1 Why

Covariance has some painful disadvantages. There is no standard scale with which we could compare and say, the number obtained is high correlation. When we measure, say a distance of 10m, we do not just have the measure 10, we also *understand the size* of it because we have a standard scale for 1m. This allows us to compare with another distance, say 15m, and accurately understand the difference between them. This type of **standardization** or normalization is missing in our Covariance value.

Further, it is highly unit dependent as we are just multiplying two RVs of different units (the 3rd factor probability we multiply with, anyway is unitless). This means, if units change, our measure also could drastically change. Imagine the last example. If X and Y, the deductibles were in cents, then they just scale by 100 times in the summation. Note what this leads to.

$$\begin{aligned}
\text{Cov}(X, Y) &= \sum_x \sum_y (100x - 17500)(100y - 12500)p(x, y) \\
&= (100)(100) \sum_x \sum_y (x - 175)(y - 125)p(x, y) \\
&= 10000(1875) \\
&= 18750000 \quad \text{cents}^2
\end{aligned}$$

Apart from a very high value, note the ugly units tag sticking with it. Though a covariance could give us a measure, this is not as useful as a unit like meters. Ideally, we would wish, our measure is units independent. Summarizing,

Covariance's main disadvantages

- Critically dependent on units of random variables being compared
- Not comparable with other covariance values

2 What

The idea to tackle the issue is by, well as said, *standardization or normalization with something*, thereby making it a ratio, due to which the units cancel out between numerator and denominator. This already suggests we need two quantities of same units of X and Y in the denominator of Covariance. Let us recall the equation of simple linear regression model between two Random variables (figure 1).

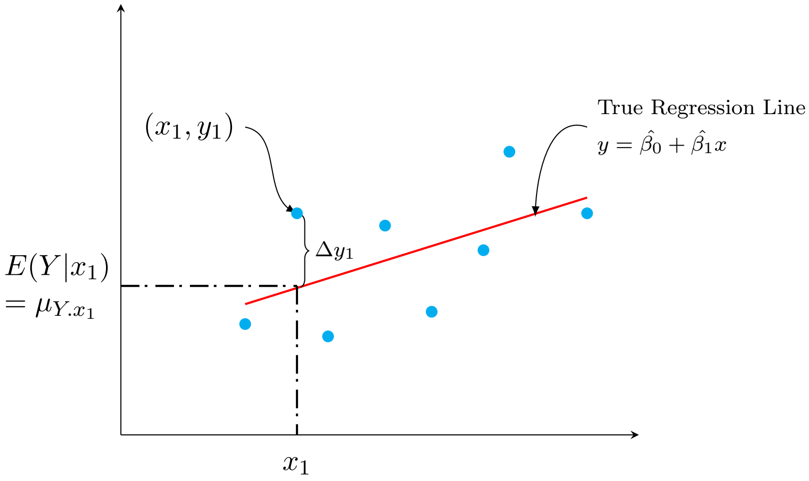

Figure 1: Recalling the regression line

The regression line is given by

$$\begin{aligned} E(Y|x) &= \beta_0 + \beta_1 x \\ \hat{Y}|x &= \hat{\beta_0} + \hat{\beta_1}x \\ \text{where} \quad \hat{\beta_1} &= \frac{\sum_i (y_i - \overline{y})(x_i - \overline{x})}{\sum_i (x_i - \overline{x})^2} \\ \hat{\beta_0} &= \overline{y} - \hat{\beta_1}\overline{x} \end{aligned} \tag{1}$$

What would it mean, when the slope $\hat{\beta_1}$ is 0 for this regression line?

$$\begin{aligned} \hat{\beta_1} &= 0 \\ \implies \hat{Y}|x &= \hat{\beta_0} = \overline{y} \end{aligned}$$

This is simply an horizontal line drawn parallel to x axis, cutting at $y = \overline{y}$. So, if such is the case, that for given sampe, $\hat{\beta_1}$ is 0, we could already say, their covariance is 0, because for any x, y remains constant at $\overline{y}$. This is illustrated in Figure 2.

