Advanced Engineering Statistics
- Section 5 -

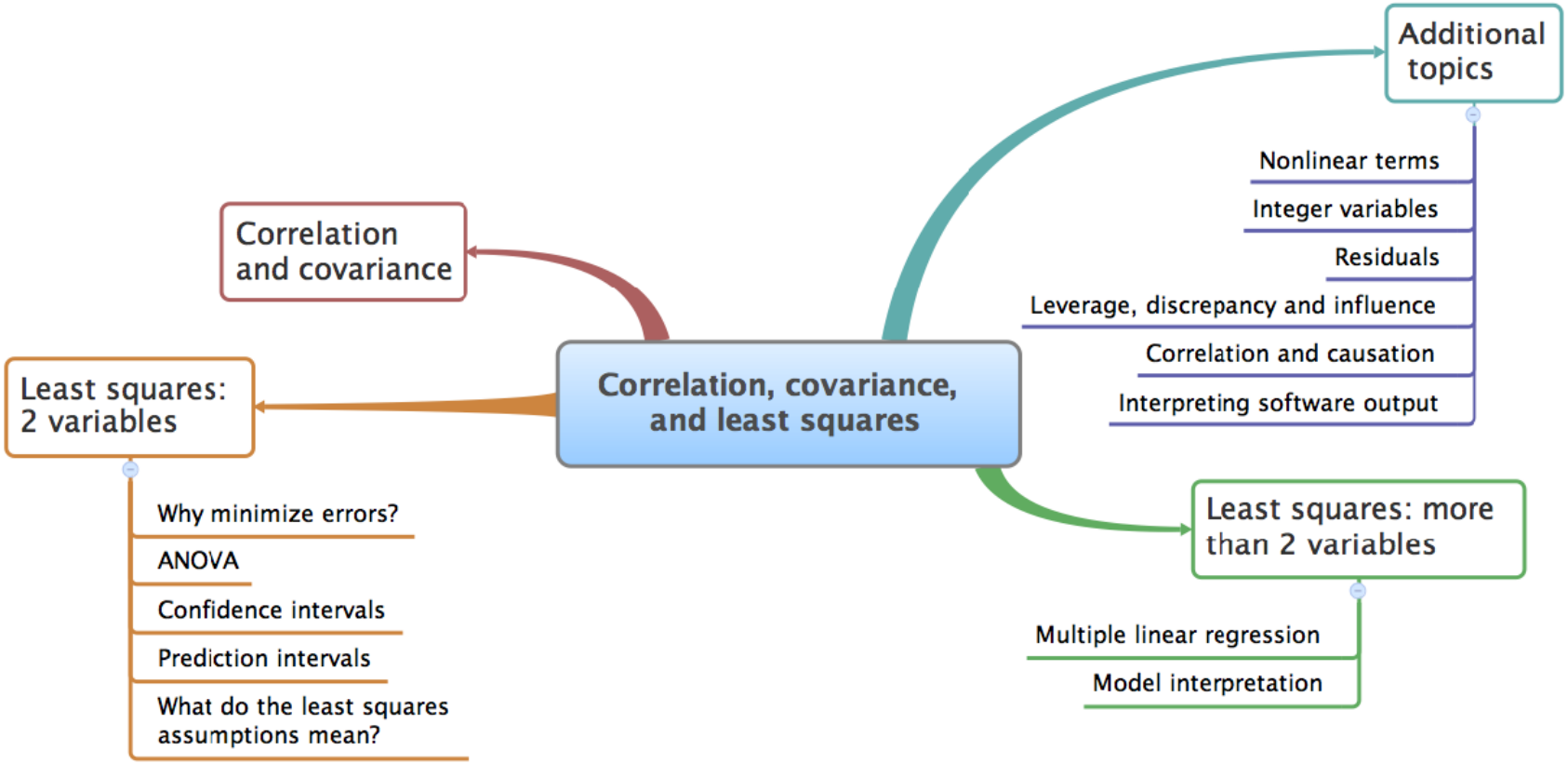
Jay Liu

Dept. Chemical Engineering

PKNU

Least squares regression

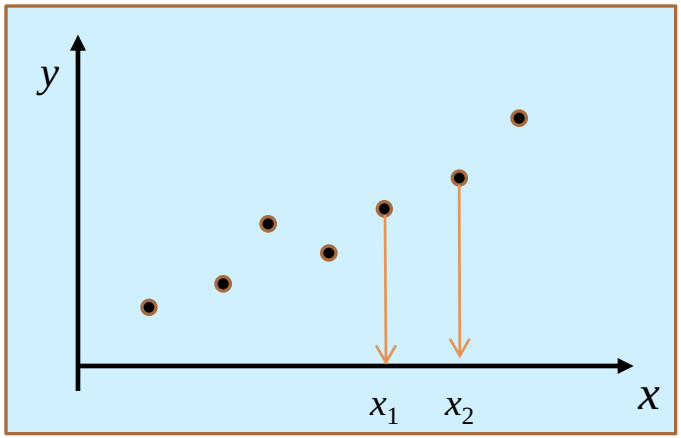
- What we will cover



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?

  - Ex. In using steam tables

- Otherwise, least squares is recommended.

Least squares - usage examples

- 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
  - y : response variables, output variables, dependent variables, …
  - x : input variables, regressor variables, independent variables, …

Review: covariance

- Consider measurements from a gas cylinder: temperature (K) and pressure (kPa).
- Ideal gas law applies under moderate condition: pV = nRT
  - Fixed volume, V = 20 × 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 = β_1T, 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: $\left(T - \bar{T}\right)\left(p - \bar{p}\right)$ *or* $T_{centered}^{T} p_{centered}$
   - 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 [K·%]

※ 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

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

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

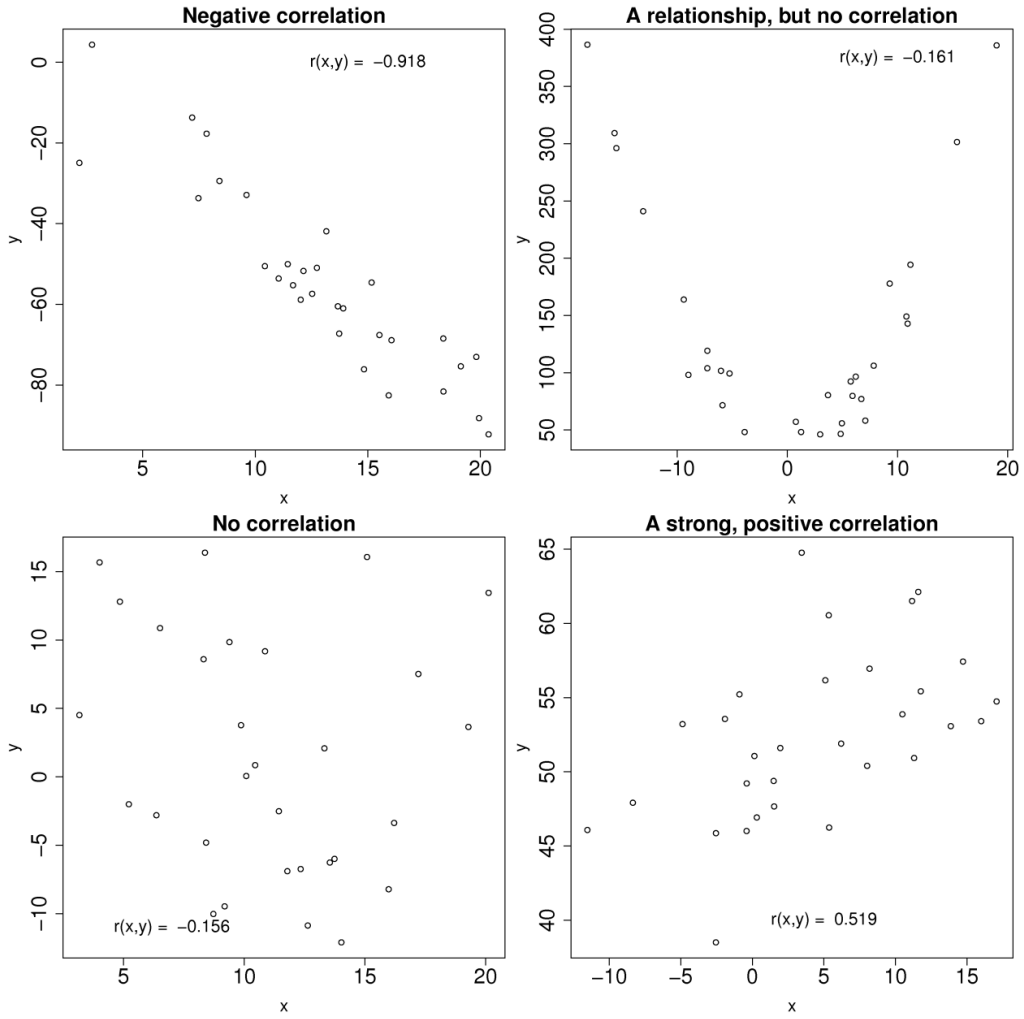
$$corr(x,y) = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y} = \frac{E\left\{(x-\bar{x})(y-\bar{y})\right\}}{\sigma_x \sigma_y}$$

