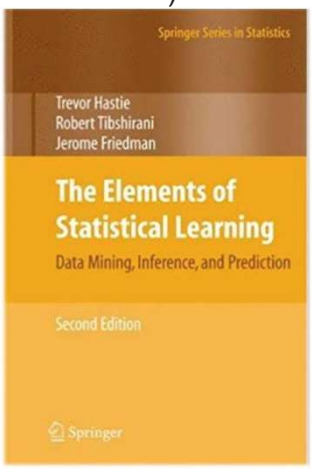
Lecture 2: Regression and statistical interpretation

Wengang Mao (Marine Technology)
Department of Mechanics and Maritime Sciences,
Chalmers University of Technology,
Goteborg, Sweden

1

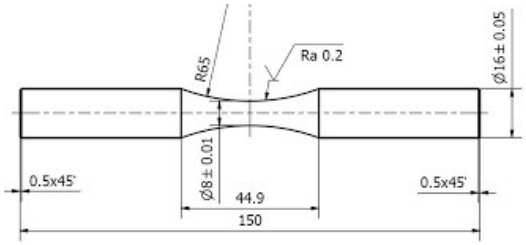
Contents of this lecture

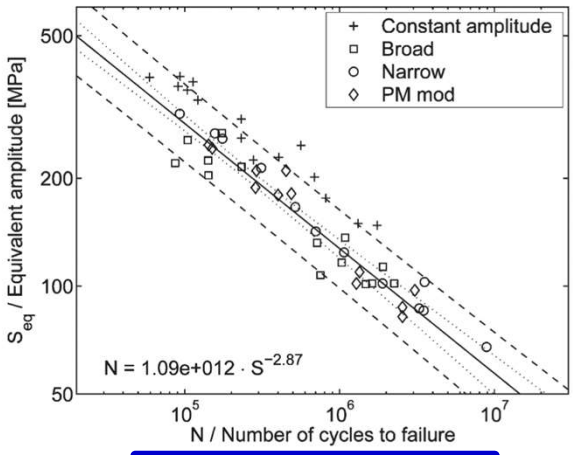
- Basics: Another way to look at regression analysis (Biased VS variance)
- Single variable regression (statistical learning)
  - Simple linear model (a special case of KNN)
  - KNN and kernel smooth
- Multivariate linear regression models
- Confidence of the regression linear models
- Introduction of the first assignment project



2

Why statistical learning? (for explicit formulas)



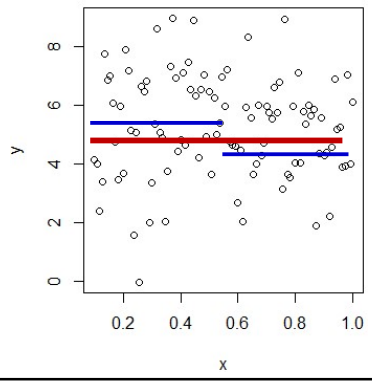


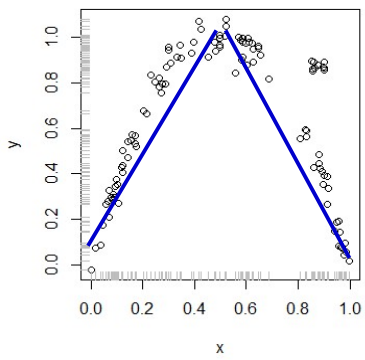
$$\log(N) = a - k\log(\Delta\sigma^{eq})$$

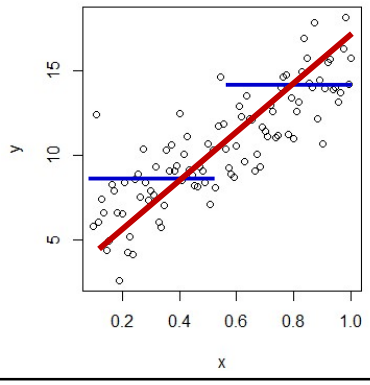
3

Regression basics

- Let's look at the simple case with one variable, i.e., ***Y = f (X)***. We can make the conditions of ***X*** by dividing into various groups
  - If the condition is made on each individual $X = x_i$ *as* $\mu(x_i) = y_i = E[Y|X = x_i]$*: high roughness, interpolation*
  - If the condition is made by splitting **X** into *j* groups: *several piecewise models* $\mu_j(X)$ *(Spline, KNN, kernel, MA)*
  - If the condition is made for the entire set of **X**, then $\mu(x)$ is a smooth model (linear, polynomial, ANN, Boost…)







4

Regression basics (mathematical def.)

