應用機器學習

Brian Chan 陳醒凡

課程目標

1. 了解基本的數據分析
2. 了解基本的機器學習(Machine Learning)方法
3. 掌握Python的基本操作和一些有用的package
4. 處理及從網上下載數據
5. 在Python上應用機器學習

今天課堂概要

Regression methods

1. Basic set-up
2. Properties of estimators
3. Assumptions of classic regression
4. Spurious regression
5. Regularization
6. Logistic regression (or do it in next class)

BASIC SET-UP

Regression analysis is a statistical technique used to describe relationships among variables.

The simplest case to examine is one in which a variable Y, referred to as the dependent or target variable, may be related to one variable X, called an independent or explanatory variable, or simply a regressor.

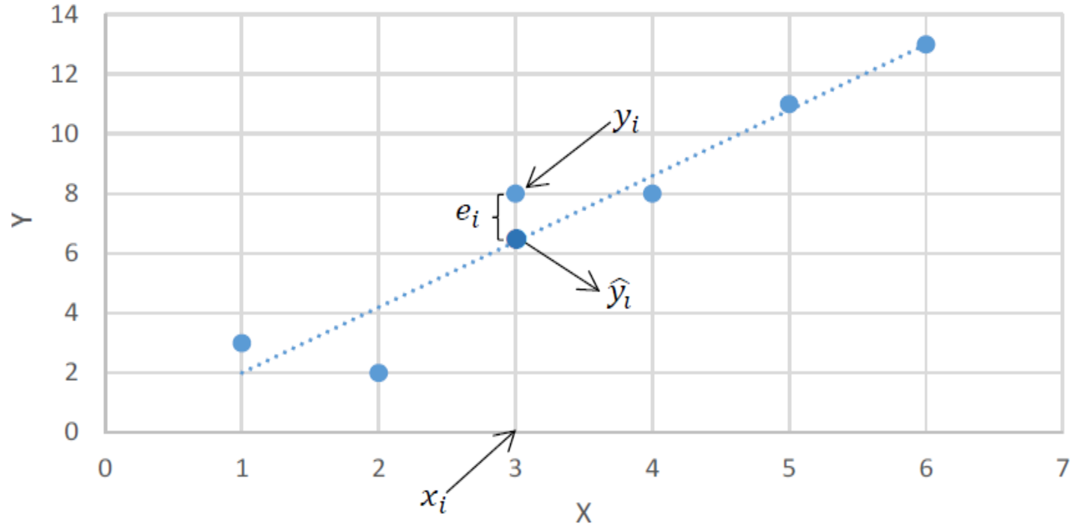
If the relationship between Y and X is believed to be linear, then the equation for a line may be appropriate: $Y = \beta_0 + \beta_1 X + \epsilon$,where β_0 is an intercept term and β_1 is a slope coefficient. ϵ be the residual noise.

https://en.wikipedia.org/wiki/Regression_analysis

BASIC SET-UP

Consider the pairs (y_i, x_i). Let $\hat{y}_i$ be the "predicted" value of y_i associated with x_i if the fitted line is used.

