Linear regression

Jitkomut Songsiri

Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University

CUEE

January 13, 2023

Outline

Multiple linear regression

Description of linear regression

- a linear relationship between variables y and x_k using a linear function:

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \triangleq x^T \beta$$

  where $y \in \mathbf{R}$, $x \in \mathbf{R}^n$, $\beta \in \mathbf{R}^n$
- y contains the measurement variables and is often called the *regressed/response/explained/dependent variable*
- x_k's are the input variables that explain the behavior of y; called the *predictor/explanatory/independent variables*
- β is the *regression coefficient*

Linear regression in matrix form

- given a data set: $\{(x_i, y_i)\}_{i=1}^m$ we can form a matrix equation

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \quad \triangleq \quad y = X\beta$$

- the matrix $X \in \mathbf{R}^{N \times n}$ is sometimes called *the design/regressor matrix*
- given y and X, one would like to estimate β that gives the linear model output match best with y
- in practice, in the presence of noise and disturbance, more data should be collected in order to get a better estimate – leading to *overdetermined* linear equations
- an exact solution to $y = X\beta$ does not usually exist; however, it can be solved by **linear least-squares** formulation

Problem statement

setting: y is linear in X but corrupted by some noise

$$y = X\beta + e, \quad X \in \mathbf{R}^{N\times n} \quad \text{with} \quad N > n$$

e is the error term

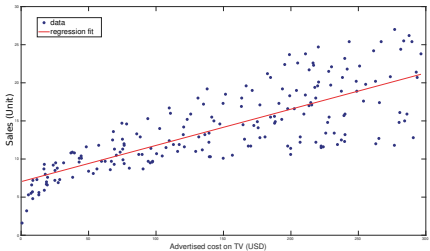
linear least-squares formulation:

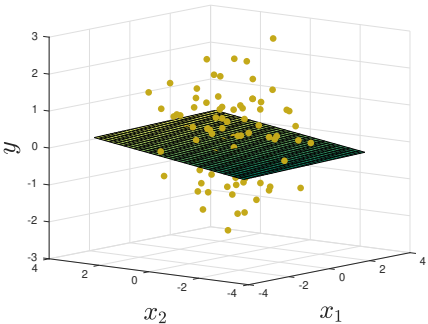
$$\underset{\beta}{\text{minimize}} \quad \|y - X\beta\|_2 = \left(\sum_{i=1}^{N}(\sum_{j=1}^{n} X_{ij}\beta_j - y_i)^2\right)^{1/2}$$

- $r = y - X\beta$ is called *the residual error*
- β with smallest residual norm $\|r\|$ is called *the least-squares solution*
- equivalent to minimizing $\|y - X\beta\|^2$

Fitting linear least-squares

left: explain the sale amount by advertising on TV





- left: sum squared distance of data points to the line is minimum (this line fits best)
- right: for two predictors, LS solution is the normal vector of hyperplane that lies closest to all data points of y

Example: data fitting

given data points $\{(t_i, y_i)\}_{i=1}^m$, we aim to approximate y using a function $g(t)$

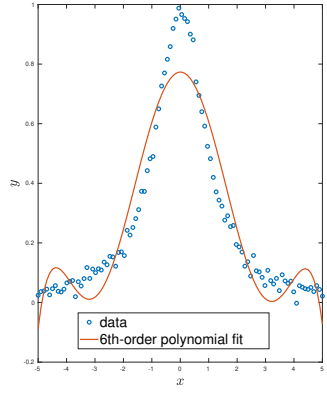
$$y = g(t) := \beta_1 g_1(t) + \beta_2 g_2(t) + \cdots + \beta_n g_n(t)$$

