Basic Statistics

Sudipta Das

Assistant Professor,
Department of Computer Science,
Ramakrishna Mission Vivekananda Educational & Research Institute

Outline I

1. Regression
   - Method of Curve Fitting

Curve Fitting: Problem I

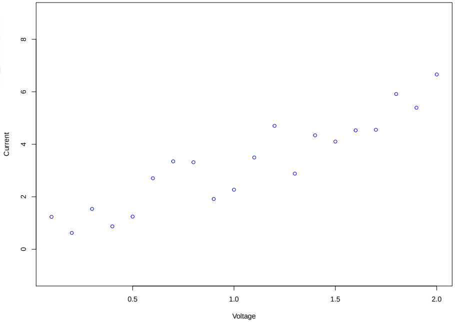
- A High School Problem in Physics:-
  - Measuring the **conductance** (=1/resistance) of a device

Curve Fitting: Problem II

- Experiment Result

Voltage(V_x)	0.1	0.2	0.3	0.4	...	1.9	2
Current(I_y)	1.236	0.622	1.537	0.873	...	5.392	6.661

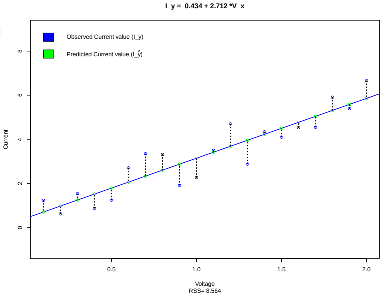
Current (I_y) vs. Voltage (V_x)



Solution I

- Intuitive Solution
  - Draw a straight line, goes through origin such that the distances of the observed points are not far from that straight line
  - Then report the slop of that straight line as the conductance of the device

Current (I_y) vs. Voltage (V_x) with Prediction Line



Solution II

- Mathematically, among numerous straight lines, passing through the origin, choose the one which **minimize**

$$\Delta = \sum_{i=1}^{n} (I_{y_i} - I_{\hat{y}_i})^2$$

  - I_{y_i} : Observed value of current in experiment at voltage V_{x_i}
  - $I_{\hat{y}_i}$: Predicted value of current at voltage V_{x_i}, calculated as

$$I_{\hat{y}_i} = c + mV_{x_i}$$

  - $(I_{y_i} - I_{\hat{y}_i})$: Error by prediction at voltage V_{x_i}

Solution III

- Question:- Which pair of (c, m) will *minimize* Δ?
- Ans:-

$$
\begin{aligned}
\frac{\delta}{\delta c}\Delta &= \frac{\delta}{\delta c}\sum_{i=1}^{n}(l_{y_i} - l_{\hat{y}_i})^2 \\
&= \frac{\delta}{\delta c}\sum_{i=1}^{n}(l_{y_i} - c - mV_{x_i})^2 = \sum_{i=1}^{n}2(l_{y_i} - c - mV_{x_i})(-1) = 0
\end{aligned}
$$

$$
\begin{aligned}
\frac{\delta}{\delta m}\Delta &= \frac{\delta}{\delta m}\sum_{i=1}^{n}(l_{y_i} - l_{\hat{y}_i})^2 \\
&= \frac{\delta}{\delta m}\sum_{i=1}^{n}(l_{y_i} - c - mV_{x_i})^2 = \sum_{i=1}^{n}2(l_{y_i} - c - mV_{x_i})(-V_{x_i}) = 0
\end{aligned}
$$

Solution IV

Therefore,

$$m = \frac{\sum_{i=1}^{n} (I_{y_i} - \bar{I}_y)(V_{x_i} - \bar{V}_x)}{\sum_{i=1}^{n} (V_{x_i} - \bar{V}_x)^2} = \frac{Cov(I_y, V_x)}{Var(V_x)}, (\text{say } \hat{\beta}_1)$$

and

$$c = \bar{I}_y - m\bar{V}_x, (\text{say } \hat{\beta}_0).$$

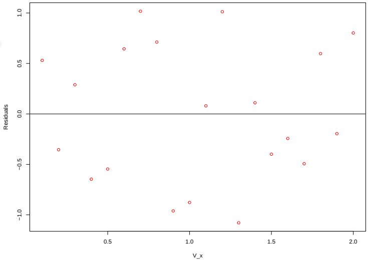
Least Square method I

- The method of finding the 'best' line by minimizing the Δ is called Least Square method
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the least square **predicted stray current** and **predicted conductance**, respectively of the **true stray current** and **true conductance** , respectively.

Least Square method II

- Outcome by the method of least square:
  - $\hat{\beta}_0$ and $\hat{\beta}_1$ are predicted values of the true stray current and the true conductance, respectively.
  - Error/Residual ($E_i = I_{y_i} - I_{\hat{y}_i}$) at each point (V_{x_i}, I_{y_i}); $i = (1, \dots, n)$.

Residual (E) vs. Voltage (V_x) with Prediction Line



Least Square method III

- We can calculate Residual Sum of Squares

$$RSS = \Delta_{min} = \sum_{i=1}^{n} E_i^2$$

- A scaled value of Δ_{min}

$$RSE = \sqrt{\frac{\Delta_{min}}{n-2}}$$

  - Lesser RSE (Residual Standard Error) better prediction

Least Square method IV

- Multiple R-squared

$$R^2 = 1 - \frac{RSS}{TSS}$$

where, $RSS = \Delta_{min}$ and

TSS (Total Sum of Squares) $= \sum_{i=1}^{n}(l_{y_i} - \bar{l}_y)^2.$

  - Higher R^2 better estimate