Linear Regression

Pritha Banerjee

University of Calcutta

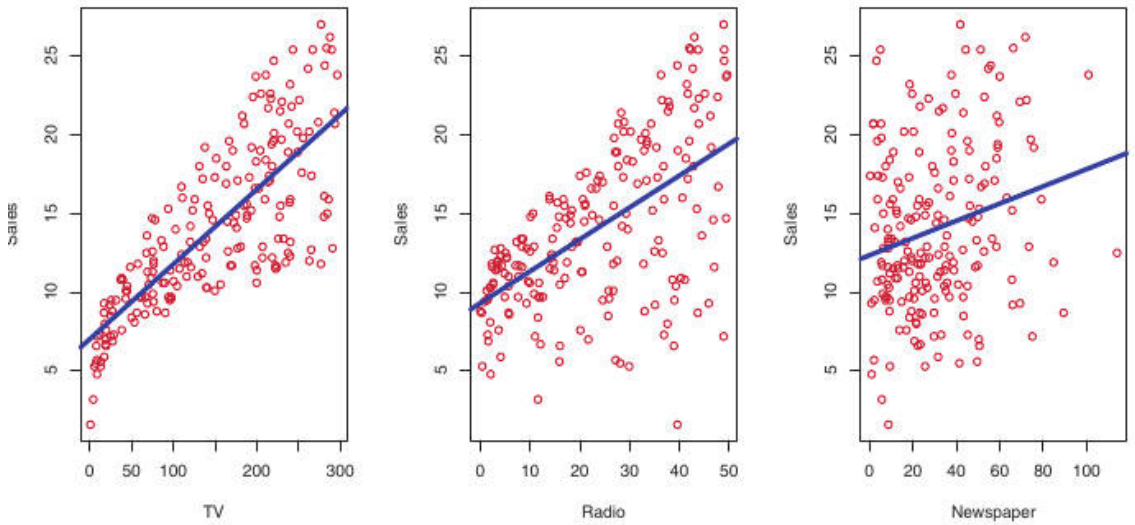
banerjee.pritha74@gmail.com

April 7, 2025

Overview

Statistical Learning

- Let Y be a quantitative response and let p be the different predictors, $X_1, X_2, \cdots, X_p$ and let us assume that there is some relationship between Y and $X = (X_1, X_2, \cdots, X_p)$, represented as $Y = f(X) + \epsilon$
- Here f (systematic information about Y) is some fixed but unknown function of $X_1, \cdots, X_p$, and ϵ is a random error term, which is independent of X and has **mean zero**.
- **Estimate f:** Want to estimate f for **prediction** and **inference**



Statistical Learning: Prediction

- Given X, predict Y as $\hat{Y} = \hat{f}(X)$, $\hat{}$ denotes an estimate
- Variability associated with ϵ affects the accuracy of prediction of Y, known as the **irreducible error**. No matter how well we estimate f , we cannot reduce the error introduced by ϵ
- For a given $\hat{f}$ and X, $\hat{Y} = \hat{f}(X)$, then $E(Y - \hat{Y})^2 = E(f(X) + \epsilon - \hat{f}(X))^2 = (f(X) - \hat{f}(X))^2 + Var(\epsilon) =$ Redicible + Irreducible; **Note** that mean of ϵ is zero, thus $E(\epsilon)^2 = Var(\epsilon)$
- Minimize the reducible error; however irreducible error will produce an upper bound on accuracy of prediction of Y

Statistical Learning: Inference

- understand how Y changes as a function of $X_1, \cdots, X_p$
- Identifying the few important predictors among a large set of possible variables
- The relationship between the response and each predictor
- Can the relationship between Y and each predictor be adequately sum-marized using a linear equation, or is the relationship more complicated?

simple linear models allow for relatively simple and interpretable inference, but may not yield as accurate predictions as others non-linear model can potentially provide quite accurate predictions for Y , but this comes at the expense of a less interpretable model

How to estimate f?

- Let x_{ij} represent the value of the j^{th} predictor, or input, for observation i, where $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, p$. Correspondingly, let y_i represent the response variable for the i^{th} observation. Then training data consist of $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ where $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$.
- Goal is to apply a statistical learning method to the training data in order to estimate the unknown function f. i.e, find function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X, Y)

Approaches:

- **Parametric:** Estimate a set of **parameters** to obtain an estimate f
- **Non-parametric:** no parameters are estimated