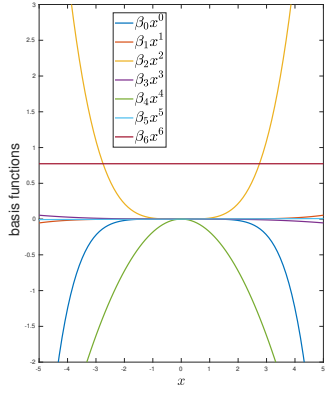
- $g_k(t) : \mathbf{R} \to \mathbf{R}$ is a basis function
  - polynomial functions: $1, t, t^2, \ldots, t^n$
  - sinusoidal functions: $\cos(\omega_k t), \sin(\omega_k t)$ for $k = 1, 2, \ldots, n$
- the linear regression model can be formulated as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} g_1(t_1) & g_2(t_1) & \cdots & g_n(t_1) \\ g_1(t_2) & g_2(t_2) & \cdots & g_n(t_2) \\ \vdots & & & \vdots \\ g_1(t_m) & g_2(t_m) & \cdots & g_n(t_m) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \quad \triangleq \quad y = X\beta$$

- often have $N \gg n$, *i.e.*, explaining y using a few parameters in the model

fitting a 6th-order polynomial to data points generated from $f(t) = 1/(1+t^2)$





- (right) the weighted sum of basis functions (x^k) is the fitted polynomial
- the ground-truth function f is nonlinear, but can be decomposed as a sum of polynomials

Closed-form solution to LS problem

setting the gradient of $\|y - X\beta\|_2^2$ gives

$$\textbf{normal equation:} \quad X^T X\beta = X^T y$$

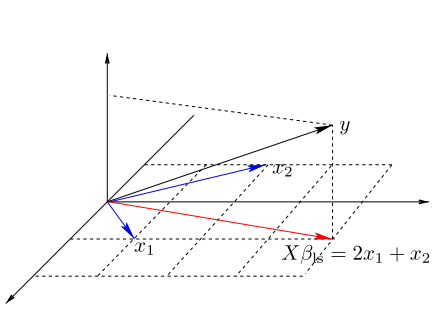
if $X \in \mathbf{R}^{N\times n}$ with $N \geq n$ is full rank, then

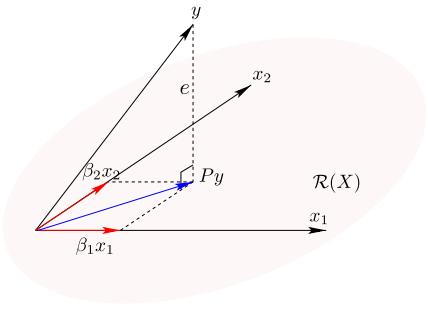
- least-squares solution can be found by solving the normal equations
- n equations in n variables with a positive definite coefficient matrix
- the closed-form solution is $\beta = (X^T X)^{-1} X^T y$ and unique
- $(X^T X)^{-1} X^T$ is a left inverse of X

note: $\mathbf{rank}(X) = n \Rightarrow \mathcal{N}(X) = \{0\} \Rightarrow X^T X \succ 0$ (hence, $X^T X$ is invertible)

in MATLAB, a LS solution is solved by `X\y`

Geometric interpretation of a LS problem

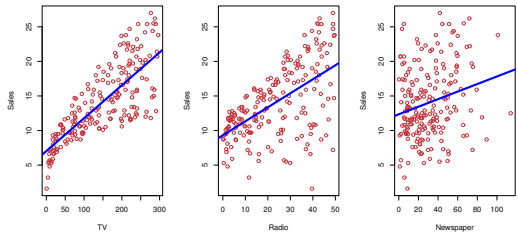




- $\|y - X\beta\|_2$ is the distance from y to $X\beta = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$
- solution β_{ls} gives the linear combination of the columns of X closest to y
- $\hat{y} = X\beta_{\text{ls}}$ is the **orthogonal projection** of y to the range of X
- $P = X(X^T X)^{-1} X^T$ is an **orthogonal projection** matrix (aka hat matrix)

Interpreting regression coefficients

advertising data: sale is explained by advertising costs in TV, radio and newspaper



given $\hat{\beta}$, a predicted output is $\hat{y} = X\hat{\beta} = x_1\hat{\beta}_1 + \cdots + x_n\hat{\beta}_n$