- 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

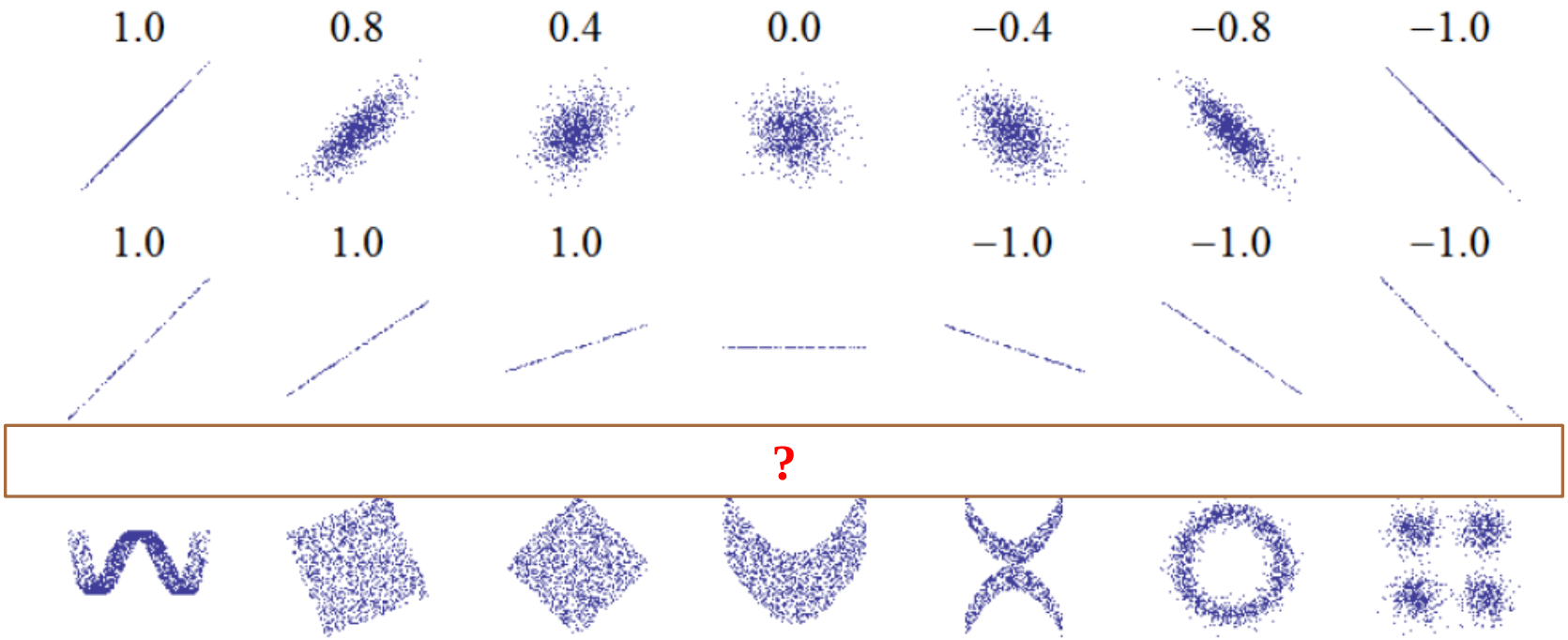
Review: correlation (cont.)



Want to find a relationship $y = f(x)$ other than the above?

Review: correlation (cont.)

- Remember!



From Wikipedia: Pearson product-moment correlation coefficient

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.
  - **Important**: error is from y, not from x.

[FYI] why minimize the sum of squares ?

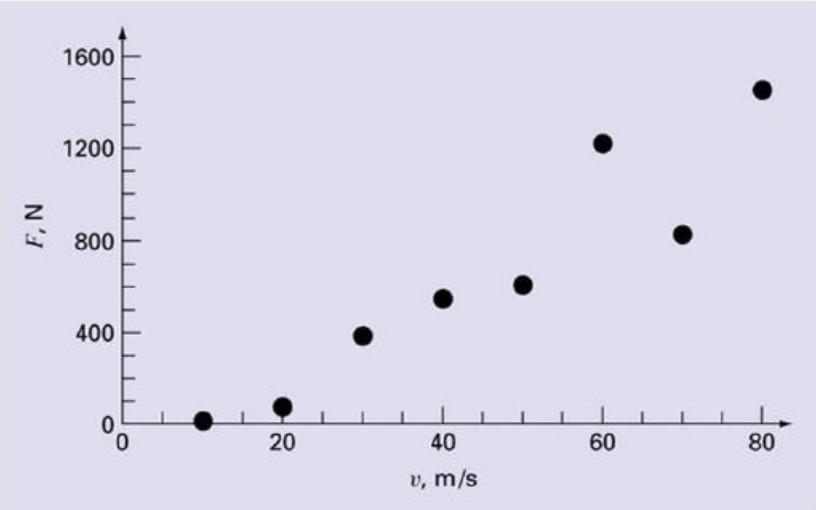
- The least squares model:
  - has the lowest possible variance for a_0 and a_1 when certain assumptions are met (more later)
  - computationally tractable by hand
  - easy to prove various mathematical properties
  - intuitive: penalize deviations quadratically
- Other forms: multiple solutions, unstable, high variance solutions, mathematical proofs are difficult

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?



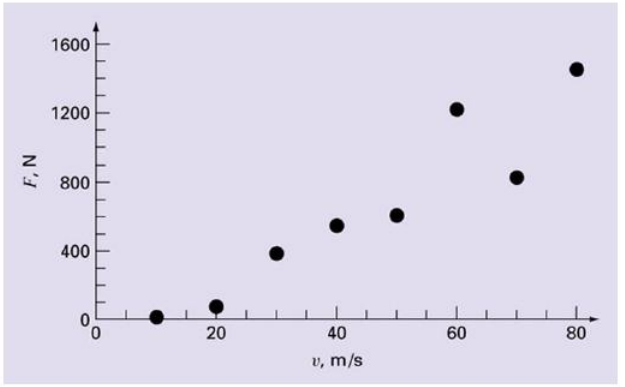
Data from wind tunnel experiments:
Drag force (F) at various wind velocities

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.)

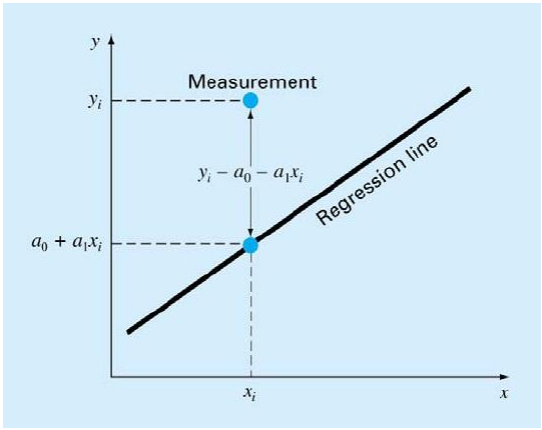
- From the plot, a linear line seems adequate.

$$y = a_0 + a_1 x$$



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

$$e_i = y_i - \hat{y}_i = y_i - a_0 - a_1 x_i$$



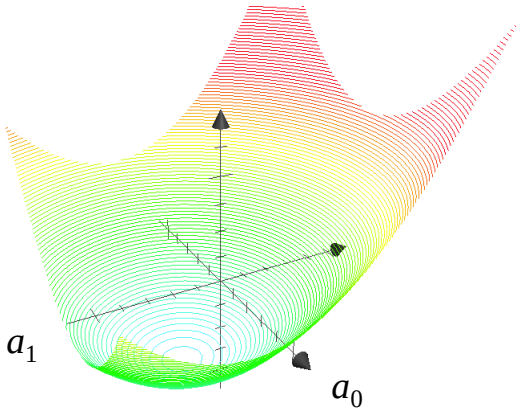
- Earlier, least squares means least sum of squares of errors. For all data points, sum of squares of errors is:

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

- We need to find model parameters a_0 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

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

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

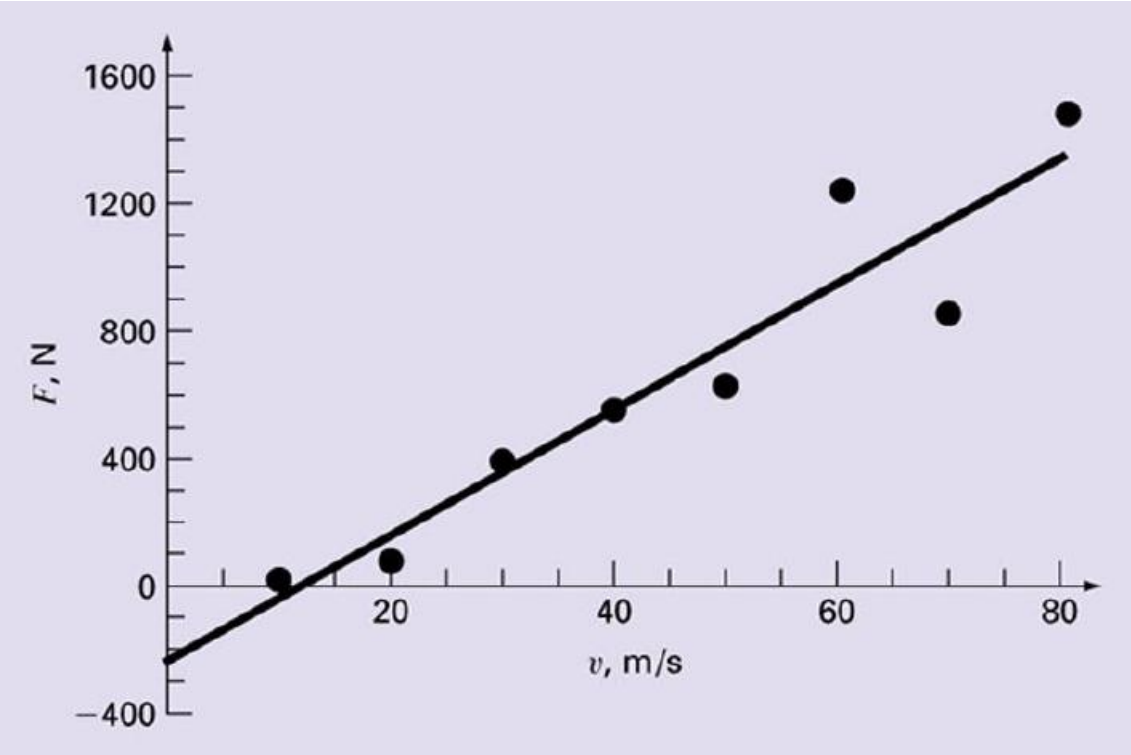
$$a_0 = 641.875 - 19.47024\,(45) = -234.2857$$

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

- This is called simple least squares.