How to estimate f? Parametric approach

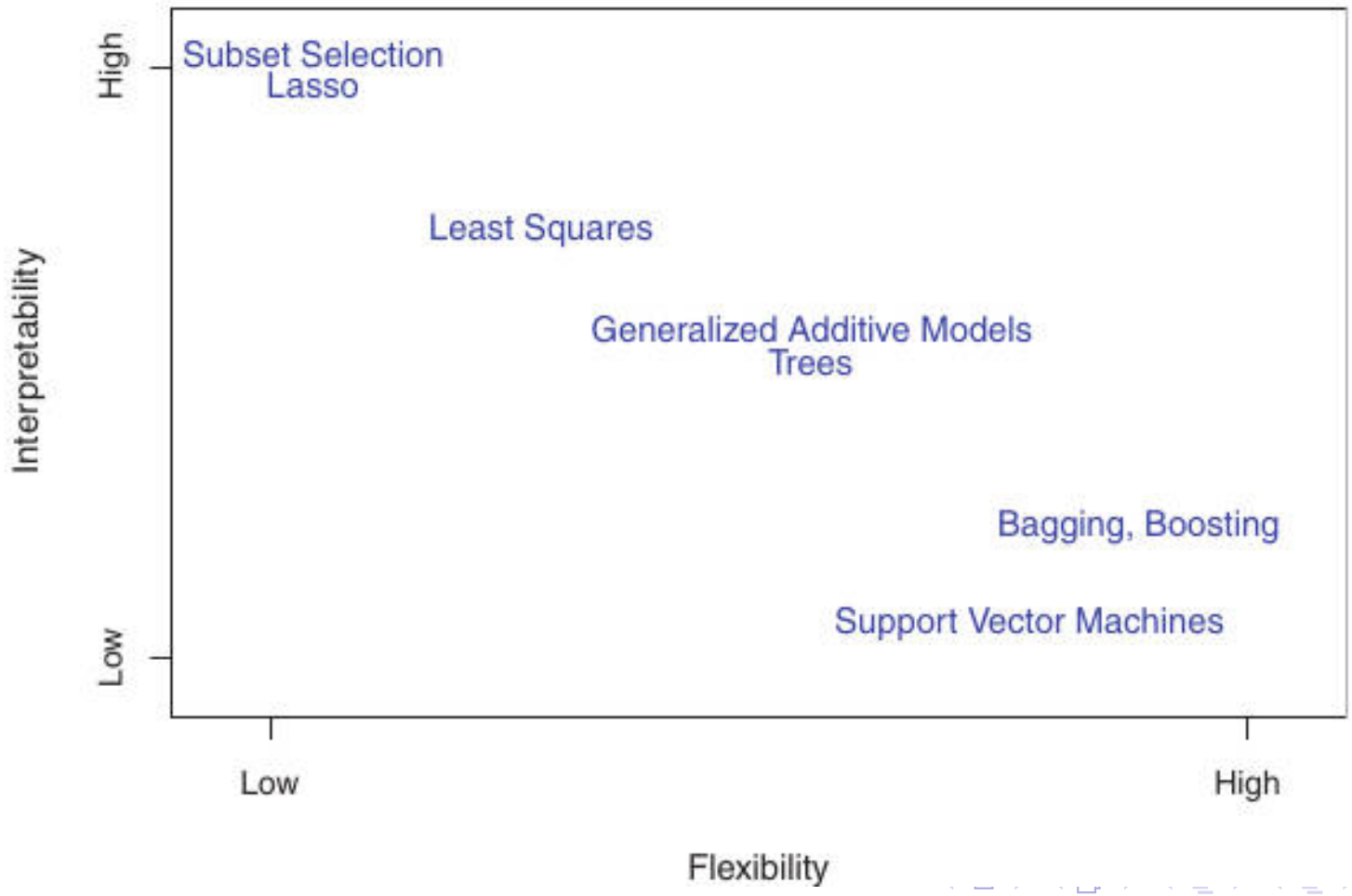
- Make an assumption about the form / shape of f,such as linear; $f(X) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$. Here **estimate** $p+1$ coefficient $\beta_0, \cdots, \beta_p$
- Use the training data to **fit or train the model** using most commonly **ordinary least square** method; obtain β_i in $f(X) \approx \beta_0 + \sum_{i=1}^{p} \beta_i X_i$.
- However the model may not match the true f; to overcome the problem, may choose to use different functional forms that leads to estimation of greater number of parameters.
- Complex model may lead to **overfitting the data**, that means they follow the errors, or noise, too closely

How to estimate f? Non-Parametric approach

- No functional form is assumed about f
- Seek an estimate of f that gets as close to the data points as possible without being too rough
- Better chance of ccurately fitting a wider range of possible shapes for f, as no functional form is assumed
- However, needs much large number of observations than parametric approaches to estimate f
- Chance of overfitting is more, which affects prediction for new observation.

Trade-off between prediction accracy and model interpretability

Restricted models, such as linear model are much more interpretable, than complex non=parametric model



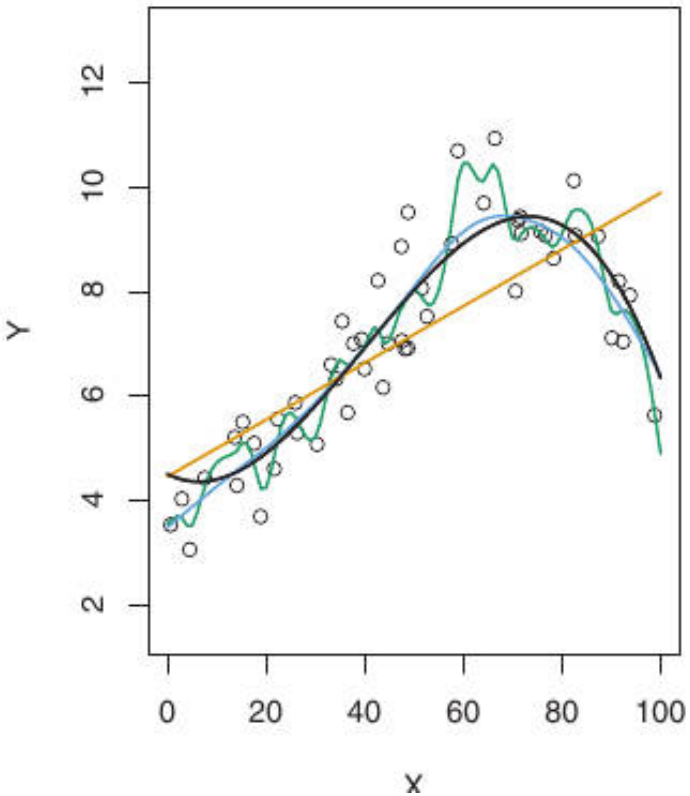
Assessing Model Accuracy

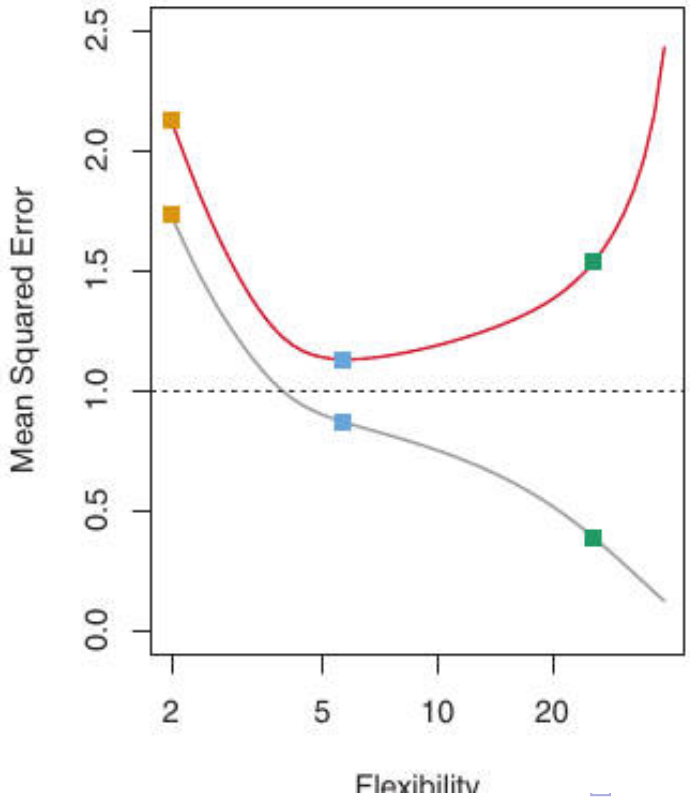
Selecting a statistical learning procedure for a specific data set: no one method dominates all others over all possible data sets. **a) Measuring quality of fit; b) Bias-Variance trade-off**