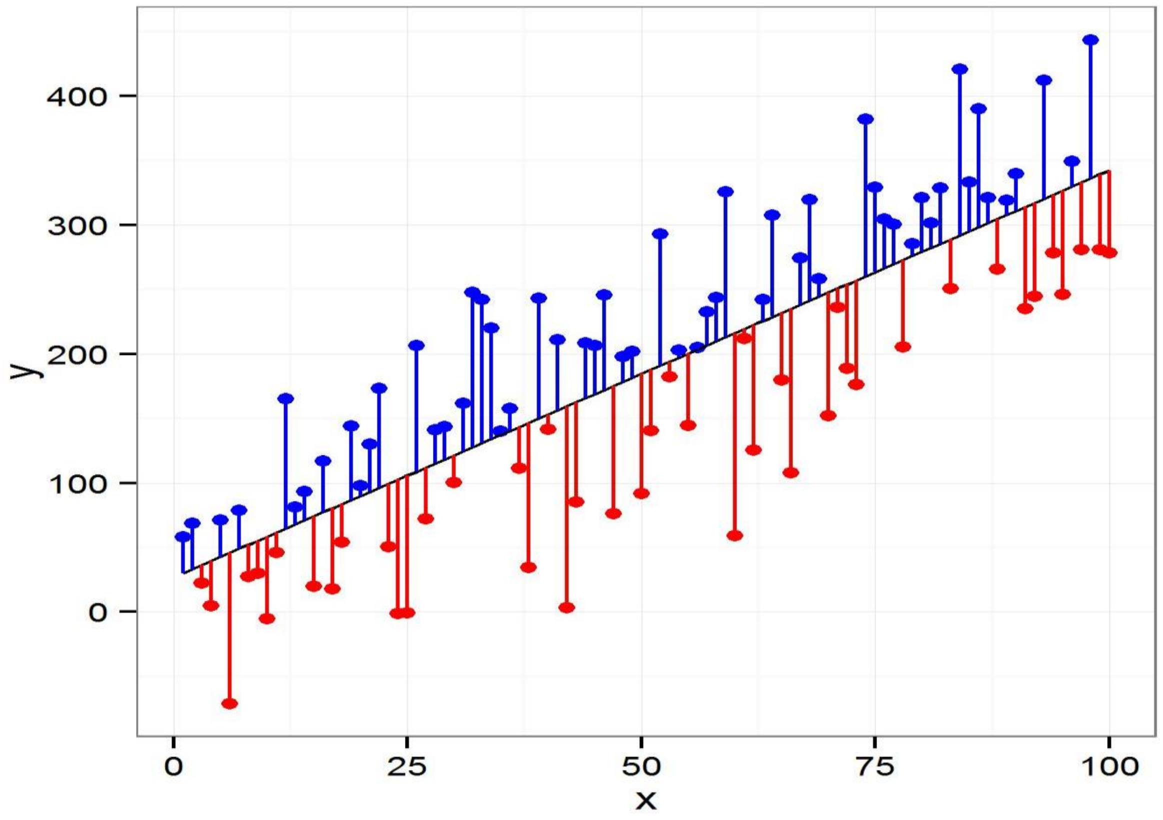
Define $\hat{e}_i = y_i - \hat{y}_i$ as the residual representina the "error" involved.



http://www.csie.ntnu.edu.tw/~u91029/Regression.html

https://stackoverflow.com/questions/47344850/scatterplot3d-regression-plane-with-residuals

Minimize the sum of the squared errors, i.e., $\sum \hat{e}_i^2 = \sum (y_i - \hat{y}_i)^2$



Minimize the sum of the squared errors, i.e., $\sum \hat{e}_i^2 = \sum (y_i - \hat{y}_i)^2$

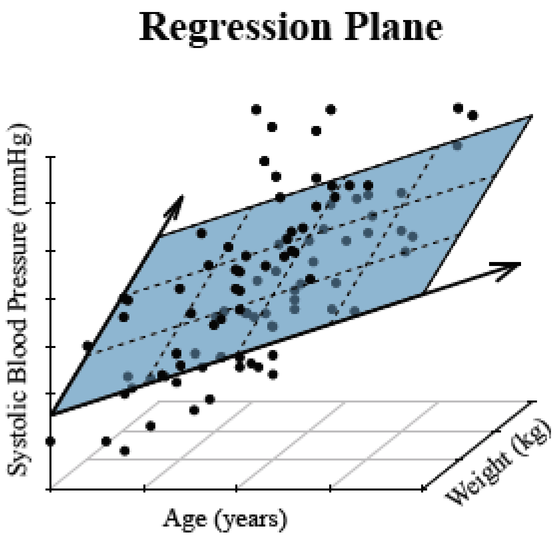


Figure 2.25: Systolic blood pressure linearly increases with age, but also with bodyweight. A line in two directions forms a plane.