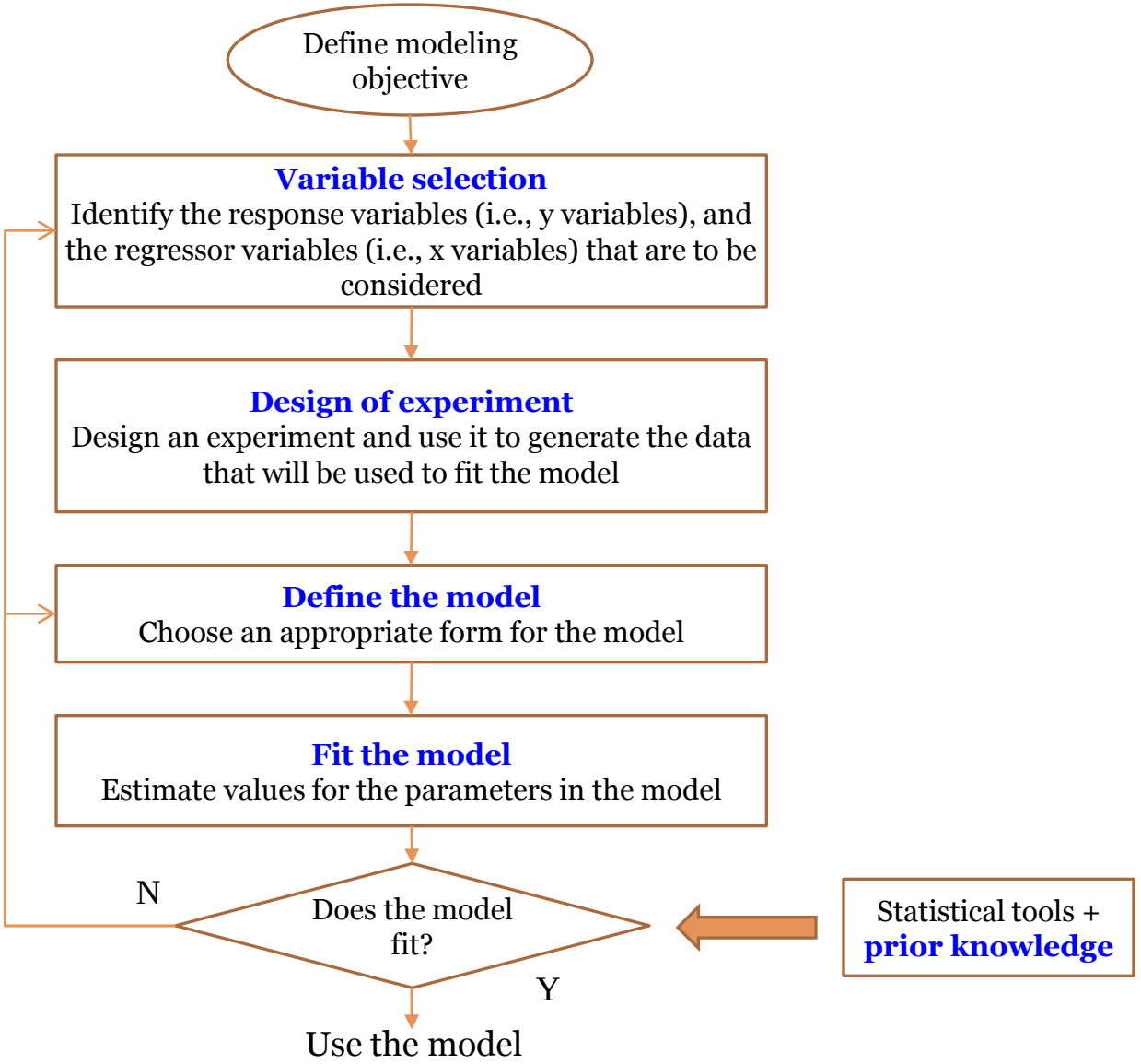
Wind tunnel example (cont.)

- Results



Is this OK with you?

General modeling procedure



Simple least squares

- Summary
  - Model form: $y = a_0 + a_1 x + e$
  - $S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$ becomes minimizes where $\frac{\partial S_r}{\partial a_0} = 0 \;\&\; \frac{\partial S_r}{\partial a_1} = 0.$
  - Rearranging and solving for a_0 and a_1

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

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

Simple least squares (cont.)

- Properties

1. $\sum_i e_i = 0$ *i.e.*, $\sum(y_i - \hat{y}_i) = 0$
2. The straight line equation passes through $(\bar{\mathrm{x}}, \bar{\mathrm{y}})$ without error
3. Prove to yourself that $\sum_i (x_i e_i) = \mathrm{x}^T \mathrm{e} = 0$
   - The residuals are uncorrelated with the input variables, x
   - There is no information in the residuals that is in the x's
4. Prove and interpret that $\sum_i (\hat{y}_i e_i) = \hat{\mathrm{y}}^T \mathrm{e} = 0$
   - The fitted values are uncorrelated with the residuals
5. Estimate of a_0 depends or a_1 : the estimates are correlated

Simple least squares (cont.)

- Questions
  1. Units of a_1 ?
     - The units of y divided by the units of x
  2. Gas cylinder example. Let $\hat{p}_i = a_0 + a_1 T_i$:
     - What is the interpretation of coefficient a_1
     - What is the interpretation of coefficient a_0
  3. How could the denominator term for a_1 equal zero? And what would that mean?
     - As long as there is variation in the x-data that we will obtain a solution

- what if our model we want to find is non-linear?

Ex. Activation energy in rate constant

$$k = k_0 e^{-E/RT}$$

➔

Linearization

- Want to model non-linear relationships between independent (x) and dependent (y) variables.

  1. Make a simple linear model through a suitable transformation.

$$y = f(x) + e \quad \rightarrow \quad y = a_0 + a_1 x + e$$

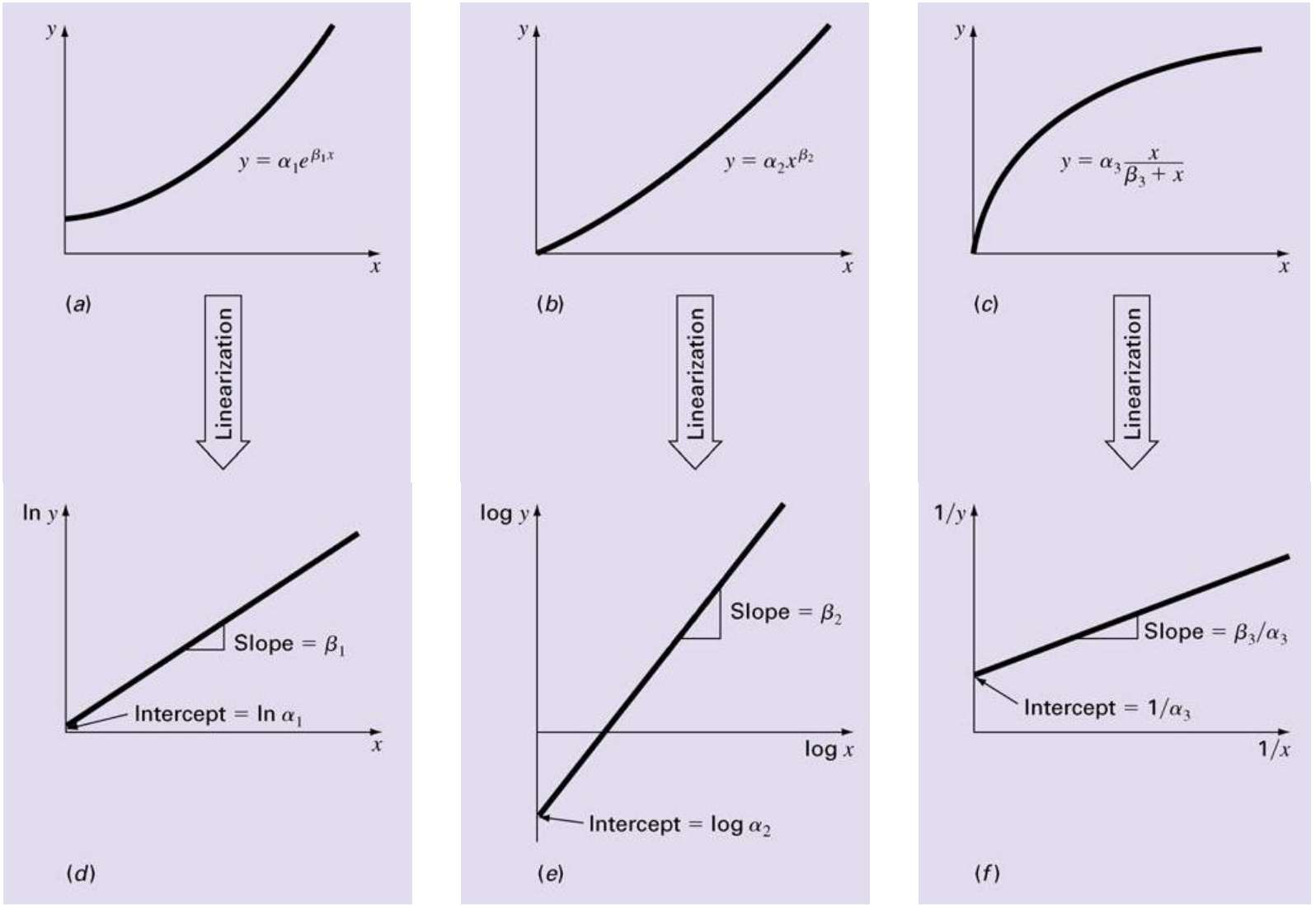
  2. Use previous results (simple least squares)

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

※Caution: nonlinear transformation also changes P.D.F of variables (and errors)

We will discuss about this in model assessment.

Linearization (Cont.)



Polynomial regression

- For quadratic form

$$y = a_0 + a_1 x + a_2 x^2 + e$$

  - Sum of squares

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 \right)^2$$

Again, S_r has a parabolic shape w.r.t a_0, a_1, and a_2. with plus signs of a_0^2, a_1^2, and a_2^2.

$$\frac{\partial S_r}{\partial a_0} = -2 \sum (y_i - a_0 - a_1 x_i - a_2 x_i^2) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum x_i (y_i - a_0 - a_1 x_i - a_2 x_i^2) = 0$$

$$\frac{\partial S_r}{\partial a_2} = -2 \sum x_i^2 (y_i - a_0 - a_1 x_i - a_2 x_i^2) = 0$$

Polynomial regression (Cont.)

- Rearranging the previous equations gives

$$\begin{aligned}(n)a_0+\left(\sum x_i\right)a_1+\left(\sum x_i^2\right)a_2&=\sum y_i\\ \left(\sum x_i\right)a_0+\left(\sum x_i^2\right)a_1+\left(\sum x_i^3\right)a_2&=\sum x_iy_i\\ \left(\sum x_i^2\right)a_0+\left(\sum x_i^3\right)a_1+\left(\sum x_i^4\right)a_2&=\sum x_i^2y_i\end{aligned}\quad\Rightarrow\quad\begin{pmatrix}n&\sum x_i&\sum x_i^2\\ \sum x_i&\sum x_i^2&\sum x_i^3\\ \sum x_i^2&\sum x_i^3&\sum x_i^4\end{pmatrix}\begin{pmatrix}a_0\\a_1\\a_2\end{pmatrix}=\begin{pmatrix}\sum y_i\\ \sum x_iy_i\\ \sum x_i^2y_i\end{pmatrix}$$

  the above equations can be solved easily. (three unknowns and three equations.)