- **Measuring quality of fit:** Mean Squared Error (MSE): how close is predicted response to the true response of a particular observation: $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$, $\hat{f}(x_i)$ is predicted response for i^{th} observation
- For **training** observations $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, yn)\}$, we obtain the estimate $\hat{f}$, then compute $\hat{f}(x_1), \hat{f}(x_2), \cdots, \hat{f}(x_n)$. Typically, MSE_{train} will be small for this.
- For an unknown **test** observation x_0, y_0 (unused observation in training set) ,how close is $\hat{f}(x_0)$ to y_0? ie, we want lowest MSE_{test}
- We want $Avg(\hat{f}(x_0) - y_0)$, average square prediction error is as small as possible for a large number of test observations.
- Small MSE_{train} may not lead to small MSE_{test}

Measuring Quality of fit

- black: true f, grey: MSE_{train}, red: MSE_{test}, dashed line: minimim possible MSE_{test}, $Var(\epsilon)$, irreducible error
- As flexibility (**degrees of freedom**) of the learning method increases,observe a monotone decrease in the MSE_{train} and a U-shape in the MSE_{test}
- Linear regression is has restricted flexibility with two degrees of freedom.
- Blue curve minimizes MSE_{test}, as it it appears to estimate f the best.
- When a method yields a small MSE_{train} but a large MSE_{test}, it is **overfitting** the data.





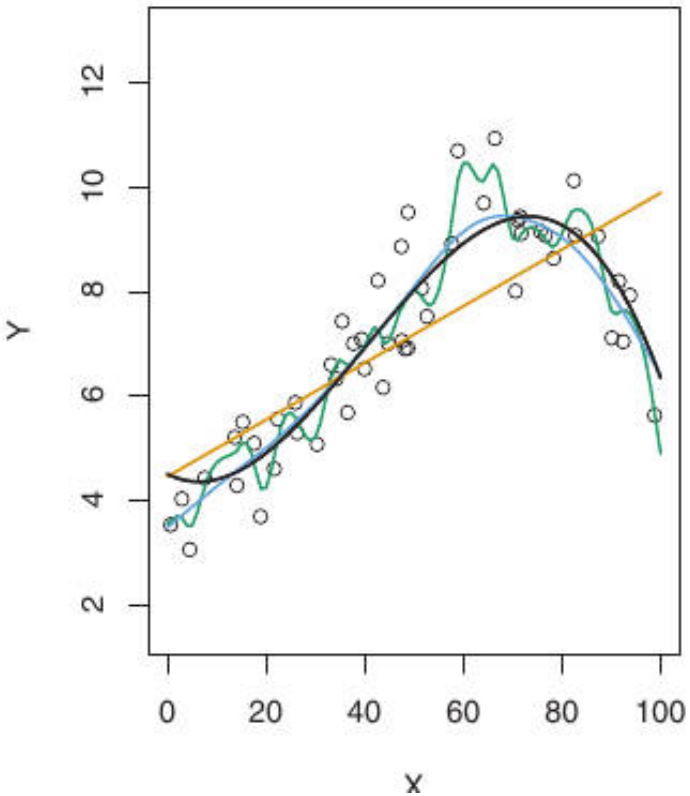
Assessing Model Accuracy: Measuring quality of fit

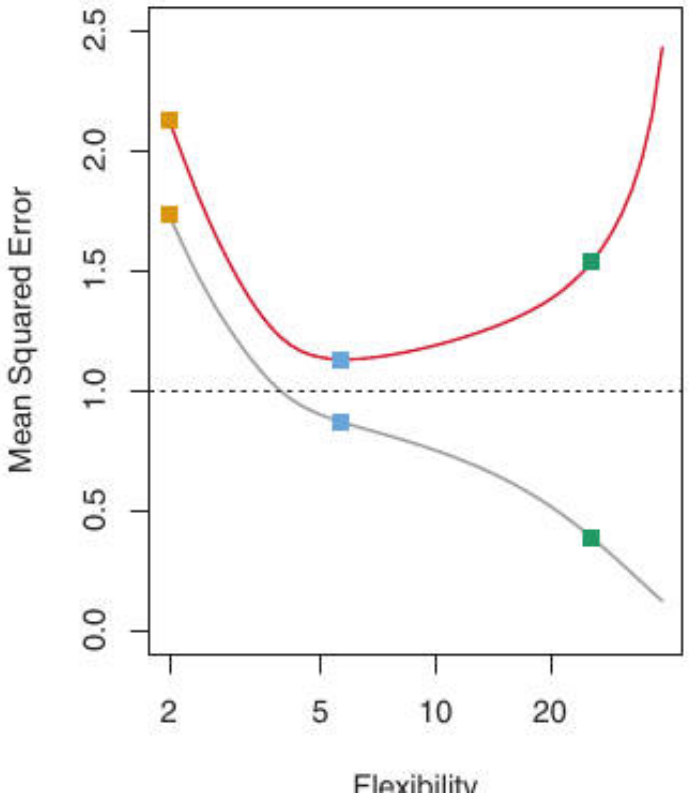
- **Expected** MSE_{test}, for a given value x_0 is: $E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$; Average test MSE is obtained by repeated estimation of f using a large number of training sets, and testing each at x_0.
- **Overall Expected** MSE_{test} is computed by averaging $E(y_0 - \hat{f}(x_0))^2$ over all possible values of x_0 in the test data set.
- **Variance of a learning method:** amount by which f would change if we estimated it using a different training data set. ie. $\hat{f}$ obtained for different training set should not vary too much; high variance means small change in training data results in large changes in $\hat{f}$
- **More flexible** methods causes **high variance**
- **Bias of a learning method:** error that is introduced by approximating a real-life problem to a simpler problem; assumption that a linear relationship exists between Y and X_is.
- **More flexible** methods results in **low bias**
- Choose a learning method that has low variance and low bias to minimize expected MSE_{test}

