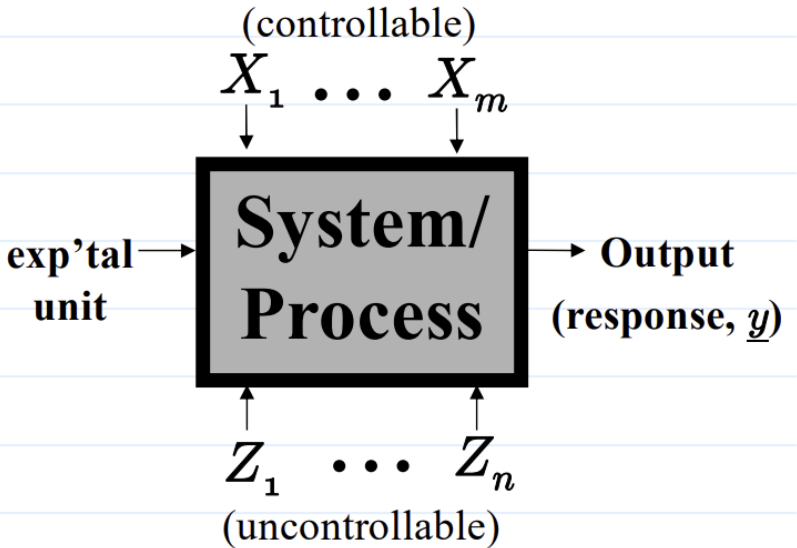
(U4284) Python程式設計
Linear Model



Speaker: 吳淳硯

python™

Some Terminologies

- A response/output/dependent variable Y is modeled or explained by p effects/functions of m predictor/input/independent variables $X_1, \cdots, X_m$

$$Y = \underbrace{\sum_{k=0}^{p} \beta_k g_k(X_1, \cdots, X_m)}_{\text{deterministic part}} + \underbrace{\varepsilon}_{\text{random part}}$$

$$E(Y|X_1, \cdots, X_m) = \sum_{k=0}^{p} \beta_k g_k\,, \qquad \text{Var}(Y|X_1, \cdots, X_m) = \sigma^2$$

- When $p = 1$, simple linear regression; $p > 1$ multiple regression
- $X_1, \cdots, X_m$: continuous, discrete (quantitative), categorical (qualitative)
  - All Quantitative ⇒ Multiple regression
  - Quantitative + Qualitative ⇒ ANCOVA
  - All Qualitative ⇒ ANOVA
- More than one Y, multivariate regression.

Rationale

- A general model for the relationship between Y and $X_1, \cdots, X_p$ is

$$Y = f(X_1, \cdots, X_p) + \varepsilon$$

  where f is unknown and arbitrary.

- Local approximation of function f may be achievable by linear model

$$f(X_1, \cdots, X_p) \approx f(a_1, \cdots, a_p) + \sum_{k=1}^{p} \left[\frac{\partial f}{\partial X_k}(a_1, \cdots, a_p) \right] (X_k - a_k)$$

$$= \beta_0 + \sum_{k=1}^{p} \beta_k X_k$$

- Predictors can be transformed and combined in any way, linear models are actually very flexible.
  Ex. $x_1 x_2, e^{x_1 + x_2}, \ldots$

- Matrix Form,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{p,1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & \cdots & x_{p,n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

MME (Method of Moment Estimation)

- Law of Large number

$$\frac{g(X_1) + \cdots + g(X_n)}{n} \to E\big(g(X)\big)$$

- X is exogenous for β if $E(X\varepsilon) = 0$. While a regressor X is exogenous if it is determined separately from Y.

- Since $E(\varepsilon) = 0$ and $E(X\varepsilon) = 0$

$$E(Y - \beta_0 - \beta_1 X) = 0 \Rightarrow \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)}{n} = 0$$

$$E(X[Y - \beta_0 - \beta_1 X]) = 0 \Rightarrow \frac{\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i)}{n} = 0$$

Obtain the normal equation

$$\begin{cases} \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i \end{cases}$$

OLS (Ordinary Least Squared)

- Consider a objective function is given by

$$\mathcal{S}(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^{\top}\boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\top}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

  find the coefficient $\boldsymbol{\beta}$ which fit the equations ″best″, in the sense of solving the quadratic minimization problem

$$\arg\min_{\boldsymbol{\beta}} \mathcal{S}(\boldsymbol{\beta}) \Leftrightarrow \frac{\partial \mathcal{S}}{\partial \boldsymbol{\beta}} = 0$$

- By calculus, $\widehat{\boldsymbol{\beta}}$ is the solution of

$$\boldsymbol{X}^{\top}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^{\top}\boldsymbol{y}$$

  assume $\boldsymbol{X}^{\top}\boldsymbol{X}$ is nonsingular then

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{y} = \boldsymbol{\beta} + (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{\varepsilon}$$
$$\Rightarrow \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

- Does $\widehat{\boldsymbol{\beta}}$ possess any optimal properties?

$$E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \qquad \mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^{\top}\boldsymbol{X})^{-1}$$

  and Gauss-Markov Theorem, for any $\breve{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{y}$ and $\|\boldsymbol{a}\| = 1$

$$\boldsymbol{a}^{\top}\left(\mathrm{Var}(\breve{\boldsymbol{\beta}}) - \mathrm{Var}(\widehat{\boldsymbol{\beta}})\right)\boldsymbol{a} \geq \mathbf{0}$$

What is likelihood?

- The parallel universe analogy for likelihood:

$$\mathcal{L}(\boldsymbol{\theta}|X_1, \cdots, X_n) = \prod_{k=1}^{n} f(X_k|\boldsymbol{\theta})$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)$

- The likelihood $\mathcal{L}$ tells us how probable it is that this particular universe generated the observed data.

- The higher the likelihood, the more plausible the universe is for explaining your observations.

MLE (Maximum Likelihood Estimation)

- Goal of MLE

  find the best parameter that make the observed data most probable.

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x})$$

- Consistency

$$\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta} \text{ in probability}$$

- Invariance

  If $\widehat{\boldsymbol{\theta}}$ is MLE of $\boldsymbol{\theta}$, then $\tau\left(\widehat{\boldsymbol{\theta}}\right)$ is the MLE of $\tau(\boldsymbol{\theta})$, where $\tau(\cdot)$ is a well-defined function.

- Asymptotic Efficiency

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \sim N(0, \mathcal{I}^{-1})$$

  where $\mathcal{I}$ is fisher information matrix

$$\mathcal{I} = E\left(-\frac{\partial^2 \ln f(X_k|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}}}\right)$$

$$= E\left[\left(\frac{\partial \ln f(X_k|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \ln f(X_k|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^{\mathsf{T}}\right]$$

What is a Gradient?

- The directional derivative measures the rate at which a function changes in a particular direction at a given point, the directional derivative of function f along a unit vector $\vec{u} = (u_1, \cdots, u_n)$.

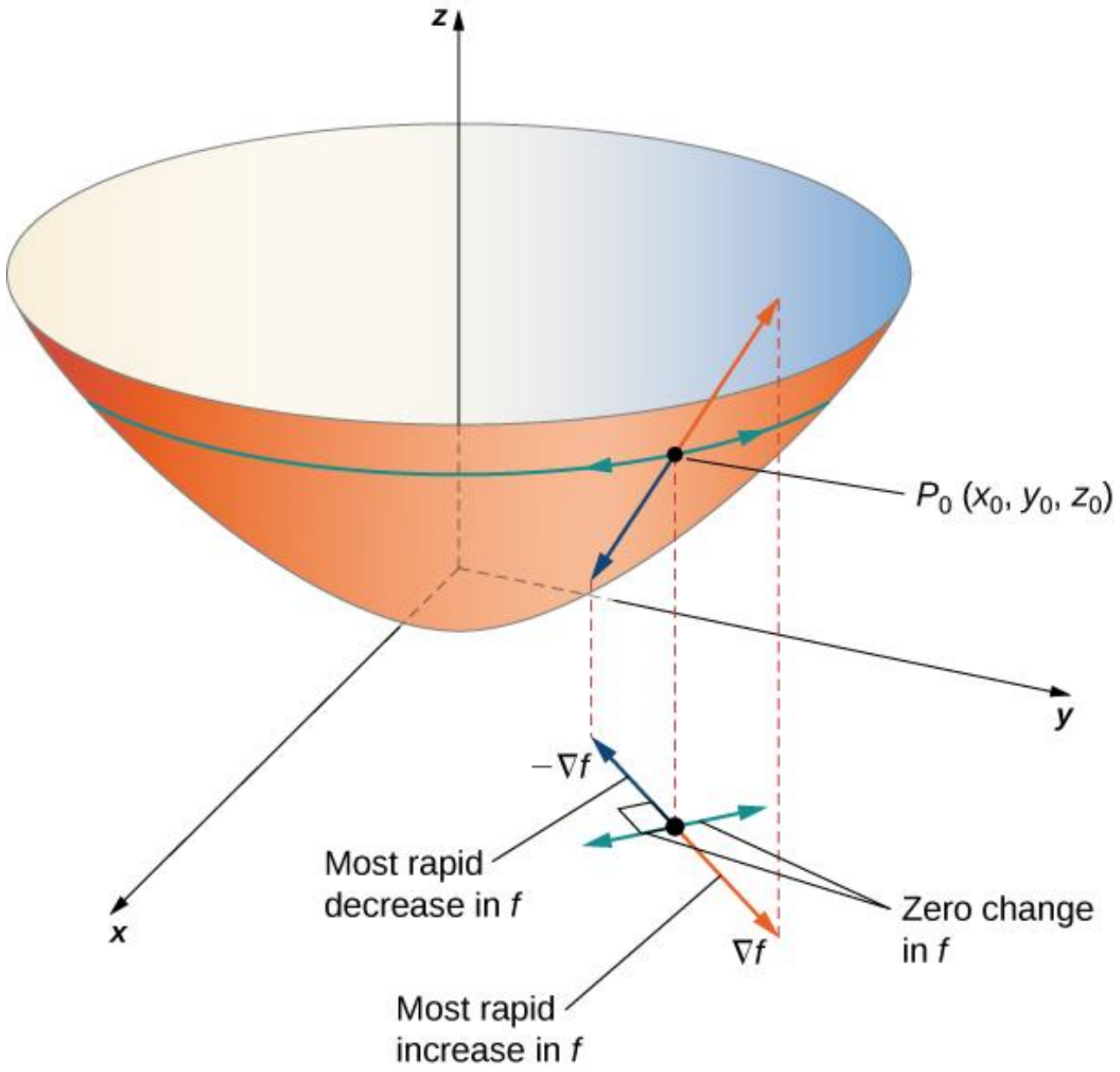
$$\mathcal{D}_{\vec{u}} f(\boldsymbol{x}) = \lim_{h \to 0} \frac{f(\boldsymbol{x} + h\vec{u}) - f(\boldsymbol{x})}{h} = \nabla f(\boldsymbol{x}) \cdot \vec{u}$$

  where $\nabla f(\boldsymbol{x})$ is gradient of f

$$\nabla f(\boldsymbol{x}) = \left[\frac{\partial f}{\partial x_1}(\boldsymbol{x}) \quad \cdots \quad \frac{\partial f}{\partial x_n}(\boldsymbol{x}) \right]^{\mathsf{T}}$$

- Properties of the Gradient
  - $\left| \begin{matrix} \nabla f(\boldsymbol{x}) \\ -\nabla f(\boldsymbol{x}) \end{matrix} \right|$ points in the direction of fasted $\left| \begin{matrix} \text{increase} \\ \text{decrease} \end{matrix} \right|$ at $\boldsymbol{x}$
  - $\left| \begin{matrix} |\nabla f(\boldsymbol{x})| \\ -|\nabla f(\boldsymbol{x})| \end{matrix} \right|$ is equal to $\left| \begin{matrix} \text{maximum} \\ \text{minimum} \end{matrix} \right| \mathcal{D}_{\vec{u}} f(\boldsymbol{x})$

Visual of Gradient



Gradient Descent

- Given α is the learning rate, consider the objective function

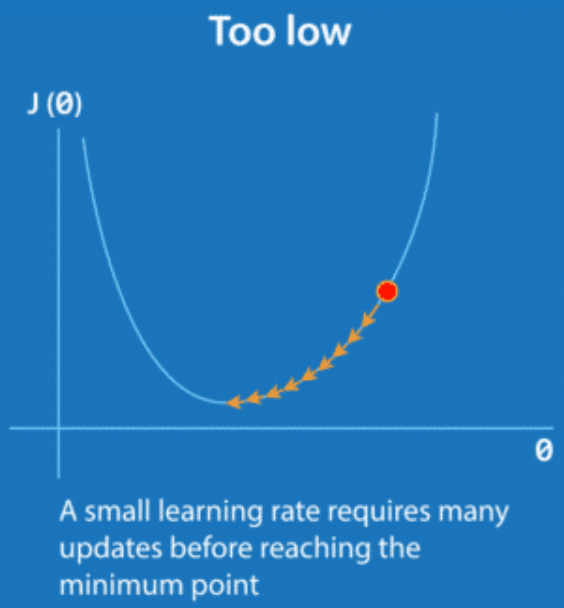
$$\mathcal{S}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

