blue lines: iterative guesses
Red line: last guess

Y=exp(-(X-B1)^2/(2*B2^2))



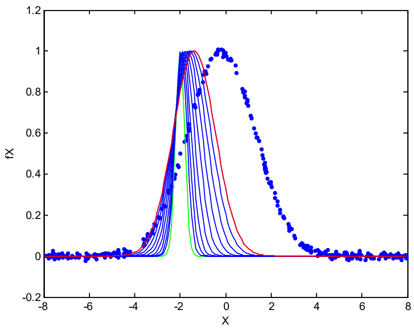


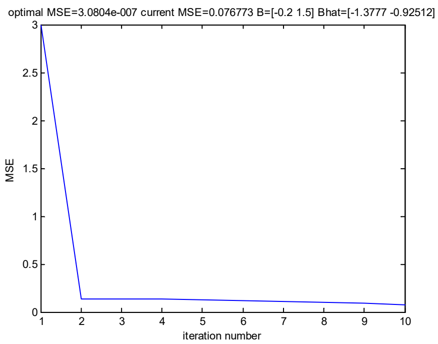
Gradient descent
Eps=0.1
Real B=[-0.2 ; 1.5]
initial B= [-2; -0.2]

Non_lin_reg2.m

Blue points: data
Green line: initial guess
Blue lines: iterative guesses
Red line: last guess

Y=exp(-(X-B1)^2/(2*B2^2))



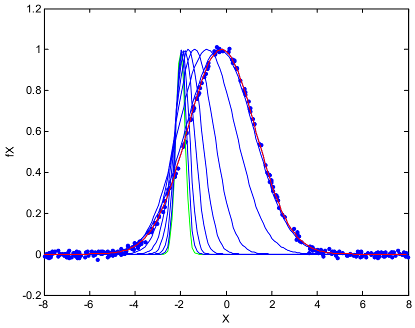


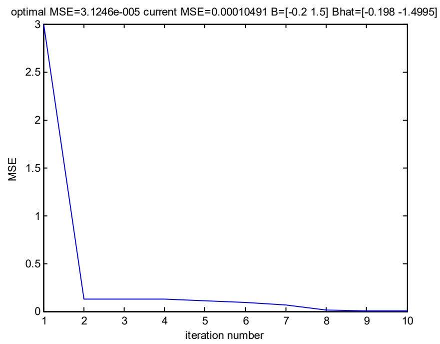
Gradient descent
Eps=0.01
Real B=[-0.2 ; 1.5]
initial B= [-2; -0.2]

Non_lin_reg2.m

Blue points: data
Green line: initial guess
Blue lines: iterative guesses
Red line: last guess

Y=exp(-(X-B1)^2/(2*B2^2))



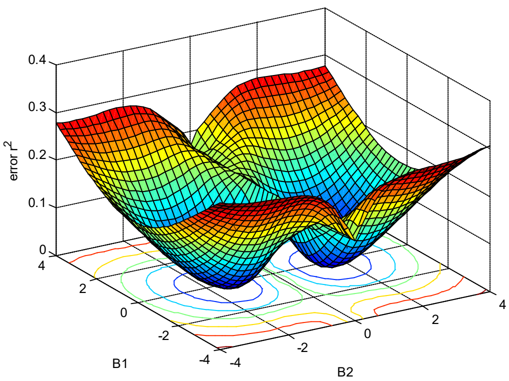


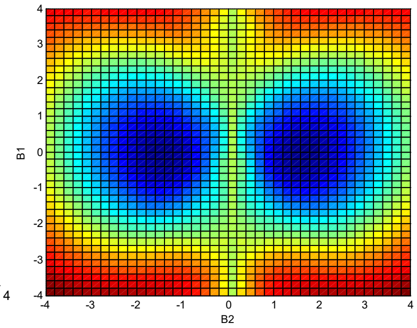
Newton raphson
Real B=[-0.2 ; 1.5]
initial B= [-2; -0.2]

Non_lin_reg2.m

Y=exp(-(X-B1)^2/(2*B2^2)) b vs. error

- N=200; B=[-0.2 ; 1.5]