Bias-Variance Tradeoff

- Choose a learning method that has low variance and low bias to minimize expected MSE_{test}
- Since $Var(\hat{f}(x_0))$ and $[Bias(\hat{f}(x_0))]^2$ are non-negative, hence expected MSE_{test} can not lie below $Var(\epsilon)$, irreducible error.
- **More flexible methods**, the variance will **increase** and the **bias** will **decrease**.
- Relative rate of change of these two quantities determines whether the MSE_{test} increases or decreases; increasing the flexibility of a class of methods, bias tends to initially decrease faster than the variance increases. Consequently, the **expected test MSE declines**.
- However, at some point increasing flexibility has little impact on the bias but starts to significantly increase the variance.Then MSE_{test} increases.

Example 1


Y
X
0
20
40
60
80
100
2
4
6
8
10
12


Mean Squared Error
Flexibility
2
5
10
20
0.0
0.5
1.0
1.5
2.0
2.5

Example 2

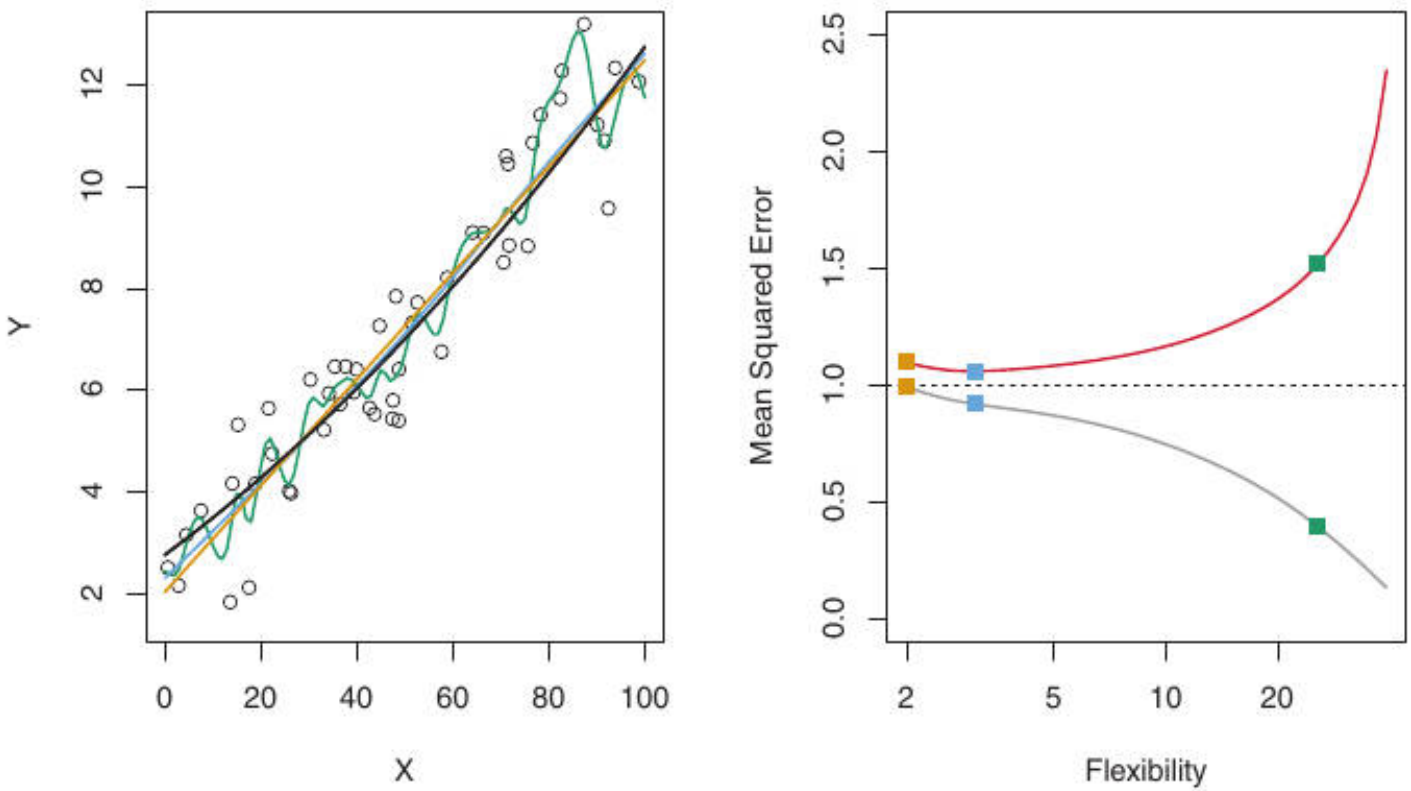


FIGURE 2.10. *Details are as in Figure 2.9, using a different true f that is much closer to linear. In this setting, linear regression provides a very good fit to the data.*