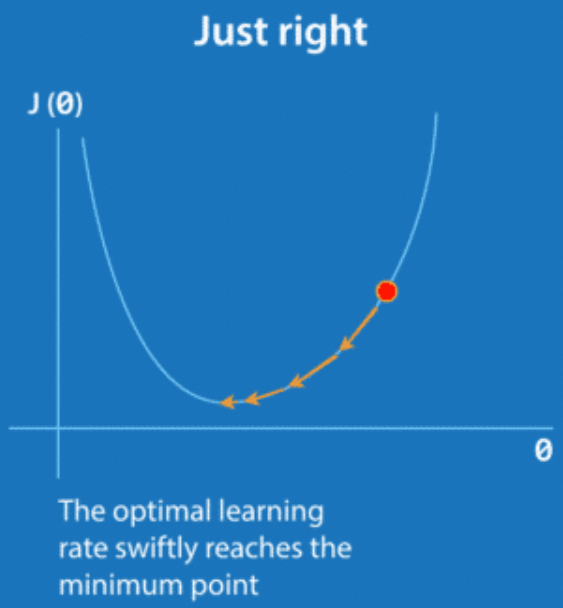
the gradient of $\mathcal{S}(\boldsymbol{\beta})$ is

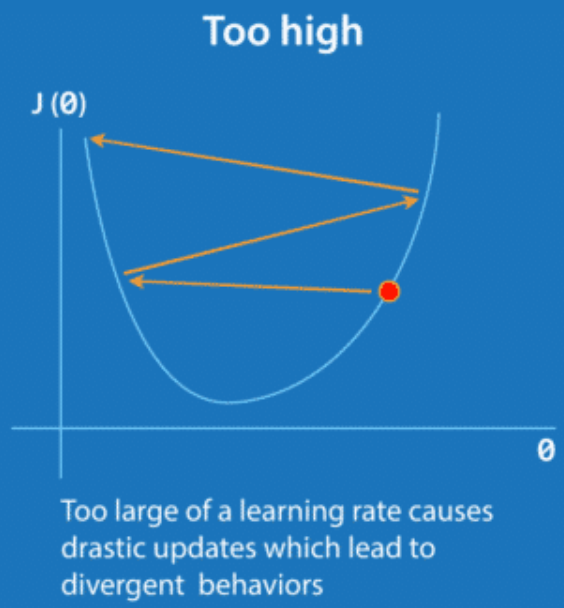
$$\nabla\mathcal{S}(\boldsymbol{\beta}) = 2\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})$$

we have the iterative form

$$\boldsymbol{\beta}^{(i)} = \boldsymbol{\beta}^{(i-1)} - \alpha\nabla\mathcal{S}(\boldsymbol{\beta})$$







GLM

- Three component of GLM (Generalized Linear Model)
  - Random component $f(\boldsymbol{y})$
  - Linear predictor $\boldsymbol{X\beta}$
  - Link function g such that $g\big(E(\boldsymbol{y})\big) = \boldsymbol{X\beta}$
- Exponential Dispersion Family

$$f(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

where

$$\mu = E(Y) = b'(\theta), \qquad \mathrm{Var}(Y) = b''(\theta)a(\phi)$$

the parameter θ is called the natural parameter, and ϕ is called the dispersion parameter.

Random Component	Link Function	Model
Normal	Identity	Regression Analysis of variance
Exponential family	Any	Generalized linear model
Binomial	Logit	Logistic regression
Multinomial	Generalized logits	Multinomial response
Poisson	Log	Loglinear

Constraint Least Square

- An intuitively appealing method to estimate a constrained linear model is to minimize the least squared criterion subject to the constraint $\boldsymbol{\mathcal{A}\beta} = \boldsymbol{c}$, by Lagrange Multiplier

$$\mathcal{S}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})^{\top}(\boldsymbol{y} - \boldsymbol{X\beta}) + \boldsymbol{\lambda}^{\top}(\boldsymbol{\mathcal{A}\beta} - \boldsymbol{c})$$

1st order condition

$$\frac{\partial \mathcal{S}}{\partial \boldsymbol{\beta}} = -\boldsymbol{X}^{\top}\boldsymbol{y} + \boldsymbol{X}^{\top}\boldsymbol{X\beta} + \boldsymbol{\mathcal{A}}^{\top}\boldsymbol{\lambda} = \boldsymbol{0}$$

$$\frac{\partial \mathcal{S}}{\partial \boldsymbol{\lambda}} = \boldsymbol{\mathcal{A}\beta} - \boldsymbol{c} = \boldsymbol{0}$$

multiplying by $\boldsymbol{\mathcal{A}}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}$

$$-\boldsymbol{\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\text{ols}} + \boldsymbol{\mathcal{A}\beta} + \boldsymbol{\mathcal{A}}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{\mathcal{A}}^{\top}\boldsymbol{\lambda} = \boldsymbol{0}$$

$$\Rightarrow \hat{\boldsymbol{\lambda}} = [\boldsymbol{\mathcal{A}}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{\mathcal{A}}^{\top}]^{-1}\left(\boldsymbol{\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{c}\right)$$

then

$$\widehat{\boldsymbol{\beta}}_{\text{cls}} = \widehat{\boldsymbol{\beta}}_{\text{ols}} - (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{\mathcal{A}}^{\top}[\boldsymbol{\mathcal{A}}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{\mathcal{A}}^{\top}]^{-1}\left(\boldsymbol{\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{c}\right)$$

Orthogonal Projection

- Let Ω be subspace of $\mathbb{R}^n$ and let $\boldsymbol{y}$ be a vector in $\mathbb{R}^n$. The expression

$$\boldsymbol{y} = \boldsymbol{y}_{\Omega} + \boldsymbol{y}_{\Omega^{\perp}}$$

where

$$\boldsymbol{y}_{\Omega} \in \Omega, \qquad \boldsymbol{y}_{\Omega^{\perp}} \in \Omega^{\perp}$$

is called orthogonal decomposition of $\boldsymbol{y}$ w.r.t Ω.

- The distance from $\boldsymbol{y}$ to Ω

$$\left\|\boldsymbol{y}_{\Omega^{\perp}}\right\| = \|\boldsymbol{y} - \boldsymbol{y}_{\Omega}\|$$

- Let A be $m \times n$ matrix with linearly independent columns and let $\Omega = C(\boldsymbol{X})$. Then the $n \times n$ matrix $\boldsymbol{X}^{\intercal}\boldsymbol{X}$ is invertible, and for all vectors $x \in \mathbb{R}^m$, we have

$$\boldsymbol{y}_{\Omega} = \boldsymbol{X}(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}\boldsymbol{X}^{\intercal}\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y}$$

and $\boldsymbol{P}$ is the orthogonal projection matrix with

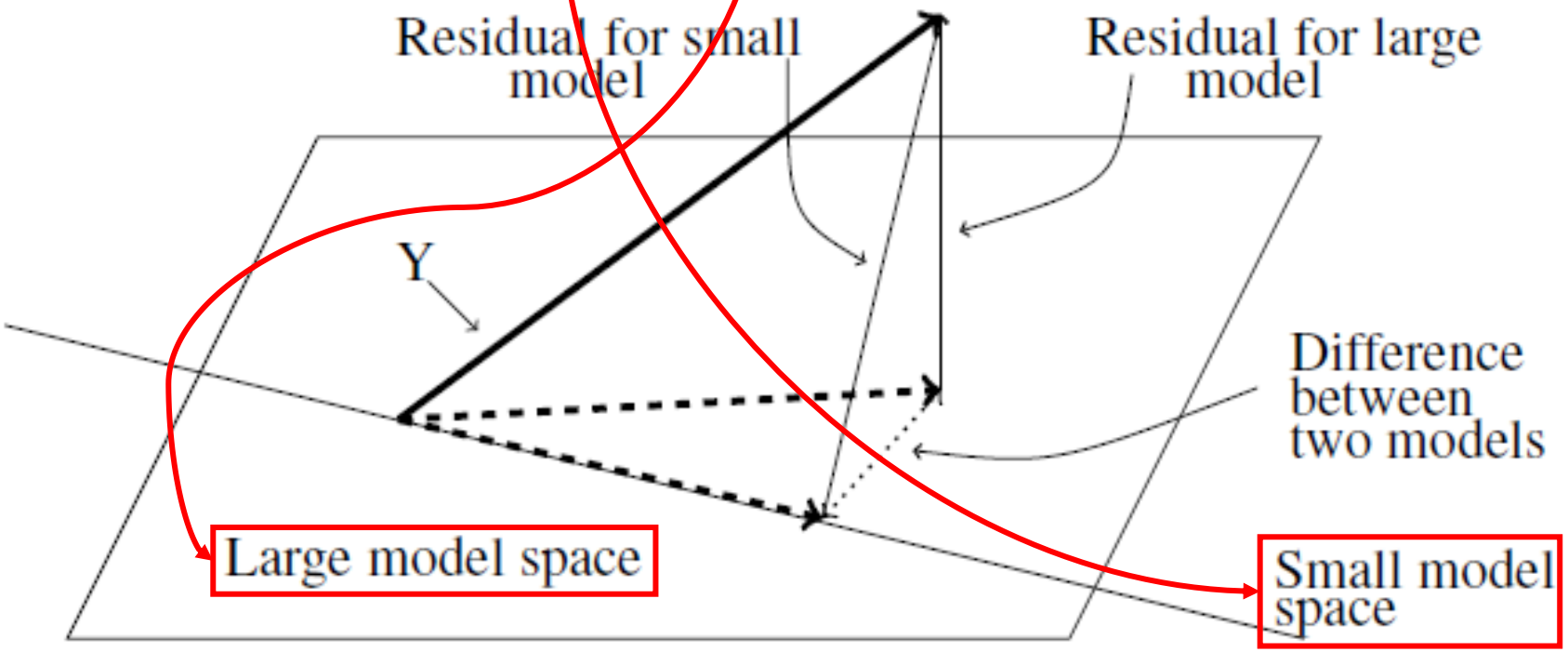
$$\boldsymbol{P}^2 = \boldsymbol{P} \text{ (idempotent)}$$

and

$$\boldsymbol{P}^{\intercal} = \boldsymbol{P} \text{ (symmetry)}$$

Hypothesis Testing - 1

- Geometric view of H_0: ω vs H_1: $\Omega \setminus \omega$



residual for large model $\triangleq \hat{\varepsilon}_\Omega$
residual for small model $\triangleq \hat{\varepsilon}_\omega$
difference between two models $\triangleq \hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega$

Hypothesis Testing - 2

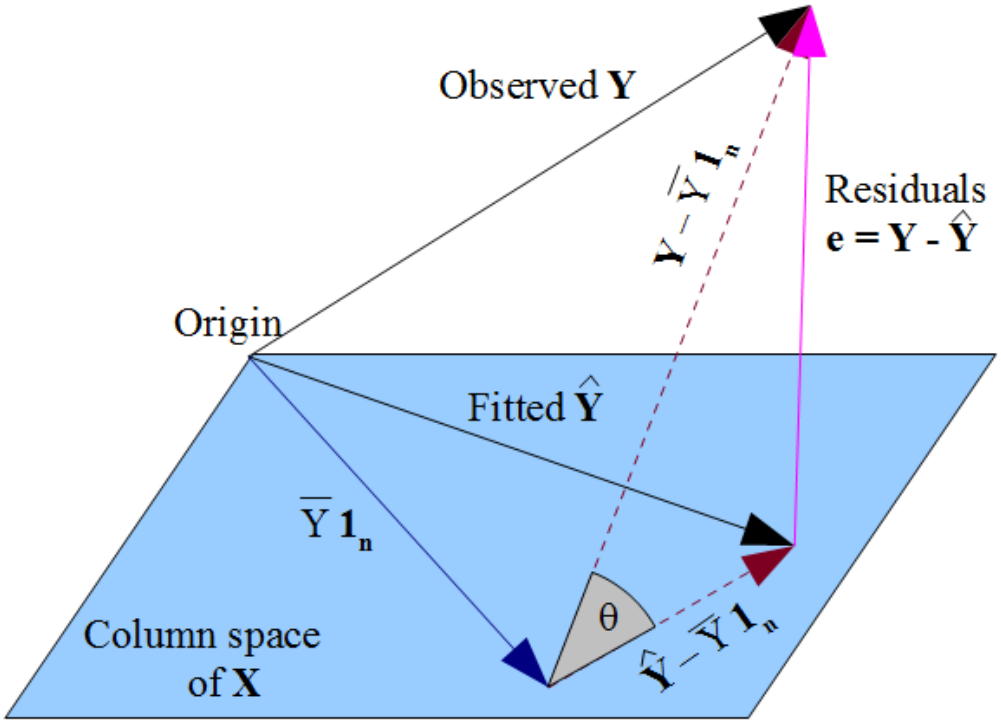
- Analysis of Variance (ANOVA)

$$\text{SST} = \mathbf{y}^\top(\boldsymbol{I} - \boldsymbol{\mathcal{J}})\boldsymbol{y} = \boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} + \boldsymbol{y}^\top(\boldsymbol{H} - \boldsymbol{\mathcal{J}})\boldsymbol{y}$$
$$= \text{SSE} + \text{SSR}$$

where

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top, \qquad \mathcal{J} = \frac{\mathbf{1}\mathbf{1}^\top}{n}$$

- Actually, ANOVA is a Pythagorean equality, as illustrate below



Hypothesis Testing - 3

- The length of the difference between two ω and Ω is
$$\|\hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega\|^2 = \|\hat{\varepsilon}_\omega\|^2 - \|\hat{\varepsilon}_\Omega\|^2 = \mathrm{RSS}_\omega - \mathrm{RSS}_\Omega$$
suggest divide with length of residual of Ω for scaling
$$\tan^2(\theta) = \frac{\|\hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega\|^2}{\|\hat{\varepsilon}_\Omega\|^2} = \frac{\mathrm{RSS}_\omega - \mathrm{RSS}_\Omega}{\mathrm{RSS}_\Omega}$$
- Why not divided by RSS_ω?