- β_j is the average effect on y of a one unit increase in x_j, holding all other predictors fixed
- in real data, predictors can have correlation (when x_1 changes then x_2 cannot be assumed constant)

Properties of LS estimate

Analysis of the LS estimate

assumptions:

- data generating process is $y = X\beta + u$
- u is *white noise* with zero mean and covariance matrix Σ
- the least-square estimate is given by $\hat{\beta} = \mathrm{argmin}_\beta \|X\beta - y\|_2$
- the regressor X is *deterministic*

then the following properties hold:

- $\hat{\beta}$ is an unbiased estimate of β ($\mathbf{E}\hat{\beta} = \beta$, or $\hat{\beta} = \beta$ when $u = 0$)
- the covariance matrix of $\hat{\beta}$ is given by

$$\mathbf{cov}(\hat{\beta}) = (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1}$$

short proof: we can write the LS estimate as

$$\hat{\beta} = (X^T X)^{-1} X^T y = (X^T X)^{-1} X^T (X\beta + u) = \beta + (X^T X)^{-1} X^T u$$

- since X is deterministic and u is zero mean, we have $\mathbf{E}\hat{\beta} = \beta$
- the covariance of $\hat{\beta}$ is derived by

$$\mathbf{cov}(\hat{\beta}) = \mathbf{E}[(\hat{\beta} - \mathbf{E}\hat{\beta})(\hat{\beta} - \mathbf{E}\hat{\beta})^T]$$

but $\mathbf{E}\hat{\beta} = \beta$ and that $\hat{\beta} - \mathbf{E}\hat{\beta} = (X^T X)^{-1} X^T u$, hence,

$$\begin{aligned} \mathbf{cov}(\hat{\beta}) &= \mathbf{cov}[(X^T X)^{-1} X^T u] \\ &= (X^T X)^{-1} X^T \, \mathbf{cov}(u) X (X^T X)^{-1} \\ &= (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1} \end{aligned}$$

if $\Sigma = \sigma^2 I$, then it reduces to $\mathbf{cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$

BLUE property

assumptions: u is white noise with zero mean and **unit** covariance ($\mathbf{cov}(u) = I$)

the estimator defined by

$$\hat{\beta}_{\text{ls}} = (X^T X)^{-1} X^T y$$

is the **optimum unbiased linear least-mean-squares** estimator of β

assume $\hat{\beta} = By$ is any other linear estimator of β

- require $BX = I$ in order for $\hat{z}$ to be unbiased
- $\mathbf{cov}(\hat{\beta}) = BB^T$
- $\mathbf{cov}(\hat{\beta}_{\text{ls}}) = BX(X^T X)^{-1} X^T B^T$ (apply $BX = I$)

Using $I - X(X^T X)^{-1} X^T \succeq 0$, we conclude that

$$\mathbf{cov}(\hat{\beta}) - \mathbf{cov}(\hat{\beta}_{\text{ls}}) = B(I - X(X^T X)^{-1} X^T) B^T \succeq 0$$

- BLUE property is also known as **Gauss-Markov theorem**
- the assumption that $\mathbf{cov}(u) = I$ (or could be $\sigma^2 I$) is equivalent to
  - $\mathbf{var}(u_i) = \sigma^2$ for all i, *i.e.*, the error terms have the same variance (**homoskedasticity**)
  - $\mathbf{cov}(u_i, u_j) = 0$ for $i \neq j$, *i.e.*, the error terms are uncorrelated
- the proof on the optimality use the fact that $P = X(X^T X)^{-1} X^T$ is an **orthogonal projection** matrix which have properties:
  - $P^T = P$
  - $P^2 = P$
  - $\|Px\| \leq \|x\|$ for all $x \in \mathbf{R}^n$

  these properties imply that $I - P \succeq 0$

Properties of estimation errors

under the homoskedastic assumption $u_i \sim \mathcal{N}(0, \sigma^2)$ and define

$$\hat{u} = y - X\hat{\beta}_{\text{ls}}, \quad \text{RSS} = \sum_{i=1}^{N} \hat{u}_i^2, \quad s^2 = \text{RSS}/(N-n) = \|\hat{u}\|_2^2/(N-n)$$

