Statistics 120
Scatter Plots and Smoothing

An Example – Car Stopping Distances

- An experiment was conducted to measure how the stopping distance of a car depends on its speed.
- The experiment used a random selection of cars and a variety of speeds.
- The measurements are contained in the R data set "cars," which can be loaded with the command:

```
data(cars)
```

Car Stopping Distances – Imperial Units

mph	*ft*	*mph*	*ft*	*mph*	*ft*	*mph*	*ft*
4	2	12	24	16	32	20	48
4	10	12	28	16	40	20	52
7	4	13	26	17	32	20	56
7	22	13	34	17	40	20	64
8	16	13	34	17	50	22	66
9	10	13	46	18	42	23	54
10	18	14	26	18	56	24	70
10	26	14	36	18	76	24	92
10	34	14	60	18	84	24	93
11	17	14	80	19	36	24	120
11	28	15	20	19	46	25	85
12	14	15	26	19	68		
12	20	15	54	20	32		

Car Stopping Distances – Metric Units

kph	*m*	*kph*	*m*	*kph*	*m*	*kph*	*m*
6.4	0.6	19.3	7.3	25.7	9.8	32.2	14.6
6.4	3.0	19.3	8.5	25.7	12.2	32.2	15.8
11.3	1.2	20.9	7.9	27.4	9.8	32.2	17.1
11.3	6.7	20.9	10.4	27.4	12.2	32.2	19.5
12.9	4.9	20.9	10.4	27.4	15.2	35.4	20.1
14.5	3.0	20.9	14.0	29.0	12.8	37.0	16.5
16.1	5.5	22.5	7.9	29.0	17.1	38.6	21.3
16.1	7.9	22.5	11.0	29.0	23.2	38.6	28.0
16.1	10.4	22.5	18.3	29.0	25.6	38.6	28.3
17.7	5.2	22.5	24.4	30.6	11.0	38.6	36.6
17.7	8.5	24.1	6.1	30.6	14.0	40.2	25.9
19.3	4.3	24.1	7.9	30.6	20.7		
19.3	6.1	24.1	16.5	32.2	9.8		

Question

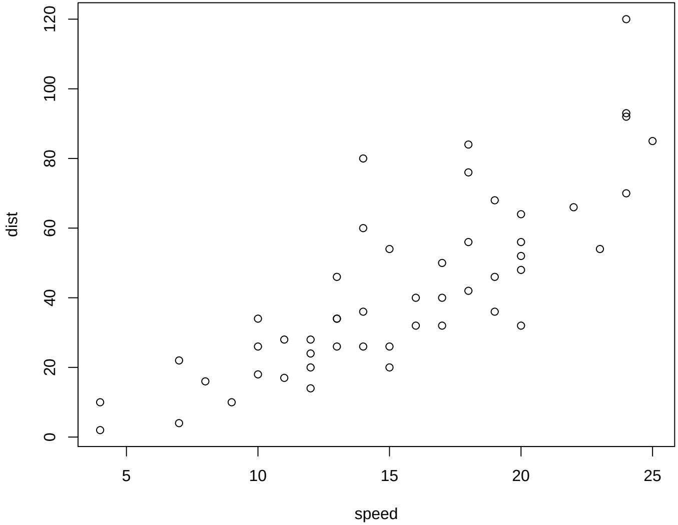
Why would anyone collect this kind of data?

Graphical Investigation

- We are going to use the value to investigate the relationship between speed and stopping distance.
- The best way to investigate the relationship between two related variables is to simply plot the pairs of values.
- The basic plot is produced with `plot`.

```
> data(cars)
> attach(cars)
> plot(speed, dist)
```

- Using default labels is fine for exploratory work, but not for publication.


dist
speed
0
20
40
60
80
100
120
5
10
15
20
25

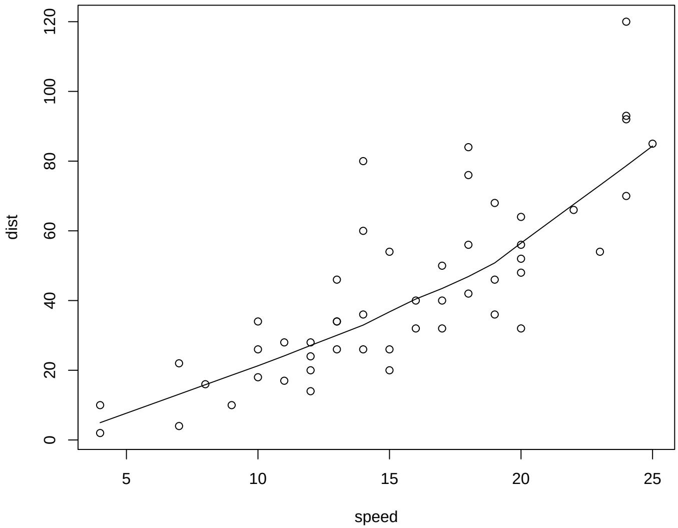
Comments

- There is a general trend for stopping distance to increase with speed.
- There is evidence that the variability in the stopping distances also increases with speed.
- It is difficult to be more precise about the form of the relationship by just looking at the scatter of points.