  Orthogonality and central χ^2 under ω.
- Under H_0, $\hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega = (\boldsymbol{H}_\Omega - \boldsymbol{H}_\omega)\boldsymbol{y} \sim N(0, (\boldsymbol{H}_\Omega - \boldsymbol{H}_\omega)\sigma^2)$
$\mathrm{RSS}_\omega - \mathrm{RSS}_\Omega \perp \mathrm{RSS}_\Omega$
$(\mathrm{RSS}_\omega - \mathrm{RSS}_\Omega)/\sigma^2 \sim \chi^2_{df_\omega - df_\Omega}$, $\mathrm{RSS}_\Omega/\sigma^2 \sim \chi^2_{df_\Omega}$
where $df_\omega = \dim(\omega^\perp)$ and $df_\Omega = \dim(\Omega^\perp)$.
- General form
$$F = \frac{(\mathrm{RSS}_\omega - \mathrm{RSS}_\Omega)/(df_\omega - df_\Omega)}{\mathrm{RSS}_\Omega/df_\Omega} \sim F_{df_\omega - df_\Omega, df_\Omega}$$

Hypothesis Testing - 4

- Test of all predictors

$$H_0: \beta_1 = \cdots = \beta_p$$
$$\Omega: y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$
$$\omega: y = \beta_0 + \varepsilon$$

- Test one predictor

$$H_0: \beta_k = 0$$
$$\Omega: y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \cdots + \beta_p X_p + \varepsilon$$
$$\omega: y = \beta_0 + \beta_1 X_1 + \cdots + \cancel{\beta_k X_k} + \cdots + \beta_p X_p + \varepsilon$$

- Test pair of predictors

$$H_0: \beta_i = \beta_j = 0$$
$$\Omega: y = \beta_0 + \beta_1 X_1 + \cdots + \beta_i X_i + \cdots + \beta_j X_j + \cdots + \beta_p X_p + \varepsilon$$
$$\omega: y = \beta_0 + \beta_1 X_1 + \cdots + \cancel{\beta_i X_i} + \cdots + \cancel{\beta_j X_j} + \cdots + \beta_p X_p + \varepsilon$$

- Test a subspace/subset

$$H_0: \beta_i = \beta_j$$
$$\Omega: y = \beta_0 + \beta_1 X_1 + \cdots + \beta_i X_i + \cdots + \beta_j X_j + \cdots + \beta_p X_p + \varepsilon$$
$$\omega: y = \beta_0 + \beta_1 X_1 + \cdots + \beta_i (X_i + X_j) + \cdots + \cancel{\beta_j X_j} + \cdots + \beta_p X_p + \varepsilon$$

Hypothesis Testing - 5

- Linear Hypothesis Test

$$H_0: \boldsymbol{R\beta} = \boldsymbol{c} \text{ vs } H_1: \boldsymbol{R\beta} \neq \boldsymbol{c}$$

$$H_0: 2\beta_1 - \beta_2 = \beta_2 - 2\beta_3 + 3\beta_4 = \beta_1 - \beta_4 = 0$$

$$H_0: \begin{bmatrix} 0 & 2 & -1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 3 \\ 0 & 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

- Let $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2 I)$

$$F = \frac{(\boldsymbol{R\beta} - \boldsymbol{c})^{\intercal}\left(\boldsymbol{R}(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}\boldsymbol{R}^{\intercal}\right)^{-1}(\boldsymbol{R\beta} - \boldsymbol{c})/\boldsymbol{q}}{\text{RSS}/(n-k-1)}$$

  - If $H_0: \boldsymbol{R\beta} = \boldsymbol{c}$ is false then

$$F \sim F_{q,n-k-1,\boldsymbol{\lambda}}$$

    where $\lambda = (\boldsymbol{R\beta} - \boldsymbol{c})^{\intercal}\left(\boldsymbol{R}(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}\boldsymbol{R}^{\intercal}\right)^{-1}(\boldsymbol{R\beta} - \boldsymbol{c})/2\sigma^2$

  - If $H_0: \boldsymbol{R\beta} = \boldsymbol{c}$ is true then

$$F \sim F_{q,n-k-1}$$

Goodness-of-fit

- R^2 is a measure of the goodness of fit of a model. In regression, R^2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points.

$$\Omega: y = \beta_0 + \beta_1 g_1(\boldsymbol{x}) + \cdots + \beta_p g_p(\boldsymbol{x}) + \varepsilon$$

$$\omega: y = \beta_0 + \varepsilon$$

$$R^2 = 1 - \frac{\mathrm{RSS}_\Omega}{\mathrm{RSS}_\omega} = \left(\frac{\mathrm{RSS}_\omega - \mathrm{RSS}_\Omega}{\mathrm{RSS}_\omega}\right) = [\mathrm{Cor}(y, \hat{y})]^2 \in [0,1]$$

- Interpretation of R^2: Proportion of total variation in y that can be explained by the effects $g_1, \cdots g_p$
  - Increase in the No. of $\boldsymbol{X}$ increase the value of R^2, R^2 alone can't be used as a meaningful comparison of models with very different No. of $\boldsymbol{X}$.
  - R^2 does not make any sense if an intercept is not in.
  - R^2 does not indicate whether the model $E(y) = \boldsymbol{X\beta}$ is correct.

Confidence Interval and region - 1

- An interval/region estimation provides
  - Plausible values for parameter.
  - Uncertainty in parameter estimator.
  - Information about its length and the values it covers may be helpful.
  - Information related to testing.
- Confidence region for $\boldsymbol{R\beta}$ where $\boldsymbol{R}$ is full rank matrix

$$\frac{\left(\boldsymbol{R\hat{\beta}}-\boldsymbol{R\beta}\right)^{\top}(\boldsymbol{R}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{R})^{-1}\left(\boldsymbol{R\hat{\beta}}-\boldsymbol{R\beta}\right)}{d\hat{\sigma}^{2}} \leq F_{d,n-p}(\alpha)$$

- Confidence region for $\boldsymbol{\beta}$

$$\boldsymbol{R}=\boldsymbol{I}_{p\times p}$$

Confidence region of β_i, β_j

$$\boldsymbol{R}=\begin{bmatrix} 0 & \cdots & 0 & \overbrace{1}^{i^{\text{th}}} & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & \underbrace{1}_{j^{\text{th}}} & 0 & \cdots & 0 \end{bmatrix}$$

Confidence Interval and region - 2

- Confidence interval for prediction of mean response at $\boldsymbol{x}_0$

$$\boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}} \pm t_{n-p}\left(\frac{\alpha}{2}\right)\left\{\hat{\sigma}\sqrt{\boldsymbol{x}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_0}\right\}$$

- What is the danger of extrapolation?
  The fitted model may approximate the true model very badly or wrongly and you don't have any information about it.
  - Interpolation: $\boldsymbol{x}_0$ lie within the range of $\boldsymbol{X}$
  - Extrapolation: $\boldsymbol{x}_0$ lie outside the range of $\boldsymbol{X}$

- C.I. for prediction of future observation at $\boldsymbol{x}_0$

$$\boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}} \pm t_{n-p}\left(\frac{\alpha}{2}\right)\left\{\hat{\sigma}\sqrt{1+\boldsymbol{x}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_0}\right\}$$

Orthogonality

- Consider the two models
  - Model 1: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$
  - Model 2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

  In general, $\hat{\beta}_1$ in the two models are not identical. If fitted model is Model 1 and true model is Model 2 then

$$E(\hat{\beta}_1) = \beta_1 + (X_1^\top X_1)^{-1} X_1^\top X_2 \beta_2$$

- An exception:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

$$\boldsymbol{X}^\top\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1^\top\boldsymbol{X}_1 & \boldsymbol{X}_1^\top\boldsymbol{X}_2 \\ \boldsymbol{X}_2^\top\boldsymbol{X}_1 & \boldsymbol{X}_2^\top\boldsymbol{X}_2 \end{bmatrix}$$

when X_1 and X_2 are orthogonal ($\boldsymbol{X}_1 \perp \boldsymbol{X}_2 \Rightarrow \boldsymbol{X}_1^\top\boldsymbol{X}_2 = 0$).

$$(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \begin{bmatrix} (\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{X}_2^\top\boldsymbol{X}_2)^{-1} \end{bmatrix}$$

then the estimation $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ independent

Identifiability

- $\boldsymbol{\beta}$ is called unidentifiable when $\boldsymbol{X}^\top\boldsymbol{X}$ is singular, which means the normal equation $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top\boldsymbol{y}$ has infinite solution.
- Unidentifiable means
  - Insufficient data to estimate the parameters of interest.
  - More parameters than are necessary to model the data.
- When does unidentifiability happen?

  Observational data
  - Same predictor measured in different scales and both in model.
  - $X_1 + X_2 = X_3$ or $X_1 + X_2 + X_3 = c$ and all three in model.
  - $\boldsymbol{X}$ is supersaturated. $(p > n)$

  Experimental data

$$y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

  - Over-parameterized: some constraint must be imposed on.

$$\sum_{i=1}^{r} \alpha_i = 0$$

Interpreting

- What does $\widehat{\boldsymbol{\beta}}$ mean?

  Some β_i have physical interpretation, especially those from conceptual model. Usually β_i do have such physical interpretation. The sign of $\widehat{\boldsymbol{\beta}}$ indicate direction of the relationship between the term and the response.