- Data (*n* observations ***X***, ***Y***) for the regression (supervised statistical learning)
  - Inputs: *p* independent variables (RVs) $\boldsymbol{X} = [X_1, X_2, \ldots, X_p]$
  - Outputs: dependent random variable ***Y***
- Prediction of outputs $\hat{\boldsymbol{Y}}$

$$\boldsymbol{X}^T = [X_1, \quad X_2, \quad \ldots \quad X_p] = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{p1} \\ x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix}$$

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \hat{\boldsymbol{Y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = f(Y|\boldsymbol{X})$$

- Model regression cost/loss function:

$$MSE(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = E[(Y - f(X))^2]$$

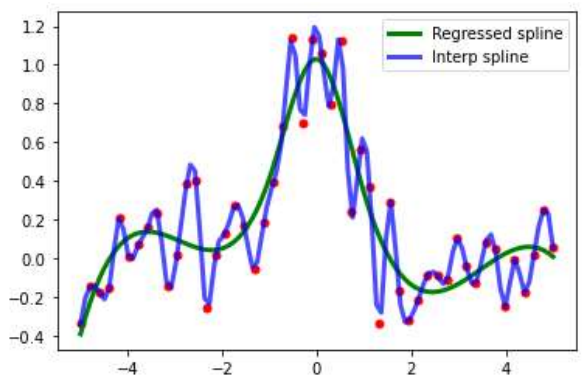
How to choose the best model for the prediction $\hat{\boldsymbol{Y}} = f(Y|\boldsymbol{X})$?

- Cost/loss function
- Objective of modelling
- Constraints

5

Basis: Is a complex model always better (1)?

- Roughness VS smoothness (interpolation VS regression)



NB: Spline fitting/regression is to choose "optimal" number of data points as knots to estimate the splines' parameters!

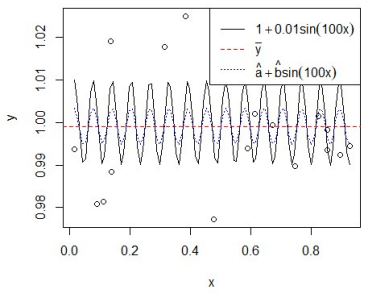
```
# Generate data with uncertainties
import numpy as np
x = np.linspace(-5, 5, 50)
y = np.exp(-x**2) + 0.2 *
np.random.randn(50)

# Spline regression
from scipy.interpolate import
make_lsq_spline, BSpline
t = [-1, 0, 1]
k = 3
t = np.r_[(x[0],)*(k+1),
        t,
        (x[-1],)*(k+1)]
spl = make_lsq_spline(x, y, t, k)

# Spline interpolation
from scipy.interpolate import
make_interp_spline
spl_i = make_interp_spline(x, y)
```


6

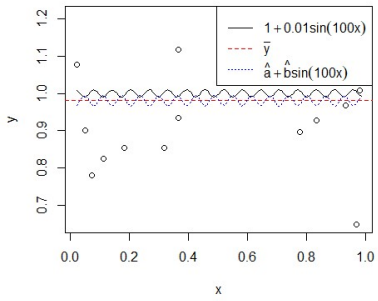
Basis: Is a complex model always better (2)?

- We have measured a series of data (x_i, y_i) from a "real" model $Y = 1 + 0.01\sin(100 * X) + \varepsilon$, where ε is the measurement noise. When ε varies in different ways, different models may be more suitable for the data.