Example 3

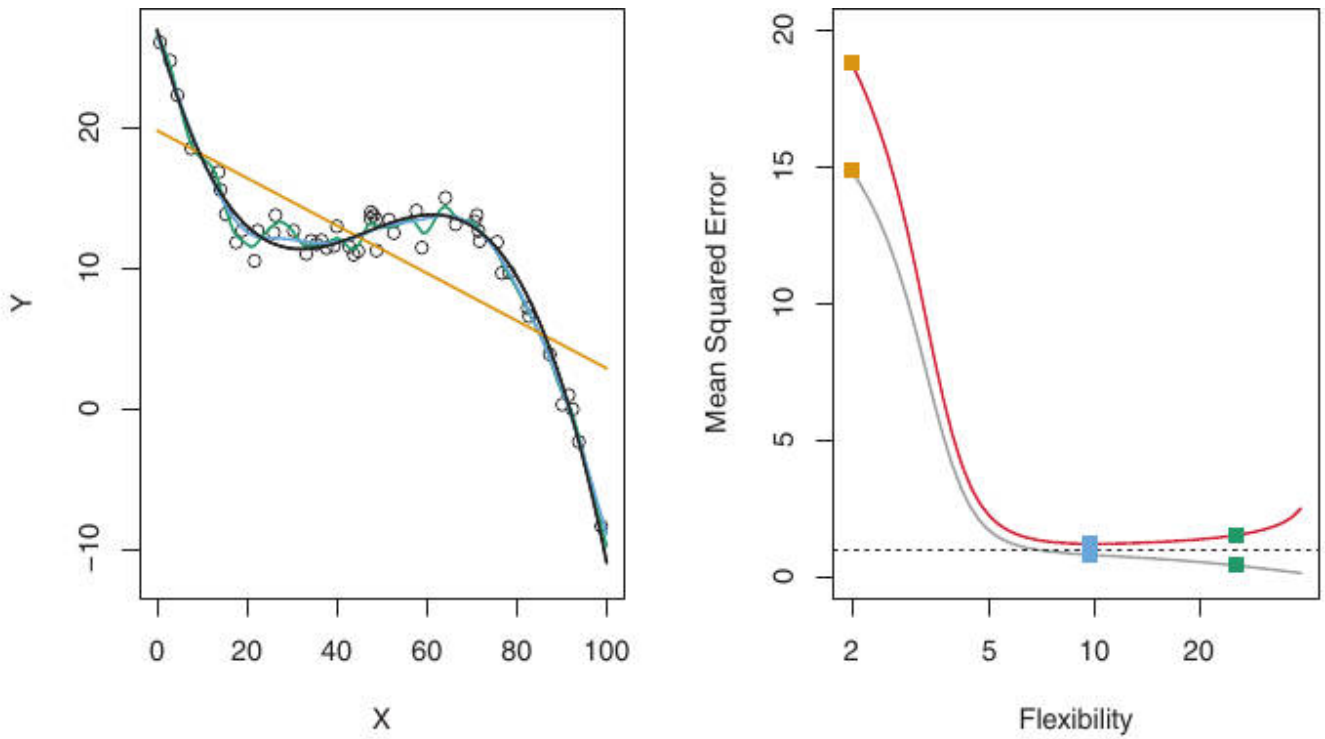
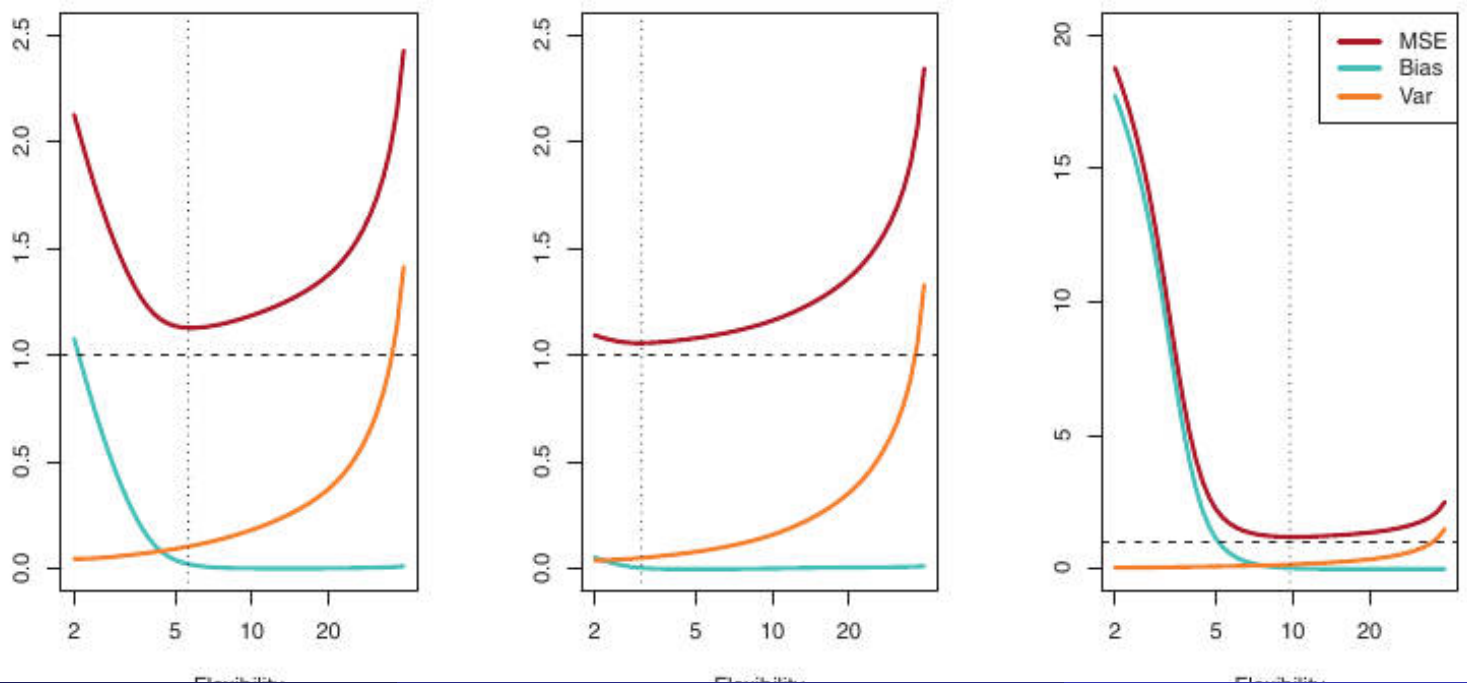


FIGURE 2.11. *Details are as in Figure 2.9, using a different f that is far from linear. In this setting, linear regression provides a very poor fit to the data.*