- A naïve interpretation (causality statement)
  - **A unit increase in X_i will cause an average change of $\hat{\beta}_i$ in Y.**
- An alternative interpretation
  - **A unit increase in X_i with all the other(specified) terms held constant will be associated with an average change $\hat{\beta}_i$ in Y.**
- An interpretation from prediction viewpoint
  - Regrading the parameters and their estimates as fictional quantities, and concentrating on prediction enable a rather cautious interpretation of $\hat{\beta}$

    given $(g_{1,0}, \cdots, g_{i,0}, \cdots, g_{p,0}) \rightarrow \hat{y}_0$

    observe $(g_{1,0}, \cdots, g_{i,0} + 1, \cdots, g_{p,0}) \rightarrow \hat{y}_0 + \hat{\beta}_i$

Causality ≠ Association

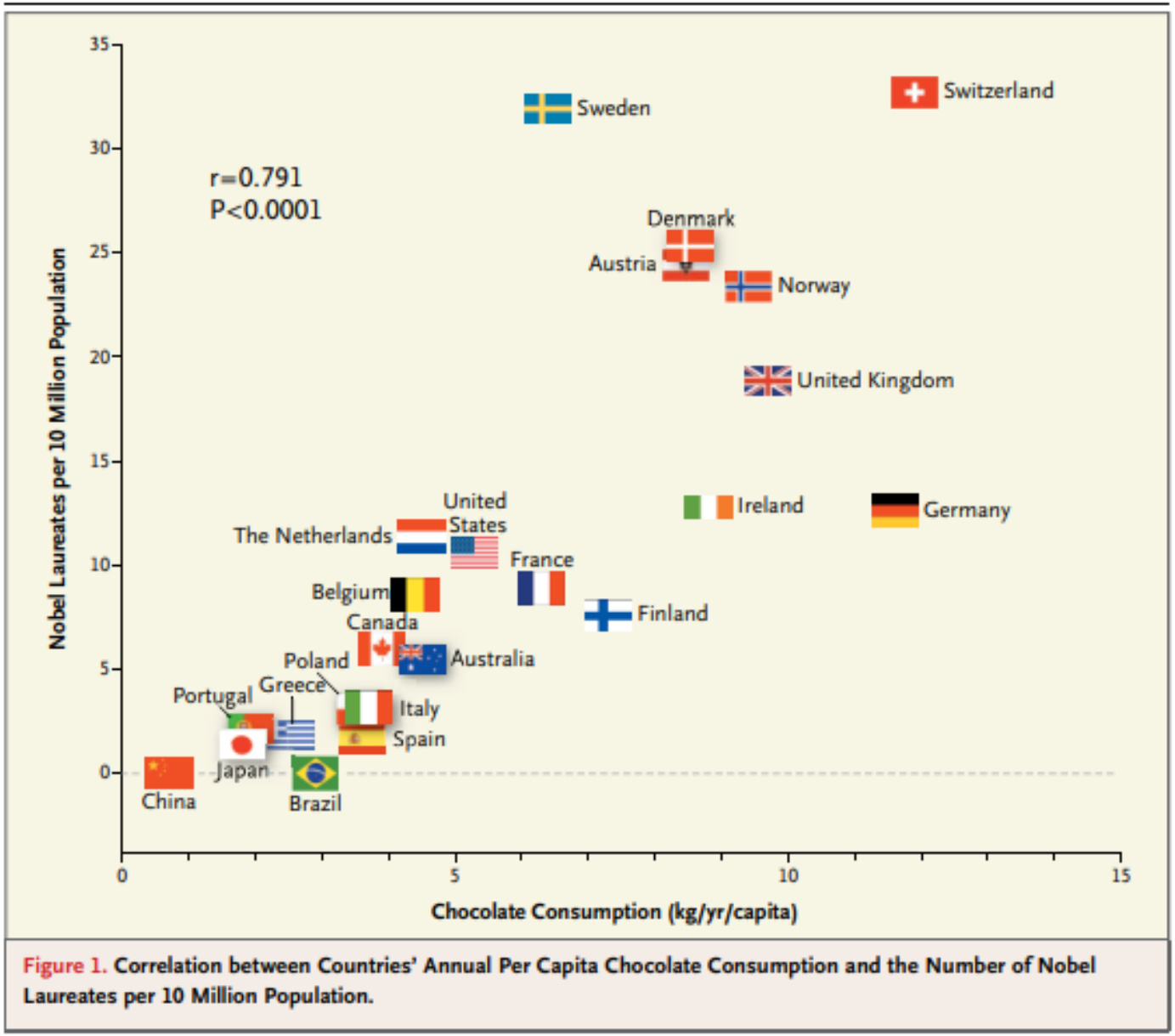
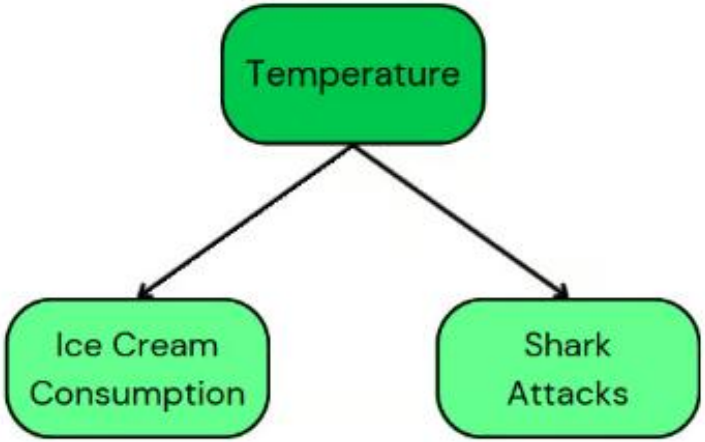
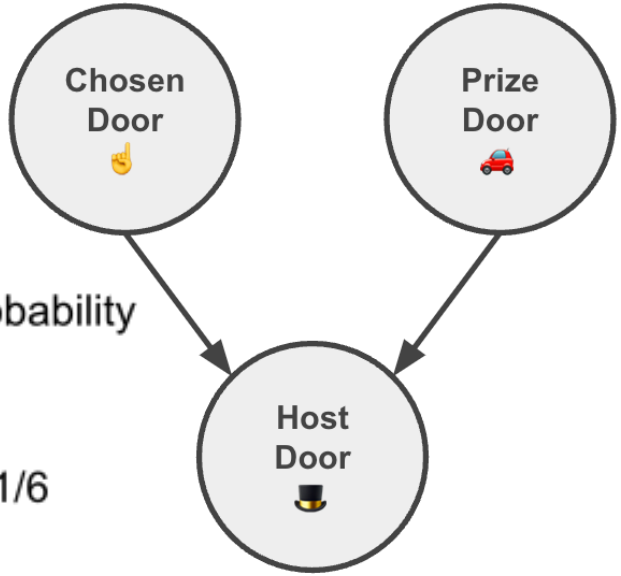


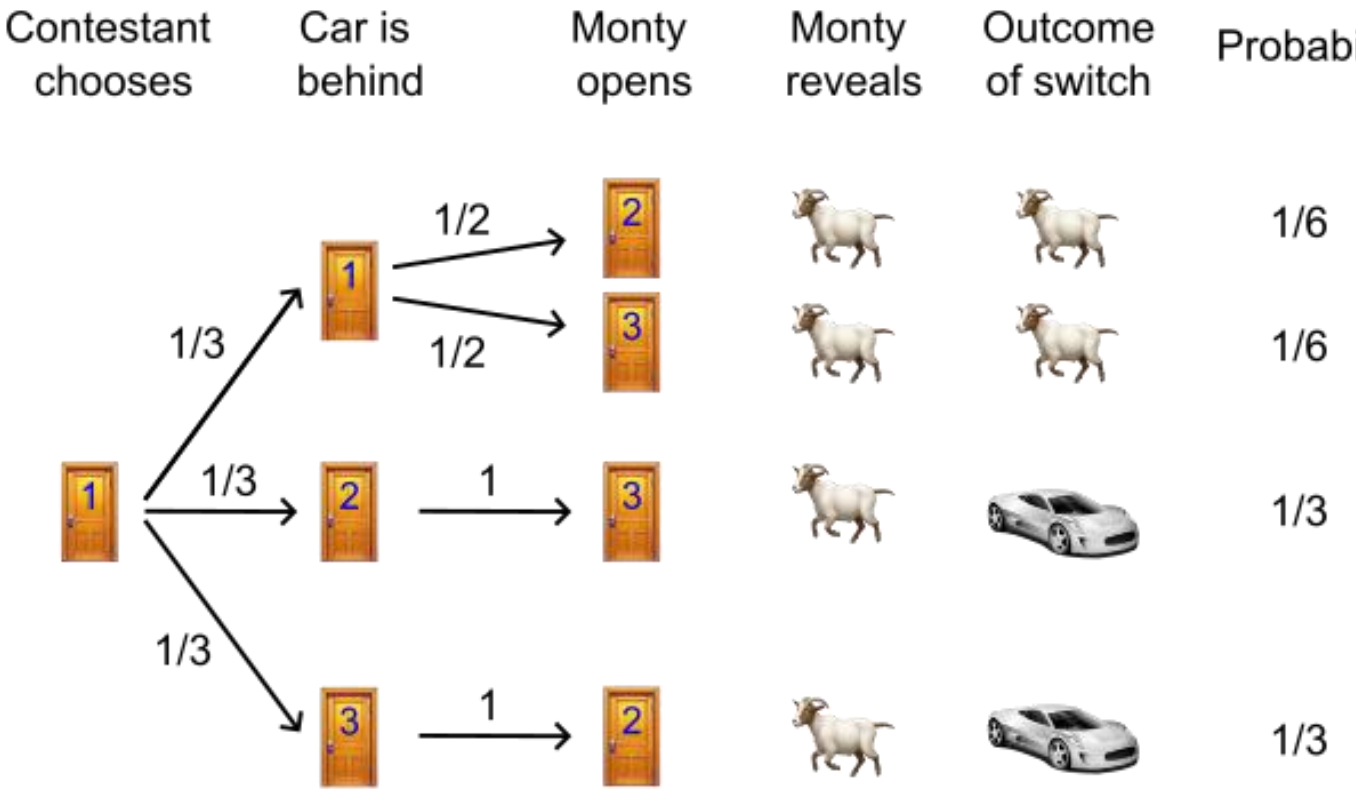
Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Common Cause



Collider – Monty Hall Problem





What can go wrong?

- Source and quality of the data
  - Not random sample. Ex. Biased sample, sample of convenience.
  - Important predictors may not have been observed.
  - Observational data often make causal conclusion problematic.
  - The range of $\boldsymbol{X}$ and qualitative nature of some predictors may limit effective predictions, it's unsafe to extrapolate.
- Error component $\boldsymbol{\varepsilon}$
  - May have unequal variance.
  - May be correlated.
  - May not be normally distributed.
- Structural component $\boldsymbol{X\beta}$
  - Errors in $\boldsymbol{X}$. (Measurement error)
  - Series collinearity in $\boldsymbol{X}$.
- Many theory rests on the assumption that the model is correct.
- Experimenter Bias.

GLS (Generalized Least Square)

- What if $\mathrm{Var}(\boldsymbol{\varepsilon}) \neq \sigma^2 \boldsymbol{I}$?
  - Time series correlation. $\varepsilon_t \sim \mathrm{ARMA}(p, g)$
  - Repeated measurement model.
  - Spatial correlation.
  - Nested errors.
- Consider $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma}$ is known but σ^2 is unknown. Since $\boldsymbol{\Sigma}$ is symmetric and positive definite ($\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} > 0$), by Cholesky decomposition $\boldsymbol{\Sigma} = SS^\top$.
- Multiply both side of $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\boldsymbol{y}^* = S^{-1}\boldsymbol{y} = S^{-1}\boldsymbol{X}\boldsymbol{\beta} + S^{-1}\boldsymbol{\varepsilon} = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$$

  then

$$\mathrm{Var}(S^{-1}\boldsymbol{\varepsilon}) = S^{-1}\sigma^2\boldsymbol{\Sigma}(S^{-1})^\top = \sigma^2 S^{-1}(SS^\top)(S^{-1})^\top = \sigma^2 \boldsymbol{I}$$

- Find $\boldsymbol{\beta}$ that minimize

$$\mathcal{S}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

$$\Rightarrow \widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{y}$$

Lack of fit

- What fit is appropriate?

  X too simple

  - Not enough to explain the mean structure in data.
  - Lack of fit. (under-fitting)
  - $\hat{\sigma}^2$ over-estimated.

  X too complex

  - Will explain the variation caused by errors, in addition to the mean structure.
  - Overfit the data.
  - $\hat{\sigma}^2$ under-estimated.

- What statistic carry the information about lack of fit or overfit?
  - $\hat{\sigma}^2$

- $H_0: E(\boldsymbol{y}) = \boldsymbol{X\beta}$ is correct **vs** $H_1: E(\boldsymbol{y}) = \boldsymbol{X\beta}$ is too simple
  - Compare $\hat{\sigma}^2$ to σ^2
    σ^2 is known **vs** σ^2 is unknown

Diagnostics

- Check model assumptions to suggest further improvement after fitting. The building of an empirical model is an iterative process. During process, it required to check whether the current fitted model is consistent with data.

- What assumptions need to be checked?

$$\text{model: } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

  - Error structure: error independent, homoscedasticity, normally distributed.
  - Mean structure: whether $E(\boldsymbol{y}) = \boldsymbol{X\beta}$ is a correct structure.
  - Unusual observations: where some observations do not fit the model.

- Two types of diagnostic techniques
  - Numerical vs Graphical

Leverage

- Recall residual

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}$$

- Leverage, consider the diagonal of $\boldsymbol{H}$, let

$$h_i = \boldsymbol{H}_{i,i}$$

  - $\boldsymbol{x_i}$ whose h_i is $\left|\begin{matrix}\text{large}\\ \text{small}\end{matrix}\right|$ then $\text{Var}(\hat{\varepsilon}_i)$ $\left|\begin{matrix}\text{small}\\ \text{large}\end{matrix}\right|$

    $\Rightarrow \left|\begin{matrix}\text{fitted model has to force to fit close to } y_i\\ \text{in this } \boldsymbol{x}_i\text{, model cannot fit so well}\end{matrix}\right|$
  - h_i roughly determines how close $(\boldsymbol{x}_i, y_i)$ to the surface $(\boldsymbol{x}_i, \hat{y}_i)$
  - h_i corresponds to Mahalanobis distance.