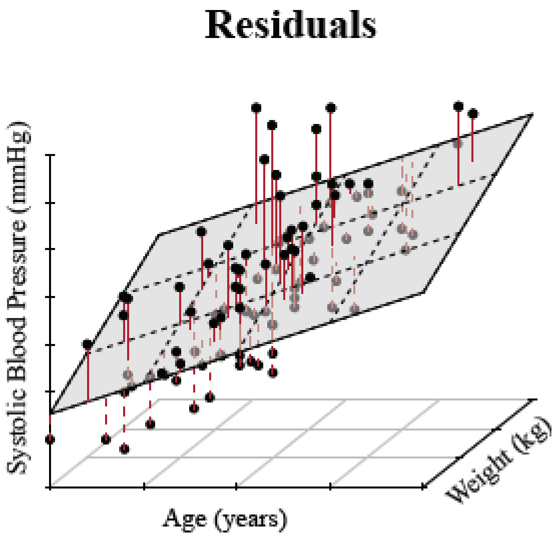


Figure 2.26: The residuals of figure 2.25 are the vertical distances to the plane. Negative residuals are indicated by dashed linepieces.

PROPERTIES OF ESTIMATORS

Closed-form solution of $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})\,(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad \textbf{(1)}$$

Properties of the estimators:

$$E[\hat{\beta}_0] = \beta_0 \quad \text{and} \quad E[\hat{\beta}_1] = \beta_1 \qquad \textbf{(2)}$$

$$var[\hat{\beta}_0] = \sigma^2_{\beta_1} \frac{\sigma^2_x}{n} \quad \text{and} \quad var[\hat{\beta}_1] = \frac{\sigma^2_\epsilon}{n\sigma^2_x} \quad \textbf{(3)}$$

http://web.hku.hk/~pingyu/0701/Ch03_MLR%20Estimation_ver1.pdf

STATISTICS OF REGRESSION

Key statistics:

R square (adjusted-R)

p-value (t-statistics)

* Concept of hypothesis test in statistics

https://en.wikipedia.org/wiki/Coefficient_of_determination

The Least Squares Approach

Output 1.2: SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.815274956
R Square	0.664673254
Adjusted R Square	0.661251553
Standard Error	27270.25391
Observations	100

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.44459E+11	1.44459E+11	194.252262	5.49151E-25
Residual	98	72879341312	743666748.1		
Total	99	2.17338E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	- 50034.6065	7422.677496	-6.740776032	1.0951E-09	-64764.6684	-35304.544
X Variable 1	72.8203802	5.22480275	13.93744102	5.4915E-25	62.45192918	83.1888312

ASSUMPTION OF CLASSIC REGRESSION

Some of classical assumptions for regression analysis include:

1. The independent variables are measured with no error.

2. The independent variables (predictors) are linearly independent, i.e. it is not possible to express any predictor as a linear combination of the others.

3. The errors are uncorrelated, that is, the variance–covariance matrix of the errors is diagonal and each non-zero element is the variance of the error.

4. The variance of the error is constant across observations (homoscedasticity).

Assumption of OLS regression

TRUE MODEL AND ESTIMATED VALUE

True model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

β_0 and β_1 are unknown parameters.

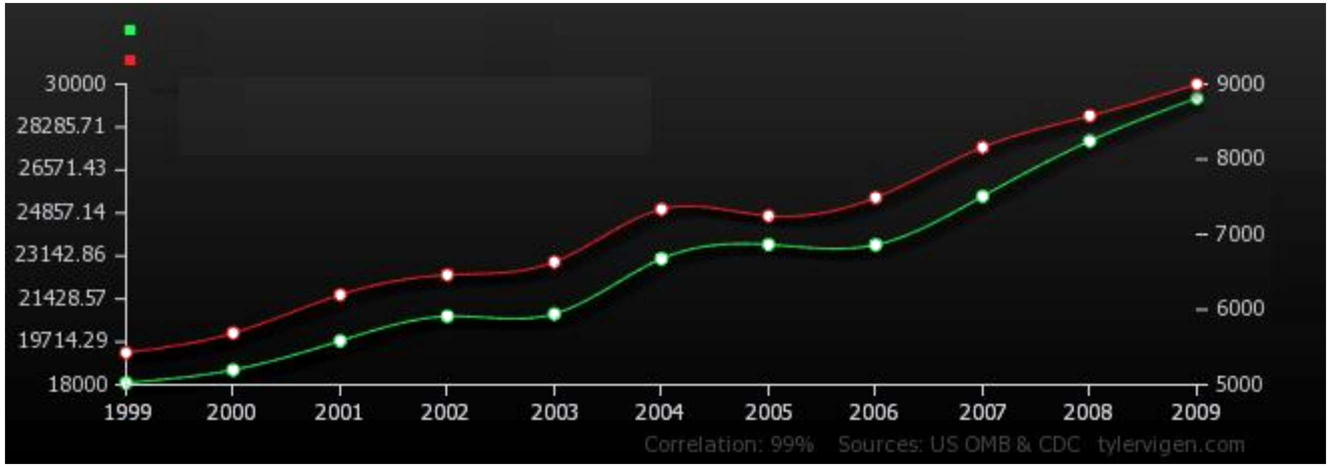
Estimated model: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated parameters which can be calculated by the closed-form solution (1) or gradient-descent method.