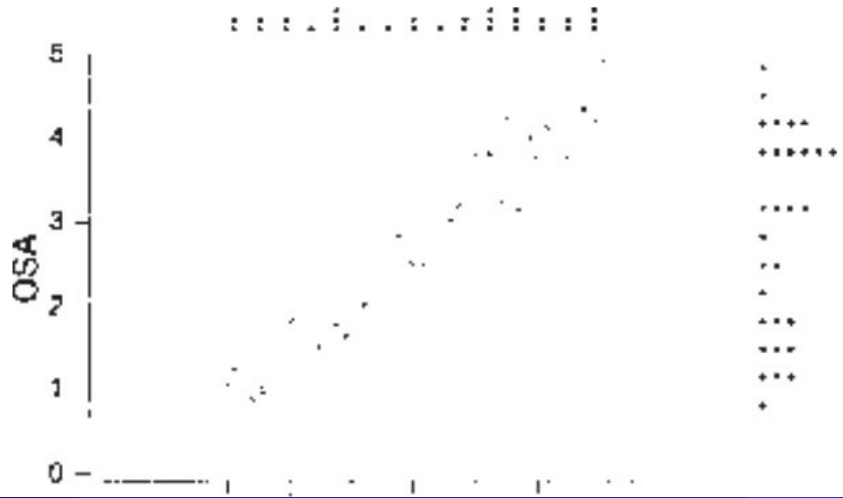
Bias-Variance Tradeoff

- In all three cases, the **variance increases** and the **bias decreases** as the method's **flexibility increases**. However, the flexibility level corresponding to the optimal test MSE differs considerably among the three data sets, because the squared bias and variance change at different rates in each of the data sets.
- **Ex 1:** bias initially decreases rapidly, sharp decrease in the **expected test MSE**.
- **Ex 2:** true f is close to linear, so a small decrease in bias as flexibility increases, and the test MSE only declines slightly before increasing rapidly as the variance increases.
- **Ex 3:** as flexibility increases, a dramatic decline in bias because the true f is very non-linear; also very little increase in variance as flexibility increases. Consequently, the test MSE declines substantially before experiencing a small increase as model flexibility increases.



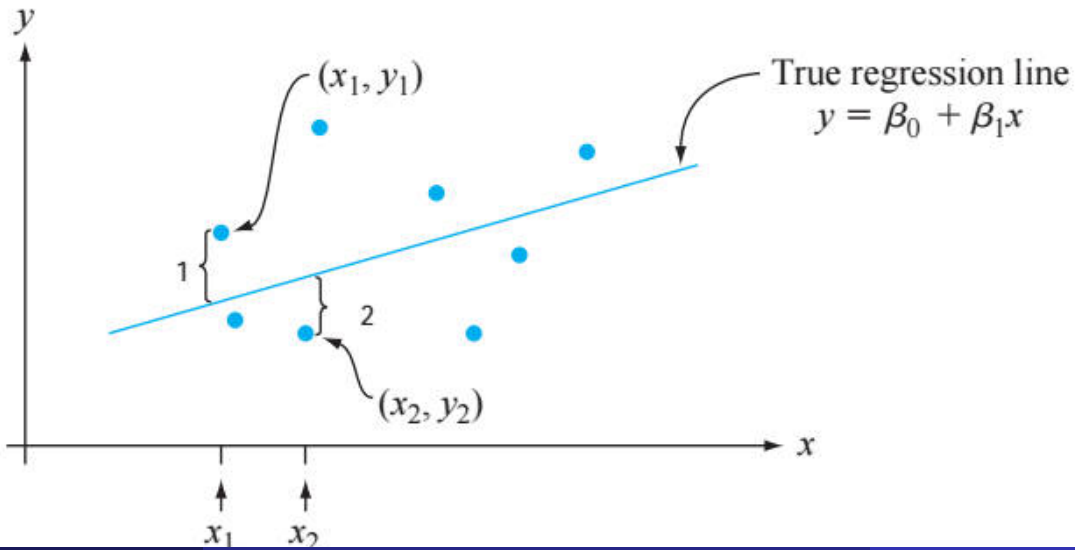
Simple Linear Regression model

- In $y = \beta_0 + \beta_1 x$, if the x and y are not deterministically related, then for a fixed value of x, there is uncertainty in the value of y.
- **independent/ predictor explanatory/ regressor/ input variable:** variable x fixed by experimenter
- **dependent/ response/ predicted/ regressand/ output variable:** For fixed x, Y is the random variable and y the observed value
- Let $x_1, x_2 \cdots x_n$ denote values of independent variable for which observations are made, and let Y_i and y_i , respectively, denote the random variable and observed value associated with x; bivariate data is n pairs $(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)$.



Linear Probabilistic model (1)

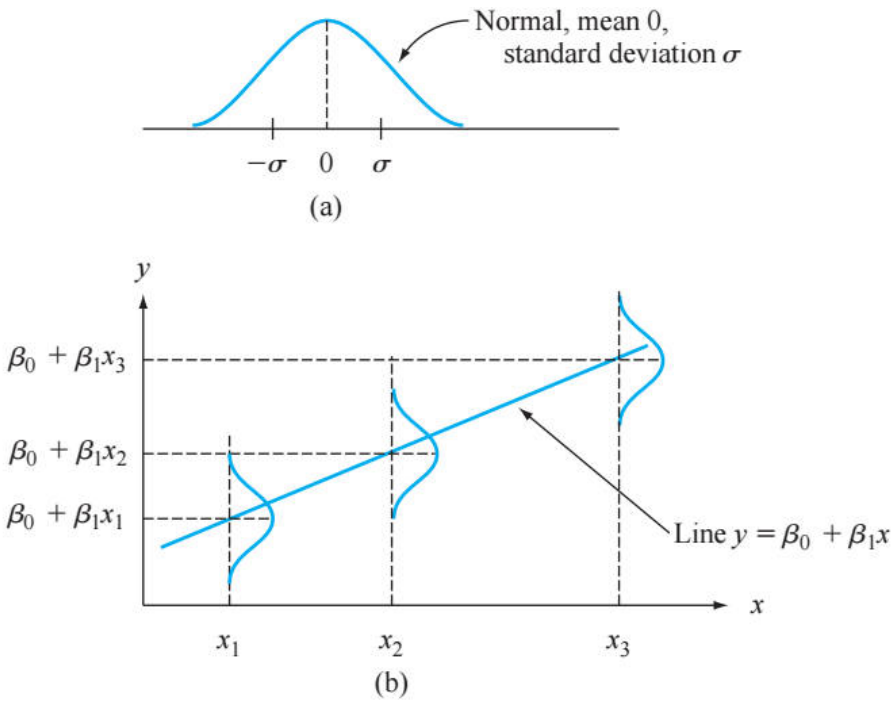
- Generalization of $y = \beta_0 + \beta_1 x$ to a probabilistic model assumes that the E(Y) is a linear function of x, but for fixed x the variable Y differs from its expected value by a random amount.
- **Simple Linear Regression Model:** There are parameters β_0, β_1, and σ^2, such that for any fixed value of x, the dependent variable is a random variable related to x through the model equation $Y = \beta_0 + \beta_1 x + \epsilon$, where ϵ is a random variable (**random deviation/ random error term**), assumed to be normally distributed with $E(\epsilon) = 0$ and $V(\epsilon) = \sigma^2$