- What h_i are too large?

$$\sum_{i=1}^{n} h_i = p, \qquad \frac{1}{n} \le h_i \le 1 \ \forall i \Rightarrow \text{large leverage} \gg \frac{p}{n}$$

$$\Rightarrow \text{rule of thumb: if } h_i > \frac{2p}{n}$$

Residuals - 1

- (Internally) studentized residual r_i
  - Because $\mathrm{Var}(\hat{\varepsilon}_i) = (1 - h_i)\sigma^2$ then

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_i}}, \qquad \mathrm{Var}(r_i) \approx 1$$

- (Externally) jackknife residual t_i
  - Exclude i^{th} observation and re-compute the estimates to get $\hat{\beta}_{(i)}$ and $\hat{\sigma}_{(i)}$, where (i) denotes that i^{th} case has been excluded.
  - Since

$$\mathrm{Var}\left(\mathrm{y_i} - \hat{y}_{(i)}\right) = \sigma^2\left[1 + \boldsymbol{x}_i^{\top}\left(\boldsymbol{X}_{(i)}{}^{\top}\boldsymbol{X}_{(i)}\right)^{-1}\boldsymbol{x}_i\right]$$

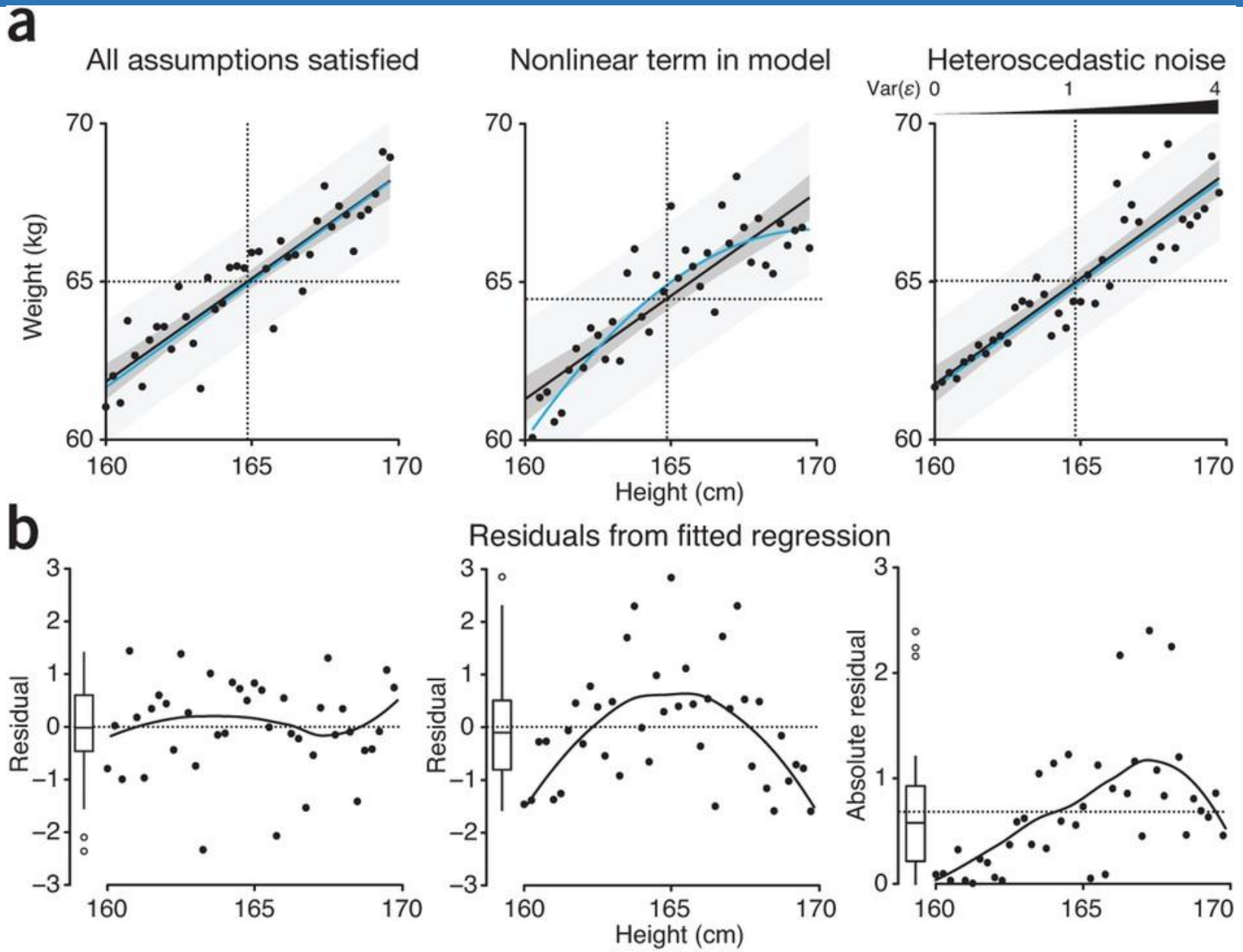
$$t_i = \frac{\mathrm{y_i} - \hat{y}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{1 + \boldsymbol{x}_i^{\top}\left(\boldsymbol{X}_{(i)}{}^{\top}\boldsymbol{X}_{(i)}\right)^{-1}\boldsymbol{x}_i}}$$

$$= r_i\sqrt{\frac{n - p - 1}{n - p - r_i^2}}$$

Residuals - 2

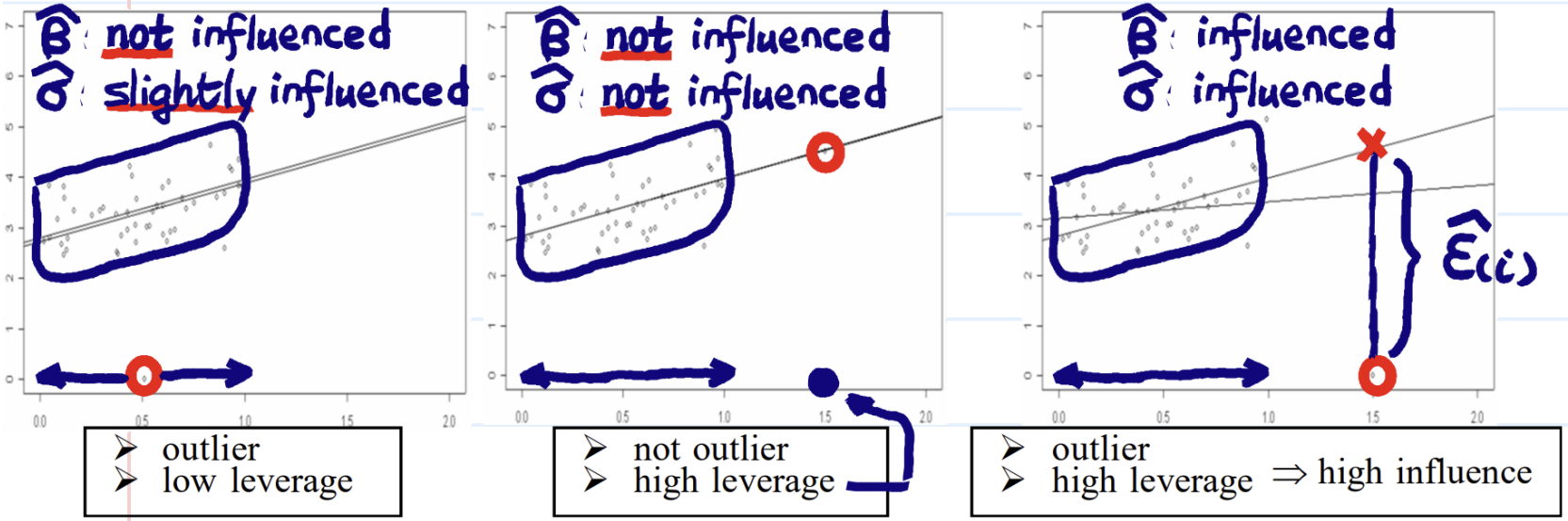
- Test for outliers
  - Given a specific case i, conclude an outlier if
$$|t_i| > T_{n-p-1}\left(\frac{\alpha}{2}\right)$$
  - In practice, a few (or all) t_i will be tested ⇒ Multiple testing
- Bonferroni Correction
  - Test need to adjust the significant level of the test accordingly.
  - Consider the test
    H_0: no outlier in the n obs **vs** H_1: at least one outlier
$$\alpha^* = P\left(\bigcup_{i=1}^{n} R_i \middle| H_0\right) \leq \sum_{i=1}^{n} P(R_i|H_0) = n\alpha$$
  - Conclude an outlier if
$$|t_i| > t_{n-p-1}\left(\frac{\alpha}{2n}\right)$$
  - It is conservative, tends not to label points as outlier.

Residual Plot



Influential observation

- An influential point may or may not be an outlier and may or may not have large leverage but it will tend to have at least one of those two properties.



- Cook's distance (scale and unit free)

$$D_i = \frac{\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)^{\top} (\boldsymbol{X}^{\top}\boldsymbol{X}) \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)}{p\hat{\sigma}^2} = \left(\frac{1}{p}\right) r_i^2 \left(\frac{h_i}{1-h_i}\right)$$

- It is a combination of residual and leverage. If assume $\boldsymbol{X}$ is normal, can do a test on D_i.

Variance Stabilizing transformation - 1

- It's better try to understand the cause of non-constant variance before taking any remedies.
  - Large response have more room to vary.
  - Response constrained to lie between a maximum and a minimum.
  - Response from Poisson or binomial distribution,…
- Remedies for non-constant variance
  - Weighted least squares
  - Find a transformation h such that $\mathrm{Var}\big(h(y)\big)$ is a constant.

$$h(y) \approx h\big(E(y)\big) + h'\big(E(y)\big)\big(y - E(y)\big)$$

$$\mathrm{Var}(h(y)) \approx \big\{h'\big(E(y)\big)\big\}^2 \mathrm{Var}(y) = c \Rightarrow h'\big(E(y)\big) \propto \frac{1}{\sqrt{\mathrm{Var}(y)}}$$

$$h\big(E(y)\big) = \int \frac{1}{\sqrt{\mathrm{Var}(y)}} d\big(E(y)\big) = \int \frac{1}{\sqrt{g(\mu)}} d(\mu)$$

  - Ex. $\mathrm{Var}(y) \propto [E(y)]^2 \Rightarrow g(\mu) = \mu^2$

$$h(\mu) = \int \frac{1}{\sqrt{\mu^2}} d\mu = \ln(\mu) \Rightarrow \text{suggest } h(y) = \ln(y)$$

Variance Stabilizing transformation - 2

TABLE 5.1 Useful Variance-Stabilizing Transformations

Relationship of σ^2 to $E(y)$	Transfonnation
$\sigma^2 \propto$ constant	$y' = y$ (no transformation)
$\sigma^2 \propto E(y)$	$y' = \sqrt{y}$ (square root; Poisson data)
$\sigma^2 \propto E(y)[1 - E(y)]$	$y' = \sin^{-1}\left(\sqrt{y}\right)$ (arcsin; binomial proportions $0 \le y_i \le 1$)
$\sigma^2 \propto [E(y)]^2$	$y' = \ln(y)$(log)
$\sigma^2 \propto [E(y)]^3$	$y' = y^{-1/2}$ (reciprocal square root)
$\sigma^2 \propto [E(y)]^4$	$y' = y^{-1}$(reciprocal)

- Other transformation
  - Box-Cox transformation (make the errors as nearly like normal)

$$t_\lambda(y) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \ln(y), & \text{if } \lambda = 0 \end{cases}$$

choose λ to fit data well using maximum likelihood

$$H_0: \lambda = \lambda_0 \textbf{ vs } H_1: \lambda \neq \lambda_0$$

the likelihood ratio test

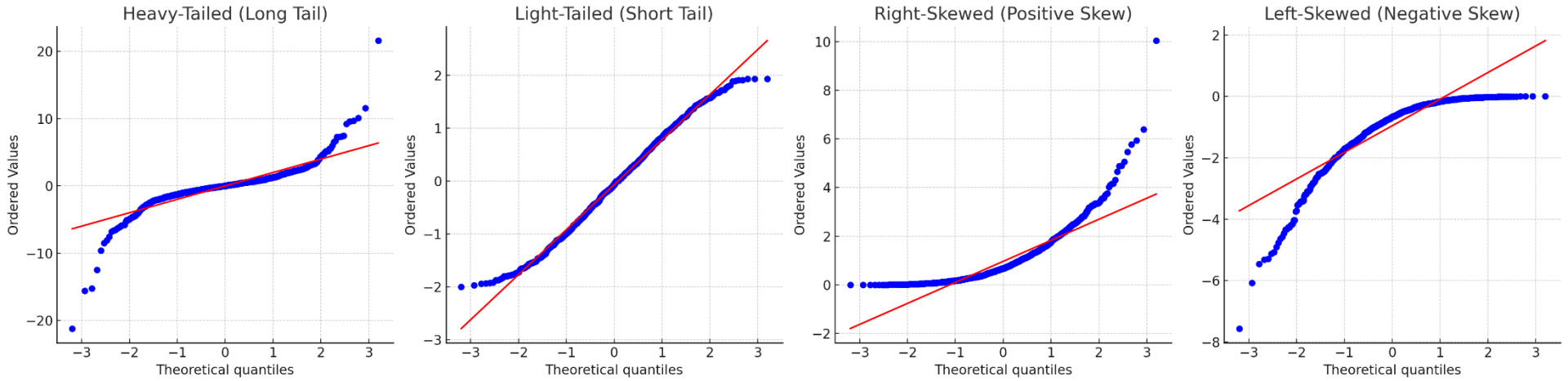
$$-2\left[\mathcal{L}(\lambda_0) - \mathcal{L}\left(\hat{\lambda}\right)\right] \sim \chi^2_{(1)} \text{ under } H_0$$

Normality Test

- It is used to determine if data is well-modeled by normal distribution and to compute how likely it is for a random variable underlying the data to be normal.
  - Bell curve shape.
  - Symmetric around the $x = \mu$.
  - Log-concave.
- Graphical methods
  - Normal probability plot.
  - Quantile-Quantile plot (Q-Q plot).
- Numerical methods
  - D'Agostino's K-squared Test.
  - Jarque-Bera Test.
  - Anderson-Darling Test.
  - Kolmogorov-Smirnov Test.
  - Shapiro-Wilk Test.