- For general polynomials

$$y=a_0+a_1x+a_2x^2+\cdots+a_mx^m+e$$

- From the results of two cases ($y=a_0+a_1x$ & $y=a_0+a_1x+a_2x^2$)

$$\frac{\partial S_r}{\partial a_0}=\frac{\partial S_r}{\partial a_1}=\cdots=\frac{\partial S_r}{\partial a_m}=0$$

  we need to solve (m+1) linear algebraic equations for (m+1) parameters.

Multiple least squares

- Consider when there are more than two independent variables, x_1, x_2, …, x_m. ➔ regression plane.

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_m x_m + e$$

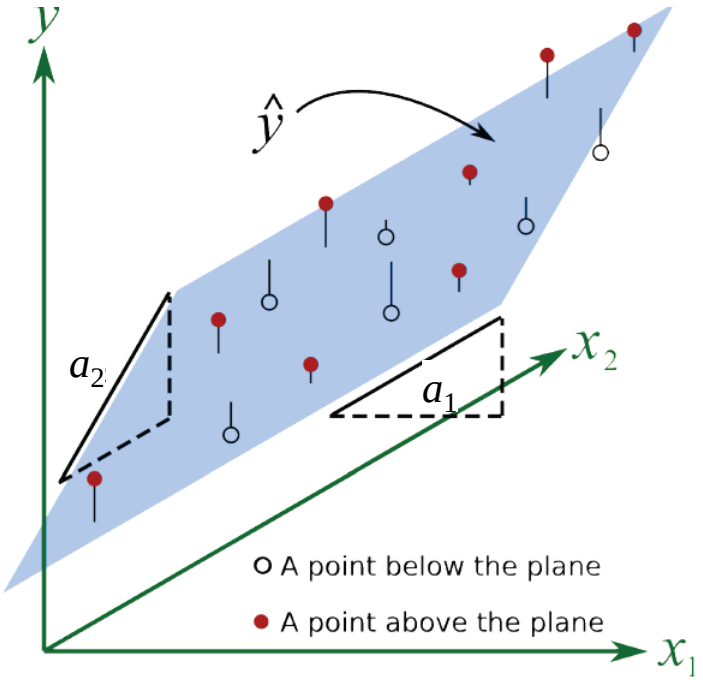
- For 2-D case, $y = a_0 + a_1 x_1 + a_2 x_2$.
  - Again, S_r has a parabolic shape w.r.t a_0, a_1.

$$S_r = \sum (y_i - a_0 - a_1 x_{1,i} - a_2 x_{2,i})^2$$

$$\frac{\partial S_r}{\partial a_0} = -2\sum (y_i - a_0 - a_1 x_{1,i} - a_2 x_{2,i}) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum x_{1,i} (y_i - a_0 - a_1 x_{1,i} - a_2 x_{2,i}) = 0$$

$$\frac{\partial S_r}{\partial a_2} = -2\sum x_{2,i} (y_i - a_0 - a_1 x_{1,i} - a_2 x_{2,i}) = 0$$



Multiple least squares (Cont.)

- Rearranging and solve for a_0, a_1 and a_2 gives

$$\begin{pmatrix} n & \sum x_{1,i} & \sum x_{2,i} \\ \sum x_{1,i} & \sum x_{1,i}^2 & \sum x_{1,i}x_{2,i} \\ \sum x_{2,i} & \sum x_{1,i}x_{2,i} & \sum x_{2,i}^2 \end{pmatrix} \begin{Bmatrix} a_1 \\ a_2 \\ a_3 \end{Bmatrix} = \begin{Bmatrix} \sum y_i \\ \sum x_{1,i}y_i \\ \sum x_{2,i}y_i \end{Bmatrix}$$

- For an m-dimensional plane,

$$y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m + e$$

- Same as in general polynomials,

$$\frac{\partial S_r}{\partial a_0} = \frac{\partial S_r}{\partial a_1} = \cdots = \frac{\partial S_r}{\partial a_m} = 0$$

we need to solve (m+1) linear algebraic equations for (m+1) parameters.

General least squares

- The following form includes all cases (simple least squares, polynomial regression, multiple regression)

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + \cdots + a_m z_m + e$$

where $z_0, z_1, \ldots, z_m$: $m+1$ different functions

Ex. Simple and multiple least squares

$$Z_0 = 1, Z_1 = x_1, Z_2 = x_2, \cdots, Z_m = x_m$$

polynomial regression

$$Z_0 = x^0 = 1, Z_1 = x^1, Z_2 = x^2, \cdots, Z_m = x^m$$

- Same as before,

$$\frac{\partial S_r}{\partial a_0} = \frac{\partial S_r}{\partial a_1} = \cdots = \frac{\partial S_r}{\partial a_m} = 0$$

we need to solve (m+1) linear algebraic equations for (m+1) parameters.

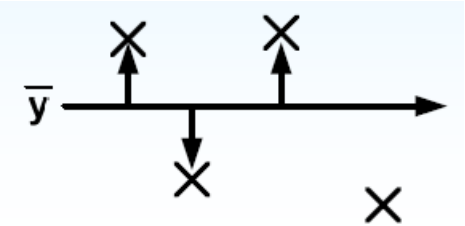
Quantification of errors

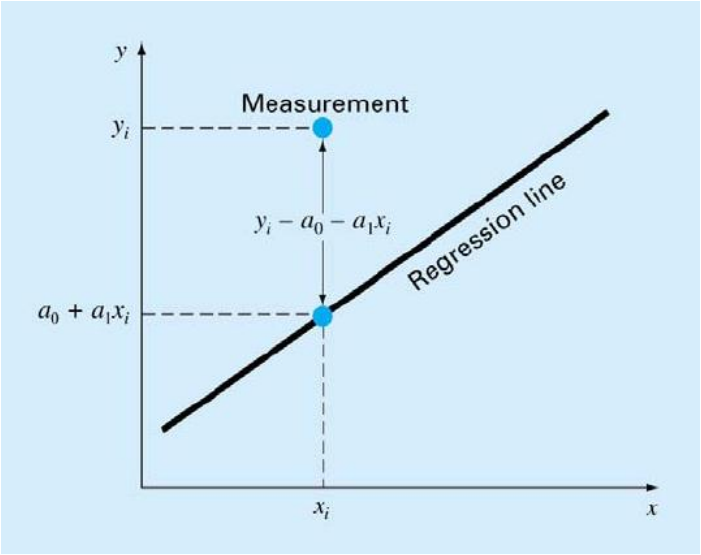
$$S_t = \sum (y_i - \bar{y})^2$$

Total sum of squares around the mean for the response variable, *y*

$$S_r = \sum e_i^2 = \sum \left(y_i - a_0 z_{0,i} - a_1 z_{1,i} - \cdots - a_m z_{m,i}\right)^2$$

Sum of squares of residuals around the regression line





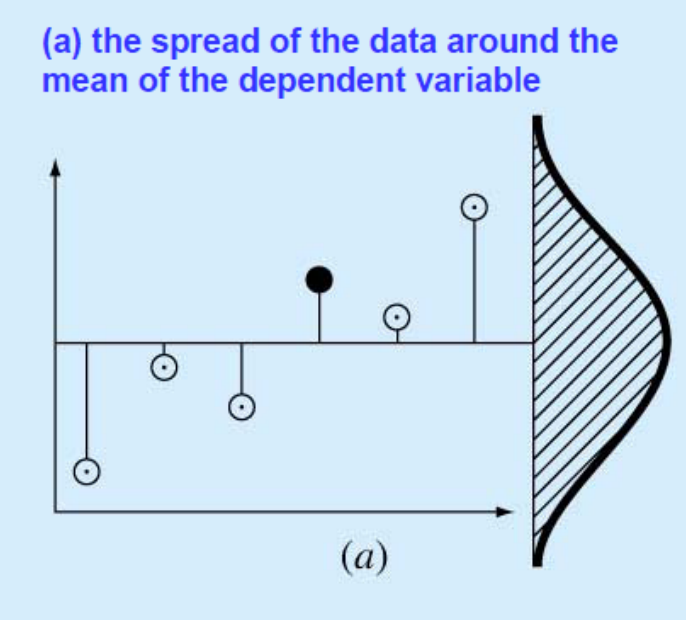
Quantification of errors (Cont.)

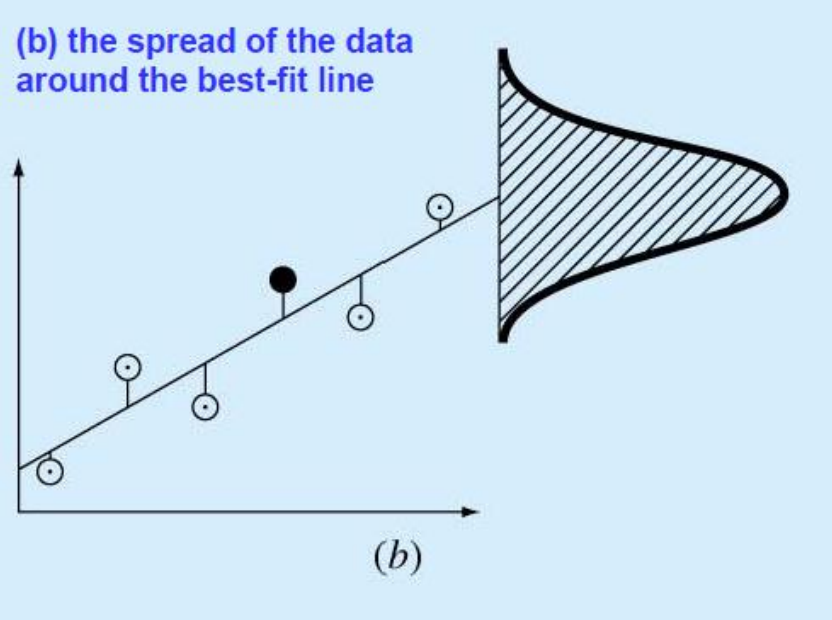
$$S_y = \sqrt{\frac{1}{n-1}\sum (y_i - \bar{y})^2} = \sqrt{\frac{S_t}{n-1}}$$

Standard deviation of y

$$S_{y/x} = \sqrt{\frac{S_r}{n-(m+1)}}$$

Standard error of predicted y (S_E)
→ quantify appropriateness of regression





Quantification of errors (Cont.)

