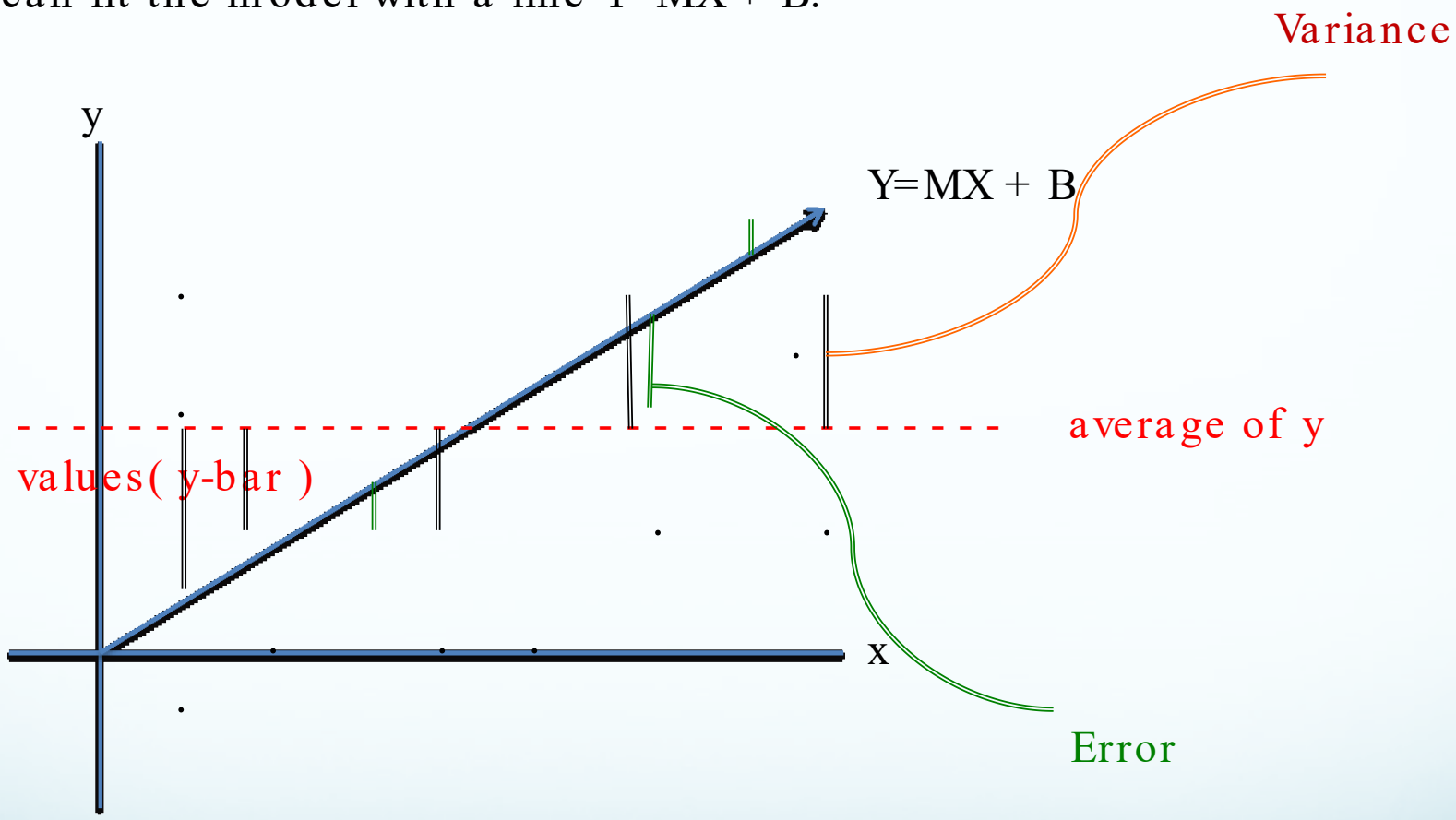
Section 6.3 Linear Regression

In this section we explore a statistical methodology for minimizing the sum of the squared deviations, called linear regression.

Objectives:

1. To illustrate the basic linear regression model and its assumptions
2. To define and interpret the statistic R^2 .
3. To illustrate a graphical interpretation for the fit of the linear regression model by examining and interpreting the residual scatterplots.

Given a set of points $\{(x_i , y_i)\}_{i=1}^{N}$ and suppose we can fit the model with a line Y=MX + B.