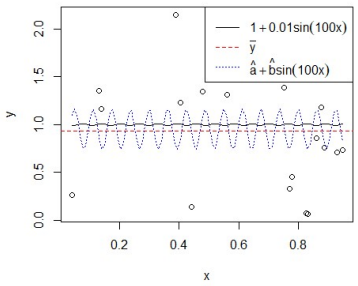
ε ~ N(0, 0.01)



ε ~ N(0, 0.1)



ε ~ N(0, 0.5)

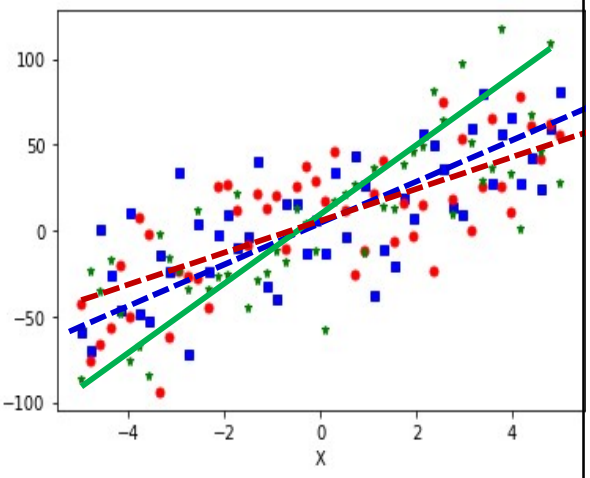


(Not necessarily the most complex model: we need to find a trade-off between criteria!)

Regression basics: bias and variance

- If the MSE is chosen as the cost function, the objective of model regression is to find *f(x)* that can minimize MSE(*f*) as follows:

$$
\begin{aligned}
MSE(f) &= E[(Y - f(X))^2] \\
&= E[E[(Y - f(X))^2]|X] \\
&= E[V(Y - f(X)|X) + (E[Y - f(x)|X])^2 \\
&= E[V(Y|X) + (E[Y - f(x)|X])^2]
\end{aligned}
$$



Regression basics: bias VS variance

- Since the data (*X*, *Y*) are observed/simulated as "random", the regressed model *f* is used for the new prediction *f(x), which is also a random variable*

$$f(x) = \mu(x) = E[Y|X = x] \xrightarrow{\text{yields}} Y = \hat{\mu}(X) + \epsilon$$

  - where ϵ is some noise variables, e.g., normally distributed random variable!

$$\begin{aligned} MSE(\hat{\mu}(x)) &= E\left[(Y - \mu(X))^2 | X = x\right] \\ &= E[E[(Y - \hat{\mu}(x))^2 | X = x, \hat{\mu}(x) = \mu(X)] | X = x] \\ &= E[\sigma^2(x) + (\mu(X) - \hat{\mu}(x))^2 | X = x] \\ &= \sigma^2(x) + (\mu(X) - E[\hat{\mu}(x)])^2 + (E[\hat{\mu}(x)^2] - (E[\hat{\mu}(x)])^2) \\ &= \sigma^2(x) + Bias(\hat{\mu})^2 + V(\hat{\mu}) \end{aligned}$$

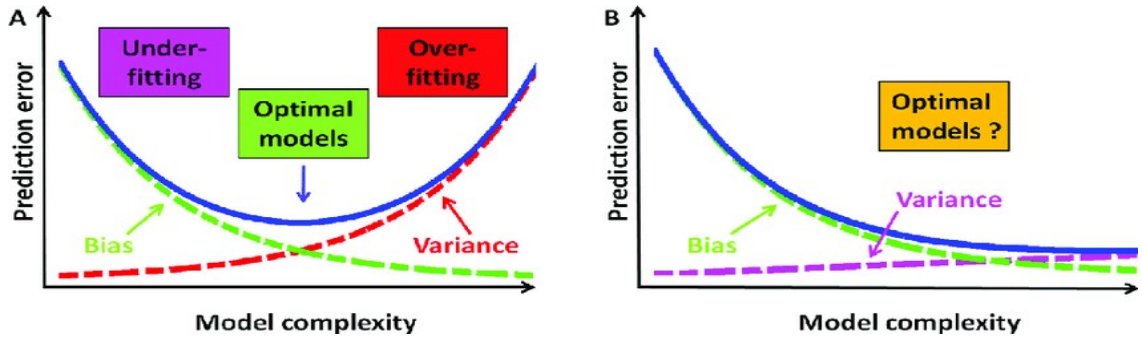
9 2023-04-10

9

Bias VS variance (accuracy - robust)

How complex should a model/regression be?

