16. SIMPLE LINEAR REGRESSION I

Often, we wish to study the relationship between two variables:

- GMAT Scores and Grade Point Averages of First-Year Business School Students.
- Current CPI, Unemployment in 2 Years.
- Movie Budget, Total Gross.
- Height, Salary.
- Advertising Expenditure, Profits

**Purpose:**

1. To describe and understand how the variables are related. For example, an advertiser can use relationship between click-through rate and demographic variables to determine if they should bid show an ad to an online visitor.

2. To forecast a new observation. If we know the relationship between current CPI and future unemployment, then we can try to forecast the unemployment.

3. To adjust or control a process. If we know how advertising expenditure influences profits, we can adjust our advertising budget to (hopefully) improve profits.

- The first thing to do is to make a **scatterplot**: a graphical display of each data point using two axes to represent the two variables.

  If one variable is seen as causing or influencing the other, it is called X and defines the horizontal axis.

  Some common names for X are:
  Predictor Variable, Explanatory Variable, Regressor, Feature.

  The variable that might respond or be influenced is called Y and defines the vertical axis. Some common names for Y are:
  Response Variable, Dependent Variable.

Our data take the form \((x_1, y_1), \ldots, (x_n, y_n)\), where each X value is paired with its corresponding Y value.

Even if the scatterplot shows a relationship between the variables, this does not in any way prove that X *causes* Y. For example, it would be ludicrous to argue that business students with A averages got them *because* they did well on the GMAT.

Sometimes (but not always) our scatterplot will show a **linear relationship**: The points seem to be bunched randomly around a straight line. Ideally, the amount of scatter does not depend on X and there are no extreme outliers. In this case, the methods to be studied here can be safely applied.
**Eg:** Rotten Tomatoes Critics and Audience Score for Top 10 Grossing Movies of 2013:

<table>
<thead>
<tr>
<th>Movie</th>
<th>Critics Score</th>
<th>Audience Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Hunger Games:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catching Fire</td>
<td>89</td>
<td>90</td>
</tr>
<tr>
<td>Iron Man 3</td>
<td>79</td>
<td>79</td>
</tr>
<tr>
<td>Frozen</td>
<td>89</td>
<td>86</td>
</tr>
<tr>
<td>Despicable Me 2</td>
<td>74</td>
<td>85</td>
</tr>
<tr>
<td>Man of Steel</td>
<td>55</td>
<td>76</td>
</tr>
<tr>
<td>Gravity</td>
<td>97</td>
<td>80</td>
</tr>
<tr>
<td>Monsters University</td>
<td>78</td>
<td>82</td>
</tr>
<tr>
<td>The Hobbit:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Desolation of Smaug</td>
<td>74</td>
<td>86</td>
</tr>
<tr>
<td>Fast &amp; Furious 6</td>
<td>36</td>
<td>50</td>
</tr>
<tr>
<td>Oz The Great and Powerful</td>
<td>59</td>
<td>56</td>
</tr>
</tbody>
</table>

The scatterplot shows a strong positive linear association, but not a perfect relationship.

**Eg:** Calories per 12-ounce serving and the Percentage of Alcohol for 101 US domestic beers.

Overall, we see a positive linear association but there are at least two strong outliers. To identify them in Minitab, hold the cursor over the data point.

It is also possible that a scatterplot will show no relationship, i.e., the points seem randomly distributed, with no particular tendency to move up or down as X changes.

**Eg:** Does the Stock Market have "momentum"? That is, is the Market likely to keep going up today because it went up yesterday? To study this, consider the daily excess returns on the Market (see Handout 1), and make a scatterplot of Today's return (Y) against Yesterday's return (X).

"I don’t know what the S&P 500 will return, and I am the chairman of the index committee which runs it."

--- David Blitzer, chief economist at Standard & Poor’s Corp.
Clearly, stock returns are not very predictable.

The scatterplot shows virtually no relationship between \( X \) and \( Y \). This, of course, would be consistent with the Efficient Market Hypothesis, i.e., that daily prices are a random walk.

“No relationship” gives a "linear relationship with a slope of zero". But if the slope is zero, today's return is useless for predicting tomorrow's. That is, \( Y \) does not depend in any meaningful way on \( X \).

Another possibility is that the scatterplot shows a **nonlinear relationship**: The points are bunched around a curve rather than a straight line. Since there are so many different kinds of curves that can be drawn, the analysis is more complex. For example, consider the height of children (\( Y \)) and their ages (\( X \)).

Top 4 outliers:
Oct 19-Oct 20, 1987 (Crash),