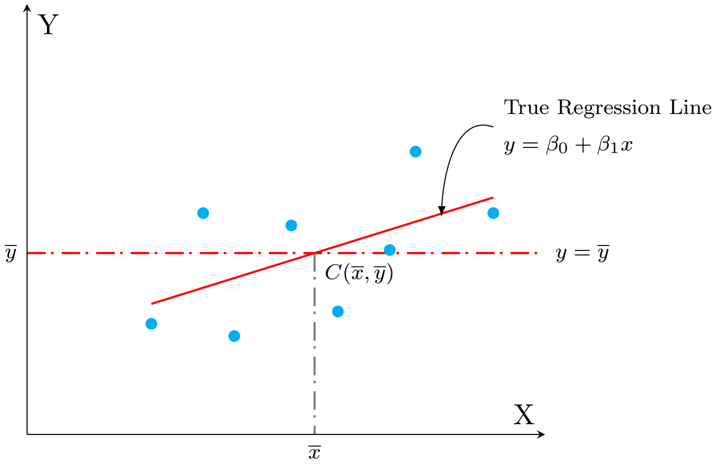

Figure 2: When the slope is zero..

Note in case of regression line, we took a variable X and evaluated the relationship of another variable Y via $E(Y|x)$. Thus naturally the reversed case is also possible that is $E(X|y)$. This is simply achieved by reversing the variables in regression line equation 1

$$\begin{aligned} E(X|y) &= \beta_2 + \beta_3 y \\ \hat{X}|y &= \hat{\beta}_2 + \hat{\beta}_3 y \\ \text{where} \quad \hat{\beta}_3 &= \frac{\sum_i (y_i - \overline{y})(x_i - \overline{x})}{\sum_i (y_i - \overline{y})^2} \\ \hat{\beta}_2 &= \overline{x} - \hat{\beta}_3 \overline{y} \end{aligned} \tag{2}$$

Again, when $\beta_3 = 0$, that is slope of regression line $E(X|y)$ is 0, we get,

$$\begin{aligned} \hat{\beta}_3 &= 0 \\ \implies \hat{X}|y &= \hat{\beta}_2 = \overline{x} \end{aligned}$$

Figure 3 illustrates plotting of both lines, along with zero correlation lines.

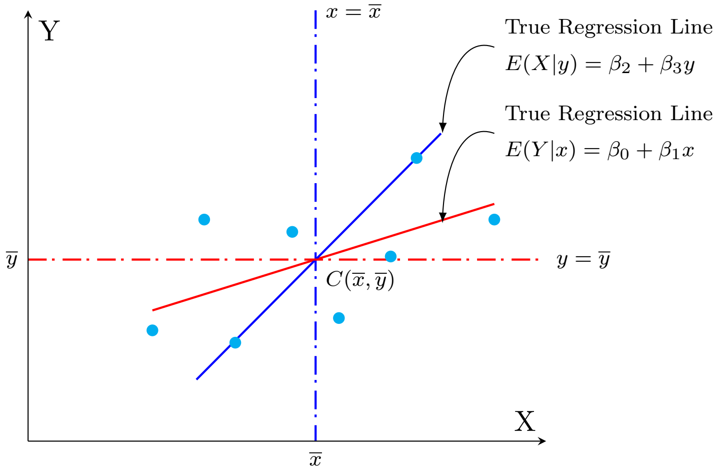

Figure 3: Two possible regression lines $E(Y|x), E(X|y)$

Summarizing, in current case of regression, we have,

- $E(Y|x)$ gives Y variation which is not same as variation indicated by $E(X|x)$
- $y = \overline{y}$ indicates zero variation of Y for any x, and $x = \overline{x}$, vice versa.