- Model is used for prediction (not to purely describe the data)
- Data contains uncertainties
- More complex models are associate with high uncertainties (variance)
- Should be a trade-off between bias and variance



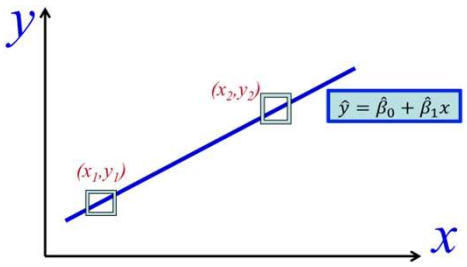
10 2023-04-10

10

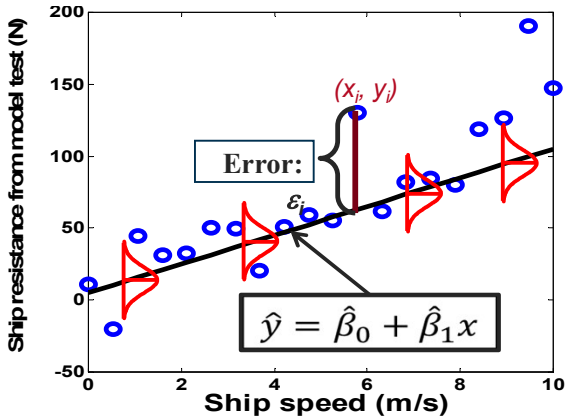
Single variable regression (statistical learning)

- Simple linear model (a special case of KNN)
- KNN and kernel smooth

Physical model VS ML model



- X: speed
- Y: resistance/power measured
- Both variable may contain errors (random variables)



Linear model (start from simple)

- Simple linear regression model (with only one variable)

$$Y = \mu(x) = \beta_0 + \beta_1 X$$

$$MSE(\mu(x)) = E[Y - \beta_0 - \beta_1 X]$$
$$= E\big[E[Y - \beta_0 - \beta_1 X|X]\big]$$
$$= E[V(Y|X)] + E[(E[Y - \beta_0 - \beta_1 X]|X)^2]$$

$$\beta_0 = \mathrm{E[Y]} - \beta_1 E[X]$$

$$\beta_1 = \frac{Cov(X,Y)}{V(X)}$$

$$\frac{\partial MSE}{\partial \beta_0} = -E[2(Y - \beta_0 - \beta_1 X)] = 0$$

$$\frac{\partial MSE}{\partial \beta_1} = E[XY] - \beta_1 E[X^2] + (E[Y] - \beta_1 E[X])E[X] = 0$$

$$\mu(x) = \mathrm{E[Y]} + \frac{Cov(X,Y)}{V(X)}(x - E[X])$$

Linear regression from data

- When we have a series of data for X = [x_1, x_2,…,x_n], Y=[y_1, y_2,…,y_n], then the parameters can be estimated by:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\hat{V}(X)} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- The new prediction becomes:

$$\hat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{y} - \hat{\beta}_1(x - \bar{x})$$
$$= \sum_{i=1}^{n} \frac{1}{n}\left(1 + \frac{(x-\bar{x})(x_i-\bar{x})}{\hat{V}(X)}\right) y_i$$
$$= \sum_{i=1}^{n} y_i w(x_i, x)$$

The linear regression model is simply a weighted average, also known as the linear smoother. The conditional expectation E[Y|X] is a special case of the linear smoother.

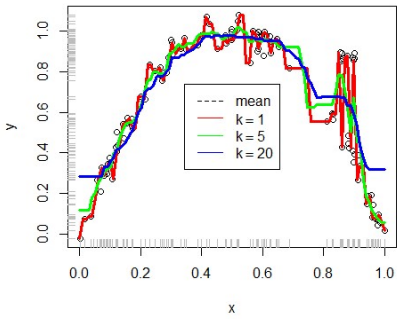
K-Nearest Neighbour (KNN) regression

For the KNN, the regression model is estimated by $\hat{\mu}(x) = \sum_{i=1}^{n} y_i \hat{w}(x_i, x)$

- First, we need to define the neighbourhood (1 nearest neighbour, or K-nearest neighbour)
- Then, we need to define the smooth function (weights), which is often described by certain probability density function, such as uniformly distribution (mean), Gaussian, exponential

$$\hat{w}(x_i, x) = \begin{cases} 1, & x_i \text{ nearest neighbor of } x \\ 0, & \text{otherwise} \end{cases}$$

$$\hat{w}(x_i, x) = \begin{cases} \frac{1}{k}, & x_i \text{ one of the k nearest neighbor of } x \\ 0, & \text{otherwise} \end{cases}$$



