Covariance and Correlation

Parthiban Rajendran

November 14, 2018

1 Examples

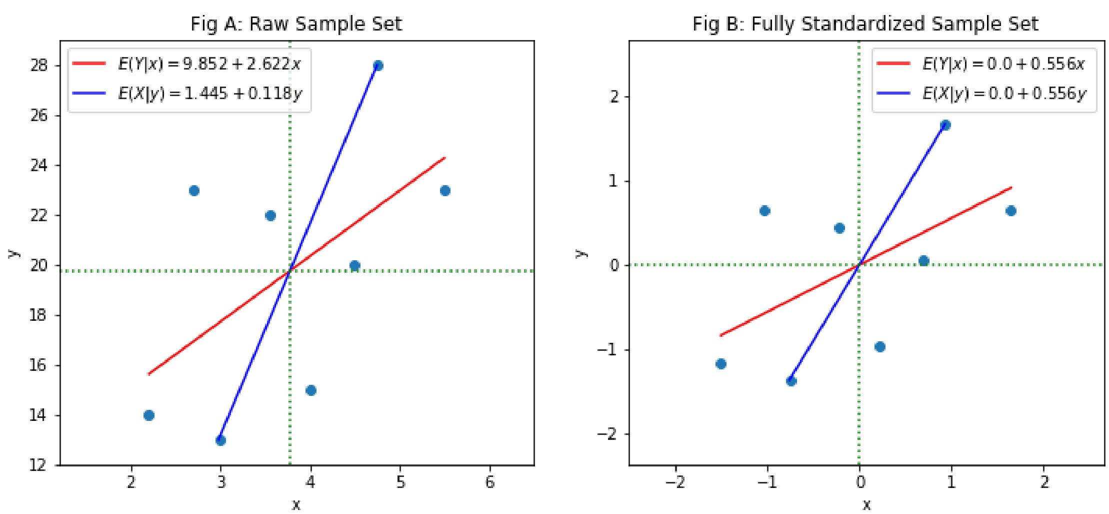
1.0.0.1 Example 1: A single simple sample set

Assume below is the given sample set. Let us plot both the direct simple regression line and standardized one to note the differences.

X	Y
2.2	14
2.7	23
3	13
3.55	22
4	15
4.5	20
4.75	28
5.5	23

```
In[17]: x_i = [2.2, 2.7, 3, 3.55, 4, 4.5, 4.75, 5.5]    # a sample set
        y_i = [14, 23, 13, 22, 15, 20, 28, 23]

        fig, axr = plt.subplots(1,2, figsize=(12,5))
        plot_regs(x_i, y_i, axr[0], std=False, label='Fig A: Raw Sample Set')
        plot_regs(x_i, y_i, axr[1], std=True, std_full=True, label='Fig B: Fully Standardized
        Sample Set')
        plt.show()
```



Note the regression line equations in both figures. In Figure B, as expected, both lines get the same slope which is *standardized covariance* $Cov(X_s, Y_s)$. Note the value of the common slope. It is positive and less than 1, this tells both sample sets are related linearly to an extent.