Experiment: simulate the model for β_0 and β_1. Then estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ by regression.

demo_lasso_by_simulation.py

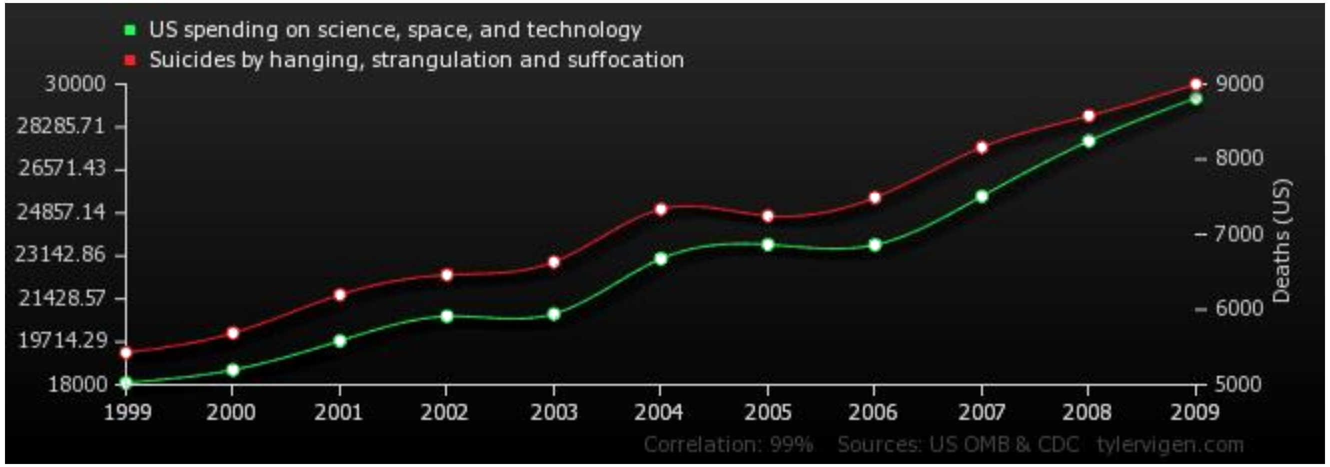
EXAMPLE: REGRESSION



main_spurious.py

https://www.tylervigen.com/spurious-correlations

EXAMPLE: SPURIOUS REGRESSION



https://www.tylervigen.com/spurious-correlations

SPURIOUS REGRESSION

例一:統計研究發現，冰淇淋銷量最高的時候，就是公共泳池的溺水事故發生得最多的時候。

例二:荷蘭的統計數字顯示，在一連串的春季中，鸛鳥巢的數目與人類嬰兒出生數目之間呈現正相關。

例三:高度民主、注重法治的國家大多富裕繁榮，可見制度對經濟有決定性的影響。
性。
例四:「夏以妺喜，殷以妲己，周以褒姒，三代所由亡也」（晉·杜預《左傳》注）

https://zh.wikipedia.org/wiki/%E5%81%BD%E9%97%9C%E4%BF%82

例一:統計研究發現，冰淇淋銷量最高的時候，就是公共泳池的溺水事故發生得最多的時候。
然而，有可能熱浪造成冰淇淋銷量和公共泳池的溺水事故增多。若視冰淇淋的銷量或遇溺事故為對方的成因，可能就被偽關係誤導了。

