공정 모형 및 해석
최소 자승법

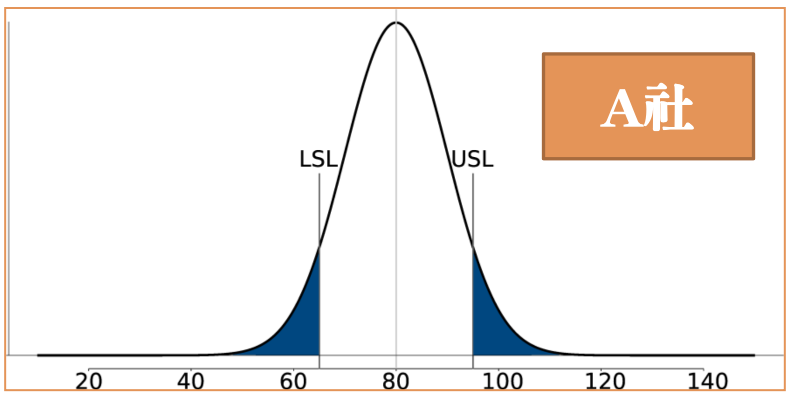
Jay Liu

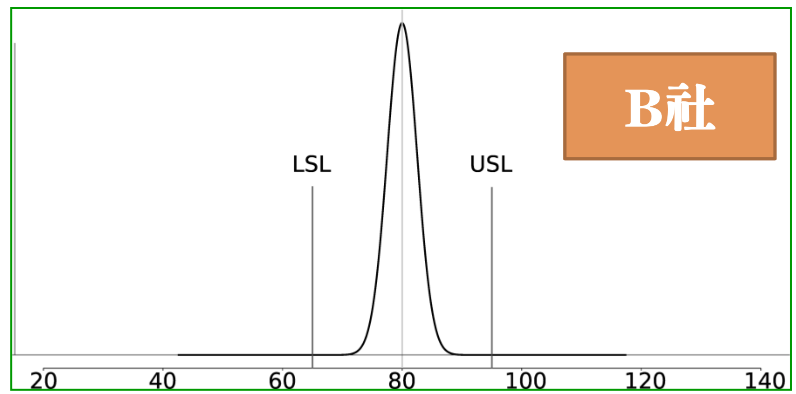
Dept. Chemical Engineering

PKNU

[FYI]Process capability (공정능력)

- Suppose you need to choose a raw material supplier among company A and company B. You received a database containing quality of a raw material from each company and plotted them with spec. limits (LSL and USL) that you product requests. Which one would you choose?





- How to quantify this capability?
- Which statistics are useful in describing this capability?

[FYI]Process capability (Cont.)

- C_p (or PCR, process capability ratio)

$$C_P = \frac{USL - LSL}{6\sigma}$$

- C_{pk} (or PCR_k) for one-sided limit

$$C_{pk} = \min\left(\frac{USL - \mu}{3\sigma}, \frac{\mu - LSL}{3\sigma}\right)$$