Scatterplot Smoothing

- One way to try to uncover the nature of the relationship is to add a line which conveys the basic trend in the plot.
- This can be done using a technique known as scatterplot smoothing.
- R has a smoothing procedure called LOWESS which can be used to add the trend line.
- LOWESS is a relatively complicated procedure, but it is easy to use.

```
plot(speed, dist)
lines(lowess(speed, dist))
```



dist
0
20
40
60
80
100
120
speed
5
10
15
20
25

Conclusions

- The “smooth” confirms that stopping distance increases with speed, but it gives us more detail.
- The relationship is not of the form

$$y = a + bx$$

  but has an unknown mathematical form.
- If we are just interested in determining the stopping distance we can expect for a given speed this doesn’t matter.
- We can just read the answer off the graph.

Turning a Smooth into a Function

- It is useful to have a a computational procedure for "reading off the results" from the lowess curve. This can be done by fitting a spline curve through the points returned by lowess.

```
> z = lowess(speed, dist)
> u = !duplicated(z$x)
> f = splinefun(z$x[u], z$y[u])
```

- The function `f` can now be used to do the lookup of values on the curve.

```
> f(10:12)
[1] 21.28031 24.12928 27.11955
```

Mathematical Modelling

- While the curve obtained by the LOWESS lets us read off the kind of stopping distance we can expect for a given speed, it does not help understand why the relationship is the way it is.
- It is possible to use the data to try to fit a well defined mathematical curve to the data points. This suffers from the same difficulty.
- It is much better to try to understand the mechanism which produced the data.

Conservation of Energy

- A moving car has kinetic energy associated with it.
- The kinetic energy is dissipated as work is done against friction during breaking.
- When the car comes to rest the Kinetic energy dissipated equals work done.

Equations from Physics

Thanks to Isaac Newton (and others) we know the following.

$$\text{Kinetic Energy} = \tfrac{1}{2}mv^2$$

where m is the mass of the car and v is the car speed.

$$\text{Work Done} = F \times d$$

where F is the frictional force and d is the distance travelled.

When the car comes to a halt, all the kinetic energy has been dissipated as work done against the frictional force.

Conservation of Energy

Because energy is conserved, we can equate right-hand side of the previous equations.