Q-Q plot

- Assessing normality assumption of $\boldsymbol{\varepsilon}$
  - Sort the data $\hat{\varepsilon}_{(1)} \leq \hat{\varepsilon}_{(2)} \leq \cdots \leq \hat{\varepsilon}_{(n)}$
  - Plot $\hat{\varepsilon}_{(i)}$ against $\Phi^{-1}\left(\frac{i}{n+1}\right)$ where $\Phi(\cdot)$ is cdf of $N(0,1)$
- If the residuals are normal distributed, an approximately straight-line relationship will be observed.
- Non-normality:
  - long-tail
  - short-tail
  - asymmetric



Collinearity

- A challenge arises when $\boldsymbol{X}^\top \boldsymbol{X}$ is close to singular but not exactly so. It leads to imprecise estimates of $\boldsymbol{\beta}$. The standard errors are inflated so that T test may fail to reveal significant factors.
- It can be detected in several ways:
  - VIF (Variance Inflation factor)

$$\mathrm{Var}(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1 - R_j^2}\right)\left(\frac{1}{\sum_i (g_{ij} - \bar{g}_j)^2}\right)$$

  - Let $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ be eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$. Zero eigenvalues denote exact collinearity, while the presence of some small eigenvalues indicates collinearity. The condition number κ measures the relative sizes of the eigenvalues and is defined

$$\kappa = \sqrt{\lambda_1 / \lambda_p}$$

    where $\kappa > 30$ is consider large. Other condition numbers, $\sqrt{\lambda_1 / \lambda_i}$ are also worth considering.

PCA - 1

- Let Σ be the covariance matrix associated with the $\boldsymbol{X} = (X_1, \cdots, X_p)$, and Σ have the eigenvalue-eigenvector pairs $(\lambda_1, e_1), \cdots, (\lambda_p, e_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Then the i^{th} principal component (PC) is given by

$$Z_i = e_i^{\top}\boldsymbol{X} = e_{i,1}X_1 + \cdots + e_{i,p}X_p, \qquad i = 1, \cdots, p$$

with these choices

$$\text{Var}(Z_i) = e_i^{\top}\Sigma e_i = \lambda_i, \qquad i = 1, \cdots, p$$
$$\text{Cov}(Z_i, Z_j) = e_i^{\top}\Sigma e_j = 0, \qquad i \neq j$$

Then

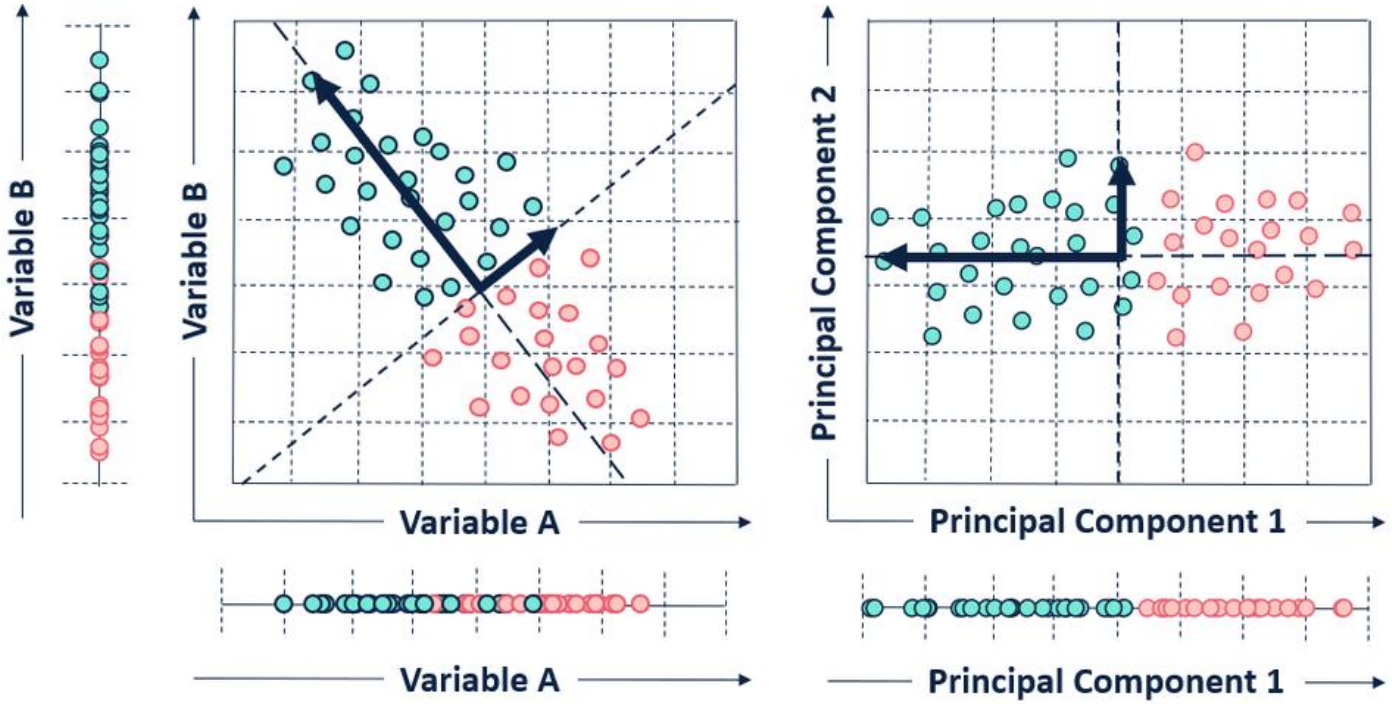
$$\sigma_{1,1} + \cdots + \sigma_{p,p} = \sum_{i=1}^{p} \text{Var}(X_i) = \lambda_1 + \cdots + \lambda_p = \sum_{i=1}^{p} \text{Var}(Z_i)$$

- Proportion of total population variance due to i^{th} PC is

$$\frac{\lambda_i}{\lambda_1 + \cdots + \lambda_p}, \qquad i = 1, \cdots, p$$

PCA - 2

- For a geometric interpretation of PCA, suppose we have two features A and B. The 1st PC is a line through the widest part; the 2nd PC is the line at right angles to the 1st PC.
- In other words, the 1st PC goes through the fattest part of the ″football ″ and the 2nd PC through the next fattest part of the ″football ″ and orthogonal to the 1st PC; and so on.



PCA - 3

- How many PCs to retain?
  - Screen plot.
  - Amount of total variance explained (say 70% ~ 90%).
  - Cutoff point 1 or 0.7.
- What are the distributions of the $(\hat{\lambda}_1, \cdots, \hat{\lambda}_p)$ and $(\hat{e}_1, \cdots, \hat{e}_p)$?
  - Let $\Lambda = \text{diag}\{\lambda_1, \cdots, \lambda_p\}$ of Σ, then $\sqrt{n}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \sim N(\mathbf{0}, 2\Lambda^2)$
  - Let

$$E_i = \lambda_i \sum_{\substack{k=1 \\ k \neq i}}^{p} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} e_k e_k^\top$$

    then $\sqrt{n}(\hat{e}_i - e_i) \sim N(\mathbf{0}, E_i)$
  - A large sample $100(1-\alpha)\%$ confidence interval for λ_i is thus

$$\frac{\hat{\lambda}_i}{\left(1 + Z_{\alpha/2}\sqrt{2/n}\right)} \leq \lambda_i \leq \frac{\hat{\lambda}_i}{\left(1 - Z_{\alpha/2}\sqrt{2/n}\right)}$$

Serial correlation

- When a time order is available
  - Plot $\hat{\varepsilon}$ against time
  - Plot $\hat{\varepsilon}_{t+1}$ against $\hat{\varepsilon}_t$, when t related to time
  - Use formal tests (Ex. Durbin-Watson)

$$\mathrm{DW} = \frac{\sum_{t=2}^{T}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{T}\hat{\varepsilon}_t^2} \in [0,4]$$

    - positive correlated $\Rightarrow$ DW $\rightarrow 0$
    - negative correlated $\Rightarrow$ DW $\rightarrow 4$
    - under null $\Rightarrow$ DW ≈ 2
    - Null distribution depends on $\boldsymbol{X}$
  - Rewrite the test statistic by lag one sample autocorrelation

$$\mathrm{DW} \approx 2(1-\hat{\rho}_1), \qquad \hat{\rho}_1 = \frac{\sum_{t=1}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_{t+1}}{\sum_{t=1}^{T}\hat{\varepsilon}_t^2}$$

    The DW statistic gets smaller as the serial correlation increase.

Dummy

- Categorical (qualitative) predictors
  - Nominal **vs** ordinal.
  - What properties can we explore for qualitative predictor?

    category $i \rightarrow y_{i,j}$, $\quad \mu_i = E(y_{i,j})$

    can only study difference between μ_i.
- One dichotomous predictor (Two categories)

$$d(C) = \begin{cases} 1, & \text{if } C = c_1 \\ 0, & \text{if } C = c_2 \end{cases}$$

Ex. $y = \beta_0 + \beta_1 d + \varepsilon$

$$\begin{cases} \mu_1 = E(y|d=0) = \beta_0 \\ \mu_2 = E(y|d=1) = \beta_0 + \beta_1 \end{cases} \Rightarrow \begin{cases} \beta_0 = \mu_1 \\ \beta_1 = \mu_2 - \mu_1 \end{cases}$$

- One polytomous predictor (More than two categories)
  - For k categories, dummy variables $d_1, \cdots, d_{k-1}$ are needed to depict the difference between categories.