- s^2 is an unbiased estimate for σ^2
- $(N-n)s^2/\sigma^2 \sim \chi^2(N-n)$ (require Gaussian assumption of u_i)

an estimate of covariance and standard error of $\hat{\beta}$ are

$$\mathbf{cov}(\hat{\beta}) = s^2(X^TX)^{-1}, \quad \text{SE}(\hat{\beta}_k) = \sqrt{\mathbf{cov}(\hat{\beta})_{kk}}$$

- using more samples gives smaller $\mathbf{cov}(\hat{\beta})$
- if predictors are highly correlated, the covariance is big

Accuracy of the model

R^2 is based on the decomposition of the total sum of squares (TSS)

$$\underbrace{\sum_i (y_i - \bar{y})^2}_{\text{TSS}} = \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{\text{RSS}} + \underbrace{\sum_i (\hat{y}_i - \bar{y})^2}_{\text{ESS}} + 2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

TSS (total), RSS (residual) and ESS (explained) sum of squares

for OLS, the last term on RHS is zero if the model has a constant term, so

$$\text{TSS} = \text{RSS} + \text{ESS}$$

R^2 is defined as

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

R^2 is between 0 and 1 and it measures the proportion of variability in Y that can be explained using X

Variable selection

Hypothesis testing

a significance test of regression coefficients involves

$$H_0 : \beta_k = 0 \quad \text{versus} \quad H_1 : \beta_k \neq 0$$

we compute a t-statistic given by

$$T = \frac{\hat{\beta}_k}{\mathsf{SE}(\hat{\beta}_k)} = \frac{\hat{\beta}_k}{\sqrt{s^2[(X^T X)^{-1}]_{kk}}} \quad \sim t_{N-n}$$

and compute the probability of observing any value equal to $|T|$ or larger

$$p\text{-value} \quad = P(t_{N-n} \geq |T|)$$

if p-value $< \alpha$ (a given significance level) then we reject H_0 (x_k is significant)

Results on advertising data

run simple regression versus multiple regression

Simple regression of `sales` on `radio`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	9.312	0.563	16.54	< 0.0001
`radio`	0.203	0.020	9.92	< 0.0001

Simple regression of `sales` on `newspaper`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	12.351	0.621	19.88	< 0.0001
`newspaper`	0.055	0.017	3.30	0.00115

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`radio`	0.189	0.0086	21.89	< 0.0001
`newspaper`	-0.001	0.0059	-0.18	0.8599

- top left: a 1,000 USD increase in radio ad budget is associated with an average increase in sales by around 203 units
- top right: a 1,000 USD increase in newspaper budget is associated with an average increase in sales by around 55 units
- bottom: the coefficient of newspaper is by contrast close to zero in multiple regression — p-value is high and newspaper is not significant

when examining the correlation matrix of predictors and response

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

- note that the correlation between radio ad and newspaper is 0.35
- markets with high newspaper ad tend to also have high radio ad
- multiple regression shows that newspaper ad is not directly associated with sales
- however, when running a simple regression, newspaper is a surrogate for radio ad and get credit for explaining sales

Deciding important predictors

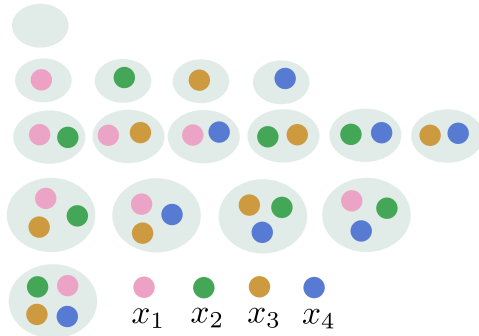
which predictors should be used to explain the response ?

common methods in variable selection:

- best subset selection: consider all possible model candidates
- forward selection: searching starts from a null model
- backward selection: searching starts from a dense model
- shrinkage method (regularization techniques)