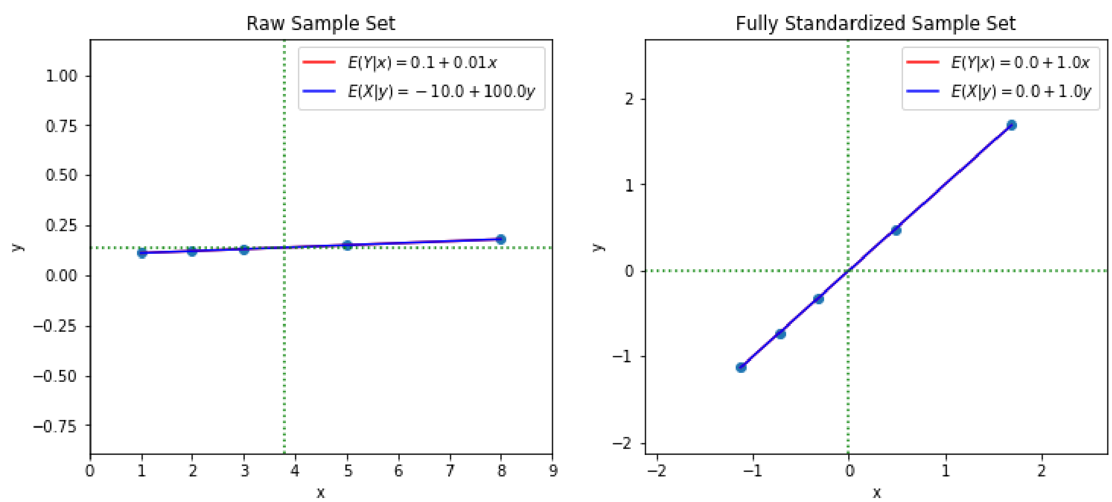
1.0.0.2 Example 2: Wikipedia Sample set

Let us try a perfectly covarying example. This is taken from Wikipedia's Pearson Correlation Coefficient article [1]

X	Y
1	0.11
2	0.12
3	0.13
5	0.15
8	0.18

```
In[18]: x_i = [1,2,3,5,8]    # a sample set
        y_i = [0.11,0.12,0.13,0.15,0.18]

        fig, axr = plt.subplots(1,2, figsize=(12,5))
        plot_regs(x_i, y_i, axr[0], std=False, label='Raw Sample Set')
        plot_regs(x_i, y_i, axr[1], std=True, std_full=True, label='Fully Standardized Sample
        Set')
        plt.show()
```