Coding Scheme

- Consider $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \varepsilon$ with 4 categories c_1, c_2, c_3, c_4.
  - Treatment coding (c_1 as a reference)

$$\begin{cases} \mu_1 = E(y|d_1 = 0, d_2 = 0, d_3 = 0) = \beta_0 \\ \mu_2 = E(y|d_1 = 1, d_2 = 0, d_3 = 0) = \beta_0 + \beta_1 \\ \mu_3 = E(y|d_1 = 0, d_2 = 1, d_3 = 0) = \beta_0 + \beta_2 \\ \mu_4 = E(y|d_1 = 0, d_2 = 0, d_3 = 1) = \beta_0 + \beta_3 \end{cases} \Rightarrow \begin{cases} \beta_0 = \mu_1 \\ \beta_1 = \mu_2 - \mu_1 \\ \beta_2 = \mu_3 - \mu_1 \\ \beta_3 = \mu_4 - \mu_1 \end{cases}$$

  - Sum coding (β_0 represent overall mean)

$$\begin{cases} \mu_1 = E(y|d_1 = -1, d_2 = -1, d_3 = -1) = \beta_0 - \beta_1 - \beta_2 - \beta_3 \\ \mu_2 = E(y|d_1 = 1, d_2 = 0, d_3 = 0) = \beta_0 + \beta_1 \\ \mu_3 = E(y|d_1 = 0, d_2 = 1, d_3 = 0) = \beta_0 + \beta_2 \\ \mu_4 = E(y|d_1 = 0, d_2 = 0, d_3 = 1) = \beta_0 + \beta_3 \end{cases}$$

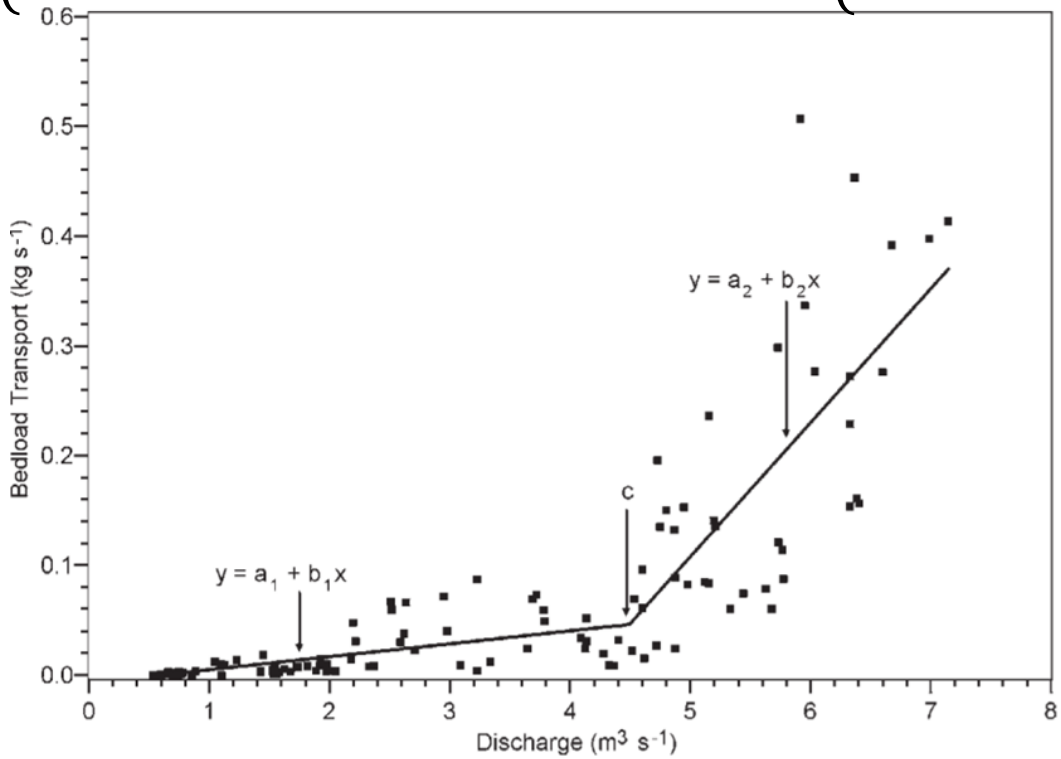
$$\Rightarrow \begin{cases} \beta_0 = \dfrac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4} = \bar{\mu} \\ \beta_1 = \mu_2 - \bar{\mu} \\ \beta_2 = \mu_3 - \bar{\mu} \\ \beta_3 = \mu_4 - \bar{\mu} \end{cases}$$

Piecewise Regression

- Sometimes we have reason to believe that different linear regression models apply in different region of data.
- Consider a simple case

$$y = \beta_0 + \beta_1 b_L(x) + \beta_2 b_R(x) + \varepsilon$$

$$b_L(x) = \begin{cases} c - x, & \text{if } x < c \\ 0, & \text{ow} \end{cases}, \qquad b_R(x) = \begin{cases} x - c, & \text{if } x > c \\ 0, & \text{ow} \end{cases}$$



Polynomial regression

- Polynomial regression: The relationship between response and predictors is smooth, but not a straight line.
  - 1 predictor with general degree d
    $$y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots \beta_d X^d + \varepsilon$$
  - 2 predictors with degree $d = 2$
    $$y = \beta_0 + (\beta_1 X_1 + \beta_2 X_2) + \left(\beta_{1,1} X_1^2 + \beta_{2,2} X_2^2\right) + \beta_{1,2} X_1 X_2 + \varepsilon$$
  - m predictors with degree $d = 2$
    $$y = \beta_0 + \left(\sum_{i=1}^{m} \beta_i X_i\right) + \left(\sum_{i=1}^{m} \beta_{i,i} X_i^2\right) + \left(\sum_{1 \le i \le j \le m} \beta_{i,j} X_i X_j\right) + \varepsilon$$
- Increasing degree $d \Rightarrow$ model may have too many parameters
  $$1 + m + m + \binom{m}{2} = \frac{(m+1)(m+2)}{2}$$
- Polynomial terms can cause numerical instability and collinearity. It is hard to fit jump function.

Inner Product Space

- An inner product space is a vector space $\mathcal{V}$ over the field $\mathcal{F}$ together with an inner product, that is, a map
$$\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \to \mathcal{F}$$
that satisfies the three properties $\forall\, x, y, z \in \mathcal{V}$ and $a, b \in \mathcal{F}$
  - Symmetry $\langle x, y \rangle = \langle y, x \rangle$
  - Linearity $\langle ax + by, z \rangle = a\langle x, z \rangle + b\langle y, z \rangle$
  - Positive definiteness $\langle x, x \rangle > 0$ if $x \neq 0$
- Every inner product space induces a norm, that is defined by
$$\|x\| = \sqrt{\langle x, x \rangle}$$
with this norm, every inner product space becomes a normed vector space. Every general property of normed vector spaces applies to inner product spaces.
- Ex. (Euclidean Vector Space)
$\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a inner product on $\mathbb{R}^n$ if and only if exist a symmetric positive-definite matrix M such that
$$\langle x, y \rangle = x^\top M y, \qquad \forall x, y \in \mathbb{R}^n$$
if M is identity matrix then $\langle \cdot, \cdot \rangle$ is the dot product.

Orthogonal Polynomial

- An orthogonal polynomial sequence $\{P_n\}_{n=0}^{\infty}$ is a family of polynomials such that

$$\langle P_n, P_m \rangle = 0, \qquad m \neq n$$

Usually the sequence is required to be orthonormal, namely

$$\langle P_n, P_n \rangle = 1$$

- It is common to define the inner product of functions f and g w.r.t a nonnegative weight function σ over an interval $[a, b]$

$$\langle f, g \rangle_\sigma = \int_a^b f(x)g(x)\sigma(x)dx\,, \qquad \|f\| = \sqrt{\langle f, f \rangle_\sigma}$$

Various polynomial sequences

  - Hermite polynomials $\Leftrightarrow \langle f, g \rangle_\sigma = \int_{-\infty}^{\infty} f(x)g(x)e^{-x^2}dx$
  - Legendre polynomials $\Leftrightarrow \langle f, g \rangle_\sigma = \int_{-1}^{1} f(x)g(x)\,dx$
  - Chebyshev polynomials $\Leftrightarrow \langle f, g \rangle_\sigma = \int_{-1}^{1} f(x)g(x)\frac{1}{\sqrt{1-x^2}}dx$
  - Laguerre polynomials $\Leftrightarrow \langle f, g \rangle_\sigma = \int_0^{\infty} f(x)g(x)e^{-x}\,dx$

Legendre Polynomials

- Let $\{B_n(x), n \in \mathbb{N}\}$ be a linearly independent sequence of $C([a,b])$. Then an orthogonal basis of functions, $\{P_n(x), n \in \mathbb{N}\}$ can be found and is given by

$$P_0(x) = B_0(x)$$

$$P_n(x) = B_n(x) - \sum_{j=0}^{n-1} \frac{\langle B_n, P_j \rangle_\sigma}{\langle P_j, P_j \rangle_\sigma} P_j(x), \qquad n = 1,2,\cdots$$

- Ex. Consider $B_n(x) = x^n, x \in (-1,1)$ and $\sigma(x) = 1$

$$P_0(x) = B_0(x) = 1$$

$$P_1(x) = B_1(x) - \frac{\langle B_1, P_0 \rangle_\sigma}{\langle P_0, P_0 \rangle_\sigma} P_0(x) = x - \frac{\langle x, 1 \rangle_\sigma}{2} 1 = x$$

$$P_2(x) = B_2(x) - \frac{\langle B_2, P_0 \rangle_\sigma}{\langle P_0, P_0 \rangle_\sigma} P_0(x) - \frac{\langle B_2, P_1 \rangle_\sigma}{\langle P_1, P_1 \rangle_\sigma} P_1(x)$$

$$= x^2 - \frac{\langle x^2, 1 \rangle_w}{2} 1 - \frac{\langle x^2, x \rangle_w}{2/3} x = x^2 - \frac{1}{3}$$

$$\Rightarrow \text{orthogonal set} \left\{1, x, x^2 - \frac{1}{3}\right\}$$

Relation to moments

- The orthogonal polynomials P_n can be expressed in terms of moments

$$\mu_n = \int_a^b x^n \sigma(x)\, dx$$

as follow

$$P_n(x) = c_n \det \begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_n \\ \mu_1 & \mu_2 & \cdots & \mu_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{n-1} & \mu_n & \cdots & \mu_{2n-1} \\ 1 & x & \cdots & x^n \end{bmatrix}$$

where the c_n are arbitrary (depend on the normalization of P_n).

- Ex. Consider the Legendre Polynomial

$$P_2(x) = c_n \det \begin{bmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ 1 & x & x^2 \end{bmatrix} = c_n \det \begin{bmatrix} 2 & 0 & 2/3 \\ 0 & 2/3 & 0 \\ 1 & x & x^2 \end{bmatrix}$$

$$= c_n \frac{4}{3}\left(x^2 - \frac{1}{3}\right)$$

Model Selection - 1

- Testing-Based Procedures
  - May miss optimal model because of its one-at-a-time adding/dropping.
  - α-values (α-to-enter and α-to-remove) should not be treated too literally: because multiple testing occurring.
  - Removal of less significant terms tend to increase the significance of the remaining terms ⇒ it may lead to overstate the importance of the remaining terms.
  - For prediction purpose, testing-based procedure tends to pick smaller models than desired.
- Criterion-Based Procedures
  - AIC and BIC

    $\text{AIC} = -2(\text{maximized loglikelihood}) + 2\text{p}$

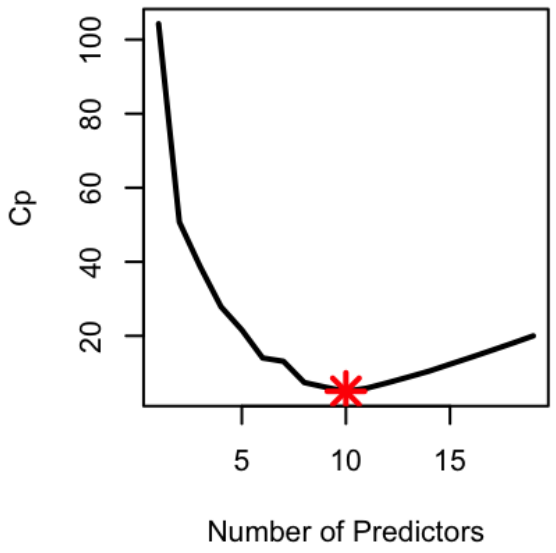
    $\text{BIC} = -2(\text{maximized loglikelihood}) + \log(n)\text{p}$

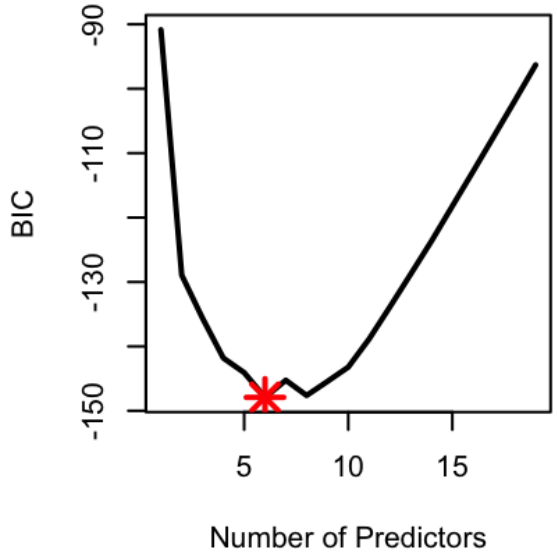
    For regression,

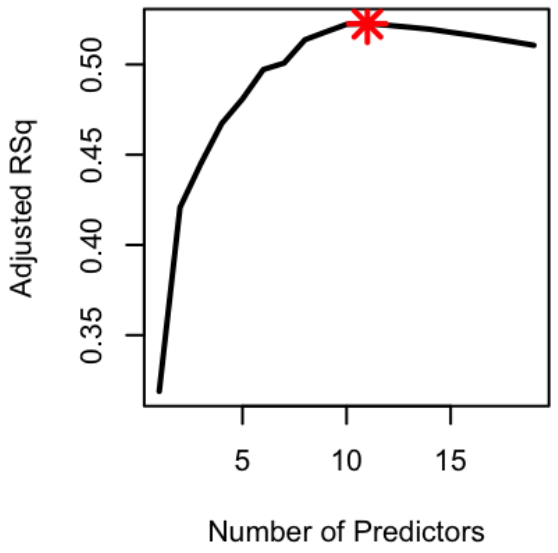
    $$-2(\text{maximized loglikelihood}) = n\log\left(\frac{\text{RSS}_{\mathcal{M}_p}}{n}\right) + \text{const}$$

Model Selection - 2

- Best Subset Selection:
  - Let $\mathcal{M}_0$ denote the which contains no predictor.
  - For $k = 1, \cdots, p$
    - Fit all $\binom{p}{k}$ models that contain exactly k predictors.
    - Pick the best among these $\binom{p}{k}$ models and call it $\mathcal{M}_k$.
  - Select a single best model from among $\mathcal{M}_0, \mathcal{M}_1, \cdots \mathcal{M}_k$.

