7. Correlation and Regression

7.1 Motivation

**Exercise:** Algebraically expand the expression $(X - \mu_X)(Y - \mu_Y)$, and use the properties of mathematical expectation given in 3.1. This motivates an alternate formula for $s_{xy}$. 

**Exercise:**
For the sake of simplicity, let us assume that the **predictor variable** $X$ is nonrandom (i.e., deterministic), and that the **response variable** $Y$ is random. (Although, the subsequent techniques can be extended to random $X$ as well.)

**Example**: $X =$ fat (grams), $Y =$ cholesterol level (mg/dL)

Suppose the following sample of $n = 5$ data pairs (i.e., points) is obtained and graphed in a **scatterplot**, along with some accompanying summary statistics:

<table>
<thead>
<tr>
<th>$X$</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>210</td>
<td>200</td>
<td>220</td>
<td>280</td>
<td>290</td>
</tr>
</tbody>
</table>

$\bar{X} = 80 \quad s^2_x = 250$

$\bar{Y} = 240 \quad s^2_y = 1750$

**Sample Covariance**

$$s_{xy} = \frac{1}{5-1} \left[ (60 - 80)(210 - 240) + (70 - 80)(200 - 240) + (80 - 80)(220 - 240) + (90 - 80)(280 - 240) + (100 - 80)(290 - 240) \right] = 600$$

As the name implies, the **variance** measures the extent to which a single variable varies (about its mean). Similarly, the **covariance** measures the extent to which two variables vary (about their individual means), *with respect to each other*. 
Ideally, if there is no association of any kind between two variables $X$ and $Y$ (as in the case where they are independent), then a scatterplot would reveal no organized structure, and covariance = 0; e.g., $X =$ adult head size, $Y =$ IQ. Clearly, in a case such as this, the variable $X$ is not a good predictor of the response $Y$. Likewise, if the variables $X =$ age, $Y =$ body temperature ($^\circ$F) are measured in a group of healthy individuals, then the resulting scatterplot would consist of data points that are very nearly lined up horizontally (i.e., zero slope), reflecting a constant mean response value of $Y = 98.6^\circ$F, regardless of age $X$. Here again, covariance = 0 (or nearly so); $X$ is not a good predictor of the response $Y$. See figures.

However, in the preceding “fat vs. cholesterol” example, there is a clear “positive trend” exhibited in the scatterplot. Overall, it seems that as $X$ increases, $Y$ increases, and inversely, as $X$ decreases, $Y$ decreases. The simplest mathematical object that has this property is a straight line with positive slope, and so a linear description can be used to capture such “first-order” properties of the association between $X$ and $Y$. The two questions we now ask are…

1) How can we measure the strength of the linear association between $X$ and $Y$?

   **Answer:** Linear Correlation Coefficient

2) How can we model the linear association between $X$ and $Y$, essentially via an equation of the form $Y = mX + b$?

   **Answer:** Simple Linear Regression

*Caution:* The covariance can equal zero under other conditions as well; see Exercise in the next section.
Before moving on to the next section, some important details are necessary in order to provide a more formal context for this type of problem. In our example, the response variable of interest is cholesterol level \( Y \), which presumably has some overall probability distribution in the study population. The mean cholesterol level of this population can therefore be denoted \( \mu_Y \) – or, recall, expectation \( E[Y] \) – and estimated by the “grand mean” \( \bar{Y} = 240 \). Note that no information about \( X \) is used.

Now we seek to characterize the relation (if any) between cholesterol level \( Y \) and fat intake \( X \) in this population, based on a random sample using \( n = 5 \) fat intake values (i.e., \( x_1 = 60, x_2 = 70, x_3 = 80, x_4 = 90, x_5 = 100 \)). Each of these fixed \( x_i \) values can be regarded as representing a different amount of fat grams consumed by a subpopulation of individuals, whose cholesterol levels \( Y \), conditioned on that value of \( X = x_i \), are assumed to be normally distributed. The conditional mean cholesterol level of each of these distributions could therefore be denoted \( \mu_{Y|x_i} \) – equivalently, conditional expectation \( E[Y|X = x_i] \) – for \( i = 1, 2, 3, 4, 5 \). (See figure; note that, in addition, we will assume that the variances “within groups” are all equal (to \( \sigma^2 \)), and that they are independent of one another.) If no relation between \( X \) and \( Y \) exists, we would expect to see no organized variation in \( Y \) as \( X \) changes, and all of these conditional means would either be uniformly “scattered” around – or exactly equal to – the unconditional mean \( \mu_Y \); recall the discussion on the preceding page. But if there is a true relation between \( X \) and \( Y \), then it becomes important to characterize and model the resulting (nonzero) variation.