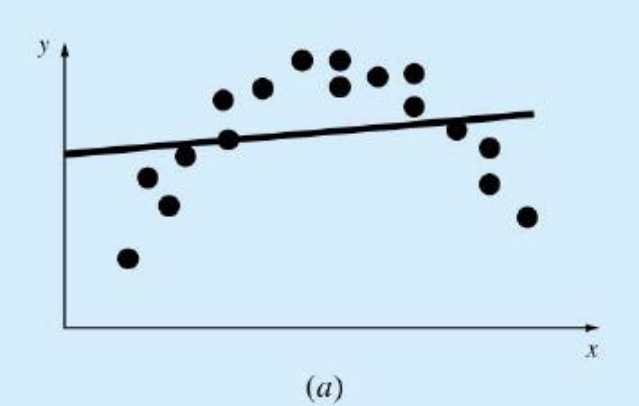
- Coefficients of determination, R^2

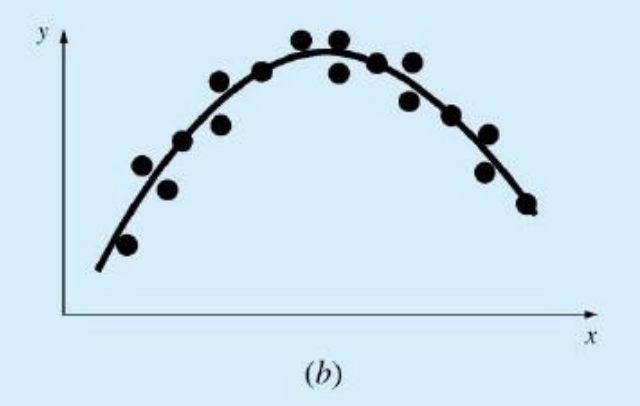
$$R^2 = \sqrt{\frac{S_t - S_r}{S_t}}$$

The amount of variability in the data explained by the regression model.

$R^2 = 1$ when $S_r = 0$: perfect fit (a regression curve passes through data points)

$R^2 = 0$ when $S_r = S_t$: as bad as doing nothing



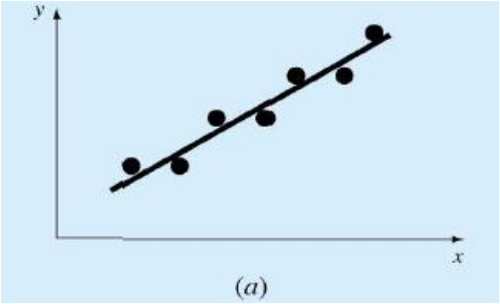


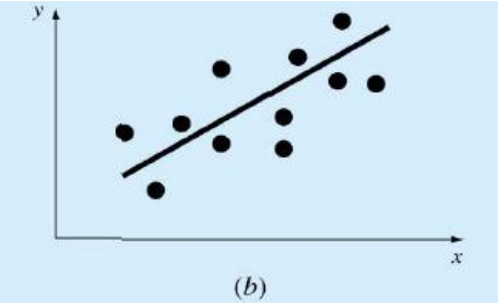
It is evident from the figures that a parabola is adequate.
R^2 of (b) is higher than that of (a)

Quantification of errors (Cont.)

- **Warning!** : $R^2 \approx 1$ **does not guarantee** that the model is adequate, nor the model will predict new data well.
  - It is possible to force R^2 to be one by adding as many terms as there are observations.
  - S_r can be big when variance of random error is large.

  (Usual assumption on error is that error is random is unpredictable)





Practice using Excel

(1) Wind tunnel example with higher polynomials

(2) Simple regression with increasing random noise

Confidence intervals - coefficients

- Coefficients in the regression model have confidence interval.

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + \cdots + a_m z_m + e$$

  - Why? They are also statistics like $\bar{x}$ & s. That is, they are numerical quantities calculated in a sample (not entire population). They are estimated values of parameters.

Statistic that we want to find its confidence interval

$$statistic \pm A \times \sigma_{statistic}$$

Value that depends on P.D.F of the statistic & confidence level α

Standard error of the statistic

statistic	A	$\sigma_{statistic}$
$\bar{x}$	$z_{\alpha/2}$	$\sigma_x / \sqrt{n}$
$\bar{x}$	$t_{\nu,\alpha/2}$	$s_x / \sqrt{n}$

※The standard error of a statistic is the standard deviation of the sampling distribution of that statistic

Confidence intervals – coefficients (cont.)

- Matrix representation of GLS

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + \cdots + a_m z_m + e$$

$$\mathbf{y} = \mathbf{Z}\mathbf{a} + \mathbf{e}$$

- matrix of the calculated values of the basis functions at the measured values of the independent variable
- observed valued of the dependent variable
- unknown coefficients
- residuals

$$\mathbf{Z} = \begin{bmatrix} Z_{01} & Z_{11} & \cdots & Z_{m1} \\ Z_{02} & Z_{12} & \cdots & Z_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{0n} & Z_{1n} & \cdots & Z_{mn} \end{bmatrix}$$

$$\mathbf{y}^T = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}$$

$$\mathbf{a}^T = \begin{bmatrix} a_0 & a_1 & \cdots & a_m \end{bmatrix}$$

$$\mathbf{e}^T = \begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix}$$

m+1: number of coefficients
n: number of data points

Confidence intervals – coefficients (Cont.)

- Example

Fitting quadratic polynomials to five data points

x	-1.0	-0.5	0.0	0.5	1.0
y	1.0	0.5	0.0	0.5	2.0

$$y = a_0 + a_1 x + a_2 x^2 + e$$

$$\mathbf{y} = \mathbf{Za} + \mathbf{e}$$

$$\begin{bmatrix} 1.0 \\ 0.5 \\ 0.0 \\ 0.5 \\ 2.0 \end{bmatrix} = \begin{bmatrix} 1 & -1.0 & 1.0 \\ 1 & -0.5 & 0.25 \\ 1 & 0.0 & 0.0 \\ 1 & 0.5 & 0.25 \\ 1 & 1.0 & 1.0 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

Three unknowns
Five equations

Can you solve this problem?

Confidence intervals – coefficients (Cont.)

- Solutions

$$\mathbf{y} = \mathbf{Z}\mathbf{a} + \mathbf{e}$$

Sum of squares of errors

$$S_r = \sum e_i^2 = \mathbf{e}^T\mathbf{e} = (\mathbf{y} - \mathbf{Z}\mathbf{a})^T(\mathbf{y} - \mathbf{Z}\mathbf{a})$$

$$\frac{\partial S_r}{\partial \mathbf{a}} = 0 \longrightarrow (\mathbf{Z}^T\mathbf{Z})\mathbf{a} = \mathbf{Z}^T\mathbf{y}$$

Called "normal equations"

1. LU decomposition or other methods to solve L.A.E

$$(\mathbf{Z}^T\mathbf{Z})\mathbf{a} = \mathbf{Z}^T\mathbf{y} \quad \Rightarrow "\mathbf{A}\mathbf{x} = \mathbf{b}"$$

2. Matrix inversion

$$(\mathbf{Z}^T\mathbf{Z})\mathbf{a} = \mathbf{Z}^T\mathbf{y} \quad \Rightarrow \mathbf{a} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}$$

computationally not efficient, but statistically useful

Confidence intervals – coefficients (Cont.)

- Matrix inversion approach

$$\mathbf{a} = \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{y}$$

Denote Z_{ii}^{-1} as the diagonal element of $\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}$

Confidence interval of estimated coefficients

$$a_{i-1} \pm t_{n-(m+1),\alpha/2}\sqrt{S_{y/x}^2 Z_{ii}^{-1}}$$

$t_{n-(m+1),\alpha/2}$ Student t statistics

$S_{y/x} = \sqrt{\dfrac{S_r}{n-(m+1)}}$ Standard error of estimate

Confidence intervals – coefficients (Cont.)

- For a linear model,

C.I. for a_1(slope)

$$a_1 \pm t_{n-(m+1),\alpha/2}\, S_{y/x} \sqrt{\frac{1}{\sum (x_i - \bar{x})^2}}$$

C.I. for a_o (intercept)

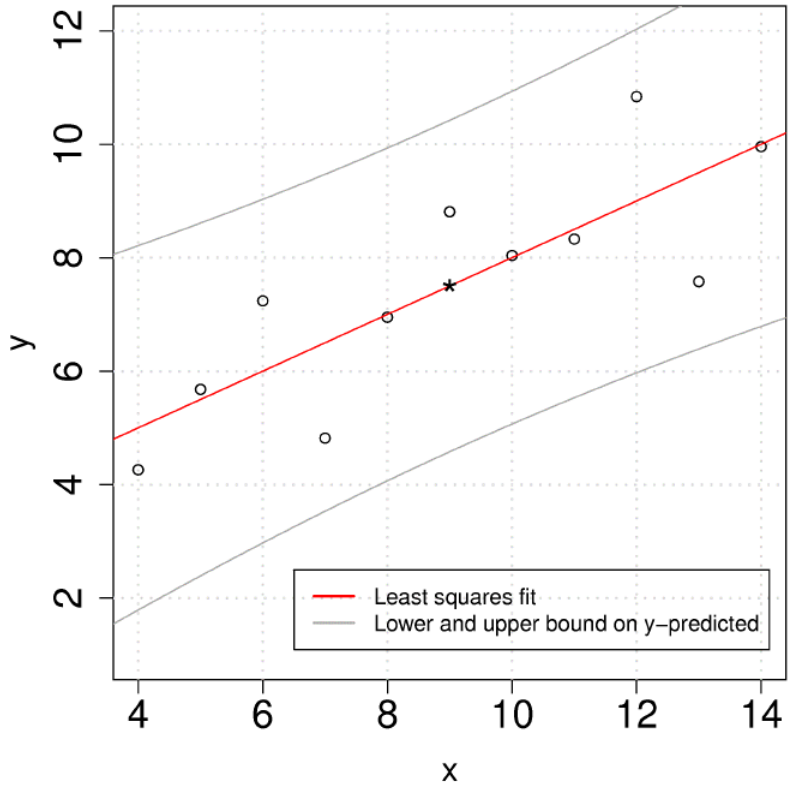
$$a_0 \pm t_{n-(m+1),\alpha/2}\, S_{y/x} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

What if confidence intervals contain zero?