μ and σ: calculated from data

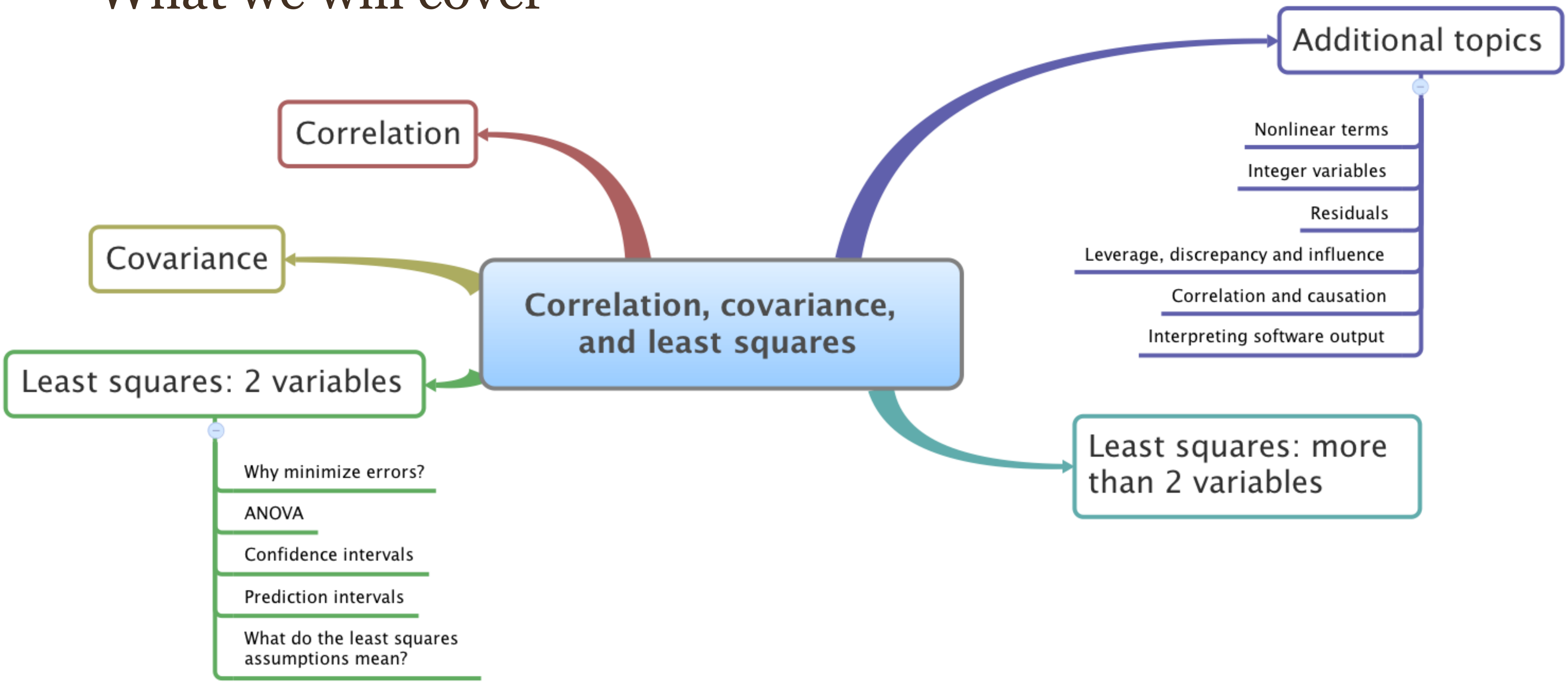
- In general, C_p (or C_{pk}) = 1.33 is minimum requirement

※Stat > quality tools > capability analysis

※Note: Cpk and Cp are only useful for a process which is stable

Least squares regression (최소자승회귀법)

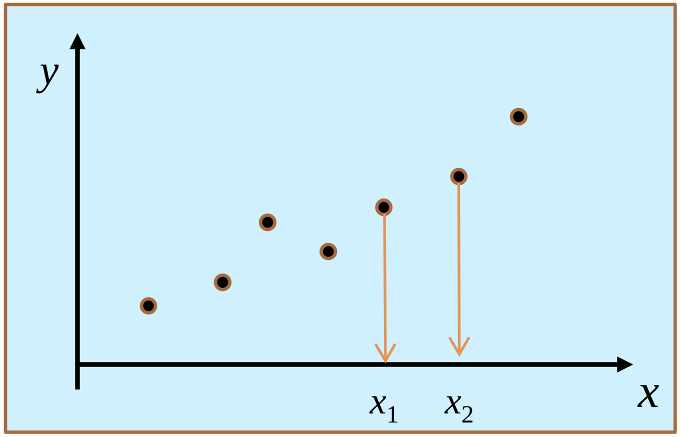
- What we will cover



Box, G.E.P., Use and abuse of regression, *Technometrics*, 8 (4), 625-629, 1966

[FYI]Least squares vs. interpolation

- Given the data, there are two choices when we want to know the value of y at $x = (x_1 + x_2)/2$



x	y
…	…
…	…
x_1	y1
x_2	y_2
…	…

  - least squares? or interpolation?
- Interpolation is recommended when data are subject to negligible experimental error (or noise)
  - Ex. In using steam tables
- Otherwise, least squares is recommended.

