Analysis of Random Walks in a Grid City

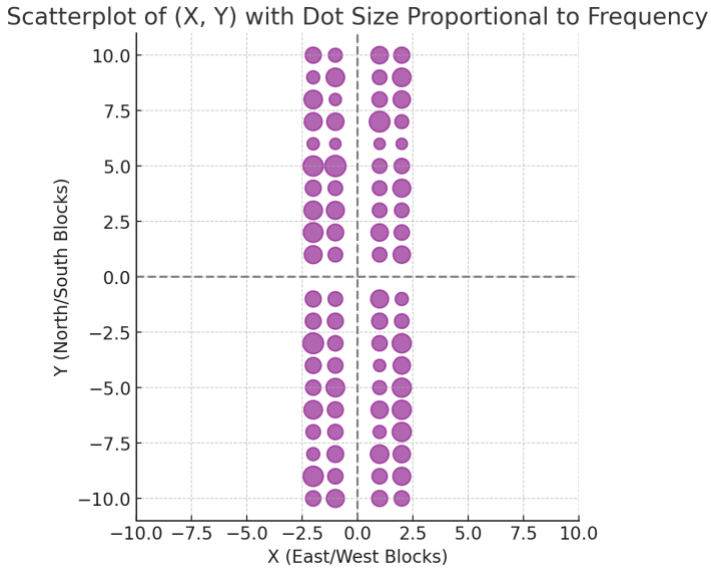
An example to show why the coefficient of determination doesn't mean what everybody thinks it means.

1. Description of the Simulation

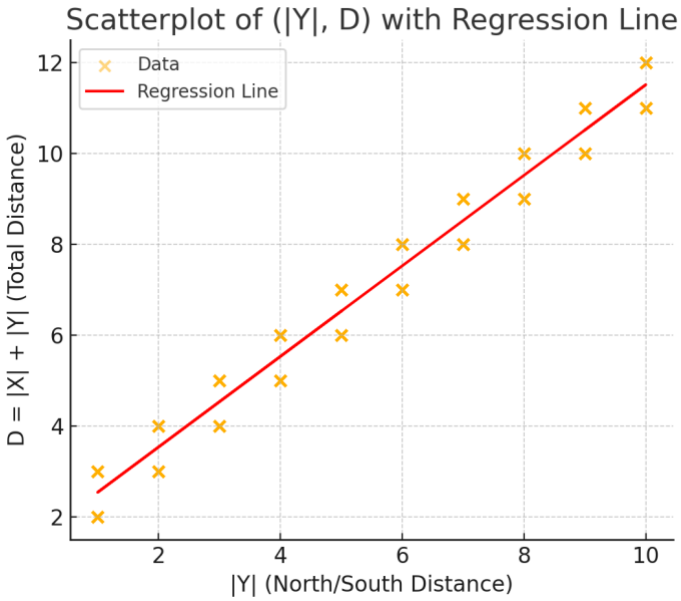
In this simulation, a person walks in a grid-based city. Each day, they randomly choose:

- A number of blocks east or west (X), uniformly from -2, -1, 1, 2 (excluding 0).
- A number of blocks north or south (Y), uniformly from -10 to -1 and 1 to 10 (excluding 0).

The total walking distance D is defined as the sum of the absolute values of X and Y:
$D = |X| + |Y|$.



2. Relationship Between |Y| and D



- Correlation coefficient (R): 0.985
- R^2 (Variance explained by |Y|): 0.970
- Standard deviation of |X|: 0.499 blocks
- Standard deviation of |Y|: 2.862 blocks
- Standard deviation of D: 2.896 blocks

3. Interpreting R^2: A Caution

While it's true that "97% of the variance in D is explained by |Y|," the unit of variance is blocks squared. This can be misleading because people don't walk in squared blocks — they walk in blocks. Using R^2 to interpret real-world uncertainty often exaggerates the clarity of predictions. If you know how far you have want north/south (|Y|) you still have about ½ block uncertainty in your total distance D.

4. Real-World Uncertainty

The standard deviation of D, which measures actual variation in how far someone walks, is 2.90 blocks. The standard deviation of |X| is 0.50 blocks — about 17.2% of the total variation. This means that even if you knew exactly how far someone walked north/south, there's still some uncertainty — about 17% of the total variation — due to the unpredictable east/west component of the walk. So |knowing Y leaves about 17% of the variation unexplained not 3%, by any common sense interpretation.

5. Interpreting Variation with Mean Absolute Deviation

Shouldn't it be 20% exactly? I think it should but you would have to change the way you measure variation. While standard deviation and variance are common ways to measure variation, they rely on squaring differences, which can make interpretation less intuitive. An alternative is to use the mean of the absolute deviation from the mean — a more direct and understandable measure of variation in real-world units (blocks).

In this simulation:

- The mean absolute deviation of D is 2.49 blocks.
- The mean absolute deviation of |X| is 0.50 blocks.
- The mean absolute deviation of |Y| is 2.50 blocks.

This shows that |X| contributes about 0.50 out of 2.49 blocks of average deviation in total distance — or roughly 20%. Similarly, |Y| contributes about 80%. These percentages offer a more intuitive breakdown of where variability in walking distance comes from — without resorting to squared units like variance.

6. The Reduction in Error Statistic: A Better Measure of Explained Variation

While the coefficient of determination (R^2) is widely used to measure how much of the variation in a response variable is explained by a predictor, it works with squared deviations — which are in units like blocks^2. This can lead to misleading conclusions when applied to real-world quantities measured in linear units like blocks.

A more interpretable alternative is the Reduction in Error (RE) statistic, which is defined as:

$$RE = 1 - (\text{Residual SD} / \text{SD of response})$$

This statistic tells us how much the standard deviation (i.e., real-world uncertainty) is reduced by knowing the predictor.

In this case:

- SD of total distance D = 2.90 blocks
- Residual SD after accounting for |Y| = 0.68 blocks
- RE = 1 – (0.68 / 2.90) ≈ 0.7652 or 76.5%

This means that knowing |Y| reduces the actual uncertainty in how far someone walks by about 76.6% — a more grounded and practical interpretation than the R^2 value of 97%.