Correlation and Regression

.. the relationship between two or more variables

ex: relationship between linear velocity; displacement

$$x = x_0 + v_0 t$$

.. deterministic relationship;
predicts outcome perfectly

.. many relationships in engineering are not deterministic!

ex: gas consumption and vehicle weight

.. sure, they're related; can't predict one from the other by itself

this course: one independent predictor variable (x)
one linearly-related response (y)

.. technique of modelling and exploring relationships:
regression analysis

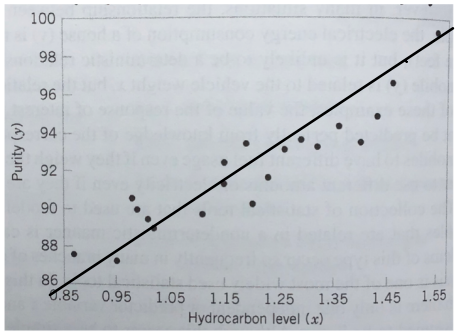
Simple Linear Regression

∴ let's look at some data;

purity of oxygen concentration (y)
vs. hydrocarbon level (x)

TABLE 11.1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x(%)	Purity y(%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33



↑ Scatter diagram; excellent place to start!

∴ already looks like a strong linear relationship

.. in fact, we could "eyeball" a straight line through these points; probably pretty good!

regression line:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

(β_0: intercept; β_1: slope; ε: random error term)

simple linear regression model

.. this is a better way to express the regression line:

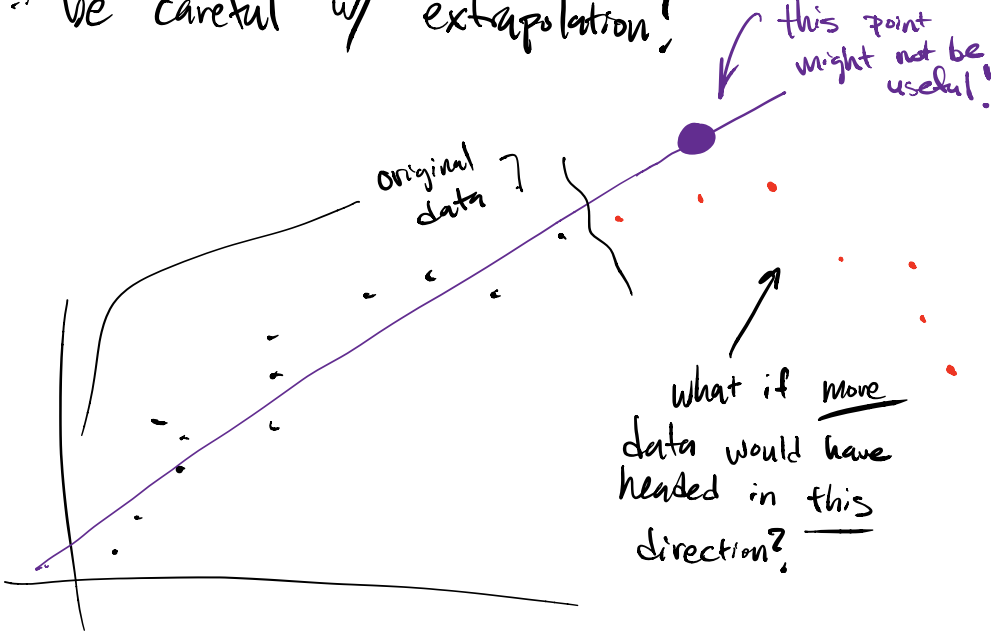
$$E(Y|_x) = \mu_{Y|x} = \beta_0 + \beta_1 x$$

.. demonstrates that each point predicted by line is really a mean value; actual data points would have probability distribution associated with them, about that mean value

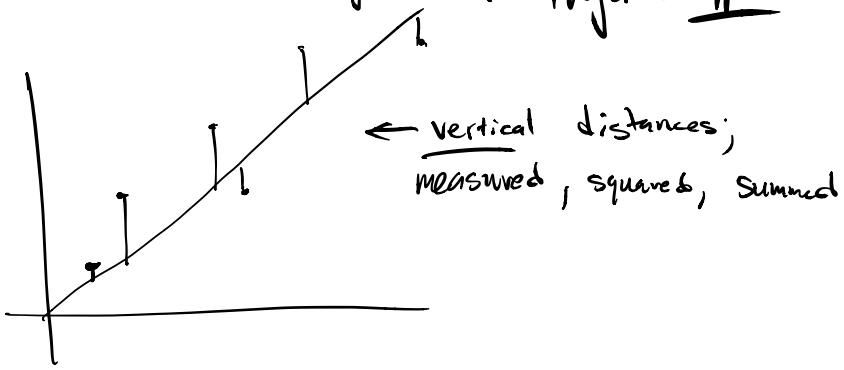
Caveats of regression:

1.) avoid developing statistically-significant relationships between non-causal variables; i.e. shear strength of spot welds vs. # of empty parking spaces in visitor lot!

2.) use caution when using a regression line outside the range of original data ∴ be careful w/ extrapolation!



- we need to determine slope and intercept
- what's the best way to do this?
- Gauss proposed minimizing the sum of the squares of the errors
- "method of least squares"
- graphical error analysis in Physics II:



- difference between actual data point and the y-value predicted by regression line at that value of x is what we square (and sum)