Least squares - usage examples (사용 예)

- Quantify relationship between 2 variables (or 2 sets of variables):
  - Manager: How does yield from the lactic acid batch fermentation relate to the purity of sucrose?
  - Engineer: The yield can be predicted from sucrose purity with an error of plus/minus 8%
  - Manager: And how about the relationship between yield and glucose purity?
  - Engineer: Over the range of our historical data, there is no discernible relationship.

Least squares - usage examples

- Two general applications
  - Predictive modeling – usually when an exact model form is unknown.
    - Modeling data trends in order to predict future y values
  - Simulation – usually when parameters in the model are unknown.
    - Getting parameter values in the known model form (e.g., calculate activation energy from reaction data)
- Terminology (용어)
  - y : response variables, output variables, dependent variables,
  - x : input variables, regressor variables, independent variables

Review: covariance (공분산)

- Consider measurements from a gas cylinder: temperature (K) and pressure (kPa).
- Ideal gas law applies under moderate condition: $pV = nRT$
  - Fixed volume, $V = 20 \times 10{-}3m3 = 20$ L
  - Moles of gas, $n = 14.1$ mols of chlorine gas, (1 kg gas)
  - Gas constant, $R = 8.314$ J/(mol.K)
- Simplify the ideal gas law to: $p = \beta_1 T$, where

$$\beta_1 = \frac{nR}{V}$$

Review: covariance (Cont.)

	Cylinder temperature (K)	Cylinder pressure (kPa)	Room humidity (%)
	273	1600	42
	285	1670	48
	297	1730	45
	309	1830	49
	321	1880	41
	333	1920	46
	345	2000	48
	357	2100	48
	369	2170	45
	381	2200	49
Mean	327	1910	46.1
Variance	1320	43267	8.1

Review: covariance (Cont.)

- Formal definition:

$$\mathrm{cov}(x,y) = E\left\{\left(x-\bar{x}\right)\left(y-\bar{y}\right)\right\} \quad \text{where } E(z) = \bar{z}$$

1. Calculate deviation variables: $T-\bar{T}$ and $p-\bar{p}$
   - Subtracting off mean centers the vector at zero.
2. Multiply the centered values: $\left(T-\bar{T}\right)\left(p-\bar{p}\right)$
   - 16740 10080 5400 1440 180 60 1620 5700 10920 15660
3. Calculate the expected value (mean): 6780
4. Covariance has units: [K.kPa]

c.f) Covariance between temperature and humidity is 202

※ Covariance with itself is the variance:

$$\mathrm{cov}(x,x) = V(x) = E\left\{\left(x-\bar{x}\right)\left(x-\bar{x}\right)\right\}$$

Review: correlation (상관관계)

Q: Which one (pressure and temperature) has stronger relationship with temperature?

- Covariance depends on units: e.g. different covariance for grams vs kilograms
- Correlation removes the scaling effect:

$$corr(x,y) = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} = \frac{E\left\{\left(x-\bar{x}\right)\left(y-\bar{y}\right)\right\}}{\sigma_x \sigma_y}$$