Confidence intervals – prediction

- C.I for predicted y, $\hat{y}_i$

$$\hat{y}\big|_{x_0} \pm t_{n-(m+1),\alpha/2} S_{y/x} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$



Model assessment

- When we do not know the model form, we have to assess the model before use it after we fit a regression model.
  - However, in order to assess the model and make inferences about the parameters and predictions from the model, we will have to employ statistics and make some assumptions about the nature of the disturbance.
- Tools for model assessment
  - $S_{y/x}$, R^2 (quantitative) (→ Do not use)
  - Residual Plots (qualitative)
  - Normal probability chart (qualitative or quantitative)
  - Test for lack of fit (quantitative)
    - This is used when the dataset includes replicates. It is based on analysis of variance (ANOVA).

Model assessment - assumptions

- What is the most desirable errors in regression ?

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + \cdots + a_m z_m + e$$

- Assumptions on error
  - Error is additive $y = a_0 + a_1 x_1 + e$ ~~$y = (a_0 + a_1 x_1)e$~~
  - The variance of the error is constant and is not related to values of the response or values of the regressor variables.
  - There is no error associated with the values of the regressor variables.
  - Error is a random variable with Gaussian distribution $N(0,\sigma^2)$ (σ^2 usually unknown)

Model assessment – residual plots

- Recall the assumptions on error
  - Error is not related to the values of response or regressor variables.

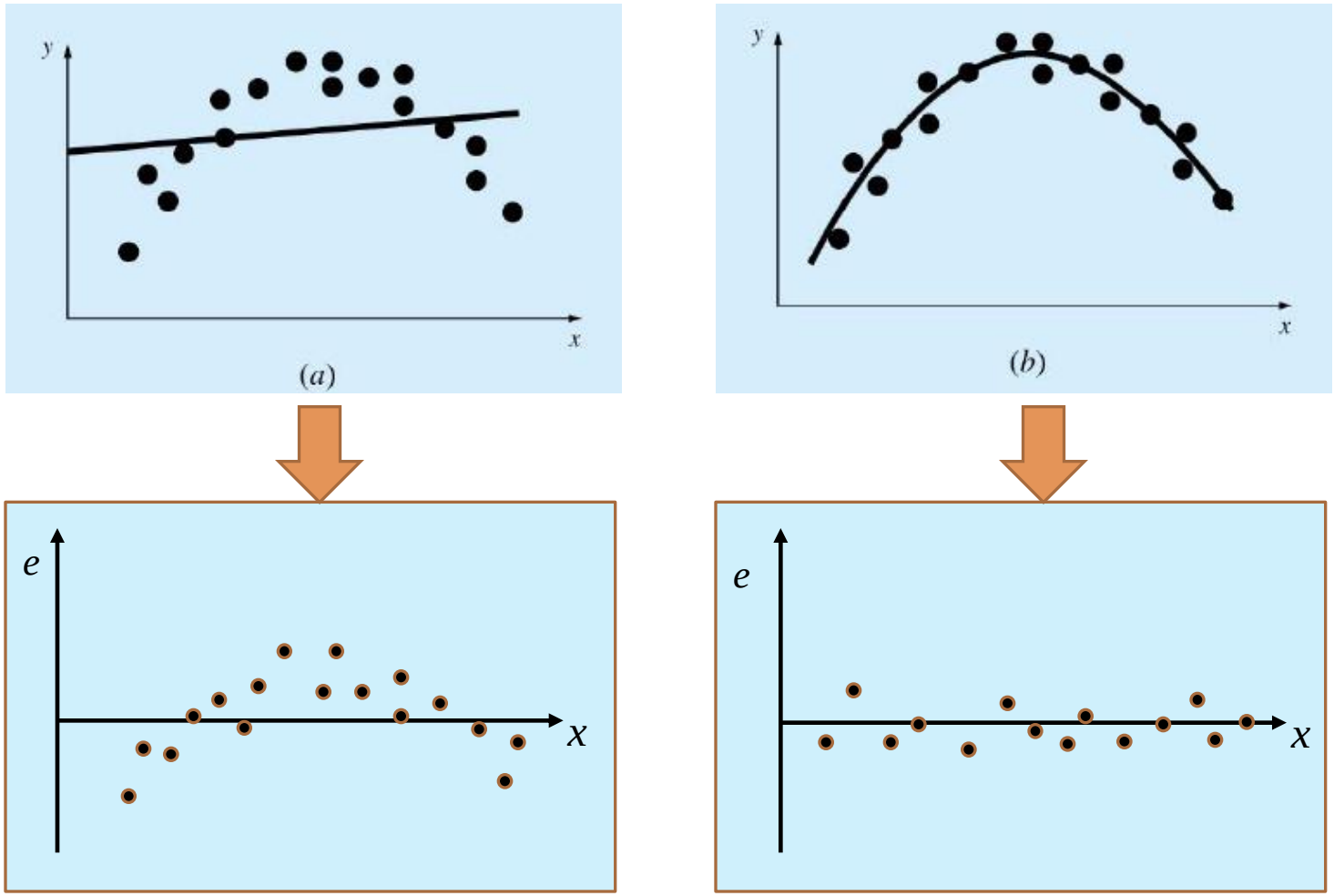
Then, assumptions will not be valid if the model is wrong.

- Following residual plots will reveal this.
  - Residuals vs. regressor variables
  - Residuals vs. fitted y values ($\hat{y}_i$)
  - Residuals vs. "lurking" variables (i.e. time or order)

➔ These plots will show "some patterns" when a model is inadequate.

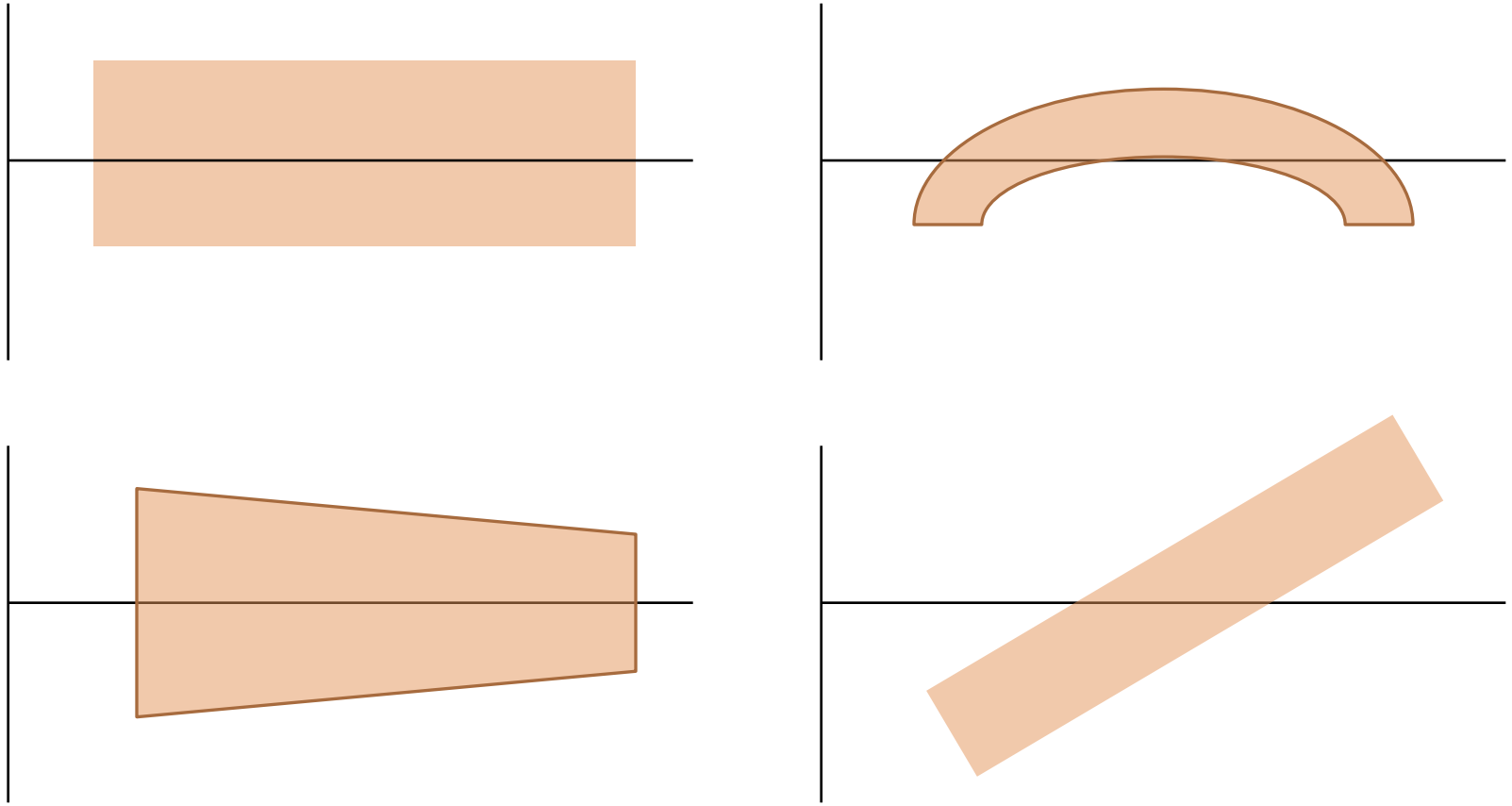
Model assessment – residual plots (con't)

- Examples



Model assessment – residual plots (con't)

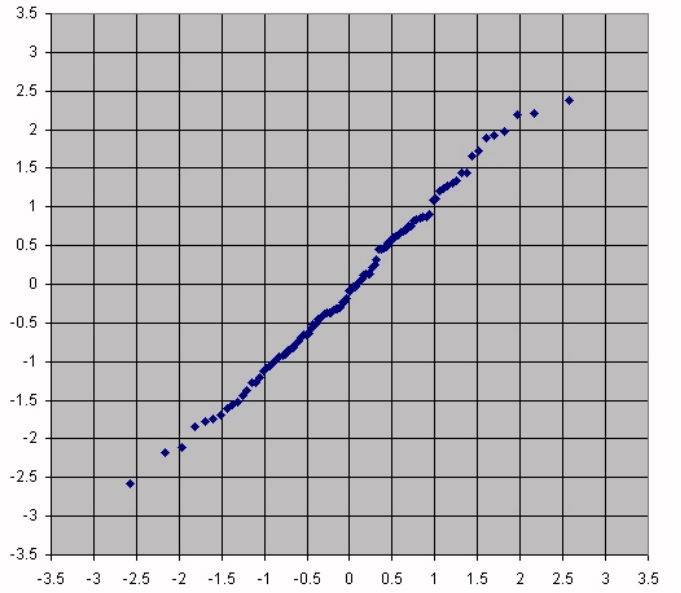
- Examples of residual plots



Model assessment – normal probability plot

- Recall the assumptions on error
  - Error is a random variable with Gaussian distribution $N(0, \sigma^2)$ (σ^2 usually unknown)