. Data (x_i , y_i)

Then by assessing the model and its fit with the data set, we can compute the following:

1. Sum of Square Error
2. Total Variance (Sum of squared total variance)
3. Sum of the Squared Residuals
4. R^2 - Value

Definitions:

- Sum of Squared Error (SSE)

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

Where ŷ is the value of the constructed model.

- Sum of Squared Total Variance (SST)

$$\sum_{i=1}^{N}(y_i - \bar{y})^2$$

Where,

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

- Sum of Squared Residuals (SSR)

SSR = SST – SSE.

The R^2 Value (R^2- Test)

By definition,

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

*** Remarks:

1. The smaller the error is, the closer R^2 is to 1.
2. R^2 is unit-less, and independent of scale of the data set.
3. How well the model fits depends on R^2- value, the closer to 1 (100%) the better the fit is.
4. When the model fits well, SST > SSE. But SSR can be negative.
5. Also, when the model fits well, R^2 $\varepsilon[0,1]$.

Example 1:

Consider the Data given by the table,

Ponderosa Pine Data

Find the R^2 value based on the data.

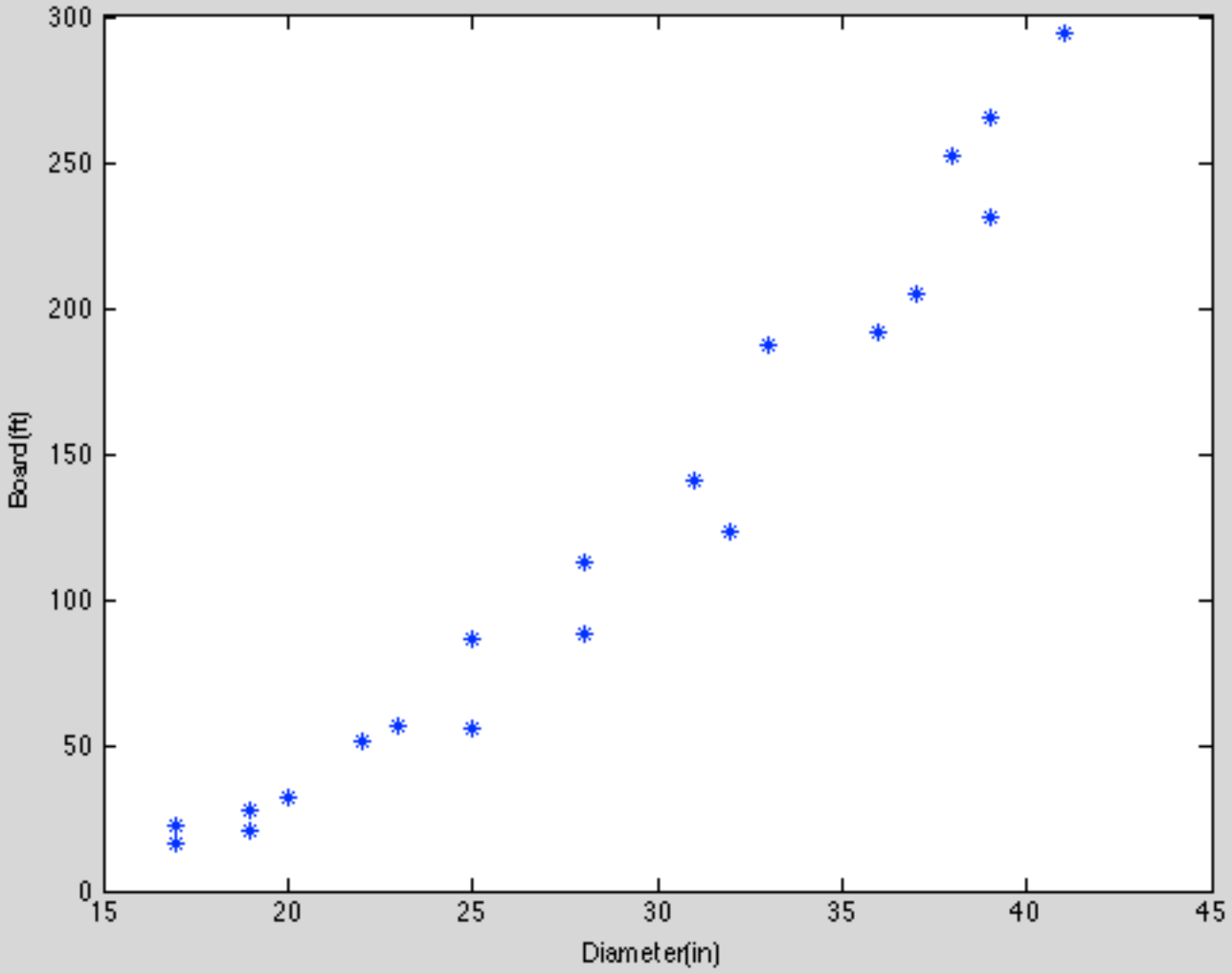
Diameter (in)	36	28	28	41	19	32	22	38	25
Board (ft)	192	113	88	294	28	123	51	252	56