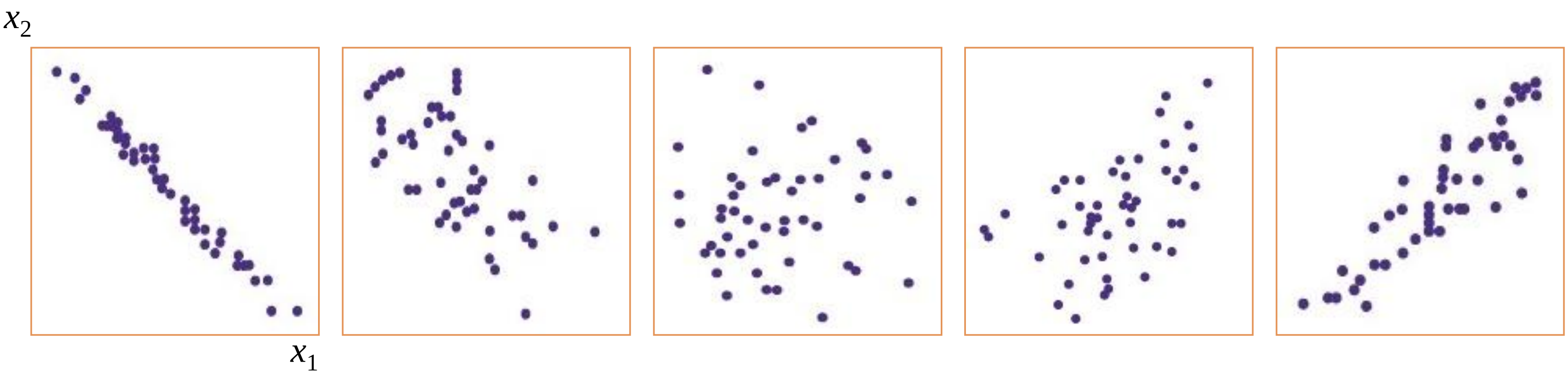
- Divides by the units of x and y: dimensionless result

$$-1 \le corr(x,y) = \rho_{xy} \le 1$$

- Gas cylinder example:
  - corr(temperature, pressure) = 0.997
  - corr(temperature, humidity) = 0.380

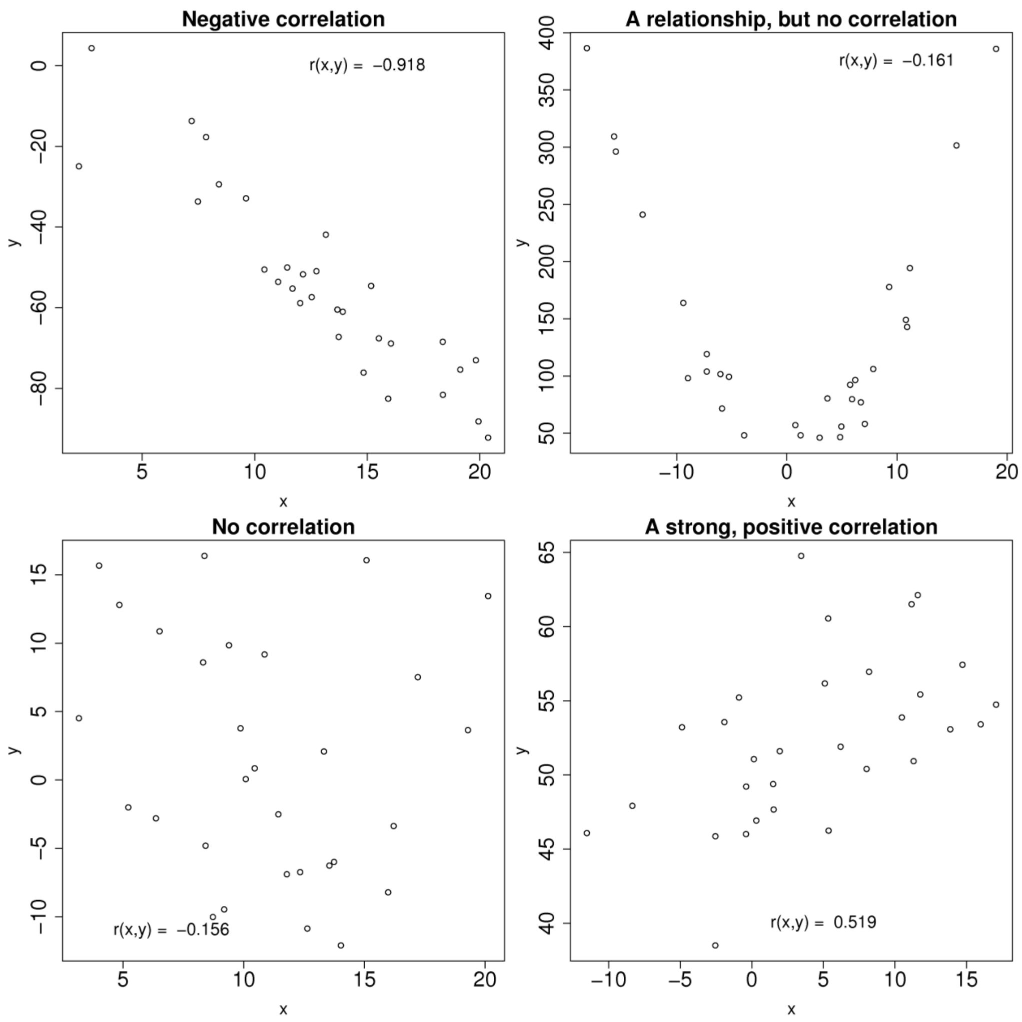
Review: correlation (cont.)