What we need is a single unified quantitative measure for reducing the disadvantages of Covariance. Note that we are dealing with samples, so our formula for *unbiased* sample covariance and variance, as referenced in Zaiontz [1], would be

$$\mathrm{cov}(X,Y) = \frac{1}{N-1}\sum_i^N (x_i - \overline{x})(y_i - \overline{y})$$

$$\mathrm{var}(X) = s_x^2 = \frac{1}{N-1}\sum_i^N (x_i - \overline{x})^2$$

$$\mathrm{var}(Y) = s_y^2 = \frac{1}{N-1}\sum_i^N (y_i - \overline{y})^2$$

Using them in the slopes, we get,

$$\hat{\beta_1} = \frac{\sum_i (y_i - \overline{y})(x_i - \overline{x})}{\sum_i (x_i - \overline{x})^2} = \frac{\frac{1}{N-1}\sum_i (y_i - \overline{y})(x_i - \overline{x})}{\frac{1}{N-1}\sum_i (x_i - \overline{x})^2}$$
$$= \frac{\mathrm{cov}(X,Y)}{s_x^2}$$

Similary for $\hat{\beta_3}$. Summarizing, now we have, slopes in terms of sample covariance and variances,

$$\hat{\beta_1} = \frac{\mathrm{cov}(X,Y)}{s_x^2} \quad , \quad \hat{\beta_3} = \frac{\mathrm{cov}(X,Y)}{s_y^2} \tag{3}$$

Thus,

$$\hat{Y}|x = \hat{\beta_0} + \frac{\mathrm{cov}(X,Y)}{s_x^2}x$$
$$\hat{X}|y = \hat{\beta_2} + \frac{\mathrm{cov}(X,Y)}{s_y^2}y$$

Now, covariance is symmetric. X is as covariant with Y as Y is with X. Check the formula again.

$$\mathrm{cov}(X,Y) = \frac{1}{N-1}\sum_i^N (x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{N-1}\sum_i^N (y_i - \overline{y})(x_i - \overline{x}) = \mathrm{cov}(Y,X)$$

However, as we saw, this cannot be said for $\hat{Y}|x$ and $\hat{X}|y$. But imagine below form for a moment.

$$\hat{Y}|x = 0 + \frac{\mathrm{cov}(X,Y)}{1}x$$
$$\hat{X}|y = 0 + \frac{\mathrm{cov}(X,Y)}{1}y$$

If we some how magically make the y-intercept of $\hat{Y}|x$, and x-intercept of $\hat{X}|y$ go away, and make the variance 1, we could have a symmetry effect for both $\hat{Y}|x$ and $\hat{X}|y$. This could be done by *standardizing* the sample set. Recall during Z transformation, we did the same. By shifting the sample set or distribution to its mean, and scaling by the standard deviation, we essentially achieve a standard distribution which could be comparable to any other standardized distribution (Recall Z scores). Such a standardized distribution will have 0 mean and variance as 1.

Lemma

For a population described by RV, $X(\mu, \sigma^2)$

$$
\begin{aligned}
Z &= \frac{X-\mu}{\sigma} \\
E(Z) &= E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}\big(E(X)-\mu\big) = \frac{1}{\sigma}(\mu-\mu) = 0 \\
\mathrm{Var}(Z) &= \mathrm{Var}\left(\frac{X-\mu}{\sigma}\right) = \mathrm{Var}\left(\frac{X}{\sigma}-\frac{\mu}{\sigma}\right) = \mathrm{Var}\left(\frac{X}{\sigma}\right) = \frac{1}{\sigma^2}\mathrm{Var}(X) = \frac{\sigma^2}{\sigma^2} = 1
\end{aligned}
$$

