В изложении доцента БГТУ Сахарова В.Ю.

Пусть произведено n взаимно независимых экспериментов с двумерной случайной величиной (X, Y). Получены пары чисел (X_1, Y_1), (X_2, Y_2), … , (X_n, Y_n). Если построить в Microsoft Office Excel точечную диаграмму, и она будет напоминать эллипс, называемый «эллипс рассеяния», то можно визуально определить прямую линию, на которой лежит большая полуось этого эллипса. Допустим, эта прямая имеет уравнение $y=kx+b$. Фактически это означает, что случайные величины X и Y связаны корреляционной зависимостью: $Y=kX+b+\varepsilon$. Где сумма ($b+\varepsilon$) это некоторая случайная величина. Последнее предложение можно сформулировать несколько по другому: где ε это случайная величина с математическим ожиданием равным нулю. Число k называют «коэффициентом регрессии» Y на X. Если эллипс рассеяния выродился в отрезок, то это означает, что $\varepsilon=0$ и случайные величины X и Y связаны не корреляционной зависимостью, а функциональной: $Y=kX+b$.

Возникает вопрос, как найти параметры k и b не визуально, а с помощью расчётов. А для этого сначала нужно решить, по какому принципу следует предпочесть одну прямую вида $y=kx+b$ другой. Естественно хотелось бы, чтобы разница между Y и суммой ($kX+b$) была бы поменьше. Для каждой пары чисел (X_i, Y_i), где $i=1,2,3,\ldots,n$, можно найти расстояние между числами Y_i и (kX_i+b), как модуль их разности $|Y_i-(kX_i+b)|$. Но таких модулей разностей будет много (а именно n). И проблема в том, что все такие разности одновременно при одних и тех же значениях k и b сделать равными нулю не удастся (разумеется, при $\varepsilon=0$ такое всё-таки возможно). Вспомним, как находится расстояние между двумя точками в многомерном эвклидовом пространстве. Это квадратный корень из суммы квадратов разностей координат точек. Если применить формулу Эвклида к рассматриваемому случаю, то расстояние в n-мерном пространстве между точкой (Y_1, Y_2, … , Y_n) и точкой (kX_1+b , kX_2+b , … , kX_n+b) будет равно $\sqrt{\sum_{i=1}^{n}\left(Y_i-\left(kX_i+b\right)\right)^2}$. Интуитивно ясно, что чем меньше это расстояние, тем лучше. Разумеется, если кто-то не согласен с таким способом нахождения наилучшего приближения, то ему придётся высказать свою идею о том, что такое хорошо при нахождении k и b.

Поскольку квадратный корень является возрастающей функцией, то для того, чтобы он был поменьше, достаточно минимизировать подкоренное выражение, являющееся функцией двух аргументов: $f(k,b)=\sum_{i=1}^{n}\left(Y_i-\left(kX_i+b\right)\right)^2$. Этот метод работы называется «метод наименьших квадратов». Согласно теоремам математического анализа функция двух переменных будет достигать минимума в той точке, где обе её частные производные равны нулю. Найдём эти производные:

$$\frac{\partial f(k,b)}{\partial k}=\frac{\partial}{\partial k}\left(\sum_{i=1}^{n}\left(Y_i-\left(kX_i+b\right)\right)^2\right)=\sum_{i=1}^{n}\frac{\partial}{\partial k}\left(\left(Y_i-\left(kX_i+b\right)\right)^2\right)=$$

$$=\sum_{i=1}^{n}2\left(Y_i-\left(kX_i+b\right)\right)\frac{\partial}{\partial k}\left(Y_i-\left(kX_i+b\right)\right)=\sum_{i=1}^{n}2\left(Y_i-\left(kX_i+b\right)\right)\left(-X_i\right)=2\sum_{i=1}^{n}X_i\left(\left(kX_i+b\right)-Y_i\right)=$$

$$=2\sum_{i=1}^{n}X_i\left(kX_i+b-Y_i\right)=2\sum_{i=1}^{n}\left(k\left(X_i\right)^2+bX_i-X_iY_i\right)=2\left(\sum_{i=1}^{n}k\left(X_i\right)^2+\sum_{i=1}^{n}bX_i-\sum_{i=1}^{n}X_iY_i\right)=$$

$$=2\left(k\sum_{i=1}^{n}\left(X_i\right)^2+b\sum_{i=1}^{n}X_i-\sum_{i=1}^{n}X_iY_i\right)$$

$$\frac{\partial f(k,b)}{\partial b}=\frac{\partial}{\partial b}\left(\sum_{i=1}^{n}\left(Y_i-\left(kX_i+b\right)\right)^2\right)=\sum_{i=1}^{n}\frac{\partial}{\partial b}\left(\left(Y_i-\left(kX_i+b\right)\right)^2\right)=$$

$$= \sum_{i=1}^{n} 2\left(Y_i - \left(kX_i + b\right)\right)\frac{\partial}{\partial b}\left(Y_i - \left(kX_i + b\right)\right) = \sum_{i=1}^{n} 2\left(Y_i - \left(kX_i + b\right)\right)\left(-1\right) = 2\sum_{i=1}^{n}\left(\left(kX_i + b\right) - Y_i\right) =$$

$$= 2\sum_{i=1}^{n}\left(kX_i + b - Y_i\right) = 2\left(\sum_{i=1}^{n} kX_i + \sum_{i=1}^{n} b - \sum_{i=1}^{n} Y_i\right) = 2\left(k\sum_{i=1}^{n} X_i + bn - \sum_{i=1}^{n} Y_i\right)$$

Теперь приравняем эти обе частные производные нулю

$$\begin{cases} 2\left(k\sum_{i=1}^{n}\left(X_i\right)^2 + b\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i Y_i\right) = 0 \\ 2\left(k\sum_{i=1}^{n} X_i + bn - \sum_{i=1}^{n} Y_i\right) = 0 \end{cases}$$

$$\begin{cases} k\sum_{i=1}^{n}\left(X_i\right)^2 + b\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i Y_i = 0 \\ k\sum_{i=1}^{n} X_i + bn - \sum_{i=1}^{n} Y_i = 0 \end{cases}$$

$$\begin{cases} k\sum_{i=1}^{n}\left(X_i\right)^2 + b\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} X_i Y_i \\ k\sum_{i=1}^{n} X_i + bn = \sum_{i=1}^{n} Y_i \end{cases}$$

Полученную линейную систему двух уравнений относительно двух переменных можно решить, например, сразу записав ответы по формулам Крамера:

$$\begin{cases} k = \dfrac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \cdot \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n}\left(X_i\right)^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \\ b = \dfrac{\sum_{i=1}^{n}\left(X_i\right)^2 \cdot \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_i \cdot \sum_{i=1}^{n} X_i Y_i}{n\sum_{i=1}^{n}\left(X_i\right)^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \end{cases}$$

Результаты вычисленные по этим формулам, можно сравнить с теоретическими, взятыми из литературы:

$$\begin{cases} k = \dfrac{\rho_{XY} \cdot \sigma(Y)}{\sigma(X)} \\ b = EY - \dfrac{\rho_{XY} \cdot \sigma(Y) \cdot EX}{\sigma(X)} \end{cases}$$