例二:荷蘭的統計數字顯示，在一連串的春季中，鸛鳥巢的數目與人類嬰兒出生數目之間呈現正相關。
兩者之間未必有因果關係。事實上，它們都和數据觀測之前9個月的天氣相關。

例三:高度民主、注重法治的國家大多富裕繁榮，可見制度對經濟有決定性的影響。
然而，有可能是其他的因素同時導致了民主、法治和富裕，像是根植文化的工作倫理或民族性。

例四:「夏以妹喜，殷以妲己，周以褒姒，三代所由亡也」（晉·杜預《左傳》注）
然而，有可能朝代滅亡和寵幸美女是因為別的因素，如君王本身的性格傾向等，所造成的。若將美女的出現與朝代的滅亡視為對方的成因，可能就被偽關係誤導了。

https://zh.wikipedia.org/wiki/%E5%81%BD%E9%97%9C%E4%BF%82

https://www.reed.edu/economics/parker/312/tschapters/S13_Ch_2.pdf

MULTIVARIATE VERSION

True model: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_n x_{n,i} + \epsilon_i$

β_i's are unknown parameters.

Estimated model: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \cdots + \hat{\beta}_n x_{n,i}$

$\hat{\beta}_i$'s are estimated parameters which can be calculated by the closed-form solution (1) or gradient-descent method.

REGULARIZATION

Regularization

This is a form of regression, that constrains/ regularizes or shrinks the coefficient estimates towards zero. In other words, *this technique discourages learning a more complex or flexible model, so as to* ***avoid the risk of overfitting*****.**

A simple relation for linear regression looks like this. Here Y represents the learned relation and β *represents the coefficient estimates for different variables or predictors(X).*

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|.$$

https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a

EXAMPLE:

Quadratic form:

$$f(x) = (x - \mu)^T \Sigma (x - \mu)$$

Quadratic form with L_1:

$$f(x) = (x - \mu)^T \Sigma (x - \mu) + \lambda |x|_1,$$

where $\lambda > 0$.