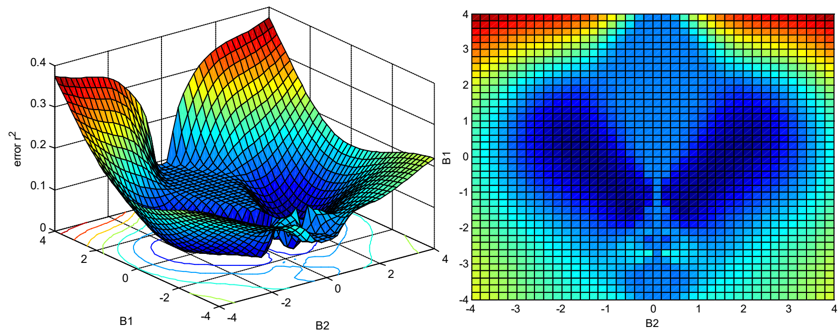


Why symmetric?

Non_lin_reg3.m

Sample size effect
Y=exp(-(X-B1)^2/(2*B2^2))
b vs. error

- N=10; B=[-0.2 ; 1.5]

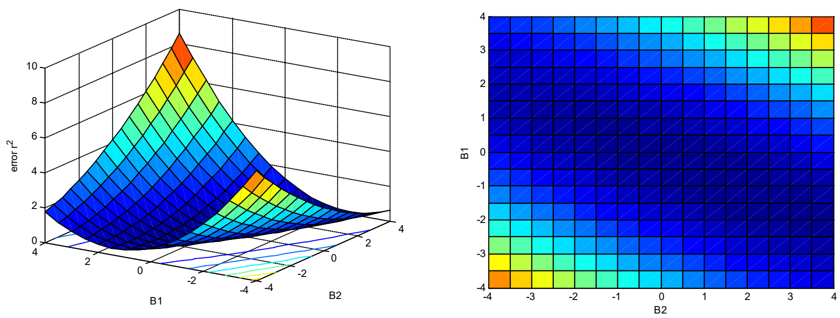


It is more difficult. Why?

Non_lin_reg3.m

Y= b1*x+b2*x^3 (linear according to B's)
b vs. error

- N=10; B=[-1 ; 1]



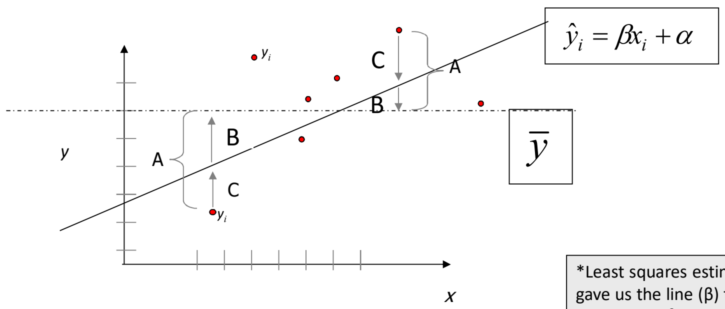
It is quadratic.

Non_lin_reg4.m

Modeling Interactions

- Statistical Interaction: When the effect of one predictor (on the response) depends on the level of other predictors.
- Can be modeled (and thus tested) with cross-product terms (case of 2 predictors):
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$

Regression Picture



*Least squares estimation gave us the line (β) that minimized C^2

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

A^2 B^2 C^2

SS_{total}
Total squared distance of observations from naïve mean of y
Total variation

SS_{reg}
Distance from regression line to naïve mean of y
Variability due to x (regression)

$SS_{residual}$
Variance around the regression line
Additional variability not explained by x—what least squares method aims to minimize

R^2=SSreg/SStotal

References

- http://www.columbia.edu/~so33/SusDev/Lecture3.pdf
- http://www.msu.edu/~fuw/teaching/Fu_Ch11_linear_regression.ppt
- http://www.holehouse.org/mlclass/10_Advice_for_applying_machine_learning.html
- http://www.imm.dtu.dk/~pcha/LSDF/NonlinDataFit.pdf
- http://fourier.eng.hmc.edu/e176/lectures/NM/node34.html
- http://fourier.eng.hmc.edu/e176/lectures/NM/node21.html
- http://www.kenbenoit.net/courses/ME104/logmodels2.pdf
- http://stattrek.com/regression/linear-transformation.aspx
- http://en.wikipedia.org/wiki/Data_transformation_(statistics)