Linear Probabilistic model (2)

- Implication of $Y = \beta_0 + \beta_1 x + \epsilon$: Let x^* is a particular value of x, then
  - $E(Y|x^*) = \mu_{Y.x^*}$ = Expected value of Y when x has value x^*
    $= E(\beta_0 + \beta_1 x^* + \epsilon) = \beta_0 + \beta_1 x^* + E(\epsilon) = \beta_0 + \beta_1 x^*$
- $V(Y|x^*) = \sigma^2_{Y.x^*}$ = Variance of Y when x has value x^*
  $= V(\beta_0 + \beta_1 x^* + \epsilon) = V(\beta_0 + \beta_1 x^*) + V(\epsilon) = 0 + \sigma^2 = \sigma^2$
- True regression line $y = \beta_0 + \beta_1 x$ is thus the line of mean values;
- height above any particular x value is the expected value of Y for that value of x
- The slope β_1 is the expected change in Y associated with a 1- unit increase in the value of x.
- The amount of variability in the distribution of Y values is the same at each different value of x(homogeneity of variance).
- For fixed x, Y has normal distribution

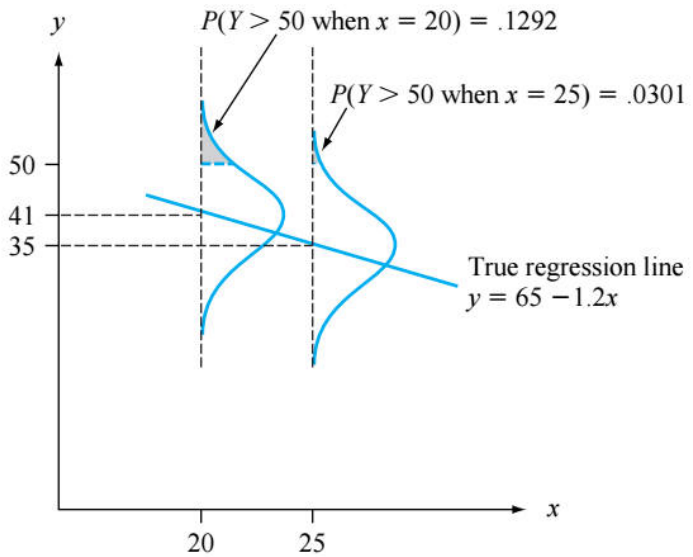
Linear Probabilistic model (3)



- σ^2 determines extent of each normal curve spreading out about its mean value (the height of the line).
- small $\sigma^2 \Rightarrow (x, y)$ falls near true regression line, otherwise it falls far off from the line.

Linear Probabilistic model (4)

- Example: x : applied stress, y : time to failure follows $y = 65 - 1.2x$ and $\sigma = 8$
- for all (x, y) points of population, magnitude of deviation from true regression line is about 8.
- For a fixed $x = 20$, $E(Y|x^*) = 65 - 1.2.20 = 41$
- So, $P(Y > 50|x = 20) = P(Z > \frac{50-41}{8}) = 1 - \Phi(1.13) = .1292$
- Even though we expected Y to decrease when x increases by 1 unit, it is not unlikely that the observed Y at $x+1$ will be larger than the observed Y at x.



Principle of Least Squares

- From sample data $(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)$ the model parameters β_0, β_1, σ^2 and the true regression line itself can be estimated.
- y_i is the observed value of Y_i, where $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and the n deviations are independent rv's and $Y_1, Y_2, \cdots, Y_n$ are independent as ϵ_i's are independent.
- **Principle of Least Squares:** The vertical deviation of (x, y) from line $y = b_0 + b_1 x$ is $y_i - (b_0 + b_1 x_i)$. Then the sum of squared vertical deviations from $(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)$ to $y = b_0 + b_1 x$ is

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

The point estimates of b_0 and b_1, denoted by β_0 and β_1 are the **least squares estimates**, that minimizes $f(b_0, b_1)$; the estimated **regression line** or **least squares line** is $y = \hat{\beta}_0 + \hat{\beta}_1 x$

Computing $\hat{\beta}_0$, $\hat{\beta}_1$ (1)

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum 2(y_i - b_0 - b_1 x_i)(-1) = 0 - (1)$$

$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum 2(y_i - b_0 - b_1 x_i)(-x_i) = 0 - (2)$$

Rearrangement gives the following system of **normal equations**:

$$nb_0 + (\sum x_i)b_1 = \sum y_i$$

$$b_0 = \frac{(\sum y_i - b_1 \sum x_i)}{n} = \bar{y} - b_1 \bar{x} - (3)$$

$$(\sum x_i)b_0 + (\sum x_i^2)b_1 = \sum x_i y_i$$

$$(\sum x_i)(\bar{y} - b_1 \bar{x}) + (\sum x_i^2)b_1 = \sum x_i y_i$$

$$\sum x_i y_i - b_1 \sum x_i^2 - \sum x_i(\bar{y} - b_1 \bar{x}) = 0$$

$$\sum x_i \left[y_i - \bar{y} + b_1 \bar{x} - b_1 x_i\right] = 0$$

$$\sum (y_i - \bar{y}) + b_1 \sum (\bar{x} - x_i) = 0$$

$$b_1 = \frac{\sum (y_i - \bar{y})}{\sum (x_i - \bar{x})}$$

$$b_1 = \frac{\sum (y_i - \bar{y}) . \sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} - (4)$$

Computing $\hat{\beta}_0$, $\hat{\beta}_1$ (2)

$$\hat{\beta}_0 = b_0 = \bar{y} - b_1\bar{x} \cdots \quad from \quad (1)$$

$$\hat{\beta}_1 = b_1 = \frac{\sum(y_i - \bar{y}).\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} - (5)$$

$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} - (6)$$

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} - (7)$$

Use the estimates for

- a point estimate of the expected value of Y when $x = x^*$ or
- a point prediction of Y value that will result from a single new observation made at $x = x^*$. The **danger of extrapolation** is that the fitted relationship (a line here) may not be valid for x values much beyond the range of the sample data.