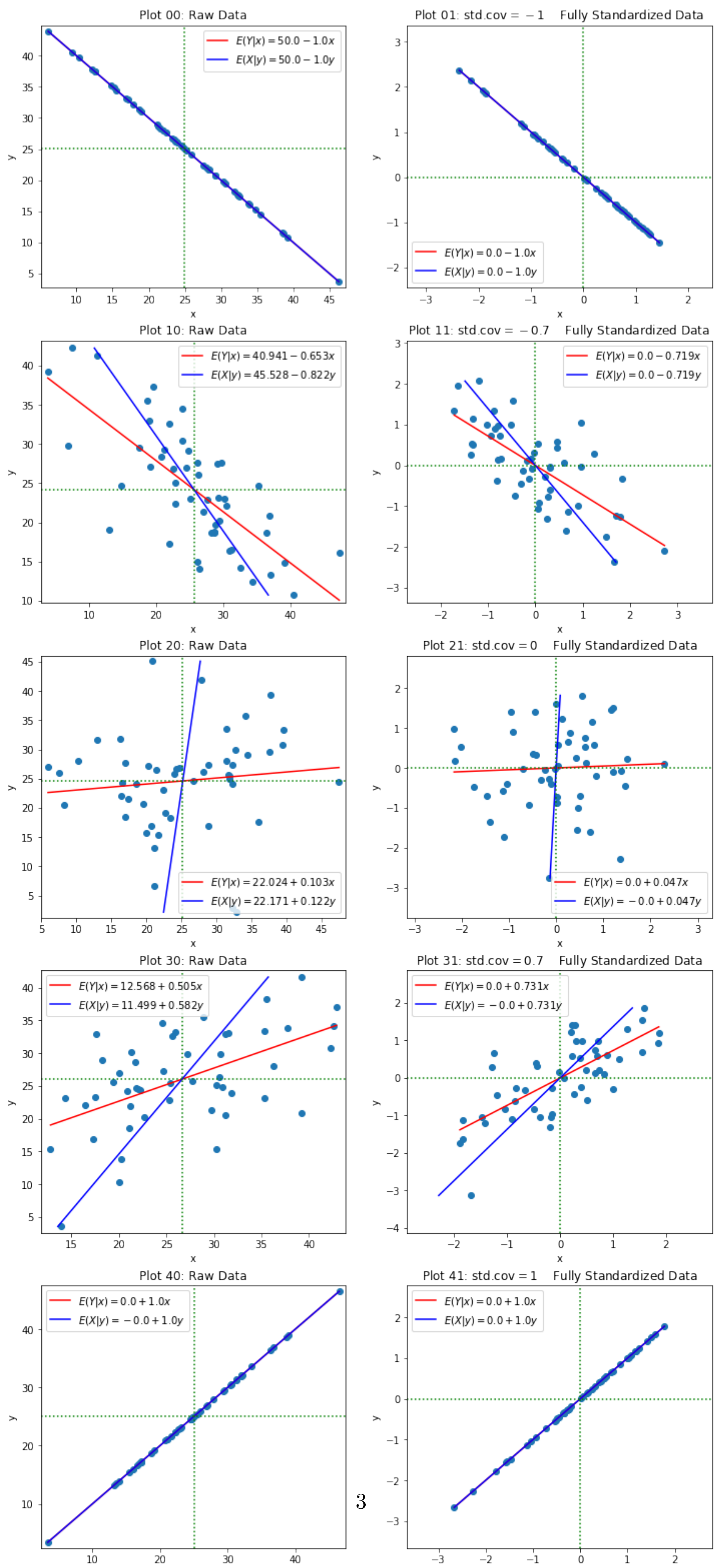


Aha! When the dataset is perfectly linearly related, we get the *standardized covariance* slope as 1. Ain't we getting somewhere?

1.0.0.3 Example 3: With different linear relationships

To test the different values of standardized covariance, we shall generate different datasets, that has perfect linearity in both directions (positive and negative), and also some what in the middle, including no linearity.

[1] https://en.wikipedia.org/wiki/Pearson_correlation_coefficient


Plot 00: Raw Data
E(Y|x) = 50.0 − 1.0x
E(X|y) = 50.0 − 1.0y
Plot 01: std.cov = −1   Fully Standardized Data
E(Y|x) = 0.0 − 1.0x
E(X|y) = 0.0 − 1.0y
Plot 10: Raw Data
E(Y|x) = 40.941 − 0.653x
E(X|y) = 45.528 − 0.822y
Plot 11: std.cov = −0.7   Fully Standardized Data
E(Y|x) = 0.0 − 0.719x
E(X|y) = 0.0 − 0.719y
Plot 20: Raw Data
E(Y|x) = 22.024 + 0.103x
E(X|y) = 22.171 + 0.122y
Plot 21: std.cov = 0   Fully Standardized Data
E(Y|x) = 0.0 + 0.047x
E(X|y) = −0.0 + 0.047y
Plot 30: Raw Data
E(Y|x) = 12.568 + 0.505x
E(X|y) = 11.499 + 0.582y
Plot 31: std.cov = 0.7   Fully Standardized Data
E(Y|x) = 0.0 + 0.731x
E(X|y) = −0.0 + 0.731y
Plot 40: Raw Data
E(Y|x) = 0.0 + 1.0x
E(X|y) = −0.0 + 1.0y
Plot 41: std.cov = 1   Fully Standardized Data
E(Y|x) = 0.0 + 1.0x
E(X|y) = 0.0 + 1.0y
x
y

Note carefully.

- When the given dataset is perfectly negatively linearly related (Plot 00,01), $\mathrm{cov}(X_s, Y_s) = -1$
- When the given dataset is somewhat negatively linearly related (Plot 10,11), $-1 < \mathrm{cov}(X_s, Y_s) < 0$
- When the given dataset is totally not linearly related (Plot 20,21), $\mathrm{cov}(X_s, Y_s) = 0$
- When the given dataset is somewhat positively linearly related (Plot 30,31), $0 < \mathrm{cov}(X_s, Y_s) < 1$
- When the given dataset is perfectly positively linearly related (Plot 40,41), $\mathrm{cov}(X_s, Y_s) = 1$

Thus, not only that our standardized covariance got rid of units, but also retains value between ± 1, perfectly reflective of the linear relationship in the dataset. Thus we observe empirically via examples the range of standardized covariance.