Diameter	17	31	20	25	19	39	33	17	37	23	39
Board	16	141	32	86	21	231	187	22	205	57	265

Plot the Data using Matlab

```
%Data-
%Diameter:
d=[36, 28, 28, 41, 19, 32, 22, 38, 25, 17, 31, 20, 25, 19, 39,
    33, 17, 37,  23, 39];
%Board:
v=[192, 113, 88, 294, 28, 123, 51, 252, 56, 16,  141, 32, 86,
    21, 231, 187, 22, 205, 57, 265];
plot(d,v,'*');
xlabel('Diameter (in)'); ylabel('Board (ft)');
```

Plot the Data



Example 1 cont.

Due to the non-linear behavior of the data, possible choices for the model include:

1. $v=a*d^3$ → 3. $v= a*d^2$ (1 parameter fit)
2. $v= a*d^3 + b$ → 4. $v= a*d^2 + b$ (myLineFit)

Which one will fit better??

We can use MatLab to compare the R^2 values.
The closer the R^2 value is to 1, the better the fit is.

In MatLab

```
%Two-Paremeter Fit to v=ad^3 + b
%From given formula
a1=sum(d.^3.*v)/sum(d.^6);
v1=a1*d.^3;
Rsq1=calcRSQ(v,v1);

X=d.^3; Y=v1;
[a,b]=myLineFit(X,Y);
xx=linspace(min(d),max(d),100);
yy=a.*xx.^3+b;

%Now to do the parameter fit to v=ad^2+b
%We will use the same data d,v
a2=sum(d.^2.*v)/sum(d.^4);
v2=a2*d.^2;
Rsq2=calcRSQ(v,v2);

X2=d.^2; Y2=v2;
[a2,b2]=myLineFit(X2,Y2);
xx2=linspace(min(d),max(d),100);
yy2=a2.*xx2.^2+b2;
```

In MatLab Cont.

```
%For the One-Parameter to v=ad^3
%We will use the same data d,v
a3=sum(d.^3.*v)/sum(d.^6);
v3=a3*d.^3;
Rsq3=calcRSQ(v,v3);

X3=d.^2; Y3=v3;
[a3,b3]=myLineFit(X3,Y3);
xx3=linspace(min(d),max(d),100);
yy3=a3.*xx3.^2;

%For the One-Parameter fit to v=ad^2
%We will use the same data d,v
a4=sum(d.^2.*v)/sum(d.^4);
v4=a4*d.^2;
Rsq4=calcRSQ(v,v4);

X4=d.^2; Y4=v4;
[a4,b4]=myLineFit(X4,Y4);
xx4=linspace(min(d),max(d),100);
yy4=a4.*xx4.^2;
```

In MatLab Cont.