Then, errors will fall onto a straight line (y = x) in a normal probability plot. (especially useful when the number of data points is large)



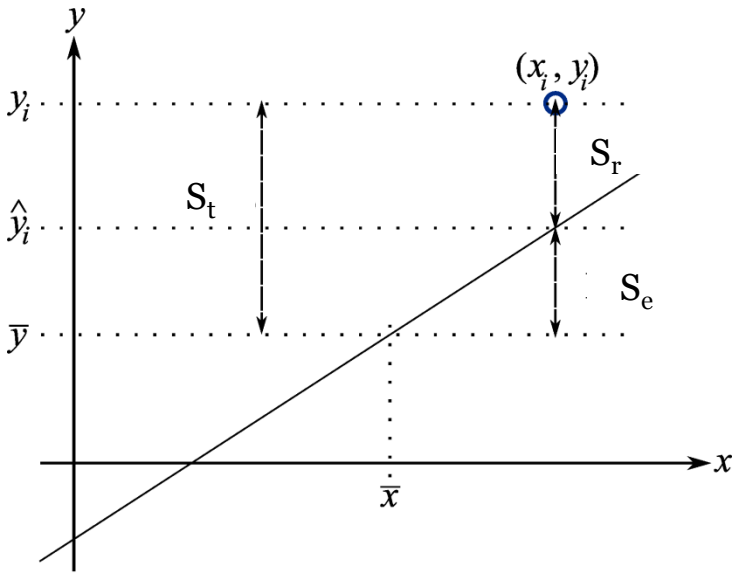
Normal probability plot

Alternatively, normality test can be used.

Model assessment – ANOVA (Test for lack of fit)

- The variance breakdown

$$S_t = \sum (y_i - \bar{y})^2$$



$$S_r = \sum (y_i - \hat{y}_i)^2$$

$$S_{Reg} = \sum (\hat{y}_i - \bar{y})^2$$

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

$$(y_i - \bar{y})^2 = (\hat{y}_i - \bar{y})^2 + 2(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) + (y_i - \hat{y}_i)^2$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

Really? Prove to yourself

$$\therefore S_t = S_{Reg} + S_r$$

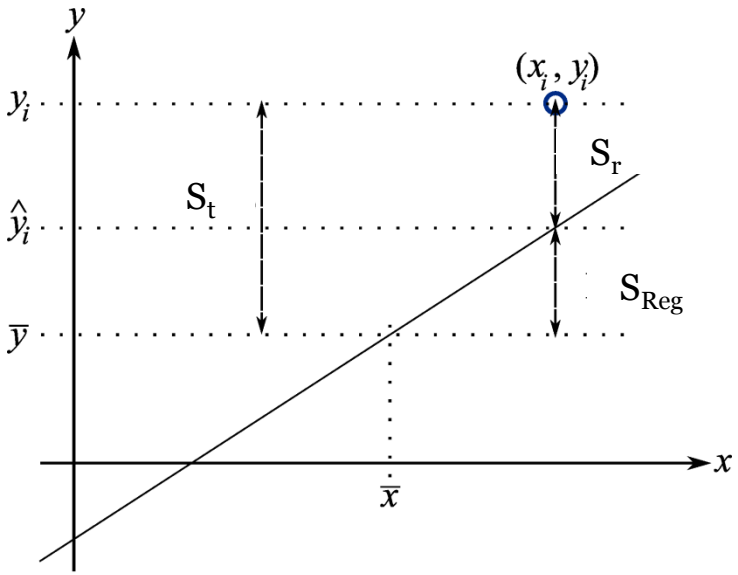
Model assessment – ANOVA (Test for lack of fit)

- The variance breakdown

$$S_t = S_{\mathrm{Re}g} + S_r$$

  - Ratio of S_{Reg}/S_r follows F distribution when corrected with degree of freedom.
  - If regression is **not** meaningful, the ratio (S_e/S_r) is small and $S_t \fallingdotseq S_r$.

$$S_t = \sum (y_i - \bar{y})^2$$



$$S_r = \sum (y_i - \hat{y}_i)^2$$

$$S_{\mathrm{Re}g} = \sum (\hat{y}_i - \bar{y})^2$$

Model assessment – ANOVA (Test for lack of fit)

- Analysis of variance (ANOVA) - just a tool to show the breakdown of variability in the y vector:
  1. doing nothing, no model: implies $\hat{y} = \overline{y}$
  2. the model: $\hat{y}_i = a_o + a_1 x_i$
  3. how much variance is left over in the errors e_i
  - All 3 add up to the total variance:
  - Variance = deviation from the mean
  - Total variance of y = model's variance + error variance

S_t $S_{\mathrm{Re}g}$ S_r

Model assessment – ANOVA (Test for lack of fit)

ANOVA Table

Source of Var.	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	S_{Reg}	p	$MS_{Reg}=S_{Reg}/p$	MS_{Reg}/ MS_E
Residual error	S_r	n-p	$MS_E=S_r/(n-p)$	
Total	S_t	n-1		

Compare F_0 to the critical value $F_{p,n-p;\alpha}$

What we are doing is a *test of hypothesis*.
We are testing the hypothesis:

$H_0 : \beta_0 = \cdots = \beta_p = 0$

H_1 : at least one parameter is not equal to zero.

[FYI]Meaning of a p-value in hypothesis test

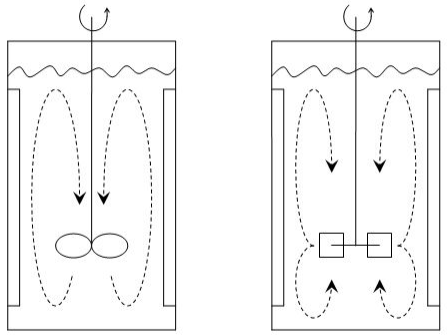
- A measure of how much evidence we have against the null hypothesis.
  - Null hypothesis (H_0) represents the hypothesis of no change or no effect.
  - Much research involves making a hypothesis and then collecting data to test that hypothesis. Then researchers will collect data and measure the consistency of this data with the null hypothesis.
  - A small p-value is evidence against the null hypothesis while a large p-value means little or no evidence against the null hypothesis.
  - Traditionally, researchers will reject a null hypothesis if the p-value is less than 0.05 ($\alpha = 0.05$).
  - → p-value can mean that the possibility that you can be wrong when rejecting the null hypothesis.

Integer variables in the model

- Integer variables 0 and 1 can represent qualitative variables.
  - Example: raw material from Spain, India, or Vietnam
    - $y = a_0 + a_1x_1 + \ldots + a_kx_k + r_1d_1 + r_2d_2 + r_3d_3$
    - $d_1 = 1$ and $d_2 = 0$ and $d_3 = 0$ for Spain
    - $d_1 = 0$ and $d_2 = 1$ and $d_3 = 0$ for India
    - $d_1 = 0$ and $d_2 = 0$ and $d_3 = 1$ for Vietnam
- Often called indicator variables for this reason

Integer variables in the model

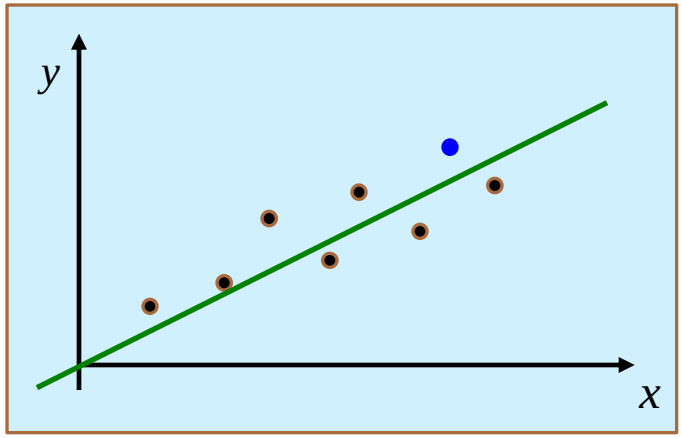
- Example
  - Want to predict yield when two different impeller used. Yield = *f*(temperature, impeller type)
  - Build two different models (one for axial, one for radial)
  - Build one model using indicator variable. $y = a_0 + a_1 T + rd$
    - $y = a_0 + a_1 T + rd_i$
    - $d_i = 0$ for axial, $d_i = 1$ for radial

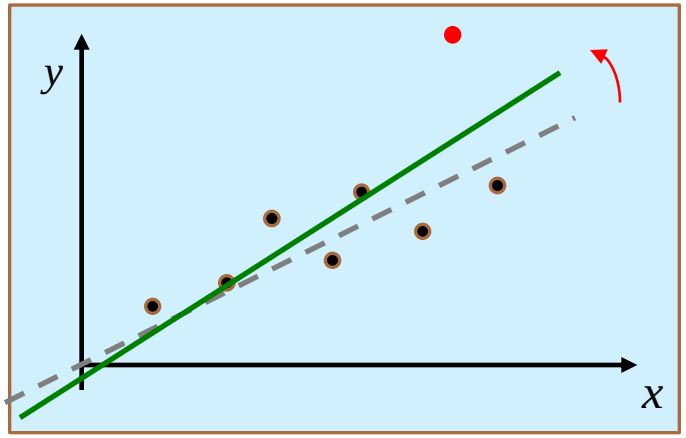


axial radial

Leverage effect

- Unusual observations influence the model parameters and our interpretation





Outliers have an over-proportional effect on resulting regression curves.

- To avoid the leverage effect,
  - Remove outliers before regression (but do not delete without investigation)
  - Use different S_r (no longer least squares)

Causal relation and correlation

- Causal relation
  - Cause and effect relation
    - Has physical/chemical/engineering meanings
  - x and y are **not** interchangeable
    - Direction exists.
- Correlation
  - (Linear) relationship between two variables
  - No physical/chemical/engineering meanings.
    - Average height of 20's men vs. year
  - x and y are interchangeable

Advanced topics

- Testing of least-squares models

Gold standard: use an independent testing data set

- $\text{RMSEP} = \sqrt{\frac{1}{n}\sum_i^n \left(y_{\text{new},i} - \hat{y}_{\text{new},i}\right)^2}$
- RMSEE is the same as RMSEP, but for building the model
- The RMSEE $\approx S_E$ = standard error
- Or, use some other measure of "closeness"

Advanced topics - Testing of least-squares models

Example:

- Need to build a predictive model for product viscosity using 3 x-variables
- Observations: one per day, from 2006 and 2007 (730 observations)

Which situation is better?