Best subset selection

consider $x_1, x_2, \ldots, x_p$ as p predictors



S_k: the model class that each contains k predictors (S_0 has only constant term)
there are $\binom{p}{k}$ sub-models in S_k and no. of all possible sub-models is $\sum_{k=1}^{p} \binom{p}{k} = 2^p$

we would like to pick the 'best' model according to some model selection criterion

steps in variable selection

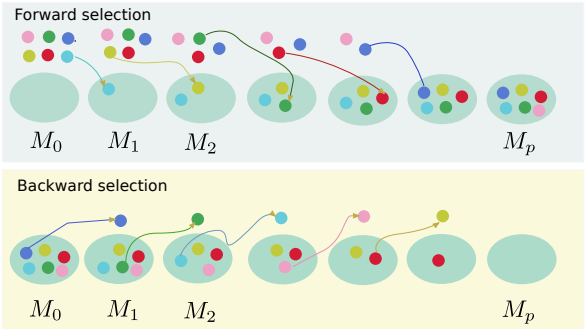
1. for $k = 1, \ldots, p$
2. for $j = 1, \ldots, \binom{p}{k}$
   1. fit all 'p choose k' sub-models that contain k predictors
   2. pick the best among $\binom{p}{k}$ models and call it M_j
   3. here 'best' is defined as having the smallest RSS on training data
3. select a single best model among $M_0, M_1, \ldots, M_p$ using *cross-validated* prediction error, AIC, BIC or adjusted R^2

step 3 is one of the two approaches to obtain the best model having *a low test error*

- *indirectly* estimate test error by *adjusting* training error to account for bias due to overfitting (here, using model selection score instead)
- *directly* estimate the test error, using a validation set/CV approach

Stepwise selection

when p is large, the best subset selection suffers from looking in a large search space



- stepwise selection explores over a a more *restricted* set of models
- forward selection starts from a null model, while backward selection starts from a full model

Forward stepwise selection

we start to consider the **null** model and add more predictors one at a time

1. let M_0 be the *null model* which contains no predictors
2. for $k = 0, \ldots, p$
   1. consider all $p - k$ models that augment the predictors in M_k with **one** additional predictor
   2. choose the best among these $p - k$ models and call it M_{k+1}
   3. the best model here is to have the smallest RSS or largest R^2
3. select a single best model among $M_0, M_1, \ldots, M_p$ using cross-validated AIC, BIC or adjusted R^2

- there are $\sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$ models involved in this algorithm (much less than 2^p)
- it may fail to find the best possible model out of all 2^p models

Backward stepwise selection

we start to consider the **full** model and remove more predictors one at a time

1. let M_0 be the *full model* which contains all p predictors
2. for $k = p, p-1 \ldots, 1$
   1. consider all k models that contain all but one of the predictors in M_k, for a total of $k-1$ predictors
   2. choose the best among these k models and call it M_{k-1}
   3. the best model here is to have the smallest RSS or largest R^2
3. select a single best model among $M_0, M_1, \ldots, M_p$ using cross-validated AIC, BIC or adjusted R^2

- there are $\sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$ models involved in this algorithm (much less than 2^p)
- it may fail to find the best possible model out of all 2^p models

Softwares and practical issues

Softwares

MATLAB: Statistical and machine learning toolbox

- `fitlm`: linear regression fit
- `stepwiselm`: stepwise regression (users can select criterion to add/remove terms)

Python modules:

- `statmodels`
- `scikit-learn`: linear model

Practical issues

- colinearity: two or more predictors are closely related
- correlation of error terms: error is not likely white
- non-constant variance of error terms: violate the homoskedastic assumption
- outliers: some data points are far from others

References

Figures and examples are taken from the first two references (ISLR, ESL)

1. Chapter 3 in T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, Second edition, 2009
2. Chapter 3 in G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Application in R*, Springer, 2013
3. Appendix in W.H. Greene, *Econometric Analysis*, Prentice Hall, 2008