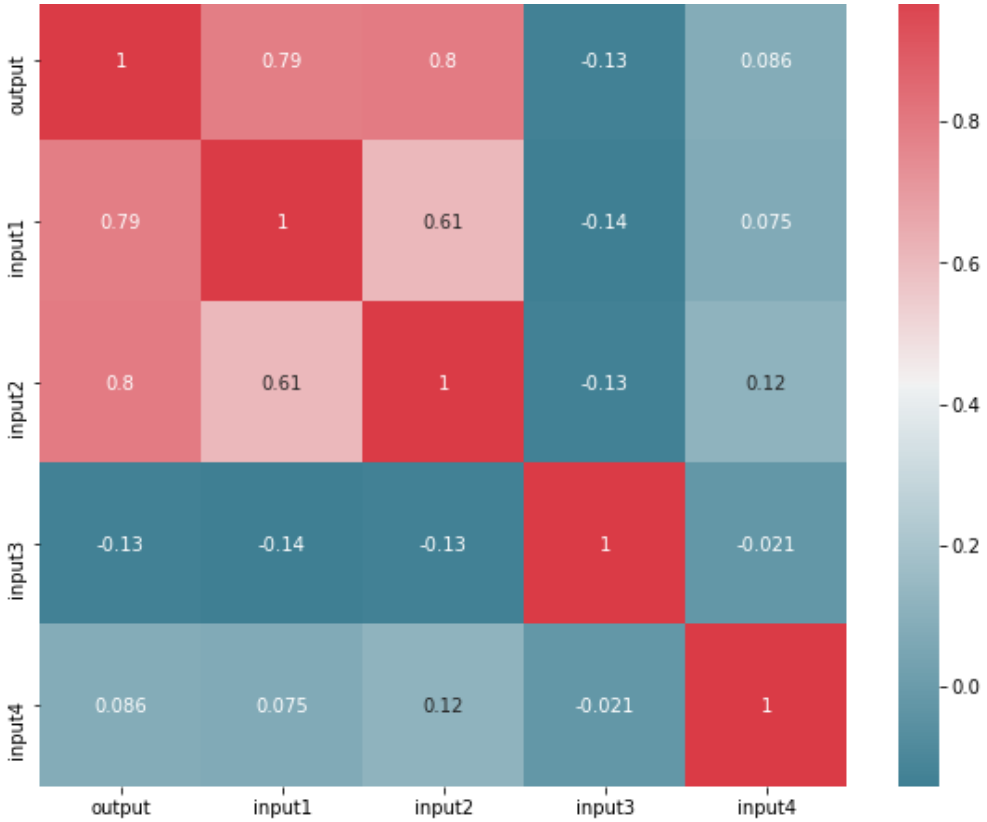






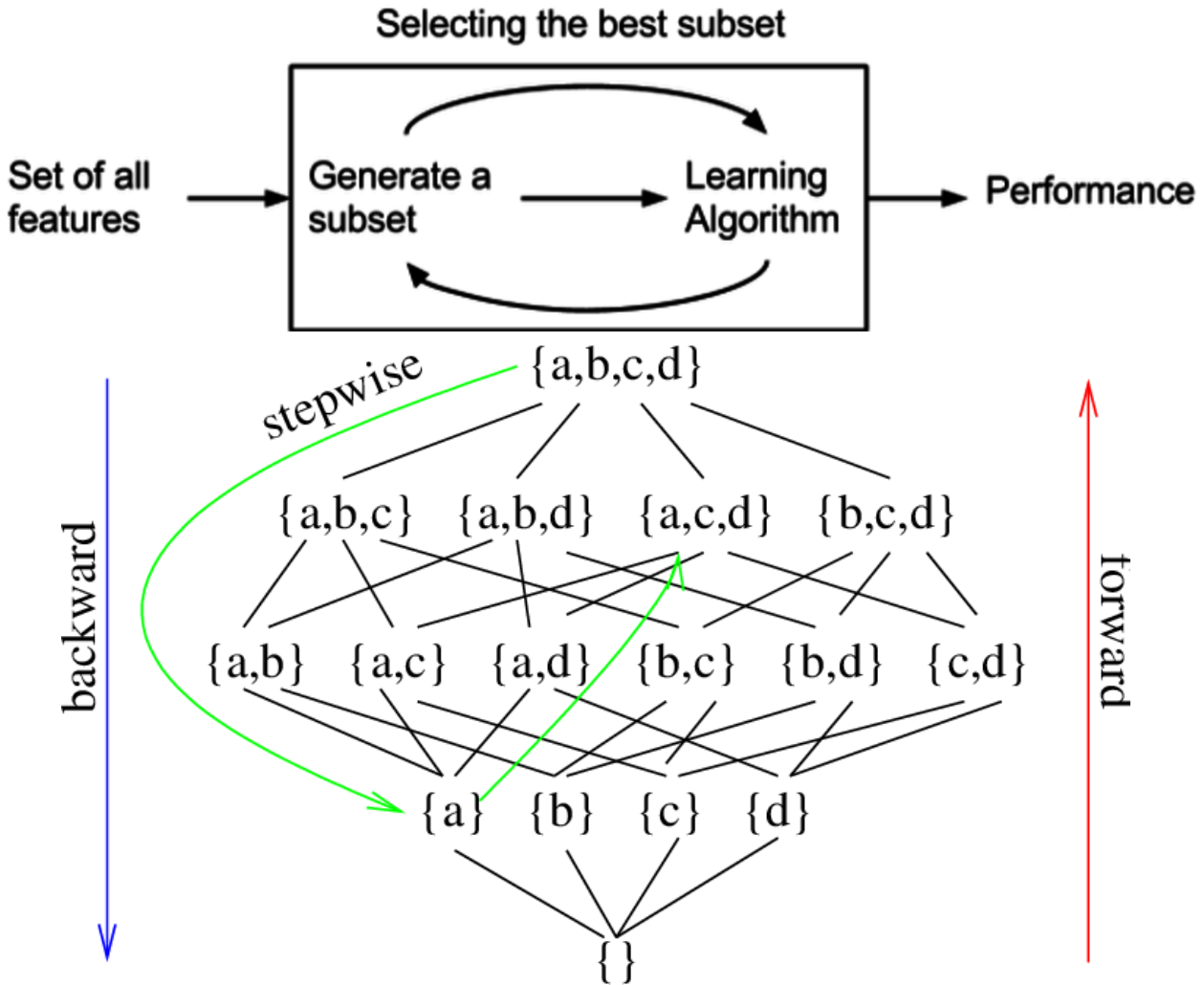
Model Selection - 3

- Three types of selection procedure
  - Filter Mode (Select Best K): Apply a measure to assign a scoring to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset.



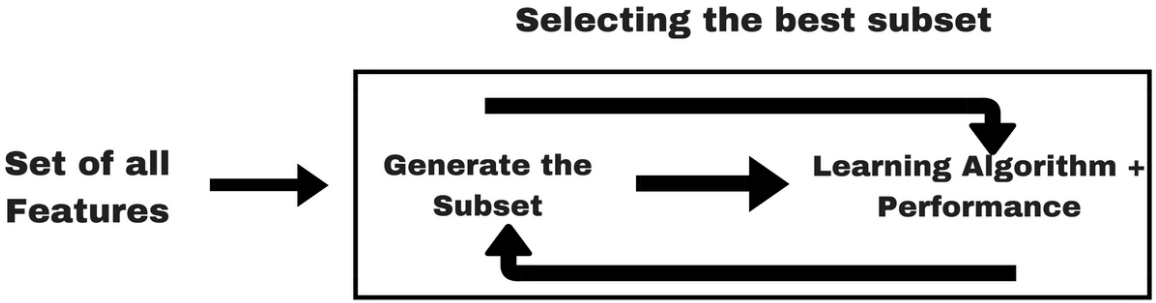
Model Selection - 4

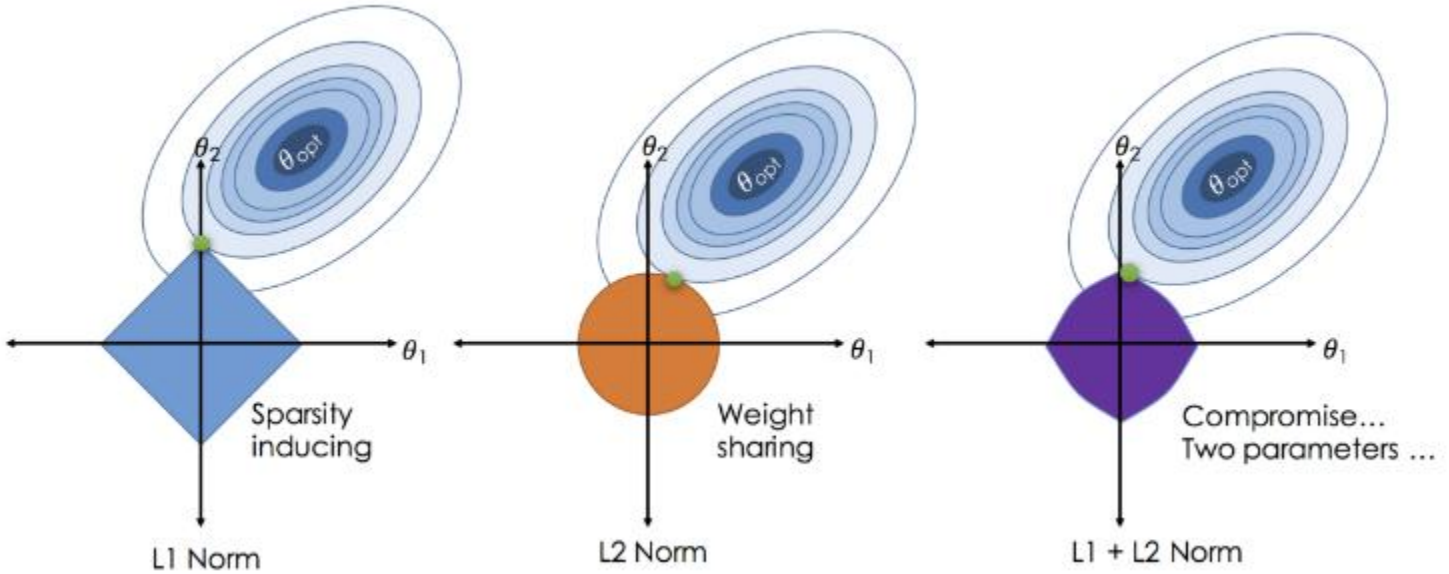
- Wrap Mode (Select from Model): Consider the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations.



Model Selection - 5

- Embedded Mode: A catch-all group of techniques which perform feature selection as part of the model construction process. Embedded methods combines the advantageous aspects of both Filter and Wrapper methods.





Shrinkage

- Choose the β that minimize
$$\mathcal{S}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\top}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|^2$$
for some choice of $\lambda > 0$.
- Ridge regression minimize $\mathcal{S}$ under ℓ^2 norm
$$\|\boldsymbol{\beta}\|_{\ell^2} = \sum_{i=1}^{p} \beta_i^2$$
Lasso regression minimize $\mathcal{S}$ under ℓ^1 norm
$$\|\boldsymbol{\beta}\|_{\ell^1} = \sum_{i=1}^{p} |\beta_i|$$
- One can show Ridge regression has analytic solution of $\boldsymbol{\beta}$
$$\widehat{\boldsymbol{\beta}}_{\ell^2} = (\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^{\top}\boldsymbol{y}$$
Lasso regression has no close form formula.
  - As $\lambda \searrow 0$, the Ridge and Lasso estimate become the OLS estimate
  - As $\lambda \nearrow \infty$, the Ridge and Lasso estimate shrink to 0.