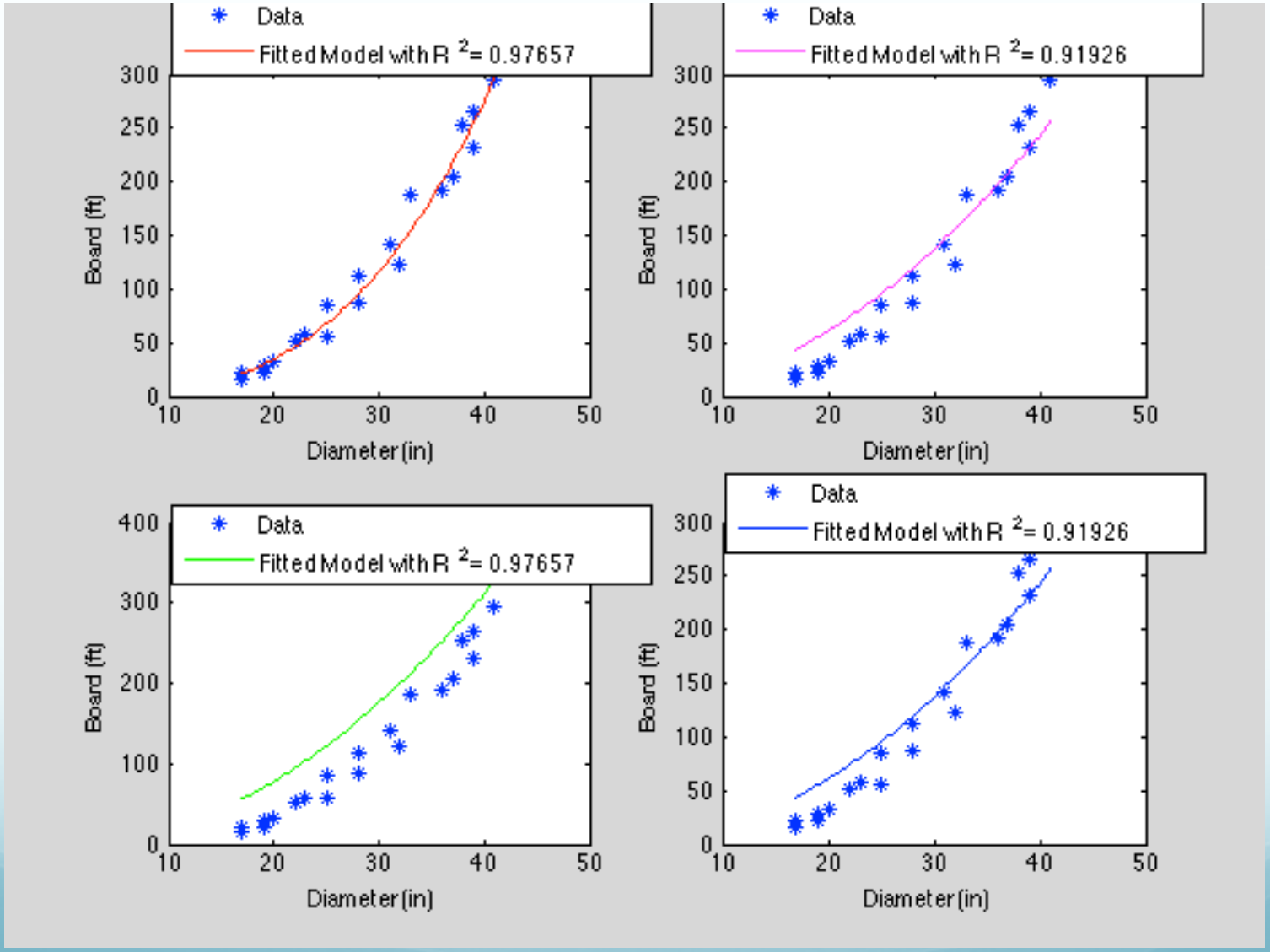
```
%To plot the individual plots we can use subplots-
subplot(2,2,1); plot(d,v,'*',xx,yy,'r'); xlabel('Diameter (in)'); ylabel('Board
    (ft)');legend('Data', ['Fitted Model with R^2= ' num2str(Rsq1)]);

subplot(2,2,2); plot(d,v,'*',xx2,yy2,'m'); xlabel('Diameter (in)'); ylabel('Board
    (ft)');legend('Data', ['Fitted Model with R^2= ' num2str(Rsq2)]);

subplot(2,2,3); plot(d,v,'*',xx3,yy3,'g'); xlabel('Diameter (in)'); ylabel('Board
    (ft)');legend('Data', ['Fitted Model with R^2= ' num2str(Rsq3)]);

subplot(2,2,4); plot(d,v,'*',xx4,yy4,'b'); xlabel('Diameter (in)'); ylabel('Board
    (ft)');legend('Data', ['Fitted Model with R^2= ' num2str(Rsq4)]);
```

Subplots With R^2 Values



R^2 value function (in MatLab cont.)

```
%The R2 value calculating function-

function [Rsq, SSE, SST]= calcRsq(y,yhat)
    SSE= sum((y-yhat).^2);
    ybar= mean(y);
    SST= sum((y-ybar).^2);
    Rsq= 1-(SSE/SST);
end
```

Graphical Analysis

- Remember the Remarks about the R^2 value.
- Which of the models had its R^2 Value closest to 1?
- That model is the best fit!

Summary

- ✓ Look at data and plot the data.
- ✓ Compute the sum of the square error
- ✓ Compute the total variance
- ✓ Compute the Squared Residuals
- ✓ Find the R^2 value (or R^2 test)

Homework §6.3 #1, 2