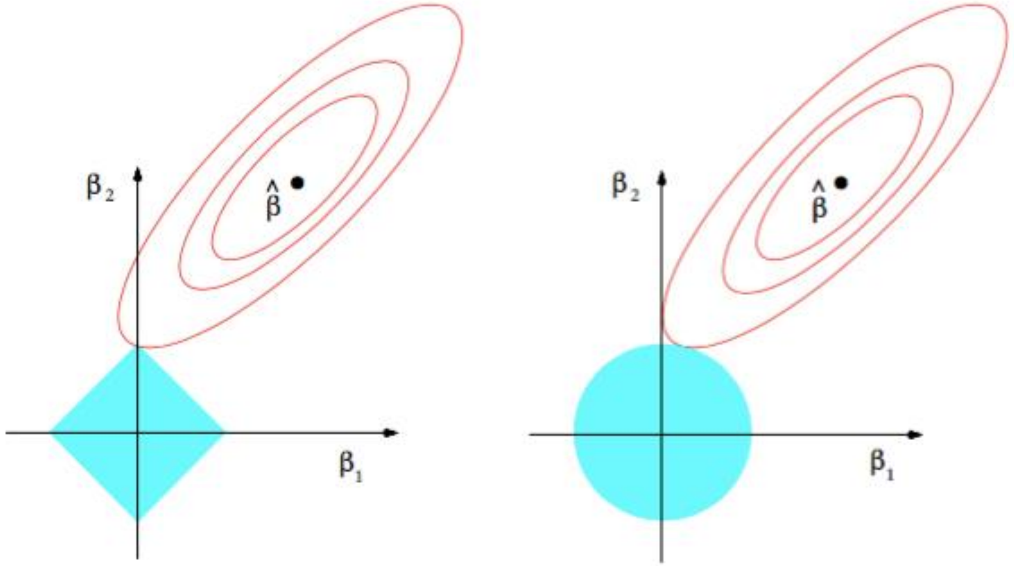
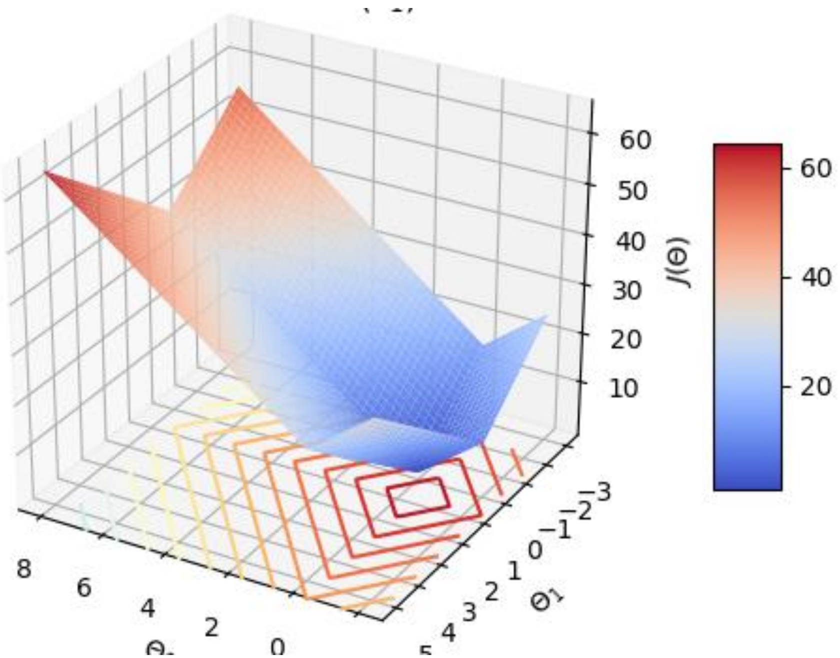