- Which one has highest/lowest/negative/positive correlation?
- Which one has (almost) no correlation?



- What does that mean if correlation of two variables is -1/+1?

Review: correlation (cont.)



Least squares? Least squares regression?

- *Regression* is the act of choosing the "best" values for the unknown parameters in a model on the basis of a set of measured data.
- Linear regression is the special case where the model is linear in the parameters. A straight line has the form:

$$y = a_0 + a_1 x + e$$

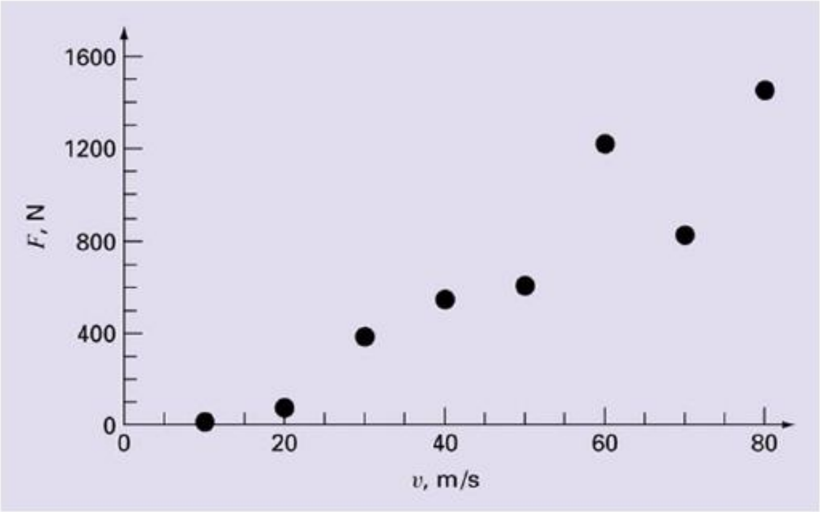
- There are many possible ways to define the "best" fit. However, the most commonly used measure for bestness is the sum of squared residuals.
  - Least sum of squares of errors → least squares in short.

Least squares (regression)

- It is the basis for :
  - DOE (Design of Experiments)
  - Latent variable methods
- We consider only 2 (sets of) variables : x and y (or x's and y)
  - Simple least squares
  - Multiple least squares
  - Generalized least squares

Simple least squares

- Wind tunnel example
  - How can we find the best line that describe the following data?