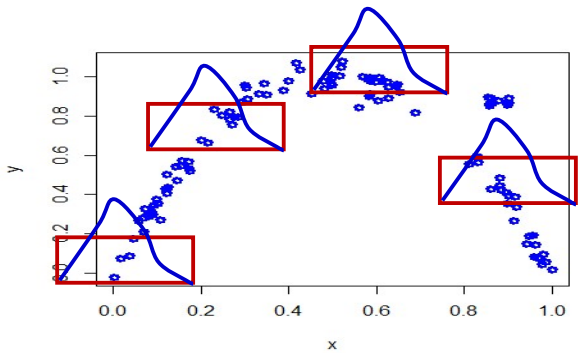
15

15

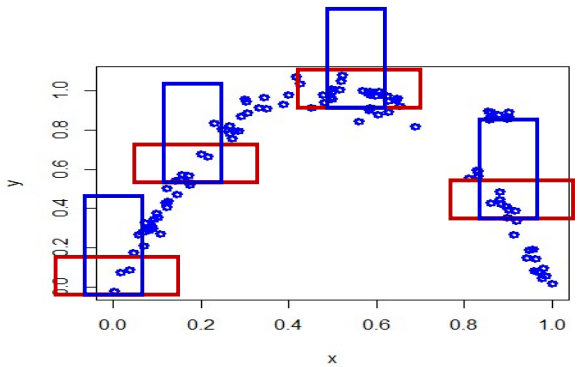
K-Nearest Neighbour (KNN) regression

Two basic issues to consider in the KNN method

- Kernels for the smooth: Box (uniform distribution) and Gaussian (normal probability)
- Width (how many k nearest neighbour should be considered)



Effect of shape of the kernel



Effect of width of the kernel

16

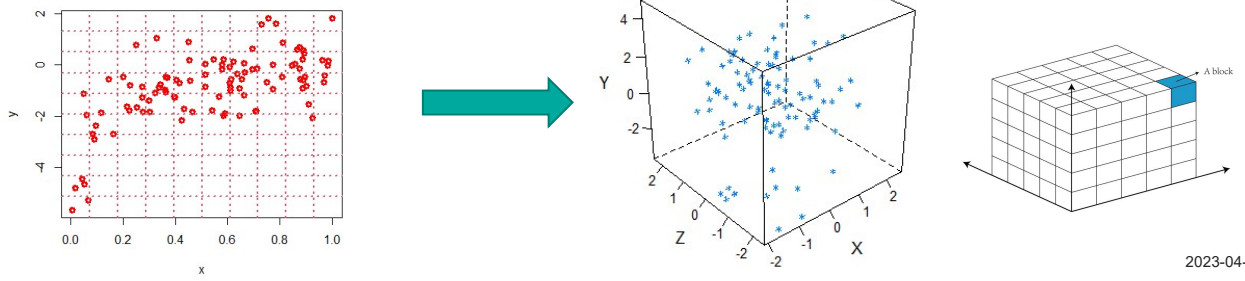
16

A more general case:

More than one independent variables → multivariate models

KNN method for multivariate models

- For a single input variable model, it might be efficient to discrete the 2-dimensional space into well defined KNN groups for the prediction
- When the dimension of input variables increase, the total number of the KNN groups will increase exponentially according to $O(n^p)$



Linear multivariate regression

When the dimension increase, a mathematically formulated model seems work efficient.

- Let $\boldsymbol{X} = [1,\ X_1,\ \ X_2,\ \ \ldots\ \ X_p]$, it becomes,

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\beta}$$

- If the least square method is used for regression, the RSS is:

$$RSS = \sum_{i=1}^{n} (y_i - X_i^T\beta)^2 = (\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta)$$

- Differentiating wrt β, we get the normal equations and estimations as

$$\boldsymbol{X^T}(\boldsymbol{y} - \boldsymbol{X}\beta) = 0 \quad \Longrightarrow \quad \hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X^T}\boldsymbol{y}$$

Uncertainties in the regression models

- Since the data observed are coming from random variables, both the model parameters and new predictions based on the estimated model are associated with uncertainties.
- For the estimated multivariate linear model, the prediction of Y under observation $\boldsymbol{X} = \boldsymbol{x}$:

$$\beta = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \qquad \hat{\beta} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y} \xrightarrow{\text{yields}} \boldsymbol{Y} = \boldsymbol{x}\hat{\beta} + \varepsilon$$