Other forms of function f

- use a regression function involving more than a single independent variable (multiple linear regression)
  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$
- replace X by a non linear function of X (non linear regression)
  $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p + \epsilon$

Given N **training data** $X^T = (X_1, X_2, \cdots, X_p$ and y, where $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$, $i = 1, \cdots, N$, determine $\beta = (\beta_0, \beta_1, \cdots, beta_p)^T$ minimizing residual sum of squares (least squares)

Estimating the β_i, the regression coefficients (1)

- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p$, p coefficients of regression, y is predicted as: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$
- p parameters(regression coefficients) estimated using **least square** approach. Minimize
  $RSS = \sum(y_i - \hat{y}_i)^2 = \sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - \cdots - \hat{\beta}_p x_p)^2$
- Let mean shifted values

$$y_{n\times 1} = \begin{bmatrix} y_1 - \bar{y} & y_2 - \bar{y} & \cdots & y_n - \bar{y} \end{bmatrix}^\top$$

$$X_{n\times p} = \begin{bmatrix} x_{1,1} - \bar{x}_1 & x_{2,1} - \bar{x}_2 & \cdots & x_{p,1} - \bar{x}_p \\ x_{1,2} - \bar{x}_1 & x_{2,2} - \bar{x}_2 & \cdots & x_{p,2} - \bar{x}_p \\ \cdots & \cdots & \cdots & \cdots \\ x_{1,n} - \bar{x}_1 & x_{2,n} - \bar{x}_2 & \cdots & x_{p,n} - \bar{x}_p \end{bmatrix}$$

- Mean shifting helps excluding β_0; Let

$$\beta_{p\times 1} = \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_p \end{bmatrix}^\top$$

$$\epsilon_{n\times 1} = \begin{bmatrix} \epsilon_1 & \epsilon_2 & \cdots & \epsilon_p \end{bmatrix}^\top$$

Estimating the β_i, the regression coefficients (2)

- **Without mean shifting:** Let X be $N \times (p+1)$ matrix with **each row an input vector with 1 in first position**, and y be the N vector of outputs in training set.
- The Linear model in Matrix form $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ with $E(\epsilon) = 0$ and $V(\epsilon) = \sigma^2 I$, I : identity matrix, ϵ_i and ϵ_j are uncorrelated for $i \neq j$
- Minimize $RSS(\beta) = \epsilon^T \epsilon = (y - X\beta)^T.(y - X\beta)$
- by differentiating and equating to zero, gives normal linear equations

$$RSS(\beta) = (y^T - \beta^T X^T)(y - X\beta) \tag{1}$$

$$= y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta \tag{2}$$

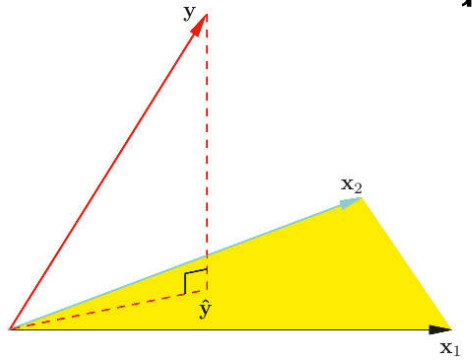
$$\frac{\partial RSS}{\partial \beta} = -X^T y - y^T X + 2\beta X^T X = -2X^T(y - X\beta) \tag{3}$$

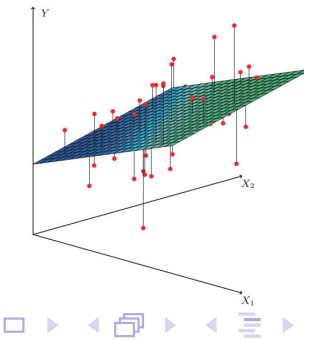
$$\frac{\partial_2 RSS}{\partial \beta \partial \beta^T} = 2X^T X \tag{4}$$

- X has full column rank p, and thus $(X^T X)$ is positive definite and thus it's inverse exists. Set

Geometric and algebraic views of linear regression

- $\frac{\partial RSS}{\partial \beta} = -2X^T(y - X\beta) = 0$
- **The coefficient vector:** unique solution $\hat{\beta} = (X^T X)^{-1} X^T Y$; (for mean shifted values: $\hat{\beta}_0 = \bar{y} - \bar{X}^T \hat{\beta}$.)
- The predicted values at an input vector x_0 given by $\hat{f}(x_0) = (1 : x_0)^T \hat{\beta}$
- The fitted values at training inputs are $\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y$, $\hat{y}_i = \hat{f}(x_i)$, $H = X(X^T X)^{-1} X^T$ is called hat matrix as it puts the hat on y. H is called projection matrix
- Let column vectors of X be x_0, x_1, x_p, with $x_0 = 1$, they span column space of X, a subspace of $\mathbb{R}^N$
- Minimize $RSS(\beta) = ||y = X\beta||^2$ implies choosing a $\hat{\beta}$ such that $y - \hat{y}$ is orthogonal to this subspace (first derivative equating to zero).





When columns of X are not linearly Independent

- X is not full rank, $\hat{\beta}$ not unique, fitting can be controlled by regularization
- Use truncated Singular Value Decomposition (SVD) of $X = U\Sigma V^T$ to compute $(X^T X)^{-1}$ (Moore-Penrose Pseudoinverse). i.e, consider only the singular values and corresponding vectors that correspond to the actual rank of X
- if X has rank k, use the first k columns of U, the first k rows of V^T, and the first k singular values in Σ to calculate the pseudoinverse and the hat matrix.