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

<table>
<thead>
<tr>
<th>mph</th>
<th>ft</th>
<th>mph</th>
<th>ft</th>
<th>mph</th>
<th>ft</th>
<th>mph</th>
<th>ft</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2</td>
<td>12</td>
<td>24</td>
<td>16</td>
<td>32</td>
<td>20</td>
<td>48</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>12</td>
<td>28</td>
<td>16</td>
<td>40</td>
<td>20</td>
<td>52</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>13</td>
<td>26</td>
<td>17</td>
<td>32</td>
<td>20</td>
<td>56</td>
</tr>
<tr>
<td>7</td>
<td>22</td>
<td>13</td>
<td>34</td>
<td>17</td>
<td>40</td>
<td>20</td>
<td>64</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>13</td>
<td>34</td>
<td>17</td>
<td>50</td>
<td>22</td>
<td>66</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>13</td>
<td>46</td>
<td>18</td>
<td>42</td>
<td>23</td>
<td>54</td>
</tr>
<tr>
<td>10</td>
<td>18</td>
<td>14</td>
<td>26</td>
<td>18</td>
<td>56</td>
<td>24</td>
<td>70</td>
</tr>
<tr>
<td>10</td>
<td>26</td>
<td>14</td>
<td>36</td>
<td>18</td>
<td>76</td>
<td>24</td>
<td>92</td>
</tr>
<tr>
<td>10</td>
<td>34</td>
<td>14</td>
<td>60</td>
<td>18</td>
<td>84</td>
<td>24</td>
<td>93</td>
</tr>
<tr>
<td>11</td>
<td>17</td>
<td>14</td>
<td>80</td>
<td>19</td>
<td>36</td>
<td>24</td>
<td>120</td>
</tr>
<tr>
<td>11</td>
<td>28</td>
<td>15</td>
<td>20</td>
<td>19</td>
<td>46</td>
<td>25</td>
<td>85</td>
</tr>
<tr>
<td>12</td>
<td>14</td>
<td>15</td>
<td>26</td>
<td>19</td>
<td>68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>15</td>
<td>54</td>
<td>20</td>
<td>32</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Car Stopping Distances – Metric Units

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

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

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

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

\[
\begin{align*}
&z = \text{lowess}(\text{speed}, \text{dist}) \\
&u = \neg \text{duplicated}(z[,]) \\
&f = \text{splinefun}(z[,]u, z[,]u)
\end{align*}
\]

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

\[
\begin{align*}
&f(10:12) \\
&[1] 21.28031 24.12928 27.11955
\end{align*}
\]
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.

\[
\text{Kinetic Energy} = \frac{1}{2}mv^2
\]

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.

\[ F \times d = \frac{1}{2}mv^2 \]

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) vs. Speed-squared (MPH^2)

The graph shows the relationship between the stopping distance of a car and its speed-squared. As the speed-squared increases, the stopping distance also increases, indicating a direct correlation between the two variables. The data points are scattered, but a trend line is drawn to illustrate the general relationship. 
Car Stopping Distances

Speed (MPH) vs. Square Root Stopping Distance (Feet)

- The graph illustrates the relationship between speed and the square root of stopping distance for cars.
- The data points show a general trend of increasing stopping distance with increasing speed.
- A linear model has been fitted to the data, indicating a positive correlation between speed and stopping distance.

This diagram helps in understanding the impact of speed on the braking distance of cars, which is crucial for safety considerations.
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.
Step 1: Find the neighbours of the point.
Step 2: Determine weights for the neighbours.
Step 3: Fit a straight line (using the weights).
Steps 4: Use the line to assign a “fitted value.”
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.

  ```r
  > lines(lowess(nhtemp, f = 2/3))
  > lines(lowess(nhtemp, f = 1/4))
  ```
Temperatures in New Haven

<table>
<thead>
<tr>
<th>Time (Year)</th>
<th>Temperature (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1910</td>
<td>48</td>
</tr>
<tr>
<td>1920</td>
<td>49</td>
</tr>
<tr>
<td>1930</td>
<td>50</td>
</tr>
<tr>
<td>1940</td>
<td>51</td>
</tr>
<tr>
<td>1950</td>
<td>52</td>
</tr>
<tr>
<td>1960</td>
<td>53</td>
</tr>
<tr>
<td>1970</td>
<td>54</td>
</tr>
</tbody>
</table>

$f = 2/3$
Temperatures in New Haven

<table>
<thead>
<tr>
<th>Time (Year)</th>
<th>Temperature (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1910</td>
<td>48</td>
</tr>
<tr>
<td>1920</td>
<td>49</td>
</tr>
<tr>
<td>1930</td>
<td>50</td>
</tr>
<tr>
<td>1940</td>
<td>51</td>
</tr>
<tr>
<td>1950</td>
<td>52</td>
</tr>
<tr>
<td>1960</td>
<td>53</td>
</tr>
<tr>
<td>1970</td>
<td>54</td>
</tr>
</tbody>
</table>

\[ f = \frac{1}{4} \]