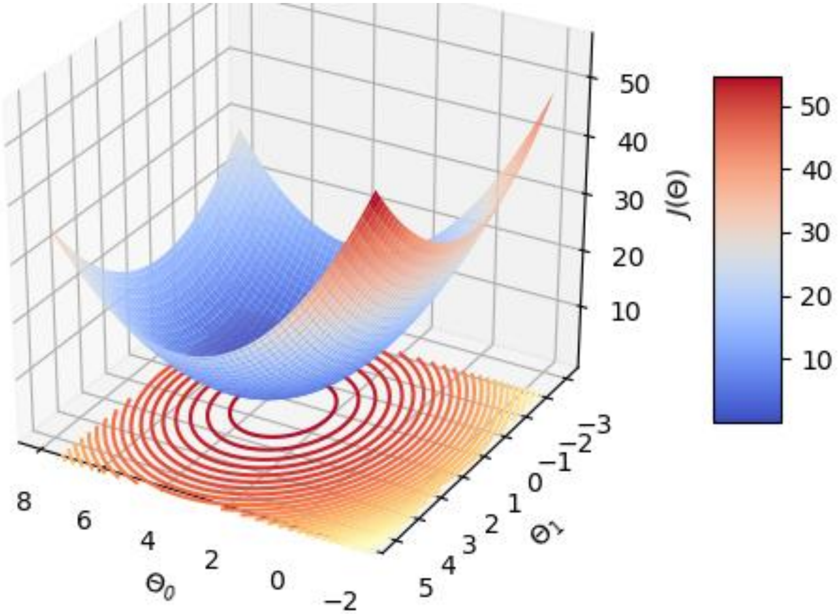
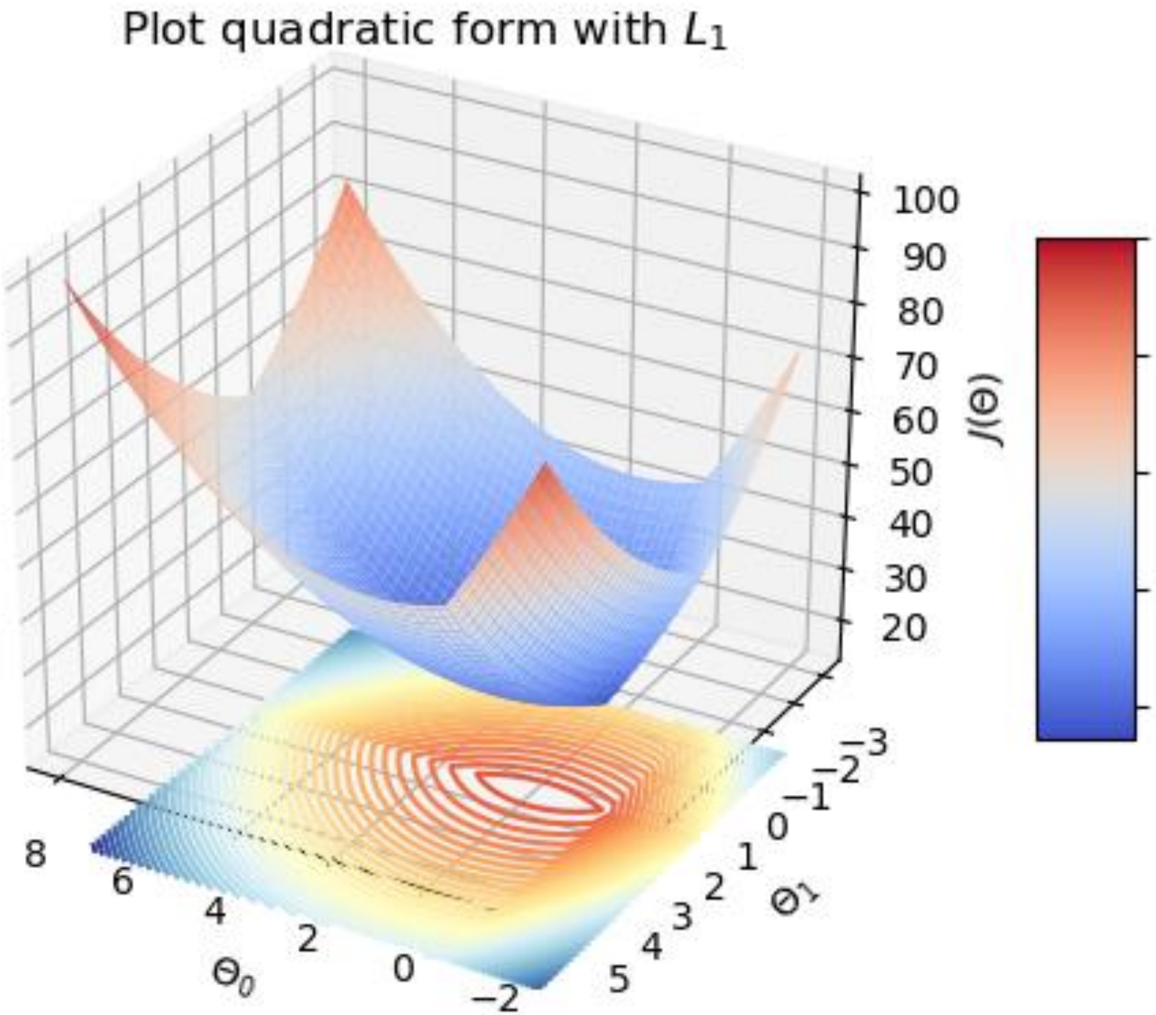