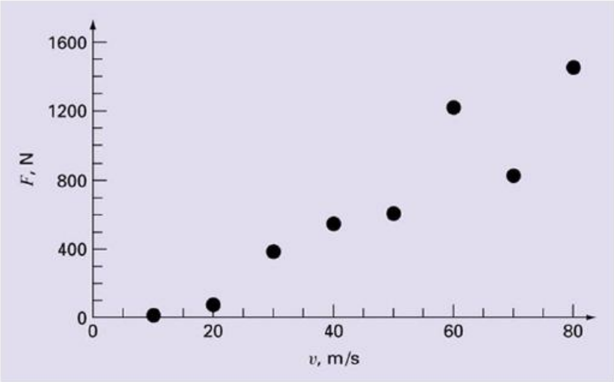
Data from wind tunnel experiments:
Drag force (F) at various wind velocities

v (m/s)	10	20	30	40	50	60	70	80
F (N)	25	70	380	550	610	1220	830	1450

Wind tunnel example (cont.)

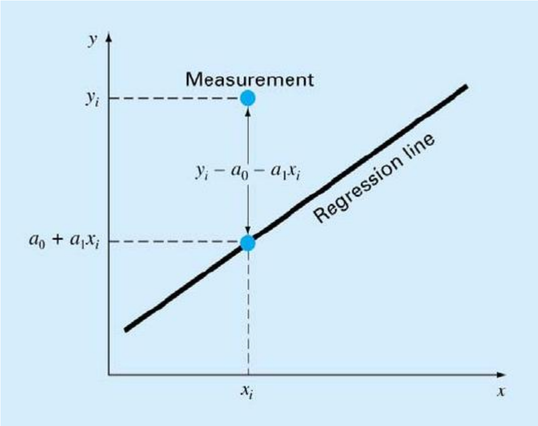
- From the plot, a linear line seems adequate.

  $y = a_o + a_1 x + e$



- At a data point (x_i, y_i), error between the line and the point is: (see the figure on the right)

  $e_i = y_i - a_o - a_1 x_i$



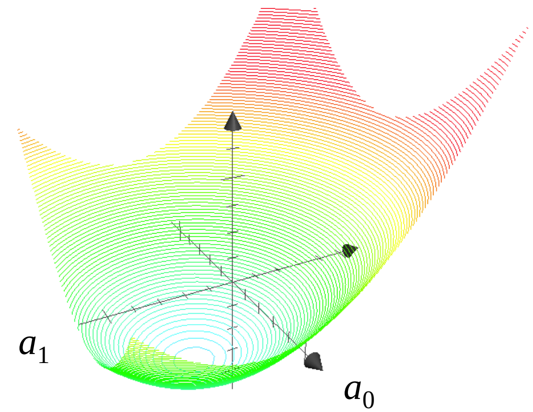
- Earlier, least squares means least sum of squares of errors. For all data points, sum of squares of errors is:

  $$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

- We need to find model parameters a_o and a_1 that minimize S_r.
  - "Least squares"

Wind tunnel example (cont.)

- How to find model parameters?
  - Take a look at Sr. $S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$
  - S_r is a parabolic function w.r.t a_0 and a_1 and sign of a_o^2 and a_1^2 are plus.
  - S_r becomes minimum where



$$\frac{\partial S_r}{\partial a_0} = 0 \ \& \ \frac{\partial S_r}{\partial a_1} = 0.$$

$$\frac{\partial S_r}{\partial a_0} = -2\sum (y_i - a_0 - a_1 x_i) \qquad \frac{\partial S_r}{\partial a_1} = -2\sum [(y_i - a_0 - a_1 x_i) x_i]$$

$$0 = \sum y_i - \sum a_0 - \sum a_1 x_i \qquad 0 = \sum x_i y_i - \sum a_0 x_i - \sum a_1 x_i^2$$

  - Rearranging and solving for a_0 and a_1

$$n a_0 + \left(\sum x_i\right) a_1 = \sum y_i \qquad \left(\sum x_i\right) a_0 + \left(\sum x_i^2\right) a_1 = \sum x_i y_i$$

$$a_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \qquad a_0 = \bar{y} - a_1 \bar{x}$$

Wind tunnel example (cont.)