Standardizing our sample set

Applying the same principles to our sample set, if we transform as follows,

$$X_s = \frac{X-\overline{x}}{s_X} \quad , \quad Y_s = \frac{Y-\overline{y}}{s_Y}$$

where s_X, s_Y are the standard deviation of X and Y respectively, then, we have new samples set (X_s, Y_s), where

$$
\begin{aligned}
\overline{x_s} &= \overline{y_s} = 0 \\
s_{X_s} &= s_{Y_s} = 1
\end{aligned}
$$

The new standardized set gives rise to new regression lines as follows.

$$
\begin{aligned}
\hat{Y_s}|x_s &= \hat{\beta_{0s}} + \frac{\mathrm{cov}(X_s, Y_s)}{s_{X_s}^2} x_s \\
\hat{X_s}|y_s &= \hat{\beta_{2s}} + \frac{\mathrm{cov}(X_s, Y_s)}{s_{Y_s}^2} y_s
\end{aligned}
$$

Using equations 1, and 2 we get,

$$
\begin{aligned}
\hat{\beta_{0s}} &= \overline{x_s} - \hat{\beta_{1s}}\overline{y_s} = 0 - \hat{\beta_{1s}}(0) = 0 \\
\hat{\beta_{2s}} &= \overline{y_s} - \hat{\beta_{3s}}\overline{x_s} = 0 - \hat{\beta_{3s}}(0) = 0
\end{aligned}
$$

Using that, and since $s_{X_s} = s_{Y_s} = 1$, we finally get new regression lines as,

$$
\begin{aligned}
\hat{Y_s}|x_s &= \mathrm{cov}(X_s, Y_s)x_s \\
\hat{X_s}|y_s &= \mathrm{cov}(X_s, Y_s)y_s
\end{aligned}
$$

Figure 4 illustrates the resultant regression lines. One could notice both these lines are symmetric because they both have same slope with respect to their independent axis.

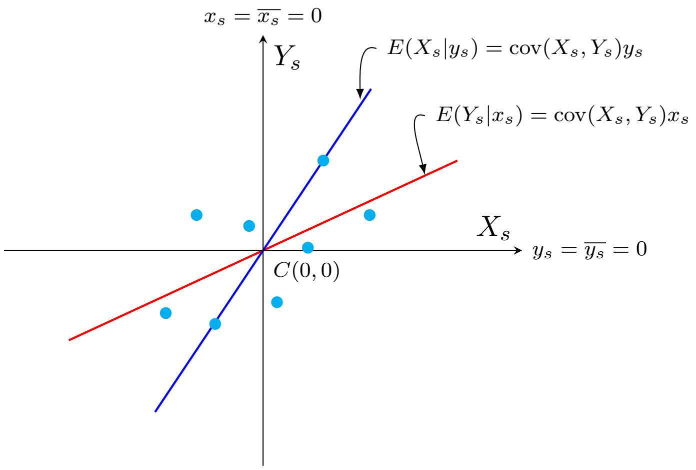

Figure 4: Two standardized regression lines $E(Y_s|x_s), E(X_s|y_s)$

The new *standardized* sample covariance $\text{cov}(X_s, Y_s)$ has very useful properties we have been longing so far.

- $\text{cov}(X_s, Y_s)$ would be now unitless and would vary between ±1 as we would observe shortly
- the covariance is now made symmetric, that is X_s is as covariant with Y_s as Y_s is with X_s
- this does not mean, the new regression lines are same. They just have same slope meaning they are *symmetric*

All the above points would become evident, once we observe a detailed example.

Covariance of Standardized Sample Sets

By standardizing the sample set, we are able to achieve interesting *symmetric* regression lines of same slope

$$\begin{aligned} \hat{Y_s}|x_s &= \text{cov}(X_s, Y_s)x_s \\ \hat{X_s}|y_s &= \text{cov}(X_s, Y_s)y_s \end{aligned} \tag{4}$$

where $\text{cov}(X_s, Y_s)$ is unitless and varies between ±1

References

[1] C. Zaiontz. Basic concepts of correlation. 2013. URL http://www.real-statistics.com/correlation/basic-concepts-correlation/.