$$F \times d = \tfrac{1}{2} m v^2$$

Ignoring constants, this says that

$$d \propto v^2$$

or

$$\sqrt{d} \propto v.$$

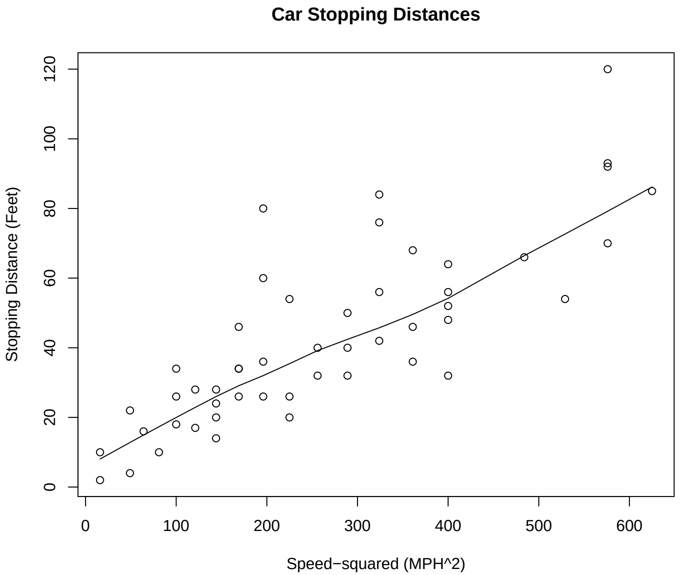
Using Plots

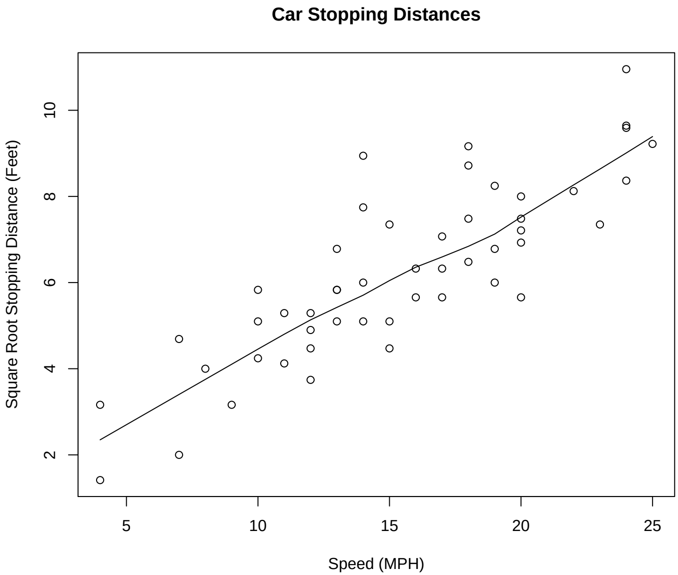
- We can check whether these are really the underlying relationships with scatterplots.
- Either plot distance against speed-squared or plot the square-root of distance against speed.

Producing the Plots

```
> plot(speed^2, dist,
       main = "Car Stopping Distances",
       xlab = "Speed-squared (MPH^2)",
       ylab = "Stopping Distance (Feet)")
> lines(lowess(speed^2, dist))

> plot(speed, sqrt(dist),
       main = "Car Stopping Distances",
       xlab = "Speed (MPH)",
       ylab = "Square Root Stopping Distance (Feet)")
> lines(lowess(speed, sqrt(dist)))
```


Car Stopping Distances
Stopping Distance (Feet)
0
20
40
60
80
100
120
0
100
200
300
400
500
600
Speed−squared (MPH^2)



Car Stopping Distances
Square Root Stopping Distance (Feet)
2
4
6
8
10
5
10
15
20
25
Speed (MPH)

Conclusions

- Both the plots indicate that there is close to a straight line relationship between speed-squared and distance.
- From a statistical point-of-view, the second plot is preferable because the scatter of points about the line is independent of speed. (I.e. it is possible to compare apples with apples).
- The straight line of *best fit* to the plot of square-root distance versus speed is:

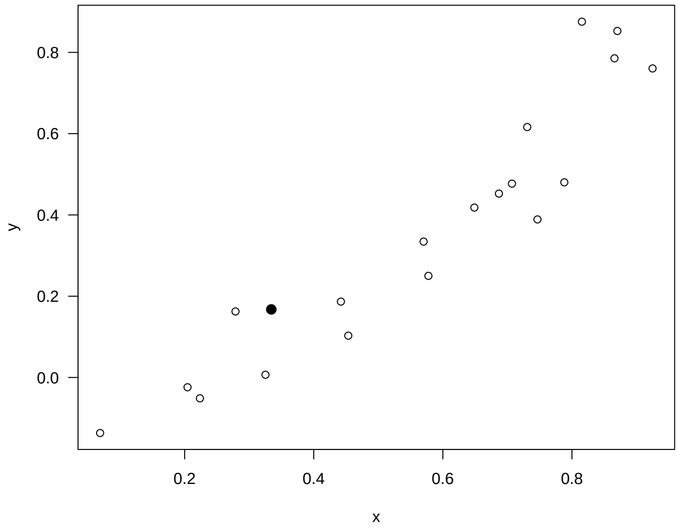
$$\sqrt{d} = 1.28 + 0.32 \times v$$

- Dropping the intercept, the best fit is:

$$\sqrt{d} = 0.4 \times v$$

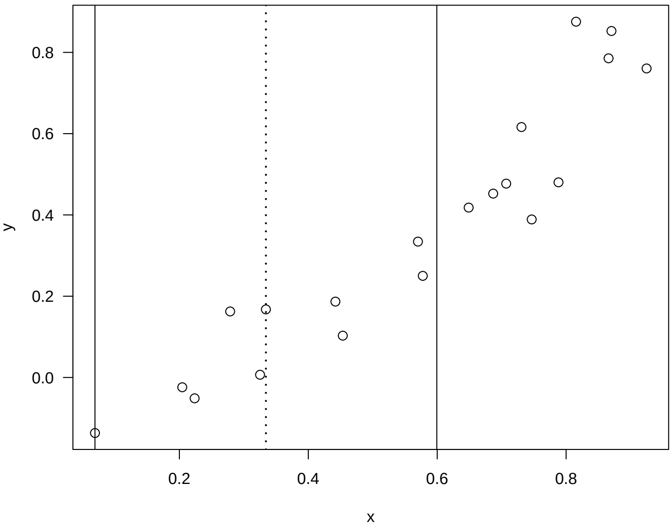
How Lowess Works

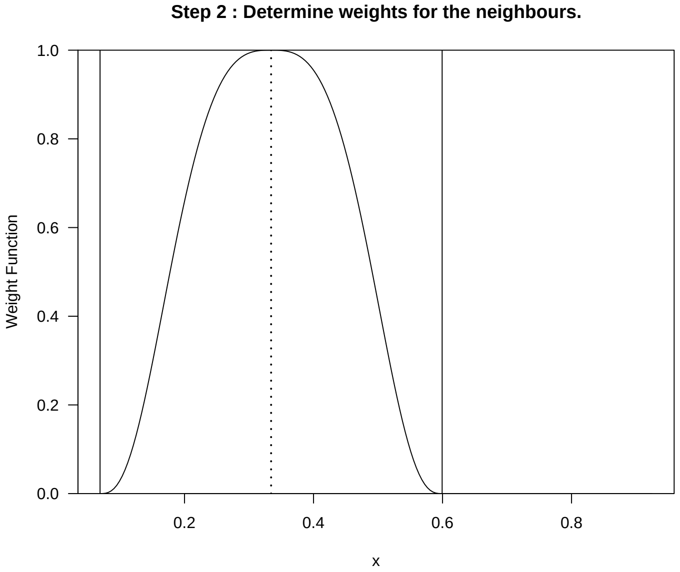
- It is worth spending a little time to see how lowess works.
- We'll consider how to get an estimate of the lowess curve at just one location in a scatter plot.
- We will compute the value of the lowess curve at the 6th point in the following plot.
- The lowess procedure does this for every point in the plot.



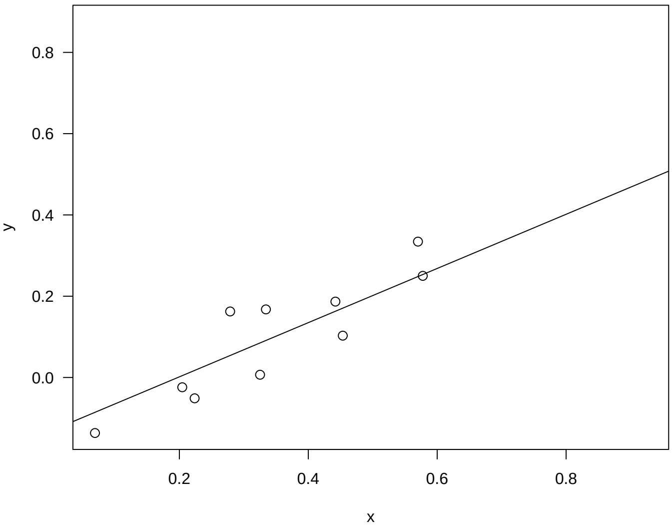
0.8
0.6
0.4
0.2
0.0
y
0.2
0.4
0.6
0.8
x

Step 1 : Find the neighbours of the point.