- The conditional mean and variances of new prediction under observation $\boldsymbol{X} = \boldsymbol{x}$:
- Under the condition of observation $\boldsymbol{X} = \boldsymbol{x}$, the model parameters also behave randomly:

$$E[\boldsymbol{Y} - \boldsymbol{X}\beta] = 0$$
$$E[\boldsymbol{Y} - \boldsymbol{X}\beta|\boldsymbol{X} = \boldsymbol{x}] \neq 0$$
$$V(\boldsymbol{Y} - \boldsymbol{X}\beta|\boldsymbol{X} = \boldsymbol{x}_1) \neq V(\boldsymbol{Y} - \boldsymbol{X}\beta|\boldsymbol{X} = \boldsymbol{x}_2)$$

$$\hat{\beta} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{Y} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T(\boldsymbol{x}\beta + \varepsilon) = \beta + (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\varepsilon$$

$$E\left[\hat{\beta}|\boldsymbol{X} = \boldsymbol{x}\right] = \beta + \left(\boldsymbol{x}^T\boldsymbol{x}\right)^{-1}\boldsymbol{x}^T E[\varepsilon] = \beta$$

$$V\left(\hat{\beta}\middle|\boldsymbol{X} = \boldsymbol{x}\right) = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T V(\varepsilon|\boldsymbol{X} = \boldsymbol{x})\boldsymbol{x}\left(\boldsymbol{x}^T\boldsymbol{x}\right)^{-1}$$

Model uncertainties

Let's move to the single input/regressor (1-dimensional) regression model to understand the uncertainties!

21

Confidence interval of the model

Confidence interval of new observation

$$\hat{y}_p \pm t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\sum_{i=1}^{n}(x_p-\bar{x})^2}\right]}$$

Confidence Interval of fitted values

$$\hat{y}_i \pm t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]}$$

1 — without 1

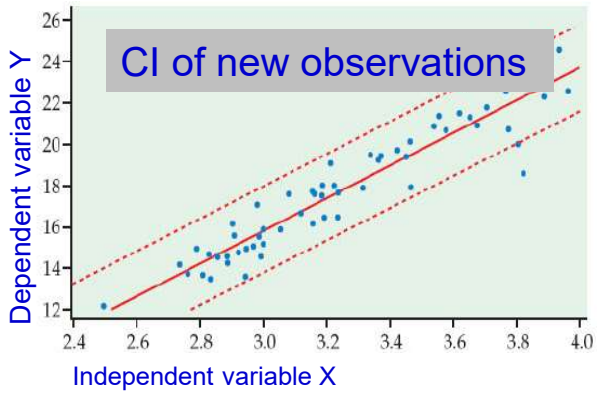
Used to estimate the value of $\boldsymbol{y = \hat{\beta}_0 + \hat{\beta}_1 x + \varepsilon}$

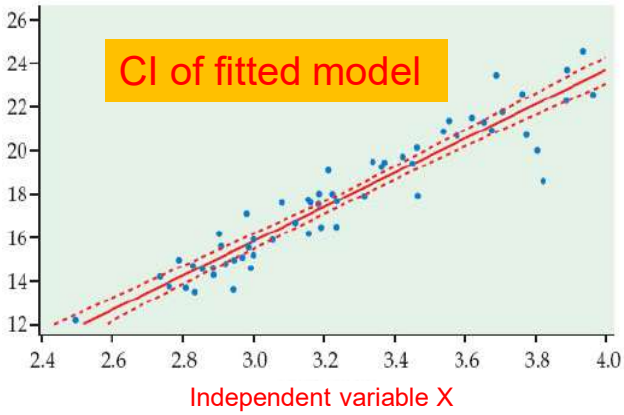
Used to estimate the ***mean value*** of y $\mathrm{E}[y] = \hat{\beta}_0 + \hat{\beta}_1 x$

The **confidence interval estimate** of the expected value of y will be **narrower** than the **prediction interval** for the same given value of x and confidence level. This is because there is less error in estimating a mean value as opposed to predicting an individual value.

22

CI for the regressed model and new prediction





2023-04-10

23

Introduction of the first assignment project

2023-04-10

24

25