- Calculations

v (m/s)	10	20	30	40	50	60	70	80
***F* (N)**	25	70	380	550	610	1220	830	1450

i	x_i	y_i	x_i^2	$x_i y_i$
1	10	25	100	250
2	20	70	400	1,400
3	30	380	900	11,400
4	40	550	1,600	22,000
5	50	610	2,500	30,500
6	60	1,220	3,600	73,200
7	70	830	4,900	58,100
8	80	1,450	6,400	116,000
Σ	360	5,135	20,400	312,850

Wind tunnel example (cont.)

- Calculations

$$\bar{x} = \frac{360}{8} = 45 \qquad \bar{y} = \frac{5{,}135}{8} = 641.875$$

$$a_1 = \frac{8(312{,}850) - 360(5{,}135)}{8(20{,}400) - (360)^2} = 19.47024$$

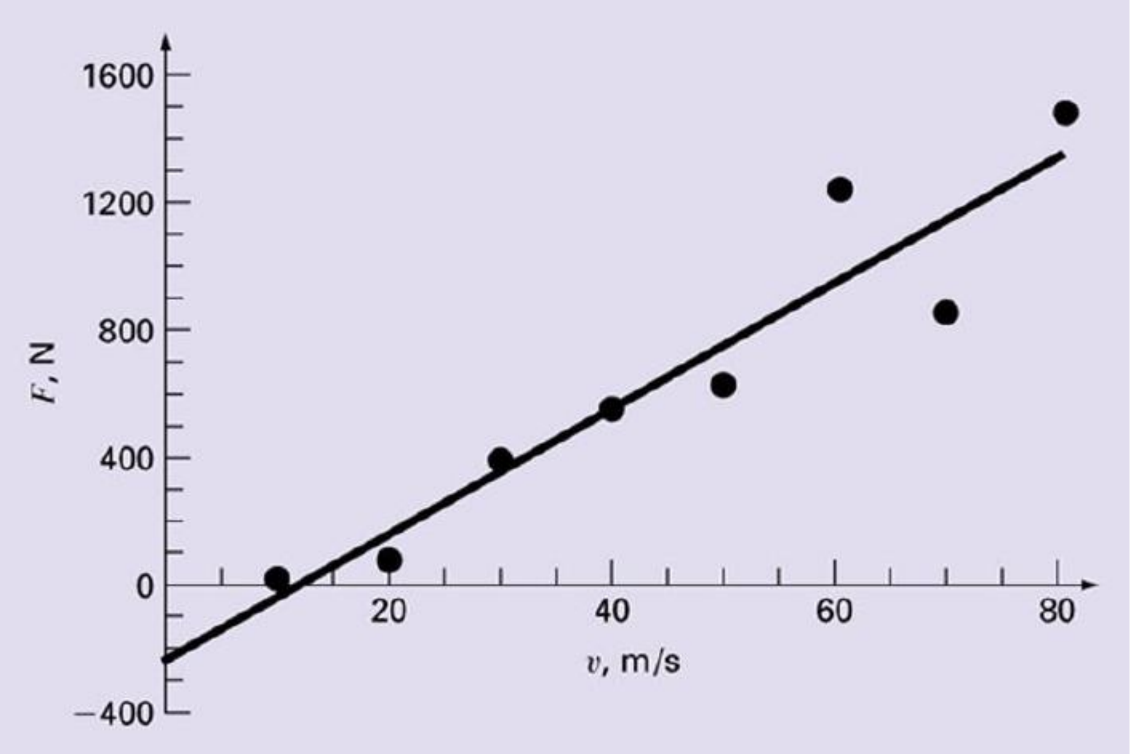
$$a_0 = 641.875 - 19.47024(45) = -234.2857$$

$$F = -234.2857 + 19.47024\,v$$

  - This is called simple least squares.

Wind tunnel example (cont.)

- Results



Is this OK with you?

General modeling procedure