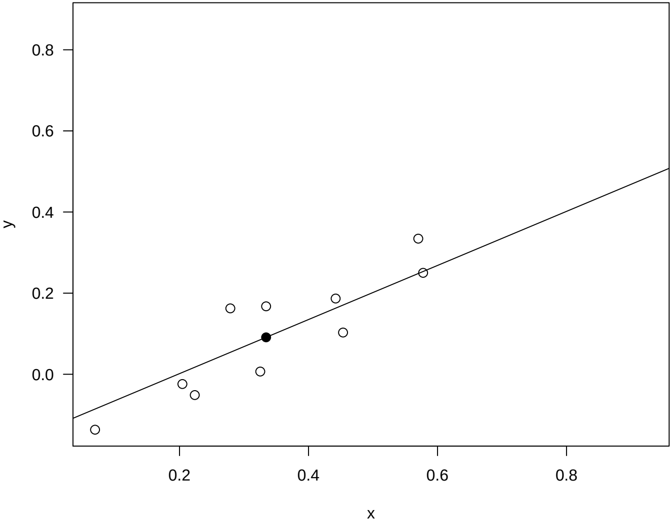


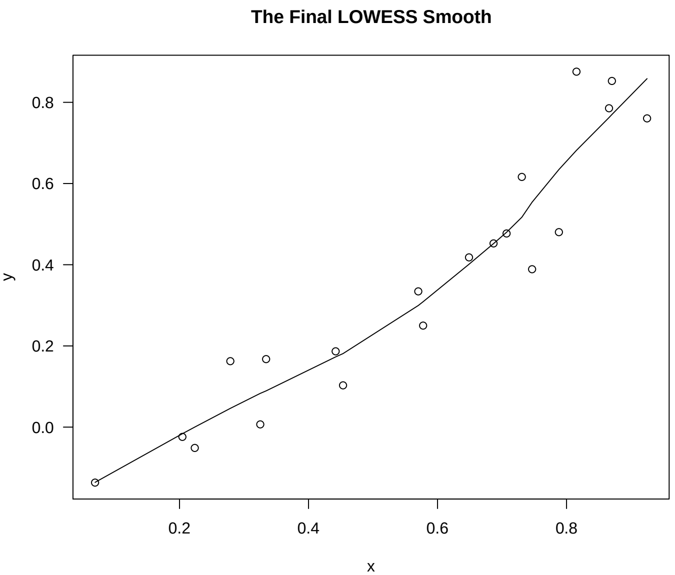

Step 2 : Determine weights for the neighbours.
Weight Function
1.0
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
x

Step 3 : Fit a straight line (using the weights).



Steps 4 : Use the line to assign a ‘‘fitted value.''

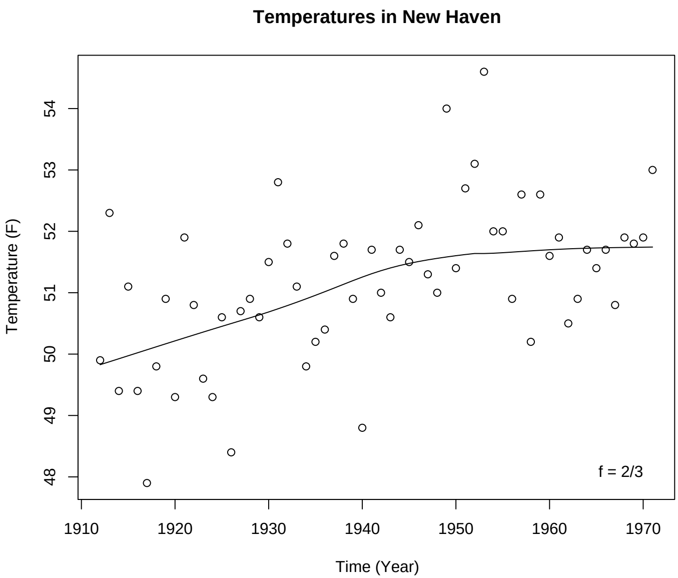


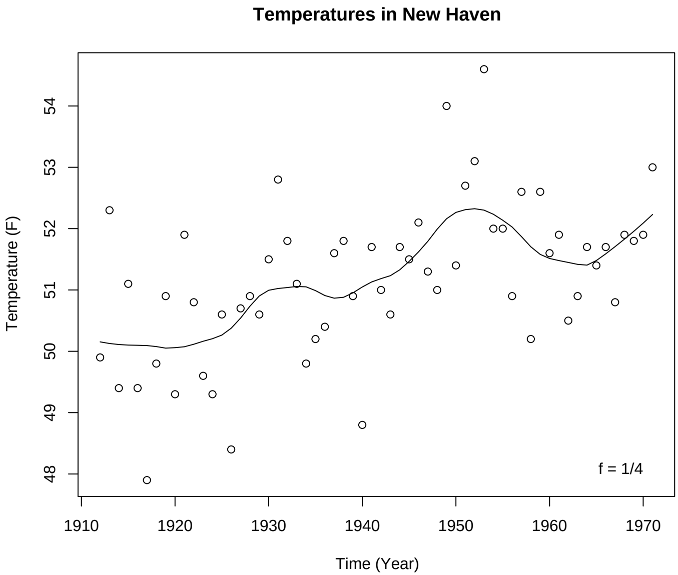

The Final LOWESS Smooth
0.8
0.6
0.4
0.2
0.0
y
0.2
0.4
0.6
0.8
x

Controlling The Amount of Smoothing

- The amount of smoothing in lowess is controlled by and optional argument called “`f`.”
- This gives the fraction of the data which will be used as “neighbours” of a given point, when computing the smoothed value at that point.
- The default value of `f` is $2/3$.
- The following examples will show the effect of varying the value of `f`.

```
> lines(lowess(nhtemp, f = 2/3))
> lines(lowess(nhtemp, f = 1/4))
```


Temperatures in New Haven
Temperature (F)
48
49
50
51
52
53
54
1910
1920
1930
1940
1950
1960
1970
Time (Year)
f = 2/3



Temperatures in New Haven
Temperature (F)
48
49
50
51
52
53
54
1910
1920
1930
1940
1950
1960
1970
Time (Year)
f = 1/4