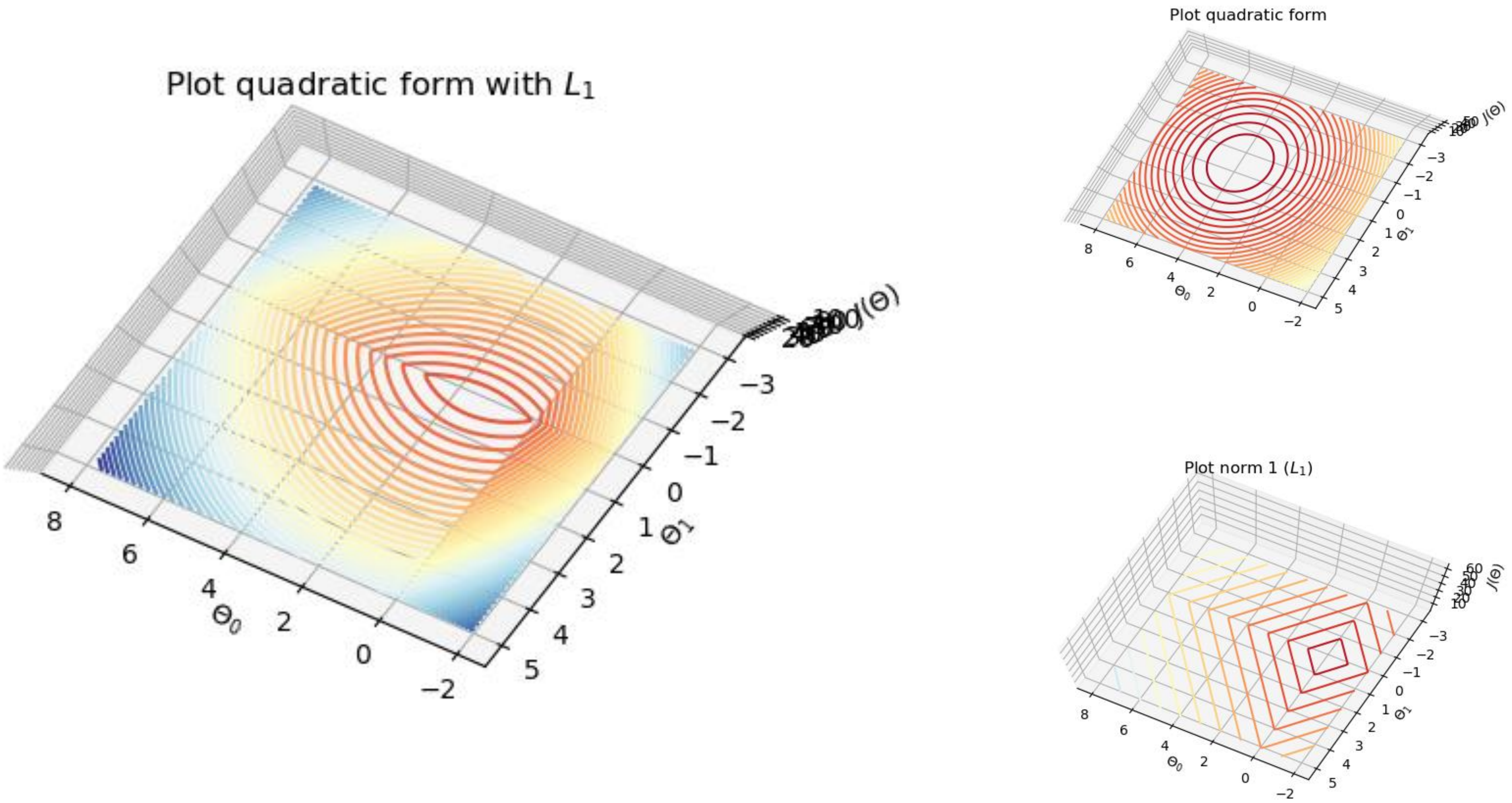
FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

from *The elements of statistical learning by Hastie, Tibshirani and Friedman.*


J(Θ)
Θ1


J(Θ)
Θ0
Θ1


Plot quadratic form with L_1
J(Θ)
Θ0
Θ1



Plot quadratic form with L_1
Θ_0
Θ_1
$J(\Theta)$
Plot quadratic form
Θ_0
Θ_1
$J(\Theta)$
Plot norm 1 (L_1)
Θ_0
Θ_1
$J(\Theta)$

SOME REMARKS:

Dummy variables

Variable selection

Check assumptions of regression

Data snooping

Transformation of model (e.g. log-log, log-linear)

Hypothesis tests

Taylor's expansion

http://web.hku.hk/~pingyu/0701/Ch10_Basic%20Time%20Series_ver1.pdf

ASSAULT RIFLES & CARBINES

CODE

main_spurious.py (US spending and suicide number)

demo_lasso_by_simulation.py (demo the manifold of lasso regression)

main_boston.py (case study)

Code\Ridge-Regression-master\main.py (lasso for the case)

下一課...

分類法:

1. 非線性迴歸 (Nonlinear regression)
2. 支援向量機 (Support vector machine)
3. 線性回歸的應用例子 (續上課堂四)