- **A**
  - Observations: 1, 3, 5, 7, ... 729 to build
  - Observations: 2, 4, 6, 8, ... 730 to test
- **B**
  - Observations: 1 to 365 (2006 data) to build
  - Observations: 366 to 730 (2007 data) to test

Advanced topics

- Correlated *x*'s
  - MLR solution

$$\mathbf{y} = \mathbf{Za} + \mathbf{e}$$

$$S_r = \sum e_i^2 = \mathbf{e}^T\mathbf{e} = (\mathbf{y} - \mathbf{Za})^T(\mathbf{y} - \mathbf{Za})$$

$$\frac{\partial S_r}{\partial \mathbf{a}} = 0 \longrightarrow (\mathbf{Z}^T\mathbf{Z})\mathbf{a} = \mathbf{Z}^T\mathbf{y}$$

$$\Rightarrow \mathbf{a} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}$$

  - When two or more x's are correlated, $(\mathbf{Z}^T\mathbf{Z})^{-1}$ becomes nearly singular, i.e., ill-conditioned.

Advanced topics – correlated *x*'s

- High/no correlation between x_1 and x_2

$$\mathbf{Z}^T\mathbf{Z} = \begin{bmatrix} 1.0000 & 0.9999 \\ 0.9999 & 1.0000 \end{bmatrix}$$

$$\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1} = \begin{bmatrix} 5000.25 & -4999.75 \\ -4999.75 & 5000.25 \end{bmatrix}$$

$$\mathbf{Z}^T\mathbf{Z} = \begin{bmatrix} 1.0000 & 0.0 \\ 0.0 & 1.0000 \end{bmatrix}$$

$$\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1} = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}$$

- What if very small (measurement) noises added to x's

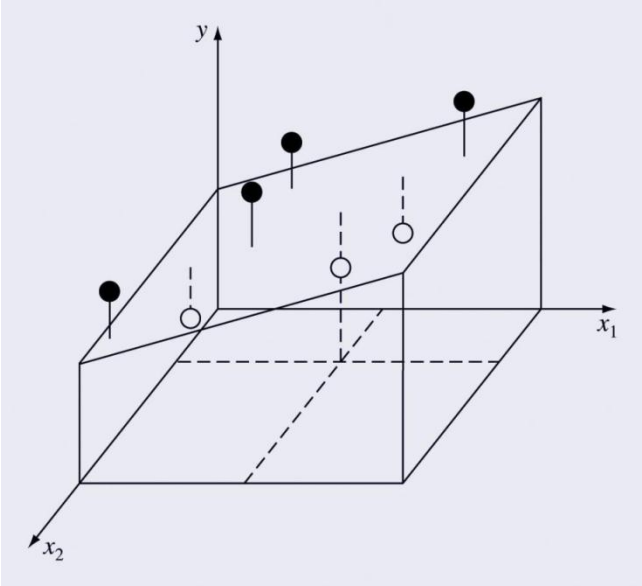
$$\mathbf{Z}^T\mathbf{Z} = \begin{bmatrix} 1.0001 & 0.9999 \\ 0.9999 & 1.0000 \end{bmatrix}$$

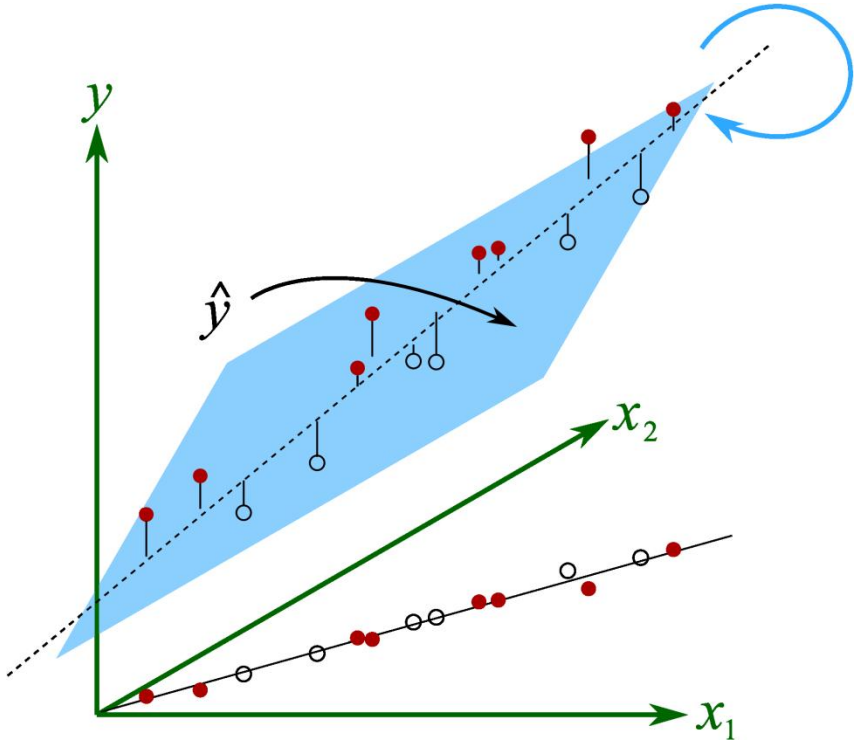
$$\mathbf{Z}^T\mathbf{Z} = \begin{bmatrix} 1.0000 & 0.0 \\ 0.0 & 1.0001 \end{bmatrix}$$

What will happen to your MLR model?

Advanced topics – correlated *x*'s

- If high correlation among x's:
  - unstable solutions for **a**
  - predictions uncertain also
- Geometrically speaking





High correlation between x_1 and x_2

Advanced topics – correlated *x*'s

- Remedies?
  - Use selected x variables → stepwise regression
  - Use ridge regression
  - Use multivariate methods (will not be covered in this lecture)

Advanced topics

- What we want to know:
  - How do we select the form of the model? Which variables should be included? Should we include transformations of the regressor variables?
  - ....
- What we want:
  - We would like to build the “best” regression model
  - We would like to include as many regressor variables as is necessary to adequately describe the behaviour of y. At the same time, we want to keep the model as simple as possible.

➔ Stepwise regression start off by choosing an equation having the single best x variables and the attempts to build up with subsequent additions of x’s one at a time as long as these additions are worthwhile.

Advanced topics – stepwise regression

- Procedure
  1. Add a x variable to the model (the variable that is most highly correlated with y).
  2. Check to see whether or not this has significantly improved the model. One way is to see whether or not the confidence interval for the parameter includes zero. (of course you can use hypothesis test) If the new term or terms are not significant, remove them from the model.
  3. Find one of the remaining x variables that is highly correlated with the residuals and repeat the procedure.

Advanced topics

- Ridge regression (Hoerl, 1962; Hoerl & Kennard, 1970a,b)
  - A modified regression method specifically for ill-conditioned datasets that allows all variables to be kept in the model.
  - This is possible by adding *additional* information to the problem to remove the ill-conditioning.
  - The objective function to minimize: $(\mathbf{y}-\mathbf{Xb})^T(\mathbf{y}-\mathbf{Xb})$
  - Least squares estimates have no bias but large variance, while ridge regression estimates have small bias and small variance.

Advanced topics – ridge regression

- Procedure

1. Mean center and scale all x's to unit variance

$$\mathbf{f}_i = \frac{\mathbf{x}_i - \bar{\mathbf{x}}}{s_{x_i}}$$

2. Rewrite the model as:

$$\mathbf{y} - \bar{\mathbf{y}} = a_1 s_{x_1}\left(\frac{\mathbf{x}_1 - \bar{\mathbf{x}}}{s_{x_1}}\right) + \cdots + a_p s_{x_p}\left(\frac{\mathbf{x}_p - \bar{\mathbf{x}}}{s_{x_p}}\right) + \varepsilon$$

$$\tilde{\mathbf{y}} = b_1\mathbf{f}_1 + \cdots + b_p\mathbf{f}_p + \varepsilon$$

$$\mathbf{Y} = \mathbf{Fb} + \boldsymbol{\varepsilon}$$

Advanced topics – ridge regression

3. The objective function to use for the optimization is to minimize

$$J = \left(\mathbf{Y} - \mathbf{Fb}\right)^T \left(\mathbf{Y} - \mathbf{Fb}\right) + k\mathbf{b}'\mathbf{b}$$

Therefore, $\mathbf{b}^* = \left(\mathbf{F}^T\mathbf{FZ} - k\mathbf{I}\right)^{-1}\mathbf{F}^T\mathbf{Y}$

4. Solve the optimization problem in Step 3 for several values of k between 0 and 1 and choose that value of k at which the estimates of b see to stabilize. Otherwise, choose k by validation on new data.

Advanced topics

- Non-linear regression
  - General form of a non-linear regression model

$$y = f(\mathbf{x}, \mathbf{a}) + \varepsilon$$

  - In a linear model, $f(\mathbf{x}, \mathbf{a}) = \mathbf{x}^T \mathbf{a}$. In a non-linear model, *f*() woulds have any form. E.g.,

$$y = \frac{e^{-\beta_1 x}}{1 + \beta_2 x} + \varepsilon$$

  - Remember that nonlinear transformation also changes P.D.F of variables (and errors)? What does this mean?

Advanced topics – non-linear regression

- The approach is exactly the same as for linear models
- We use the same objective function:

$$
\begin{aligned}
S_r &= \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i, \mathbf{a}))^2
\end{aligned}
$$

All we need is to minimize S over **a**. but how?

Advanced topics – non-linear regression

- The big difference between linear and nonlinear regression is that in general, the optimization problem for a nonlinear model does not have an exact analytical solution.
- Therefore, we have to use a numerical optimization algorithm such as:
  - Gauss-Newton
  - Steepest Descent
  - Conjugated Gradients
  - Any other optimization algorithm

Advanced topics – non-linear regression

- When using an optimization algorithm to solve nonlinear regression problems, one needs to be able to specify:
  1. an expectation function (i.e. the form of the model)
  2. Data
  3. starting guesses for **a**
  4. stopping criteria
  5. possibly other "tuning" parameters associated with the optimization algorithm

Advanced topics – non-linear regression

- Problems with Numerical Optimization
  - Failure to converge
  - Finding only a local minimum and not the global minimum
  - Requires good starting guesses for the parameters
  - Can be sensitive to the choice of convergence criteria and other "tuning parameters" of the algorithm
  - Sometimes requires specification of the derivatives of the model with respect the the parameters.